BEURTEILERINNEN IN DEN KOPF GESCHAUT. WIE DAS VERFAHREN DES LAUTEN DENKENS IM RAHMEN VON BEURTEILUNGSSCHULUNGEN EINGESETZT WERDEN KANN.[1]

Arras, Ulrike/Marks, Daniela/Zimmermann, Sonja, TestDaF-Institut, Hagen

0 Problemaufriss und Begründungszusammenhang

Ulrike Arras, Daniela Marks und Sonja Zimmermann sind Referentinnen für Testentwicklung am TestDaF-Institut, das seit 2001 die standardisierte Prüfung Test Deutsch als Fremdsprache erstellt und weltweit administriert. Ihr Arbeitsgebiet umfasst neben Fragen der Testerstellung und Leistungsbeurteilung die Durchführung von Schulungen und Fortbildungen auf dem Gebiet der Leistungsmessung. Vorliegender Beitrag wurde auf der XIV. IDT in Jena in der Sektion I2 Qualitätsentwicklung und Qualitätssicherung in institutionellen Zusammenhängen vorgetragen.

Die Beurteilung von Prüfungsleistungen stellt ein zentrales Problem bei der Qualitätssicherung eines Tests dar. Zwar sind Schulungen und Kalibrierungen[2] der BeurteilerInnen gerade bei standardisierten Tests grundlegend, jedoch wissen wir trotz kriterienorientierten Vorgehens und der Operationalisierung von Beurteilungsmaßstäben wenig darüber, wie die BeurteilerInnen bei ihrer Arbeit verfahren, mit welchen Strategien sie die Beurteilung einer Leistung bewerkstelligen und welche Faktoren die Wahrnehmung und damit das Urteil beeinflussen. Daher besteht ohne diese Erkenntnisse die Gefahr mangelnder Validität der Beurteilung.

Um Einblicke in die Vorgehensweisen und Maßstäbe der Beurteilung zu erhalten, wurden in einer empirischen Studie mittels introspektiver Verfahren – wie sogenannter Laut-Denken-Protokolle und retrospektiver Interviews – Daten im Kontext der Prüfung Test Deutsch als Fremdsprache (TestDaF) erhoben. Beim Laut-Denken-Verfahren waren die an der Studie teilnehmenden BeurteilerInnen gehalten, möglichst alle ihre Gedanken und Überlegungen während einer konkreten Arbeit (hier die Beurteilung schriftlicher Leistungen) zu verbalisieren, um so Einblicke in die bei der Beurteilung ablaufenden kognitiven Prozesse zu gewinnen.

Ein wesentlicher Befund bestand darin, dass die BeurteilerInnen das Verfahren des Lauten Denkens für sich selbst und für ihre Beurteilung als hilfreich erachteten. Dies führte zu der Frage, ob introspektive Verfahren bei der Schulung von BeurteilerInnen dienlich sein könnten, insbesondere weil sie dabei helfen, verzerrende Faktoren wie individuell geprägte Beurteilungsstrategien, subjektive Theorien[3], aber auch akute persönliche Befindlichkeiten während der Beurteilungsarbeit bewusst zu machen. Die Bewusstmachung dieser Faktoren ist von zentraler Bedeutung bei Beurteilungsschulungen, denn sie ermöglicht eine kritische Reflexion und darauf aufbauend eine mögliche Revision des Beurteilungsverhaltens, was wiederum ausschlaggebend ist für eine valide, an gemeinsamen Kriterien orientierte Leistungsbeurteilung.

Um zu überprüfen, ob das Laute Denken als integraler Bestandteil von Beurteilungsschulungen eingesetzt werden kann, wurden in einer Pilotstudie das Verfahren erprobt, und mittels Fragebogen Daten zur Handhabbarkeit und zum Nutzen dieses aufwendigen Verfahrens erhoben.

Im Folgenden soll zunächst die genannte Studie zu Beurteilungsstrategien kurz vorgestellt werden, um sodann Erkenntnisse aus der erwähnten Pilotstudie zur Einsetzbarkeit des Lauten Denkens in Beurteilungsschulungen zu diskutieren und praktische Vorschläge zu unterbreiten. Zunächst jedoch soll kurz der Rahmen der genannten Untersuchungen skizziert werden, nämlich die Prüfung TestDaF und der Prüfungsteil Schriftlicher Ausdruck, auf dessen Grundlage die Daten erhoben wurden.

1 Der TestDaF

Die Prüfung TestDaF ist eine seit 2001 eingesetzte Deutschprüfung für ausländische Studierende, die in Deutschland ein Hochschulstudium aufnehmen wollen. Es handelt sich um eine standardisierte Prüfung auf den Kompetenzstufen B2 und C1 des Gemeinsamen europäischen Referenzrahmens für Sprachen (GER). Getestet werden die vier Kompetenzbereiche Leseverstehen, Hörverstehen, Schriftlicher Ausdruck und Mündlicher Ausdruck in vier separaten Subtests.

Die zu testenden Kompetenzen umfassen Sprachhandlungen, die für den akademischen Kontext relevant sind, so beispielsweise beim Prüfungsteil Schriftlicher Ausdruck das Abfassen eines längeren, diskursiven Textes anhand von Vorgaben wie statistische Daten und Statements o. Ä[4].

Die Prüfung wird zentral im TestDaF-Institut in Deutschland erstellt, dezentral in lizenzierten Testzentren in aller Welt durchgeführt und unmittelbar danach zentral in Deutschland ausgewertet. Die Qualitätssicherung ruht dabei auf mehreren Säulen:

- Die Testerstellung liegt in der Hand geschulter TestautorInnen, die die Testaufgaben anhand vorgegebener Testspezifikationen konzipieren.
- Die Testaufgaben werden vor ihrem Einsatz im Rahmen einer TestDaF-Prüfung in einem umfassenden Evaluationsprozess erprobt, testmethodisch analysiert und ggf. revidiert.
- Die Durchführung erfolgt weltweit anhand vorgeschriebener Durchführungsmodalitäten.
- Die Beurteilung der Leistungen geschieht analog des für jeden Prüfungsteil vorgesehenen Testformats: Im Falle geschlossener Items des Leseverstehens und in Teilen des Hörverstehens werden die Punktscores ermittelt und das Ergebnis einer der drei Kompetenzstufen zugeordnet[5]. Die Teilnehmerantworten zu den offenen Itemformaten der Subtests Schriftlicher Ausdruck und Mündlicher Ausdruck werden von geschulten BeurteilerInnen beurteilt. Hierzu führt das TestDaF-Institut verschiedene Maßnahmen durch, die ein größtmögliches Maß an Validität

und Reliabilität der Urteile ermöglichen, nämlich vorgegebene Beurteilungskriterien in Form skalierter Deskriptoren, Schulungen und testsatzspezifische Kalibrierungen, mit denen die BeurteilerInnen zu jedem Testsatz erneut auf die Beurteilungsarbeit vorbereitet werden.

- Herzstück der Beurteilungsarbeit in den produktiven Prüfungsteilen sind skalierte Deskriptoren, die bei der Beurteilung anzulegen sind. Insgesamt werden im Schriftlichen Ausdruck neun Einzelaspekte weitgehend unabhängig voneinander und gleichgewichtet beurteilt[6]. Die Ermittlung eines fairen Durchschnitts für die Prüfungsteilnehmenden erfolgt dann im TestDaF-Institut mittels Multifacetten-Rasch-Analysen. Bei diesem Verfahren wird der Einfluss von unterschiedlichen Variablen auf die Beurteilung – beispielsweise die Strenge oder Milde der beurteilenden Person – erfasst und ausgeglichen. Zudem ermöglicht dieses Verfahren auch eine individuelle Rückmeldung zur Beurteilungsleistung an die BeurteilerInnen (s. dazu für den Kontext TestDaF: Eckes (2004, 2005); für den Kontext DSD: Eckes/Weiss-Motz/Whelan-Mostofizadeh (2009).

1.1 Prüfungsteil Schriftlicher Ausdruck

Im Prüfungsteil Schriftlicher Ausdruck müssen die Prüfungsteilnehmenden eine komplexe Aufgabe bearbeiten, die verschiedene kognitive Operationen und Schreibhandlungen umfasst[7]. Die Prüflinge schreiben einen diskursiven Text anhand von Vorgaben in Form von statistischen Daten, die in einem Diagramm aufbereitet sind, und Statements zu einem gesellschaftlich oder (bildungs-) politisch relevanten und aktuellen Thema. Die Daten sind zu beschreiben, das Thema bzw. Die Leitfrage ist zu diskutieren. Ferner soll der Text einen klaren Gedankengang und eine Textstruktur aufweisen, die dem Rezipierenden die Orientierung erleichtert[8]. Für die Bearbeitung der Aufgabe stehen 60 Minuten zur Verfügung.

2 Untersuchung zu Beurteilungsstrategien

Die Studie zur Eruierung spezifischer Beurteilungsstrategien bei der Beurteilung schriftlicher Leistungen im Kontext der Prüfung TestDaF basiert auf introspektiven Daten, die mittels Laut-Denken-Protokollen erhoben wurden. In vier Fallstudien beurteilten BeurteilerInnen dieselben acht authentischen schriftlichen Leistungen aus dem Prüfungsteil Schriftlicher Ausdruck unter den Bedingungen des Lauten Denkens[9]. Es handelt sich um langjährige und mehrfach geschulte Beurteilerinnen von TestDaF-Prüfungsleistungen, die somit sehr gut vertraut sind mit dem Beurteilungsverfahren, den Beurteilungsmaßstäben sowie den Anforderungen des Prüfungsteils. Die vier Beurteilerinnen wurden aufgefordert, die Beurteilung der Leistungen nach dem gewohnten TestDaF-Beurteilungsverfahren vorzunehmen. Die Daten der Laut-Denken-Protokolle wurden aufgezeichnet, transkribiert und sodann segmentiert und kodiert, um sie einer Analyse zugänglich zu machen.

Zudem wurden am Tag nach den Laut-Denken-Sitzungen noch retrospektive Interviews durchgeführt. Dies sollte zusätzliche Aufschlüsse etwa hinsichtlich problematischer Stellen in den Laut-Denken-Protokollen erbringen. Außerdem gab es den Beurteilerinnen die Möglichkeit, Rückmeldung zum Verfahren zu geben. Die retrospektiven Interviews wurden ebenfalls transkribiert und einer Inhaltsanalyse unterzogen.

2.1 Erkenntnisse für zukünftige Beurteilungsschulungen

Die Untersuchung zeigte, dass die BeurteilerInnen vielfältige Strategien bei der Beurteilung der schriftlichen Texte einsetzten, die sich aus verschiedenen, sich ergänzenden und aufeinander aufbauenden Einzelhandlungen zusammensetzen[10].

Besonders zu erwähnen ist, dass die Handlungen sehr stark von den Beurteilungskriterien bestimmt sind – bei einer kriterienorientierten Beurteilung nicht überraschend, da sie zu den institutionellen Faktoren gehören, die den BeurteilerInnen vorgegeben werden. Sie sind daher das zentrale Referenzwerk der Beurteilung und erhöhen durch ihre Anwendung die Reliabilität und Validität der Beurteilung.

Zudem zeigte sich, dass – wie einleitend bereits angedeutet – alle vier an der Studien beteiligten Beurteilerinnen in den retrospektiven Interviews unabhängig voneinander vorbrachten, das introspektive Verfahren als gewinnbringend für sich selbst wahrzunehmen, weil es sie dazu brachte, das eigene Verhalten, die eigenen Vorlieben und Befindlichkeiten bewusst zu machen und zu reflektieren. Trotz aller Grenzen des Verfahrens (s. dazu 3.2) bietet Introspektion die Möglichkeit zur Selbstreflexion, zur Eruierung individuell geprägter Strategien und nicht zuletzt zur Rekonstruktion subjektiver Theorien, die der Beurteilungsarbeit zugrunde liegen, denn Beurteilungsstrategien sind bewusstseinsfähig. Diese Fähigkeit ist ein wichtiger Faktor bei der Professionalisierung. Dies hat zunächst vor allem die Handlungsforschung formuliert[11].

Eine Konsequenz der Untersuchung war somit die Überlegung, inwiefern das Untersuchungsverfahren für Schulungen nutzbar gemacht werden kann.

Im Folgenden nun sollen Erkenntnisse aus einer Pilotstudie berichtet werden, die das Laute Denken als Trainingsmethode entwickelt und in Beurteilungsschulungen im Kontext TestDaF implementiert hat.

3 Pilotstudie zur Integration des Lauten Denkens in Beurteilungsschulungen

3.1 Vorgehen

BeurteilerInnen, die schriftliche und mündliche Leistungen von TestDaF-Prüfungsteilnehmenden bewerten, müssen einmal im Jahr an einer sogenannten Kalibrierungssitzung teilnehmen. Im Rahmen dieser Schulungen werden Teilnehmerleistungen beurteilt, die sich von den sonst üblichen Texten unterscheiden und somit z. T. Schwierigkeiten bereiten. Dazu gehören u. a. heterogene Leistungen, also Leistungen, die beispielsweise inhaltlich gut, aber sprachlich fehlerhaft sind, aber auch solche Leistungen, die zunächst allein durch externe Faktoren (unleserliche Handschrift, Korrekturen im Text, ungewohnte Stimmlage) Einfluss auf die Beurteilung haben können.

Durch die Diskussion in Kleingruppen erhalten die BeurteilerInnen Rückmeldung zu ihrem individuellen Bewertungsverhalten und können sich so in der Gruppe bezüglich ihrer unterschiedlichen Beurteilungsmaßstäbe verorten.

Daher war die erste Überlegung, das Laute Denken während einer solchen Schulung einzusetzen, indem z. B. die BeurteilerInnen in Paaren arbeiten: eine Person beurteilt, eine andere notiert Auffälligkeiten und bespricht diese danach mit der Person, die die Beurteilung vorgenommen hat. Problematisch bei diesem

Vorgehen erschienen jedoch mehrere Aspekte: Zum einen wäre für dieses Verfahren ein enormer Zeitaufwand notwendig gewesen, v. a. bei einer so großen Personengruppe – im Durchschnitt nehmen etwa 40 Personen an einer solchen Schulung teil.

Zum anderen würde dadurch die Rückmeldung dann nicht mehr zentral gesteuert, sondern auf der Grundlage der Beobachtungen Einzelner erfolgen, was ggf. ein Validitätsproblem darstellen könnte. Des Weiteren ist fraglich, ob sich die BeurteilerInnen dazu bereit erklärt hätten, das Verfahren im Beisein anderer durchzuführen.

Deshalb wurde alternativ ein anderes Verfahren gewählt, bei dem die BeurteilerInnen im Vorfeld der Schulung Leistungen zur Beurteilung erhalten und das Laute Denken für sich alleine durchführen. Jede Person erhielt zwei zu beurteilende Leistungen, wovon eine analog zum Vorgehen bei Arras (2007) als Übung zum Vertrautmachen mit dem Verfahren diente.

Zusätzlich wurde ein Begleitschreiben verschickt, in dem die Gründe für das Verfahren dargelegt und den BeurteilerInnen ihre Aufgabe näher erläutert wurde. Die BeurteilerInnen erhielten außerdem eine Leerkassette für die Aufnahme sowie einen Fragebogen, der retrospektiv – und nach Anhören des Lauten Denkens – ausgefüllt werden sollte. Der Fragebogen erhob Daten zum Verfahren selbst, zu den Problemen, die bei der Beurteilung auftraten sowie zum subjektiv wahrgenommenen Nutzen des Verfahrens für die eigene Beurteilungsarbeit.

Insgesamt wurden an 111 geschulte Beurteiler Unterlagen für das Laute Denken versendet. Der Rücklauf lag mit 86 zurückgesendeten Beurteilungsunterlagen und Fragebogen bei knapp 80 %. Nur rund 70 % der beteiligten Personen nahmen sich aber selbst beim Lauten Denken auf und werteten diese Aufnahme mithilfe des Fragebogens aus[12], darunter waren 13 BeurteilerInnen, die in der Regel nur Leistungen aus dem Mündlichen Ausdruck beurteilen. Diese Personen waren zwar in der Beurteilung schriftlicher Leistungen geschult, hatten aber teilweise jahrelang keine Beurteilungen für diesen Prüfungsteil mehr vorgenommen.

Die Fragebogen und Aufnahmen des Lauten Denkens wurden anschließend im TestDaF-Institut ausgewertet.

3.2 Ergebnisse

Bei der Auswertung der Fragebogen standen folgende Aspekte im Vordergrund: Wie waren die einzelnen BeurteilerInnen während des Lauten Denkens dem für sie ungewohnten Verfahren gegenüber eingestellt? Was fiel ihnen schwer? Wo gab es Probleme bei der Anwendung der Kriterien? Sowie die Frage danach, ob das Verfahren als hilfreich und für Schulungen generell einsetzbar betrachtet werden kann.

Dabei zeigte sich, dass die BeurteilerInnen dem Verfahren zunächst skeptisch gegenüberstanden, im Laufe des Vorgehens jedoch diese Vorbehalte ablegten. Dennoch betonten viele, dass sie sich durch das Verfahren verunsichert fühlten. So beschrieb eine Beurteilerin ihre Einstellung gegenüber dem Lauten Denken während der Anwendung des Verfahrens folgendermaßen: „vorher: Abwehr! Dann ging es leichter als erwartet, aber ich fühlte mich die ganze Zeit beobachtet".

Die geäußerte Verunsicherung mag auch damit zusammenhängen, dass eine Beurteilung unter den Bedingungen einer Introspektion ein ungewohntes Verfahren ist, das sich stark von der Vorgehensweise einer Beurteilung unter normalen Bedingungen unterscheidet.

So hat das Verfahren Grenzen, zu denen gehört, dass sich die BeurteilerInnen auf mehrere Aspekte gleichzeitig konzentrieren mussten – was vielen schwer fiel: den Text laut lesen, alle Handlungen und Gedanken verbalisieren, Notizen auf dem Bewertungsbogen machen und die Bewertungskriterien heranziehen. Dadurch fühlten sich viele von der eigentlichen Beurteilung abgelenkt. Als weitere Einschränkung in Bezug auf die Validität des Laut-Denken-Verfahrens merkten einige BeurteilerInnen an, dass sie Schwierigkeiten hatten, alle Gedanken zu verbalisieren. Aus diesen Gründen ist das Laute Denken auch nur schwer für die Beurteilung mündlicher Leistungen anzuwenden, da BeurteilerInnen nicht gleichzeitig sprechen und die Äußerung eines Kandidaten hören und beurteilen können.

Hinsichtlich der Anwendung der Beurteilungskriterien wurde wiederholt geäußert, dass einige Kriterien schwer voneinander zu trennen seien. Dies bezog sich v. a. auf die Kriterien Lesefluss und Gedankengang.

Diese holistischen Kriterien beziehen sich auf die Gesamtwirkung des Textes, ohne dabei einzelne Aspekte (z. B. wie häufig syntaktische oder lexikalische Fehler auftreten) zu analysieren.

Da sich beide Kriterien auf das Verstehen des Textes beziehen, ist die Zuordnung einzelner Textstellen, an denen das Verständnis beeinträchtigt ist, zu einem der beiden Kriterien erfahrungsgemäß schwierig.

Auch in Bezug auf die Formulierung der einzelnen Deskriptoren wurden Probleme deutlich: Wie lässt sich beispielsweise rein quantitativ ein „breites Spektrum" im Wortschatz von einem „begrenzten Spektrum" abgrenzen?[14].

Insgesamt wertete ein Drittel der BeurteilerInnen, die das Verfahren selbst durchgeführt hatten, das Laute Denken als nicht hilfreich für ihre Beurteilungsarbeit[15]. Gründe dafür waren vor allem die bereits o. g. Konzentrationsprobleme sowie ein erhöhter Zeitaufwand. Mehr als die Hälfte betrachtete die Introspektion jedoch als nützliches Verfahren bei der Beurteilung. Dies wurde v. a. damit begründet, dass die Beurteilung stärker reflektiert würde, wie folgende Aussagen belegen:

> „Man reflektiert das eigene Verhalten stärker. ‚Bauchentscheidungen' werden minimiert."

> „Man macht sich Gedanken, revidiert nochmals eine Einstufung, weil man gezwungen ist, die Begründung zu formulieren und man geht nicht zu sehr ‚nach Gefühl'."

Die restlichen BeurteilerInnen haben die Nützlichkeit des Verfahrens nicht einheitlich bewertet. So merkte eine Beurteilerin auf die Frage, ob das Verfahren für sie hilfreich in Bezug auf ihre Beurteilungsarbeit gewesen sei, an:

> „Auswertung könnte hilfreich sein, wenn man mir z. B. sagt: ‚an dieser Stelle haben Sie „falsch" gedacht'"

Sicherlich ist eine solche individuelle Rückmeldung zum Beurteilungsverhalten wünschenswert und würde den persönlichen Nutzen sowie die Akzeptanz des Verfahrens zusätzlich erhöhen, bei einer Zahl von über hundert BeurteilerInnen ist sie jedoch nur mit großem Aufwand zu realisieren.

Neben den Reflexionen der BeurteilerInnen in den Fragebogen, wurden auch die Laut-Denken-Aufnahmen im TestDaF-Institut ausgewertet.

Diese Auswertung ermöglichte es, den Weg vom Lesen des Textes zur Einstufung der Leistung nachvollziehbar zu machen. In der Regel stehen der Institution lediglich die Beurteilungsbogen mit stichwortartig festgehaltenen Begründungen für die Einstufungen zur Verfügung. Welche Entscheidungen aber einer Einstufung vorausgingen, ist nicht erkennbar. Die Aufnahmen ermöglichen also – unter Berücksichtigung der oben genannten Grenzen des Verfahrens – einen Blick „in den Kopf" der BeurteilerInnen zu werfen und den Beurteilungsweg transparent zu machen.

Bei der Auswertung der Aufnahmen wurde u. a. deutlich, dass sich einige der BeurteilerInnen nicht an das vorgeschriebene Verfahren zur Beurteilung hielten[16]. So lautet die Vorgabe des TestDaF-Instituts, die schriftliche Leistung nach einem ersten Lesen des Textes zunächst anhand der drei holistischen Kriterien einzustufen (*Lesefluss, Gedankengang und Textaufbau*). Erst beim zweiten (und ggf. dritten) Lesen des Textes wird die Leistung anhand der sechs analytischen Kriterien zur inhaltlichen Umsetzung der Aufgabe und sprachlichen Realisierung beurteilt. [17]

Äußerungen wie die folgenden belegen, dass dieses Verfahren nicht konsequent eingehalten wurde:

> „[Der Gedankengang] Ist so 4 bis 5. Ich muss das nachher beurteilen, wenn ich die Grafik beurteile."

> „[Textaufbau] Mit Fragezeichen 3. Also zwischen 3 und 4. Ich werde das dann später noch mal ansehen. So beim nochmaligen Lesen wird das dann vielleicht deutlicher."

In einem Fall wurde die Leistung beim Textaufbau erst nach der Beurteilung anhand der anderen acht Kriterien eingestuft:

> „Ja, das letzte, wo ich mich entscheide, ist der Textaufbau. Ob ich den mit 3 oder unter 3 bewerte. 3 – ‚Text weist Brüche auf', und unter 3 – ‚Text ist nicht klar strukturiert'. Okay, dann erscheint mir das doch eher eine 3 zu sein."

Wie schon bei der Auswertung der Fragebogen, so wurde auch in den Aufnahmen deutlich, bei welchen Kriterien BeurteilerInnen vermehrt Anwendungsprobleme hatten. Sei es durch die vagen Formulierungen der Deskriptoren oder durch unklare Abgrenzungen zwischen den Niveaustufen:

> Einstufung des Kriteriums Argumentation:
> „[Zitat aus dem Teilnehmertext] Okay das ist n ja in Ansätzen ein Argument, aber ähm [räuspern] es wird ja gar nicht begründet. Ja doch es wird dann begründet: [Zitat aus dem Text] Aber ich denke das ist insgesamt, ja, nee, ich denke dann also Argumentation nicht unter 3 sondern 3. Also in Ansätzen ist ja da. Ähm, tja, ja das ist jetzt hier sind natürlich wieder die Kriterien recht vage. Für 3 im argumentativen Teil werden *Standpunkte/ Überlegungen deutlich und ggf.* das könnte hier dann genauso zutreffen [unverständlich] werden *Standpunkte nicht oder nur in Ansätzen verdeutlicht.* Das ist dann ja also: Sie werden deutlich/sie werden in Ansätzen deutlich. Wo ist da der Unterschied? Das ist die übliche Schwierigkeit mit den Bewertungskriterien. So Argumentation also dann 3, weil hier doch wenigstens eine persönliche Wertung deutlich wird. Okay."

Des Weiteren fielen bei der Auswertung der Aufnahmen Aussagen auf, die deutlich machen, dass auch subjektive Theorien bei der Urteilsfindung einbezogen und Einstufungen nicht ausschließlich mithilfe der Deskriptoren getroffen wurden – ein Befund, der sich ebenfalls bei Arras (2007: 442ff.) findet.

Die Hochschultauglichkeit wird oftmals vor dem Hintergrund der beruflichen Erfahrung in wissenschaftspropädeutischen oder studienbegleitenden Kursen interpretiert und nicht in Bezug auf das Testkonstrukt, wobei diese antizipierten Anforderungen auch über das zu erwartende Niveau hinausgehen können, wie folgende Äußerung eines Beurteilers belegt:

> „Punkte der Aufgabenstellung: Entspricht nicht einem hochwissenschaftlichen Text, auf keinen Fall."

Von keinem TestDaF-Prüfling wird aber für eine Einstufung auf TDN 5 ein Text erwartet, der den kulturspezifischen wissenschaftssprachlichen und fachlichen Kriterien im deutschen Hochschulkontext genügt. Diese Leistungsbeschreibung läge über dem erwarteten Niveau und deckt sich nicht mit dem Deskriptor. Zudem erlaubt die Aufgabenstellung und die Bearbeitungszeit von 60 Minuten keine Umsetzung, die diesen Ansprüchen genügen würde, da sie so konzipiert ist, dass Prüflinge aus ganz unterschiedlichen Fachrichtungen ihre Kompetenzen unter Beweis stellen können. An einer bestimmten Disziplin ausgerichtete Konventionen können hier nicht angewendet werden.

Diese Erkenntnisse aus den Aufnahmen liefern wichtige Hinweise für die regelmäßigen Schulungen der TestDaF-BeurteilerInnen. So wird bei kommenden Erstschulungen und den jährlichen Kalibrierungsveranstaltungen verstärkt auf die Einhaltung des Beurteilungsverfahrens geachtet, um das Prozedere möglichst einheitlich zu halten. Kriterien, bei denen deutlich wurde, dass Schwierigkeiten in der Anwendung auftreten, werden eingehender diskutiert und anhand von Beispielen veranschaulicht.

3.3 Weitere Auswertungen

Neben den Erkenntnissen zu den individuellen und gemeinsamen Beurteilungsstrategien, den Erkenntnissen zum Beurteilungsverfahren, zu den Schwierigkeiten bei der Beurteilung sowie zur Erwartungshaltung der BeurteilerInnen hinsichtlich der Leistungen, die im Zentrum des Forschungsinteresses standen, ermöglichen die Aufnahmen aber auch weitere Analysen. So lässt sich auswerten, wie viel Zeit verschiedene BeurteilerInnen für die Bewertung desselben Textes benötigen. Die Zeitspanne der Beurteilungen variierte in der Studie enorm[19]. Eventuell lassen sich aus der benötigten Zeit Rückschlüsse auf die Erfahrung als Beurteiler von TestDaF-Prüfungsleistungen ziehen. Zu vermuten ist, dass erfahrenere Beurteiler, d. h. diejenigen, die gut mit dem Bewertungsverfahren, den Kriterien und Deskriptoren vertraut sind, ihre Einstufungen schneller und sicherer treffen. Eine weitere Hypothese lautet, dass die langjährigen BeurteilerInnen durch die jährlichen Kalibrierungen und die Vielzahl an vorgenommenen Beurteilungen, treffsichere Einstufungen vornehmen und weniger stark vom Mittel der Bewertungen abweichen[20].

4 Ausblick

Die Rückmeldungen der BeurteilerInnen, die an dieser Studie teilgenommen haben, waren überwiegend positiv. Insbesondere wurden die Erkenntnisse hervorgehoben, die durch die Bewusstmachung des eigenen Beurteilungsverhaltens gewonnen

wurden. Auch der (gefühlte) Zwang, seine Einstufungen zu begründen und dadurch „Bauchentscheidungen" zu vermindern, wurden als positiv bewertet[21].

Wünschenswert wäre, das Verfahren nicht wie im oben beschriebenen Sinne anzuwenden, wo jeder Beurteiler bzw. jede Beurteilerin für sich allein arbeitet und Verbaldaten dabei aufnimmt. Denn diese Vorgehensweise ermöglicht zwar die Selbstreflexion, erschwert aber individuelle Rückmeldungen der Testinstitution an die BeurteilerInnen. Rückmeldungen beschränken sich daher im Wesentlichen auf allgemeine Erkenntnisse und Probleme.

Kleinere Institutionen könnten das Laut-Denken-Verfahren beispielsweise bei gemeinsamen Kalibrierungssitzungen der BeurteilerInnen anwenden: Eine Person bewertet laut denkend einen Text, zwei oder mehr BeurteilerInnen hören zu und notieren Auffälligkeiten.

Danach wird gemeinsam ausgewertet, welche – vielleicht bereits automatisierten Strategien – unerwünscht sind und welche effektiven Strategien weiter ausgebaut werden sollten. Auch subjektive Theorien können hierdurch bewusst gemacht und hinterfragt werden. Aus beiden Vorgehensweisen (Durchführung allein oder in der Gruppe) resultiert die Fähigkeit, das eigene Handeln zu reflektieren. Dies führt zur weiteren Professionalisierung der BeurteilerInnen und dient damit auch der Qualitätssicherung innerhalb einer Institution.

Anmerkungen

1 Bei vorliegendem Beitrag handelt es sich um die aktualisierte Fassung unseres gemeinsamen Vortrags auf der IDT 2009 in Jena/Weimar.
2 Unter Kalibrierungen verstehen wir in diesem Zusammenhang regelmäßige Beurteilertrainings, die der Qualitätssicherung dienen.
3 Zu subjektiven Theorien im Kontext Fremdsprachenlernen und -lehren, s. insbesondere Grotjahn (1998). Zu subjektiven Theorien im Kontext der hier behandelten Beurteilung von Prüfungsleistungen s. Arras (2009).
4 Eine genaue Beschreibung der Prüfung findet sich beispielsweise in Arras (2006) und auf der Internetseite des TestDaF-Instituts unter www.testdaf.de.
5 Der TestDaF differenziert drei Leistungsstufen: TestDaF-Niveaustufe 3, 4 und 5. Unterhalb von TDN 3 erfolgt keine weitere Differenzierung. Ein Ergebnis unterhalb von TDN 3 besagt lediglich, dass das für den Hochschulkontext erforderliche sprachliche Eingangsniveau nicht erreicht wurde. Die Verortung am Gemeinsamen europäischen Referenzrahmen (s. o.) zeigt, dass die TestDaF-Niveaustufen das Leistungsspektrum B2-C1 abdecken. TDN 5 ist im oberen C1-Bereich angesiedelt, TDN 3 hingegen deckt den unteren B2-Bereich ab. TDN 4 liegt zwischen den beiden Kompetenzstufen B2 und C1; s. auch Kecker/Eckes (erscheint).
6 Eine detaillierte Darstellung des Beurteilungsverfahrens bei schriftlichen Leistungen im Kontext TestDaF findet sich in Arras (2007: 68ff.).
7 Auf der Internetseite des TestDaF-Instituts sind Modellsätze einsehbar, die den Aufbau sowie das Format der einzelnen Prüfungsteile zeigen. TestDaF-Musterprüfungen sind zudem im Handel erhältlich (Hueber-Verlag).
8 Zu den Anforderungen s. ausführlich Arras (2007: 41ff.).
9 Vorgeschaltet war die Beurteilung eines Textes als Übung, um die vier Beurteilerinnen mit dem Verfahren des Lauten Denkens vertraut zu machen und mögliche Hemmnisse zu überwinden. Diskutiert wird das Verfahren u. a. bei Arras (2007), Green (1998) und insbesondere auch Lumley (2005).
10 Zu diesen Strategien gehört beispielsweise die Konstituierung einer Erwartungshaltung aufgrund äußerer Faktoren (wie z. B. Handschrift, Korrekturen im Text), oder aber auch der ständige Abgleich der Leistung mit den Deskriptoren (v. a. bei nicht eindeutigen Leistungen). Andere Strategien basieren auf individuellen Faktoren, insbesondere (berufliche und kulturelle) Erfahrungen, damit zusammenhängend subjektive Theorien. S. dazu Arras (2007: 217ff.).
11 S. vor allem Altrichter/Posch (1998) sowie für den in der anglophonen Welt verwendeten Begriff action research Reason/Bradbury (2002). Der Begriff des „reflektierten Praktikers" weist darauf hin, dass eine Person, eine Lehrkraft oder wie im vorliegenden Fall eine Beurteilerin oder ein Beurteiler, ihr bzw. sein Handeln reflektiert und begründet, ggf. auch begründet revidiert.
12 Dies war zum großen Teil durch die technischen Schwierigkeiten wie z. B. die Beschaffung eines geeigneten Mikrofons zur Aufnahme mit dem Kassettenrekorder oder am PC bedingt.
13 Die Verbalisierung auch von sichtbaren Handlungen (beispielsweise das Notieren von Ergebnissen auf dem Beurteilungsbogen oder die Heranziehung der skalierten Deskriptionen) war notwendig, da nur Tonaufnahmen gemacht wurden und keine Vertreterin des TestDaF-Institut bei der Beurteilung anwesend war und somit keine Beobachtungen bzw. Auffälligkeiten notiert werden konnten.
14 Diese Wahrnehmungen stützen Befunde der Studie von Arras, in der die Beurteilerinnen ebenfalls auf Bewertungsschwierigkeiten hinweisen, die sich aus der Struktur sowie der sprachlichen und inhaltlichen Gestaltung der Beurteilungskriterien ergeben Arras (2007: 400ff.)
15 Von den insgesamt 13 Personen, die ausschließlich mündliche Leistungen beurteilen, wertete erwartungsgemäß mehr als die Hälfte das Verfahren als nicht dienlich für eigene Beurteilungen.
16 Dieser Befund weicht ab von den Ergebnissen der Studie von Arras (2007). Die vier Beurteilerinnen dort nahmen die Beurteilungen weitgehend nach dem vorgegebenen Verfahrensweg vor. Dies kann damit begründet werden, dass Arras selbst bei der Durchführung des Lauten Denkens anwesend war und somit als Kontrollinstanz seitens der Beteiligten wahrgenommen wurde.
17 Vgl. Arras (2007, S. 90 ff.).
18 S. hierzu die Diskussion bei Arras (2007: 42ff.)
19 Dabei ist aber zu beachten, dass viele Beurteiler rückmeldeten, die Einstufungen des Textes hätte mehr Zeit in Anspruch genommen als beim üblichen Verfahren. Das mag damit zusammenhängen, dass die Beurteiler ihre Gedanken verbalisieren und strukturieren mussten und sich gezwungen sahen, ihre Einschätzungen zu begründen (s. o.).
20 Dies lässt sich im TestDaF-Institut aber auch anhand der Vergleichsbeurteilungen untersuchen, die BeurteilerInnen mit jedem Korrekturpaket erhalten. Es handelt sich dabei um drei Teilnehmertexte, die alle an einem Testereignis Beteiligten beurteilen müssen und deren Einstufungen wesentlichen Einfluss auf die Berechnung der Strenge bzw. Milde eines Beurteilers bzw. einer Beurteilerin hat.
21 Einschränkend muss allerdings darauf hingewiesen werden, dass der zeitliche Aufwand für die Introspektion sowohl für die BeurteilerInnen selbst als auch für die auswertende Institution groß ist.

Literaturverzeichnis

Altrichter, H./Posch, P. (1998): Lehrer erforschen ihren Unterricht. Eine Einführung in die Methoden der Aktionsforschung. 3., durchgesehene und erweiterte Auflage. Bad Heilbrunn: Klinkhardt.

Arras, U. (erscheint): „What's on a rater's mind? Die Erforschung von Beurteilungsstrategien und ihr Bewusstmachung durch Schulungsmaßnahmen als Voraussetzungen für die Testvalidität". In: Dokumentation der AILA 2008 in Essen, ZfAL.

Arras, U. (2009): „Subjektive Theorien als Faktor bei der Beurteilung von Prüfungsleistungen". In: Berndt, A./ Kleppin, K. (Hrsg.): Sprachlehrforschung: Theorie und Empirie. Festschrift für Rüdiger Grotjahn. Frankfurt: Peter Lang: 169-179..

Arras, U. (2009): „Wie es zu einer Beurteilung kommt. Ein Forschungsbericht zu Strategien bei der Beurteilung schriftlicher Leistungen im Kontext der Prüfung TestDaF". In: Hunstiger, A./Koreik, U. (Hrsg.): Chance Deutsch: Schule - Studium – Arbeitswelt. Beiträge der 34. Jahrestagung DaF 2006. Materialien Deutsch als Fremdsprache, Band 78. Göttingen: Universitätsverlag: 179-196.

Arras, U. (2007): Wie beurteilen wir Leistung in der Fremdsprache? Strategien und Prozesse bei der Beurteilung schriftlicher Leistungen in der Fremdsprache am Beispiel der Prüfung Test Deutsch als Fremdsprache (TestDaF). Giessener Beiträge zur Fremdsprachendidaktik. Tübingen: Narr.

Arras, U. (2006): Der TestDaF. Konzept und Prinzipien des standardisierten Tests Deutsch als Fremdsprache. In: Fòrum – Anuari de l'Associació de Germanistes de Catalunya, No. 12. Akten des sechsten Kongresses des Katalanischen Deutschlehrer- und Germanistenverbandes (A.G.C.) in Tarragona, April 2005: 39-52.

Eckes, T./Weiss-Motz, F./Whelan-Mostofizadeh, S. (2009): Ermittlung fairer Ergebnisse im Prüfungsteil/ Schriftliche Kommunikation des Deutschen/ Sprachdiploms. In: Deutsche Lehrer im Ausland,/ 56, 15-22.

Eckes, T. (2005): Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. In: Language Assessment Quarterly 2/3: 197-221.

Eckes, T. (2004): Beurteilerübereinstimmung und Beurteilerstrenge. Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im „Test Deutsch als Fremdsprache" (TestDaF). In: Diagnostica, 50/2: 65-77.

Europarat/Rat für kulturelle Zusammenarbeit (2001): Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen. Berlin et al.: Langenscheidt.

Grotjahn, R.: Testen im Fremdsprachenunterricht. Aspekte der Qualitätsentwicklung. In: PRAXIS Fremdsprachenunterricht, Heft 1/2009: 4-8.

Green, A. J. K. (1998): Verbal Protocol Analysis in Language Testing Research. A Handbook. Cambridge Cambridge University Press.

Grotjahn, R. (1998): Subjektive Theorien in der Fremdsprachenforschung: Methodologische Grundlagen und Perspektiven. In: Fremdsprachen Lehren und Lernen 27: 33-59.

Kecker, G./Eckes, T. (erscheint). Putting the Manual to the test: The TestDaF–CEFR linking project. In W. Martyniuk (Ed.), Linking tests to the CEFR: Case studies and reflections on using the Council of Europe's draft Manual for relating language examinations to the CEFR. Cambridge: Cambridge University Press.

Lumley, T. (2005): Assessing Second Language Writing. The Rater's Perspective. Frankfurt/Main: Peter Lang.