ZfAL
Zeitschrift für Angewandte Linguistik

## Aus dem Inhalt:



**Sonderdruck**

Heft 50 August 2009 PETER LANG

ZfAL
50/2009

INHALTSVERZEICHNIS



**Ulrike Arras**

# What's on a rater's mind? Die Erforschung von Beurteilungsstrategien und ihre Bewusstmachung durch Schulungsmaßnahmen als Voraussetzungen für die Testvalidität

The construction of valid test tasks and the reliable assessment of language competence form an important part of fair language testing. It is therefore insufficient to develop evaluation criteria which reflect the test construct, but also important to make raters aware of their assessment strategies through special trainings. To develop an awareness for assessment that helps raters to reflect and by that means provide a proof of their behaviour as well as their individual and thus subjective reasons for evaluating competences in a certain way is of central importance for test validity.
This article describes an empirical study on assessment strategies of raters of the TestDaF (Test of German as a Foreign Language). The TestDaF is a test which measures language competences of foreign students who intend to study at German higher education institutions. The research is based on case studies using introspective methods such as thinking aloud protocols and retrospective interviews. The rich data show that raters use a wide range of complex assessment strategies to get valid results. The aim of test institutions as well as teacher trainings should be to help raters to build an awareness for their individual assessment strategies to meet validity criteria.

## 1. Problemaufriss

Wenn wir schriftliche, oder auch mündliche, Kompetenzen in der Fremdsprache mit Hilfe offener Aufgabenformate messen wollen, ist es zur Sicherstellung der Testvalidität erforderlich, die Beurteilungsarbeit einer näheren Betrachtung zu unterziehen. Das bedeutet: wir müssen die Person der Bewerterin bzw. des Bewerters in den Blick nehmen, untersuchen, wie sie arbeitet, wie sie ihre Entscheidungen fällt und worauf sie diese gründet. Die Forschungslage

zeigt, dass wir noch relativ wenig darüber wissen, wie die Beurteilerinnen und Beurteiler bei ihrer Arbeit vorgehen, worauf sie achten, welche psychischen und psychosozialen Faktoren mitspielen bei der schwierigen und nicht zuletzt anstrengenden Aufgabe, Leistungen in der Fremdsprache angemessen und fair zu beurteilen. Die empirischen Forschungsarbeiten auf diesem Gebiet machen deutlich, dass die Beurteilung selbst eine black box zu sein scheint.[1] Um Licht in dieses Dunkel zu bringen, habe ich in einer empirischen Untersuchung anhand von vier Fallbeispielen mein Augenmerk auf die beurteilende Person gerichtet, indem ich anhand von vier Fallbeispielen die Prozesse und Strategien bei der Bewertung schriftlicher TestDaF-Leistungen untersucht habe. Meine Ausgangsfrage war also: „*What's on a rater's mind?*"[2]

## 2. Der TestDaF

Die Untersuchung erfolgte im Kontext der Prüfung Test Deutsch als Fremdsprache, einer seit 2001 administrierten standardisierten Sprachprüfung, die sich in erster Linie an internationale Studierende wendet, die ein Studium an einer deutschen Hochschule absolvieren möchten und dazu einen entsprechenden Sprachnachweis benötigen.[3] Der TestDaF wird zentral in Deutschland entwickelt, dezentral weltweit in lizenzierten Testzentren durchgeführt und die Ergebnisse wiederum zentral in Deutschland durch geschulte BewerterInnen sowie durch Rasch-analytische Berechnungen ermittelt.[4]
Der TestDaF misst Deutschkenntnisse auf fortgeschrittenem Niveau, das Leistungsspektrum umfasst ungefähr die Stufen des Referenzrahmens B2 und C1. Die vier Fertigkeiten, Leseverstehen, Hörverstehen, Schriftlicher Ausdruck und Mündlicher Ausdruck werden weitgehend unabhängig voneinander in separaten Subtests geprüft. Der Subtest Schriftlicher Ausdruck erfordert das Abfassen eines längeren schriftlichen Textes anhand einer komplexen Textstellungsaufgabe. Getestet werden die für den akademischen Kontext relevanten Schreibhandlungen Beschreiben und Argumentieren. Dabei sind die in einem Diagramm präsentierten statistischen Daten zu beschreiben; außerdem muss eine Leitfrage, ein Problem oder ähnliches diskutiert werden, indem Argumente für und wider eine Fragestellung vorgebracht, Fremdmeinungen paraphrasiert werden und eine eigene Stellungnahme begründet formuliert wird, auch die Darstellung eigenkultureller Phänomene ist dabei einzubringen. Der Text muss kohärent und schlüssig aufgebaut sein. Eine Vorgabe zum Textumfang existiert nicht. Für diesen Prüfungsteil stehen 60 Minuten zur Verfügung.[5]

[1] Wichtige Erkenntnisse erbrachten allerdings vor allem die neueren, auf introspektiven Verfahren basierenden Studien von *Cumming/ Kantor/ Powers* (2001/ 2002) sowie von *Lumley* (2005).
[2] Diese Leitfrage der Studie wurde inspiriert durch den Titel von *Brigitte Stemmers* Arbeit (1991) „*What's on a C-test taker's mind? Mental processes in C-test taking*". Sie geht darin den kognitiven und mentalen Prozessen bei der Bearbeitung von C-Tests nach.
[3] Eine ausführliche Darstellung des Testformats sowie der Testziele findet sich z. B. in *Arras* (2006); Modellprüfungen sind auf der Webseite des TestDaF-Instituts einsehbar: www.testdaf.de.
[4] Dieses Verfahren und seine Anwendung im Kontext des TestDaF wird beispielsweise in *Eckes* (2004) dargestellt.



## 3. Bewertungsverfahren

Das TestDaF-Institut wendet verschiedene Maßnahmen zur Qualitätssicherung bei der Beurteilung schriftlicher (und mündlicher) Leistungen an. Da es sich beim TestDaF um eine standardisierte Sprachprüfung handelt und die Gleichhaltung der Schwierigkeit gewährleistet werden muss, ist ein kriterienorientiertes Bewertungsverfahren erforderlich. Die Bewertung erfolgt daher anhand von Bewertungskriterien in Form skalierter Deskriptoren, wobei im Falle des Prüfungsteils Schriftlicher Ausdruck insgesamt neun Einzelaspekte der schriftlichen Leistung einer Niveaustufe zugeordnet werden müssen. Die neun Einzelaspekte sind gleichgewichtet. Die Bewertung erfolgt sowohl über holistische als auch über analytische Kriterien.[6] Die endgültige Stufenzuordnung erfolgt im TestDaF-Institut, indem anhand der neun Einzelbeurteilungen unter Anwendung der bereits genannten Multifacetten-Rasch-Analysen ein fairer Durchschnitt ermittelt wird.[7] Daneben stellt das TestDaF-Institut für jedes Testereignis aufgabenspezifische Kalibrierungen in Form von musterhaften Beurteilungen authentischer Leistungen zur Verfügung, an denen sich die BeurteilerInnen orientieren. Dies ist erforderlich, da die Aufgaben selbst zwar standardisiert sind, dennoch bei unterschiedlichen Aufgaben Varianz hinsichtlich Komplexität und Anforderungen nicht ausgeschlossen werden kann. Um die Gleichhaltung der Schwierigkeit zu gewährleisten, ist es unumgänglich, die Anforderungen aufgabenspezifisch festzulegen bzw. zu justieren, also ein testsatzspezifisches *bench marking* vorzunehmen. Dies erfolgt anhand

[5] Eine ausführliche Darstellung der Aufgabenstellung sowie der Anforderungen s. vor allem *Arras* (2007).
[6] Zur Diskussion holistisch versus analytisch s. *Arras* (2007).
[7] S. exemplarisch *Eckes* (2004) und *Eckes* (2005b).

von Leistungen aus der Erprobungsphase der Aufgabe. Hierbei werden im TestDaF-Institut in einem Gremium Leistungen unterschiedlichen Niveaus gemeinsam bewertet und als Modell für die verschiedenen Niveaustufen den BeurteilerInnen zur Verfügung gestellt. Nur so kann gewährleistet werden, dass der Prüfungsteil Schriftlicher Ausdruck über verschiedene Testereignisse hinweg weitgehend gleich schwierig bleibt. Denn die Schwierigkeit einer Aufgabe ist nicht allein an den Schwierigkeitsfaktoren der Aufgabenstellung selbst festzumachen. Vielmehr entsteht die Gesamtschwierigkeit stets aus beiden Aspekten: Anforderungen der Aufgabenstellung und Bewertungsmaßstäbe. Eine komplexe Aufgabe kann durch die Justierung der Bewertungsmaßstäbe dasselbe Schwierigkeitsniveau erhalten wie eine Aufgabe mit weniger komplexen Anforderungen, bei der die Bewertungsmaßstäbe angehoben werden. Vor diesem Hintergrund ist also eine aufgabenspezifische Justierung und Kalibrierung der Maßstäbe notwendig.

Alle BewerterInnen schriftlicher (und mündlicher) TestDaF-Leistungen durchlaufen ein System an Schulungen und Kalibrierungen, dessen Ziel es ist, ein hohes Maß an Interrater-Reliabilität zu gewährleisten, indem innerhalb der *peer group* die Beurteilungsmaßstäbe operationalisiert und ihre Anwendung trainiert werden. Damit bestehen die Maßnahmen zur Qualitätssicherung der Beurteilungen neben den rasch-analytischen Berechnungen zum einen aus aufgabenspezifischen Kalibrierungen und zum anderen aus regelmäßigen Beurteilungsschulungen. Außerdem erhalten die BewerterInnen Rückmeldungen zu ihrem individuellen Strenge- und Konsistenzmaß.[8]

## 4. Das Untersuchungsdesign

Kern der empirischen Untersuchung zum Rater-Verhalten sind Laut-Denken-Protokolle: Die vier Versuchspersonen bewerteten acht schriftliche Leistungen unter Laut-Denken-Bedingungen.[9] Die acht schriftlichen Leistungen aus dem Prüfungsteil „Schriftlicher Ausdruck", die im Kontext der Prüfung TestDaF entstanden waren (diese Prüfung wurde nach ihrem Einsatz 2005 vom TestDaF-Institut als Modellprüfung 1 veröffentlicht), sind von unterschiedlicher Qualität und unterschiedlicher soziokultureller Herkunft. Bei den vier Versuchspersonen handelt es sich um erfahrene, wiederholt geschulte

[8] S. *Eckes* (2004: 511f.), s. auch die Diskussion bei *Elders* et al. (2005) zu Auswirkungen von individuellen Rückmeldungen hinsichtlich des Rater-Profils.

[9] Zum Verfahren s. *Arras* (2007) sowie *Green* (1998), s. auch *Lumley* (2005).



TestDaF-Bewerterinnen, die sich aus Interesse für die Untersuchung zur Verfügung stellten. Da das Laut-Denken-Verfahren für die Versuchspersonen ungewohnt war, wurde eine Übungsphase vorgeschaltet, bei der ein Text zu Übungszwecken unter Laut-Denken-Bedingungen bewertet wurde.

Zur Datentriangulation wurden zudem retrospektive (semistrukturierte) Interviews am Tag nach der Bewertungsarbeit durchgeführt, um gezielt nach Phänomenen, Schwierigkeiten und Problemlösestrategien fragen zu können. Die introspektiven Daten der Protokolle sowie der Interviews wurden transkribiert, die Laut-Denken-Protokolle außerdem segmentiert und kodiert, um sie einer Analyse zugänglich zu machen.

Der Untersuchung vorgeschaltet waren außerdem zwei kleinere Vorstudien, bei denen u. a. das Laut-Denken-Verfahren erprobt wurde, um die Angemessenheit des Verfahrens zu prüfen und das Design der Studie zu optimieren.

Leitfragen der Untersuchung waren vor allem:

- Wie gehen die Bewerterinnen bei der konkreten Bewertungsarbeit vor?
- Welche Prozesse durchlaufen die Versuchspersonen bei der Bewertungsarbeit?
- Welche Faktoren determinieren dabei ihr Vorgehen?
- Wie werden die Beurteilungsinstrumente, vor allem die Bewertungskriterien in Form skalierter Deskriptoren eingesetzt? Mit anderen Worten: Wie erfolgt deren Operationalisierung?
- Welche individuell und interindividuell relevanten Strategien sind beobachtbar?

Die der Untersuchung übergeordnete Frage lautete: Welche Maßnahmen können Testinstitutionen treffen, um die BewerterInnen angemessen und effektiv zu schulen, so dass sie in der Lage sind, valide Leistungsbeurteilungen vorzunehmen? Dieses Anliegen entstammte meiner langjährigen Aufgabe im TestDaF-Institut, neben der Entwicklung von Testaufgaben, auch die BeurteilerInnen von TestDaF-Prüfungsleistungen zu schulen und zu betreuen.

## 5. Einige zentrale Befunde

Im Folgenden sollen einige Befunde referiert werden, die nicht zuletzt für die Frage nach geeigneten Maßnahmen zur Schulung von BeurteilerInnen von Bedeutung sein können.

Insgesamt betrachtet erweist sich die Beurteilungsarbeit als hochkomplexe Handlung, die sich aus verschiedenen, sich ergänzenden und aufeinander aufbauende Einzelhandlungen prozesshaft gestaltet. Die Versuchspersonen bedienen sich bei der Beurteilungsarbeit aus einem ganzen Arsenal an Strategien und Strategiencluster, wobei diese Strategien zum einen von institutionellen Faktoren geprägt sind. Institutionelle Faktoren sind insbesondere die Bewertungskriterien, also die in einem Kriterienraster präsentierten skalierten Deskriptoren. Sie erweisen sich als Herzstück der Beurteilungsarbeit. Ihre zentrale Rolle zeigt sich nicht zuletzt in der Häufigkeit, mit der auf sie referiert wird, sei es, indem sie wörtlich zitiert werden oder paraphrasiert präsent sind. Sie sind weitgehend internalisiert. Zum anderen basieren die Strategien auf individuellen Faktoren, insbesondere Erfahrungen, damit zusammenhängend Erwartungen und subjektive Theorien. Die meisten beobachteten Strategien und Strategiencluster gründen sowohl auf institutionelle als auch individuelle Faktoren.[10] Dies sei am Beispiel des Testkonstrukts erläutert, welches die Beurteilungsarbeit prägt. Einerseits ist das Testkonstrukt institutionell definiert: Das TestDaF-Institut beschreibt und exemplifiziert das Testkonstrukt für die BeurteilerInnen in den Bewertungsrichtlinien, in den kalibrierten Beurteilungen, die zu jedem Testereignis aufgabenspezifisch entwickelt werden (s. o.), durch Schulungen usw. Andererseits verfügen die BeurteilerInnen offensichtlich, das zeigen die Daten sehr deutlich, über individuell recht klare Vorstellungen beispielsweise darüber, was eine Studentin oder ein Student an schriftlichen Kompetenzen mitbringen muss, um (erfolgreich) an einer deutschen Hochschule zu studieren. Diese Vorstellungen gehen v. a. auf eigene Erfahrungen zurück, und zwar sowohl hinsichtlich der eigenen Studienerfahrung und der eigenen akademischen Sozialisation, als auch hinsichtlich beruflicher Erfahrungen im Kontext Deutsch als Fremdsprache, denn bei den TestDaF-BeurteilerInnen handelt es sich um erfahrene DaF-Lehrkräfte, die über Lehr- und damit auch Prüfungserfahrung mit ausländischen Studierenden an deutschen Hochschulen und Bildungseinrichtungen verfügen. Neben den Beurteilungskriterien scheint mithin das Testkonstrukt die wichtigste Arbeitsgrundlage bei der Beurteilung zu sein.

Zusammenfassend kann festgehalten werden, dass es sich bei der Beurteilungsarbeit um einen hochkomplexen Prozess handelt, der zum einen auf institutionell geprägten Strategien (beispielsweise der genannte Rekurs auf das

[10] Eine Darstellung und Systematisierung der im Rahmen der Studie beobachteten Strategien und Strategiencluster findet sich in *Arras* (2007) und *Arras* (erscheint).



Beurteilungsinstrumentarium), zu einem ganz erheblichen Teil jedoch auch auf individuell geprägten Strategien und Strategiencluster basiert.

## 6. Schlussfolgerungen: *assessment awareness* dank Introspektion

Diese Beobachtung zusammen mit der Erkenntnis, dass Beurteilungsstrategien bewusstseinsfähig sind und den Versuchspersonen im Rahmen der Studie durch das Laut-Denken-Verfahen zugänglich gemacht wurden, führt zur Frage, inwiefern introspektive Verfahren für Schulungszwecke nutzbar gemacht werden können. Denn die Versuchspersonen geben in den retrospektiven Interviews an, dass ihnen das Verfahren geholfen hat, sich ihr eigenes Verhalten und ihre Handlungen bewusst zu machen, zu begründen und nachvollziehen zu können. Damit sind die Beurteilungsstrategien und –prozesse nicht allein aus der Außenperspektive zugänglich gemacht worden, also aus der Sicht der Forscherin, sondern auch den Versuchspersonen selbst sind die Handlungen, die psychischen und auch physischen Befindlichkeiten metakognitiv zugänglich und somit analysierbar gemacht worden. Einige Ausschnitte aus den Interviews mögen diese Selbstbeobachtung belegen:

Versuchsperson K7 gibt im retrospektiven Interview an: (RIK7, ab ca. 0:15):

> Man weiß gar nicht, was einen alles beeinflusst. Das ist im Unterbewusstsein, das ist schwierig. Deshalb ist das auch mit dem Laut-Denken. Man weiß gar nicht, dass man das denkt, eigentlich.

Am Ende des Interviews greift sie den Gedanken erneut auf (RIK7, ab ca. 47:20):

> Ah ich danke dir auch, das war sehr interessant. Ja, da lernt man schon ne Menge, denn ich mein, wenn man sonst normalerweise dasitzt, man überlegt sich zwar schon, aber man reflektiert ja nicht so genau, warum hab ich das jetzt so gesagt und eh. Wenn ich so für mich denke, na ja an sich könnte man ne 3, was bedeutet das, ne? Also das ist schon. Oder dass man eben genau reflektiert: Warum mach ich jetzt das, warum mach ich das? Seine Mechanismen ein bisschen hinterfragt.

Auch Versuchsperson K6 macht dank des Laut-Denken-Verfahrens Beobachtungen an sich selbst und ihrer Beurteilungsarbeit. Auf die Frage, ob sie das Verfahren gestört habe, gibt sie an (RIK6, ab ca. 0:35):

> Überhaupt nicht, also ich hab mich so gefühlt wie immer eigentlich. Also am Anfang bei diesem Null-Durchgang, da war so ne Hemmschwelle, weil es etwas ungewohnt war, Gedanken auszusprechen bzw. überhaupt sich bewusst zu werden, was man sich da für Gedanken macht. Oder es gibt ja verschiedene Speichermöglichkeiten, was da als Hintergedanken noch so verborgen war. Aber dann war es eigentlich so wie immer, ne?

K5 schließlich gibt folgende Beobachtungen zu Protokoll (RIK5, ab ca. 0:30):

> Also nervig fand ich das nicht. Das war in Ordnung. Mit ist nur aufgefallen, dass ich offensichtlich mehr denke, wenn ich laut denke. Also hatte ich jetzt ganz subjektiv mal den Eindruck. Also ich hab das Gefühl, [.] ich zerpflücke die Texte nicht so analytisch, wenn ich eh alleine bin, erstens Mal. Das macht was aus. Obwohl ich das jetzt nicht störend empfunden hab. Das hab ich nicht als Kontrolle empfunden, dass du da gesessen hast. Aber irgendwie ist die Rechtfertigung, der Rechtfertigungszwang größer, sag ich mal, wenn man da mit jemandem sitzt, dem man das noch dazu erklären soll, wie man zu der Meinung kommt, als wenn man da alleine sitzt. Ich glaub da dreht sich das dann eher so um kritische, wirklich um die kritischen Punkte. Da würd ich dann nur begründen, ehm, an einzelnen Stellen, also wo ich, wo ich große Abweichungen sehe oder wo ich wirklich überhaupt nicht klar komme. Aber da würd ich auch schneller zu [.] Ergebnissen kommen. Und eh ja aber ich fand das nichtsdestotrotz fand ich das ganz interessant. Also weil wenn man das mal laut hört, dann wird es einem halt auch bewusster, was man tut.

Die Daten zeigen ein hohes Maß an Reflexionsfähigkeit seitens der Versuchspersonen. Alle vier kommentieren, dass ihnen das Verfahren des Lauten Denkens bewusst gemacht hat, welchen Prozessen und Einflüssen sie selbst bei der Urteilsfindung unterworfen sind. Wenn wir davon ausgehen, dass das Ziel von Schulungsmaßnahmen sein sollte, die BeurteilerInnen zu professionalisieren, indem sie in die Lage versetzt werden, ihre eigenen Handlungen, ihre Beurteilungsstrategien und ihren Beurteilungsprozess wahrzunehmen und kritisch zu reflektieren, dann sollte eine solche *assesment awareness* oder Beurteilungsbewusstheit genuiner Teil der Aus- und Fortbildung aller mit Leistungsmessung befasster Personen sein, seien es Personen, die – wie im Falle des TestDaF – für ein Testsystem arbeiten, seien es Personen, die im schulischen oder außerschulischen Kontext im Rahmen von Fremdsprachenunterricht Leistungen beurteilen müssen. Die Bewusstmachung eigener



Beurteilungsstrategien, gerade auch in der Auseinandersetzung mit anderen BeurteilerInnen der *peer group* kann zur Justierung der eigenen Handlungen und Maßstäbe und damit zur Erhöhung der Reliabilität sowie der Konstruktvalidität führen. „Diese Sensibilisierung oder *assessment awareness* für eigene (aber auch fremde) Beurteilungsstrategien ist Voraussetzung für die Reflexion des eigenen Handelns und damit Voraussetzung für die Optimierung oder Revision von Beurteilungsstrategien. Beurteilungsbewusstheit bezieht das Wissen über Abläufe und Determinanten der Beurteilung, also auch (unerwünschte) die Beurteilung beeinflussende Effekte, ein und erfordert damit auf Seiten der BeurteilerInnen die Bereitschaft, sich dem Phänomen Beurteilung auch theoretisch zu nähern und auf einer Metaebene die Tätigkeit des Beurteilens zu kommunizieren. Auf dieser Basis kann dann auch analysiert werden, welche Handlungen ggf. wünschenswert, welche Handlungen hingegen problematisch sind und daher verändert werden sollten. Diese Reflexion der eigenen Bewertungspraxis kann dank des Laut-Denken-Verfahrens in der Handlung selbst erfolgen".[11] Ich möchte daher anregen, introspektive Verfahren zur Bewusstmachung der eigenen Beurteilungsstrategien in die Aus- und Weiterbildung zu integrieren. Praktisch können introspektive Verfahren im Rahmen von Schulungsmaßnahmen systematisch eingesetzt werden, die Verbaldaten können aufgenommen, also konserviert und somit einer Analyse (eigen-initiiert oder fremd-initiiert, alleine oder in der *peer group*) zugänglich gemacht werden. Zunächst sollte es darum gehen, das Verfahren des Lauten Denkens sinnvoll einzusetzen. Dies erfordert ein gewisses Training und selbstredend auch eine entsprechend von Vertrauen geprägte Atmosphäre. Die Analyse selbst kann zunächst darin bestehen, die eigenen oder fremden Handlungen und Strategien zu differenzieren, sie aufzulisten und zu kategorisieren. In diesem Zusammenhang ist es entscheidend, die Gründe für bestimmte Handlungen zu eruieren, also möglichst genau zu analysieren, warum an welcher Stelle welche Strategie zum Einsatz kommt, mit welchen anderen Strategien sie Cluster bildet, welche Erwartungen, subjektiven Theorien und Entscheidungen aus welchen Gründen zugrunde liegen. An dieser Stelle werden vermutlich überraschende Erkenntnisse ans Tageslicht gelangen, denn die Verbaldaten der Laut-Denken-Protokolle ebenso wie der Interviews der vorliegenden Studie zeigen, dass sich die Versuchspersonen an neuralgischen Punkten der Bewertungsarbeit Aspekten bewusst werden, die vermutlich ohne

[11] *Arras* (2007:449), s. hierzu auch den Ansatz der Aktionsforschung von Altrichter und *Posch* (1998).

introspektive Verfahren verschüttet geblieben wären.[12] Ein Beispiel soll dieses Phänomen illustrieren: Wir beobachten, dass die BewerterInnen dazu neigen, Leistungen vornehmlich in den mittleren Leistungsstufen einzuschätzen, also im Fall des TestDaF die Einstufungen TDN3 und TDN4 vorzunehmen, TDN5 sowie unter TDN3 hingegen eher zu meiden. Dieses als Zentraltendenz bekannte Phänomen zeigt sich auch im Beurteilungsverhalten der vorliegenden Studie. Problematisch hierbei ist, dass die Leistungsstufen gleich breit sein sollten, immerhin werden auf der Basis der Einstufungen mathematische Operationen vorgenommen. Wenn jedoch Leistungen zu wenig differenziert eingestuft werden, also sowohl stärkere als auch schwächere Leistungen im mittleren Leistungsbereich eingestuft werden, so liegen verzerrte Resultate vor. Versuchsperson K7 wird sich dieses Problems bewusst, sie reflektiert kritisch die Gründe für ein solches Verhalten und findet schließlich die Ursache in ihrer schulischen Sozialisation, also in einer Zeit, die Jahrzehnte zurückliegt, die aber nichtsdestotrotz ihre Auswirkungen auf das Heute und auf ihr berufliches Handeln hat (RIK7, ca. 1:25):

> Grad an dem, eh [.] das mit der 5, ne? Man hat vielleicht doch die Tendenz, dass man [...] Ja, ich denk, das ist auch so, das ist schon etwas, was vielleicht schon im Unterbewusstsein ist, und so richtig internalisiert, von der Schule schon her. Also als ich in der Schule war, da war es praktisch unmöglich, ne Eins zu kriegen. Das war so was Seltenes, eigentlich, ne Eins gab es nicht. Heute hat sich das geändert. Heute gibt es eigentlich Einsen genug. Aber zu meiner Zeit, ne Eins war so eine fantastische Leistung, die hat man nur selten erreicht. Und das ist vielleicht so etwas, was so ein bisschen in einem drin ist.

Introspektion – im vorliegenden Fall sowohl über das Verfahren des Lauten Denkens als auch mit Hilfe retrospektiver Betrachtung – erlaubt uns, unsere Handlungen und Strategien bei der Beurteilungsarbeit bewusst zu machen und metakognitiv zu reflektieren mit dem Ziel, wünschenswerte Strategien beizubehalten, unerwünschte Maßstäbe und Strategien hingegen zu verwerfen oder zu optimieren. Dies wiederum ist Voraussetzung dafür, dass wir uns innerhalb der Gruppe der BeurteilerInnen auf einheitliche Maßstäbe einigen können und die Bewertungskriterien angemessen operationalisieren, m. a. W.: dass wir Leistungen reliabel, valide und somit fair beurteilen. Erst wenn wir uns darüber im Klaren sind, warum wir welche Beurteilung vornehmen, können wir

[12] Die Palette der im Rahmen der Studie beobachteten Strategien sowie ein Versuch ihrer Systematisierung befindet sich in *Arras* (2007) sowie in Arras (erscheint).



sicher sein, dass wir das messen, was wir vorgeben zu messen. Reflexion und mithin Kontrolle der eigenen Handlungen, der individuellen und institutionell geprägten Beurteilungsstrategien ist Voraussetzung für die Validität eines (Schreib-)Tests.

## Literatur

*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): Wie es zu einer Beurteilung kommt. Ein Forschungsbericht zu Strategien bei der Beurteilung schriftlicher Leistungen im Kontext der Prüfung TestDaF, In: Materialien Deutsch als Fremdsprache.

*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. Akten des sechsten Kongresses des Katalanischen Deutschlehrer- und Germanistenverbandes (A.G.C.), Tarragona, April 2005, 39-52. <http://www.tinet.org/~asgc2/Forum_2005/Autors/Arras/arras04.html> (26.06.08)

*Arras, U.* (erscheint): Wie es zu einer Beurteilung kommt. Ein Forschungsbericht zu Strategien bei der Beurteilung schriftlicher Leistungen im Kontext der Prüfung TestDaF, in Materialien Deutsch als Fremdsprache.

*Arras, U./ Grotjahn, R.* (2003): TestDaF: Einige aktuelle Entwicklungen, in: *Katzorke, H.* (Ed.) Fremdsprachen an Hochschulen: Integration – Interdisziplinarität – Internationalität. Dokumentation der 22. Arbeitstagung 2002. Bochum: AKS-Verlag, 40-50.

*Cumming, A./ Kantor, R./ Powers, D.* (2001): Scoring TOEFL Essays and TOEFL 2000 prototype writing tasks: An investigation into rater's decision making and development of a prelimnary analytic framework, TOEFL Monograph Series, MS-22. Princton, NJ: ETS.

*Cumming, A./ Kantor, R./ Powers, D.* (2002): Decision making while rating ESL/EFL writing tasks. A descriptive framework, in: *Modern Language Journal* 86/1, 67-96.

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

*Eckes, T.* (2005a): 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.* (2005b): Analyse und Evaluation sprachproduktiver Prüfungen beim TestDaF. In: *Kühn, I./ Lehker, M./ Timmermann, W.* (Eds.): Sprachtests in der Diskussion, Frankfurt: Lang. 60-93.

*Eckes, T.* (2004): Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen, in: *Wolff, A./ Ostermann, T./ Chlosta, C.* (Eds): Integration durch Sprache. Regensburg: FaDaF, 485-518.

*Eckes, T.* (2003): Qualitätssicherung beim TestDaF: Konzepte, Methoden, Ergebnisse, in: *Fremdsprachen und Hochschule*, 69, 43–68.

*Elder, C/ Knoch, U./ Barkhuizen, G./ Randow, J., von* (2005): Individual feedback to enhance rater training: does it work? Language Assessment Quarterly Vol. 2/3, 175-196.

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

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

*Kleppin, K./ Tönshoff, W* (2000): Autonomiefördernde Strategievermittlung als Gegenstand und Verfahren in der Ausbildung von Fremdsprachenlehrern, in: *Helbig, B./ Kleppin, K./ Königs F G* (Eds): Sprachlehrforschung im Wandel. Beiträge zur Erforschung des Lehrens und Lernens von Fremdsprachen. Festschrift für Karl-Richard Bausch zum 60. Geburtstag, Tübingen: Stauffenburg, 113-128.

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

*Lumley, T./ McNamara, T F* (1995): Rater characteristics and rater bias: Implications for training, Language Testing, 12/1, 54-71.

*TestDaF-Institut* (2005): Musterprüfung 1, Ismaning: Hueber.

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

<www.testdaf.de>



Adresse der Verfasserin

Ulrike Arras
ulrike.arras@gmail.com