Agnieszka Hunstiger und
Uwe Koreik (Hg.)

Chance Deutsch

Schule – Studium - Arbeitswelt

34. Jahrestagung des Fachverbandes
Deutsch als Fremdsprache 2006
an der Leibniz Universität
Hannover

Materialien
Deutsch als Fremdsprache
Band 78

Universitätsverlag Göttingen
2009

Wie es zu einer Beurteilung kommt. Ein Forschungsbericht zu Strategien bei der Beurteilung schriftlicher Leistungen im Kontext der Prüfung TestDaF.

Ulrike Arras, Hagen

1 Problemaufriss

Die valide und reliable Beurteilung von Prüfungsleistungen stellt neben der Erstellung geeigneter Testaufgaben und der objektiven Prüfungsdurchführung eine weitere Säule der Qualitätssicherung eines Tests dar. Zwar nehmen Schulung und Kalibrierung der BeurteilerInnen von Prüfungsleistungen, die anhand von offenen Items erbracht wurden, gerade bei standardisierten Tests wie dem TestDaF eine zentrale Rolle ein. Jedoch wissen wir trotz kriterienorientierten Vorgehens und der Operationalisierung von Beurteilungsmaßstäben und -verfahren im Grunde wenig darüber, wie die BeurteilerInnen bei ihrer Arbeit vorgehen, mit welchen Strategien sie die Beurteilung einer (schriftlichen oder mündlichen) Leistung bewerkstelligen, worauf sie ihr Augenmerk richten und welche Textfaktoren, aber auch von der konkreten Leistung unabhängige Faktoren die Wahrnehmung und damit das Urteil beeinflussen. Gründe genug also, den Blick auf die Beurteile-

rInnen selbst zu lenken. Die hier vorzustellende Untersuchung möchte zur Erhellung der Frage beitragen, wie es überhaupt zu einer Beurteilung fremdsprachlicher Prüfungsleistungen kommt, indem sie die BeurteilerInnen, ihr (individuelles) Verhalten bei der Beurteilung schriftlicher Prüfungsleistungen zum Forschungsgegenstand macht. Eine explorativ-interpretative Untersuchung kann zum Verständnis dieses Gegenstands beitragen, indem mit Hilfe eines Mehrmethodendesigns und introspektiver Verfahren Einblicke in die Strategien und Prozesse der Beurteilungsarbeit ermöglicht werden.

2 Der TestDaF

Der TestDaF (Test Deutsch als Fremdsprache) ist ein seit 2001 vom TestDaF-Institut administriertes Testsystem für ausländische StudienbewerberInnen. Die Prüfung testet Deutschkenntnisse auf fortgeschrittenem Niveau, die für ein Studium an einer deutschsprachigen Hochschule relevant sind: Damit fungiert der TestDaF als Nachweis ausreichender Sprachkenntnisse, um ein Hochschulstudium aufzunehmen. Die vier Fertigkeiten Lesen, Hören, Schreiben und Sprechen werden getrennt in je eigenen Subtests gemessen und die Leistungen entsprechend getrennt auf einem Zeugnis ausgewiesen. Dank seiner Standardisierung und Orientierung an wichtigen Referenzsystemen wie dem Gemeinsamen europäischen Referenzrahmen für Sprachen des Europarats und der Skala der *Association of Language Testers in Europe* (ALTE)[1] ist ein TestDaF-Zeugnis jedoch auch für berufliche Zwecke, insbesondere für akademische Berufsfelder, von Nutzen. Die Prüfung wird weltweit durchgeführt, deshalb kann sie bereits im Heimatland abgelegt werden und vereinfacht dadurch den Zugang zu einem Hochschulstudium in Deutschland (Althaus 2004, Arras 2005).

3 Der Prüfungsteil Schriftlicher Ausdruck: Aufgabenformat und Beurteilungsverfahren

Die Überprüfung der schriftlichen Ausdrucksfähigkeit erfolgt anhand lediglich einer Texterstellungsaufgabe (Arras/Grotjahn 2003). Es handelt sich um eine direkte Erfassung der Fähigkeit auf der Basis eines offenen Itemformats, wobei jedoch schriftlich bzw. grafisch präsentierte Vorgaben die Aufgabe steuern. Die Prüfungsteilnehmenden sollen zeigen, ob sie in der Lage sind, zu einem bestimmten Thema einen zusammenhängenden und klar aufgebauten, diskursiven Text zu schreiben. Gefordert werden im Wesentlichen zwei Schreibhandlungen, die für den akademischen Kontext von besonderer Bedeutung sind: Das Beschreiben und

[1] Die Niveaustufen-Beschreibungen des Gemeinsamen europäischen Referenzrahmens für Sprachen s. www.coe.int sowie Europarat (2001). Die Niveaustufen-Beschreibungen der ALTE s.: www.alte.org.

Zusammenfassen von statistischen Daten, die in Form einer Grafik oder einer Tabelle präsentiert werden, sowie das Argumentieren, indem beispielsweise zu einer Frage oder einem Problem begründet Stellung genommen werden soll und dabei i.d.R. unterschiedliche Meinungen zu paraphrasieren und zu berücksichtigen sind[2].

Ein standardisierter Test, der weltweit abgenommen wird, muss sich eines kriterienorientierten Beurteilungsverfahrens bedienen. Das bedeutet: Ausschlaggebend bei der Bewertung von Leistungen ist nicht die (durchschnittliche) Leistung der Gesamtgruppe, i. e. alle Teilnehmenden an einem Prüfungsereignis, sondern die durch die TestDaF-Niveaustufen ausgewiesene Leistung selbst. Aus diesem Grunde werden die Leistungen aus den Prüfungsteilen zur Erfassung der produktiven Fähigkeiten, d. h. die schriftlichen und mündlichen Texte, zentral beurteilt und zwar von BeurteilerInnen, die eigens durch das TestDaF-Institut geschult werden und regelmäßig an Kalibrierungsveranstaltungen teilnehmen. Das wichtigste Instrument der Beurteilung sind die Bewertungskriterien in Form von skalierten Deskriptoren (s. Beurteilungsraster im Anhang). Das standardisierte Bewertungsverfahren sieht vor, dass die individuelle Prüfungsleistung hinsichtlich vorgegebener, das Testkonstrukt widerspiegelnder Aspekte mit den Deskriptoren des Beurteilungsrasters abgeglichen wird. Diese Kriterien erfassen zum einen die Gesamtwirkung bei der Rezeption eines Textes. Es handelt sich um eine eher holistische Erfassung der Leistung. Zum anderen werden die sprachliche und die inhaltliche Umsetzung der jeweiligen Aufgaben beurteilt, was ein eher analytisches Vorgehen bei der Bewertung erfordert. Um die Schwierigkeit der verschiedenen Aufgaben über verschiedene Testsätze bzw. Testereignisse hinweg konstant zu halten, bedarf es darüber hinaus einer testsatzspezifischen Kalibrierung. Die entsprechenden Instrumente bestehen aus zwei Teilen: Zum einen werden die Anforderungen aufgabenspezifisch festgehalten. Das bedeutet, zu jeder Aufgabe wird anhand von Texten aus den Erprobungen eruiert, welche Leistungen bzw. Anforderungen die Aufgabe elizitiert und welche Maßstäbe bei der Umsetzung der jeweiligen Aufgabe anzulegen sind. Hiermit wird also versucht, die Schwierigkeit über verschiedene Testereignisse und Aufgaben konstant zu halten, 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 entsprechenden begründeten Einstufungen werden schriftlich fixiert und stellen eine weitere Orientierungshilfe für alle BeurteilerInnen dar, die für die Bewertung schriftlicher bzw. mündlicher Leistungen aus einem

[2] Modellaufgaben zum Prüfungsteil Schriftlicher Ausdruck sind auf der Internetseite des TestDaF-Instituts unter www.testdaf.de einsehbar. Die im Rahmen der Untersuchung verwendete Aufgabe ist mittlerweile als „Musterprüfung 1“ veröffentlicht worden (TestDaF-Institut 2005).

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.

Da trotz Schulung, Kalibrierung und anderen Monitoring-Maßnahmen 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 einzelnen Beurteilerin bzw. des einzelnen Beurteilers 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 (Eckes 2003, 2004). Die Erfassung verschiedener Determinanten – also die Leistungseinstufungen hinsichtlich der verschiedenen Aspekte wie sie im Kriterienraster vorgegeben sind, die Aufgabenschwierigkeit sowie die individuelle Strenge und Konsistenz der BeurteilerInnen – ermöglicht schließlich eine faire endgültige Stufenzuweisung. Die mannigfaltigen Faktoren, die bei der Beurteilung einer (schriftlichen) Prüfungsleistung wirksam werden, können in einem Modell zusammengestellt werden.

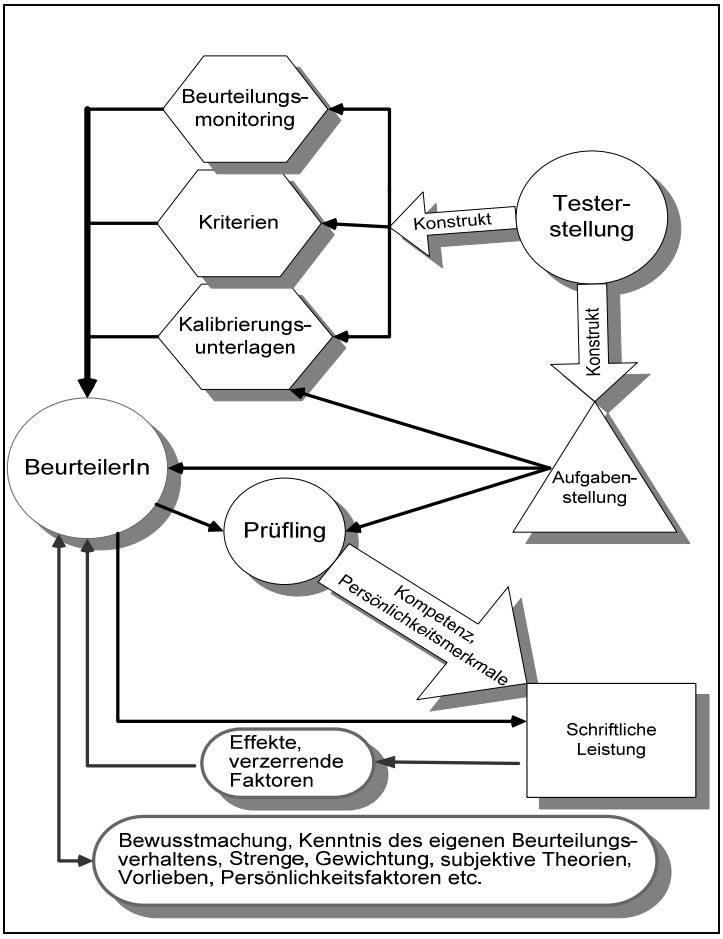

Abbildung: Faktorenmodell Schreibtest

Wie das Modell zeigt, spielen nicht allein die institutionell geprägten Faktoren wie Testkonstrukt, Kalibrierung, Schulung etc. eine Rolle. Vielmehr wird die Beurteilung durch Persönlichkeitsfaktoren sowie subjektive Theorien u. ä. Faktoren gesteuert. Es sind nicht zuletzt diese Faktoren, welche die hier vorzustellende empirische Studie zu eruieren sucht.

4 Fragestellung

Dass die Bewertung unserer Leistung nicht unbedingt unserer subjektiv empfundenen Selbsteinschätzung entspricht und auch nicht stets einer wie auch immer geregelten objektiven Bewertung standhält, ist eine Erfahrung, die wir spätestens

in der Schule machen müssen, wo nicht nur das Testen, sondern auch Klagen über ungerechte Noten an der Tagesordnung sind. Es ist daher überraschend, dass im deutschsprachigen Raum das Problem (subjektive) Bewertung (fremdsprachlicher Leistungen) nur sporadisch Gegenstand der Fachliteratur ist. Selbst in Fachzeitschriften, in denen Lehrkräfte selbst zu Wort kommen, finden sich nur gelegentlich Ansätze zur kritischen Auseinandersetzung damit. In der Regel werden dann jedoch lediglich Einzelaspekte wie die „Abmilderung" von oder Kritik an Fehlerquotienten (Drexel-Andrieu/Kahl 1991, Schnädter 1991) behandelt. Eine Ausnahme stellt die Studie von David (1991) dar: Schriftliche Abiturprüfungen im Fach Englisch wurden von verschiedenen Englischlehrkräften bewertet. Das Ergebnis ist deprimierend, denn die Lehrkräfte kommen zu völlig unterschiedlichen Beurteilungen, so dass erhebliche Zweifel an der Zulässigkeit dieses Prüfungsverfahrens gerechtfertigt erscheinen, zumal es sich bei der Abiturprüfung immerhin um einen High-Stakes-Test handelt, der zudem den Anspruch erhebt, in Ansätzen standardisiert zu sein. Tatsächlich erfüllt zumindest die untersuchte Prüfung wichtige Testgütekriterien nicht.

Es befremdet, dass in Deutschland – auch angesichts solcher, freilich singulärer Studien – der Gegenstand Testen und Prüfen (im Kontext Fremdsprachenerwerb) erst seit kurzem Beachtung findet[3]. In der englischsprachigen Fachliteratur hingegen werden unterschiedliche Aspekte der Testerstellung und -auswertung seit Langem behandelt. Insbesondere im Kontext der großen – standardisierten – Testsysteme zur englischen Sprache mit langer Tradition werden seit den 60er Jahren zunehmend die Probleme der Bewertung schriftlicher, später auch mündlicher, Leistungen erforscht. Mehrere empirische Studien vergleichen beispielsweise die Bewertung von Prüfungsleistungen durch verschiedene Gruppen von BeurteilerInnen („lay" vs. „expert", „native speaker" vs. „non-native speaker"); sie kommen freilich zu recht unterschiedlichen Ergebnissen (Überblick s. Charney 1984, Lumley 2005). Tendenziell ist jedoch festzuhalten, dass vor allem mangelnde Interrater-Reliabilität (s. die o. g. Studie von David), aber auch mangelnde Intrarater-Reliabilität Anlass zur Besorgnis geben. Den Untersuchungen gemeinsam ist, dass sie im Wesentlichen das Ergebnis der Bewertung fokussieren. Die gleichen Arbeiten werden von einer Versuchs- sowie von einer Kontrollgruppe bewertet, die Ergebnisse werden sodann miteinander verglichen. Wie jedoch bewertet wird, welche – insbesondere auch individuellen – Strategien oder auch subjektiven Theorien und Erfahrungen die BeurteilerInnen leiten, inwiefern Persönlichkeitsfaktoren, Erwartungen, Vorlieben und Idiosynkrasien, eine Leistungsbewertung mitbestimmen – und vor allem warum – bleibt weitgehend im Dunkeln.

Solche Fragen sind seit Beginn der 90er Jahre mehr und mehr in das Interesse der Forschung gerückt. Es sind wiederum die großen englischsprachigen Testsys-

[3] Für den schulischen Bereich exemplarisch Freudenstein 1994; einen historischen Forschungsüberblick zu subjektiven Fehlerquellen bei der Zensurvergabe liefern Ingenkamp 1995 und Kieweg 2001. Zum Rater-Verhalten im Kontext TestDaF s. Eckes (in Vorbereitung).

teme, die in den letzten Jahren zunehmend den Blick auf die mentalen Prozesse der Beurteilungsarbeit gerichtet haben – mit entsprechend methodischen Konsequenzen. Das bedeutet: Nicht mehr allein das Reliabilitätsproblem steht im Mittelpunkt, vielmehr wird nunmehr auch Fragen zur Validität von Bewertungsverfahren und -skalen nachgegangen. Arbeiten wie die von Weigle (1994), Milanovic et al. (1996), Cumming et al. (2002) sowie von Lumley (2005) erheben Daten mittels introspektiver Methoden, um das Beurteilungsverhalten, die spezifischen sowie individuell geprägten Beurteilungsstrategien und Entscheidungsprozesse zu analysieren.

Die vorliegende Untersuchung versteht sich als Teil dieser Entwicklung, um zur Theoriebildung des Gegenstandes Beurteilung von Prüfungsleistungen beizutragen. Sie widmet sich in erster Linie den bei der Beurteilung schriftlicher Leistungen beobachtbaren Strategien und geht der Frage nach, in welchen Schritten die Beurteilungsarbeit verläuft. Die folgenden Aspekte stehen dabei im Vordergrund:

- Wie gestaltet sich der individuelle sowie durch das Bewertungsverfahren vorgegebene Prozess der Rezeption und Urteilsfindung?
- Welche Strategien werden hierbei eingesetzt?
- Auf welche Probleme treffen die BeurteilerInnen und mittels welcher Strategien lösen sie diese?
- Gibt es Unterschiede in der Herangehensweise sowie bei der Interpretation der Leistungen und der Beurteilungsmaßstäbe?

Die Untersuchung erfolgt anhand eines spezifischen Aufgabentyps im Kontext einer High-Stakes-Prüfung mit entsprechend spezifischem Testkonstrukt und klar definiertem Testziel. Auch das Beurteilungsverfahren sowie die Beurteilungsinstrumente sind – wie erwähnt – analog den speziellen Anforderung der Prüfung konzipiert. Insofern scheint zunächst Vorsicht geboten bei der Generalisierung der Ergebnisse und der Übertragbarkeit bestimmter Befunde auf andere Kontexte. Allerdings lassen sowohl die verwendeten Methoden als auch die aufgefundenen Strategien und Beurteilungsprozesse gerade den Schluss zu, dass das Beurteilungsinstrumentarium ebenso wie seine Operationalisierung von so zentraler Bedeutung für das Beurteilungsverhalten ist, dass Aussagen möglich sind, die sich nicht allein auf den Kontext TestDaF beziehen, sondern auch allgemeine, bei der Konzipierung von Testsystemen und der Betreuung von BeurteilerInnen relevante Aspekte benennen.

5 Zum Untersuchungsdesign

Angesichts der Komplexität des Problems Leistungsbeurteilung liegt ein Mehrmethodendesign nahe, um durch Triangulation zur Verbesserung der Validität beizutragen (Flick 2002). Dies ist insbesondere vor dem Hintergrund erforder-

lich, dass die Studie im Wesentlichen auf Fallstudien gründet. Im Folgenden werden die einzelnen Schritte der Untersuchung kurz skizziert.

Die Untersuchung gliedert sich in zwei Vorstudien und die Hauptuntersuchung:

- Bei Vorstudie I handelt es sich um eine Fragebogenerhebung. TestDaF-BeurteilerInnen des Subtests Schriftlicher Ausdruck (39 Frauen, 11 Männer, Rücklauf ca. 50%) wurden mit Hilfe von offenen Fragen um eine Evaluierung bzw. kritische Rückmeldung zu den Schulungsmaßnahmen, den Beurteilungskriterien sowie zu den testsatzspezifischen Kalibrierungsmaterialien gebeten. Die Daten wurden einer Inhaltsanalyse unterzogen.
- Die Vorstudie II wurde im Anschluss an Vorstudie I mit vier Versuchspersonen durchgeführt und gliederte sich in zwei Teile: ein problemzentriertes, halbstrukturiertes und leitfadenorientiertes Interview und ein Laut-Denken-Protokoll während der Beurteilung einer schriftlichen Prüfungsleistung. Ziel dieses Teils war nicht zuletzt, das Laut-Denken-Verfahren zu erproben und seine Anwendbarkeit im Kontext Beurteilung zu prüfen.

Dank der beiden Vorstudien konnten erste Erkenntnisse sowie methodische Konsequenzen in das Design der Hauptstudie einfließen. Dies betraf die folgenden Entscheidungen:

- Auswahl und Anzahl der Versuchspersonen
- Auswahl der Aufgabenstellung
- Auswahl, Anzahl und Anordnung der zu beurteilenden schriftlichen Leistungen
- Prozedere und Durchführung der introspektiven Verfahren, also der Laut-Denken-Protokolle und der retrospektiven Interviews

Die Hauptuntersuchung erfolgte anhand von vier Fallbeispielen: Vier Versuchspersonen, es handelt sich um geschulte und erfahrene Beurteilerinnen von TestDaF-Prüfungsleistungen, beurteilten zunächst per Laut-Denken-Verfahren acht schriftliche Leistungen. Als Warming-up bzw. zu Trainingszwecken wurde ein Text vorgeschaltet, um das Laut-Denken-Verfahren zu trainieren und Rückmeldungen zu ermöglichen. Denn während der eigentlichen Laut-Denken-Sitzungen sollte seitens der Forscherin nach Möglichkeit nicht eingegriffen werden. Die Protokolle wurden unmittelbar nach Abschluss der Beurteilungsarbeit auditiv einer Grobanalyse unterzogen, so dass am Folgetag im Rahmen der introspektiven Interviews auf Basis der Protokolle problemzentriert das semistrukturierte Interview durchgeführt werden konnte. Diese Interviews wurden transkribiert und einer Inhaltsanalyse unterzogen. Die Laut-Denken-Protokolle wurden

angesichts der Fragestellung der Studie, die nicht allein die Strategien, sondern auch die Prozesshaftigkeit der Beurteilungsarbeit fokussiert, transkribiert, segmentiert und zu Analysezwecken anhand eines eigens entwickelten Kodiersystems kodiert[4]. Die Codes ermöglichen eine detaillierte Analyse der einzelnen Beurteilungsschritte und zeigen, in welchen Kontexten bestimmte Strategien auftreten und in welchen Abfolgen einzelne Strategien angeordnet sind.

6 Vorläufige Befunde

Die Verbaldaten aus den Laut-Denken-Protokollen haben eine Fülle an Handlungen und Strategien beobachtbar gemacht, deren Kontext und Motivierung durch die retrospektiven Interviews genauer betrachtet werden können, Erkenntnisse aus den beiden Vorstudien zu spezifischen Beurteilungsproblemen und -strategien validieren zudem die Befunde. Im Folgenden sollen erste Befunde aufgezeigt werden. Dabei können beobachtete Strategien lediglich skizziert werden, auf ausführliche Belege mit Hilfe der Verbaldaten muss weitgehend verzichtet werden. Im Wesentlichen sollen folgende Aspekte dargelegt werden:

- Der Beurteilungsvorgang, also die Phasierung der Beurteilungsarbeit im Kontext des vorgegebenen Beurteilungsverfahrens
- Individuell und institutionell determinierte Beurteilungsstrategien
- Die Funktion und Anwendung der Deskriptoren, auch individuell geprägte Interpretationen und Strategien des Ausgleichens, des Abwägens u. Ä.

Dabei lassen sich Strategien auf der Makroebene von solchen auf der Mikroebene der Beurteilungsarbeit unterscheiden. Zunächst zeigt sich, dass sich der Bewertungsablauf auf Makroebene sehr stark an dem standardisierten Verfahren orientiert, so dass folgende Phasen differenziert werden können:

- Erste Wahrnehmung und Identifizierung der zu beurteilenden Leistung
- Lesedurchgang (zur Beurteilung holistischer Aspekte)
- Lesedurchgang (zur Beurteilung analytischer Aspekte)
- Resümee.

Der Ablauf der Beurteilungsarbeit kann wie folgt dargestellt werden:

[4] Zur Aufbereitung und Analyse von Verbalprotokollen s. beispielsweise Green 1998.

Erste Wahrnehmung und Identifizierung:	1. Lesedurchgang	2. Lesedurchgang	Resümée
- Konstituierung einer ersten Erwartung - unter Einbezug von Faktoren wie Handschrift, Text länge etc.	- totales Lesen - ggf. wiederholtes Lesen - unter Einbezug des korrigierenden Lesens, Sinn-Rekonstruktion - holistische Beurteilung mittels Deskriptoren - ...	- scannen - paraphrasieren - ggf. erneut totales Lesen - Beurteilung analytischer Aspekte mittels Desk riptoren - ...	- Einbezug des Testkonstrukts - Einbezug der „real world“ - ...

Diese vier Phasen sind unterschiedlich lang. Den größten Raum nehmen die eigentlichen Lesephasen, also der erste sowie der zweite Lesedurchgang ein. Vor allem in diesen Phasen zeigt sich auf der Mikroebene eine breite Palette an unterschiedlichen Strategien, die der Rezeption der Leistung sowie der Entscheidungsfindung in Bezug auf das Urteil dienen. Die Strategien, die in den vier verschiedenen Phasen eingesetzt werden, lassen sich wie folgt gruppieren:

1. Zu den Rezeptionsstrategien, im Wesentlichen also jene des Lesens und der Wahrnehmung des Textes, können folgende Strategien gezählt werden:
   - Identifizierung der zu beurteilenden Leistung
   - Totales Lesen
   - Scannen und Search Reading
   - Paraphrasieren
   - Korrigierendes Lesen
   - Interpretierendes Lesen

2. Strategien, die die Beurteilungsinstrumente sowie das institutionalisierte Testkonstrukt einbeziehen; hier sind insbesondere zu nennen:
   - Abgleich zwischen Leistung und Deskriptoren
   - Ausgleichen
   - Abwägen
   - Einbezug der Aufgabenstellung
   - Gewichten von Aspekten der Leistung

3. Strategien, die auf eigene Erfahrungen und auf subjektive Theorien (auch subjektive Testkonstrukte) rekurrieren:
- Konstituierung der Erwartung(en)
- Konstituierung des ersten Eindrucks
- Einbezug externer Faktoren wie Handschrift, Textlänge etc.
- Einbezug des Testkonstrukts (auch vor dem Hintergrund eigener – beruflicher etc. – Erfahrungen)
- Entwicklung einer Vorstellung von der Person des Prüflings

4. Eine weitere, für die Beurteilungsarbeit wichtige Strategiengruppe liegt in Form von Strategien zur Arbeitsorganisation vor. Hierzu zählen im Wesentlichen weitgehend individualisierte Notizformen („Gedächtnisstützen") sowie die Arbeitseinteilung und das Zeitmanagement.

Eine weitere Kategorisierung bietet sich hinsichtlich der Referenz der eingesetzten Strategien an:
- Strategien, die sich der zur Verfügung stehenden Beurteilungsinstrumente bedienen
- Strategien, die sich auf das übergeordnete Testkonstrukt beziehen

So referieren Strategien wie das Abwägen auf die Deskriptoren des Beurteilungsrasters, während bei Strategien wie dem Resümieren Bezug zum zugrunde gelegten Testkonstrukt genommen wird. Zu der Kategorie der verfahrensgesteuerten Strategien zählen solche, die im Kontext des Leseprozesses eingesetzt werden sowie jene Strategien, die auf der Anwendung der Deskriptoren beruhen. Es handelt sich um Strategien, die durch das (standardisierte) Beurteilungsverfahren geprägt sind. Hierzu zählen Faktoren wie das Lesen, die Anwendung der Kriterien, die Berücksichtigung des Testkonstrukts, die Einbeziehung von Erfahrungen und Erkenntnissen aus Kalibrierungen. Es handelt sich um Strategien, die von allen Versuchspersonen angewendet werden, wenn auch in unterschiedlich starkem Ausmaß. Zu den individuell geprägten Strategien zählen beispielsweise Strategien, die der Arbeitsorganisation dienen, etwa individuelle Notizformen oder Maßnahmen zum Zeitmanagement, aber auch solche Strategien, die von Faktoren der Persönlichkeit geprägt sind wie der Einbezug von Erfahrungen, Vorlieben bei der Beurteilung von Leistungen u. Ä.

Die Strategien interagieren oder korrespondieren miteinander und ergänzen sich, etwa wenn die Strategie des Abwägens die Strategie des wiederholten Lesens einzelner Passagen notwendig macht, um Belege für oder gegen ein vorläufiges Urteil vorzubringen. Damit formieren sich Strategiengruppierungen. Das bedeutet: Strategien ergeben sich aus bereits zuvor verwendeten Strategien. Beispielsweise besteht eine enge Verbindung zwischen der Strategie des Abwägens und der Strategie der Verifizierung oder Revision des Urteils. Zu dieser Gruppe gesellt sich

auch eine spezifische Lesestrategie, nämlich die des Scannens zur Suche nach Belegen für oder gegen eine (vorläufige) Entscheidung. Solche Strategiencluster liegen auch hinsichtlich des Einbezugs der Beurteilungskriterien vor: So erfolgt die Einstufung eines Einzelaspekts prozesshaft mittels Aussage zur Qualität i. d. R. unter Zuhilfenahme der Deskriptoren in paraphrasierter Form, oder indem der geeignet erscheinende Deskriptor zitiert wird.

Diese meist sehr dichte Abfolge verschiedener Strategien soll anhand eines Auszugs aus einem der Protokolle illustriert werden. In dem zitierten Auszug beurteilt die Versuchsperson die Qualität der geforderten Schreibhandlung Beschreiben statistischer Daten, die in einer Grafik präsentiert werden[5].

Transkript
Beschreibung der Grafik. [räuspert sich] [.] Ist ein bisschen quer mit der Grafik. Ist NICHT GANZ KLAR gedacht. ALSO KLAR, 4 IST KLAR UND FOLGERICHTIG WIEDERGEGEBEN. [..] *Während die Anteile der Eltern Studierende unter 24 ..... noch groß sind betragen sie ziehmlich wenig*. Es ist NICHT KLAR ausgedrückt. Es ist irgendwie *Anteile der Wohngemeinschaft nimmt nicht so auffällig ab.* [!] Was ist das, nimmt nicht so auffällig ab? Was soll uns das sagen? Das ist irgendwo nonsense. Und die Schwankung dazwischen ist auch Quatsch. Dann so ein paar Zahlen aneinander gereiht. [..] Eh ist nicht so richtig schön in Zusammenhang. ALSO KLAR UND FOLGERICHTIG, ne 4, würde ich da nicht geben, sondern da tendiere ich eher zu der 3. Schreibe ich jetzt mal hin. *Die Informationen der Grafik werden überwiegend aufzählend wiedergegeben.* Ja, ÜBERWIEGEND AUFZÄHLEND. Es ist zwar ein bisschen ein Versuch, so quer zu denken, aber es ist [...] ja es ist auch nicht natürlich, auch nicht, nicht total AUFZÄHLEND, ABER ÜBERWIEGEND AUFZÄHLEND, kann man, könnte man vertreten. [...] *Die Informationen der Grafik werden folgerichtig, klar und folgerichtig wiedergegeben.* Ne 4, das ist es nicht. ES IST NICHT KLAR UND FOLGERICHTIG. ES IST KLAR, NICHT [!] KLAR, was da *der Anteil der Wohngemeinschaft nimmt nicht so auffällig ab wie die obengenannte* Was soll das? Gut, bleibe ich bei der 3.

[5] Bei den im Transkript in Kapitälchen gesetzten Äußerungen handelt es sich um Äußerungen, die direkt (fett) oder in paraphrasierter Form Bezug zu den Deskriptoren nehmen (s. Beurteilungsraster im Anhang). Wenn die Versuchsperson die zu beurteilende schriftliche Leistung liest, ist diese Äußerung kursiv gesetzt. Aus Gründen der besseren Lesbarkeit sind die Zeitleiste des Protokolls sowie die Codes hier nicht abgebildet.

Der Auszug zeigt zum einen, wie intensiv die Deskriptoren bei der Urteilsfindung herangezogen werden (Kapitälchen). Sie werden meist paraphrasiert, nur manchmal direkt zitiert (Kapitälchen fett), vermutlich dann, wenn ein direkter Vergleich zwischen dem Wortlaut der Deskriptoren und dem Eindruck der Leistung erforderlich ist. Damit erweisen sich die Deskriptoren als zentrales Instrument bei der konkreten Beurteilungsarbeit. Sie sind von den vier Versuchspersonen derart verinnerlicht, dass Formulierungen und Wortmaterial der Deskriptoren verwendet werden. Mit Hilfe des sprachlichen Materials aus den Deskriptoren wird schließlich eine Beurteilung formuliert. Oder aber das sprachliche Material der Deskriptoren wird in Form einer an den zu beurteilenden Text gerichteten Frage bzw. in Form einer Antwort verwendet, eine Entscheidung herbeizuführen. Die zuletzt genannte Vorgehensweise zeigt sich im folgenden Auszug:

Transkript
Und Korrektheit. Ja, da sind also manchmal Fehler. Frage, SIND DIE FEHLER SO, DASS ES MEIN VERSTEHEN BEEINTRÄCHTIGT? Ja, ich glaube fast, DIE FEHLER SIND SO, DASS ES MEIN VERSTEHEN MANCHMAL BEEINTRÄCHTIGT.

Zum anderen wird die Verbindung der einzelnen Strategien zu Clusters deutlich. Die Urteilsfindung erfolgt im steten Wechsel zwischen der genannten Referenz auf die Deskriptoren (und zwar jenen der angrenzenden Niveaustufen), der Überprüfung der Leistung, der Suche nach Belegen und dem Abwägen.

Die hier genannten Strategien können auch unter dem Gesichtspunkt der Referenz kategorisiert werden:

- Strategien, die durch das vorgegebene Beurteilungsverfahren bestimmt werden (Lesen, Abgleichen)
- Strategien, die durch die Person bzw. durch individuelle Faktoren der Beurteilerin geprägt sind (z. B. solche, die nur von bestimmten Versuchspersonen verwendet werden, von anderen nicht)
- Strategien, die sowohl auf individuelle Faktoren der Versuchspersonen als auch auf den standardisierten Bewertungsverlauf zurückzuführen sind (z. B. die Erwartungen an die Leistung, die wiederum bestimmt wird zum einen durch die Berufserfahrung, zum anderen aber auch durch Faktoren wie Kalibrierungen)

Die meisten der in der vorliegenden Untersuchung eruierten Strategien weisen sowohl Einflüsse aus dem persönlichen oder beruflichen Erfahrungsschatz als auch aus dem weitgehend standardisierten Beurteilungsverfahren auf. Danach entsteht eine Teilmenge durch das Zusammentreffen von institutionell geprägten

Strategien, etwa solchen, die auf das Testkonstrukt oder auf das Beurteilungsinstrumentarium referieren, und jenen, die von Persönlichkeitsfaktoren geprägt sind und auf individuellen Erfahrungen, subjektiven Theorien u. Ä. beruhen.

7 Erste Schlussfolgerungen

Der Forschungsstand, insbesondere auch die auf der Basis von introspektiven Daten entwickelten Modelle und Systematisierungen der Beurteilungsstrategien und -prozesse zeigen, dass die Bewertung (schriftlicher) Leistungen eine komplexe kognitive Handlung, bestehend aus verschiedenen, sich ergänzenden und aufeinander aufbauenden Einzelhandlungen, darstellt. Diese wiederum sind geürägt durch Faktoren wie persönliche Erwartungen und Erfahrungen, aber auch durch institutionelle Faktoren wie das Beurteilungsinstrumentarium und vor allem auch das Training, durch welches Testkonstrukt und Maßstäbe transportiert werden. Beurteilen stellt damit eine kontinuierliche „problem-solving activity" dar, bei der die BeurteilerInnen die Informationen und Begriffe der Beurteilungsmaßstäbe interpretieren „and then reconcile this interpretation with the specifics of the text. Thus, evaluating writing when using a scoring rubric is a constructive activity." (DeRemer 1998: 13). Angesichts der Komplexität und der teils individuellen Prägung der Beurteilungsstrategien erscheint eine Auseinandersetzung mit den sie determinierenden Faktoren, ja eine Bewusstmachung des eigenen Handelns während der Beurteilung von besonderer Bedeutung. Diese vorläufige Schlussfolgerung stellt ein starkes Argument dar für die Bedeutung von Schulungen. Die Auseinandersetzung mit dem kriterienorientierten Beurteilen und den eigenen Beurteilungsstrategien, die auf subjektiven Theorien, auf eigenen (persönlichen und beruflichen) Erfahrungen und anderen, auch affektiven und sozialen, Faktoren beruhen, trägt zur Professionalisierung der Beurteilungsarbeit bei, und zwar unabhängig davon, ob es sich um BeurteilerInnen handelt, die Leistungen im Kontext eines speziellen Tests beurteilen, oder ob es um angehende oder fortzubildende (Fremd-)SprachenlehrerInnen geht, die für ein breites Spektrum an Leistungsmessung – verschiedene Formen des *classroom assessment* bis hin zu (standardisierten) High-Stakes-Prüfungen wie insbesondere das Abitur – vorbereitet und trainiert werden müssen. Trotz der eingeschränkten Generalisierbarkeit der anhand von vier Fallbeispielen eruierten Befunde können einige praktische Vorschläge vor allem für die Optimierung von Trainings gemacht werden. Eine Möglichkeit für Schulungsmaßnahmen kann in der Verwendung des Laut-Denken-Verfahrens liegen. Das Verfahren macht die eigenen Verhaltensweisen und Strategien bei der Beurteilung schriftlicher Leistungen bewusst und somit zugänglich für Reflexion. Damit kommt dem Verfahren ein Lerneffekt zu, dessen sich Schulungen bedienen können. Dieser Ansatz wird auch in anderen Untersuchungen zu Problemlösestrategien vertreten. So kommt Dominowski (1998: 43) zu dem Schluss, dass das

Verfahren des Lauten Denkens gezielt für die Effektivierung von Problemlösestrategien eingesetzt werden sollte:

> Having people simply think aloud while working on a problem can provide useful information about problem-solving processes; task performance typically is not changed. Asking more specific questions [...] can elicit additional information and might affect problem solving, presumably for the better. Asking people to focus on their own problem solving, to explain what they are trying to do, promotes metacognitive processing and leads to more effective problem solving, even when the questions are no longer asked. (Dominowski 1998: 43).

Diese Überlegungen werden durch Daten der Untersuchung gestützt. Alle vier Beurteilerinnen der Studie, aber auch die vier Versuchspersonen der Vorstudie II geben nach den Laut-Denken-Sitzungen an, dass das Verfahren, während der Bewertung die kognitiven und emotionalen Prozesse zu verbalisieren, ihnen interessante Aufschlüsse über sich selbst und ihr Verhalten gab. Die vier Versuchspersonen der Hauptstudie wurden im retrospektiven Interview am Tag nach den Laut-Denken-Sitzungen einleitend um Kommentierung des Laut-Denken-Verfahrens selbst gebeten. Hier merkten alle Beurteilerinnen an, dass dieses Verfahren zwar anstrengend und zeitintensiv war, dass sie jedoch ganz persönlich davon profitiert hatten, denn das Laute Denken, v. a. im Beisein einer weiteren Person, führe zu intensiver Reflexion des eigenen Verhaltens. Im Gegensatz zu Erkenntnissen aus der Erforschung von Lern- und anderen Problemlösestrategien liegt im Falle der Beurteilungsstrategien jedoch ein gewichtiger Unterschied vor: das Gebot der Standardisierung. Im Falle beispielsweise des Fremdsprachlernens geht es darum,

> dem einzelnen Lerner zu helfen, aus dem präsentierten Strategienspektrum die zum jeweiligen Lerntyp und der Aufgabe passenden Strategien *bewusst* auszuwählen, praktisch zu erproben und selbst zu evaluieren, damit er so schrittweise sein Strategienrepertoire erweitern und modifizieren kann (Kleppin/Tönshoff 2000: 113, Hervorhebung im Original).

Geht es im Falle von Lernstrategien also um Individualisierung bzw. um Diversifikation, so ist es Anliegen eines standardisierter Tests wie dem TestDaF, ein gewisses Maß an Standardisierung auch in Form von Beurteilungsverfahren zu gewährleisten, um das Testkonstrukt und die Validität zu berücksichtigen. Ein standardisiertes Beurteilungsverfahren schränkt damit die Freiheit in der Auswahl an Beurteilungsstrategien notgedrungen ein. Aus diesem Grunde erscheint die Bewusstmachung eigener, aber auch das Angebot an bislang nicht eingesetzten Strategien sowie ihrer unter Berücksichtigung des Testkonstrukts begründeten Evaluation und Revision besonders naheliegend.

Literatur

Althaus, Hans-Joachim: „Der TestDaF". In: DAAD (ed.): *Die internationale Hochschule: Ein Handbuch für Politik und Praxis*, Band 8. Bielefeld: Bertelsmann 2004, 80-87.

Arras, Ulrike: „Der TestDaF. Konzept und Prinzipien des standardisierten Tests Deutsch als Fremdsprache". In: *Fòrum – Anuari de l'Associació de Germanistes de Catalunya. Akten des sechsten Kongresses des Katalanischen Deutschlehrer- und Germanistenverbandes (A.G.C.)*, Tarragona, April 2005, Número 12.

Arras, Ulrike / Grotjahn, Rüdiger.: „TestDaF: Einige aktuelle Entwicklungen". In: Katzorke, Heidrun (ed.): *Fremdsprachen an Hochschulen: Integration – Interdisziplinarität – Internationalität. Dokumentation der 22. Arbeitstagung 2002*. Bochum: AKS-Verlag 2003, 40-50.

Charney, Davida: "The validity of using holistic scoring to evaluate writing: a critical overview", *Research in the Teaching of English* 18 (1984), 65-81.

Cumming; Alister / Kantor, Robert / Powers, Donald E.: „Decision making while rating ESL/EFL writing tasks. A descriptive framework", *Modern Language Journal* 86/1 (2002), 67-96.

David, Reinhard: „Einheitliche Prüfungsanforderungen in der Abiturprüfung Englisch? Eine Betrachtung nach einer Vergleichskorrektur", *Die Neueren Sprachen* 90/6 (1991), 624-635.

DeRemer, Mary L.: „Writing assessment: Raters' elaboration of the rating task", *Assessing Writing* 5/1 (1998), 7-29.

Dominowski, Roger L.: „Verbalization and problem solving". In: Hacker, Douglas J./ Dunlosky, John / Graesser, Arthur C. (Hrsg.): *Metacognition in Educational Theory and Practice*. Mahwah, NJ, London: Erlbaum (1998), 25-45.

Drexel-Andrieu, Irène / Kahl, Detlev: „Empfehlungen zur Fehlerbewertung im schriftlichen Abitur", *Die Neueren Sprachen* 90/6 (1991), 686-687.

Eckes, Thomas: „Qualitätssicherung beim TestDaF: Konzepte, Methoden, Ergebnisse", *Fremdsprachen und Hochschule*, 69 (2003), 43-68.

Eckes, Thomas: „Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen". In: Wolff, Armin / Ostermann, Torsten / Chlosta, Christoph (Hrsg): *Integration durch Sprache*. Regensburg: FaDaF 2004, 485-518.

Eckes, Thomas (in Vorbereitung): „Rater types in writing performance assessments: a classification approach to rater variability", *Language Testing*.

Europarat: *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. Berlin et al.: Langenscheidt 2001.

Flick, Uwe: *Qualitative Sozialforschung. Eine Einführung.* 6. Auflage. Reinbek bei Hamburg: Rowohlt 2002.

Freudenstein, Reinhold: „Fremdsprachentests für die Schule. Ein Plädoyer für die objektivierte Leistungsmessung", *PRAXIS* 41/4 (1994), 339-347.

Green, Alison J. K.: *Verbal Protocol Analysis in Language Testing Research. A Handbook*. UCLES, Cambridge University Press 1998.

Ingenkamp, Karlheinz (Hrsg.): *Die Fragwürdigkeit der* Zensurengebung. Texte und Untersuchungsberichte. 9. Auflage. Weinheim, Basel: Beltz 1995.

Kieweg, Werner: Evaluation fremdsprachlicher Leistungen im schulischen Kontext. In: *FluL* 30 (2001), 65-86.

Kleppin, Karin / Tönshoff, Wolfgang: „Autonomiefördernde Strategievermittlung als Gegenstand und Verfahren in der Ausbildung von Fremdsprachenlehrern". In: Helbig, Beate / Kleppin, Karin / Königs Frank G. (Hrsg.): *Sprachlehrforschung im Wandel. Beiträge zur Erforschung des Lehrens und Lernens von Fremdsprachen. Festschrift für Karl-Richard Bausch zum 60. Geburtstag.* Tübingen: Stauffenburg, 2000, 113-128.

Kvale, Steinar: *InterViews. An introduction to qualitative research interviewing*. Thousand Oaks et al.: SAGE Publications 1996.

Lumley, Tom: Assessing Second Language Writing. The Rater's Perspective. Frankfurt/Main: Peter Lang 2005.

Milanovic, Michael / Saville, Nick / Pollitt, Alastair / Cook, A. (1996): „Developing rating scales for CASE: theoretical concerns and analyses". In: Cumming, Alister / Berwick, Richard (Hrsg.): Validation in Language Testing. Clevedon: Multilingual Matters 1996, 15-38.

Schnädter, Herbert: „Der Fehlerindex – ein zuverlässiger Bewertungsfaktor? Zur Korrekturpraxis im Fach Französisch", Die Neueren Sprachen 90/6 (1991), 636-652.

Steinke, Ines: Kriterien qualitativer Forschung. Ansätze zur Bewertung qualitativ-empirischer Sozialforschung. Weinheim, München: Juventa 1999.

Steinke, Ines: „Gütekriterien qualitativer Forschung". In: Flick, Uwe / Kardorff, Ernst v. / Steinke, Ines (Hrsg.): Qualitative Forschung. Ein Handbuch. 2. Auflage. Reinbek bei Hamburg: Rowohlt 2003, 319-331.

TestDaF-Institut: Musterprüfung 1. Ismaning: Hueber 2005.

Weigle, Sara C.: „Effects of training on raters of ESL compositions", Language Testing 11/2 (1994), 197-223.

Internet-Seiten

www.alte.org

www.coe.int

www.testdaf.de