KOLLOQUIUM FREMDSPRACHENUNTERRICHT

Herausgegeben von Daniela Caspari,
Lars Schmelter, Karin Vogt und Nicola Würffel

BAND 48

Zu Qualitätssicherung und Peer Review der vorliegenden Publikation

Die Qualität der in dieser Reihe erscheinenden Arbeiten wird vor der Publikation durch alle vier Herausgeber der Reihe geprüft.

Notes on the quality assurance and peer review of this publication

Prior to publication, the quality of the work published in this series is reviewed by all four editors of the series.

Karin Aguado / Lena Heine / Karen Schramm (Hrsg.)

Introspektive Verfahren und Qualitative Inhaltsanalyse in der Fremdsprachenforschung

PETER LANG
EDITION

Introspektive Verfahren in der Sprachtestforschung

Ulrike Arras

Traditionally, research in language testing is conducted using quantitative methods, such as psychometrics and statistics. Since the 1980s, however, an increasing number of qualitative research methods, such as introspection and verbal protocols, have been implemented. This is based on the concern that quantitative methods are unable to discover and explain cognitive processes and factors which influence testing and assessment satisfactorily. By and large, two areas have since been investigated using introspective methods. Firstly, the field of test-taking strategies: Are the tests we develop able to elicit those language behaviours which we aim to measure? Secondly, the field of rating test products such as oral and written texts: What kind of strategies and cognitive processes can be observed in raters? On which aspects of the text do raters base their assessment and why do they do so? Thus, introspective methods in the context of language test quality have two benefits: We learn about how tests function and whether we can measure that which we sought to. Additionally, the findings might be used for our language behavior as learners and our assessment behaviour as teachers and raters. Both aspects are crucial for test validity. This article aims to give an overview of the state of the art in language testing with regard to introspective methods and discusses its potential and limitations.

1 Begründungszusammenhang

Tests und Sprachtests sind traditionell vor allem die Spielwiese quantitativer Forschungsmethoden. Die Reklamation von Testgütekriterien wie z. B. die Reliabilität, die mit Hilfe statistischer Verfahren kontrolliert werden kann, gehört entsprechend zu den Errungenschaften der Qualitätssicherung auf dem Gebiet der Sprachtestforschung. Einen Überblick über forschungsmethodologische Entwicklungen auf dem Gebiet des Sprachtestens der letzten Dekaden verdanken wir Banerjee/Luoma (1997), Bachman (2000) sowie Lumley/Brown (2005). Demnach ist die Sprachtestforschung bis in die 1970er Jahre geprägt von quantitativen Methoden.

Nichtsdestotrotz bleibt das Unbehagen, allein mit quantitativen Verfahren ein so zentrales Problem des Fremdsprachenunterrichts wie die Leistungsmessung zu erforschen. Daher sind Entwicklungen erfreulich, die sich mittlerweile in der Sprachtestforschung abzeichnen und die mit Hilfe qualitativer Methoden spezifische Strategien und Prozesse im Kontext Sprachtest sowie die Produkte von Leistungsmessung (d. h. schriftliche und mündliche Prüfungsleistungen) bzw. deren Merkmale zu erforschen suchen. Lumley/Brown (2005: 833) fassen diese Entwicklung in ihrem Überblick zu Methoden in der Sprachtestforschung folgendermaßen zusammen:

> There has been a move from positivistic research focusing on properties of tests and scores toward a broader and more critical examination of a wide range of validity issues embracing construct definition as well as language testing practice and policy.

Seit den 1990er Jahren lässt sich somit eine Hinwendung zur Erforschung von Faktoren beobachten, die die Performanz beeinflussen. Bachman (2000) zählt folgende Forschungsgebiete auf:

- Merkmale, die den Prozess des Testens determinieren, unter Einschluss von Aspekten wie der Bewertung von Prüfungsleistungen durch BeurteilerInnen;
- Prozesse und Strategien der Prüflinge bei der Bewältigung der Testaufgaben;
- Merkmale der Prüflinge selbst.

Der Einbezug von Faktoren wie etwa Persönlichkeitsmerkmalen der Prüflinge sowie *impact*-Faktoren in die Erforschung von Sprachtests macht qualitative Forschungsmethoden erforderlich. Hierbei bieten sich introspektive Verfahren an, die Einblicke in kognitive Prozesse erlauben. Exemplarisch genannt seien die auf introspektiven Daten bzw. *verbal protocols* beruhenden Arbeiten zur Analyse der *test-taking processes* bei Hörverstehenstests etwa von Buck (1991) oder jüngst von Rossa (2012), die Analysen der Beurteilungsprozesse und -strategien von Lumley (2005) und Arras (2007) zu schriftlichen Prüfungsleistungen sowie jüngst die Arbeiten von Ducasse (2010) und May (2011) zu den Prozessen bei mündlichen Prüfungen.

Interessant ist hierbei vor allem die Entwicklung, Daten außerhalb des eigentlichen Tests in die Interpretation einzubeziehen (Lumley/Brown 2005: 849). Denn damit stellen sich der Sprachtestforschung interdisziplinäre Fragen etwa aus angrenzenden Gebieten wie den Sozialwissenschaften und der Psychologie. Dies hängt auch mit dem veränderten bzw. erweiterten Validitätsbegriff (s. insb. Messick 1989) zusammen.

> As in other areas of educational measurement, validity is no longer viewed as a set of measurable phenomena (based on traditional categories of construct, content, predictive and concurrent validities) but as an integrated and holistic judgment that is based on the collection of evidence from a number of different areas [...]. In this revised conceptualization, the main concern is the validity of the inferences made about test-takers on the basis of their scores, that is, the meaning of scores. (Lumley/Brown 2005: 840f.)

Konsequenterweise geht es darum, sowohl das Produkt von Leistungsmessung, also konkrete schriftliche oder mündliche Leistungen, als auch den Leistungsprozess, etwa die kognitiven und strategischen Handlungen, die diesen Leistungen zugrunde liegen, zu verstehen. Und hierbei sind qualitative Daten erforderlich (ebd.). Folgende Entwicklungen sind erkennbar:

- Mehr und mehr werden heute auch externe Merkmale in die Sprachtestanalyse einbezogen, indem etwa der soziale und (bildungs)politische Kontext sowie die Instrumentalisierung und der *Impact* von Sprachtests in den Fokus geraten (Shohamy 2001).
- Bei den qualitativen Methoden zur Erforschung von Sprachtests wird zum einen die Diskursanalyse eingesetzt, etwa zur Analyse mündlicher Leistungen (McNamara/Hill/May 2002). Zum anderen werden mittels Introspektion erhobene Daten, in erster Linie Lautes Denken in Form von *verbal protocols*, hinsichtlich kognitiver Prozesse analysiert.[1]

Einen Überblick zum Einsatz qualitativer Methoden in der Sprachtestforschung legen Banerjee und Luoma (1997) vor. Sie führen an, dass bereits in den 1990er Jahren zunehmend qualitative Ansätze in der Sprachtestforschung berücksichtigt werden. Die Arbeiten zeigen nach Bachman (2000: 7) zudem, dass verschiedene Ansätze Anwendung finden und miteinander kombiniert werden, etwa „expert judgments, introspective and restrospective verbal reports, observations, questionnaires and interviews, as well as text analysis, conversational analysis and discourse analysis."

Introspektive Verfahren im Kontext Sprachtestforschung konzentrieren sich im Wesentlichen auf zwei Problembereiche, erstens auf die Eruierung von Teststrategien seitens der Prüflinge und zweitens auf die Analyse der Strategien zur Beurteilung von Prüfungsleistungen seitens der BewerterInnen. Der erste Aspekt erfasst also den Umgang mit Testaufgaben, also die kognitiven Prozesse, die bei der Bearbeitung von Testaufgaben zu beobachten sind, der zweite die Interpretation der Ergebnisse, also die Beurteilung der Prüfungsleistungen. Beides sind zentrale Anliegen jedweden Testens, denn zum einen wird fokussiert, ob das, was man mit bestimmten Testaufgaben elizitieren will, etwa bestimmte Lesestrategien, diese auch tatsächlich hervorruft. Zum anderen wird betrachtet, wie die Produkte eines Tests, also z. B. ein schriftlich fixierter Text, der aufgrund einer bestimmten Aufgabenstellung verfasst wird, rezipiert und beurteilt werden. Beide Fragestellungen sollen im Folgenden genauer skizziert werden.

2 Introspektion zur Erforschung von Teststrategien seitens der Prüflinge

Die Leitfragen zur Erforschung von Teststrategien lauten etwa: Wie gehen Prüflinge mit Testaufgaben um? Wie lösen sie die Aufgaben, wie gehen sie konkret vor? Welche Strategien setzen sie ein? Und sind diese Strategien zielführend im Sinne des Testkonstrukts? Zugrunde gelegt ist dabei die Frage

[1] Nützliche Literatur hierzu s. insbesondere Ericsson/Simon (1993) sowie speziell für die Belange der Sprachtestforschung Green (1998).

danach, ob die Aufgabe das elizitiert, was auch tatsächlich gemessen werden soll, ein nachgerade zentrales Anliegen von Sprachtests.

Ein Überblick über Studien auf diesem Gebiet stammt von Cohen (1998). Er fasst Untersuchungen zum Problem Teststrategien zusammen, die sich introspektiver Verfahren (*verbal report techniques*, *qualitative methodologies*, ebd.: 107) bedienen. Das entscheidende Moment sieht Cohen (ebd.: 92) in der Bewusstmachung. Eine Handlung kann demnach erst dann als Strategie bezeichnet werden, wenn sie bewusst ausgewählt wurde (s. auch Grotjahn 1997). Dies gilt gerade auch für Teststrategien:

> *Language use strategies* are mental operations or processes that learners consciously select when accomplishing language tasks. These strategies also constitute *test-taking strategies* when they are applied to tasks in language tests. [...] Test-taking strategies will be viewed as those test-taking processes that the respondents have selected and of which they are conscious, at least to some degree. In other words, the notion of strategy implies an element of selection. Otherwise the processes would not be considered strategies. (Cohen 1998: 92, Hervorhebungen i. O.)

Während sich frühere Studien mittels introspektiver Verfahren vor allem dem Effekt von Itemtypen auf die Teststrategie widmen,[2] konzentrieren sich jüngere Arbeiten auf spezifische Teststrategien je nach Fertigkeitsbereich. Zu nennen sind etwa Buck (1991) und Rossa (2012) zum Hörverstehen sowie Grupa (1999) zum Seh-Hörverstehen (zit. nach Grotjahn 2005). Cohen (1998) untersucht integrative Testaufgaben, die Lesen und Schreiben (*summarizing*), also kognitiv hoch komplexe Handlungen, erfassen. Ebenfalls zum Leseverstehen, aber anhand von *Multiple Choice*-Items liegt eine entsprechende Studie von Dollerup, Glahn und Hansen (1994) vor. Auch Wus (1998) Arbeit untersucht Teststrategien anhand von *Multiple Choice*-Items, allerdings zum Hörverstehen. Cohen/Upton (2006) verwenden ebenfalls introspektive Verfahren (Laut-Denken-Protokolle) zur Überprüfung des revidierten Leseverstehenstests beim *TOEFL*.[3]

Zusammenfassend kann festgehalten werden, dass introspektive Verfahren Einblicke in die Funktionsweise von Aufgaben gewähren: Was löst ein bestimmtes Item-Format aus? Welche (beispielsweise Lese-)Strategie wird durch ein bestimmtes Aufgabenformat initiiert? Diese Einsicht erweist sich als zentral für die Sicherstellung der Testqualität. Denn nur wenn Aufgaben das elizitieren, was der Test vorgibt zu messen, ist der Test valide. Wird hingegen eine Aufgabe allem Anschein nach korrekt gelöst (d. h. entspricht die Lösung dem vorgegebenen Lösungsschlüssel), beruht die Lösung jedoch auf Strategien, die mit dem Test nicht beabsichtigt werden und die somit nicht Teil des Testkonstrukts

[2] Eine frühe empirische Arbeit auf diesem Gebiet verdanken wir Stemmer (1991) speziell zum Format C-Test.

[3] Eine Übersicht über die Forschungsliteratur zu Teststrategien unter Verwendung von Verbaldaten s. Cohen/Upton (2006: 11).

sind, so liegt eingeschränkte Validität vor. Bei bestimmten Aufgabenformaten, etwa geschlossenen Item-Formaten wie *Multiple Choice*, kann die Bewältigung der Aufgabe nicht direkt beobachtet werden, vielmehr liegt mit der Lösung lediglich ein Indiz vor. Wir schließen beispielsweise bei einer Leseverstehensaufgabe mit *Multiple Choice*-Aufgaben von der korrekten Lösung auf die angemessene Rezeption und Verarbeitung des Lesetextes, also auf die intendierte kognitive Operation. Welche konkreten Leseverstehensstrategien aber tatsächlich angewendet wurden oder ob der Text überhaupt gelesen wurde, können wir nicht mit Sicherheit sagen. Cohen (1998: 107) schlussfolgert in diesem Zusammenhang: „Verbal report can help us see what items are actually testing, aiding us in making decisions about which items to keep and which to throw out." Er schlägt somit vor, Verbaldaten in die Erprobung von Testaufgaben einzubeziehen, denn „if a respondent has legitimate reasons for marking an item wrong, then the item needs to be rewritten".[4] Dem entgegen stehen allerdings sicherlich Praktikabilitätserwägungen. Denn die Erhebung introspektiver Daten bei der Erprobung neuer Testaufgaben stellt einen erheblichen Aufwand dar. Zumindest sind jedoch solche Verfahren bei der Konzipierung neuer Sprachtests oder neuer Testformate dringend anzuraten.

3 Introspektion zur Erforschung von Bewertungsstrategien seitens der BeurteilerInnen sprachlicher Prüfungsleistungen

Gleichermaßen zentral für die Testvalidität ist die Frage danach, wie Prüfungsleistungen beurteilt werden. Welche Prozesse sind zu beobachten? Worauf achten die BeurteilerInnen? Wovon lassen sich die BeurteilerInnen bei ihren Wahrnehmungen und Urteilen leiten? Inwiefern beziehen sie bereitgestellte Beurteilungsinstrumente, etwa Bewertungskriterien, ein? Und welche subjektiven Theorien liegen den Entscheidungen zugrunde? Introspektion vermag auch hier Aufschluss zu geben, inwiefern die Beurteilung im Sinne des Testkonstrukts erfolgt. Erst wenn, etwa durch Bewusstmachung und Reflexion, klar wird, worauf die BeurteilerInnen ihr Urteil gründen, kann sichergestellt werden, dass jene Aspekte beurteilt werden, die mit der Sprachprüfung erfasst werden sollen. Eine schriftliche Deutschprüfung beispielsweise, die vorgibt, kommunikative Kompetenzen zu messen, muss sicherstellen, dass bei der Beurteilung das Augenmerk nicht allein auf grammatische Korrektheit gerichtet wird.

Auf dem genannten Gebiet entstanden in den letzten Jahren einige empirische Arbeiten, die mittels Introspektion versuchten, den Beurteilungsprozess und die

[4] S. exemplarisch die Verwendung introspektiver Daten zur Validitätssicherung des neuen TOEFL-Lesetests (d. h. der neue internetbasierte iBT-TOEFL) bei Cohen/Upton (2007) und Cohen/Upton (2006).

Strategien bei der Beurteilung schriftlicher Prüfungsleistungen zu eruieren (Überblick bei Arras 2007: 121ff.). Zu nennen ist insbesondere die Studie von Lumley (2005), die auch meine eigene Arbeit (Arras 2007) beeinflusst hat.[5] In beiden empirischen Studien werden die BeurteilerInnen schriftlicher Prüfungsleistungen in den Blick genommen. Lumleys Untersuchung liegt ein Englischtest für EinwanderInnen in Australien zugrunde, die Studie von Arras entstand im Kontext Test Deutsch als Fremdsprache (TestDaF), einer Sprachprüfung für den Zugang zu Hochschulen in Deutschland. Insgesamt zeigen solche introspektiven Untersuchungen, wie komplex das Beurteilungsverhalten tatsächlich ist. Etliche Strategien lassen sich differenzieren. Die Beurteilungsarbeit selbst erweist sich als stark prozesshaft, gesteuert sowohl durch die Beurteilungsinstrumente, etwa Bewertungskriterien, aber insbesondere auch durch Persönlichkeitsmerkmale und subjektive Theorien, individuelle Erfahrungen und Erwartungen an die Prüfungsleistungen. Dies wiederum wirft Fragen auf zu speziellem Schulungsbedarf für BewerterInnen (s. dazu genauer Abschnitt 6).

Mittlerweile sind auch Arbeiten entstanden, die die Beurteilung mündlicher Leistungen in der Fremdsprache mittels introspektiver Verfahren bzw. anhand von *verbal protocols* untersuchen. Dies ist methodisch ungleich schwieriger, denn das Untersuchungsdesign muss dem Problem Rechnung tragen, dass kaum gleichzeitig gehört und gesprochen werden kann, d. h. die mündliche Leistung muss auditiv wahrgenommen werden und die eigenen Gedanken und Emotionen sollen laut verbalisiert werden. Hierzu haben nun Ducasse (2010) sowie May (2011) empirische, auf verbalen Daten beruhende Arbeiten vorgelegt. Sie untersuchen, wie mündliche Leistungen beurteilt werden, und zwar anhand von Daten aus Prüfungssituationen, wie sie oft, auch im universitären Kontext, durchgeführt werden: die Paarprüfung in einer *face-to-face*-Situation. Im Gegensatz zu Einzelprüfungen, bei denen eine durchaus problematische kommunikative Situation vorliegt, die ein hierarchisches Gefälle zwischen PrüferIn und Prüfling aufweist, kommt in einer Paarprüfung hinzu, dass die beiden Prüflinge gehalten sind zu interagieren.

4 Einige methodologische Überlegungen

Introspektive Verfahren (Lautes Denken, Retrospektion) liefern qualitative verbale Daten, die Aufschluss geben sollen über kognitive Prozesse, Strategien, aber auch emotionale Aspekte, Einstellungen, Motivation, Befindlichkeiten etc. seitens der ProbandInnen (vgl. hierzu insb. Schnell in diesem Band). Die Stärke

[5] Eine der ersten Arbeiten auf diesem Gebiet stammt von Huot (1988, 1993); allerdings untersucht er die Beurteilung muttersprachlicher (Englisch-)Prüfungsleistungen. Die Übertragbarkeit auf den Kontext Beurteilung fremdsprachlicher Leistungen ist nicht unproblematisch, s. auch Arras (2007: 125).

introspektiver Daten liegt gerade darin, Einblick zu gewähren in Befindlichkeiten und bewusstseinsfähige Daten, die strategischen Entscheidungen zugrunde liegen. Introspektive Verfahren ermöglichen diesen Zugriff. Sie bringen jedoch auch einige gravierende Probleme und Einschränkungen mit sich, die bei der Entwicklung des Forschungsdesigns sowie bei der Interpretation der Daten und nicht zuletzt bei der Generalisierbarkeit der Ergebnisse berücksichtigt werden müssen.[6] Sie seien im Folgenden kurz skizziert:

Zum einen ist zu beachten, dass das Datenerhebungsverfahren die Daten beeinflusst: Wir müssen davon ausgehen, dass introspektive Verfahren spezifische Verhaltensweisen (Strategien, Prozesse) initiieren, die auf das Verfahren selbst zurückzuführen sind und damit den Verhaltensweisen unter realen Bedingungen nicht entsprechen. So erfordert beispielsweise die Beurteilung einer schriftlichen Prüfungsleistung unter Laut-Denk-Bedingungen das Vorlesen des Textes. Da diese interimsprachlichen Texte fehlerhaft sind, werden diese fehlerhaften Textstellen entsprechend verbalisiert. Vermutlich führt die doppelte Präsentation der sprachlichen Fehler (visuell über das Lesen und auditiv über das Hören) zu erheblicher Irritation, wenn nicht (wie in der Studie von Arras 2007 beobachtet) offensichtlich eine Vermeidungsstrategie eingesetzt wird: Die ProbandInnen verbalisieren fehlerhafte Textstellen korrigiert; d. h. sprachlich fehlerhafte Textstellen werden offensichtlich automatisiert korrigiert, es handelt sich sozusagen um ein „Sich-zurecht-Lesen", vermutlich eine kooperative Verstehensstrategie den interimssprachlichen, teils stark fehlerhaften Texten gegenüber.[7]

Daneben stellt sich folgende Frage: Sind die Versuchspersonen überhaupt in der Lage zu verbalisieren, was mental präsent ist? Prinzipiell ist zu berücksichtigen, dass Persönlichkeitsfaktoren, etwa Extroversion vs. Introversion, akute Stimmungen, Erfahrung, Motivation, vor allem auch das Ausmaß an Bewusstheit für das eigene Handeln sowie die Fähigkeit zur Selbstreflexion etc. die Qualität und auch den Umfang des Datenmaterials bestimmen.[8]

[6] Hilfreich in diesem Zusammenhang sind vor allem Greens (1998) kritische Einschätzungen introspektiver Verfahren bei der Erforschung von Beurteilungsstrategien (s. insb. Lumley 2005: 67ff. sowie Arras 2007: 140ff.). Rossa (2012) schlägt in einem ähnlichen Zusammenhang (Rekonstruktion von Teststrategien in Laut-Denken-Versuchssituationen) den Begriff ‚Konfabulation' vor, ein aus der Medizin bzw. Psychologie entlehnter Begriff, der ein Verhalten beschreibt, bei dem Informationen verbalisiert werden, die – bewusst oder unbewusst – nicht der Realität entsprechen, mit dem Ziel, gegenüber einer anderen Person oder auch sich selbst gegenüber Gedächtnislücken zu kaschieren bzw. antizipierten Verhaltenserwartungen zu entsprechen.

[7] Ausführlich hierzu Arras (2007: 270ff. und 469ff.). S. auch Lumley (2005: 283ff.).

[8] Eine Klassifizierung von Lernertypen in diesem Kontext nimmt Grotjahn (2007) vor. Zu Persönlichkeitsfaktoren bei BeurteilerInnen s. Arras (2007: 445ff. und 467ff.).

Zudem stellen das bei Ericsson/Simon (1993) diskutierte Phänomen der *heeded information* oder *heeded thoughts* sowie der latente Erinnerungsfehler ein Problem dar. So sind verbalisierte erinnerte Informationen nicht zwangsläufig zuverlässig, denn die Erinnerung kann trügen. Die Autoren gehen von *error in recall from LTM* (*long term memory*) aus (Ericsson/Simon 1993: 258).

Dieses Phänomen ist besonders kritisch, wenn die ProbandInnen nicht in ihrer Erstsprache verbalisieren, beispielsweise wenn man eine mehrsprachige internationale LernerInnen-Gruppe hinsichtlich ihrer Test- oder Lernstrategien untersucht. Grundsätzlich sind Verbaldaten in der Erstsprache zu erheben, um zu verhindern, dass Informationen (Überlegungen, Befindlichkeiten etc.) nicht verbalisiert werden, allein weil sie nicht in der Zielsprache formuliert werden können und um den ohnehin problematischen Umstand zu relativieren, dass die Verbalisierung selbst Schwierigkeiten verursacht. Für die ProbandInnen sollten die Bedingungen der Untersuchungsdurchführung, also auch die Verbalisierung selbst, so einfach wie möglich gestaltet werden.

Dieser Punkt wirft Fragen bei Untersuchungen heterogener Gruppen auf, etwa der Erfassung von *test-taking strategies* in Gruppen von Prüflingen unterschiedlicher Herkunft, wie sie oftmals in Kursverbänden, auch zur Vorbereitung einer Sprachprüfung, zusammengesetzt sind. Wenn die ProbandInnen in ihrer Erstsprache introspektive Daten liefern, dann sollte sichergestellt werden, dass die Auswertung von *native speakers* vorgenommen wird. Oder aber die Versuchsgruppe sollte hinsichtlich der Herkunftssprache homogen organisiert sein, etwa indem nur eine bestimmte Zielgruppe untersucht wird, z. B. nur chinesische Lernende oder nur Lernende mit Französisch als Ausgangssprache.

Problematisch ist darüber hinaus die Frage nach der antizipierten sozialen Erwünschtheit. Das bedeutet, inwiefern filtern die Versuchspersonen bei der Introspektion, so dass nur solche Informationen preisgegeben, also verbalisiert werden, die als erwünscht oder seitens der Forscherin/des Forschers als erwartet antizipiert werden? Inwiefern möchten die ProbandInnen damit auch einem bestimmten Bild entsprechen, insbesondere wenn ein soziales Gefälle oder gar eine Abhängigkeit vorliegt (etwa Lernende gegenüber ihren forschenden LehrerInnen)? Diese Fragen beeinträchtigen direkt die Validität der Daten. Das bedeutet, introspektiv ermittelte qualitative Daten sind ggf. nur eingeschränkt valide. (Diskussion zur Nutzung von *verbal protocol analysis* im Kontext Sprachtestforschung insbesondere bei Green 1998).

Ein weiteres Problem bei der Verwendung von Verbaldaten (introspektive Daten, Lautdenkenprotokolle) ist die Repräsentanz von Informationen im Gedächtnis. Wird beispielsweise untersucht, wie BeurteilerInnen mündliche Leistungen beurteilen, so sind wir auf retrospektive Daten angewiesen, denn die mündlichen Leistungen müssen zunächst gehört (im Falle von Videomitschnitt auch visuell aufgenommen) werden und können erst retrospektiv, etwa nach Ablauf

der Prüfungssequenz, von BeurteilerInnen kommentiert werden (methodische Diskussion dieses Sachverhalts s. bei Ericsson/Simon 1987, Green 1998, s. auch May 2011: 51f.).

Auf alle Fälle erscheint es notwendig, ProbandInnen gut in die Erhebungsweise introspektiver Daten einzuführen. Sie sollten Gelegenheit erhalten, das Verfahren zuvor zu üben. Doch auch ein vorhergehendes Training garantiert keine zuverlässigen Daten.[9]

Aufwendig bei introspektiven Daten ist nicht allein deren Erhebung, sondern vor allem auch ihre Auswertung. Meist werden alle Daten transkribiert, wobei sich die Komplexität der Transkription an der Fragestellung orientiert. Gleiches gilt für die Aufbereitung der Daten etwa mittels Segmentierung, Kodierung etc.[10]

Darüber hinaus sei auf methodologische Unterschiede aufmerksam gemacht, etwa die Differenzierung von introspektiven und retrospektiven Daten.[11] Arras (2007: 143ff.) unterscheidet zudem selbstinitiierte von fremdinitiierter Retrospektion. So sind im Rahmen von Lautdenkprotokollen selbstinitiierte retrospektive Daten zu beobachten, bei denen die Äußerungen der Versuchspersonen ohne äußeren Stimulus erfolgen. Als Beispiel sei eine Stelle aus einem Protokoll angeführt, an der die Versuchsperson Schwierigkeiten bei der Einschätzung eines Aspekts der Aufgabenumsetzung erläutert, indem sie Bezug nimmt auf eine zuvor beurteilte Leistung: „Also für mich ist es immer so ein bisschen schwierig, wie begründen die jetzt ihre selektive Auswahl, ne? Also das hatte ich vorhin schon mal bei einem Kandidaten, ich weiß nicht mehr, welche Nummer das war, 3 oder 4" (Arras 2007: 144). Hingegen liegen fremdinitiierte Daten vor, wenn die Versuchspersonen etwa im Anschluss an die Beurteilungsarbeit bzw. die Bearbeitung einer Testaufgabe aufgefordert wird, ihre Handlung zu kommentieren.

5 Diskussion und Perspektiven für die weitere Forschung

Trotz der skizzierten Einschränkungen sollte auf introspektive Verfahren in der Sprachtestforschung nicht verzichtet werden. Im Folgenden seien deshalb einige Überlegungen und Forschungsfragen angeführt.

- Inwiefern lassen sich introspektive Verfahren durch neue Technologien optimieren bzw. ergänzen? Können etwa bildgebende Verfahren die Interpretation der verbalen Daten unterstützen? Wie kann beispielsweise die Beobachtung von Augenbewegungen die Erfassung der kognitiven Prozesse bei der Bearbeitung eines Leseverstehenstests unterstützen? Oder können Video- und Audioaufnahmen von mündlichen Prüfungssituationen die Beurteilung retrospektiv genauer nachvollziehbar machen?
- Inwiefern lassen sich introspektive Verfahren systematisch in die Qualitätssicherung von Sprachtests einbeziehen, etwa bei der Erprobung neuer Testaufgaben, wie Cohen (1998) vorschlägt, oder wenigstens bei der Konzipierung neuer Sprachtests und neuer Testformate, wie oben bereits angedeutet? Welche Praktikabilitätserwägungen limitieren dieses Vorgehen?
- Inwiefern lassen sich introspektive Verfahren zu Schulungszwecken oder im Rahmen von Aktionsforschung nutzen? Denn die Verfahren haben über das eigentliche Anliegen, qualitative Daten zu wissenschaftlichen Zwecken zu erheben, hinaus Nutzen, nämlich als Initiation für Reflexion. Hierbei eröffnen sich die in den folgenden beiden Unterkapiteln beleuchteten Perspektiven.

5.1 Reflexion von Teststrategien

Introspektion/Lautes Denken wird eingesetzt zur Initiation von Reflexion eigener Teststrategien (oder allgemeiner Sprachstrategien, etwa Lesestrategien, Hörverstehensstrategien, Strategien zur Organisation und Realisierung eines Redebeitrags oder eines schriftlichen Textes) mit dem Ziel, hilfreiche Strategien zu identifizieren und zu trainieren bzw. weniger hilfreiche Strategien zu verwerfen oder zu optimieren. Wie also lassen sich introspektive Verfahren bei Prüflingen bzw. Lernenden nutzen, um eine Reflexion der eigenen Teststrategien (etwa im Rahmen der Vorbereitung auf eine Sprachprüfung) in Gang zu setzen? Zugrunde liegt die Annahme, dass die Externalisierung innerer Rede mittels Introspektion/Retrospektion den Grad an Metakognition erhöht und damit Reflexionsfähigkeit initiiert wird, was wiederum zur Optimierung von Teststrategien führen soll. Hierbei kann einerseits differenziert werden zwischen spezifischen Lösungsstrategien und -techniken, die für die Bewältigung einer Aufgabe im Rahmen einer Testsituation eingesetzt werden, und allgemeinen sprachverarbeitenden Strategien und Techniken andererseits. So sind beispielsweise bestimmte Techniken im Umgang mit (fremdsprachlichen) Lesetexten, etwa die Visualisierung bzw. das Markieren von Schlüsselwörtern, das Exzerpieren und Paraphrasieren zentraler Aussagen eines Textes, relevant für die Bearbeitung eines Leseverstehenstests, aber ebenso relevant für das Lesen und die Verarbeitung von Literatur allgemein, also Teil allgemeiner Lesestrategien. Für den Lernprozess bedeutet dies wiederum, dass im Unterricht erworbene und bewusstgemachte Textverarbeitungsstrategien ebenso für die Anforderungen in einer Testsituation nutzbar sind. Und umgekehrt: Die in einer Testsituation nützlichen Strategien und Techniken können u. U. auch für die Text- bzw. Sprachverarbeitung außerhalb einer Testsituation dienlich sein (s. auch den Begriff *strategic competence* bei Bachman/Palmer 1996). Der Unterschied liegt

[9] Diskussion hierzu insb. bei Ericsson/Simon (1993), s. auch Lumley (2005).

[10] Praktisch orientierte Ausführungen hierzu s. insb. Green (1998).

[11] Hinweise hierzu etwa bei Banerjee/Luoma (1997: 277) sowie bei Cohen (1987).

im Wesentlichen in der Aufmerksamkeitssteuerung. Sie wird im Falle eines Leseverstehenstests gemeinhin durch Items geregelt: Die Informationssuche im Lesetext erfolgt dann anhand des im Item formulierten Problems, etwa wenn ein *Multiple Choice*-Item eine bestimmte Detailinformation aus dem Text fokussiert. Ähnlich verhält es sich mit Lernstrategien, die für die Lösung von Testaufgaben nutzbar gemacht werden können bzw. auf Testsituationen übertragen werden. So können im Unterricht eingesetzte Übungen zur Erweiterung des Wortschatzes (Sammeln von Synonymen, Paraphrasierung von Texten und Textpassagen) auch für die Bearbeitung von Leseverstehenstests dienlich sein. Dies zeigt sich auch in diversen Übungsmaterialien zur Testvorbereitung, die systematisch Synonyme, Umformulierung, Paraphrasierung etc. beispielsweise zur Prüfungsvorbereitung Leseverstehen einsetzen.[12]

Als reine *test-taking strategies* hingegen können solche Strategien (oder auch Techniken) bezeichnet werden, die dem Zeitmanagement oder der Selbstorganisation dienen. Als Beispiel soll eine Technik angeführt werden, die gelegentlich in Testvorbereitungsmaterial genannt wird, nämlich solche Items visuell markieren, die bereits beantwortet bzw. gelöst sind (etwa durchstreichen) oder die später noch einmal überprüft werden müssen (etwa mit einem Fragezeichen versehen), mit dem Ziel, den Überblick zu bewahren und zeitökonomisch vorzugehen (Lodewick 2010: 45f.).

Was den Einsatz introspektiver Daten zu Lernzwecken anbelangt, so ist mittlerweile auch gefordert worden, Introspektion einzusetzen, um Lernenden ihre eigenen Lernstrategien bewusst zu machen und dadurch ggf. eine Optimierung der Lernstrategien zu initiieren. Hierzu führt Beyer (2005) einige Studien aus dem Kontext Spracherwerb an, die den Einsatz von Introspektion als positiv für den Lernprozess bzw. für die Entwicklung und Bewusstmachung von Lernstrategien nachweisen.[13] Den gewinnbringenden Effekt von Introspektion im Fremdsprachenunterricht fasst sie folgendermaßen zusammen:

- „Erhebung und Thematisierung von Aufgabenlösungsstrategien;
- Aufmerksamkeitsrichtung und Bewusstmachung gezielter fremdsprachlicher Strukturen;
- Erhebung von metasprachlichem Lernerwissen;
- Validierung von Tests;
- Erstellung von Lernerprofilen zur besseren Orientierung des Unterrichts an den individuellen Lernerbedürfnissen" (Beyer 2005: 20)

[12] Exemplarisch für den Subtest Leseverstehen der Prüfung Test Deutsch als Fremdsprache (TestDaF) s. Roche (2005: 42ff.), Kniffka/Gutzat (2003: 19, 23), Lodewick (2010: 68f.), ebenso die „Hinweise und Tipps" zur Prüfungsvorbereitung auf den Internetseiten des TestDaF-Instituts: http://www.testdaf.de/teilnehmer/tn-vorbereitung_tipp.php.

[13] Zur Relevanz der Bewusstmachung individueller Lernstrategien für den Lernprozess s. auch die Beiträge in Rampillon/Zimmermann (1997).

Insbesondere für den Kontext Testen ermöglichen uns Introspektion bzw. Selbstreport herauszufinden, welche Teststrategien ungeeignet sind, etwa weil der Prüfling falsche Annahmen macht in Bezug auf das, was von ihm gefordert ist und wie seine Leistungen schließlich beurteilt werden. So berichten etwa Kleppin/Reich (2009: 96) von der *test-taking strategy* eines arabischen Studenten, der bei einem Hörverstehenstest, bei dem gefordert ist, Kurzantworten zu notieren, bestrebt ist, als Antwort möglichst alles aufzuschreiben, was er hört. Denn er geht davon aus, dass der Beurteiler bzw. die Beurteilerin sich die korrekte Antwort aus dem Material aussucht, das er als Antwort notiert: „Wenn ich viel schreibe, dann gibt es kein Problem. Der Prüfer wird die zusätzlichen Antworten löschen" („löschen" meint hier vermutlich „nicht berücksichtigen". Und „zusätzlich" bedeutet sicher überflüssig im Sinne der Fragestellung). Der Prüfling geht also davon aus, dass sich die BeurteilerInnen sozusagen die richtige Antwort aus den Aufzeichnungen zusammensuchen. Erst die Selbstbeobachtung und Bewusstmachung sowie schließlich die Verständigung über die Testziele, das Anliegen der Sprachprüfung und schließlich auch über das entsprechende Vorgehen bei der Leistungsbewertung vermögen dem besagten Lerner zu verdeutlichen, dass diese Strategie keinen Nachweis über die zu messende Sprachkompetenz erbringt und somit auch in einem validen Test sinnlos ist.

5.2 Reflexion von Beurteilungsstrategien

Introspektion/Verbalisierung der *inner speech* bzw. Lautes Denken bietet sich auch für die Schulung von BeurteilerInnen (fremd-)sprachlicher (Prüfungs-) Leistungen an. Denn in einer empirischen Studie zum Beurteilungsverhalten konnte festgestellt werden, dass die Versuchspersonen (es handelte sich um erfahrene Beurteilerinnen, die wiederholt geschult und mit dem Beurteilungsinstrumentarium bestens vertraut waren) durch die Methode des Lauten Denkens angeregt wurden, ihr eigenes Beurteilungsverhalten zu reflektieren, zu hinterfragen, zu begründen. Dieses Verhalten, das für die Validität ihrer Urteile erforderlich ist, erfolgt unter normalen Arbeitsbedingungen, also das Beurteilen ‚für sich' ohne Verbalisierung und ohne die als sozial erwünscht wahrgenommene ‚Rechtfertigung' der Urteile bzw. der Urteilsschritte, vermutlich nicht (s. hierzu die Ausführungen und Vorschläge in Arras 2007: 462f.). Auch hierbei ist jedoch zu berücksichtigen, dass introspektive Verfahren sehr aufwendig sind, so dass sie sich für reguläre Schulungen ggf. als nicht praktikabel erweisen. Allerdings können diese Verfahren als Anregung bzw. als Projekt zur Integration in Schulungskonzepte dienen:

- Schulung von BeurteilerInnen im Kontext von Sprachprüfungen, sowohl standardisiert als auch weitgehend nicht standardisiert mit dem Ziel der Kalibrierung, d. h. um auf der Basis von Bewusstmachung einheitliche Beurteilungsmaßstäbe innerhalb der Gruppe der BeurteilerInnen zu erzielen, wobei diese Maßstäbe in Einklang mit dem Testkonstrukt stehen müssen.
- Aus- und Weiterbildung von Fremdsprachen-Lehrkräften, also Einbindung introspektiver Verfahren in grundständige Studiengänge, in Lehramtsstudiengänge, in Aufbaustudiengänge mit dem Ziel, typische Verzerrungen bei der Beurteilung, etwa Zentraltendenz oder Positionseffekt, bewusst zu machen.

6 Introspektive Verfahren im Kontext Schulung und Training

Wie also können introspektive Verfahren über reine Forschungszwecke hinaus auch in Schulungen eingesetzt werden? Im Folgenden sollen praktische Überlegungen zu möglichen Szenarien im Kontext Schulung skizziert werden:

- In Schulungen von Lehrkräften bzw. BeurteilerInnen sollen schriftliche oder mündliche (Prüfungs-) Leistungen alleine oder in der Arbeitsgruppe beurteilt werden, dabei sollen die Beurteilungsstrategien und möglicherweise subjektiv geprägten Maßstäbe reflektiert werden.
- Beim Training von Lernenden, die sich auf eine Sprachprüfung vorbereiten, sind für die anstehende Prüfung relevante Testaufgaben alleine oder in der Arbeitsgruppe zu bearbeiten bzw. zu lösen. Dabei sollen unterschiedliche Vorgehensweisen wahrgenommen und kritisch überprüft werden.[14]

Introspektive Verfahren im Rahmen von Schulungen können als Paararbeit oder als Arbeit in der Kleingruppe in folgenden Szenarien konzipiert werden:[15] Eine Person im Paar oder in der Kleingruppe löst die entsprechende Aufgabe unter Lautdenk-Bedingungen. Die Verbaldaten werden aufgezeichnet (Video- und/ oder Audioaufnahme). Die PartnerInnen hören zu, ohne das Wort zu ergreifen oder Rückfragen zu stellen, sie machen sich aber ggf. Notizen zu Vorgehen, Prozesshaftigkeit, Problemen, Widersprüchen etc. (am besten in einem dazu entwickelten Arbeitsblatt). Sodann werden die aufgezeichneten Verbaldaten gemeinsam gehört und reflektiert. Es besteht die Möglichkeit zu Rückfragen, zu retrospektiver Begründung bzw. Rechtfertigung für bestimmte Strategien, Arbeitsschritte, Entscheidungen. Am Ende sollte ein schriftlich (ggf. anhand eines Arbeitsblattes) fixiertes Arbeitsergebnis stehen, das gemeinsam im Paar oder in der Kleingruppe erarbeitet wird. Folgende Punkte können dabei relevant sein:

[14] S. auch die Vorschläge zu Reflexion und *peer*-Beobachtung bei Kleppin/Reich (2009: 109f.).

[15] S. auch das in einer Pilotstudie verwendete Verfahren, dokumentiert in Arras/Marks/ Zimmermann (2009).

- Welche Schwierigkeiten traten bei der Lösung der Aufgabe oder bei der Observation auf?
- Worauf gründen bestimmte Strategien? Etwa: Warum wurde eine Testaufgabe auf eine bestimmte Weise gelöst? Warum wurde eine bestimmte sprachliche Leistung so und nicht anders beurteilt? Welche Annahmen und subjektiven Theorien liegen zugrunde? Worauf sind diese zurückzuführen? (s. auch Arras 2009b).
- Wie haben sich die Beteiligten während der verschiedenen Phasen (Lösen der Aufgabe, Besprechung im Paar/in der Kleingruppe) gefühlt?

7 Schlussfolgerungen

Um die Validität eines Sprachtests sicherzustellen, ist es empfehlenswert, Daten zu erheben, die nachvollziehbar machen, wie die Prüflinge die Aufgaben bewältigen und die BeurteilerInnen schließlich Leistungen bewerten. Denn erst wenn die Testaufgaben die erwünschten kognitiven Prozesse, etwa bestimmte Verstehensstrategien bei einem Lese- oder einem Hörverstehenstest elizitieren, und BeurteilerInnen Prüfungsleistungen in Bezug auf das Testziel einschätzen, kann von einer validen Testaufgabe im Sinne des zugrundeliegenden Testkonstrukts ausgegangen werden. Die Beantwortung der folgenden Fragen trägt daher wesentlich zur Konstruktvalidität bei:

Inwiefern korrespondieren die von den Prüflingen in der Auseinandersetzung mit der Testaufgabe aktivierten mentalen Prozesse und Strategien mit den im Testkonstrukt theoretisch angenommenen mentalen Operationen und Teilkompetenzen? Und inwiefern führen unterschiedlich ausgeprägte Prozesse zu unterschiedlichen Testleistungen, erkennbar in unterschiedlichen Messergebnissen (s. Rossa 2012)? Schließlich: Inwiefern spiegeln die Beurteilungsstrategien und -kriterien sowie die schlussendlichen Einschätzungen der Testleistung seitens der BeurteilerInnen das Testkonstrukt wider?

Um nun Einblicke in die genannten Prozesse zu gewinnen, erscheint Introspektion sinnvoll, denn sie bildet die kognitiven Prozesse seitens des Prüflings bei der Bewältigung von Testaufgaben und seitens der Beurteilerin oder des Beurteilers bei der Bewertungsarbeit ab. Trotz des nicht unerheblichen Aufwands, den introspektive Forschungsmethoden mit sich bringen, und der nach wie vor begrenzten Aussagekraft dieser Daten sollten entsprechende Begleituntersuchungen Bestandteil der Qualitätssicherung eines Sprachtests sein. Darüber hinaus geben introspektive Daten ggf. Aufschluss über erfolgreiche und weniger erfolgreiche Teststrategien, was wiederum für didaktische und curriculare Zwecke sowie für Beurteilungstrainings verwendet werden kann.

Introspektive Verfahren erlauben also Einblicke, die über das primäre Anliegen eines Sprachtests hinausgehen: Wir erfahren, wie Sprache verarbeitet wird, welche Strategien dabei nützlich oder aber weniger nützlich sind, was wiederum didaktische Konsequenzen nach sich ziehen kann. Zu unterscheiden ist hierbei zwischen Teststrategien, die dem Testkonstrukt angemessen sind, und jenen, die lediglich zu einem korrekten Testergebnis führen, jedoch den intendierten kognitiven Prozessen und Verstehensstrategien nicht entsprechen. So sollten in einem Leseverstehenstest, der vorgibt, akademische Sprachverwendung zu messen, tatsächlich adäquate Lesestrategien zu beobachten sein, etwa kognitive Fähigkeiten, wie Orientierung im Text, Ausnutzung von Schlüsselwörtern etc. Der Einsatz von – evtl. sogar erfolgreichen, weil zum korrekten Testergebnis führenden – *test wiseness tricks* (Cohen/Upton 2006: 117) untergräbt hingegen die Validität eines Sprachtests. So führt etwa *lucky guessing*, beispielsweise ein zufällig korrekt gelöstes *multiple-choice*-Item, zwar zu einem guten Testergebnis, der Nachweis über die diesem Item zugrundeliegende Fähigkeit ist jedoch nicht erbracht worden.

Ähnlich verhält es sich mit der Frage nach der validen Beurteilung von Prüfungsleistungen. Introspektive Daten erlauben Einblick in die Strategien, Prozesse und begründeten Einschätzungen sprachlicher Leistungen. Die Bewusstmachung und Reflexion der subjektiv geprägten Beurteilungsfaktoren scheint nicht zuletzt für die Professionalisierung der BeurteilerInnen zentral, denn dies bildet die Grundlage für eine begründete Veränderung und Optimierung des Beurteilungsverhaltens, was wiederum von zentraler Bedeutung für die Testqualität ist.

Grundsätzlich sollte unser Ziel sein, zu Selbstreflexion zu animieren. Die Ausbildung von Reflexionsfähigkeit wird gemeinhin als grundlegende Fähigkeit betrachtet, um selbstständig zu lernen, zu arbeiten, Handlungen und Strategien zu überprüfen und begründet zu verwerfen oder beizubehalten.[16] Dies gilt sowohl für Lernende und potenzielle Prüflinge als auch (und insbesondere!) für Lehrende und somit potenzielle BeurteilerInnen sprachlicher Leistungen.

Insgesamt betrachtet nehmen introspektive Verfahren heute einen wichtigen Platz in der Sprachtestforschung ein, insbesondere bei der Erforschung der Validität von Testaufgaben und Beurteilungsverfahren. Sie ergänzen quantitativ erhobene psychometrische Daten. Damit kann das von Bachman formulierte Plädoyer, „to utilize quantitative and qualitative approaches in a complementary fashion", (Bachman 2000: 7) für eine ganzheitliche Sicht auf Sprachtests und ein besseres Verständnis der damit verbundenen Prozesse realisiert werden.

[16] Für den Bereich Fremdsprachenlernen s. die Beiträge in Rampillon/Zimmermann (1997), dort insbesondere Grotjahn. Für die Aus- und Weiterbildung s. Konzepte wie die Aktionsforschung (exemplarisch Altrichter/Posch 1998 sowie Reason/Bradbury 2002).

Für beide Bereiche – Erforschung der Beurteilungsstrategien und Erforschung von *test-taking strategies* – gilt, dass noch etliche Aspekte einer genaueren Untersuchung und Systematisierung bedürfen, nicht zuletzt, um Konsequenzen für die Praxis abzuleiten, sei es für Unterricht, für Aus- und Weiterbildung von Lehrkräften oder für die Testkonzeption. Der vorliegende Beitrag hat versucht hierzu Anregungen zu geben.[17]

Literatur

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

Arras, Ulrike (2007). *Wie beurteilen wir Leistung in der Fremdsprache? Strategien und Prozesse bei der Beurteilung schriftlicher Leistungen in der Fremdsprache am Beispiel Test Deutsch als Fremdsprache (TestDaF)*. Tübingen: Narr.

Arras, Ulrike (2009a). What's on a rater's mind? Die Erforschung von Beurteilungsstrategien und ihre Bewusstmachung durch Schulungsmaßnahmen als Voraussetzungen für die Testvalidität. *Zeitschrift für Angewandte Linguistik (ZfAL)* 50: 33-45.

Arras, Ulrike (2009b). Subjektive Theorien als Faktor bei der Beurteilung von Prüfungsleistungen. In: Berndt, Annette / Kleppin, Karin (Hrsg.): *Sprachlehrforschung: Theorie und Empirie: Festschrift für Rüdiger Grotjahn*. Frankfurt am Main: Peter Lang. 169-179.

Arras, Ulrike / Marks, Daniela / Zimmermann, Sonja (2009). BeurteilerInnen in den Kopf geschaut. Wie das Verfahren des Lauten Denkens im Rahmen von Beurteilungsschulungen eingesetzt werden kann. *DaF-Brücke. Zeitschrift für Deutschlehrerinnen und Deutschlehrer Lateinamerikas* 11: 5-9.

Bachman, Lyle F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing* 17: 1-42.

Bachman, Lyle F. / Palmer, Adrian S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Banerjee, Jayanti / Luoma, Sari (1997). Qualitative approaches to test validation. In: Clapham, Caroline / Corson, David (eds.): *Encyclopedia of Language and Education*. Dordrecht: Kluwer. 275-287.

Beyer, Sabine (2005). Introspektive Verfahren im fremdsprachlichen Unterricht. *Deutsch als Fremdsprache 42*: 18-22.

Buck, Gary (1991). The testing of listening comprehension: An introspective study. *Language Testing* 8: 76-91.

[17] Zu Forschungsdesiderata auf dem Gebiet Beurteilungsstrategien s. Arras 2007. Zu Forschungsdesiderata auf dem Gebiet *test-taking strategies* s. jüngst Kleppin/Reich 2009. Dort auch Versuche, *test-taking strategies* zielgruppengerecht zu kategorisieren.

Cohen, Andrew D. (1987). Using verbal reports in research on language learning. In Færch, Claus / Kasper, Gabriele (eds.): *Introspection in second language research*. Clevedon: Multilingual Matters. 82-95.

Cohen, Andrew D. (1998). Strategies and processes in test-taking and SLA. In: Bachman, Lyle / Cohen, Andrew (eds.): *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press. 90-111.

Cohen, Andrew D. / Upton, Thomas A. (2006). *Strategies in responding to the New TOEFL reading tasks* (=TOEFL Monograph Series Report No. 33). Princeton, NJ: Educational Testing Service. Online: http://www.ets.org/Media/Research/pdf/RR-06-06.pdf. (letzter Aufruf [29.07.2012]).

Cohen, Andrew D. / Upton, Thomas, A. (2007). 'I want to go back to the text': Response strategies on the reading subtest of the new TOEFL. *Language Testing* 24: 209-250.

Dollerup, Cay / Glahn, Ester / Hansen, Carsten R. (1994). 'Sprogtest': A smart test (or how to develop a reliable and anonymous EFL reading test). *Language Testing*, 11/1: 65-81.

Ducasse, Ana Maria (2010). *Interaction in Paired Oral Proficiency Assessment in Spanish: Rater and Candidate Input into Evidence Based Scale Development and Construct Definition.* Frankfurt am Main: Peter Lang.

Ericsson, K. Anders / Simon, Herbert A. (1987). Verbal reports on thinking. In: Færch, Claus/ Kasper, Gabriele (eds.). *Introspection in second language research.* Clevedon: Multilingual Matters. 24-53.

Ericsson, K. Anders / Simon, Herbert A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge: Bradford. 2nd, revised ed.

Green, Alison (1998). *Verbal protocol analysis in language testing research: A handbook.* Cambridge: Cambridge University Press.

Grotjahn, Rüdiger (1997). Strategiewissen und Strategiegebrauch: Das Informationsverarbeitungsparadigma als Metatheorie der L2-Strategieforschung. In: Rampillon, Ute / Zimmermann, Günter (Hrsg.): *Strategien und Techniken beim Erwerb fremder Sprachen.* Ismaning: Hueber. 33-76.

Grotjahn, Rüdiger. (2005). Testen und Bewerten des Hörverstehens. In: Ó Dúill, Micheál / Zahn, Rosemary / Höppner, Kristina D. C. (Hrsg.): *Zusammenarbeiten: Eine Festschrift für Bernd Voss*. Bochum: AKS-Verlag. 115-144.

Grotjahn, Rüdiger (2007). Lernstile/Lernertypen. In: Bausch, Karl-Richard u. a. (Hrsg.): *Handbuch Fremdsprachenunterricht*. Tübingen: Francke. 326-331. 5. Auflage.

Grupa Paul A. (1999). *The role of digital video media in second language listening comprehension.* Ph.D. thesis, University of Melbourne.

Huot, Brian (1988). *The validity of holistic scoring: a comparison of the talk-aloud protocols of expert and novice holistic raters.* Unpublished dissertation. Indiana: University of Pennsylvania.

Huot, Brian (1993). The influence of holistic scoring procedures on reading and rating student essays. In: Williamson, Michael M. / Huot, Brian (eds.). *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Cresskill, NJ: Hampton Press. 206-236.

Kleppin, Karin / Reich, Astrid (2009). Test-Taking Strategien. In: Berndt, Annette / Kleppin, Karin (Hrsg.): *Sprachlehrforschung: Theorie und Empirie: Festschrift für Rüdiger Grotjahn.* Frankfurt am Main: Peter Lang. 95-112.

Kniffka, Gabriele / Gutzat, Bärbel (2003). *Training TestDaF. Material zur Prüfungsvorbereitung*. Berlin: Langenscheidt.

Lodewick, Klaus (2010). *TestDaF-Training 20.15. Vorbereitung auf den Test Deutsch als Fremdsprache*. Göttingen: Fabouda.

Lumley, Tom (2005). *Assessing second language writing: The rater's perspective.* Frankfurt am Main: Peter Lang.

Lumley, Tom / Brown, Annie (2005). Research methods in language testing. In: Hinkel, Eli (ed.): *Handbook of research in second language teaching and learning*. Mahwah, N.J.: Lawrence Erlbaum. 833-855.

May, Lyn (2011). *Interaction in a paired speaking test: The rater's perspective.* Frankfurt am Main: Peter Lang.

McNamara, Tim / Hill, Kathryn / May, Lynette May (2002): Discourse and Assessment. In: *Annual Review of Applied Linguistics*, Volume 22 / March 2002, pp 221-242, Cambridge University Press.

Messick, Samuel (1989). Validity. In: Linn, Robert L. (ed.): *Educational measurement*. New York: Macmillan. 13-103. 3rd ed.

Rampillon, Ute / Zimmermann, Günther (Hrsg.) (1997). *Strategien und Techniken beim Erwerb fremder Sprachen*. Ismaning: Hueber.

Reason, Peter W. / Bradbury, Hilary (eds.) (2002). *Handbook of action research*. London: Sage Publications.

Roche, Jörg-Matthias (Hrsg.) (2005). *Fit für den TestDaF: Tipps und Übungen*. Ismaning: Hueber.

Rossa, Henning (2012). *Mentale Prozesse beim Hörverstehen in der Fremdsprache. Eine Studie zur Validität der Messung sprachlicher Kompetenzen.* Frankfurt am Main: Peter Lang.

Shohamy, Elana (2001). *The power of tests: A critical perspective on the use of language tests*. New York: Pearson.

Stemmer, Brigitte (1991). *What's on a C-test taker's mind? Mental processes in C-test taking*. Bochum: Universitätsverlag Dr. N. Brockmeyer.

TestDaF-Institut: *Hinweise und Tipps zur Vorbereitung*. Online: http://www.testdaf.de/teilnehmer/tn-vorbereitung_tipp.php (letzter Aufruf [29.07.2012]).

Wu, Yi'an (1998). What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing* 15: 21-44.