1.2 Zur Konstruktion von Sprachdiagnoseverfahren

Nun muss sich jede Sprachbeobachtung - auch die vorliegende - notwendigerweise aufgrund der zur Verfügung stehenden Zeit und finanziellen Mittel auf die Beobachtung einiger Aspekte beschränken. Außerdem kann auch mit methodologisch seriösen und gründlichen Untersuchungen immer nur ein gewisser Ausschnitt des Sprachverhaltens erfasst werden. Angesichts der Komplexität von Sprache sollten bei Untersuchungen daher aber möglichst markante Faktoren ausgewählt werden, die für die sprachliche Entwicklung der Altersstufe relevant sind. Außerdem sind die erfassten Fähigkeiten genau zu beschreiben und als Teilbereiche von Sprache zu kennzeichnen. Wenn bei Untersuchungen nur das Hörverständnis, der schriftliche Ausdruck, der phonetisch/ prosodische oder morphosyntaktische Perfektionsgrad von Sprachäußerungen, die funktionalen Anwendungsbereiche oder bei Bilingualen etwa ein angemessener Sprachwechsel beobachtet werden, so mag das in gewissen Zusammenhängen durchaus legitim sein. Aber eine Beschränkungen unterliegende Sprachbeobachtung kann niemals den Anspruch erheben, die Sprachkompetenz eines Menschen vollständig zu erfassen. Nun bleibt aber bei jedem Datenerhebungsverfahren der Eindruck vom Sprachverhalten eines Individuums zwangsläufig lückenhaft, d.h. auch mit noch so umfangreichen Beobachtungen können immer nur relativ objektive, zuverlässige und gültige Aussagen über die untersuchten Sprachfähigkeiten getroffen werden.

Da Untersuchungen im sprachlichen Bereich angesichts der im letzten Abschnitt erläuterten Komplexität von Sprache immer nur eine Stichprobe der Sprachkompetenz erfassen können, kann die inhaltsbezogene Validität einer Sprachbeobachtung nur mit einer angemessenen Auswahl wesentlicher sprachlicher Aspekte einigermaßen erreicht werden. Doch die Validität der bei einem Sprachdiagnoseverfahren zu gewinnenden Ergebnisse wird noch von weiteren Kriterien beeinflusst, die bei der Konstruktion des Erhebungsverfahrens ebenfalls berücksichtigt werden müssen. Daher sollen in diesem Abschnitt neben der Erläuterung allgemeiner Anforderungen an Sprachdiagnoseverfahren vor allem der Einfluss der Erhebungsbedingungen und besondere Schwierigkeiten näher erörtert werden, die bei Sprachbeobachtungen der in dieser Untersuchung maßgeblichen Altersstufe von ca. achtjährigen Kindern auftreten.

Der an ein Sprachdiagnoseverfahren zu stellende Anspruch hängt natürlich auch vom Zweck der mit einem Erhebungsverfahren beabsichtigten Sprachdiagnose ab. Bei einem informellen Test zur Lernkontrolle, bei dem die sprachlichen Leistungen einzelner Schüler innerhalb der Lerngruppe miteinander verglichen werden, mag es ausreichen, wenn ein Lehrer anhand von ihm selbst entworfener Aufgaben etwa zum

neu eingeführten Wortschatz oder einer grammatischen Regel am Ende einer Unterrichtseinheit die Lernergebnisse überprüft und aufgrund der nur von ihm vorgenommenen Auswertung den aktuellen Lernerfolg beurteilt. Auch eine partielle Sprachdiagnose, die als Basis für eine zielgerichtete Unterrichtsplanung dienen soll, kann ein ausreichend sensibilisierter Lehrer nach einem von ihm oder kompetenten Instanzen aufgestellten Kriterienkatalog weitgehend selbständig planen, durchführen und auswerten. Zweckmäßig wird allerdings auch hierbei schon die Hilfe eines Kollegen sein, da ein Lehrer als teilnehmender Beobachter zumindest bei großen Lerngruppen mit der gleichzeitigen Wahrnehmung von Unterrichtsaufgaben und Sprachbeobachtungen überfordert sein dürfte. Insoweit Ergebnisse einer Sprachdiagnose nur lerngruppeninterne Auswirkungen haben, wird sich ein seiner Verantwortung bewusster Pädagoge zwar auch bei solchen Verfahren um gültige, zuverlässige und objektive Beobachtungen bemühen, aber seine Erhebungen müssen keinen strengen wissenschaftlichen Ansprüchen genügen.

Wenn jedoch aus den Ergebnissen von Sprachdiagnoseverfahren selektierende Entscheidungen abgeleitet werden sollen, sind angesichts der möglichen individuellen oder allgemeinen Konsequenzen höhere Anforderungen an das Verfahren zu stellen. Höhere Maßstäbe gelten selbstverständlich auch bei geplanten Veröffentlichungen und wissenschaftlichen Untersuchungen, bei denen vom Forscher eine ethische Haltung erwartet wird, die sich sowohl auf ein methodologisch korrektes Vorgehen bei der Datengewinnung erstreckt wie auf Bedenken hinsichtlich eines möglichen Missbrauchs bei einer Veröffentlichung der empirischen Erkenntnisse.[23] Außerdem sollte linguistische Forschung jedenfalls bei Kindern nicht als Selbstzweck betrieben werden und ein verantwortungsbewusster Umgang nicht nur mit den Daten, sondern auch mit den jungen Probanden selbstverständlich sein.

Bei der schulischen Anwendung von Sprachdiagnoseverfahren, die über die Planung der Lerninhalte, der im Unterricht zu steuernden Lernprozesse oder eine binnendifferenzierende Gruppeneinteilung hinausgeht, verlangt die pädagogische Verantwortung – angefangen bei Gruppeneinteilungen von schulinternen Förderkursen bis hin zu Entscheidungen über die weitere Schullaufbahn – ein ernsthaftes Bemühen um zumindest annähernd gültige Aussagen. Daher sollten bei Sprachdiagnosen in Form von Sprachstandserhebungen bei der Beobachtung, Beschreibung und Bewertung des komplexen Sprachverhaltens die grundlegenden wissenschaftlichen Gütekriterien der Objektivität, Reliabilität und Validität weitgehend beachtet werden.

[23] Siehe Atteslander, *Methoden der empirischen Sozialforschung,* Berlin 1969, S.41-46

Zur Erfassung der sprachlichen Kompetenz konkurrieren derzeit verschiedene Verfahrenstypen, die jeweils den Anspruch erheben, sich an den erwähnten Gütekriterien zu orientieren. Bei den vorliegenden Verfahren ist grundsätzlich zwischen Sprachstandstests und Beobachtungsverfahren zu unterscheiden. Auf den ersten Blick bestechen häufig die praktischen Vorzüge der Sprachtests. Genaue Anleitungen mit vorstrukturiertem Untersuchungsmaterial und Auswertungsschema scheinen vom Testanwender keine besonderen Kenntnisse und Vorbereitungen zu fordern. Die quantifizierten Daten werden in der Regel nur bei einer einmaligen Anwendung erhoben, weshalb sie nur eine relativ geringe Zeitdauer für Durchführung und Auswertung in Anspruch nehmen. Testverfahren stehen im Ruf, ökonomisch und objektiv zu sein, während Beobachtungsverfahren dagegen allgemein als aufwändig und subjektiv gelten.

Zur Einschätzung der tatsächlichen Aussagekraft der verschiedenen Ansätze erscheinen jedoch einige Bemerkungen zur Testmethodik angebracht. Nach Lienert, der in seinem für den deutschen Sprachraum maßgeblichen Standardwerk zur Testtheorie neben den Hauptgütekriterien der Objektivität, Reliabilität und Validität auch die Nebenkriterien der Vergleichbarkeit, Normierung, Ökonomie und Nützlichkeit ausführlich dargestellt hat, ist ein Test zu verstehen als ein wissenschaftliches Routineverfahren

> *„zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung.“* [24]
> *„Standardisierte Tests müssen wissenschaftlich entwickelt, hinsichtlich der wichtigsten Gütekriterien untersucht und unter Standardbedingungen durchführbar sein.“* [25]

Allerdings lassen genauere Analysen von vorliegenden Testverfahren zur Messung des Sprachstands, deren Scheingenauigkeit, Ökonomie versprechende Konstruktion und optisch wirksame Datenpräsentation unerfahrene Testanwender oft beeindrucken, hinsichtlich der Gütekriterien erhebliche Zweifel aufkommen. Einige Testverfahren berufen sich zwar ausdrücklich auf die klassischen Gütekriterien, weisen ihre Einhaltung aber nur unvollkommen, unvollständig oder gar nicht nach.

So gibt Steinert in seiner Dissertationsveröffentlichung zum Allgemeinen Deutschen Sprachtest (ADST)[26] zwar sehr ausführlich die Ergebnisse seiner statistischen Berechnungen zu den Variablen Region, Geschlecht, Schicht, Alter, Schulklasse und

[24] Lienert, *Testaufbau und Testanalyse*, Weinheim 1969³, S.7
[25] Lienert, a.a.O., S.21

[26] Steinert, Allgemeiner Deutscher Sprachtest (ADST), Kürten 1975 (Dissertation) sowie als Test: Braunschweig u. Göttingen 1978

Schultyp an, aber eine Überprüfung der Güte des ADST wurde offensichtlich nur hinsichtlich der Reliabilität in angemessener Weise durchgeführt. Die Angaben zur Objektivität erschöpfen sich in bloßen Annahmen darüber, warum nach Meinung des Verfassers eine objektive Durchführung, Auswertung und Interpretation bei der Anwendung des ADST gesichert sei. Die ziemlich hohen Korrelationen (mit einem Durchschnittswert von $r_V = 0.85$) zwischen den 24 Testsegmenten, 6 Sprachebenen und 4 Sprachfertigkeiten mit dem Gesamtergebnis werden als Hinweis auf eine hinreichende statistische Validität des ADST genannt.[27] Doch bestätigen diese hohen Korrelationen eigentlich nur die Homogenität des Sprachtests und belegen noch nicht, dass die Testaufgaben insgesamt tatsächlich eine allgemeine Sprachfähigkeit erfassen. Statt einer Überpüfung der inhaltlichen Validität, die sich erübrige, weil sie *„ja ohnehin logisch evident“* [28] sei, verweist Steinert lediglich auf die *„eindeutige Zielsetzung“*[29] und ihren Bezug zur Taxonomie sprachlicher Leistungen seines Doktorvaters Messelken. Eine übliche Überprüfung anhand eines äußeren Kriteriums fand anscheinend nicht statt. Da sich die Validität eines Verfahrens aber nur unvollkommen mit der statistisch berechneten internen Validität begründen lässt, erweist sich die Aussage des Verfassers, der ADST sei ein valides Verfahren zur Messung einer allgemeinen deutschen Sprachkompetenz, als reine Behauptung.

Bezüglich der inhaltlichen Validität des ADST bestehen arge Zweifel. Die in der deutschen empirischen Sprachdidaktik häufig zitierte Taxonomie von Messelken[30] erfasst zwar beim Hören, Sprechen, Lesen und Schreiben jeweils viele sprachliche Leistungen auf den linguistischen Ebenen Textematik, Lexematik, Syntagmatik, Morphematik, Phonematik und Prosodie, aber berücksichtigt längst nicht alle bei Aussagen über eine allgemeine Sprachkompetenz wesentlichen sprachlichen Aspekte. Im Grunde wird nur der organisatorische, d.h. vor allem der formale Aspekt der Sprachkompetenz bei gewisser Beachtung der Semantik unter ziemlicher Vernachlässigung des pragmatischen Sprachgebrauchs erfasst. Steinert hat bei der Entwicklung seines Sprachdiagnoseverfahrens aber noch nicht einmal die Ausführungen Messelkens zu den linguistischen Ebenen ausreichend aufgegriffen, so dass sein Test höchstens über den Wortschatz und formale sprachliche Leistungen der damit untersuchten Probanden Aufschluss geben kann. *„Sämtliche inhaltlichen, funktionalen und kommunikativen Aspekte der Sprache“* werden *„ausgeblendet“.*[31]

---

[27] Steinert 1975 a.a.O., S.33
[28] Steinert 1975, a.a.O., S.33
[29] Steinert 1975, a.a.O., S.30
[30] Messelken, Empirische Sprachdidaktik, Heidelberg 1971
[31] Neuland, Sprachtests. Möglichkeiten und Grenzen standardisierter Sprachleistungsmessung, in: *Diskussion Deutsch*, Heft 65, Jahrgang 13, 1982, S.274

Bei eingehender Überprüfung der Testitems und der Erhebungsbedingungen lässt sich auch keines der für die Sprachstandserhebung von Grundschülern gegenwärtig in Deutschland angebotenen Sprachdiagnoseverfahren, die für Erhebungen im Bereich des Zweitspracherwerbs des Deutschen durch Migrantenkinder entwickelt wurden, als standardisiert bezeichnen[32], was im folgenden Abschnitt 1.3 anhand eines kritischen Vergleichs einer Auswahl von gebräuchlichen Verfahren näher dargestellt werden soll.

Abgesehen von der Testkonstruktion ist die Aufbereitung der Daten mangels genauer Angaben oder ohne statistische Kenntnisse oft schwer zu durchschauen. Die Gefahr einer Fehlinterpretation soll kurz an einem Beispiel erläutert werden. Wenn in einem Bericht über eine Untersuchung zur Sprachkompetenz die herausgefundene Korrelation zwischen Hörverständnis und Qualität des Unterrichts von r_{tet} = .40 als *„starker Zusammenhang“* [33] hervorgehoben wird, ohne darauf hinzuweisen, dass bei einer solchen nach statistischen Maßstäben eher niedrigen Korrelation lediglich ein Zusammenhang von 16% zwischen den Variablen besteht, 84% der Varianz hingegen von anderen Faktoren abhängen, bietet die Angabe der Korrelationshöhe trotz ihrer Genauigkeit eine irreführende Information.

Der schwerwiegendste Einwand gegen die meisten Sprachtests richtet sich gegen die bevorzugte Auswahl leicht überprüfbarer Teilbereiche von Sprache und das unseriöse Schlussfolgern von ungeeigneten Indikatoren auf allgemeinere Sprachebenen. Beliebt sind Untersuchungen zum Hörverständnis und zum Wortschatz - oft auch noch beschränkt auf einfach visualisierbare Wortarten wie Nomen, Adjektive und Präpositionen unter Vernachlässigung der Verben, Zeit- und Ortsadverbien und sonstiger in syntaktischen Strukturen häufig auftretender Partikel - oder das Abfragen nur formaler, mitunter ziemlich willkürlich ausgewählt wirkender grammatikalischer Strukturen, die für die Sprachkompetenz des Probanden keineswegs markante Merkmalsausprägungen darstellen.[34]. Vor allem die kommunikative Verwendung der Sprache wird selten berücksichtigt, weil sie empirisch nur mit hohem Aufwand zu beobachten ist.

Da psychometrische Testverfahren überwiegend auf auszählbaren Phänomenen beruhen, klammern sie meistens wesentliche Aspekte der Sprachverwendung aus. Selbst wenn über das isolierte Abfragen von mehr oder weniger relevanten

[32] Siehe Apeltauer, *Gesteuerter Zweitspracherwerb*, Voraussetzungen und Konsequenzen für den Unterricht, München 1987, S.39 - Die Situation hat sich trotz einiger Überarbeitungen seit 1987 nicht wesentlich verändert.

[33] Doyé, *Eine Untersuchung zum Hörverstehen der Schülerinnen und Schüler der Staatlichen Europa-Schule Berlin,* in: Göhlich (Hrsg.) 1998, S.64

[34] Siehe dazu ausführlichere Erörterungen im 3. Abschnitt im Zusammenhang mit allgemeinen sprachlichen Aspekten bzw. altersgemäßen Merkmalen von Sprachkompetenz.

Teilbereichen von Sprachkompetenz hinaus auch die mündliche oder schriftliche Produktion von Texten in das Testverfahren aufgenommen wurde, werden solche komplexen und aussagekräftigen Sprachleistungen in der Auswertung wiederum oft nur auf quantitativ zu erfassende Einzelaspekte reduziert. So wird beispielsweise beim Allgemeinen Deutschen Sprachtest (ADST)[35] statt der Betrachtung des Inhalts etwa hinsichtlich der Kohärenz der Aussage oder der syntaktischen Gestaltung nur der Anteil der Wörter pro Satz oder gar der Buchstaben pro Wort ausgezählt.

Die in Hinblick auf Quantifizierbarkeit konstruierten psychometrischen Testverfahren kommen mitunter zu nahezu absurden Aufgabenstellungen, die sich natürlich auch auf die Glaubwürdigkeit der Auswertung auswirken. So ist z.B. bei der kurz PI genannten Düsseldorfer *Sprachstandsmessung bei Schulanfängern*[36] bei der Aufgabengruppe zur Pluralbildung, die für Schulanfänger einen ohnehin noch nicht sonderlich geeigneten Indikator für Sprachkompetenz darstellt, vorgesehen, dass auf das Zeigen einer Abbildung hin zunächst bei der richtigen Nennung des Begriffs im Singular, z.B. *Buch*, ein Punkt vergeben wird. Sollte das Kind den Begriff nicht kennen, soll der Lehrer ihn nennen und danach zur Pluralbildung auffordern. Dabei wird offensichtlich erwartet, dass ein Kind, das kurz zuvor noch nicht einmal das deutsche Wort *Buch* kannte, auf die ungewöhnliche Pluralform *Bücher* kommen kann, denn nur bei korrekter Pluralbildung darf der Lehrer einen Punkt vergeben.

Insgesamt ist leider festzustellen, dass die sprachlichen Bereiche, die bei den – für kleine Kinder zudem allzu formell und asymmetrisch ablaufenden – Testverfahren ausgewählt werden, und oftmals auch die dabei dargebotenen Bildimpulse zur Überprüfung der Sprachkompetenz wenig geeignet erscheinen. *„Unklar ist daher, was mit diesen ... Verfahren erfasst wird: Ratefähigkeit, Wahrnehmungsfähigkeit oder sprachliche Fertigkeiten.“*[37] Da viele dieser quantifizierenden Sprachtests weder testmethodisch noch linguistisch ausgereift sind, es also oft fraglich ist, was mit diesen Tests „gemessen“ wird, gibt ihr Einsatz höchstens dem in diesem besonderen Bereich ausgebildeten Personal einen gewissen Aufschluss über einzelne Aspekte einer Sprachkompetenz. Auf jeden Fall ist eine schematische Anwendung zu vermeiden, da sie zu gravierenden Fehleinschätzungen führen kann. Eine unkritische Anwendung im schulischen Bereich wäre mit pädagogischer Verantwortung unvereinbar. Bevor solche Verfahren Pädagogen ohne eine spezielle Aus- oder Fortbildung zur Durchführung mit Schülern empfohlen werden, müssten sie dringend überarbeitet werden.

---

[35] Steinert, Allgemeiner Deutscher Sprachtest (ADST), 1975 u. 1978, a.a.O.
[36] Pädagogisches Institut der Landeshauptstadt Düsseldorf (Hrsg.), *Sprachstandsmessung bei Schulanfängern*, überarbeitete Neufassung (von 1980), Düsseldorf 1982
[37] Apeltauer 1987, a.a.O., S.40

Bei der vorliegenden Untersuchung erübrigte sich die Entscheidung zwischen einem Test- oder Beobachtungsverfahren schon aufgrund der Tatsache, dass für die Erfassung einer deutsch/ italienisch bilingualen Sprachkompetenz, zumal in dieser Altersstufe, bislang keinerlei ausgearbeitetes Verfahren vorliegt. Es galt daher sowieso ein eigenes Verfahren zu entwerfen, wobei ich mich entsprechend der vorstehenden Kritik an quantifizierenden Sprachtestverfahren und der folgenden Argumente für die Entwicklung eines strukturierten Beobachtungsverfahrens entschied.

Bekanntlich kann kein Verfahren ohne eine ausreichende Objektivität zu zuverlässigen und gültigen Ergebnissen führen. Nun gelten Beobachtungsverfahren im Allgemeinen gegenüber den „objektiven“ Testverfahren als weniger seriös und werden oft als subjektiv abklassifiziert, womit ihnen von vornherein eine persönlich gefärbte und zufällige Datenerhebung, Auswertung und Interpretation unterstellt wird. Daher erscheint es angebracht, zunächst auf diese pauschale Fehlbeurteilung von Beobachtungsverfahren einzugehen und eine differenziertere Einschätzung der auch von Beobachtungsverfahren zu erreichenden Objektivität zu empfehlen. Bei Beobachtungsverfahren können sehr wohl intersubjektiv überprüfbar wissenschaftlich abgesicherte Daten anhand von systematischen Wahrnehmungen erhoben werden, wenn vor Beginn der Beobachtung möglichst genaue Erhebungsbedingungen vereinbart werden. Außerdem sollte sowohl während der Beobachtung als auch bei der Auswertung weitgehend strukturiertes Material eingesetzt werden, dessen Entwicklung an möglichst aktuellen wissenschaftlichen Erkenntnissen orientiert ist. Unter den genannten Voraussetzungen können auch bei Beobachtungsverfahren bei der Datengewinnung und deren Analyse annähernd objektive Bedingungen hergestellt werden, die hinsichtlich der Zuverlässigkeit der Mess- bzw. Beobachtungsergebnisse den Anforderungen an standardisierte Testverfahren durchaus entsprechen. Übrigens wird in englischsprachigen Publikationen zur Testmethodik die konzeptuelle Nähe der Objektivität zur Reliabilität betont, indem sie häufig nur als Unterbegriff der Reliabilität aufgeführt wird.

Natürlich sind die bei einer Beobachtung feststellbaren Ergebnisse unter anderem abhängig von der Vorgehensweise des dem Probanden gegenüberstehenden, datengewinnenden Beobachters. Dieser nicht zu leugnende Einfluss der Persönlichkeit des Untersuchers, seines Sprachgebrauchs und der gewählten Sprachebene auf die Sprachäußerungen des Probanden macht sich aber außer bei rein schriftlichen Verfahren auch bei Testverfahren bemerkbar. Selbst wenn bei Testverfahren die engen Instruktionen genau eingehalten werden, kann sich doch schon der mehr oder weniger freundliche Tonfall der Anweisungen auf die Qualität und Quantität der Antworten auswirken. Bei einem Beobachtungsverfahren lässt sich die Gefahr einer

die Objektivität beeinträchtigenden Beeinflussung während der Beobachtung durch die gegenseitige Kontrolle durch mehrere geschulte Beobachter während der Durchführung des Verfahrens minimieren. Eine subjektive Auswertung lässt sich durch die Verwendung von Sprachaufzeichnungen vermeiden, die nach einem strukturierten Auswertungsraster entsprechend verbindlich festgelegter Bewertungskriterien von mehreren Beobachtern unabhängig voneinander analysiert werden. Referenzfehlern bei der beurteilenden Interpretation von Verhaltensweisen kann gleichfalls durch unabhängiges Rating mehrerer Beobachter anhand möglichst niedriginferent definierter Schätzskalen vorgebeugt werden.

Unter Beachtung der vorgenannten Bedingungen sind Beobachtungsverfahren nicht zwangsläufig weniger objektiv als herkömmliche Testverfahren, bieten aber gerade durch ihre spezifischen Vorteile mehr Chancen hinsichtlich einer größeren Validität, dem wichtigsten Gütekriterium für jedes Untersuchungsinstrumentarium. Ein Erhebungsverfahren ist bekanntlich nur dann als valide oder gültig zu bezeichnen, wenn es tatsächlich Verhaltensmerkmale erfasst, die es zu messen oder zu beobachten vorgibt. Bei einer Sprachdiagnose sollte also die vom Verfahren erfasste Stichprobe der Sprachkompetenz einen repräsentativen Indikator für das sprachliche Vermögen eines Menschen darstellen. Da Sprache üblicherweise in Situationen angewendet wird, die sprachliches Verhalten erfordern, müssen zur Sprachdiagnose geeignete Verfahren folglich Erhebungsbedingungen herstellen, die zu möglichst natürlichem Sprachgebrauch anregen, ohne die sprachlichen Äußerungen des Probanden zu beeinflussen, was Beobachtungsverfahren eher leisten können als Testverfahren.

Optimal wäre die Beobachtung während natürlicher Kommunikationssituationen, wenn der Sprecher oder Schreiber aufgrund seiner Interessen oder Bedürfnisse frei seine Themen, seine Interaktionspartner und damit auch das angemessene Sprachregister wählen könnte. Eine solche ideale Beobachtungssituation lässt sich in der Praxis aber nur bei Einzelbeobachtungen anwenden, ansonsten erfordern Vergleichbarkeit und Ökonomie doch eine gewisse Steuerung des möglichen Sprachverhaltens. Aber bei Beobachtungsverfahren können immerhin Situationen und Impulse geboten werden, die natürlichen Kommunikationssituationen nahe kommen. Bei jüngeren Kindern können Spielsituationen mit Gleichaltrigen sehr ergiebig sein, wenn die Spielregeln und/oder das Spielmaterial entsprechend der Untersuchungsanordnung ausgewählt wird. In strukturierten Interviews kann ein erfahrener Beobachter immer auch individuell auf den Probanden eingehen, ohne den Untersuchungsbereich aus dem Auge zu verlieren, aber auch ohne für die Untersuchung vielleicht sehr wertvolle freie Äußerungen zu unterbinden.

Eine wesentliche Voraussetzung für das Beobachten natürlicher Sprachäußerungen ist selbstverständlich eine gewisse Vertrautheit zwischen dem Beobachter und dem Probanden, die bei externen Beobachtern durch mehrmalige Hospitationen mit eingeplanten Interaktionen hergestellt werden kann. Bei Testverfahren wird in der Regel in einer asymmetrischen Situation ein Kind mit einem unbekannten Erwachsenen konfrontiert, was zwangsläufig die Validität der Stichprobe herabsetzt, denn gegenüber Fremden ist der Sprachgebrauch immer kontrollierter, wenn nicht bei extremer Unsicherheit sogar eine Sprachverweigerung auftritt. Da in einer vertrauten Sprachdomäne meistens mit anspruchsvolleren sprachlichen Mitteln mehr und flüssiger gesprochen wird als in einer unvertrauten Situation,[38] verfälscht die bei fremden Beobachtern häufig eintretende Sprachvermeidung den Eindruck von der potentiellen Sprachkompetenz. Um die Beobachtungssituation vertrauter zu gestalten, sollten bei Untersuchungen mit Kindern mindestens bis zum Alter von 12 Jahren möglichst befreundete Gleichaltrige durch Partner- oder Gruppenaufgaben miteinbezogen werden.

Auf jeden Fall setzt die Erhebung valider Beobachtungen von Sprachverhalten altersgemäße Untersuchungsanordnungen voraus. Da selbst bei Erwachsenen, die sich freiwillig einer Überprüfungssituation unterwerfen, oftmals stressbedingte Verhaltensänderungen und somit nicht unbedingt repräsentative Ausschnitte ihres Sprachvermögens zu beobachten sind, erscheint eine formale Testatmosphäre bei Schulanfängern und auch noch bei 8-jährigen Kindern wenig angebracht, deren Sprachverhalten in der Muttersprache oder gar in einer Zweitsprache beobachtet werden soll. Bei jüngeren Kindern sind Partner- oder Gruppenaktivitäten sowie spielerische Aufgaben gegenüber Einzelbefragungen zu bevorzugen. Das schriftliche Antworten auf Testfragen oder selbst das Ankreuzen von Multiple-Choice-Aufgaben scheiden wegen der noch unvollkommenen Lese- und Schreibfähigkeiten in dieser Altersstufe selbstverständlich noch aus. Auch das in Sprachdiagnoseverfahren für Schulanfänger häufig anhand von Bildgeschichten verlangte zusammenhängende Erzählen überfordert diese Alterstufe, in der noch eine dialogische Gesprächsstruktur angemessener ist. Daher sollten Bildgeschichten bis zum 3. Schuljahr, ab dem erst die zusammenhängende Wiedergabe eines Handlungsablaufs allgemein geübt wird, höchstens in einzelnen Schritten nacheinander angeboten werden.

Allerdings lässt sich auch bei strukturierten Beobachtungsverfahren das sogenannte Beobachterparadoxon nicht gänzlich vermeiden, das alle empirischen Verfahren gleichermaßen betrifft. Auch Beobachtungsverfahren kommen nicht umhin, zur

[38] Siehe Apeltauer 1987, a.a.O., S.37

Beobachtung eines möglichst natürlichen Verhaltens Methoden anzuwenden, die natürliches Verhalten gerade verhindern. Schon die zur unabhängigen Auswertung durch mehrere Personen unvermeidliche Aufzeichnung des Sprachverhaltens macht den Probanden auf eine gewisse Überprüfungssituation aufmerksam. Dieser Störfaktor einer natürlichen Kommunikation kann jedoch gemindert werden, indem keine aufwendige Apparatur eingesetzt wird oder die Kinder vor der eigentlichen Untersuchung ausreichend Gelegenheit erhalten, sich an die Aufzeichnung ihres Verhaltens zu gewöhnen.

Im Allgemeinen ist die Verwendung von Tonrecordern wohl einer noch auffälligeren Videoaufzeichnung vorzuziehen, obwohl diese wiederum wegen des gleichzeitigen Erfassens nonverbalen Verhaltens von großem Vorteil sein kann. Auf jeden Fall ist der Einsatz von Tonaufzeichnungen günstiger als ein zwangsläufig nur ungenaues Protokollieren der Sprachäußerungen. Vor allem gleichzeitiges Protokollieren während der Untersuchung durch den Untersuchungsleiter, wie es bei etlichen Testverfahren vorgesehen ist, beeinflusst die Datenerhebung erheblich und verzerrt dadurch die Ergebnisse, indem es wegen der Überforderung der Untersucher sowohl die Wahrnehmungs- wie die Urteilsfähigkeit nachweislich beeinträchtigt. So wurden bei der nachträglichen empirischen Prüfung des Düsseldorfer Verfahrens PI neben anderen Mängeln bei knapp einem Viertel aller Messungen Durchführungsfehler sowie eine nur äußerst geringe Bewertungsobjektivität festgestellt, was auf das gleichzeitige Protokollieren und Auswerten zurückzuführen ist. Insgesamt erschienen einer der Autorinnen des Verfahrens, die an der Überprüfung beteiligt war, die Mängel als so gravierend, dass sie sogar eine Revision des Verfahrens verwarf. Leider kommt dieses Verfahren trotz dieses negativen Befundes immer noch häufig bei Kindern nichtdeutscher Muttersprache zum Einsatz.[39]

Bei Untersuchungen mit jüngeren Kindern sind Bildimpulse zur Anregung von sprachlichen Äußerungen oft unerlässlich. Da sie weniger als ältere Probanden durch verbale Impulse in ein Gespräch verwickelt werden können, werden Kindern häufig Bilder vorgelegt, deren Beschreibung den Gebrauch bestimmter Strukturen nahe legt.

[39] Auch zur 1998 durchgeführten Sprachstandmessung der 1594 Schulanfänger im Berliner Bezirk Wedding wurde das Düsseldorfer Verfahren PI (siehe Literaturverzeichnis unter: Pädagogisches Institut der Landeshauptstadt Düsseldorf) trotz der in Fachkreisen spätestens seit 1988 allgemein bekannten erheblichen Zweifel an der Validität eingesetzt, womit hier keineswegs der von Berliner Lehrern seit Jahren geforderte und nach den niederschmetternden Ergebnissen endlich anerkannte Förderbedarf geleugnet werden soll. Allerdings erscheint Kritik angebracht, dass ausgerechnet dieses fragwürdige Verfahren zur Untersuchung ausgewählt wurde, von dessen Revision selbst eine der Autorinnen anlässlich der bei einer empirischen Prüfung gewonnenen Erkenntnisse Abstand nahm. - Siehe: Boos-Nünning/ Gogolin, *Sprachdiagnose bei ausländischen Schulanfängern: Resultate der empirischen Prüfung eines*

Ob sich darüber hinaus jedoch auch ein erhofftes, einer natürlichen Kommunikation nahekommendes spontanes Gespräch ergibt, hängt sehr vom Geschick des Beobachters ab. Statt das Bild wie vorgesehen nur als Sprechanlass zu benutzen, tendieren ungeübte Beobachter dazu, durch verengende sprachliche Impulse ausschließlich auf einer Bildbeschreibung zu beharren. Dabei wäre das Aufgreifen spontaner Äußerungen zu Themen, die das Kind persönlich interessieren, sehr wünschenswert. Sofern aber das geführte Gespräch beim gegebenen Bildkontext verbleibt, können die vorwiegend benennenden oder beschreibenden Äußerungen hinsichtlich einer Analyse der verwendeten syntaktischen Strukturen nur wenig ergiebig sein,

> denn *„Kinder verzichten auf eine Versprachlichung von Informationen, von denen sie – da ja beide Gesprächspartner das Bild vor Augen haben – voraussetzen, daß sie die testende Person kennt. Deshalb sind ihre Äußerungen oft kurz, sprunghaft, von geringer Komplexität und auf die lexikalischen Mittel beschränkt, die der Bildkontext unmittelbar veranlaßt.“*[40]

Leider werden bei den vorliegenden Test- wie Beobachtungsverfahren gleichermaßen oftmals wenig ansprechende und dilettantische Bildvorlagen angeboten, die mehr die Ratefähigkeit des Kindes überprüfen als sprachliche Äußerungen evozieren. Die mangelhafte Qualität der Bildimpulse beeinträchtigt aber auf nicht unerhebliche Weise das zu beobachtende Sprachverhalten. Daher sollte das als Bildimpuls verwendete Bildmaterial sowohl anregend als auch eindeutig sein. Besonders bei der Überprüfung der Kenntnisse von lokalen Präpositionen führt eine zu flächige Darstellung der Bildvorlagen oft zu missverständlichen Interpretationen. Ob die auch möglichen Antworten aber als richtig anerkannt werden, obliegt der subjektiven Entscheidung des Auswerters.

Sollten nicht genügend finanzielle Mittel für die professionelle grafische Gestaltung zur Verfügung stehen, sind auf jeden Fall reale Objekte vorzuziehen, was natürlich einigen Aufwand für den Beobachter verursacht und unter Verzicht auf das verbale Erfassen von Aktionen zwischen handelnden Personen zur Reduktion auf leblose Gegenstände und ihre Eigenschaften zwingt.[41] Bei Untersuchungen zum Zweitspracherwerb sind

---

„Sprachtests“, in: *Deutsch lernen,* Zeitschrift für den Sprachunterricht mit ausländischen Arbeitnehmern, 13. Jg, Hefte 3-4/ 1988, S.3-71

[40] Boos-Nünning/ Gogolin, a.a.0., S.58

[41] Sonst käme es zu absurden Anforderungen an die Beobachter wie bei dem „verschlimmbessernden“ Ansatz einer Überarbeitung des Düsseldorfer Verfahrens durch Una M. Röhr-Sendlmeier, die zur Ausmerzung der unbefriedigenden Abbildungsschwächen vorschlägt, ca. 80 Gegenstände bereitzuhalten, z.B. *„ein Haus halb offen (z.B. aus Bausteinen, Lego oder Pappe) mit Tür, Klingel (z.B. Weihnachtsglöckchen), Fenster, Treppe, Dach, Zimmer mit Stuhl, Tisch, Spielnahrung und -gedeck im ersten Stock (...); 2 kleine Tassen, 2 Gabeln, 2 Portionen Eis: groß und klein (...), 1 Vogel, 1 Katze, 4 Fische, einer an der Angel.“* - Zitat nach Boos-Nünning / Gogolin, a.a.O., S.64

zudem die Bildimpulse hinsichtlich kulturspezifischer Inhalte zu überprüfen. In Frage kommt der bewusste Einsatz kulturspezifischen Materials eventuell bei oft jedoch vernachlässigten Untersuchungen der muttersprachlichen Kompetenz. Ansonsten sind neutrale Bereiche vorzuziehen. Der in Sprachuntersuchungen bei Schulanfängern nichtdeutscher Herkunft häufig verwendete Bereich Familie eignet sich keinesfalls zur Beobachtung von Sprachkompetenz in einer noch unvollkommen beherrschten Zweitsprache, da gerade im familiären Bereich die Muttersprache dominiert und die Zweitsprache noch weitgehend funktionslos ist.

Nach den Ausführungen zur Objektivität und Validität, nach denen strukturierte Beobachtungsverfahren als mindestens so geeignet erscheinen wie Testverfahren, soll nun die Reliabilität aufgegriffen werden. In Hinblick auf dieses Kriterium gelten Testverfahren den situationsabhängigeren Beobachtungsverfahren insbesonders als überlegen. Unter Reliabilität versteht man den Grad der Genauigkeit oder Zuverlässigkeit, mit der ein Verfahren ein bestimmtes Persönlichkeitsmerkmal erfasst. Im Idealfall würde ein Verfahren bei wiederholten Messungen bzw. Beobachtungen desselben Merkmals unabhängig von der Situation dasselbe Ergebnis erbringen.

Im sprachlichen Bereich lassen sich die in der Empirie gebräuchlichen Verfahren zur Überprüfung der Reliabilität aber gerade bei Testverfahren kaum anwenden. Eine Überprüfung durch Testwiederholung scheidet aus, weil sich schon beim ersten Anwenden ein gewisser Übungseffekt einstellt. Die Anwendung von Paralleltests scheitert angesichts der ohnehin schon großen Anzahl notwendiger Einzelerhebungen zur Erfassung des komplexen Sprachverhaltens in der Praxis. Eher möglich erscheint bei Sprachtests noch die Aufteilung des Tests in zwei vergleichbare Hälften, womit sich die sogenannte Testhalbierungszuverlässigkeit oder bei entsprechender Aufgliederung in zahlreiche Einzeltests die interne Konsistenz des Verfahrens bestimmen lässt. Bei Beobachtungsverfahren können Erhebungen zu denselben sprachlichen Aspekten mit verschiedenen Untersuchungsanordnungen quasi parallel ermöglicht werden, wodurch sich aber nicht nur die Reliabilität des Verfahrens, sondern auch die zu seiner Durchführung erforderte Zeit insgesamt erhöht. Außerdem führen bei einer anderen Untersuchungsanordnung natürlich auch weitere Faktoren wie z.B. eine höhere Motivation durch die andere Aufgabenstellung oder eine gleichzeitig spezifisch geforderte Sprachleistung zu abweichendem Sprachverhalten. Jedenfalls ergaben sich entgegen der Erwartung zwischen den parallelformähnlichen Untersuchungsanordnungen der Pilotfassung des an der SESB benutzten Verfahrens außer beim mündlichen Sprachgebrauch (in der starken Sprache) nur relativ niedrige Interkorrelationen.

Zwar bildet eine ausreichende Reliabilität neben der objektiven Datenerhebung eine wichtige Voraussetzung für die Gültigkeit eines Verfahrens, aber eine hohe Reliabilität garantiert nicht unbedingt auch eine hohe Validität des Verfahrens. In der Tat besteht zwischen diesen beiden Gütekriterien eine nicht auflösbare Antinomie. Wie jedes Untersuchungsverfahren einem unvermeidlichen Standardmessfehler unterliegt, können sich auch Sprachstandserhebungen der geforderten Reliabilität immer nur annähern. Bei genauerer Analyse erweisen sich im sprachlichen Bereich die angeblich zuverlässigeren Sprachtests wegen ihrer einmaligen Anwendung im Vergleich mit Beobachtungsverfahren als weniger zuverlässig, denn eine einmalige Erhebung kann niemals, und zwar unabhängig von der Qualität des Erhebungsverfahrens, einen repräsentativen Ausschnitt der komplexen und relativ labilen Sprachkompetenz erfassen.

Jedes Verfahren wird bei einer einzelnen Erhebung immer nur eine zufällige Stichprobe eines möglichen Sprachverhaltens darstellen. Ein Mensch wird je nach seiner Befindlichkeit an einem bestimmten Tag in einer gewissen Situation jeweils nur einen relativ begrenzten Ausschnitt aus dem ganzen Komplex seines potentiellen Sprachvermögens zeigen. Auch Sprachstandsbeobachtungen, die innerhalb des Entwicklungsprozesses des Spracherwerbs die Sprachkompetenz in einem zeitlich begrenzten Stadium analysieren, kommen daher hinsichtlich Validität und Reliabilität nicht umhin, mehrere Beobachtungen innerhalb eines gewissen Zeitraums zu erheben.[42] Da sich die bei Beobachtungsverfahren erhobenen Beobachtungen aber sowieso meistens über einen längeren Zeitraum erstrecken, in dessen Verlauf mehrere Erhebungen vorgesehen sind, erfüllen Beobachtungsverfahren das Kriterium der Reliabilität möglicherweise eher als einmalig durchgeführte Testverfahren, deren Vorteil letztlich nur in der größeren Ökonomie liegt. In Anbetracht der durch den geringeren Personal- und Zeitaufwand eingebüßten Qualität erscheint die größere Ökonomie allerdings als recht zweifelhafter Vorzug gegenüber einem sicherlich mehr Validität bietenden *strukturierten* Beobachtungsverfahren.

Um bei Beobachtungen aussagekräftige Daten sammeln zu können, müssen die sprachlichen Verhaltenskategorien schon während der Beobachtung möglichst in operationale Beschreibungen von äußerlich wahrnehmbarem Verhalten aufgegliedert werden. Intuitive grobe Einschätzungen wie *spricht korrekt* oder *hat eine mangelhafte Aussprache* haben nur einen geringen Informationswert, weil sie zwangsläufig subjektive Schlussfolgerungen ausdrücken. Statt hochinferenter Beurteilungen präzisieren niedriginferente Beschreibungen wie *unterscheidet deutlich bei langen und*

[42] Apeltauer 1987, a.a.O., S.37

kurzen Vokalen oder *verwendet auffallend oft Füllpartikel wie äh, hm usw.* z.B. die Beobachtungen phonetischer bzw. prosodischer Merkmale des Sprachgebrauchs. Ein strukturiertes Beobachtungs- und Auswertungsraster ist Voraussetzung für eine objektive Beurteilung des Sprachstandes. Im Gegensatz zur quantitativen Auswertung, die sich in der Auszählung von korrekten oder fehlerhaften Leistungen erschöpft, ermöglichen operationalisierte Beschreibungen darüber hinaus aber auch eine qualitative Auswertung der Beobachtungen, da die Defizite und Fähigkeiten kategorisiert werden können. Eine qualitative Analyse bietet dem Pädagogen außer der Sprachdiagnose zugleich eine brauchbare Grundlage zur gezielten Planung fördernder Maßnahmen.

Zur Beurteilung der Wahrnehmungen müssen die Ergebnisse der Beobachtung außerdem mit vergleichbaren Leistungen in Beziehung gesetzt werden. Dabei können die Ergebnisse je nach Wahl des Vergleichsmaßstabs[43] entweder mit einer individuellen Bezugsnorm (z.B. wie in einem Lernentwicklungsbericht mit dem Lernfortschritt des Schülers in einem bestimmten Zeitraum), einer sozialen Bezugsnorm (bezugsgruppenorientierter Vergleichsmaßstab oder Realnorm, bei klasseninternen Lernkontrollen z.B. die Leistungsverteilung aller Mitschüler oder bei standardisierten Testverfahren z.B. eine Eichstichprobe, klassisches Beispiel: die Normalverteilung bei Intelligenztests) oder einer idealtypischen Bezugsnorm (kriterienbezogenener Vergleichsmaßstab oder Sollnorm, z.B. ein Lehrzielkatalog) verglichen werden, wobei eine kriterienorientierte Bezugsnorm gegenüber einem bezugsgruppenorientierten Vergleichsmaßstab als sachlicher gilt.

Je nach dem Zweck des Verfahrens und der Verfügbarkeit einer als gesichert geltenden absoluten Norm kann der eine oder der andere Maßstab die Ansprüche an eine Vergleichbarkeit der Ergebnisse eher erfüllen. Auf jeden Fall sollte die einem Verfahren beigefügte Anleitung zur Auswertung oder ein Untersuchungsbericht genaue Angaben über die verwendete Vergleichsnorm enthalten, bei einer Eichstichprobe z.B. Angaben über die soziokulturellen Faktoren und die Anzahl der einbezogenen Probanden, damit Beurteilung und Interpretation der Ergebnisse nachvollzogen werden können. Sprachstandsdiagnosen im Zweitsprachbereich sollten

[43] Bei der Benennung der Vergleichsmaßstäbe wird hier auf die im deutschen Sprachraum in der pädagogischen Diagnostik übliche Unterscheidung zwischen normorientierten und kriterien- bzw. lehrzielorientierten Maßstäben verzichtet, da sich schließlich alle Vergleichsmaßstäbe auf Normen beziehen. Dieser missverständliche Sprachgebrauch ist nach Doyé auf einen Übersetzungsfehler eines ursprünglich in englischer Sprache veröffentlichten Artikels (R.Glaser, *„Instructional Technology and the Measurement of Learning Outcomes“,* in: American Psychologist 1963, S.519-521) zurückzuführen. Im Englischen wird der Begriff der Norm abweichend vom Deutschen nur für soziale Wertmaßstäbe verwendet, ansonsten wird von *criterions* oder *standards* gesprochen. Siehe: Doyé, *Die Feststellung von Ergebnissen des Englischunterrichts*, Hannover u. Dortmund 1981, S.52

sinnvollerweise durch Vergleichsdaten über die muttersprachliche Sprachkompetenz ergänzt werden, was von vielen vorliegenden Sprachdiagnoseverfahren leider noch vernachlässigte Beobachtungen zur Herkunftssprache voraussetzt.

Zum Schluss dieses Abschnittes sollen die Anforderungen, die an Aussagen über die Sprachkompetenz von Kindern treffende Sprachdiagnoseverfahren zu stellen sind, noch einmal zusammengefasst werden. Dabei sollte sich die Entwicklung eines geeigneten Verfahrens meines Erachtens zwar nicht an den Konstruktionsprinzipien psychometrischer Sprachtestverfahren orientieren, aber mit gewissen Abweichungen bzw. Ergänzungen doch die von Lienert für wissenschaftliche Untersuchungsverfahren dargestellten Gütekriterien der Objektivität, Reliabilität, Validität, Vergleichbarkeit, Normierung, Ökonomie und Nützlichkeit anstreben.

Demnach sollten bei der Entwicklung von Sprachdiagnoseverfahren für die Zielgruppe von Kindern im Primarbereich beachtet werden:

- empirisch objektive und zuverlässige Methoden bei der Datenerhebung
- Auswahl von für die jeweilige Altersstufe gültigen und markanten Sprachleistungen
  d.h. auch: bei Zweisprachigen unter Einbezug beider Sprachen
- pädagogisches Verantwortungsbewusstsein

Bei dem kritischen Vergleich einiger gebräuchlicher Verfahren aus dem Bereich der Sprachstandsdiagnostik im Deutschen als Zweitsprache, der anhand dieser Kriterien im nächsten Abschnitt erfolgt, kann zwar keines der geprüften Verfahren in allen Bereichen überzeugen, aber immerhin erscheinen die aufwendigeren *strukturierten* Beobachtungsverfahren in Bezug auf eine valide Einschätzung der Sprachkompetenz als eine annehmbare Alternative zu den rein psychometrischen Testverfahren.