technische universität
dortmund

Testmanual Delfin 4 - Teil 2: Messgüte

Lilian Fried
unter Mitarbeit von Eva Briedigkeit, Patrick Isele und Rabea Schunder

2009

Die Etablierung pädagogischer Sprachdiagnostik sowohl im Früh-, als auch im Elementarbereich ist Teil umfassender politisch-pädagogischer Reformbestrebungen, welche sich dem Ziel unterordnen, die Qualität der Sprachbildung bzw. –förderung in unseren Kindertageseinrichtungen weiterzuentwickeln (vgl. Fried 2008). Das gelingt allerdings nur, wenn die Diagnoseverfahren spezifischen Qualitätsstandards genügen. Solche Standards betreffen zum einen die sprachtheoretische Fundierung und zum anderen die Messgüte.

Im ersten Teil des Testmanuals ist bereits dargestellt worden, wie die evidenzbasierte sprachtheoretische Basis von Delfin 4 beschaffen ist. Der hiermit vorgelegte zweite Teil des Testmanuals zeigt nun auf, wie bei der Konstruktion von Delfin 4 mit Hilfe spezifischem lingustischem Fachwissen sowie mit adäquaten teststatistischen Methodologien gewährleistet wurde, dass messmethodische Standards erfüllt sind.

Ausgangspunkt ist das im ersten Teil des Testmanuals zu Delfin 4 erläuterte Sprachkompetenzmodell (Konstruktionsrational). Nachfolgend wird nun dargelegt, wie dieses Modell in Aufgaben bzw. Items übersetzt, diese zu Testformen zusammengefügt sowie auf ihre Messgüte hin empirisch geprüft und revidiert worden sind.

1. Testaufbau

Die Aufgabenkonstruktion ging vom Sprachkompetenzmodell aus. Dabei orientierte sich die „Übersetzung“ der einzelnen Komponenten dieses Modells in Testaufgaben vorwiegend am Erkenntnsstand zu solchen Aufgabenformen, welche einerseits von vierjährigen Kindern bereits bewältigt werden können und andererseits nachweislich reliable bzw. valide Messungen von deren Sprachleistungen erbringen können. Die Auswahl geeigneter Testaufgaben musste dabei den unterschiedlichen Funktionen von Stufe 1: „Besuch im Zoo“ (BiZ) und Stufe 2: „Besuch im Pfiffikus-Haus“ (BiP) angepasst werden.

Stufe 1: Besuch im Zoo (BiZ)

Angesichts der Grobscreeningfunktion von Stufe 1 BiZ, dessen Aufgaben jedes Kind durchlaufen muss, galt es, einen zeitökonomischen Gruppentest zu entwickeln, mit dem bis zu vier Kinder gleichzeitig in ausgewählten Aufgabenbereichen eingeschätzt werden können.

Derzeit werden bei der Feststellung der Sprachleistungen von vierjährigen Kindern Gruppentests eher selten eingesetzt. Meist kommen sie erst bei Kindern zum Einsatz, die schon über schriftsprachliche Fähigkeiten verfügen.[1] Aufgrund des jungen Alters der Zielkinder unseres Verfahrens können jedoch ihre Leistungen nur mündlich erfasst werden. Bei einem Gruppentestverfahren muss also in besonderer Weise sichergestellt werden, dass jedes Kind genügend Raum für die Bearbeitung der Aufgaben hat und möglichst wenig durch die anderen Kinder beeinflusst oder beeinträchtigt wird.

Daher wurde das Verfahren in eine „Rahmenhandlung“ eingebettet, welche wiederum in eine brettspielähnliche Form übersetzt wurde, um Kindern und Testleiterinnen die Anwendung zu vereinfachen. So ist für alle nachvollziehbar, dass stets nur ein Kind an der Reihe ist. Didaktisch unterstützt wird dieses „Nacheinander“ durch das Setzen von Spielfiguren. Die Aufgaben sind in Textform auf Aktionskarten festgehalten (pro Kind und Aufgabentyp eine Karte). Das Kind, das einen Spielzug ausführt, gelangt an eine

[1] Vgl. zu den gängigen Verfahren zur Sprachstandsbestimmung bei Kindergartenkindern Fried (2004);

bestimmte Stelle des Spielplans, der ein spezieller Aufgabentyp zugeordnet ist. Die Testleiterin kann Anleitungstext und Aufgaben von der entsprechenden Aktionskarte ablesen. Diese Anordnung soll den Testleiter/innen das Auswendiglernen der Aufgaben oder ein gleichzeitiges Hantieren mit einem weiteren Schriftstück während der Durchführung ersparen.

Um zu vermeiden, dass Kinder die Lösungen der Aufgaben schon hören, bevor sie selbst agieren, entschieden wir uns dafür, vier möglichst gleichwertige Aufgabensätze zu entwickeln. Da sich bei den Erprobungen unsere Vermutung bestätigte, dass das viermalige Stellen der gleichen Aufgabe (wenn auch mit unterschiedlichem Inhalt) irritierend und ermüdend auf die Kinder wirkt, veränderten wir die Reihenfolge, in der den Kindern die Aufgaben jeweils gestellt werden sollten. Um dies bei der Durchführung nicht künstlich erscheinen zu lassen, startet jedes Kind mit seiner Spielfigur an einem anderen Ort auf dem Spielplan und erledigt dementsprechend unterschiedliche Aufgaben.

Die folgende Übersicht zeigt, in welcher Reihenfolge die Kinder die Aktionskarten bearbeiten:

Runde	Delfinbecken	Giraffengehege	Tigerkäfig	Elefantengehege
1	blaue Spielfigur	grüne Spielfigur	gelbe Spielfigur	rote Spielfigur
2	rote Spielfigur	blaue Spielfigur	grüne Spielfigur	gelbe Spielfigur
3	gelbe Spielfigur	rote Spielfigur	blaue Spielfigur	grüne Spielfigur
4	grüne Spielfigur	gelbe Spielfigur	rote Spielfigur	blaue Spielfigur

Tabelle 1: Bearbeitung der Aktionskarten

An der Durchführung des Testverfahrens sind zwei Personen beteiligt: Eine der Beiden Testleiter/innen führt das Verfahren mit den Kindern durch, die andere hält die Leistungen der Kinder fest. Um die Protokollierung zu erleichtern, wurde ein Protokollheft entwickelt, mit dem alle vier Kinder während eines Testdurchgangs eingeschätzt werden können, ohne dass mit mehreren Bogen hantiert werden muss. In diesem Heft entspricht die Reihenfolge der Aufgaben pro Kind der Testdurchführung.

Stufe 2: Besuch im Pfiffikus-Haus (BiP)

Das Feinscreening von Stufe 2, das nur bei den Kindern zum Einsatz kommt, deren Sprachleistungen in Stufe 1 potentielle Risiken indizierte, wurde dagegen von vornherein als Einzeltestverfahren konzipiert. Auch hier soll die thematische Rahmung (Pfiffikus-Haus) die Motivation des Kindes unterstützen und die Aufgaben in einen nachvollziehbaren Sinnzusammenhang betten. Um die Anwendung für die Testleiter/innen zu vereinfachen, ist die Gestaltung der Materialien der Stufe 1 und 2 ähnlich. So sind auch bei dem Einzeltestverfahren die Aufgaben auf Karten abgedruckt, die das Kind in einer vorgegeben Reihenfolge ziehen und bearbeiten soll. Anders als bei Stufe 1 sind hier für diejenigen Aufgaben, welche in Stufe 1 nicht auftauchen, weitere Bildmaterialien erforderlich, deren Einsatz durch die Rahmenhandlung gesteuert wird.

2. Testaufgaben

Die Testaufgaben von Delfin 4 sollen Sprachentwicklungsaspekte erfassen, die nach Erkenntnissen der internationalen Spracherwerbstheorie und –forschung als curriculumrelevant, entwicklungssensitiv und risikoindizierend gelten.[2]

Dieser Anspruch wurde durch die Entwicklung eines theoretisch und empirisch gestützten Sprachkompetenzmodells gewährleistet, das ein differenziertes Verständnis des Aufbaus, der Entwicklung und somit auch der Diagnose und Förderung von bereichsspezifischen Kompetenzen vermittelt. Ein derartiges Modell benennt konkrete Anforderungen, die in spezifischen Sprachbereichen von einem Kind auf einer bestimmten Entwicklungsstufe bewältigt werden können. Die somit ermittelten Teilkompetenzen ermöglichen es zu erfassen, was ein Kind einer bestimmten Altersgruppe sprachlich zu leisten vermag bzw. was ihm noch nicht zugänglich ist.

Das Sprachkompetenzmodell beschränkt sich dabei auf Teilbereiche, die mit bewährten Elizitierungstechniken erfasst werden können. Das Modell wird also unter Beachtung des Erkenntnisstands zur Testkonstruktion in einen Itempool übersetzt.

[2] Vgl. dazu die Ausführungen im Arbeitsbericht „Sprachtheoretischer Hintergrund“.

(Teil-)Aufgabenbereiche	Teilaufgaben	Itempool Stufe 1	Itempool Stufe 2
Artikulation	• An-, In-, Endlaute bilden	-	(X)*
Lexik-Semantik (Quantität und Qualität des aktiven/ passiven Wortschatzes)	• Wortverständnis	**-**	**X**
	• Wortproduktion	**-**	**X**
	• Begriffsklassifikation	**-**	**X**
Morpho-Syntax (Satzverstehen, Satzproduktion, grammatische Formen bilden)	• Ausführung von Handlungsanweisungen	**X**	
	• Produktion von Mehrwortsätzen	**X**	**X**
	• Pluralbildung	**-**	**X**
Pragmatik (Textverstehen, Textproduktion)	• Bildbeschreibung	**X**	**-**
	• Bilderzählung	**-**	**X**
Metasprachliche Fähigkeiten/ Arbeitsgedächtnis (sprachlich-akustische Verarbeitungsprozesse; Phonembewusstheit)	• Lautgedächtnis	**X**	**X**
	• Silben erkennen	**(X)****	**-**
	• Reime identifizieren/bilden	**(X)****	**-**

Tabelle 2: Vom Modell zum Test
* Nur beiläufig, nicht durch spezifischen Untertest erfasst;
** Nach den ersten Erprobungen verworfene Untertests;

Für den Gruppentest der Stufe 1 wurden die Teilaufgaben ausgewählt, die laut Forschungsstand mit nur wenigen Items reliabel, diskrimativ valide und ohne vielfältige (Bild-)Materialien umgesetzt werden können. Aufgabenbereiche bzw. Teilaufgaben, die nur mit einer größeren Anzahl von Items und durch bildunterstützte Elizitierungstechniken erfasst werden können, bleiben dem Einzeltestverfahren der Stufe 2 vorbehalten.

Da mit jedem Aufgabenbereich jeweils spezifische Sprachkompetenzbereiche erfasst werden sollen, und die Itemkonstruktion dementsprechend variiert, werden die Schritte der Itemkonstruktion jeweils der Logik eines Aufgabenbereichs folgend dargestellt. Bei dem Gruppentest-Verfahren der Stufe 1 handelt es darüber hinaus um die Konstruktion

von Aufgaben, die jeweils in vierfacher Menge parallel konstruiert worden sind, damit sie im Anspruch, Schwierigkeitsgrad bzw. bezüglich ihrer Trennschärfe möglichst ähnlich sind (vier Parallel-Tests). Dies war bei der Konstruktion der Items der Stufe 2 nicht erforderlich. Aufgrund dieser Besonderheit wird im Folgenden die Entwicklung der Items von Stufe 1 und Stufe 2 getrennt voneinander beschrieben.

3. Aufgabenpool

Konstruktion der Aufgaben zu Stufe 1

Stufe 1: Aufgabenbereich Morpho-Syntax; Untertest HA

Mit der Aufgabe „Handlungsanweisungen ausführen (HA)“ wird überprüft, inwiefern ein Kind schon Sätze unterschiedlicher Komplexität zu verstehen vermag. Daher gibt dieser Untertest Auskunft über das erworbene grammatische Regelwissen, indem die Kinder aufgefordert werden, vorgegebene Sätze in Handlungen umzusetzen. Diese sog. Manipulationsaufgaben eignen sich durch ihren hohen Aufforderungscharakter besonders als Einstieg in das Testverfahren.[3]

Bei der Konstruktion der angebotenen Sätze wurde darauf geachtet, die Komplexität der Sätze zu variieren. Von den in den Sätzen enthaltenen Wörtern kann aufgrund von Listen zum Alterswortschatz angenommen werden, dass sie den Kindern bekannt sind, so dass hier das Verstehen der grammatischen Struktur und nicht die Wortschatzleistung im Vordergrund steht. Bei der Konstruktion der Sätze wurde darauf geachtet, dass die Handlungen vom Kind nur ausgeführt werden können, wenn es die Struktur des Satzes verstanden hat (Mehrfachaufträge). Dabei werden dem Kind zunächst grammatisch einfachere, dann komplexere Instruktionen gegeben, die zu den entwicklungssensitiven und risikoindizierenden Aufgaben gehören und dazu beitragen können, zwischen sprachunauffälligen und möglicherweise sprachauffälligen Kindern unterscheiden zu können.[4] Um während der Testdurchführung nicht mit vielen zusätzlichen Materialien hantieren zu müssen, werden die Aufgaben lediglich mit der Spielfigur durchgeführt, die das Kind laut Instruktion auf dem Spielplan bewegen soll.

[3] Manipulationsaufgaben haben sich in der Sprachstandsdiagnostik schon lange bewährt (vgl. HSET (Grimm/Schöler 1991); WET (Kastner-Koller/Deimann 1998);
[4] Vgl. Fredrick u.a. 1984; Hirsh-Pasek/Golinkoff 1991;

	Anweisung	Leistung
1.	**Stelle deine Figur auf** [unbestimmter Artikel + Substantiv (Tiername)]**.**	Einzelauftrag ausführen (Einzelnomen erkennen)
2.	**Stelle deine Figur neben** [bestimmter Artikel + Substantiv (Bezeichnung für Mensch)] mit [bestimmter Artikel + Substantiv (kein Lebewesen, kein Mensch)]**.**	Zweifachauftrag ausführen (Nominalgruppe bestehend aus Nomen und Attribut erkennen)
3.	**Stelle deine Figur auf das Dach** [Phrase im Genitiv: bestimmter Artikel + Substantiv (Gebäude/Behausung) + Attribut]**.**	Dreifachauftrag ausführen (Nominalgruppe bestehend aus attributivem Adjektiv, Nomen und Präpositionalgruppe erkennen)
4.	**Stelle deine Figur zu** [bestimmter Artikel + Substantiv (Lebewesen) + Attribut (Relativsatz)**.**	Zweifachauftrag ausführen (Nomen mit Relativsatz erkennen)
5.	**Zeige mit dem Finger auf** [unbestimmter Artikel + Substantiv (Tiername)] und stelle deine Figur zurück auf dein Feld bei den Delfin-Karten.	Dreifachauftrag mit mehreren Handlungsschritten (Verbsemantik und Zeitverhältnisse der Handlungsschritte erkennen)

Tabelle 3: Konstruktion HA Stufe 1 (Pilot 2007)

Die Reihenfolge der Items und der Inhalt der jeweiligen Handlung variiert von Kind zu Kind unter Beibehaltung der syntaktischen Struktur der Instruktion. So soll die Wahrscheinlichkeit verringert werden, dass sich die Kinder beobachtete Handlungsmuster einprägen können. Anhand der Aufgaben des ersten Kindes (blau) können die anderen Aufgaben der ursprünglichen Reihenfolge zugeordnet werden.

	blau	grün	rot	gelb
1.	Stelle deine Figur auf einen Delfin. (Nr. 1)	Stelle deine Figur auf eine Giraffe. (Nr. 1)	Stelle eine Figur auf einen Elefanten. (Nr. 1)	Stelle deine Figur auf einen Tiger. (Nr. 1)
2.	Stelle deine Figur neben den Mann mit dem Fisch. (Nr. 2)	Stelle deine Figur neben die Frau vor dem Giraffengehege. (Nr. 4)	Stelle deine Figur neben die Frau mit dem Regenschirm. (Nr. 2)	Stelle deine Figur neben das Kind mit dem Teddy. (Nr. 3)
3.	Stelle deine Figur auf das Dach des Häuschens am Zooeingang. (Nr. 3)	Stelle deine Figur auf das Dach des Schlangenhäuschens. (Nr. 2)	Stelle deine Figur auf das Dach vom Zelt mit dem Elefanten. (Nr. 4)	Stelle deine Figur auf das Dach vom Tigerkäfig. (Nr. 2)
4.	Stelle deine Figur zu den Kindern, die auf dem Pony reiten. (Nr. 4)	Stelle deine Figur zu dem Zebra, das die Blume frisst. (Nr. 3)	Stelle deine Figur zu dem Elefanten, der die Blume frisst. (Nr. 3)	Stelle deine Figur zu dem Tiger, der auf dem Baum liegt. (Nr. 4)
5.	Zeige mit dem Finger auf einen Seehund und stelle deine Figur zurück auf dein Feld bei den Delfin-Karten. (Nr. 5)	Zeige mit dem Finger auf ein Zebra und stelle deine Figur zurück auf dein Feld bei den Delfin-Karten. (Nr. 5)	Zeige mit dem Finger auf einen Löwen und stelle deine Figur zurück auf dein Feld bei den Delfin-Karten. (Nr. 5)	Zeige mit dem Finger auf einen Papagei und stelle deine Figur zurück auf dein Feld bei den Delfin-Karten. (Nr. 5)

Tabelle 4: Aufgaben HA Stufe 1 (Pilot 2007)
* In Klammern ist angegeben, welche Nummer das Item in der Pilotierungsform erhalten hat.

Evozierungsmethodik:

Die Anweisungen werden dem Kind mit möglichst normaler Intonation vorgesprochen, so dass kein besonderer Hinweis auf die Lösung der Aufgabe gegeben wird. Zu Beginn muss dem Kind klar gemacht werden, dass es wichtig ist, sich erst die gesamte Aufgabe anzuhören, bevor es agiert.

Stufe 1: Aufgabenbereich Phonemgedächtnis; Untertest (KN)

Bei dem Nachsprechen von Kunstwörtern handelt es sich um eine schon seit vielen Jahren eingesetzte, bestens erforschte und bewährte Methode der Sprachstandsdiagnostik. Sie hat sich als besonders reliabler und ökonomischer Aufgabentyp etabliert.[5] Bei dieser Aufgabe werden dem Kind Wörter vorgesprochen, die es unmittelbar nachsprechen soll. Vor der Konstruktion sichteten wir die einschlägigen Verfahren aus dem deutschsprachigen und angloamerikanischen Raum, in denen die Aufgabe „Kunstwörter nachsprechen“ zum Einsatz kommt (u. a. Mottier (1951); Gathercole u.a. (1994); HASE (2001/02); BISC; SETK).[6] Aufgrund des hohen Aufwands, der mit der Konstruktion und Überprüfung der Messgüte von Testaufgaben verbunden ist, wird in einigen der deutschen Verfahren bei der Auswahl der Testitems auf schon vorhandene Verfahren zurückgegriffen.[7] So sind in HASE die Kunstwörter aus IDIS - Inventar diagnostischer Informationen bei Sprachentwicklungsauffälligkeiten (Schöler 1999) verwendet worden. Im BISC - Bielefelder Screening (Jansen u.a. 1999) nutzten die Autoren die als Prädiktorgruppe P2 bezeichnete Aufgabengruppe zur Prüfung auditiver Differenzierungsleistungen und sprechmotorischer Koordination von Tiedemann, Faber und Kayshra (1985). Grimm entwickelte für den SSV/SETK (2000) eigene Kunstwörter und stützte sich bei der Konstruktion auf Gathercole u.a. (1994).

Vor allem zwei Kriterien sind dem aktuellen Forschungsstand folgend für die Konstruktion der Kunstwörter, deren Nachsprechen zur Auslotung des kindlichen Arbeitsgedächtnisses dienen soll, entscheidend; und zwar die Silbenzahl[8] und die Wortähnlichkeit[9]. Der Schwierigkeitsgrad der Nachsprechaufgaben hängt bedeutsam mit diesen beiden Faktoren zusammen.

---

[5] So z.B. Gray (2003); Reuterskiöld-Wagner./Sahlén/Nyman (2005); Bishop/North/Donlan (1996); Gathercole et al. (1994);
[6] The Children’s Test of Nonword Repetition (CNRep);
[7] Selbst in den Beobachtungsbogen SISMIK (Ulich, Mayr) hat die Aufgabe „Kunstwörter nachsprechen“ Eingang gefunden. Hier werden als Beispiele für viersilbige Unsinnswörter zwei Beispiele aus IDIS bzw. HASE (FODEKINA; RIBANELU) angeführt (die Autoren geben als Quelle den Mottier-Test (1951) an - hier scheint es zu einer Verwechslung gekommen zu sein);
[8] Vgl. Snowling u.a. (1991); Grey (2003); Radeborg u.a. 2006;
[9] Vgl. Masterson u.a. (2005);

Folgt man den mehr oder weniger expliziten Ausführungen zur Aufgabenkonstruktion der oben genannten Verfahren, wird einerseits deutlich, dass das Kriterium der „Wortähnlichkeit“ schwieriger zu kontrollieren ist, als das der Silbenzahl. Zum anderen zeigt sich, dass manche Verfahren durch das Nachsprechen nicht nur Aufschluss über die Kapazität des Arbeitsgedächtnisses, sondern auch über die artikulatorische Leistungsfähigkeit des Kindes erhalten wollen.

Da es unmöglich ist, mit einer nur geringen Anzahl von Aufgaben die gesamte Artikulationsfähigkeit eines Kindes einzuschätzen und für die Tests von Delfin 4 möglichst wenige, aber sehr aussagekräftige Aufgaben benötigt werden, sollten die konstruierten Aufgaben nur zur Auslotung des Arbeitsgedächtnisses, nicht jedoch zur Erfassung der Artikulationsfähigkeit dienen.[10]

Für den deutschen Sprachraum ist der Mottier-Test (1951)[11] lange in Verwendung und daher gut dokumentiert. Hierbei handelt es um einen Test, der sowohl bei Kindern im Vorschul- als auch Schulalter eingesetzt wird für den mittlerweile auch Normwerte jüngeren Datums vorliegen.[12] Kritisiert werden vor allem konzeptionelle Schwächen der Itemkonstruktion und die hohe, unökonomische Itemanzahl.[13]

Daher war es unser Anliegen, die Items für die Aufgabe „Kunstwörter nachsprechen“ der Stufe 1 nach durchgängigen Prinzipien zu konstruieren, und dabei die geringste Zahl noch trennscharf messender Aufgaben zu eruieren.

Nach ersten Erprobungen in der Praxis zeigte sich, dass acht Items pro Kind eine noch zu bewältigende Menge im Rahmen des gesamten Testverfahrens darstellten. Diese Items sollten nach ansteigendem Schwierigkeitsgrad angeordnet werden.[14]

---

[10] Bei mangelnder Leistung in diesem Bereich lässt das Verfahren allerdings keinen Rückschluss darauf zu, in welchem Teilbereich eventuell eine Funktionsschwäche vorliegen könnte. Einzig bei guten Leistungen lässt sich feststellen, dass das Arbeitsgedächtnis altersgemäß entwickelt ist.

[11] Ein Zusatz des Zürcher-Lesetests (ZLT), zuletzt teststatistisch an Vier- bis Sechsjährigen Kindern überprüft von Reh/Kiese-Himmel (2007) und Risse/Kiese-Himmel (2008);

[12] Vgl. Seibert et. al. (2001);

[13] Vgl. Linder/Grissemann (1968); Welte (1981); Brunner/Schöler (2001/2002);

[14] Im Vorfeld waren sowohl mehr als auch weniger Aufgaben pro Kind erprobt worden (vgl. Arbeitsbericht „Erprobungen“).

So sollte die Silbenzahl von Wort zu Wort steigen und der Grad der „Wortähnlichkeit“ kontrolliert werden. Die „Wortähnlichkeit“ eines Kunstwortes hängt eng mit dem Bau und der Betonung der natürlichen Sprache zusammen. So erscheinen uns Kunstwörter „wortähnlich“, die den morphologischen Prinzipien der deutschen Sprache folgen. Ein typisches deutsches Wort besteht aus: (Präfix)–Stamm –(Suffix).

Stämme sind Einheiten mit lexikalischem, d.h. bedeutungstragendem Inhalt. Prä- bzw. Suffixe sind Morpheme mit grammatischer Funktion, aber ohne bedeutungstragenden Inhalt. Grammatische Morpheme werden je nach Wortart des zu bildenden Wortes an andere Stämme gebunden. Man unterscheidet dabei Derivationsmorpheme- und Flexionsmorpheme. Viele mehrsilbige Wörter sind nach Wortbildungsmustern gebildet (häufig durch Komposition (lexikalisches Morphem + lexikalisches Morphem) oder Derivation: lexikalisches Morphem + Wortbildungsmorphem), z.B:

lexikalisches M. + lexikalisches M.: Haus-schuh
lexikalisches M. + lexikalisches M. + Flexionsm.: Haus-schuh-e
Derivationsm. + lexikalisches M.+ Derivationsm. + Flexionsm.: un-trink-bar-e

Deutsches Wort (Wortart)	Kunstwort (1)*	Kunstwort (2)**
un-trink-bar-es (Adjektiv)	un-knurk-bar-es	bilknurktu
Vor-les-er-in (Nomen agentis)	Vor-knurk-er-in	kiknurkmolt
zer-brech-e (Verb)	zer-knurk-e	moltknurki

Tabelle 5: Konstruktion von Kunstwörtern nach unterschiedlichen Prinzipien

* Kunstwort in Analogie zu bekannten Wörtern gebildet: Wortstamm entspricht keinem uns bekannten lexikalischen Morphem des Deutschen; grammatische Morpheme lassen jedoch einen Rückschluss auf die Wortart des gebildeten Kunstwortes zu.

** Im Kunstwort sind keine uns bekannten deutschen Morpheme enthalten.

Kunstwörter, die grammatische Morpheme an Pseudostämme binden, erscheinen uns daher „wortähnlicher“, als Kunstwörter ohne grammatische Morpheme.[15] Daher verzichteten wir in den Pseudowörtern ganz auf grammatische Morpheme. Da es im Deutschen jedoch kaum mehr als zweisilbige Wörter ohne grammatische Morpheme gibt, steigt die „Wortunähnlichkeit“ unserer Kunstwörter mit der Silbenlänge.[16]

Durch die besonderen Bedingungen des Gruppentests dürfen die Kunstwörter nicht völlig gleich sein, um einen „Übungseffekt“ auszuschließen (das vierte Kind hätte sonst die Wörter schon drei Mal gehört). Sie sollten aber hinsichtlich der Artikulationsanforderungen möglichst ähnlich sein.[17] So wurden in einem ersten Schritt Laute ausgewählt, die dreieinhalbjährige Kinder nach aktuellem Forschungsstand ohne Schwierigkeiten produzieren können.[18]

[15] In Verfahren, bei denen die Wortähnlichkeit bei der Itemkonstruktion berücksichtigt wird, wird diese häufig durch die Einschätzung von Erwachsenen aufgrund unklarer Prinzipien erhoben (vgl. Grimm; ...). Bei näherer Betrachtung der Kunstwörter zeigt sich meist, dass die „wortähnlichen“ Pseudowörter grammatische, manchmal sogar lexikalische Morpheme beinhalten (z.B. Toschlander; Defsal; Pristobierichkeit).

[16] Im Deutschen sind Wortstämme in der Regel höchstens zweisilbig (z.B. Arbeit, Ausnahme: Dreisilber auf Schwa wie Forelle, Banane).

[17] Aufgrund von Koartikulationsbedingungen können die Anforderungen nie wirklich identisch sein, jedoch sollte ausgeschlossen werden, dass ein Kind ein Wort nachsprechen muss, das für Vierjährige prinzipiell schwierig ist – unabhängig von seiner Gedächtnisleistung.

[18] Vgl. Fox/Dodd 2005;

Der Silbenaufbau musste sehr strikten Konstruktionsregeln folgen, um zu gewährleisten, dass die vier Kinder zwar jeweils andere Wörter nachsprechen sollen, die jedoch möglichst ähnlich im Schwierigkeitsgrad sind:

- Vorgegeben werden zwei-, drei- und viersilbige Wörter[19]. Die Silbenzahl pro Wort nimmt hierbei zu.
- Die Häufigkeit von Konsonant-Vokal-Folgen und die Position der Vokale über In- und Auslautpositionen variieren systematisch.
- Die Wörter enthalten sowohl offene, als auch geschlossene Tonsilben.

Um vier jeweils möglichst gleichwertigen Aufgaben konstruieren zu können, erfolgt ein systematischer Wechsel der Vokale an verschiedenen Positionen innerhalb eines Kunstwortes. von Kind zu Kind bleiben die Vokale gleich – die Konsonanten wechseln, allerdings wechselt in jedem Fall die Artikulationsstelle der Konsonanten.[20]

Für das Deutsche typische Konsonantencluster wurden (mit einer Ausnahme) aus zwei Gründen vermieden: Zum einen steigt mit bestimmten Clustern die Wortähnlichkeit (z.B. Perbst), zum anderen fällt die Artikulation vielen jungen Kindern unserer Zielgruppe noch schwer.[21]

Ausgeschlossen werden sollten darüber hinaus aus den oben beschriebenen Gründen grammatische oder lexikalische Morpheme des Deutschen.

[19] Diese Silbenzahl der Kunstwörter entspricht dem Forschungsstand zur Leistungsfähigkeit von Drei- bis Vierjährigen (im SETK sind die Fünfsilber den Vier- bis Fünfjährigen Kindern vorbehalten). In einem ersten Entwicklungsschritt wurden auch Ein- und Fünfsilber konstruier, jedoch nach den Erprobungen in der Praxis wieder verworfen (s. Arbeitsbericht „Erprobungen“);

[20] Die ersten Erprobungen von Kunstwörtern, die z.B. Zischlaute und Konsonantencluster enthielten, ergaben, dass viele Kinder noch Schwierigkeiten bei der Artikulation dieser Laute hatten (s. Arbeitsbericht „Erprobung“).

[21] Vgl. Fox/Dodd 2005;

	Silben	Silbenaufbau	Auswahl-Konsonanten/Vokale
1.	2	K-V-K-V	[k,p,t,f]–[a:]–[n,l,m,r]–[i]
2.	2	K-V-K-V-K	[k,p,t,f]–[e:]–[b,g,v,d]–[u]–[k,p,t,f]
3.	2	K-V-K-K-V-K	[b,g,v,d]–[u:]–[r]–[k,p,t,f]–[i]–[k,p,t,f]
4.	3	K-V-K-V-K-V	[n,l,m,r]–[i:]–[k,p,t,f]–[o:]–[k,p,t,f]–[a]
5.	3	K-V-K-V-K-V	[b,g,v,d]–[a:]–[k,p,t,f]–[e:]–[n,l,m,r]–[o]
6.	3	K-K-V-K-V-K-V	[k]–[l]–[e:]–[b,g,v,d]–[i:]–[b,g,v,d]–[a]
7.	4	K-V-K-V-K-V-K-V	[k,p,t,f]–[u:]–[k,p,t,f]–[a:]–[n,l,m,r]–[o:]–[n,l,m,r]–[i]
8.	4	K-V-K-V-K-V-K-V	[b,g,v,d]–[o:]–[n,l,m,r]–[i:]–[n,l,m,r]–[e:]–[b,g,v,d]–[u]

Tabelle 6: Konstruktion KN Stufe 1 (Pilotversion 2007)

	grün	blau	rot	gelb
1.	T**a**mi	F**a**ri	P**a**ni	K**a**li
2.	K**e**butt	P**e**duff	T**e**wuck	F**e**gupp
3.	G**u**rfipp	W**u**rpitt	B**u**rtiff	D**u**rbick
4.	Rif**o**pa	Lip**o**ta	Mit**o**ka	Nik**o**fa
5.	Wak**e**mo	Daf**e**lo	Gap**e**ro	Bat**e**no
6.	Kleg**i**ba	Kled**i**ga	Kleb**i**wa	Klew**i**da
7.	Putar**o**li	Fupal**o**mi	Kufan**o**ri	Tukam**o**ni
8.	Bolim**e**du	Gonil**e**wu	Dorin**e**gu	Womir**e**bu

Tabelle 7: Kunstwörter Stufe 1 (Pilotversion 2007)

Evozierungsmethodik:

Da aus der einschlägigen Forschung bekannt ist, dass Dreijährige aufgrund mangelnder linguistischer Bewusstheit nur bedingt dazu zu bewegen sind, etwas nachzusprechen, wurde die Anweisung in der Aufgabe so umgestaltet, dass die Kinder aufgefordert wur-

den, der auf dem Spielplan abgebildeten Giraffe einen Namen zu geben.[22] Dies sollte dazu beitragen, die hohe Quote von Kindern, die laut Berichten in der Literatur bei diesem Aufgabentyp die Mitarbeit verweigern, zu senken.

Zur Standardisierung der Aussprache der Kunstwörter wurde die Betonung der Wörter markiert und die Schreibung den gängigen Regeln der deutschen Orthographie gemäß gestaltet.[23] Auf eine CD zur Präsentation der Aufgaben verzichteten wir, da jüngere Kinder eine höhere Aufmerksamkeitsspanne zeigen, wenn anwesende Personen die Aufgaben präsentieren, als wenn diese vom Band kommen.[24]

Stufe 1: Aufgabenbereich Morpho-Syntax; Untertest SN

Da die Analyse spontansprachlicher Äußerungen als äußerst aufwändiges Verfahren gilt, wird zur Erfassung morpho-syntaktischer Fähigkeiten in der Sprachstandsdiagnostik häufig auf eine weitere Nachsprechaufgabe zurückgegriffen; nämlich auf das Nachsprechen von Sätzen. Damit prüft man, inwieweit ein Kind in der Lage ist, erworbene grammatische Kenntnissysteme für die Wiedergabe von Sätzen zu nutzen. Dieser Aufgabentyp gilt als besonders hoch alterskorreliert und risikoindizierend.[25] Für den Itempool wurden Sätze unterschiedlicher Länge und unterschiedlich komplexer Satzbaupläne entwickelt. Enthalten sein sollten Haupt- und Nebensatzkonstruktionen, Sätze mit Präpositionalphrasen, mehrteiligen Prädikaten und Inversionen (flexibler Besetzung des Vorfeldes)[26]. Dabei sollten einige der Sätze Wortfolgen enthalten, die ohne stützenden Sinnzusammenhang miteinander verbunden sind.[27]

[22] Grimm (2000);
[23] Phonem-Graphem-Korrespondenzen;
[24] Vgl. Radeborg et al. (2006, S.190). Schöler/Schäfer (2004) und Schöler (1999) gehen zwar davon aus, dass schon minimale Veränderungen in der Vorgabe Einfluss auf die Leistungen der Kinder haben, dennoch scheinen die Nachteile beim Einsatz einer CD im Rahmen von BiZ/BiP zu überwiegen.
[25] Vgl. Schöler/Schäfer 2004; Schöler u.a. 1997; Stokes u.a. 2006;
[26] Schöler 1999;
[27] Vgl. Grimm/Weinert 2000;

	Satzbauplan	Anzahl der Wörter
1.	**Subjekt + Prädikat** Lena schläft.	**2**
2.	**Subjekt + Prädikat + Akkusativ-Objekt** Leo erzählt ein Märchen.	**4**
3.	**Subjekt + Prädikat + Ortsergänzung** Murat schläft in seinem neuen Bett.	**6**
4.	**Subjekt + Prädikat + Dativ-Objekt** Das Baby wird von Tom gebadet.	**6**
5.	**Adverbial-Ergänzung + Prädikat + Akkusativ-Objekt** Heute füttert Anna einen großen Hund.	**6**
6.	**Subjekt + Prädikat + Zeit-Ergänzung + Direktional-Ergänzung** Die Kinder rennen schnell zur Schule.	**6**
7.	**Zeit-Ergänzung + Prädikat + Subjekt + Adverbialergänzung +Direktionalergänzung** Heute rennen die Kinder schnell zur Schule	**6**
8.	**Subjekt + Prädikat + Orts-Ergänzung** Die Brille liegt unter dem Sofa.	**6**
9.	**Orts-Ergänzung + Prädikat + Subjekt + Akkusativ-Objekt** Auf dem Spielplatz hat Ina ihren Schal verloren.	**8**
10.	**Subjekt + Negation + Prädikat (zweiteilig) + Akkusativ-Objekt** Tim kann seinen Schlüssel nicht finden.	**6**

Tabelle 8: Konstruktion SN Stufe 1 (Pilotversion 2007)

	Satzbauplan (Fortsetzung)	Anzahl der Wörter
11.	**Subjekt + Prädikat (zweiteilig) + Dativ-Objekt + Akkusativ-Objekt** Tom nimmt dem Hund den Knochen weg.	**7**
12.	**Subjekt + Prädikat + Kausal-Ergänzung** Anna ist traurig, weil es heute regnet.	**9**
13.	**Subjekt + Prädikat + Orts-Ergänzung + Kausal-Ergänzung** Die Kinder spielen im Haus, weil es heute regnet.	**9**
14.	**Subjekt + Prädikat + Akkusativ-Objekt** Das lustige Eis tanzt einen Baum.	**6**
15.	**Adverbial-Ergänzung + Prädikat + Subjekt + Akkusativ-Objekt** Heute kitzelt der fleißige Hund einen Apfel.	**7**
16.	**Konditional-Ergänzung + Prädikat + Direktional-Ergänzung** Wenn die Hose singt, klettert sie über die Straße.	**9**
17.	**Orts-Ergänzung + Prädikat + Subjekt + Akkusativ-Objekt** Hinter der lustigen Wiese finden die Flaschen den Streit.	**9**

Tabelle 8: Konstruktion SN Stufe 1 Fortsetzung (Pilotversion 2007)

Für die Pilotversion Stufe 1 wurden vier Satztypen (2, 3, 9, 11 und 12) mit zwischen 6 und 9 Wörtern ausgewählt und entsprechende Items in vierfacher Form konstruiert.

	grün	blau	rot	gelb
Ü	Burak backt einen Kuchen.	Ayla malt ein Bild.	Lara liest ein Buch.	Malte singt ein Lied.
1.	Der Tisch wird von Tom gedeckt.	Die Blume wird von Tim gepflückt.	Die Suppe wird von Maria gekocht.	Der Hund wird von Igor gestreichelt.
2.	Anna ist wütend, weil sie ihr Turnzeug vergessen hat.	Kemal ist traurig, weil er seinen Ball verloren hat.	Marek ist glücklich, weil er seinen Freund getroffen hat.	Lisa ist fröhlich, weil sie ihre Brille gefunden hat.
3.	Morgens füttert das kluge Bett einen Teppich.	Abends hört der liebe Tisch einen Schrank.	Mittags riecht das dumme Sofa einen Stuhl.	Heute trinkt das schlaue Telefon einen Tisch.
4.	Wenn die Tür weint, kriecht sie in die Stunde.	Wenn der Löffel lacht, hüpft er in den Mittag.	Wenn der Teller denkt, fällt er aus dem Morgen.	Wenn die Gabel schläft, springt sie in den Abend.

Tabelle 9: Aufgaben SN Stufe 1 (Pilot 2007)

Evozierungsmethodik:

Da sich bei dieser Aufgabe gerade bei den unter vierjährigen Kindern Probleme bei der Mitarbeit zeigen können, ist auch hier eine Rahmung geschaffen worden, die die Kinder zum Nachsprechen animieren soll („Wir spielen Papagei...“).[28] Zwischen die semantisch korrekten Sätze und die Sätze mit sinnlos verbundenen Wortfolgen wurde eine Zwischeninstruktion eingefügt.[29]

Die Vor- und Nachteile des Einsatzes einer CD sind dieselben wie schon bei dem Aufgabentyp „Kunstwörter nachsprechen“ beschrieben - wir haben uns gegen die Präsentation mithilfe eines zusätzlichen Mediums entschieden, um den Beziehungsaufbau zur Testleiterin nicht zu untergraben.

Erprobt wurden sowohl die Aufgaben aus dem Item-Pool als auch die für die Pilotversion ausgewählten Items.

[28] Grimm (2001) setzt diesen Aufgabentyp aus ähnlichen Gründen im SETK erst bei vierjährigen Kindern ein;
[29] Beide Möglichkeiten der Item-Vorgabe wurden in der Praxis erprobt;

Stufe 1: Aufgabenbereich Prgmatik/Erzählen; Untertest: Bild beschreiben (BB)

Mit diesem Aufgabentyp werden Diskursfähigkeiten des Kindes erfasst, indem es aufgefordert wird, einen Bildausschnitt zu beschreiben. Ziel dieser Aufgabe ist es, herauszufinden, inwieweit ein Kind schon in der Lage ist, die Beschreibung eines Bildausschnittes zu strukturieren und auszugestalten.

Um die Gesprächssituation während der Testdurchführung aufrecht zu erhalten, muss ein Kind über pragmatische Fähigkeiten verfügen, sich auf den Zuhörer einstellen, Zugzwänge erkennen und bedienen. Darüber hinaus bietet es sich an, bei diesen spontanen Äußerungen des Kindes ausgewählte lexikalische und morpho-syntaktische Leistungen festzuhalten.

Bildbeschreibung	Leistungsbereich
1. Erzählt spontan	Pragmatik
2. Fasst das Wesentliche gleich am Anfang kurz zusammen	Pragmatik
3. Fängt nach Ermutigung/Frage an zu erzählen	Pragmatik
4. Identifiziert mindestens drei Akteure	Pragmatik
5. Beschreibt mindestens ein Ereignis	Pragmatik
6. Verknüpft mehrere Akteure und / oder Ereignisse sinnvoll	Pragmatik
7. Markiert einen Höhepunkt	Pragmatik
8. Stellt logische und / oder zeitliche Verknüpfungen her	Pragmatik
9. Kommentiert das Geschehen	Pragmatik
10. Dramatisiert	Pragmatik
11. Verknüpft mit eigenen Erlebnissen	Pragmatik
12. Beendet von sich aus	Pragmatik
13. Beginnt eine Äußerung, unterbricht und startet neu	Pragmatik
14. Verwendet Laute/Gestik/Mimik zur Überbrückung	Pragmatik
15. Bezieht den Zuhörer mit ein	Pragmatik

Tabelle 10: Aufgaben BE Stufe 1 (Pilot 2007)

Bildbeschreibung (Fortsetzung)		Leistungsbereich
16.	Zählt Dinge auf	Lexik-Semantik
17.	Macht Ortsangaben	Lexik-Semantik
18.	Kennzeichnet näher	Lexik-Semanik
19.	Benennt mindestens drei Handlungsträger	Lexik-Semantik
20.	Verwendet mindestens Dreiwortäußerungen	Morpho-Syntax
21.	Reiht mindestens zwei Sätze aneinander	Morpho-Syntax
22.	Verwendet Haupt-, Nebensatzkonstruktionen	Morpho-Syntax

Tabelle 10: Aufgaben BE Stufe 1 Fortsetzung (Pilot 2007)

Konstruktion des Aufgabenpools zu Stufe 2

Stufe 2: Aufgabenbereich Wortschatz (WS): Untertests: Wortproduktion (WP), Wortverständnis (WV) und Begriffklassifikation (BK)

Mit den Untertests dieses Aufgabenbereichs soll untersucht werden, wie der Wortschatz des Kindes beschaffen ist.

Selbst in einem sehr umfangreichen Verfahren kann nur ein Teil der individuellen Wortschatzleistung eines Kindes erfasst werden. Da in einem Subtest eines mehrteiligen Verfahrens im Gegensatz zu einem reinen Wortschatztest nur eine kleinere Anzahl von Aufgaben gestellt werden kann, ist zu bedenken, inwieweit sichergestellt wurde, dass die Auswahl der Aufgaben die Wortschatzleistung eines Drei- bis Vierjährigen Kindes repräsentieren kann. Die leichtere Abbildbarkeit konkreter Objektbegriffe darf nicht zu einer Nomen-Zentrierung des Instrumentes führen.

Es geht also weniger darum, das individuelle Spektrum der Wortschatzleistungen des Kindes aufzuzeigen, als vielmehr die Bereiche zu überprüfen, deren Entwicklungsstand Rückschlüsse auf mögliche Gefährdungen der weiteren Wortschatzentwicklungen zulässt.

Mit der Pilotierungsform der Stufe 2 „Besuch im Pfiffikus-Haus (BiP)“ sollten am Ende – schon allein aus Gründen der Testökonomie - insgesamt nicht mehr als rund 50 Zielwörter überprüft werden, die möglichst gleichmäßig mit je 26 auf die Rezeptions- und Produktionsaufgaben verteilt werden.

Zunächst ist eine vielfache Menge an Aufgaben entwickelt worden, um nach den Erprobungen in der Praxis die besonders geeigneten auswählen zu können.

Anschließend wurde eine umfangreiche Wortliste mit mehr als 500 Einträgen erstellt. In der Gesamtwortliste wurden die Wortarten wie in der folgenden Tabelle dargestellt gewichtet. Im Unterschied zur gängigen Testpraxis haben wir die Anzahl der Nomina zugunsten der Präpositionen gekürzt, weil lokale Präpositionen aufgrund ihrer Zugehörigkeit zu der geschlossenen Wortklasse der Funktionswörter nur einen kleinen, wenn auch sehr wichtigen Anteil im Lexikon des Kindes ausmachen.[30]

	Substantive	**Verben**	**Adjektive**	**Präpositionen**
Aufgabenanteil (Wortverständnis/ Wortproduktion)	55 %	30 %	10 %	5 %

Tabelle 11: Verteilung der Wortarten in den Untertests WV und WP von BiP

Die konkreten Aufgaben wurden multikriterial bestimmt. So wurde zu jeder Aufgabe angegeben, aus welcher Quelle das Wort stammt (z. B. Häufigkeitswörterbücher; Grundwortschätze; Wortlisten geprüfter Verfahren). Auch wurde jeder Aufgabe ein Schwierigkeitsgrad zugemessen, der unter Abwägung von Kriterien, wie: früh erworben, später erworben, besonders entwicklungssensitiv, bestimmt worden war. Außerdem wurde jede Aufgabe mit einer Angabe zu Vorkommen und Frequenz im kindlichen Wortschatz markiert (Kennzeichnungen: häufig bei Kindern im Vorschulalter, seltener belegt im Vorschulalter, erst im Grundschulalter belegt). Zusätzlich wurde sichergestellt, dass die Wörter sich sowohl auf natürliche Objekte und Artefakte beziehen.[31]

Dabei war es hilfreich, die Informationen, die zur Repräsentation von Wörtern im mentalen Lexikon beitragen, festzuhalten. Dies sind neben Informationen zu der Bedeutung eines Wortes auch seine morpho-syntaktischen, phonetisch-phonologischen und u. U. auch pragmatischen Eigenschaften. Dies soll in der folgenden Tabelle am Beispiel des lexikalischen Eintrags *Fensterbank* deutlich werden.[32]

[30] Vgl. die Diskussion bei Rothweiler 2001, S. 195;
[31] Vgl. Kauschke 2000; Rothweiler 2001, S. 167;
[32] Vgl. dazu ausführlicher Rothweiler 2001, S. 32 f.; Meibauer/ Rothweiler 1999, S. 11; Kiese-Himmel 2005, S. 25 ff.;

	Semantische Eigenschaften	Wortart/ Syntaktische Eigenschaften	Morphologische Eigenschaften	Phonetisch-phonologische Form
Fenster-bank	[+ konkret] [+ man-made] [- belebt] Unterkategorie von BANK	Substantiv Appellativum (zählbar) Konkretum Femininum	*Determinativ-Kompositum (Subst. + Subst.), kein Fugenelement* Plural: [Umlaut + e]	[‘fɛnstɛɐbaŋk] /’fɛnstɛrbaŋk/ 3- Silber: ‘σσσ

Tabelle 12: Informationen zum lexikalischen Eintrag eines Wortes

Damit Kinder trotz unterschiedlicher Erfahrungshintergründe möglichst gleiche Chancen haben, die Aufgaben zu lösen (Fairness), werden mit dem Wortschatz-Subtests der Stufe 2 von Delfin 4 Wörter aus verschiedenen Erfahrungsbereichen aufgegriffen, mit denen schon sehr junge Kinder in Berührung kommen.[33]

Nach einer umfassenden Erprobung des gesamten Aufgabenpools mit Kindern der entsprechenden Altersklasse, erfolgte eine empirisch begründete Auswahl der Aufgaben für die Pilotierungsform der Wortschatz-Untertests.

Evozierungsmethodik:
Die Erfassung des passiven (Wortverständnis) und des aktiven (Wortproduktion) Wortschatzes erfordert unterschiedliche Evozierungsmethoden. Bei der Entscheidung für bestimmte Vorgehensweisen haben wir uns an der einschlägigen empirischen Forschung orientiert.[34]

Untertest Wortverständnis
Ziel dieser Aufgaben ist die Einschätzung der Verständnisleistung für Wörter verschiedener Wortarten. Da im Bereich der Lexik die rezeptiven Leistungen zeitlich vor den produktiven Leistzungen erfolgen, ist es gerade bei schwachen Bearbeitungen der Aufgaben zum produktiven Wortschatz besonders wichtig, auch die Rezeption überprüft zu haben, weil sich daraus Hinweise für die Sprachförderung ableiten lassen.

[33] Vgl. Fenson et. al. 1994; Dromi 1999; zusammenfassend Szagun 2006, S. 114 ff.;
[34] Vgl. z. B. Abedi 2006; Dunn/Dunn 1997; Gleason/Fiske/Chan 2004; Reynell/Gruber 1990; Rohlfing 2002; Wiig et al. 1992; Wilken 2005;

Bewährt haben sich Wort-Bild-Zuordnungsaufgaben, bei denen das Ziel-Wort von Ablenker- oder Störbildern abgegrenzt werden muss. Zu beachten ist, dass bei dieser Methode auch die Auswahl der Ablenkerbilder nach kontrollierten Kriterien erfolgen muss.[35] Um die rezeptiven Möglichkeiten des Kindes zu erfassen, wenden wir das sog. Vier-Karten-Verfahren an.[36] Auf den Bildern befinden sich neben einer Abbildung des Ziel-Wortes zwei semantische Ablenker und ein extremes Störbild (vgl. dazu die Auflistung im Anhang). Das Kind wird aufgefordert, nach auditiver Vorgabe eines Wortes auf ein entsprechendes Bild aus der vorgegebenen Auswahlmenge zu zeigen. Bei diesen Aufgaben ist keine verbale Reaktion des Kindes notwendig.

Untertest Wortproduktion

Bei den Aufgaben, die dem Ziel dienen, die expressive Wortschatzleistung des Kindes einschätzen zu können, wird auf die bewährte Methode des Bild-Benenn-Tests zurück gegriffen (so wie im AWST-R 3-6; Kiese-Himmel 2005). Den Kindern wird eine Karte mit einer Abbildung vorgelegt, welche sie nach einer Instruktions-Frage benennen sollen. Da sich die Kinder bei dieser Aufgabe verbal äußern, ist hier auch eine Berücksichtigung artikulatorischer Fähigkeiten möglich.

Bei der Überprüfung des aktiven Wortschatzes durch einen Test sollte vorher begründet festgelegt werden, welche Äußerungen, die vom Ziel-Wort abweichen, als Antworten akzeptiert werden (z. B. Komposita, Diminutiva, regionale Elemente, usw.) und welche als unzutreffend bewerten werden sollten.

Die Formenbildung sollte im Rahmen der Wortschatz-Subtests jedoch keine Grundlage für die Bewertung der Äußerungen sein, da wir zunächst nur erfahren wollen, ob ein bestimmtes Wort bzw. Lexem im kindlichen Lexikon vorhanden ist. Im Morpho-Syntax-Subtest sieht es dann umgekehrt aus; hier ist weniger die Äußerung eines inhaltlich treffenden Lexems von Bedeutung als die gebildete grammatische Form (z. B. eine Verwendung von Pluralallomorphen wie *Tisch-e*, *Kind-er*).

[35] z.B. semantische, optische oder extreme Störbilder; vgl. dazu Rothweiler 2001; Kauschke 2002;
[36] Ahnlich den Karten-Verfahren des PPVT-III (Dunn/ Dunn 1997), bei Rothweiler 2001 oder Kauschke 2002;

Untertest Begriffsklassifikation

Bei diesem Untertest geht es um die Überprüfung der Fähigkeit zur Bildung von Begriffstaxonomien als Bestandteil des semantischen Wissens. Mit Hilfe von Bildkarten wird festgestellt, ob und wieweit die Kinder in Bezug auf die Oberbegriffe „Kleidung", „Obst" und „Spielzeug" bereits in der Lage sind, untergeordnete Elemente zu identifizieren. Die Bildersets setzen sich so zusammen, dass der Schwierigkeitsgrad der einzelnen Zuordnungen unterschiedlich hoch ist. Eine Banane ist für Kinder eine leichter identifizierbare Obstsorte als die Zitrone. Beim Erkennen von Spielzeugen muss das Kind genuine Spielzeuge von anderen Gegenständen, mit denen man auch spielen kann, unterscheiden.[37] Geprüft wird die Zuordnung von Unterbegriffen zu Oberbegriffen. Zu jedem verbal vorgegeben Oberbegriff wählt das Kind aus einem Kartenstapel die passenden Unterbegriffe aus. Jeder Kartenstapel umfasst sechs Bildkarten, die entweder ein Zielwort oder ein Störbild darstellen. Die Störbilder haben zwei Funktionen: Sie geben Hinweise, wie klar die Vorstellungen eines Kindes bezüglich eines Begriffsfeldes bereits sind; und sie dienen dazu, die Wahrscheinlichkeit zu senken, dass ein Kind zufällig die richtige Bild-Karte auswählt.

[37] Vgl. Hasselhorn 1990; Scheib/Rothmann 1994; Kauschke/Siegmüller 2002;

	Oberbegriff	Unterbegriffe	Ablenker
1.	**Kleidung**	Pullover	Obstmesser
		Schuh	Schere
		Hut	
		Gürtel	
2.	**Spielzeug**	Lego-Stein	Kissen
		Bauklotz	Koffer
		Puppe	
		Puzzle	
3.	**Obst**	Apfel	Blume
		Banane	Spielzeugkiste
		Kirsche	
		Zitrone	

Tabelle 13: Aufgaben BK Stufe 2 (Pilot 2007)

Stufe 2: Aufgabenbereich Phonembewusstheit und –gedächtnis; Untertest: Kunstwörter nachsprechen (KN)

Das Konstruktionsprinzip für die Kunstwörter der Stufe 1 wurde für die Stufe 2 beibehalten. Da hier jedoch keine vierfache Aufgaben-Menge entwickelt werden musste, konnten weitere Konsonanten ([j];[h]) hinzugenommen werden, ohne den systematischen Aufbau zu durchbrechen.

Aus dem Aufgaben-Pool wurden folgende Kunstwörter für die Pilotierungsform ausgewählt:

	Silben	Silbenaufbau	Auswahl-Konsonanten/Vokale
1.	2	K-V-K-V	[ni:bu]
2.	2	K-V-K-V-K	[fe:git]
3.	3	K-V-K-V-K-V	[lu:po:ri]
4.	3	K-V-K-V-K-V-K	[ro:bi:wam]
5.	3	K-V-K-V-K-V	[ja:ke:du]
6.	4	K-V-K-V-K-V-K-V	[ku:ta:bo:di]
7.	4	K-V-K-V-K-V-K-V	[ha:mi:fu:ko]
8.	4	K-V-K-V-K-V-K-V	[ti:wo:re:pi]
9.	4	K-V-K-V-K-V-K-V-K	[go:tu:ma:lik]

Tabelle 14: Aufgaben KN Stufe 2 (Pilot 2007)

Verworfene Untertests: Silben und Reime identifizieren (SI/RI)

Zur phonologischen Bewusstheit im weiteren Sinne zählen die Fähigkeiten, Lauteinheiten wie Silben und Reime zu erkennen. Für die Verfahren von Delfin 4 wurde ein Satz von Aufgaben entwickelt, mit der diese Fähigkeiten erfasst werden können. Die Konstruktion dieser Aufgaben wird im Folgenden nur kurz umrissen, da sich schon nach den ersten Erprobungen in der Praxis gezeigt hat, dass sehr viele Kinder den Aufgabenanweisungen nicht gefolgt sind.[38]

Untertest Silben identifizieren

Die Kinder wurden aufgefordert, ein mehrsilbiges Wort in seine einzelnen Silben durch Klatschen oder Hüpfen zu unterteilen. Ausgewählt wurden Zweisilber (z.B. Auto, Eisbär), Dreisilber (z.B. Elefant, Papagei) und Viersilber (z.B. Regenbogen, Schokolade).

[38] Vgl. Arbeitsbereicht „Erprobungen“;

Untertest Reime identifizieren

Um heraus zu finden, inwieweit die Kinder schon Reime erkennen können, wurden ihnen Sets mit jeweils drei Bildkarten präsentiert. Zwei der Bilder stellten Zielwörter dar, die sich reimten, das dritte Wort war ein sog. Ablenker (z.B. Haus - Maus; Ablenker: Brot). Um die Schwierigkeit zu erhöhen, wurden auch Ablenker ausgewählt, die im Anlaut mit einem der Zielwörter übereinstimmten (z.B. Bein - Stein; Ablenker: Ball). Des Weiteren wurde auch ganz ohne Bildkarten gearbeitet.

Offensichtlich fehlt vor allem den noch nicht ganz vierjährigen Kindern noch die notwendige linguistische Bewusstheit, um die Aufgabenstellung zu verstehen bzw. zu befolgen. Da aus der einschlägigen Literatur bekannt ist, dass diese Aufgaben besser von über Vierjährigen und Fünfjährigen gemeistert werden.

Wir haben diese Untertests deshalb nicht mit in die Verfahren aufgenommen.

<u>Stufe 2: Aufgabenbereich Morpho-Syntax; Untertest „Sätze nachsprechen (SN)“</u>

Das Konstruktionsprinzip des Untertests „Sätze nachsprechen“ blieb für die Aufgaben der Stufe 2 unverändert (s. o.). Ausgewählt wurden für das Einzelscreening sechs Sätze mit verschiedenen Satzbauplänen. Zwei dieser Sätze zählen zu den sog. sinnfreien Sätzen.

	Satzbauplan	Anzahl Wörter
1.	**Subjekt + Prädikat + Ortsergänzung** Murat schläft in seinem neuen Bett.	**6**
2.	**Subjekt + Prädikat + Dativ-Objekt** Die Katze wird von Lena gefüttert.	
3.	**Zeit-Ergänzung + Prädikat + Subjekt + Adverbialergänzung +Direktionalergänzung** Heute rennen die Kinder schnell zur Schule	**6**
4.	**Subjekt + Negation + Prädikat + Akkusativ-Objekt** Ayla kann ihren Schlüssel nicht finden.	**6**
5.	Tom nimmt dem Hund den Knochen weg.	
6.	**Subjekt + Prädikat + Kausal-Ergänzung** Anna ist traurig, weil es heute regnet.	**9**
7.	**Subjekt + Prädikat + Akkusativ-Objekt** Das lustige Eis tanzt einen Baum.	**6**
8.	**Adverbial-Ergänzung + Prädikat + Subjekt + Akkusativ-Objekt** Heute kitzelt der fleißige Hund einen Apfel.	**7**
9.	**Konditional-Ergänzung + Prädikat + Direktional-Ergänzung** Wenn die Hose singt, klettert sie über die Straße.	**9**

Tabelle 15: Konstruktion SN Stufe 2 (Pilot 2007)

Stufe 2: Aufgabenbereich Morpho-Syntax; Untertest „Pluralbildung (PB)“

Die Untersuchung des Niveaus der morphologischen Kompetenz stellt eine wichtige entwicklungsdiagnostische Aufgabe dar.[39] Zur Messung dieser Kompetenz bieten sich verschiedene Untersuchungsbereiche an wie Kasus, Numerus und Tempus. Pluralbildungen lassen sich relativ einfach untersuchen; Die Überprüfung der Pluralbildung an-

[39] Vgl. Grimm 2001, S. S. 19;

hand von Kunstwörtern gilt in der Sprachpsychologie als „Klassiker“ und wurde seit Berko (1958) zahlreich als Untersuchungstechnik eingesetzt.

Mit Kindern ab vier Jahren gilt der Einsatz von Kunstwörtern lohnenswert, da jüngere Kinder häufig Pluralformen auswendig lernen. Jüngere Kinder haben bei Kunstwörtern häufig noch Probleme, die Aufgabe zu bewältigen.[40]

Aus den Antworten der Kinder kann darauf geschlossen werden, welche Stufe der morphologischen Regelbildung (nach Bowermann 1982) das Kind erreicht hat. Da fehlerhafte Pluralbildungen durchaus auf schon erworbenes Regelwissen hinweisen können, ist bei dieser Aufgabe eine differenzierte Bewertung notwendig.

In den Itempool sind Wörter der weiter vorn bereits genannten Wortschatzliste eingegangen, die als häufige Wörter gelten und zu unterschiedlichen Deklinationstypen des Deutschen zählen. Da es im Deutschen fünf verschiedene Typen der Pluralbildung gibt, sollten alle möglichen Formen erprobt werden. Um herauszufinden, wie die jungen Zielkinder von Delfin 4 mit der Aufgabe „Pluralbildung bei Kunstwörtern“ umgehen, ist auch ein Satz entsprechender Pseudowörter entwickelt worden.

Durch das Zusammenspiel von Singular- und Pluraltypen ergeben sich im Deutschen insgesamt zehn verschiedene Deklinationstypen. Auf die Typen I, II und IX entfallen ca. 90 % der Substantive (nach Mugdan 1977, S. 97):

[40] Vgl. Grimm 2001, S. S. 19;

Deklinationstyp	Sing./Plur.-Kombination	Charakteristik (Gen. Sg./ Nom. Pl.)	Beispiel	Häufigkeit im	
				Wortschatz	im Text
I	S1/ P1	-[e]s/-e	des Tages – die Tage	22,6%	29,9%
II	S1/P2	-[e]s/-Ø	des Wagens – die Wagen	13,1%	9,3%
III	S1/P3	-[e]s/-[e]n	des Staates – die Staaten	0,8%	4,9%
IV	S1/P4	-[e]s/-er	des Bildes – die Bilder	2,3%	3,1%
V	S1/P5	-s/-s	des Uhus – die Uhus	2,4%	0,9%
VI	S2/P3	-[e]n/-[e]n	des Menschen – die Menschen	3,7%	1,6%
VII	S3/P1	-Ø/-e	der Kraft – die Kräfte	1,3%	1,3%
VIII	S3/P2	-Ø/-Ø	der Mütter –die Mütter	0,2%	0,2%
IX	S3/P3	-Ø/-[e]n	der Frau – die Frauen	52,0%	48,5%
X	S3/P5	-Ø/-s	der Oma – die Omas	0,2%	0,02%
Sonderfälle		-ns/n	des Herzens – die Herzen	0,2%	0,2%
		Sonstige		1,1%	0,8%
				100%	**100%**

Tabelle 16: Deklinationstypen (nach Gelhaus 1998, S. 236)

Demnach sind die häufigsten Deklinationstypen bei den Maskulina und Neutra Deklinationstyp I und II und bei den Feminina Deklinationstyp IX.

Als Items zur Überprüfung der Pluralbildungsfähigkeit eignen sich Kunstwörter, die dem Kind nicht bekannt sind. Ob ein bestimmter Pluraltyp sicher oder zumindest tendenziell einem Wort zugeordnet werden kann, hängt von dem Wortausgang (bei mehrsilbigen Wörtern) und von dem Genus des Wortes ab.

Wenn das Kind sehr häufig Zuordnungen übergeneralisiert, können Fehlbildungen auf schon erworbene Regelbildungen hinweisen – umgekehrt ist man nicht so sicher, weil richtig gebildete Formen auch auswendig gelernt sein können.

Itempool zur Pluralbildung:

	Singular	Plural	Deklinationstyp/ Pluraltyp	Begründung der Auswahl
1	die Hose die Eule das Auge	die Hosen die Augen	S3/P3 (Typ IX)	**sichere Zuordnung:** häufigster Deklinationstyp; alle Substantive auf Schwa (*-e*) (meist Feminina und Maskulina) bilden Plural auf *-n*
2	die Waffel die Gabel die Nudel	die Waffeln die Gabeln die Nudeln	S3/P3 (Typ IX)	**sichere Zuordnung:** häufigster Deklinationstyp; alle Feminina auf *–el* bilden den Plural auf *-n*
3	der Wagen	die Wagen	S1/P2 (Typ II)	**sichere Zuordnung:** häufiger Deklinationstyp; alle Maskulina auf *–en* bilden Null-Plural; z. T. mit Umlaut
4	die Oma das Sofa der Uhu	die Omas die Sofas die Uhus	S3/P5 (Typ X)	**sichere Zuordnung:** seltener Deklinationstyp; alle Substantive, die in der unbetonten Nebensilbe auf Vollvokal oder Diphthong enden, bilden den Plural auf *–s;* kein Umlaut
5	der Gürtel der Apfel der Vogel	die Gürtel die Äpfel die Vögel	S1/P2 (Typ II)	**sichere Zuordnung:** sehr häufiger Deklinationstyp; (fast) alle Maskulina und Neutra auf *–el* bilden Null-Plural +mit Umlaut, (Ausnahme*: Pantoffel, Stachel, Muskel*)
6	das Brot das Bein das Schaf	die Brote die Beine die Schafe	S1/ P1 (Typ I)	**Tendenz:** häufiger Deklinationstyp; ca. 74% der Einsilbigen Neutra bilden P1-Plural i.d.R. kein Umlaut (Ausnahme *Flöße*)
7	der Tisch der Bart der Stift	die Tische die Bärte die Stifte	S1/P1 (Typ I)	**Tendenz:** häufiger Deklinationstyp; ca. 89% der einsilbigen Maskulina bilden P1-Plural (oft mit Umlaut)
8	das Bett das Ohr das Herz	die Betten die Ohren die Herzen	S1/P3 (Typ III) Sonderfall	**Tendenz:** seltener Deklinationstyp; ca. 4% der einsilbigen Neutra bilden Plural mit *-en*

Tabelle 17: Konstruktion PB Stufe 2 (Pilot 2007)

Fortsetzung der Tabelle:

	Singular	Plural	Deklinationstyp/ Pluraltyp	Begründung der Auswahl
9	die Maus die Hand die Kuh	die Mäuse die Hände die Kühe	S1/P1 (Typ I)	**Tendenz:** häufiger Deklinationstyp; ca. 25% der einsilbigen Feminina bilden P1-Plural + Umlaut
10	das Huhn das Loch das Buch das Haus	die Hühner die Löcher die Bücher die Häuser	S1/P4 (Typ IV)	**Tendenz:** seltener Deklinationstyp; ca. 21% der einsilbigen Neutra bilden Plural auf *–er;* immer mit Umlaut
11	der Pémmich	die Pemmiche	S1/P1 (Typ I)	**sichere Zuordnung**: alle Substantive auf *–ich* bilden e-Plural
12	der Gímpel	die Gimpel	S1/P2 (Typ II)	**sichere Zuordnung:** (fast) alle Maskulina auf *–el* bilden Null-Plural
13	das Kíepchen	die Kiepchen	S1/P2 (Typ II)	**sichere Zuordnung:** alle Maskulina Neutra auf *–chen* bilden Null-Plural
14	die Sóbel	die Sobeln	S1/P2 (Typ II)	**sichere Zuordnung:** alle Feminina auf *–el* bilden Plural auf *–n* ohne Umlaut
15	die Lóme	die Lomen	S3/P3 (Typ IX)	**sichere Zuordnung:** alle Substantive auf *–e* bilden Plural auf *–n* ohne Umlaut
16	der Máte	die Maten	S3/P3 (Typ IX)	**sichere Zuordnung:** alle Substantive auf *–e* bilden Plural auf *–n* ohne Umlaut
17	die Bólli	die Bollis	S3/P5 (Typ X)	**sichere Zuordnung:** alle Substantive, die in der unbetonten Nebensilbe auf Vollvokal oder Diphthong enden, bilden den Plural auf *–s* ohne Umlaut

Tabelle 17: Konstruktion PB Stufe 2 Fortsetzung (Pilot 2007)

Die Praxis-Erprobungen legten nahe, die Wortauswahl nochmals zu überarbeiten, da nicht alle bildlichen Darstellungen die Zielwörter evozierten.[41] Die Erprobungen dienten auch dazu festzuhalten, welche abweichenden Pluralformen von den Kinder gebildet wurden, die einen Hinweis auf schon erworbenes Regelwissen geben können. Folgen-

[41] So bezeichneten viele Kinder das Sofa als Couch, den Uhu als Eule u.ä. (vgl. Arbeitsbericht „Erprobungen“);

de Wörter und Kunstwörter gingen in die Pilotversion für den ersten Durchgang 2007 ein:

	Wort	2-Punkte-Antwort	1-Punkt-Antwort
1	Schaf	Schafe	Schafen, Schäfe
2	Auto	Autos	Autoen
3	Bett	Betten	Better, Bette
4	Loch	Löcher	Löche, Löchern, Löchers, Loche, Lochen, Locher
5	Löffel	Löffel	Löffels, Löffeln, Löffelns
6	Nagel	Nägel	Nagels, Nageln, Nagelen, Nägels, Nägeln
7	Dopf	Döpfe	Dopfe, Dopfen, Dopfs, Dopfer, Döpfer
8	Dagel	Dägel	Dagels, Dageln, Dagelen, Dägels, Dägeln,
9	Mate	Maten	Matens, Mates
10	Pemmich	Pemmiche	Pemmichen, Pemmichs, Pemmicher
11	Sobel	Söbel	Sobels, Sobeln, Sobelen, Söbels, Söbeln

Tabelle 18: Aufgaben PB Stufe 2 (Pilot 2007)

Evozierungsmethodik:

Es werden Bildkarten eingesetzt, auf denen die Abbildung des zu bezeichnenden Gegenstandes zu sehen ist (in der Einzahl und in der Mehrzahl). Die Testleiterin benennt den abgebildeten Gegenstand oder das Objekt und fragt das Kind, wie mehrere davon genannt werden.

Stufe 2: Aufgabenbereich Pragmatik/Erzählen; Untertest „Bilderzählung (BE)“

Die Erzählungen von Dreijährigen sind häufig noch bruchstückhaft. Sie beschränken sich auf wenige Hintergrundinformationen und beschreiben einzelne Aspekte[42]. Demgegenüber konstruiert ein erheblicher Teil der Vierjährigen Erzählungen bereits als komplexes Ganzheiten, die zielgerichtet erzählt werdenen[43]. Allerdings benötigen sie oft noch Unterstützung, um ihre Erzählungen zusammenhängend wiedergeben zu können[44].

Die Altersspanne der Kinder, die mit dem Verfahren getestet werden, ist erheblich. Um dieses Leistungsspektrum erfassen zu können, haben wir a) eine vier Bilder umfassende Bildergeschichte entwickelt und b) dafür gesorgt, dass die Äußerungen zu den Bildvorlagen insofern genau erfasst werden, als die Spontanäußerungen der Kinder wörtlich protokolliert werden. Wie die Erprobungen ergeben haben, ist das leistbar, denn die Kommentierungen der Bildergeschichte durch die Kinder fallen meist kurz aus.

Konstruktion:

Diese Tatsachen wurden bei der Konstruktion und Auswertung der Erzählaufgabe berücksichtigt. Einerseits wurde die Rolle des Testleiters/der Testleiterin standardisiert (es dürfen nur vorgegebene Hilfen angeboten werden); andererseits wurden die Auswertungskriterien anhand des Forschungsstands zu den Strukturmerkmalen[45] der Erzählungen von Vierjährigen bestimmt. Die resultierende Aufgabenstandardisierung sowie die Auswertungskriterien reflektieren den aktuellen Erkenntnisstand.

[42] vgl. Berman/Slobin 1994, S. 45; Marjanovic-Umek et al. 2002;
[43] vgl. Benson 1993, S. 211ff.;
[44] z.B. Hickman 2000, S. 204; Painter 1999; Pellegrini/Galda 1990;
[45] Hausendorff/Quasthoff 1996;

Vorgabe Bild 1:

	Kennzeichnung Bild 1	Leistung
1	**erzählt von sich aus**	Zugzwang erkennen
2	**fasst das Wesentliche gleich am Anfang kurz zusammen** (bezieht sich auf den weiteren Verlauf, das unangenehme Ereignis, die überraschende, erfreuliche Lösung)	Abstract/ Summary verfassen
3	**fängt nach Ermutigung/Frage an zu erzählen**	Zugzwang erkennen
4	**weist auf den Jungen hin**	Handlungsträger identifizieren
5	**kennzeichnet den Jungen näher** (Äußeres: Kleidung, Gesichtsausdruck, Schirm usw.)	Details benennen
6	**verwendet dabei abwechslungsreiche Begriffe**	detailliert beschreiben
7	**beschreibt den Ort** (Haus, Weg, Pflanzen, Standort des Jungen vor dem Haus usw.)	Handlungsort beschreiben
8	**kennzeichnet das Ereignis näher** (Handlung: es regnet; Junge geht weg; Junge hält einen Schirm zum Schutz vor dem Regen über sich usw.)	Handlung erkennen
9	**kommentiert das Geschehen** (Inneres: „Der will spazieren gehen!"; Der ärgert sich, weil es regnet.")	Bewerten/Kommentieren/Deuten
10	**„dramatisiert" das Geschehen** („Tropf, tropf!, und der wird nass." usw.)	Aufmerksamkeit des Zuhörers fesseln
11	**verknüpft das Bild mit eigenen Erlebnissen** (z.B. „Ich auch schon nass gewest." usw.)	Kontext erweitern
12	**verwendet mindestens Dreiwortäußerungen**	Satzbildung
13	**verwendet mindestens eine Haupt-, Nebensatzkonstruktion**	Satzverknüpfung
14	**reiht mindestens zwei Sätze aneinander**	Satzreihung
15	**bezieht den Zuhörer mit ein** (z.B. „Wo geht der hin?"; „Schau mal!" usw.)	Aufmerksamkeit des Zuhörers aufrecht erhalten

Tabelle 19a: Aufgaben BE Stufe 2 – Bild 1 (Pilot 2007)

Vorgabe Bild 2:

	Kennzeichnung Bild 2	Leistung
1	**erzählt von sich aus**	Zugzwang erkennen
2	**fängt nach Ermutigung/Frage an zu erzählen**	Zugzwang erkennen
3	**bezieht sich auf den Jungen**	Handlungsträger identifizieren
4	**kennzeichnet den Jungen näher** (Äußeres usw.)	Details benennen
5	**verwendet dabei abwechslungsreiche Begriffe**	detailliert beschreiben
6	**beschreibt den Ort** (auf dem Weg, Bäume/Büsche im Hintergrund, Regenpfützen usw.)	Handlungsort beschreiben
7	**kennzeichnet das Ereignis** (Handlung: vom Wind gepeitschte Regentropfen, Büsche; Junge stemmt den Schirm zum Schutz vor dem Regen vor sich her usw.)	Handlung erkennen
8	**stellt Bezug zum 1. Bild her** („Der sollte lieber zu Hause bleiben.“ usw.)	Kohäsion herstellen
9	**kommentiert das Geschehen** (Inneres: „Der ärgert sich!“ usw.)	Bewerten/Kommentieren/Deuten
10	**„dramatisiert“ das Geschehen** („So ein Pech!“ usw.)	Aufmerksamkeit des Zuhörers fesseln
11	**verknüpft das Bild mit eigenen Erlebnissen**	Kontext erweitern
12	**verwendet mindestens Dreiwortäußerungen**	Satzbildung
13	**verwendet mindestens eine Haupt-, Nebensatzkonstruktion**	Satzverknüpfung
14	**reiht mindestens zwei Sätze aneinander**	Satzreihung
15	**bezieht den Zuhörer mit ein**	Aufmerksamkeit des Zuhörers aufrecht erhalten

Tabelle 19b: Aufgaben BE Stufe 2 – Bild 2 (Pilot 2007)

Vorgabe Bild 3:

	Kennzeichnung Bild 3	Leistung
1	**erzählt von sich aus**	Zugzwang erkennen
2	**fängt nach Ermutigung/Frage an zu erzählen**	Zugzwang erkennen
3	**bezieht sich auf den Jungen**	Handlungsträger identifizieren
4	**kennzeichnet den Jungen näher** (Äußeres)	Details benennen
5	**verwendet dabei abwechslungsreiche Begriffe**	detailliert beschreiben
6	**kennzeichnet das Ereignis** (Handlung: Wind ist stark, Schirm fliegt weg, Junge versucht den Schirm zu packen, läuft dem Schirm hinterher usw.)	Handlung erkennen
7	**stellt Bezug zum 1.und/oder 2. Bild her**	Kohäsion/ Kohärenz herstellen
8	**kommentiert das Geschehen** (Inneres: „Was die Mama sagt." usw.)	Bewerten/Kommentieren/Deuten
9	**„dramatisiert" das Geschehen** („Der schreit: Blöder Schirm!")	Aufmerksamkeit des Zuhörers fesseln
10	**verknüpft das Bild mit eigenen Erlebnissen**	Kontext erweitern
11	**verwendet mindestens Dreiwortäußerungen**	Satzbildung
12	**verwendet mindestens eine Haupt-, Nebensatzkonstruktion**	Satzverknüpfung
13	**reiht mindestens zwei Sätze aneinander**	Satzreihung
14	**bezieht den Zuhörer mit ein**	Aufmerksamkeit des Zuhörers aufrecht erhalten

Tabelle 19c: Aufgaben BE Stufe 2 – Bild 3 (Pilot 2007)

Vorgabe Bild 4:

	Kennzeichnung Bild 4	Leistung
1	**erzählt von sich aus**	Zugzwang erkennen
2	**fängt nach Ermutigung/Frage an zu erzählen**	Zugzwang erkennen
3	**bezieht sich auf den Jungen**	Handlungsträger identifizieren
4	**bezieht sich auf die Ente**	Handlungsträger identifizieren
5	**kennzeichnet den Jungen und oder die Ente näher** (Äußeres)	Details benennen
6	**verwendet dabei abwechslungsreiche Begriffe**	detailliert beschreiben
7	**kennzeichnet das Ereignis** (Handlung: Schirm hat sich mit Regen gefüllt; Ente schwimmt auf diesem Wasser usw.)	Handlung erkennen
8	**markiert das Ereignis als überraschenden Höhe-punkt** („Das hat der nicht gedacht.“ usw.)	Kohärenz herstellen
9	**stellt Bezug zum 1.und/oder 2. und/oder 3. Bild her**	Kohärenz/Kohäsion herstellen
10	**kommentiert das Geschehen** („Der freut sich jetzt!“ usw.)	Bewerten/Kommentieren/Deuten
11	**„dramatisiert“ das Geschehen** („Da sagt er: Wo kommst du her?“ usw.)	Aufmerksamkeit des Zuhörers fesseln
12	**verknüpft das Bild mit eigenen Erlebnissen**	Kontext erweitern
13	**verwendet mindestens Dreiwortäußerungen**	Satzbildung
14	**verwendet mindestens eine Haupt-, Nebensatzkon-struktion**	Satzverknüpfung
15	**reiht mindestens zwei Sätze aneinander**	Satzreihung
16	**bezieht den Zuhörer mit ein**	Aufmerksamkeit des Zuhörers aufrecht erhalten
17	**rundet die Geschichte ab** („Das war gut!“; „Jetzt ist Schluss.“ usw.)	Zugzwang erkennen

Tabelle 19c: Aufgaben BE Stufe 2 – Bild 4 (Pilot 2007)

4. Aufgabenerprobung

Ab Herbst 2006 fanden dann Praxiserprobungen von konstruierten Aufgaben, Testanweisungen und Materialien statt. Ziel war dabei zu ermitteln, welche Untertests und Aufgabenformen letztlich in die Endform von Delfin 4 (Stufe 1 und 2) übernommen werden sollen.

(In diesem Zusammenhang wurden auch Aufgaben zu den später nicht in die Endform aufgenommenen Untertests „Reime identifizieren“ und „Silben segmentieren“ angewandt. Deren Erprobung ergab jedoch, dass die Aufgabenstellungen einen größeren Teil der Kinder überforderte. Deshalb haben wir entschieden, diese Untertests nicht in die Endform aufzunehmen. Im Folgenden wird somit nicht mehr darauf Bezug genommen).

Die Vorstudien zum Instrument Delfin 4 bestehend aus „Besuch im Zoo (BiZ)“ und „Besuch im Pfiffikus-Haus (BiP)“ wurden von Herbst 2006 bis Frühjahr 2007 durchgeführt. Die Erhebungen fanden in städtischen, sowie ländlichen Regionen von Nordrhein-Westfalen statt. Es wurde darauf geachtet, die Erprobung in Gebieten mit unterschiedlich hohem Aufkommen an Kindern mit Migrationshintergrund durchzuführen. An den Praxiserprobungen nahmen insgesamt 195 Kinder teil. Dabei wurden die Materialien, Aufgaben bzw. Anweisungen meist in Kindertageseinrichtungen, aber auch in anderen Räumlichkeiten u.a. an der Universität Dortmund (jetzt Technische Universität Dortmund) eingesetzt oder im Rahmen von frei organisierten Spielgruppen angewandt. Die Erprobungen wurden in der Regel von zwei, mitunter auch von einer Wissenschaftlichen Mitarbeiterin(en) in einem ruhigen Nebenraum in Anwesenheit einer bzw. mehrerer pädagogischer Fachkräfte durchgeführt.

Ziel dieser Erprobungen war die Zusammenstellung der Instrumente „Besuch im Zoo (BiZ)“ und „Besuch im Pfiffikus-Haus (BiP)“ für die darauf folgenden Pilotierungsstudien[46]. Die konstruierten Items wurden auf

[46] Siehe Kurzbericht Pilotierungsstudie;

- Kindgemäßheit und
- Praxistauglichkeit überprüft.

Hierfür wurden die konstruierten Items und Anweisungen mit Kindern erprobt bzw. mit deren pädagogischen Fachkräften besprochen. Anhand der dabei gewonnenen Erfahrungen, wurden einzelne Aufgaben und Subtests überarbeitet und optimiert. Ziel der Praxiserprobung war die Erstellung einer Pilotierungsversion von BiZ und BiP, welche als zu einem späteren Zeitpunkt mit einer großen Stichprobe im Rahmen der Pilotierung evaluiert werden sollten. Im Verlauf der Praxiserprobungen wurden drei Aufgabentypen, nebst Anweisungen sowie die dazu gehörenden Bildmaterialien in diversen Variationen angewandt. Letztere wurden im Hinblick auf Aufforderungscharakter und Eindeutigkeit geprüft.

Ergebnisse zu den Untertests von Stufe 1 „Besuch im Zoo (BiZ)“

Beim Untertest „Handlungsanweisungen ausführen (HA)“ geht es um das Verstehen semantischer Informationen auf dem Wortlevel. Als Bildmaterial dienten hier vier Illustrationen von Tiergehegen. Jedes Gehege war für die Vorstudien auf einer DIN A4-Seite abgebildet. Erprobt wurden insgesamt 32 HA. Davon ausgewählt wurden 20.

Evaluiert wurde der Aufgabentyp zunächst mit zehn Kindern von denen drei einen Migrationshintergrund hatten. Keines dieser Kinder wies Sprachauffälligkeiten auf. Keines der Kinder verweigerte die Mitarbeit oder brach diesen Subtest vorzeitig ab.

Die Durchführung zeigte, dass diese Aufgabe für die Kinder einen großen Aufforderungscharakter hat, da sie bildgestützt durchgeführt wird und die Kinder nonverbal agieren konnten. Allerdings zeigten sich während der Erprobungen mehrfach Probleme. So waren 10% der Handlungsanweisungen für die Protokollantin nicht eindeutig zu bewerten. So erwiesen sich z. B. Formulierungen wie „Stelle deine Figur weit weg vom Wasser“ als zu vage. Für die Protokollantin war es deshalb schwer zu entscheiden, ob bzw. wann die Aufgabe gelöst war, weil es sich bei der Ortsangabe „weit weg“ um eine subjektive Empfindung handelt. Deshalb wurden die Handlungs-anweisungen nochmals gründlich überarbeitet. Im Einzelnen bedeutete dies, dass einige Formulierungen mit Zusätzen ergänzt, andere ganz gestrichen bzw. ersetzt werden mussten, um Anwei-

sungen ganz eindeutig zu gestalten. Konkret wurde verzichtet a) auf die Unterscheidung verschiedener Fortbewegungsarten mit den Püppchen (z.B. gehen und springen), b) auf die Verwendung von Adjektiven (z.B. das kleinste Kind, das größte Tier) und c) auf ungenaue Ortsangaben (z.B. weit weg). Da die Erprobungen außerdem ergeben hatten, dass die Kinder Anweisungen mit mehreren nacheinander zu vollziehenden Schritten nur teilweise durchführten, wurde die Bewertung erweitert, so dass nun mehrere Teilpunkte für eine Aufgabe erreicht werden können. Dieses Detail wurde dann im Protokollheft durch Unterstreichung der Wörter kenntlich gemacht.[47]

Die so veränderten Aufgaben wurden dann nochmals mit elf Kindern erprobt. Dabei erwiesen sich 20 als geeignet, so dass sie in die Pilotierungsversion der ersten Stufe des Verfahrens übernommen werden konnten.

Der Untertest „Kunstwörter nachsprechen (KN)“ erfasst, wie gut Kinder unbekannte lexikalische Einheiten im phonologischen Arbeitsgedächtnis behalten und im Anschluss reproduzieren können. In der Praxis kamen nach verschiedenen Kriterien erstellte Wörter zum Einsatz. Die Praxiserprobung fand mit insgesamt 30 Kindern statt. Dabei handelte es sich um 15 Mädchen und 15 Jungen. Acht Kinder dieser Stichprobe hatten einen Migrationshintergrund. Verweigerungen gab es keine. Zum Einsatz kamen Wörter aus unterschiedlich lange Lautfolgen, welche neben schwierigen Lauten, auch Konsonatencluster beinhalteten. Es handelte sich dabei sowohl um wortähnliche, als auch um wortunähnliche Fantasiewörtern. Insgesamt wurden den Kindern 130 Fantasiewörter vorgelegt. Die Erprobungen ergaben, dass das Nachsprechen von unbekannten Lautfolgen von Kindern gut aufgenommen wird und sehr zeitökonomisch ist.

Zunächst kamen bei zehn Kindern nur 70 Fantasiewörter zum Einsatz. Dabei zeigte sich, dass die häufigsten Fehler bei der Reproduktion in Form von Auslassung einzelner, schwieriger Laute oder Reduktion von Mehrsilbern gemacht wurden. Da aber der Untertest KN nicht die Aussprache, sondern das Arbeitsgedächtnis überprüfen soll, wurden - unter Berücksichtigung von Lauterwerbstabellen - weitere 60 Fantasiewörter konstruiert bzw. die vorhandenen Wörter umgearbeitet und erneut mit 20 Kindern erprobt.

[47] Die Unterstreichungen wurden ab Durchgang 2008 durch gezielte Vorgaben in der Handreichung abgelöst.

Die Ergebnisse sprachen dafür, diesen Untertest sowohl in Stufe 1, als auch in Stufe 2 aufzunehmen. Dafür wurden insgesamt 36 geeignete Fantasiewörter bestimmt.

Beim Untertest „Sätze nachsprechen (SN)“ soll ermittelt werden, wie gut es Kindern gelingt, erworbene grammatische Kenntnissysteme für die Wiedergabe von Sätzen zu nutzen. Für diesen Untertest wurden ursprünglich 40 unterschiedliche Satzkonstruktionen entwickelt. Diese beinhalteten sowohl sinnvolle Satzkonstruktionen, als auch „sinnfreie Sätze“ (Unsinns-Sätze).

Dieser Aufgabentyp wurde mit insgesamt 68 Kindern erprobt. Bei der Auswahl der Kinder wurden bewusst 20 Kinder ausgewählt, die zum Zeitpunkt der Erprobung noch keine vier Jahre alt waren, da in der Fachliteratur vereinzelt die Ansicht vertreten wird, dass erst ältere Kinder in der Lage sind, nicht nur sinnvolle Sätze, sondern auch sogenannte Unsinns-Sätze nachzusprechen. Von den 68 Kindern haben nur fünf die Unsinnssätze nicht bzw. nur teilweise nachgesprochen. Zwei davon hatten sehr schlechte Deutsch-Kenntnisse. Insgesamt lässt das den Schluss zu, dass vierjährige Kinder durchaus schon in der Lage und bereit sind, Unsinns-Sätze nachzusprechen. Das entspricht im übrigen auch den Erfahrungen von Grimm (2003), die dies ebenfalls belegen konnte. Vor diesem Hintergrund und angesichts der Tatsache, dass die Aufgabe, Sätze mit bis zu sieben Wörtern nachzusprechen, selbst für Kinder, die eine unterdurchschnittliche Sprachkompetenz aufwiesen, leicht war, wurde dieser Aufgabentyp beibehalten.

Deutlich war auch geworden, dass die Sätze, die den Kindern zum Nachsprechen angeboten werden, nicht zu kurz sein dürfen, wenn es gelingen soll, die Leistungsbreite in diesem Bereich valide zu erfassen. Das deckt sich mit Befunden einschlägiger Forschungen, nach denen Kinder kurze Sätze unter Umständen einfach memorieren, so dass die Leistung nicht ohne weiteres nur auf zugrunde liegendes syntaktisches Wissen verweist. Deshalb wurden die 20 selektierten Aufgaben so zusammengestellt, die es damit möglich ist, ein breites Leistungsspektrum abzubilden.

Nicht zuletzt ergab die Praxiserprobung, dass dieser Untertest sehr zeitökonomisch durchgeführt werden kann.

Deshalb wurde entschieden, diesen Untertest nicht nur in die Pilotierungsversion der ersten, sondern auch in die der zweiten Stufe aufzunehmen. Insgesamt wurden von den 40 Erprobungs-Sätzen 20 ausgewählt.

Bei den Erprobungen des Untertests „Bildbeschreibung (BB)/Bilderzählung (BE)“ kamen für BiZ und BiP unterschiedliche Bildmaterialien zum Einsatz. Bei BiZ dienten die Gehege auf dem Zooplan als Grundlage für die kindlichen Erzählungen; bei BiP wurden zwei Bildergeschichten bestehend aus jeweils vier Bildern als Erzählgrundlage genutzt. Alle Illustrationen waren zu diesem Zeitpunkt noch unkoloriert. Die Tiergehege von BiZ befanden sich jeweils auf einer einzelnen DIN A4-Seite. An den Vorstudien zu beiden Instrumenten nahmen insgesamt 106 Kinder teil.

Die Erprobung des Bildmaterials zur Stufe 1 erfolgte im Hinblick auf den Anregungsgrad sowie die Eindeutigkeit des Bildmaterials. Mit dieser Zielstellung haben zwei Wissenschaftliche Mitarbeiterinnen insgesamt 56 Kindern sechs verschiedene Zeichnungen vorgelegt. 32% dieser Kinder hatten einen Migrationshintergrund.
Die Kinder wurden einzeln untersucht. Es war immer eine den Kindern vertraute Erzieherin im Raum. Die Bilder wurden gemäß folgender Anweisung eingeführt.

- Wollen wir zusammen ein paar Bilder betrachten? Das ist schön.

Erstes Bild:

- Das ist das erste Bild. Schau es Dir erst einmal an.... Auf dem Bild passiert viel. Schau genau hin. Siehst Du es?... Erzähl doch mal...

(Wenn das Kind **nicht** von sich aus beginnt, auf das **herausragende Ereignis** zu zeigen (Tiger-Bild: streitende Hunde; Giraffen-Bild: Giraffe nimmt dem Wärter den Hut weg; Elefanten-Bild: Elefant bespritzt Zoobesucherin; Seehund-Bild: Wärter füttert Seehunde und Delfine).

- Schau mal, hier?

Wenn das Kind immer noch nicht reagiert:

- (Tiger-Bild:) Was machen denn die da (auf Hunde zeigen)? (Giraffen-Bild:) Was macht denn die da (auf Giraffe mit der Kappe im Maul zeigen)? usw.

Falls das Kind schon nach der ersten Äußerung stockt; es aber nicht auszuschließen ist, dass es noch etwas zu sagen hat:

- Und dann?

Wenn es dann immer noch schweigt:

- Was ist da noch?, oder: Was passiert da noch?

Weitere Bilder:

- Jetzt schauen wir uns das nächste Bild an. (Weiter, wie oben beschrieben).

Wenn ein Kind beim ersten und zweiten Bild keine Reaktion zeigt, wird abgebrochen.

- Danke, dass Du zu mir gekommen bist. Jetzt kannst Du wieder zurück gehen. Oder möchtest Du mir noch etwas erzählen?

Dokumentation der Reaktionen des Kindes:

- Für jedes Kind wird auf einem eigenen Blatt festgehalten, wie es auf die sechs Bilder reagiert hat.
- Eingangs oder abschließend wird vermerkt, ob es sich bei dem Kind um Junge oder Mädchen handelt.
- Es wird außerdem vermerkt, ob das Kind – laut Erzieherin - eine gering, durchschnittlich oder gut entwickelte Sprachfähigkeit aufweist.
- Sprachliche Äußerungen des Kindes zu den einzelnen Bildern werden wortwörtlich aufgeschrieben.

Die Bildbeschreibungsaufgaben von BiZ umfassten ursprünglich acht Tiergehege.

Bei diesem Untertest gab es keine Verweigerung. In drei Fällen wurde allerdings die Bildbeschreibung vorzeitig abgebrochen.

Die Kinder wurden dazu aufgefordert nacheinander zu allen Gehegen etwas zu erzählen. (In der späteren Pilotierungsversion bezieht sich die Erzählaufgabe nur noch auf das Gehege auf dem Zooplan, vor dem das Kind unmittelbar sitzt).

Die Erprobung der Zeichnungen ergab, dass die Entwürfe zu vier Gehegen einem Teil der Kinder nur bedingt zugänglich war, weil die dargestellten Tiere entweder nicht erkannt oder nicht richtig benannt wurden. Auf diese Zeichnungen wurde deshalb verzichtet. Die restlichen Zeichnungen wurden ebenfalls überarbeitet. Vor allem wurde das „herausragende Ereignis“ pointierter dargestellt.

Allgemein verdeutlichte die Erprobung dieses Untertests von BiZ, dass die Bildmaterialien einen großen Aufforderungscharakter für die Kinder haben und einen guten Zugang zum Erzählen ermöglichen.

Die Analyse der wörtlichen Transkripte der kindlichen Äußerungen zu den Bildern machte deutlich, dass Kinder dieser Altersgruppe nur kurz erzählen bzw. eher deskriptiv beschreiben. Das erklärt, dass es den zehn pädagogischen Fachkräften, die das Auswertungsraster an diesem Material anwandten, ohne Probleme gelang, die Aussagen der Kinder richtig zu charakterisieren.

Für BiP wurden zwei Bildergeschichten, bestehend aus je vier unkolorierten Bildern, erprobt. Dabei waren die vier Bilder gemeinsam auf einer DIN A4-Seite abgebildet. Beteiligt waren insgesamt 50 Kindern. Diese Gruppe setzte sich aus 19 Jungen und 31 Mädchen zusammen, zehn der Kinder wiesen einen Migrationshintergrund auf. Bei insgesamt sechs Kindern berichteten die Erzieher/innen von Sprachauffälligkeiten.

Eingesetzt wurden zwei Durchführungsvarianten: a) das Nacherzählen einer bildgestützten Geschichte; b) das eigenständige Erzählen einer Bildgeschichte. Um zu ermitteln, welche der beiden Geschichten sich für die Kinder interessanter bzw. ansprechender darstellte, wurden die Kinder zu Anfang gebeten, sich die Geschichte selbst auszuwählen. 44% der Kinder favorisierten die Geschichte „Das Picknick“ und 56% entschieden sich für „Die kleine Ente“.

Bei der ersten Variante „Das Picknick“ wurde den Kindern ein kurzer Text vorgelesen wurde. Im Anschluss daran wurden sie aufgefordert, die eben gehörte Geschichte mit Hilfe von vier unkolorierten Bilder nachzuerzählen. Da schon gleich am Anfang acht Kinder ihre Mitarbeit verweigerten, haben wir diese Variante sehr bald verworfen. Im Folgenden wurde somit nur noch die zweite Variante eingesetzt.

Bei der Variante „Die kleine Ente“ besteht die Herausforderung für die Kinder darin, eine Erzählung nach Vorlage einer Bildergeschichte eigenständig zu strukturieren und auszugestalten. Zunächst durften die Kinder sich die Bildergeschichte in Ruhe anschauen. Mit Hilfe von vorher festgelegten stützenden Fragen wurden sie dann ermuntert, die Bilder zu kommentieren. Dabei wurde deutlich, dass es einzelne Kinder überfordert, wenn alle Bilder gleichzeitig sichtbar auf dem Tisch liegen, weil sie dadurch animiert werden, während der Erzählung zwischen den einzelnen Bildern „hin und her zu springen“, was ihre Möglichkeiten blockiert oder einschränkt, eine „Geschichtenstruktur“ zu entwickeln. Deshalb haben wir uns entschlossen, jedem Kind zunächst die Möglichkeit zu geben, alle Bilder in Ruhe zu betrachten und ihm dann die Bilder einzeln nacheinander anzubieten.

Wie schon bei BiZ wurde auch bei BiP die Handhabbarkeit des Auswertungsrasters erprobt. Dazu wurden 15 pädagogische Fachkräfte gebeten, verschriftete Bildergeschichten von zehn Kindern anhand des Rasters zu charakterisieren. Es zeigte sich, dass die Anwendung der Analysekriterien ohne Probleme möglich war. Das Analyseraster konnte somit beibehalten werden.

Ergebnisse zu den Untertests von Stufe 2 „Besuch im Pfiffikus-Haus (BiP)“

Die Erprobung der Untertests „Kunstwörter nachsprechen (KN)“, „Sätze nachsprechen (SN)“ sowie „Bildbeschreibung (BB) bzw. „Bilderzählung (BE)“ und die daraus resultierenden Aufgabenselektionen wurden bereits weiter oben dargelegt.

Die Anwendung des Untertests „Wortverständnis (WV)“ fand mit insgesamt 65 Kindern (37 Mädchen und 28 Jungen) durchgeführt. Zwölf Kinder dieser Stichprobe zeigten Sprachauffälligkeiten und 11 Kinder hatten eine nicht-deutsche Muttersprache. Dieser Aufgabentyp zur Überprüfung des passiven Wortschatzes fiel den Kindern offenbar leicht, denn es gab weder einen Abbruch während der Durchführung, noch verweigerte eines der Kinder die Teilnahme.

Insgesamt wurden 50 Aufgaben für diesen Untertest erprobt. Diese zielten auf vier verschiedene Wortarten. Konkret erprobt wurden 20 Nomen, 15 Verben, zehn Adjektive und fünf Präpositionen. Bei der Durchführung dieser Aufgaben kamen diverse Bildmaterialien zum Einsatz. Bei den Nomen, Adjektiven und den Verben beinhaltete eine Aufgabe jeweils ein Zielwort aus einer der genannten Wortkategorien, sowie aus drei sogenannten „Ablenkern“. Auf einer DIN A4-Seite waren hierfür vier unkolorierte Illustrationen abgebildet. Bei den Präpositionen setzten sich die Aufgaben jeweils nur aus drei Bildern (einem Zielwort und zwei Ablenkern) zusammen. In diesen Fällen waren jeweils drei unkolorierte Illustrationen auf einer DIN A4-Seite.

Die Erprobungen ergaben, dass es den Kindern leicht fällt, diese Aufgaben zu lösen. So gab es keine „Abbrecher“ und alle Kinder hatten kein Problem, alle 50 Aufgaben nacheinander abzuarbeiten. Für den Untertest wurden aber – aus Gründen der Zeitökonomie –23 Aufgaben mit folgender Verteilung ausgewählt:

- 8 Nomen
- 7 Verben
- 5 Adjektive und
- 3 Präpositionen.

Mit den Aufgaben des Untertests „Wortproduktion (WP)“ soll der aktive Wortschatz der Kinder evoziert werden. Die Stichprobe bestand aus 62 Kindern (38 Mädchen und 24 Jungen). Sprachauffälligkeiten waren bei 12 Kindern bekannt, zehn Teilnehmer waren Migrantenkinder. Dieser Aufgabentyp wurde nur von einem Kind verweigert, bei keiner der Durchführungen wurde der Untertest vorzeitig abgebrochen.

Die Aufgaben zum aktiven Wortschatz sind identisch aufgebaut, wie die Aufgaben zum passiven Wortschatz. Sie zielen also ebenfalls auf vier verschiedenen Wortarten. Auch die Gewichtung der Aufgaben (Anzahl) ist parallel.

Die Bildmaterialien bestanden aus noch unkolorierten DIN A4 Seiten, auf denen sich mehrere zu benennende Abbildungen befanden. Während der Erprobung zeigte sich, dass ein Teil der Kinder durch die Vielfalt der Bilder abgelenkt wurde, was sich dahingehend äußerte, dass sie zwischen den einzelnen Bildern hin und her sprangen. Des-

halb haben wir beschlossen, für die Pilotierungsversion jeweils nur eine Abbildung auf eine Seite DIN A5-Format zu setzen.

Da es sich bei diesem Untertest um Benennaufgaben handelt, bei denen die Kinder die Zielwörter selbstständig produzieren sollen, waren klare, eindeutige Zeichnungen von besonderer Bedeutung. Bei der Erprobung wurde deutlich, dass ein Teil der Zeichnungen noch nicht eindeutig genug war, um bei allen Kindern das Zielwort zu evozieren. Beispielsweise erwiesen sich bei den Nomen-Zeichnungen folgende Zeichnungen als zu wenig eindeutig: Lichtschalter, Kurve, Unfall, Zahl und Schmuck. Das hat damit zu tun, dass diese Abbildungen ohne Kontext nur sehr bedingt zu identifizieren sind. Deshalb wurde auf diese Begriffe verzichtet.

Ansonsten zeigte sich, dass die Kinder diese Aufgaben gerne und schnell bewältigten. Deshalb wurde dieser Subtest mit insgesamt 23 Items (Zusammensetzung siehe Subtest WV) in die Pilotierungsversion übernommen.

Mit dem Untertest „Pluralbildung (PB)“ soll überprüft werden, wie weit die Entwicklung der Differenzierung zwischen Ein- und Mehrzahl bei einem Kind schon fortgeschritten ist. Neben bekannten Wörtern kommen dabei auch Unsinnswörter zum Einsatz. Bei der Bearbeitung der Aufgaben sollen die Kinder zeigen, ob sie schon die morphologische Regelbildung beherrschen.

Die Pluralbildung wurde mit 57 Kindern erprobt. Die Stichprobe setzte sich aus 25 Jungen und 32 Mädchen zusammen, von denen neun Kinder Sprachauffälligkeiten und sieben einen Migrationshintergrund aufwiesen. Es gab keine Verweigerer und auch keine Abbrüche.

Erprobt wurden 27 Items (17 sinnvolle Wörter und 10 „Unsinnswörter“). Diese wurden bildgestützt durchgeführt. Die Zeichnungen waren jeweils auf einer DIN A4-Seite angeordnet. Bei einem Teil der Aufgaben war auf der Vorlage nur eine einzelne Abbildung. Hier sollten die Kinder von der Ein- auf die Mehrzahl schließen. Bei den anderen Aufgaben befanden sich mehrere Elemente auf der Vorlage, mit deren Hilfe ebenfalls die Mehrzahl evoziert werden sollte. Für die „Unsinnswörter“ wurden Illustrationen von Phantasietieren angefertigt, die nicht zu sehr an lebende Tiere erinnern.

Den Kindern wurde zur Einführung die Einzahl des Begriffes vorgegeben. Da sich während der Erprobung zeigte, dass manche Kinder dazu neigen nicht nur die Mehrzahl der Wörter zu bilden, sondern auch die Abbildungen zu zählen, wurde die Anzahl der abgebildeten Elemente auf drei bis vier pro Seite reduziert.

Nach den Erprobungen wurde dieser Untertest mit insgesamt 11 Items (zuzüglich zwei Übungsitems) in die Pilotierungsversion übernommen. Er setzt sich aus sechs sinnvollen Wörtern und fünf „Unsinnswörtern“ mit unterschiedlichem Genus zusammen.

Beim Untertest „Begriffsklassifikation (BK)“ hatten die Kinder die Aufgabe, zu einem vorgegebenen Oberbegriff passende Unterbegriffe zu nennen. Dabei bestand die Stichprobe aus 56 Kindern. In der Praxis kamen neun Oberbegriffe zum Einsatz. Erprobt wurden für diesen Untertest zwei Varianten.

Bei der *ersten Variante* wurde den Kindern ein von den Oberbegriffen unabhängiges Beispiel gegeben. Bei dieser Version hatten die Kinder erhebliche Probleme, passende Unterbegriffe zu den Kategorien Kleidungsstücke (N = 6) und Körperteile (N = 3) zu nennen. (Nachdem wir den Begriff Kleidungsstücke durch Anziehsachen ersetzt hatten, konnten die meisten Kinder passende Unterbegriffe nennen).

Bei der *zweiten Variante* wurde den Kindern zu jedem Oberbegriff ein passendes Beispiel angeboten. Dieser Aufgabentyp wurde 16mal erprobt. Ein Kinder verweigerte die Mitarbeit und ein weiteres Kind mit nicht-deutscher Muttersprache war zwar in der Lage, die passenden Unterbegriffe im Raum zu zeigen, konnte diese aber nicht selber benennen. Ansonsten kamen die Kinder mit diesem Aufgabentyp gut zurecht.

Dessen ungeachtet haben wir noch weitere Modifikationen erprobt. Dabei variierten wir die Durchführungsart. Hier hat es sich bewährt, den Kindern Bildmaterialen zum Sortieren an die Hand zu geben. Erprobt wurde diese Variante mit neun Kindern (sieben Mädchen und zwei Jungen). Wobei es bei der Durchführung weder Verständnisprobleme noch Verweigerungen gab. Deshalb haben wir entschieden, die zweite Variante in der modifizierten Durchführungsform in die Pilotierungsversion aufzunehmen.

Die Ergebnisse der Praxiserprobung werden abschließend nochmals tabellarisch zusammengefasst.

BiZ Aufgaben		BiP Aufgaben	
HA	5	-	-
KN	8	KN	9
SN	4	SN	9
BE	21	BE	61
-	-	WV	23
-	-	WP	23
-	-	PB	11
-	-	BK	3

Tabelle 20: Anzahl der Aufgaben für BiZ / BiP

Abschließend wurden beide Stufen nochmals im Gesamtzusammenhang in der Praxis erprobt. Dabei haben wir uns auf je zwei Durchgänge beschränkt. Wichtige Evaluationskriterien waren

- die Durchführungsdauer sowie
- die Abfolge der einzelnen Subtests.

Der zeitliche Aufwand für die Durchführung der Pilotierungsversionen betrug bei BiZ 20 - 25 Minuten, für BiP ergab sich eine Dauer von rund 30 Minuten. Bezüglich der Abfolge der einzelnen Untertests wurde bei BiZ darauf geachtet, dass die Eröffnungsaufgabe für das erste Kind eine Aufgabe ist, bei der es nonverbal agieren kann. In diesem Fall handelt es sich um den Untertest HA. Im weiteren Verlauf wechseln sich die einzelnen Untertests dann ab, da es sich um ein Gruppenscreening handelt, an dem mehrere Kinder gleichzeitig teilnehmen und es sich als unvorteilhaft darstellt, wenn alle Kinder nacheinander die gleichen Untertests bearbeiten müssen.

BiP wird mit der „Zeige-Aufgabe“ zur Überprüfung des Wortverständnisses eröffnet. Auch hier schien es sinnvoll, den Kindern zur Eingewöhnung jeweils als Aufwärmaufgaben solche anzubieten, bei denen sie nonverbal agieren können (WV und BK). Hier wurde ebenfalls darauf geachtet, dass die verschiedenen Aufgabentypen einander abwechseln, um die Aufmerksamkeit der Kinder über den gesamten Zeitraum hinweg zu halten.

5. Pilotierung von BiZ und BiP

Ein Test kann nur dann eine hohe Messqualität besitzen, wenn die Testaufgaben bestimmten Kriterien genügen. Die Pilotierung dient dazu, empirisch zu bestimmen, ob bzw. wieweit das der Fall ist. Gemäß der „Klassischen Testtheorie“ (vgl. z.B. Raatz/Lienert 1998) kann diese Frage geklärt werden, indem die Schwierigkeit sowie die Trennschärfe der Aufgaben sowie die Reliabilität und Validität der Untertest bestimmt wird.

Stufe 1 „Besuch im Zoo (BiZ)“

Die Stufe 1 von Delfin 4 „Besuch im Zoo (BiZ) wurde im Zeitraum von 5. bis 16. Februar 2007 in den Regionen Gütersloh und Hamm im Rahmen einer Pilotierungsstudie empirisch evaluiert. Auf der Basis der erhobenen Daten wurden

- die empirischen Kennwerte der Aufgaben überprüft
- die Reliabilität der Untertests ermittelt,
- die Übereinstimmensvalidität der Subtests KN, HA und SN mit dem Sprachscreening für das Vorschulalter (SSV)[48]

Die Pilotierungsstichprobe umfasste insgesamt Bogen von 678 Kindern (340 Mädchen, 334 Jungen, 4 Kinder ohne Angaben zur Geschlechtszugehörigkeit). Der Hauptanteil der Erhebungen fiel mit 64,1% (N = 435) auf die Stadt Hamm, während aus Gütersloh insgesamt 243 Bogen in die Studie einflossen. Da das Screening für Kinder im Alter von rund vier Jahren entwickelt worden ist, ist die Altersspanne der Stichprobe begrenzt. Die Kinder sind im Durchschnitt exakt 4 Jahre alt (48,11 Monate); die Standardabweichung beträgt 3,92 Monate. Die Kitabesuchsdauer der getesteten Kinder variiert zwischen weniger als 1 Monat bis hin zu 42 Monaten. Durchschnittlich beträgt sie 9,63 Monate. Die Verteilung weißt jedoch zwei deutliche Spitzen auf: 64,9% besuchen die Ein-

[48] Grimm 2003;

richtung seit 6 Monaten und 14,3% seit 18 Monaten.[49] Bei kategorialer Aufteilung der Besuchsdauer wird deutlich, dass der Großteil der Stichprobe (N = 454) sich aus Kindern zusammensetzt, die zum Erhebungszeitpunkt bis zu 6 Monaten in der Einrichtung sind.

Was die Familiensprache betrifft, so zeigt die Verteilung, dass bei 504 Kindern der Stichprobe die Familiensprache Deutsch ist. Bei den Kindern mit nicht-deutscher Muttersprache wurde am häufigsten türkisch (48,7% der Teilstichprobe der Kinder mit nicht-deutscher Muttersprache/9,2% der Gesamtstichprobe), polnisch (14,5%/ 2,7%) und russisch (13,7%/2,6%) angegeben. Auf multilinguales Aufwachsen wiesen die Angaben zu zwei Kindern hin. Dort wurden jeweils drei Sprachen aufgeführt.

Die Testleitung lag bei den Erhebungen mit BiZ fast ausschließlich bei den pädagogischen Fachkräften. Insgesamt haben 642 mal Erzieher/innen und nur 26 mal Lehrkräfte die Testdurchführung moderiert. Fast immer haben die Lehrer/innen die Beobachterrolle eingenommen und somit die Aufgabe der Protokollführung übernommen. In 1,6% der Fälle wurden hierzu keine Angaben gemacht.

Bei der Durchführung von BiZ hat ein geringer Prozentsatz von 1,3% (N = 9) der Kinder die Teilnahme an dem Verfahren komplett verweigert. Die unten aufgeführten Zahlen zur Verweigerungsquote in einzelnen Untertests machen deutlich, dass die Nachsprechaufgaben von den Kindern vergleichsweise am häufigsten nicht bearbeitet wurden: 8,2% KN und 7,9% SN.[50]

Untertest	Aktive Teilnahme		Verweigerung	
	N	%	N	%
HA	650	95,7	29	4,3
KN	622	91,6	57	8,4
SN	624	91,9	55	8,1
BB	652	96,0	27	4,0

Tabelle 21: Stichprobe / Verweigerung BiZ

[49] N = 656, bei 31 Kinder wurden bezüglich der Besuchsdauer der Tageseinrichtung keine Angaben gemacht.
[50] Allerdings ist dieser Prozentsatz – gemessen an Berichten aus der Literatur – immer noch gering.

Eine zentrale Funktion der Pilotierung besteht darin, empirisch zu bestimmen, ob bzw. wieweit die bei der Praxiserprobung ausgewählten Aufgaben spezifischen Standards entsprechen. Zu diesem Zweck wurden anhand der Pilotierungsstichprobe zu BiZ einerseits die Aufgabenkennwerte „Schwierigkeitsgrad“, „Item-Fit_Werte“ sowie „Trennschärfen“ (KTT/IRT) und Reliabilitätskennwerte zu den Untertests ermittelt. Im Weiteren wird - Untertest für Untertest – über die Ergebnisse berichtet.

		KTT		IRT		
Nr.	**Aufgabe**	**Schwierigkeit**	**Trennschärfe**	**MNSQ**	**t**	**Trennschärfe**
1	HA 1	86,0	,565	0,84	-2,0	,63
2	HA 2	77,0	,628	1,00	0,0	,76
3	HA 3	72,5	,617	1,00	-0,0	,76
4	HA 4	80,0	,625	0,98	-0,2	,76
5	HA 5	64,5	,584	1,19	2,9	,84
Durchschnitt		76,0	,604			,75

Tabelle 22: Aufgabenstatistik des Untertests HA

Der Untertest „Handlungsanweisungen ausführen (HA)“ wurde von 650 Kindern bearbeitet und von 29 Kindern verweigert. Es gab fünf im Schwierigkeitsgrad variierende Aufgaben zu bearbeiten, bei denen bis zu 11 Punkte erreicht werden konnten. Die Aufgabensatistiken machen deutlich, dass die Aufgaben dieses Untertests eher leicht sind, allerdings eine gewisse Streuung aufweisen. Die MNSQ-Werte und Trennschärfen stellen durchweg zufrieden. Die Aufgaben können also beibehalten werden.

Den Untertest „Nachsprechen von Kunstwörtern (KN)“ haben 623 Kindern bearbeitet. Hier galt es, acht Kunstwörter zu reproduzieren. Für jedes korrekt nachgesprochene Kunstwort gab es einen Punkt, somit konnten insgesamt bis zu acht Punkte erreicht werden. Für Schwierigkeit und Trennschärfe der einzelnen Kunstwörter ergaben sich folgende Werte:

		KTT		IRT		
Nr.	**Aufgabe**	**Schwierigkeit**	**Trennschärfe**	**MNSQ**	**t**	**Trennschärfe**
1	KN 1	83,1	,535	0,90	-1,3	,64
2	KN 2	72,9	,546	1,01	0,2	,66
3	KN 3	63,4	,570	0,97	-0,5	,69
4	KN 4	63,0	,564	1,00	-0,0	,69
5	KN 5	61,1	,566	0,97	-0,5	,69
6	KN 6	46,7	,457	1,11	1.9	,61
7	KN 7	49,5	,502	1,05	0,9	,65
8	KN 8	44,0	,517	1,01	0,3	,66
Durchschnitt		60,5	,532			,66

Tabelle 23: Aufgabenstatistik des Untertests KN

Diese Aufgaben sind insgesamt schwieriger, beinhalten aber ebenfalls sowohl leichtere als auch anspruchsvollere Aufgaben. Die MNSQ-Werte sind in Ordnung. Die Trennschärfen sind niedriger, als beim vorherigen Untertest der Fall, streuen aber ebenfalls in einem angemessenen Bereich.

Insgesamt 625 Kinder haben den Untertest „Sätze nachsprechen (SN)“ bestehend aus zwei sinnvollen und zwei sinnfreien Sätzen mit sechs bis neun Wörtern bearbeitet. Dabei konnten sie maximal bis zu 31 Punkte erreichen. Die Aufgabenstatistik dieses Untertests stellt sich wie folgt dar:

Nr.	Aufgabe	Schwierigkeit	Trennschärfe	MNSQ	t	Trennschärfe
1	SN 1	63,7	,728	1,01	0,2	,83
2	SN 2	49,2	,750	1,14	2,0	,88
3	SN 3	40,4	,718	1,07	1,0	,83
4	SN 4	51,3	,789	1,04	0,5	,90
Durchschnitt		51,2	,746			,86

Tabelle 24: Aufgabenstatistik des Untertests SN

Hier befinden sich die Aufgaben vorwiegend auf einem mittleren Schwierigkeitsniveau. Die Trennschärfen sind sehr hoch, die MNSQ-Werte zufriedenstellend. Es können alle Aufgaben beibehalten werden.

Der Untertest „Bildbeschreibung (BB)“ wurde von 652 Kindern bearbeitet. Er besteht jeweils aus einem Bild, das anhand des Analyserasters im Hinblick auf 12 Kriterien bewertet wird. Pro erfülltem Kriterium kann ein Punkt vergeben werden. Somit ergibt sich eine Höchstpunktzahl von bis zu 12 Punkten. Nachfolgend sind Schwierigkeit sowie Trennschärfe pro Item aufgelistet. Hierbei tritt zutage, dass es den Kindern bei vielen Kriterien noch recht schwer fällt, diese zu erfüllen. Allerdings gibt es von Kriterium zu Kriterium große Unterschiede. Im Durchschnitt jedoch ist das der Untertest, bei dem die Kinder die anspruchsvollsten Leistungen bringen müssen. Die Trennschärfen sind nicht allzu gut, aber immer noch akzeptabel. Bei diesem Untertest merkt man, dass die Kinder mit der Ausbildung ihrer Erzählfähigkeit noch am Anfang stehen. Die Aufgaben bzw. Kriterien können deshalb beibehalten werden.

		KTT		IRT		
Nr.	**Aufgabe**	**Schwierigkeit**	**Trennschärfe**	**MNSQ**	**t**	**Trennschärfe**
1	Erzählt spontan	40,5	,478	1,02	0,5	,62
2	Fasst das Wesentliche gleich am Anfang kurz zusammen	13,8	,452	0,94	-0,8	,56
3	Identifiziert mindestens drei Akteure	53,1	,399	1,13	2,6	,56
4	Beschreibt mindestens ein Ereignis	65,2	,472	0,93	-1,5	,61
5	Verknüpft mehrere Akteure und/oder Ereignisse sinnvoll	18,6	,523	0,88	-1,9	,63
6	Markiert einen Höhepunkt	7,5	,243	1,08	0,7	,34
7	Stellt logische und/oder zeitliche Verknüpfungen her	7,7	,386	0,94	-0,5	,47
8	Macht Ortsangaben	25,8	,362	1,16	1,8	,51
9	Kennzeichnet näher	7,5	,275	1,06	0,6	,37
10	Verwendet mindestens eine Dreiwortäußerung	52,5	,483	0,99	-0,3	,62
11	Reiht mindestens zwei Sätze aneinander	21,0	,564	0,87	-2,4	,67
12	Verwendet Haupt-, Nebensatzkonstruktionen	8,7	,398	0,94	-0,6	,49
	Durchschnitt	26,8	,419			,54

Tabelle 25: Aufgabenstatistik des Untertests BB

Kennwerte zu den „Farben“

Abschließend zur Darstellung der Pilotierung von BiZ werden Reliabilitäts- und Aufgabenkennwerte der Untertests auch noch bezogen auf die „Püppchen-Farbe“ (Sitzposition am „Spielbrett“) des Kindes aufgelistet.

Untertest	Farbe	Zuverlässigkeit[51]	Schwierigkeit			Trennschärfe		
			Min	Max	MW	Min	Max	MW
HA	blau	0,70	59,3	82,0	71,6	0,489	0,560	0,524
	grün	0,77	63,5	89,0	74,7	0,564	0,663	0,631
	rot	0,77	66,5	92,0	81,2	0,381	0,730	0,616
	gelb	0,80	63,5	89,0	78,1	0,586	0,738	0,669
KN	blau	0,82	38,4	80,2	62,2	0,392	0,658	0,542
	grün	0,82	30.0	80,9	54,5	0,409	0,646	0,534
	rot	0,77	42,6	86,7	60,1	0,313	0,597	0,485
	gelb	0,84	41,5	85,9	65,1	0,480	0,662	0,573
SN	blau	0,87	53,6	71,2	60,5	0,718	0,761	0,738
	grün	0,85	31,8	65,2	49,0	0,652	0,737	0,712
	rot	0,85	41,4	69,3	53,9	0,680	0,797	0,715
	gelb	0,90	34,7	49,5	41,2	0,734	0,846	0,793
BB	blau	0,79	8,5	70,6	32,2	0,156	0,568	0,418
	grün	0,75	7,0	62,6	25,5	0,212	0,576	0,401
	rot	0,79	4,5	65,4	26,4	0,288	0,555	0,440
	gelb	0,77	4,7	62,0	22,8	0,195	0,542	0,403

Tabelle 26: Kennwerte der „Farben“

Reliabilität

Abschließend sollen noch die Reliabilitätskoeffizienten berichtet werden. Bei der Gesamtskala BiZ ist Croncachs Alpha r = 0,8.; bei den einzelnen Untertests liegt Cronbachs Alpha zwischen r = 0,76 und r = 0,87.

	N	Reliabilitätskoeffizienten*	
Pilotierung	678	r_{GES} = .85	r_{HA} =.76; r_{KN} = .81; r_{SN} = .87; r_{BB} = .78

Tabelle 27: Reliabilität BiZ (Gesamtskala und Unterskalen)

Bei BiZ wird im Protokollbogen dazu aufgefordert, am Ende des Verfahrens anhand der Beobachtungen bei der Bildbeschreibung (BB)[52], die Sprechprobleme und -auffälligkeiten der Kinder einzuschätzen. Hier zeigte sich, dass die pädagogischen Fachkräfte bzw. Lehrkräfte am häufigsten von Artikulationsproblemen berichten. Die

[51] Cronbachs Alpha
[52] Bei der Endversion wurde dieser Aufgabenbereich in Bild beschreiben (BB) umbenannt, da dieser Begriff eher den Fähigkeiten von etwa Vierjährigen entspricht.

folgende Auflistung zeigt die prozentuale Verteilung der festgehaltenen Sprachauffälligkeiten in absteigender Reihenfolge.

Sprachauffälligkeiten	%
Einzelne Laute fehlen oder klingen komisch.	31,8
Es werden falsche Sätze gebildet.	17,7
Das auf dem Bild dargestellte Ereignis wird nicht oder falsch erfasst.	17,1
Einzelne Äußerungen sind unverständlich.	15,7
Einzelne Lautverbindungen fehlen oder klingen komisch.	15,2
Äußerungen bleiben unbestimmt.	13,2
Es fehlen mehrfach passend Wörter.	8,9

Tabelle 28: Häufigkeit von Sprachauffälligkeiten (Pilot 2007)

Insgesamt unterstreichen diese Befunde, dass Biz sowohl auf Gesamttestebene, als auch auf Untertestebene zuverlässig misst. Das gilt umso mehr, als es sich bei dem Grobscreening um einen kurzen Test handelt.

Abschließend wird noch das Ergebnis von Korrelationsstudien, in denen einzelne Untertests von BiZ mit korrespondierenden Untertests anderer Sprachtests in Beziehung gesetzt wurden. Die ermittelten Koeffizienten indizieren das Ausmaß an Übereinstimmungsvalidität zu.

Parallel zu Delfin 4 Stufe 1 wurde mit denselben Kindern, sofern die Eltern ihr Einverständnis dazu gaben, ein weiteres diagnostisches Verfahren (SSV von Hannelore Grimm) durchgeführt, dessen Validität bereits hinreichend nachgewiesen wurde. Beim SSV handelt es sich um ein Sprachscreening für das Vorschulalter mit den Untertests Phonologisches Arbeitsgedächtnis für Nichtwörter (PGN), Morphologische Regelbildung (MR) und Satzgedächtnis (SG). Durchgeführt werden je nach Alter des Kindes jeweils zwei Untertests. Bei Kindern von 3;0 bis 3;11 Jahren finden die Untertests PGN (mit insgesamt 13 Items) und MR (mit insgesamt 10 Items) Anwendung; mit den Kindern von 4;0 bis 5;11 Jahren werden die Untertests PGN (mit insgesamt 18 Items) und SG (mit insgesamt 15 Items) bearbeitet. Die Durchführung des SSV nahm pro Kind ca. 10 Minuten in Anspruch.

Miteinander korreliert wurden:

- Delfin 4: Kunstwörter nachsprechen (KN) – SSV: Phonologisches Arbeitsgedächtnis für Nichtwörter (PGN)
- Delfin4: Handlungsanweisungen ausführen (HA) – SSV: Morphologische Regelbildung (MR)
- Delfin4: Sätze nachsprechen (SN) – SSV: Satzgedächtnis (SG).

Insgesamt wurde der **SSV** bei 436 (= 63,5% von 687) Kindern eingesetzt. Nicht zum Einsatz kam er bei 251 (= 36,5% von 687) Kindern. Im Folgenden eine Übersicht über die Verweigerungsquoten vom SSV – Gesamt sowie der einzelnen Untertests:

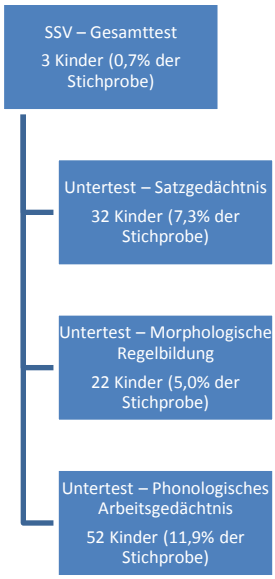


Abbildung 1: Verweigerung SSV

Im Folgenden sind die Ergebnisse der Übereinstimmungsvalidität zwischen den oben genannten Untertests von Delfin4 und SSV aufgeführt.

N	Delfin4-KN	SSV-PGN
388	0,50**	

Abbildung 2: Übereinstimmungsvalidität KN-GES/SSV-PGN
N = Stichprobengröße
** = Die Korrelation (nach Pearson) ist auf dem Niveau von 0,01 (2-seitig) signifikant

Es handelt sich um eine mittlere Übereinstimmung.

N	Delfin4-HA	SSV-MR
235	0,51**	

Abbildung3: Übereinstimmungsvalidität HA-GES/SSV-MR
N = Stichprobengröße
** = Die Korrelation (nach Pearson) ist auf dem Niveau von 0,01 (2-seitig) signifikant

Auch hier besteht eine mittlere Übereinstimmung.

N	Delfin4-SN	SSV-SG
225	0,67**	

Abbildung 4: Übereinstimmungsvalidität SN-GES/SSV-SG
N = Stichprobengröße
** = Die Korrelation (nach Pearson) ist auf dem Niveau von 0,01 (2-seitig) signifikant

Hier besteht eine hohe Übereinstimmung.

Die Unterschiede im Ausmaß der Übereinstimmung erklären sich aus der mehr oder minder gegebenen inhaltlichen Übereinstimmung zwischen den Delfin 4- und den SSV-Untertests. Insgesamt sprechen die Befunde dagfür, dass mit Delfin 4 – Stufe 1 eine valide Erfassung bestimmter Sprachfähigkeiten möglich ist.

Stufe 2: Besuch im Pfiffikus-Haus

Im Rahmen einer Pilotierungsstudie wurde BiP im April 2007 in Köln evaluiert. Auf der Basis der erhobenen Daten wurden

- die empirischen Kennwerte der Aufgaben überprüft
- die Reliabilität der Untertests ermittelt,
- die Übereinstimmensvalidität der Delfin 4-Untertests zum aktiven und passiven Wortschatz mit dem AWST-R (Aktiver Wortschatztest für drei- bis fünfjährige Kinder[53] – revidierte Form) ermittelt.

An der Pilotierung nahmen 542 Kinder teil. Für 207 Kinder war die Teilnahme an der zweiten Verfahrensstufe fakultativ, da sie laut dem Gruppenscreening „Besuch im Zoo BiZ“ zum vertiefenden Verfahren geladen wurden. Gemäß den gesetzlichen Regelungen wurden die Testungen von Lehrkräften durchgeführt. 335 Kinder wurden nach Stufe 1 als unauffällig eingestuft und nahmen freiwillig an der Pilotierung teil. Die Testungen wurden von pädagogischem Fachpersonal in den jeweiligen Kindertageseinrichtungen vorgenommen.

An der Pilotierung nahmen insgesamt 542 Kinder teil. Die Geschlechterverteilung war mit 44,5% Mädchen, 43,5% Jungen recht ausgewogen. Bei 12% der Kinder fehlt die Angabe zur Geschlechtszugehörigkeit. Die Stichprobe setzte sich aus 207 Kindern zusammen, die nach BiZ zur zweiten Verfahrenstufe eingeladen und aus 335, die nach Stufe 1 als unauffällig eingestuft worden waren. Bei letzteren wurde das Verfahren von den pädagogischen Fachkräften in den Einrichtungen durchgeführt. Die Bogen der Kinder, welche zur zweiten Stufe eingeladen worden waren, wurden - gemäß den gesetzlichen Regelungen - von Lehrkräften erhoben.

Als Familiensprache wurde in 43% der Fälle deutsch angegeben, gefolgt von türkisch mit 19,2%. 5,9% der Kinder wachsen mit 2-4 Familiensprachen auf. Bei 87 Kindern die Angabe zur Familiensprache.

[53] vgl. Kiese-Himmel 2005;

529 Kinder haben BiP bearbeitet, 2,4% der Kinder verweigerten die Durchführung des gesamten Screenings. Betrachtet man die Verweigerungszahlen (vgl. folgende Übersicht) bezogen auf die einzelnen Untertests, so wird deutlich, dass nur 2,6% der Kinder den Untertest WV zum passiven Wortschatz verweigert haben. Am häufigsten trat Verweigerung beim Untertest SN auf.

Untertest	Aktive Teilnahme		Verweigerung	
	N	%	N	%
WV	542	97,4	14	2,6
BK	525	96,9	17	3,2
KN	502	92,6	40	7,4
SN	478	88,2	64	11,8
PB	502	92,6	40	7,4
WP	502	92,6	40	7,4
BE	491	90,6	51	9,4

Tabelle 29: Verweigerungsstatistik

Am Ende des Verfahrens hatten die Testleiter die Möglichkeit Beobachtungen, die sie während der Durchführung des gesamten Verfahrens und im speziellen während der Bilderzählung gemacht haben, zu kennzeichnen. Am häufigsten wurden hier die mehrfache Bildung von unvollständigen/bruchstückhaften Sätzen, das Verwenden von falschen oder keinen Artikeln und dass die Kinder keine vollständigen Sätze verwenden, angekreuzt. Die folgende Tabelle gibt die Rangordnung der notierten Auffälligkeiten wieder.

Sprachauffälligkeiten	%
Es werden mehrfach unvollständige/bruchstückhafte Sätze gebildet	29,2
Es werden mehrfach keine oder falsche Artikel verwendet *(z.B. „Da schwimmt Ente“; „Der hat diese Regenschirm“)*	29,0
Es werden keine vollständigen Sätze verwendet	25,8
Einzelne Laute fehlen oder werden fehlgebildet *(b, k usw.)*	24,5
Die Wortordnung wird mehrfach nicht eingehalten *(Verb steht nicht an der richtigen Stelle, z.B. „Der Mann raus geht“; „…ob die ist im Regenschirm“)*	22,0
Einzelne Lautverbindungen fehlen oder werden fehlgebildet *(str, spr, schw usw.)*	20,7
Einzelne Äußerungen sind unverständlich	20,5
Es fehlen mehrfach passende Wörter *(z.B. „Quakquak“ statt Ente; „Wie heißt das?“; „Das da!“ usw.)*	19,7
Es werden mehrfach nur Einzelwörter geäußert	17,7
Es scheinen Hemmungen vorhanden, sich zu äußern	17,1
Äußerungen bleiben unbestimmt *(z.B. „So halt“)*	10,7

Tabelle 30: Angaben zu Auffälligkeiten im Sprachgebrauch

Abschließend wird nochmals ein Überblick zu den Kennwerten von BiZ gegeben.

Unter-test	Schwierigkeit			Trennschärfe		
	Min	Max	MW	Min	Max	MW
WV	34,0	89,6	68,1	,337	,596	,466
BK	58,8	73,8	67,5	,501	,584	,552
KN	65,7	87,8	73,6	,385	,456	,421
SN	71,3	79,5	75,3	,775	,853	,808
PB	25,0	69,5	38,9	,431	,656	,537
WP	17,0	88,5	67,1	,364	,697	,522
BE	29,5	35,8	32,3	,715	,802	,768

Tabelle 31: Spannbreite der Kennwerte der Untertests BiP

Nachfolgend werden die Kennwerte der Klassischen Testtheorie (KTT) sowie der Item-Response-Theorie (IRT) berichtet. Die Itemselektion erfolgte in zwei Schritten: Zuerst

wurde die Itemauswahl auf der Grundlage der KTT unter Berücksichtigung von Schwierigkeit, Trennschärfe und der internen Konsistenz getroffen. Im Anschluss erfolgten, zur Absicherung der Selektion, Analysen entsprechend der IRT mit dem Programm ConQuest. Diese ermöglichen die Prüfung des Fits eines jeden Items zum angenommenen Modell. Zur Beurteilung des Item-Fits werden die gewichteten Abweichungsquadrate (weighted mean square „MNSQ“) und T-Wert-Statistiken betrachtet. Ebenfalls wurde die Item-Diskrimination berücksichtigt.

Delfin 4 umfasst sowohl Untertests denen das dichotome als auch das ordinale Rasch-Modell zugrunde liegen. Die Untertests „Wortverständnis (WV)“ und „Kunstwörter nachsprechen (KN)“ vereinen dichotome Items. Die Aufgaben dieser Subtests können entweder richtig oder falsch gelöst werden. „Wortproduktion (WP)“, „Begriffsklassifikation (BK)“ und „Pluralbildung (PB)“ gehören zu den ordinalen Modellen, bei denen auch für Teillösungen (vollständig gelöst, teilweise gelöst und nicht gelöst) Punkte vergeben werden. Diese nennt man auch partial-credit-Modelle. Während bei den Untertests „Pluralbildung (PB)“ und „Wortproduktion (WP)“ Punkte von 0 bis 2 vergeben werden können, sind es bei „Begriffsklassifikation (BK)“ bis zu vier. Die Untertests „Sätze nachsprechen (SN)“ und „Bilderzählung (BE)“ zählen zu den Strukturgleichungsmodellen, bei denen die Punktzahlen je Aufgabe variieren.

Wortverständnis (WV)

Der Untertest „Wortverständnis (WV)“ zur Überprüfung des rezeptiven Wortschatzes wurde von insgesamt 542 Kindern bearbeitet. Der Untertest „Wortverständnis (WV)“ besteht aus 23 Items der Wortgruppen: Nomen (8), Verben (7), Adjektive (5) und Präpositionen (3). Es gab insgesamt 23 Punkte zu erreichen. Die Trennschärfekoeffizienten der einzelnen Items sind akzeptabel bis gut. Lediglich Item 19 (kariert) weist eine niedrige Trennschärfe auf. Die Analysen mit der IRT zeigen für dieses Item einen schlechten Fit an. Da der Untertest zu den leichtesten und das Item den Schwierigsten gehört, wurde es im Verfahren belassen. Zwar zeigen zwei weitere Items einen erhöhten T-Wert, da jedoch der MNSQ nicht erhöht ist und die Trennschärfekoeffizienten sowohl in der KTT als auch IRT akzeptabel sind, wird dieses nicht als problematisch empfunden.

		KTT		IRT		
Item		Schwierigkeit	Trennschärfe	MNSQ	T	Trennschärfe
1	Fledermaus	80,1	,506	0,92	-1,2	,57
2	Kapuze	62,7	,494	1,02	0,5	,56
3	Spaghetti	89,4	,401	0,96	-0,3	,46
4	Wut	65,8	,463	1,05	1,0	,54
5	Streit	74,0	,506	0,97	-0,6	,57
6	Stängel	31,1	,427	0,99	-0,1	,50
7	Küken	75,3	,474	0,98	-0,3	,54
8	Einkaufswagen	90,9	,357	0,99	-0,1	,42
9	sich freuen	72,3	,488	0,99	-0,2	,56
10	kämpfen	79,7	,277	1,19	2,7	,35
11	öffnen	69,1	,539	0,95	-0,9	,59
12	warten	70,3	,437	1,06	1,0	,51
13	ziehen	64,6	,491	1,02	0,3	,56
14	schälen	65,5	,570	0,91	-1,8	,63
15	tanken	78,2	,515	0,93	-1,1	,58
16	nass	79,1	,389	1,04	0,7	,46
17	eckig	64,5	,621	0,84	-3,3	,68
18	krumm	40,9	,398	1,07	1,4	,48
19	kariert	31,0	,187	1,32	5,5	,27
20	lustig	80,6	,329	1,15	2,1	,39
21	auf	89,0	,448	0,90	-1,1	,49
22	vor	63,7	,630	0,93	-3,6	,68
23	hinter	60,3	,558	0,92	-1,6	,62
Durchschnitt		68,6	,456			,52

Tabelle 32: Aufgabenkennwerte „Wortverständnis“ (WV)

Begriffsklassifikation (BK)

Die Sortieraufgabe umfasst drei Items (Spielzeug, Kleidung, Obst) und setzt sich jeweils aus vier Unterbegriffen und zwei Ablenkern zusammen. Pro Oberbegriff können bis zu vier Punktwerte vergeben werden. Somit ergibt sich ein Gesamtpunktwert von 12. Der Untertest zeigt durchweg gute Trennschärfekoeffizienten (Min: ,496; Max: ,525; MW: ,515) Auch die Analysen mit der IRT ergaben für alle Items eine gute Übereinstimmung mit dem Modell. Die Kennwerte der Analysen sind Tabelle 3 zu entnehmen.

		KTT		IRT		
Item		Schwierigkeit	Trennschärfe	MNSQ	T	Trennschärfe
1	Spielzeug	65,5	,496	1,05	0,7	,80
2	Kleidung	75,5	,523	0,98	-0,3	,78

3	Obst	78,7	,525	1,00	-0,0	,78
Durchschnitt		73,2	,515			,79

Tabelle 33: Aufgabenkennwerte „Begriffsklassifikation“ (BK)

Wortproduktion (WP)

Zur Überprüfung des aktiven Wortschatzes wird der Untertest „Wortproduktion (WP)“ eingesetzt. Überprüft werden 23 Items der Wortkategorien Nomen (8), Verben (7), Adjektive (5) und Präpositionen (3). Insgesamt können bis zu 46 Punkte erreicht werden. Die Trennschärfekoeffizienten der Items stellen sich gut bis hervorragend dar. Der weighted MNSQ ist nicht > 1.2. Lediglich Item 18 (stumpf) weist einen schlechten Fit zum Modell auf. Da es sich um eines der schwersten Items dieses recht leichten Untertests handelt, wurde es weiterhin im Verfahren belassen.

		KTT		IRT		
Item		Schwierigkeit	Trennschärfe	MNSQ	T	Trennschärfe
1	Schrank	77,5	,643	0,91	-1,2	,70
2	Handschuh	76,5	,510	1,08	1,1	,59
3	Schildkröte	72,5	,681	0,84	-2,4	,73
4	Gummistiefel	69,0	,399	1,18	2,0	,48
5	Hubschrauber	71,5	,577	0,99	-0,1	,63
6	Kleiderbügel	19,0	,336	1,08	1,0	,39
7	Marienkäfer	78,5	,504	1,08	1,0	,59
8	Schlüssel	90,5	,443	0,95	-0,4	,53
9	bauen	78,5	,471	1,19	2,0	,54
10	schieben	74,0	,570	0,99	-0,1	,64
11	(Seil-)springen	77,0	,461	1,18	2,2	,54
12	küssen	80,0	,469	1,12	1,4	,55
13	reiten	61,5	,746	0,71	-5,4	,78
14	pflücken	41,0	,550	0,96	-0,8	,61
15	schwimmen	83,5	,530	0,98	-0,2	,59
16	dick	64,0	,585	1,02	0,3	,65
17	dunkel	79,5	,682	0,76	-3,1	,72
18	stumpf	20,0	,198	1,27	3,6	,26
19	fröhlich	46,5	,503	1,04	0,7	,57
20	alt	47,5	,609	0,89	-2,2	,66
21	in	66,5	,626	0,94	-1,0	,68
22	neben	51,0	,639	0,86	-2,7	,69
23	hinter	54,0	,614	0,92	-1,4	,67
Durchschnitt		64,3	,537			,60

Tabelle 34: Aufgabenkennwerte „Wortproduktion“ (WP)

Sätze nachsprechen (SN)

Der Untertest Sätze nachsprechen (SN) umfasst neun Sätze unterschiedlicher Länge und Komplexität. Die Höchstpunktzahl pro Satz liegt zwischen 6 bis 9 Punkten. Insgesamt kann man 63 Punkte in der Feinauswertung erhalten, in der Grobauswertung sind höchstens 7 Punkte möglich. Die Trennschärfen der einzelnen Items sind durchgängig hervorragend. Dies spiegelt sich auch in der Reliabilität der Untertests wider. Hier konnte ein Koeffizient von 0,93 erreicht werden.

	Item	KTT Schwierigkeit	KTT Trennschärfe
1	Murat schläft in seinem neuen Bett.	76,3	,763
2	Die Katze wird von Lena gefüttert.	78,7	,779
3	Heute rennen die Kinder schnell zur Schule.	77,1	,838
4	Ayla kann ihren Schlüssel nicht finden.	79,3	,778
5	Tom nimmt dem Hund den Knochen weg.	76,6	,800
6	Anna ist traurig, weil sie ihren Ball verloren hat.	74,1	,765
7	Das lustige Eis tanzt einen Baum.	74,2	,778
8	Heute kitzelt der fleißige Hund einen Apfel.	64,4	,729
9	Wenn die Hose singt, klettert sie über die Straße.	71,9	,804
	Durchschnitt	74,7	,781

Tabelle 35: Aufgabenkennwerte „Sätze nachsprechen“ (SN)

Pluralbildung (PB)

Der Untertest setzt sich aus 11 Items zusammen. Die Kinder müssen sowohl den Plural von sechs sinnvollen Wörtern als auch von fünf Fantasietieren bilden. Pro Item sind bis zu 2 Punkte zu vergeben, somit ergibt sich für diesen Subtest eine Gesamtpunktzahl von bis zu 22 Punkten. Der Untertests ist mit einer durchschnittlichen Schwierigkeit von 38,8 einer der schwierigsten von BiP. Die Trennschärfen der einzelnen Items sind gut. Einzig Item 5 (Löffel) zeigt eine niedrige, wenn auch akzeptable Trennschärfe. Die Analysen mit der IRT weisen auf einen schlechten Fit für dieses Item hin. Sowohl der weighted MNSQ als auch der zugehörige T-Wert bewegen sich im signifikanten Bereich. Da es sich um eines der leichtesten Items des ansonsten schweren Untertests handelt, wurde das Item beibehalten. Nachfolgend sind Schwierigkeit und Trennschärfkoeffizienten der einzelnen Plurale aufgelistet.

		KTT		IRT		
Item		Schwierigkeit	Trennschärfe	MNSQ	T	Trennschärfe
1	Schaf	59,0	,504	1,09	1,3	,65
2	Auto	65,0	,557	1,00	-0,0	,70
3	Bett	41,5	,585	0,95	-0,8	,70
4	Loch	46,0	,685	0,81	-3,3	,78
5	Löffel	85,0	,221	1,26	2,8	,38
6	Nagel	23,0	,479	0,99	-0,2	,59
7	Dopf	25,5	,559	0,93	-1,1	,65
8	Dagel	13,5	,500	0,89	-1,4	,57
9	Mate	34,5	,547	1,08	1,2	,66
10	Pemmich	15,0	,441	0,96	-0,4	,53
11	Sobel	19,5	,500	0,96	-0,5	,60
Durchschnitt		38,8	,507			,62

Tabelle 36: Aufgabenkennwerte „Pluralbildung“ (PB)

Kunstwörter nachsprechen (KN)

Das Nachsprechen von Kunstwörtern besteht aus neun zwei- bis viersilbigen Items. Pro korrekt reproduziertem Kunstwort kann ein Punkt vergeben werden. Somit ergibt sich für diesen Untertest ein Gesamtpunktwert von 9 Punkten. Die Items weisen sowohl in der Klassischen als auch in der Item-Response-Theorie gute Trennschärfen auf.

		KTT		IRT		
Item		Schwierigkeit	Trennschärfe	MNSQ	T	Trennschärfe
1	nibu	91,4	,241	1,09	0,8	,39
2	fegitt	87,0	,400	1,01	0,1	,54
3	lupori	77,8	,508	0,95	-0,7	,65
4	robiwamm	70,3	,519	0,96	-0,6	,66
5	jakedu	80,7	,488	0,94	-0,7	,62
6	kutabodi	72,3	,486	1,01	0,2	,63
7	hamifuko	72,6	,553	0,91	-1,4	,68
8	tiworepi	52,1	,424	1,06	1,1	,60
9	gotumalik	66,2	,430	1,09	1,5	,60
Durchschnitt		74,5	,450			,60

Tabelle 37: Aufgabenkennwerte „Kunstwörter nachsprechen“ (KN)

Bilderzählung (BE)

Die Bilderzählung besteht aus 4 Bilditems. Beim ersten und zweiten Bild können bis zu 15 Punkte, beim dritten Bild 14 und beim letzten Bild bis zu 17 Punkte erreicht werden. Insgesamt ergibt das eine Höchstpunktzahl von 61 Punkten für den Untertest BE. Mit einer durchschnittlichen Schwierigkeit von 31,5 handelt es sich um den schwersten Untertest des Screenings. Die Trennschärfen sind insgesamt hervorragend und auch die Reliabilität liegt bei 0,92. Nachfolgend werden Schwierigkeitsgrade und Trennschärfekoeffizienten der einzelnen Items dargestellt.

		KTT	
	Item	Schwierigkeit	Trennschärfe
1	BE1	31,8	,847
2	BE2	30,1	,809
3	BE3	32,6	,821
4	BE4	31,6	,821
	Durchschnitt	31,5	,824

Tabelle 38: Aufgabenkennwerte „Bilderzählung“ (BE)

Die Pilotierungsversion von BiP umfasst somit sieben Untertests mit 82 Items. Insgesamt können bis zu 182 Punkte erreicht werden. Die Schwierigkeitsgrade der Untertests von BiP variieren von 13,5 und 91,4. Im Folgenden sind Schwierigkeiten und Trennschärfekoeffizienten der einzelnen Untertests aufgelistet.

Untertest	Schwierigkeit			Trennschärfe		
	Min	Max	MW	Min	Max	MW
WV	31,1	90,9	68,6	0,19	0,63	0,46
BK	65,5	78,7	73,2	0,49	0,52	0,51
KN	66,2	91,4	74,5	0,24	0,55	0,45
SN	64,4	79,3	74,7	0,73	0,84	0,78
PB	13,5	85,0	38,8	0,22	0,68	0,51
WP	19,0	90,5	64,3	0,20	0,75	0,54
BE	30,1	32,6	31,5	0,81	0,85	0,82

Tabelle 39: Kennwerte KTT Pilotierungsversion BiP

Die bei der internen Konsistenzschätzung resultierenden Koeffizienten der einzelnen Untertests bewegen sich zwischen 0,70 und 0,93. Die Zuverlässigkeit des Gesamttests BiP liegt bei 0,95. Damit misst Delfin 4, Stufe 2 ausgesprochen genau.

Die nachfolgende Zusammenstellung gewährt einen Vergleich der Reliabilitäts-koeffizienten der einzelnen Untertests.

Untertest	N	Reliabilität
WV	524	0,88
BK	523	0,70
KN	492	0,77
SN	428	0,93
PB	475	0,83
WP	488	0,91
BE[54]	452	0,92
BiP	373	0,95

Tabelle 40: Reliabilitätskoeffizienten von BiP

Zur Überprüfung der Übereinstimmungsvalidität wurde neben BiP der „Aktive Wort-schatztest für drei- bis fünfjährige Kinder – revidierte Version (AWST-R)“ (Kiese-Himmel 2005) eingesetzt. In den Korrelationsstudien wurden die Ergebnisse der Delfin 4 Unter-tests „Wortverständnis (WV)“ und „Wortproduktion (WP)“ jeweils mit denen des AWST-R in Beziehung gesetzt.

Von der Gesamtstichprobe haben 402 der Kinder (74,2%) zusätzlich an dem Verfahren AWST-R teilgenommen. Von diesen Kindern waren 238 nach Stufe 1 als unauffällig eingestuft worden, während die restlichen 40,8% der Stichprobe am vertiefenden Ver-fahren teilnehmen mussten. 8% darunter verweigerte die komplette Bearbeitung der Aufgaben bestehend aus der Bildbenennung von 24 Verben und 51 Substantiven. Die Stärke des Zusammenwirkens des Untertests WV von Delfin 4 mit den Verben des AWST-R sowie des Zusammenwirkens des Untertests WP mit den Substantiven des AWST-R ergibt sich aus folgender Auflistung:

[54] Die Berechnungen wurden auf der Ebene der Analysekriterien durchgeführt.

N	D4-WV AWST-R_V
368	0,77**
N	**D4-WP AWST-R_S**
357	0,80**

Tabelle 41: Korrelationsstatistik
D4 = Delfin 4
** = auf dem Niveau von 0,01 (2-seitig) signifikant.

Diese Befunde sprechen dafür, dass der Wortschatz mit den beiden Delfin-Untertests WV und WP valide gemessen werden kann.

5. Praxisevaluation von Delfin 4

Die Pilotierungen von Delfin 4 boten einen geeigneten Rahmen, um in der Praxis zu evaluieren, wie die pädagogischen Fach- und Lehrkräfte die Handhabbarkeit des Testmaterials einschätzen.

Stufe 1 „Besuch im Zoo (BiZ)“

Den Begleiter/innen und Beobachter/innen der Durchführung von BiZ wurden jeweils 3-seitige Evaluationsbogen mit Freiumschlägen zur Verfügung gestellt. (Dieser Bogen befindet sich im Anhang). Dabei wurden den an der Pilotstudie beteiligten 44 Kindertageseinrichtungen und 15 Grundschulen jeweils mehrere Evaluationsbogen übersandt, da wir nicht einschätzen konnten, wie viele Lehrer/innen und Erzieher/innen innerhalb dieser Institutionen bereits Erfahrungen mit der Durchführung von BiZ gemacht haben und das Testmaterial deshalb fundiert einschätzen können. Insgesamt 71 Evaluationsbögen wurden uns in der Zeit vom 25.2.2007 bis 5.3.2007 zurückgesandt.

Das Befragungsinstrument beinhaltet zehn geschlossene und 14 offene Fragen. Bei den geschlossenen Fragen handelt es sich um Richtig/Falsch-Aufgaben, die mit Ja oder Nein beantwortet werden können und mit einer Trichterfunktion in weitere zehn offene Fragen überleiten (z. B. „Wenn nein, warum nicht?“). Diese optionalen Fragen wurden also nur noch von einem Teil der Befragten bearbeitet. Darüber hinaus gibt es weitere vier offene Fragen, in denen nach speziellen Problemen oder Verbesserungsvorschlägen gefragt wird. Die Antworten auf die geschlossenen Fragen konnten unmittelbar quantifiziert werden, die Antworten auf die offenen Fragen mussten zunächst inhaltsanalytisch bearbeitet werden, konnten dann aber ebenfalls quantifiziert werden.

Im Folgenden werden die Erkenntnisse aufgeführt, die sich aus der Beantwortung der Evaluationsbogen ergeben haben.

Zum thematischen Rahmen:

„Hatten die Kinder Probleme mit dem thematischen Rahmen „Besuch im Zoo“? – Wenn ja, welche?“

Etwa zwei Drittel der Befragten sind der Ansicht, dass die Kinder gut mit dem Thema „Besuch im Zoo“ klar gekommen sind (66,2%). 32,4% der Befragten halten es für problematisch, dass einige der Kinder noch nie im Zoo waren, 29,6% finden es schwierig, dass manche Tiernamen den Kindern unbekannt waren. Von denjenigen, die das Thema für unproblematisch halten, fügen zwei der Befragten hinzu, dass die Kinder zwar selbst nie im Zoo gewesen seien, dass dies aber für die Durchführung kein Problem gewesen sei. In diesem Zusammenhang wird auch genannt, dass den Kindern das Thema aus Büchern und Geschichten vertraut sei. Dreimal wird festgehalten, dass die Kinder sich für dieses Thema interessierten.

Zum Material:

„Sind die Kinder mit dem Material gut klar gekommen? – Wenn nein, warum nicht?“

57,7% der Befragten geben an, dass die Kinder gut mit dem Material klar gekommen sind. Für 63,4% der Evaluierenden war der mehrteilige Spielplan ein Problem, da die Teile während der Durchführung verrutscht sind. Eine Erzieherin gab an, dass sie den Plan mithilfe von Klebebändern auf dem Tisch befestigt habe und so das Problem umgangen habe.

29,6% der Befragten sind der Ansicht, dass die Bilder auf dem Spielplan für die Kinder zu klein sind; 28,2% wünschen sich weniger Details. Die Anordnung der Spielfiguren auf dem Spielplan wird von 29,6% der Befragten als verwirrend für die Kinder empfunden, weil die Kinder ihre Püppchen nicht selbstständig setzen konnten (Startfeld zu schlecht zu erkennen (12,7%). Sie seien es gewohnt, ihre Püppchen nach rechts abzustellen (14,3%).

Ganz konkret wird einige Male genannt, dass der Zooeingang nicht gut zu erkennen sei (12,7%). Die Umrandung des Spielplans sei nicht nur zu klein, sondern es sei auch ungünstig, dass eine Farbe zwei Gehege umrahmt – die Kinder wüssten also nicht, welches der beiden „ihr“ (Start-)Gehege sei (9,9%).

Die Spielkarten werden als zu dünn und instabil für mehrere Durchführungsdurchgänge beschrieben (14,3%), da die Kinder damit spielen und sie dabei knicken.

Mehrfach ist als Problem genannt worden, dass der Plan nicht auf die Größe der Kinder abgestimmt sei (7,0%). Ihre Arme seien oft zu kurz, um selbstständig ihre Figur setzen zu können. Die Bilder seien zum Beschreiben zu weit weg bzw. wenn ein anderes Kind erzählt, müssten mehrere Kinder „über Kopf“ schauen.
Darüber hinaus ist dreimal genannt worden, dass Tiger und Löwen für Kinder in dem Alter nur schwer auseinander zu halten seien. Auch der Delfin erscheint drei der Befragten ungünstig, weil dieses Tier den Kindern nicht immer bekannt gewesen sei.

„Sind Sie selbst mit dem Material klar gekommen? – Wenn nein, warum nicht?“
Auch hier war die häufigste Nennung für Probleme der mehrteilige Spielplan (18,3%), 81,7% der Befragten sind gut mit dem Material klar gekommen.

„Haben Sie Ideen, wie man das Material noch verbessern kann?“
Am häufigsten wird ein einteiliger Spielplan gefordert (über 63,4%). 18,3% der Befragten schlagen vor, die „Spielregeln“ zu ändern (z. B. dass die Kinder nur ein „Tier“ sammeln). Ebenso häufig wird genannt, dass die Anordnung der Spielfiguren und des Weges weniger „verwirrend“ sein sollte, damit die Kinder selbstständiger agieren könnten und ihr eigenes Feld nicht so lange suchen müssten.
Allerdings erscheint einigen der Einsatz der Püppchen ganz überflüssig, da diese keine Funktion haben (8,5%).
Damit die Kind-Gehege-Zuordnung eindeutig sei, wünschten sich mehrere der Befragten, dass die Umrandung des Spielplans deutlicher ist (9,9%) und zu dem Gehege passt oder über Eck verläuft (zweimal genannt).
Einige Male wird vorgeschlagen, dass die Spielkarten stabiler sein und beschichtet oder laminiert werden sollten (14,3%). Auch wird vorgeschlagen, die Karten, die die Kinder ja nicht lesen könnten, bei der Begleiter/in zu belassen und durch Gewinnsteine zu ersetzen (3 Nennungen).
Auch wünschen sich einige der Befragten, dass bestimmte Wörter, die in den Aufgaben verwendet werden, ausgetauscht werden (die Namen sollten frei austauschbar sein, damit sie den Kindern bekannt sind; Gehege; Zooeingang; Zoo) (12,7%)
Manche der Erzieher/innen und Lehrer/innen merken an, dass die Bilder zum Ausmalen zu klein und zu aufwändig seien (5 Nennungen). (Die Bilder mussten vor der Durchführung in ausreichender Anzahl kopiert werden).

Zur Situation:

„Haben Sie die Situation als natürlich empfunden? – Wenn nein, warum nicht?“

Zwei Drittel der Befragten empfanden die Situation als nicht natürlich, 18,3% geben als Grund dafür an, dass der Test- bzw. Prüfungscharakter der Situation offensichtlich gewesen sei. In 66,2% der Fälle wird das Verhalten der Erzieherin als fremd und gekünstelt beschrieben, auch, weil sie die Anweisungen der Aufgaben-Karten nicht umformulieren könne (4 Nennungen). 45,1% der Befragten geben an, dass die fremde Person sich störend auf die Situation ausgewirkt hätte. Für 21,1% trug zur Verstärkung der „fremden“ Situation bei, dass die Durchführung in einem anderen Raum stattfand.

Vielfach wird hier beklagt, dass die üblichen Unterstützungsmechanismen bei der Durchführung fehlten (loben, ermutigen, wiederholen, u. a.) und ein individuelles Eingehen auf das Kind nicht möglich gewesen sei (19,7%). Allerdings wird von Seite der Beobachter/in mehrfach betont, dass dies von Person zu Person sehr unterschiedlich gewesen sei (vier Nennungen). Auch Erzieher/innen berichten, dass sie die Situation nur zu Beginn als künstlich empfunden hätten und mit der Zeit „lockerer“ geworden seien (insgesamt sechs Nennungen, z. B.: „Je öfter man spielte, desto sicherer wurde man.“ – „Der erste Durchgang war „krampfig“ und unnatürlich. Der zweite Durchgang war dagegen natürlicher und selbstverständlicher.“) Auch hätten sie selbst etwas dazu beigetragen, dass die Situation „natürlicher“ gewesen sei (gute Kooperation mit der Lehrerin, eigenes Verhalten überdenken).

„Haben Sie die Situation als kindgemäß empfunden? – Wenn nein, warum nicht?“

Knapp zwei Drittel der Begleiter/innen und Protokollant/innen kennzeichneten die Situation als „nicht kindgemäß“ (64,8%). Ein Drittel sprach sich ausdrücklich dafür aus, dass sie die Situation als kindgemäß erlebt hatten (33,8%).

Als größtes Problem wird von 29,6% der Evaluator/innen genannt, dass die Durchführungszeit mit über 40 Min. zu lang gedauert hätte. Auch fehlte der Spielcharakter (11,3%); das „Spiel“ sei langweilig, monoton und die Wartezeiten zwischen den Aktionen für das einzelne Kind zu lang (16,9%). Nur eine der Befragten hielt das „Redeverbot“ für problematisch, weil es dem üblichen Vorgehen in der Tageseinrichtung, mehrere Kinder ins Gespräch zu ziehen, entgegenstehe. Problematisch erscheint aber, dass die Erzieherin das Geschehen stark lenken soll und die Kinder dadurch über lange Zeit passiv und unselbstständig sein müssten (8,5%). 16,9% geben an, dass die Aufgaben für das Alter der Kinder zu schwierig seien. Hier wird besonders die Bilderzählung kriti-

siert, da die Bilder zu wenig Erzählanlässe böten und die vorgegebenen Impulse oft nur Einwort-Äußerungen hervorriefen (15,5%). Schwierig erschien in diesem Zusammenhang auch das Protokollieren, da die Kriterien für die Einschätzung der Bilderzählung für zu zahlreich und nicht eindeutig genug gehalten wurden (14,1%).

„Haben Sie die Situation als praxisgerecht empfunden? – Wenn nein, warum nicht?“

50,7% der Befragten halten die Situation für nicht praxisgerecht, 35,2% für praxisgerecht und 14,1% beantworten diese Frage nicht oder mit „teils-teils“. Die Gründe, die für diese Einschätzung genannt wurden, decken sich mit denen, die genannt wurden, wenn die Situation als nicht-kindgemäß eingeschätzt wird.

„Traten in der Situation besondere Probleme auf? – Wenn ja, welche?“

71,8% der Erzieher/innen und Lehrer/nnen gaben an, dass es während der Durchführung zu besonderen Problemen gekommen ist. Am häufigsten wird das Problem genannt, dass Kinder ihre Mitarbeit „verweigert“ haben (60,6%).

Viele der Kinder würden im Spielverlauf immer unkonzentrierter und/oder überbrückten Wartezeiten, indem sie mit den Karten spielten, darauf herumkauten, alberten oder störten (22,5%). Zwölfmal (16,9%) wird angesprochen, dass die Kinder Probleme hatten, die (Quatsch-)Sätze nachzusprechen. Auch viermal wird allerdings betont, dass die Kinder diese Aufgabe lustig fanden und gut damit zurechtgekommen sind. Fünfmal wird festgestellt, dass das gleichzeitige Protokollieren der Bilderzählung auch für einen externen Beobachter schwierig war (7,0%).

„Haben Sie Ideen, wie man die Situation noch verbessern könnte?“

Viermal wird ausdrücklich genannt, dass die (Quatsch-)Sätze „lustiger“ formuliert werden sollten. Hier werden auch konkrete Vorschläge gemacht (inhaltlicher Bezug zum Zoo, zum Kindergarten). Mehrfach werden Hinweise gegeben, wie der Spielplan überarbeitet werden könne.

Zur Durchführungsanleitung

„Haben Sie sich durch die Durchführungsanleitung gut vorbereitet gefühlt? – Wenn nein, warum nicht?“

Zwei Drittel der Evaluator/innen fühlten sich durch die Durchführungsanleitung gut vorbereitet. Diejenigen, die sich nicht ausreichend vorbereitet fühlten, gaben als Grund fast

ausschließlich Zeitmangel und die besonderen Umstände der Pilotstudie an (besonders kurze Vorbereitungszeit, Material kurzfristig nur an die Schulen geliefert). Besonders die Erzieher/innen hätten sich gerne länger mit dem Verfahren und der Durchführungsanleitung beschäftigt – einige (9,9%) hatten im Vorfeld keine Möglichkeit, die ganze Anleitung zu lesen.

„Fanden Sie die Anleitung verständlich? – Wenn nein, warum nicht?

76,1% der Befragten sind gut mit der Anleitung zurechtgekommen. 9,9% hatten keine Gelegenheit, den Text zu beurteilen. 14,1% haben die Anleitung nicht oder nur teilweise verständlich gefunden. Als Gründe werden angegeben: die Anordnung der Karten ist nicht sofort ersichtlich gewesen (vier Nennungen); die Anleitung ist zu langatmig und kompliziert, die Anleitung ist zu kurz, müsste länger und ausführlicher sein; die Anleitung ist zu abstrakt (jeweils eine Nennung).

„Haben Sie in der Anleitung Fehler entdeckt? – Wenn ja, welche?“

Vier Mal werden konkrete Fehler in der Anleitung genannt und zwar folgende: auf der letzten Delfin-Karte ist der letzte Handlungsschritt nicht eindeutig (3 Nennungen). Einmal scheint es sich um ein Missverständnis zu handeln, da kritisiert wird, dass in der Anleitung nur die Delfin-Karten beschrieben würden.

„Hat Ihnen in der Anleitung etwas gefehlt? – Wenn ja, was?“

71,8% der Befragten fanden keine Lücken in der Anleitung. In neun Fällen fehlte den Begleiter/innen oder Protokollant/innen etwas in dem Text (12,7%). Sie wünschten sich eine ausführlichere Beschreibung der Sortierung der Karten. Darüber hinaus wünschten sie sich Vorschläge, wie man als Erzieherin mit „unerwarteten“ Situationen umgehen könne und Hinweise darauf, welche Variationsmöglichkeiten erlaubt seien. 15,5% konnten diese Fragen nicht beantworten, weil sie die Anleitung nur kurz oder gar nicht vorliegen hatten.

„Haben Sie Ideen, wie man die Durchführungsanleitung noch verbessern könnte?“

Hier nannten die Befragten die Punkte, die sie oben als fehlend oder unverständlich kritisiert hatten.

An verschiedenen Stellen des Bogens bringen 11,3% der Befragten zum Ausdruck, dass sie bezweifelnd, dass ein/dieser Test das Sprachvermögen der Kinder zeigen kann.

Die vielfältigen Hinweise und Vorschläge haben wir gründlich bedacht und das Material daraufhin entscheidend überarbeitet. Natürlich haben wir alle Aspekte ausgeschlossen, deren Umsetzung eine Minderung der Messgüte des Tests mit sich gebracht hätte.

Im Folgenden soll nun noch auf die Praxisevaluation von BiP eingegangen werden.

Stufe 2 „Besuch im Pfiffikus-Haus (BiP)“

Die Praxisevaluation des Testmaterials von BiP fand im Rahmen der Pilotierung von Stufe 2 in Köln statt. Der Evaluationsbogen BiP gleicht dem Evaluationsbogen BiZ. Er umfasst vier Seiten und ist ebenfalls im Anhang abgebildet.

Die Bögen wurden an 39 Kindertageseinrichtungen sowie die ihnen zugeordneten Grundschulen versandt. Zurückgesandt wurden 52 Bögen. Diese waren durch 36 Erzieher/innen und 16 Lehrer/innen ausgefüllt worden. Die Antworten auf die Fragen werden nachfolgend aufgelistet.

1. Hatten die Kinder Probleme mit dem thematischen Rahmen „Besuch im Pfiffikus-Haus“?

N = 45	
Ja	**Nein**
8,9%	**91%**

Variable	Anteil[55]
Wenn ja, welche?	
Begriff „Pfiffikus“ musste erläutert werden	11,1%
Familie Pfiffikus als Rahmen überflüssig	4,4%

2. Sind die Kinder mit dem Material gut klar gekommen?

N = 45	
Ja	**Nein**
75,6%	**24,4%**

[55] Mehrfachnennungen waren hier und bei den folgenden Fragen möglich.

Variable	Anteil
Wenn nein, warum nicht?	
Bezug zwischen Spielplan und Aufgaben fehlte / war künstlich	48,9%
Spielplan lenkt Kinder von den Testaufgaben ab, ist überflüssig	46,7%
Störbilder „Spielzeug“ problematisch	24,4%
Abbildungen bei der Pluralbildung verleiteten Kinder eher zum Zählen	17,8%
Begriff „Memory“ verwirrend	11,1%
Spielplan zu groß	8,9%
Flur des Pfiffikus-Hauses gleicht eher einer Rumpelkammer	8,9%
Belohnungsbild überflüssig (zu viele Details, enttäuschend)	6,7%
Darstellungen insgesamt zu „comichaft“/ kitschig	6,7%

3. Wurde der Zeitrahmen von ca. 30 - 35 Min. pro Durchführung überschritten?

N = 50	
Ja	**Nein**
54,0%	**46,0%**

Variable	Anteil
Wenn ja, wie viel Zeit haben Sie für die Durchführung benötigt?	
35 Min. - 40 Min.	2,0%
40 Min. - 45 Min.	42,0%
mehr als 45 Min.	8,0%

4. Haben Sie die Situation als natürlich empfunden?

N = 51	
Ja	**Nein**
25,5%	**74,5%**

Variable	Anteil
Wenn nein, warum nicht?	
fremde Person beeinflusste die Situation	37,3%
Durchführung fand in einem fremden Raum statt	23,5%
Test-/ Prüfungscharakter war offensichtlich	19,6%
Blickkontakt und Zuwendung wurden durch die Protokollierung beeinträchtigt	11,8%
übliche Unterstützungsmöglichkeiten (wie Loben, Ermutigen, Wiederholen) fehlten	5,9%

4**. Haben Sie die Situation als kindgemäß empfunden?**

N = 49	
Ja	**Nein**
44,9%	**55,1%**

Variable	Anteil
Wenn nein, warum nicht?	
zu viele Aufgaben, zu lange Durchführungs-zeit	38,8%
freies Erzählen wurde gestoppt, auf Äuße-rungen zum Spielplan konnte fast gar nicht eingegangen werden	28,6%
Anweisungen zu lang, künstlich oder um-ständlich formuliert	20,4%
Aufgaben- und Hilfestellung bei der BE ermu-tigt Kinder nicht zum Erzählen (viele Einwort- oder Fragmentsätze werden evoziert)	16,3%
Pluralbildung von Unsinns-Wörtern zu schwierig	14,3%
Nachsprech-Aufgaben entsprechen Kindern nicht	12,2%
Kindern sollte ein Test nicht als „Spiel“ vorge-stellt werden	4,1%

5. Haben Sie die Situation als praxisgerecht empfunden?

N = 45	
Ja	**Nein**
55,6%	**44,4%**

Variable	Anteil
Wenn ja, welche?	
Einzelsituation unüblich	22,2%
nicht situationsorientiert/ der üblichen Förde-rung/ dem Bildungsauftrag entsprechend	17,8%
Methode unüblich/ Testsituationen unbekannt	4,4%

6. Traten in der Situation besondere Probleme auf?

N = 40	
Ja	**Nein**
32,5%	**67,5%**

Variable	Anteil
Wenn ja, welche?*	
Kinder verweigerten Mitarbeit	17,5%
Konzentrationsmangel, Ermüdung	15,0%

* außer materialbedingte Probleme, vgl. dazu 2.

7. Haben Sie sich durch die Durchführungsanleitung gut vorbereitet gefühlt?

N = 50	
Ja	**Nein**
88,2%	**11,8%**

Variable	Anteil
Wenn nein, warum nicht?	
zu lang	5,9%
Zeitmangel (Material zu spät erhalten)	3,9%
zu kompliziert	2,0%

8. Fanden Sie die Handanweisung verständlich?

N = 50	
Ja	**Nein**
100%	**0%**

9. Haben Sie Fehler in der Handanweisung entdeckt?

N = 50	
Ja	**Nein**
10,0%	**90,0%**

10. Hat Ihnen in der Handanweisung etwas gefehlt?

N = 49	
Ja	**Nein**
12,2%	**87,8%**

Variable	Anteil
Wenn ja, was?	
Hintergründe zu bestimmten Aufgaben (Pluralbildung; Nachsprechaufgaben)	4,1%
zusammenfassende Darstellung	4,1%
Zeitangabe auf Abdeckplatten	2,0%
Hinweise zu Bestätigungs- oder Motivationsmöglichkeiten	2,0%

11. Hatten Sie Probleme mit der Protokollierung?

N = 51	
Ja	**Nein**
54,9%	**45,1%**

Variable	Anteil
Wenn ja, welche?	
zusätzliche/r Protokollant/in notwendig	31,4%
erschwerte Blickkontakt/ Zuwendung zum Kind	31,4%
gleichzeitiges Protokollieren schwierig, Nacharbeit war nötig	13,2%

Auch diese Rückmeldungen haben wir sorgsam bedacht. Dabei haben wir die Antwortmuster so gelesen, dass ein Teil der Erzieher/innen und Lehrer/innen dringend Unterstützung und Qualifizierung benötigt, um die mit der Durchführung von Delfin 4 einhergehenden Herausforderungen gut bewältigen bzw. handhaben zu können[56]*. So gilt es, die Bewusstheit von Erzieher/innen und Lehrer/innen für die Funktionen und Anforderungen von Delfin 4 zu schärfen. Das setzt voraus, sie besser diagnostisch so zu qualifizieren, so dass sie in die Lage versetzt werden, Delfin 4 richtig einzuschätzen sowie kompetent damit auch gewinnbringend einzusetzen. Schließlich kann Delfin 4 nur dann zur Verbesserung der Praxis beitragen, wenn das Verfahren professionell angewendet wird.*

Wir selbst konnten dem innerhalb des uns vorgegebenen Rahmens zwar nicht nachhaltig genug, aber immerhin insoweit gerecht werden, als wir die Teilnehmer/innen der verschiedenen Studien zur Sicherung der Qualität von Delfin 4: BiZ und BiP durch Qualifizierungen vorbereitet haben. Diese Maßnahmen wurden im Rahmen von zwei Pilotierungen, zwei Reliabilitätsstudien sowie zwei Normierungsstudien durchgeführt und folg-

[56] Diese Einschätzung haben wir an die beteiligten Ministerien weitergegeben.

ten alle mehr oder minder dem folgenden Schema. Zunächst wurde von unserer Seite über die sprachentwicklungs-theoretischen sowie teststatistischen Hintergründe von Delfin 4 informiert. Daran anknüpfend wurde die Konstruktion des Verfahrens erläutert. Dabei lag der Schwerpunkt darauf, das Konstruktionsrational sowie die Funktion jedes einzelnen Untertests bzw. jeder Testaufgabenform zu erfassen. Der Schwerpunkt der Veranstaltungen lag aber auf der konkreten Durchführung der Untertests bzw. Erprobung der Testaufgaben in Arbeitsgruppen. Diese wurden von uns betreut, so dass die während der Durchführung bzw. der Erprobung auftretenden Fragen und Probleme diskutiert sowie erläutert werden konnten. Die insgesamt sechs Qualifizierungen waren mindestens halb-, bei Bedarf auch ganztägig.

Im Anschluss an die Praxiserprobungen konnte die Endform von Delfin 4 fertig gestellt werden. Für BiZ und BiP ergaben sich aus den Befragungsergebnissen folgende Veränderungen:

Bei Delfin 4 - Stufe 1 wurden aufgrund der Ergebnisse der Praxisevaluation hauptsächlich grafische Umgestaltungen vorgenommen sowie die Durchführung vereinfacht. Bei der Entscheidung, wie die Resultate dieser Evaluation am ehesten eingelöst werden können, unterstützte uns beratend ein Team von 12 Praktikerinnen aus dem Elementar- und Primarbereich.

Bei Delfin 4 – Stufe 2 haben wir eine Optimierung der Durchführungsökonomie vorgenommen. Das beinhaltete – neben der Vereinfachung bzw. Komprimierung der Protokollbogen – vor allem eine Kürzung der Aufgaben.

6. Empirische Aufgabenanalysen zu den Testendformen BiZ/BiP

Um sicher zu sein, dass die Messgüte von BiZ und BiP durch die Veränderungen nicht beeinträchtigt worden ist, haben wir 2008 erneut empirische Aufgabenanalysen zu den Endformen von Stufe 1 und 2 durchgeführt. Die zugrundeliegenden Berechnungen wurden a) mittels SPSS (Klassische Testtheorie KTT) und b) parallel dazu mittels ConQuest (Item-Response-Theorie IRT) durchgeführt. Über die Ergebnisse wird im Folgenden kurz berichtet.

Aufgabengüte der Testendform BiZ

Grundlage der Berechnungen waren die Ergebnisse von 1.552 Kindern (Normierungsstichprobe 2007). Die Stichprobe ist nach Sozialindex geschichtet und setzt sich aus 49,9 Prozent Mädchen und 50,1 Prozent Jungen zusammen. 85,4 Prozent der Kinder ist einsprachig. Der Hauptanteil von 73,5 Prozent der Kinder wächst mit deutscher und 11,9 Prozent mit einer anderen Muttersprache auf. 14,6 Prozent der Kinder wachsen bilingual auf. 13,7 Prozent von ihnen sprechen in den Familien deutsch und ein bis zwei weitere Sprache(n) während 0,7 Prozent zweisprachig ohne deutsch aufwachsen. Die Altersspanne liegt zwischen 42 und 62 Monaten. Im Durchschnitt sind die Kinder 4 Jahre und 1 Monat alt (SD: 4).

Im Folgenden werden die Ergebnisse der erneuten empirischen Aufgabenanalysen zur Erstversion von Delfin berichtet. Die Analysen basieren – wo möglich – sowohl auf Methoden der Klassischen Testtheorie (KTT) sowie der Item-Response-Theorie (IRT). Die dokumentierten Befunde bestätigen die bereits bei der Pilotierung ermittelte ausgeprägte Messgüte des Instruments.

Als erstes werden die Resultate der Berechnungen zu BiZ dokumentiert.

Unter-test	Schwierigkeit			Trennschärfe		
	Min	Max	MW	Min	Max	MW
HA	62,2	86,0	76,3	,527	,635	,597
KN	44,2	84,5	63,3	,425	,556	,524
SN	59,7	67,0	57,2	,751	,822	,784
BB	18,8	74,4	42,5	,440	,636	,537

Tabelle 42: Aufgabenstatistik zu den Untertests von BiZ

Ergebnisse zum Untertest HA:

		KTT	
Nr.	Aufgabe	Schwierigkeit	Trennschärfe
1	HA 1	85,7	,527
2	HA 2	77,5	,584
3	HA 3	72,5	,635
4	HA 4	83,5	,635
5	HA 5	62,2	,605
Durchschnitt		76,3	,597

Tabelle 43: Aufgabenstatistik HA (BiZ)

Ergebnisse zum Untertest KN:

		KTT		IRT		
Nr.	Aufgabe	Schwierigkeit	Trennschärfe	MNSQ	t	Trennschärfe
1	KN 1	84,5	,474	0,95	-0,8	,59
2	KN 2	74,9	,550	0,95	-1,2	,67
3	KN 3	67,4	,556	0,97	0,9	,69
4	KN 4	69,5	,548	0,97	-0,7	,68
5	KN 5	66,3	,533	1,00	0,0	,67
6	KN 6	48,7	,461	1,05	1,7	,62
7	KN 7	51,3	,478	1,04	1,2	,63
8	KN 8	44,2	,425	1,08	2,5	,59
Durchschnitt		63,3	,503			,64

Tabelle 44: Aufgabenstatistik KN (BiZ)

Ergebnisse zum Untertest SN:

		KTT	
Nr.	**Aufgabe**	**Schwierigkeit**	**Trennschärfe**
1	SN 1	67,0	,778
2	SN 2	55,9	,784
3	SN 3	46,4	,751
4	SN 4	59,7	,822
Durchschnitt		57,2	,784

Tabelle 45: Aufgabenstatistik SN (BiZ)

Ergebnisse zum Untertest BE:

		KTT		IRT		
Nr.	**Aufgabe**	**Schwierigkeit**	**Trennschärfe**	**MNSQ**	**t**	**Trennschärfe**
1	Erzählt spontan	66,1	,532	1,10	2,6	,63
2	Fasst das Wesentliche gleich am Anfang zusammen	38,6	,548	1,05	1,4	,64
3	Identifiziert mindestens drei Akteure	61,9	,509	1,13	3,3	,61
4	Beschreibt mindestens ein Ereignis	74,4	,558	0,87	-3,0	,64
5	Verknüpft mehrere Akteure und/oder Ereignisse sinnvoll	34,1	,635	0,87	-4,1	,71
6	Markiert einen Höhepunkt	24,3	,440	1,14	3,5	,54
7	Stellt logische und/oder zeitliche Verknüpfungen her	24,0	,548	0,94	-1,4	,63
8	Macht Ortsangaben	37,7	,501	1,13	3,6	,60
9	Kennzeichnet näher	18,8	,448	1,04	0,8	,54
10	Verwendet mindestens Dreiwortäußerungen	70,3	,568	0,89	-2,7	,65
11	Reiht mindestens zwei Sätze aneinander	39,9	,636	0,90	-3,2	,72
12	Verwendet Haupt-, Nebensatzkonstruktionen	20,5	,518	0,95	-1,3	,60
Durchschnitt		42,5	,537			,62

Tabelle 46: Aufgabenstatistik BB (BiZ)

Abschließend werden noch die bis dato errechneten Korrelationskoeffizienten zu BiZ berichtet.

	N	Reliabilitätskoeffizienten*	
Pilotierung	678	r_{GES} = .85	r_{HA} = .76; r_{KN} = .81; r_{SN} = .87; r_{BB} = .78
Normierung 2007	1.552	r_{GES} = .86	r_{HA} = .76; r_{KN} = .79; r_{SN} = .89; r_{BB} = .86

Tabelle 47: Reliabilitätsstatistik BiZ

Aufgabengüte der Testendform BiP

Um die Befunde der empirischen Aufgabenanalyse der Endform der zweiten Stufe vorwegzunehmen: Die Resultate ließen es gerechtfertigt erscheinen, BiP im Umfang von 20 Aufgaben zu reduzieren, ohne dass die Qualität des Instruments dadurch leidet. Aus Gründen der Testökonomie haben wir das Verfahren deshalb verkürzt. Die Endversion setzt sich nunmehr aus 62 Test- zzgl. 10 Übungsaufgaben aus sieben Untertests zusammen und ist zeitökonomisch und präzise messend. Die folgende Tabelle gewährt einen Überblick über die Aufgabenreduktion je Untertest.

Untertest	Erstversion BiP 2007	Endversion BiP 2008	Selektion
WV	23	18	5
BK	3	3	0
KN	9	7	2
SN	9	6	3
PB	11	6	5
WP	23	18	5
BE	4	4	0
Gesamt	**82**	**62**	**20**

Tabelle 48: Aufgabenreduktion Stufe 2 BiP

Die Berechnungen basierten auf einer Stichprobe von 2.278 Kinder. Diese setzten sich aus 49 Prozent Mädchen und 51 Prozent Jungen zusammen. 70,1 Prozent von ihnen wachsen monolingual in ihren Familien auf, davon sprechen 56,6 Prozent deutsch und 13,5 Prozent eine andere Muttersprache. 29,8 Prozent der Kinder wachsen bilingual auf. Von ihnen sprechen 28,6 Prozent deutsch und eine weitere Sprache und 1,2 Prozent andere Sprachen ohne deutsch. Gemäß dem Ergebnis von Delfin 4 - Stufe 1 nahmen 50,2 Prozent der Kinder freiwillig an den Erhebungen zur Stufe 2 teil[57].

In den nachfolgenden Übersichten werden zunächst die bei dieser Untersuchung ermittelten Reliabilitätskoeffizienten, daran anschließend die Aufgaben- bzw. Untertestkennwerte aufgelistet.

[57] Bei 40,6 von ihnen war das Ergebnis im grünen und bei 9,6 Prozent im roten Bereich. Für 49,8 Prozent der Kinder war die Teilnahme an der zweiten Stufe obligatorisch. Ihr Ergebnis lag im gelben Bereich.

Ergebnisse der Reliabilitätsstudien:

		Reliabilitätskoeffizienten*
N = 2.278	r_{GES} = .93	r_{WV} =.86; r_{BK} = .73; r_{KN} = .71; r_{SN} = .93; r_{PB} =.78; r_{WP} = .89; r_{BE} = .89

Tabelle 49: Reliabilitätskennwerte BiP

Ergebnisse zu den Aufgabenkennwerten der Untertests:

Unter-test	Schwierigkeit			Trennschärfe		
	Min	Max	MW	Min	Max	MW
WV	34,0	89,6	68,1	,337	,596	,466
BK	58,8	73,8	67,5	,501	,584	,552
KN	65,7	87,8	73,6	,385	,456	,421
SN	71,3	79,5	75,3	,775	,853	,808
PB	25,0	69,5	38,9	,431	,656	,537
WP	17,0	88,5	67,1	,364	,697	,522
BE	29,5	35,8	32,3	,715	,802	,768

Tabelle 50: Kennwerte nach KTT der Endversion BiP

Ergebnisse zum Untertest WV:

		KTT		IRT		
Nr.	**Item**	**Schwierigkeit**	**Trennschärfe**	**MNSQ**	**T**	**Trennschärfe**
1	Fledermaus	80,0	,503	0,95	-1,4	,56
2	Kapuze	62,3	,406	1,13	4,9	,50
3	Wut	60,1	,404	1,13	5,1	,51
4	Streit	74,2	,498	0,99	-0,4	,57
5	Stängel	34,0	,419	1,01	0,4	,52
6	Küken	78,9	,470	0,98	-0,7	,55
7	freuen	75,0	,500	0,96	-1,2	,57
8	öffnen	72,5	,484	0,99	-0,3	,57
9	warten	72,5	,337	1,18	6,2	,44
10	ziehen	68,3	,452	1,04	1,6	,55
11	schälen	64,4	,461	1,05	2,0	,55
12	tanken	79,4	,494	0,96	-1,2	,56
13	nass	82,9	,381	1,06	1,7	,45
14	eckig	60,6	,588	0,88	-5,0	,66
15	krumm	43,9	,464	0,99	-0,4	,56
16	auf	89,6	,435	0,93	-1,5	,49
17	vor	66,0	,596	0,88	-5,0	,66
18	hinter	61,7	,499	1,01	0,4	,58
Durchschnitt		68,1	,466	,55		

Tabelle 51: Kennwerte des Untertests WV (KTT und IRT)

Ergebnisse zum Untertest BK:

		KTT		IRT		
Nr.	**Item**	**Schwierigkeit**	**Trennschärfe**	**MNSQ**	**T**	**Trennschärfe**
1	Spielzeug	58,8	,501	1,08	2,5	,79
2	Kleidung	69,8	,567	1,00	0,1	,81
3	Obst	73,8	,589	0,97	-0,9	,82
Durchschnitt		67,5	,552	,81		

Tabelle 52: Kennwerte des Untertests BK (KTT und IRT)

Ergebnisse zum Untertest WP:

		KTT		IRT		
Nr.	**Item**	**Schwierigkeit**	**Trennschärfe**	**MNSQ**	**T**	**Trennschärfe**
1	Schrank	85,0	,573	0,89	-2,4	,56
2	Handschuh	88,5	,454	0,96	-0,8	,50
3	Schildkröte	79,5	,634	0,86	-3,8	,51
5	Hubschrauber	65,0	,450	1,17	5,4	,57
6	Kleiderbügel	17,0	,364	0,97	-0,8	,52
7	Marienkäfer	75,5	,547	1,03	0,8	,55
9	bauen	75,0	,468	1,16	4,4	,57
10	schieben	78,5	,522	1,01	0,4	,57
11	(Seil-)springen	81,0	,391	1,24	5,6	,44
13	reiten	67,0	,697	0,79	-7,1	,55
14	pflücken	37,0	,483	1,02	0,6	,55
15	schwimmen	84,5	,398	1,10	2,2	,56
16	dick	64,5	,550	1,05	1,6	,45
17	dunkel	84,5	,549	0,88	-2,6	,66
20	alt	48,0	,511	1,05	1,7	,56
21	in	67,5	,587	1,01	0,2	,49
22	neben	52,5	,601	0,90	-3,8	,66
23	hinter	57,5	,619	0,88	-4,4	,58
Durchschnitt		67,1	,522			,61

Tabelle 53: Kennwerte des Untertests WP (KTT und IRT)

Ergebnisse zum Untertest SN:

Nr.	Item	Schwierigkeit	Trennschärfe
1	Die Katze wird von Lena gefüttert.	79,5	,805
2	Heute rennen die Kinder schnell zur Schule.	76,7	,823
3	Ayla kann ihren Schlüssel nicht finden.	78,8	,775
4	Tom nimmt dem Hund den Knochen weg.	73,0	,791
5	Das lustige Eis tanzt einen Baum.	72,3	,800
6	Wenn die Hose singt, klettert sie über die Straße.	71,3	,853
	Durchschnitt	75,3	,808

Tabelle 54: Kennwerte des Untertests SN (KTT und IRT)

Ergebnisse zum Untertest WP:

		KTT		IRT		
Nr.	Item	Schwierigkeit	Trennschärfe	MNSQ	T	Trennschärfe
1	Schaf	60,5	,532	1,07	2,2	,71
2	Auto	69,5	,572	0,99	-0,2	,74
3	Bett	39,0	,598	0,92	-3,1	,73
4	Loch	52,0	,656	0,86	-4,7	,80
5	Dopf	25,0	,431	1,03	0,9	,59
6	Mate	41,5	,431	1,21	6,4	,64
	Durchschnitt	38,9	,537			,70

Tabelle 55: Kennwerte des Untertests PB (KTT und IRT)

Ergebnisse zum Untertest KN:

		KTT		IRT		
Nr.	Item	Schwierigkeit	Trennschärfe	MNSQ	T	Trennschärfe
1	fegitt	87,8	,403	0,96	-0,8	,56
2	lupori	74,7	,402	1,04	1,3	,61
3	robiwamm	65,7	,385	1,07	2,4	,61
4	jakedu	76,5	,446	0,98	-0,7	,63
5	kutabodi	73,6	,435	1,00	0,1	,63
6	hamifuko	67,3	,419	1,04	1,3	,62
7	gotumalik	69,3	,456	0,98	-0,6	,65
	Durchschnitt	73,6	,421			,62

Tabelle 56: Kennwerte des Untertests KN (KTT und IRT)

Ergebnisse zum Untertest BE:

Nr.	Item	Schwierigkeit	Trennschärfe
1	BE1	29,5	,715
2	BE2	31,1	,802
3	BE3	32,9	,784
4	BE4	35,8	,770
	Durchschnitt	32,3	,768

Tabelle 57: Kennwerte des Untertests BE (KTT)[58]

[58] Aufgrund der spezifischen Aufgabenstruktur war es nicht möglich, IRT-Analysen in ConQuest durchzuführen.

7. Gütekriterien von BiZ und BiP

Anhand der Endversion von Delfin 4 wurde (zum Teil erneut) geprüft, ob bei Delfin 4 die Hauptgütekriterien (Objektivität, Reliabilität, Validität) gewährleistet sind.

Objektivität

Mit Objektivität bezeichnet man das Ausmaß, in dem die Ergebnisse eines Tests unabhängig von der Person sind, die den Test durchführt. Dabei wird unterschieden zwischen: Durchführungsobjektivität, Auswertungsobjektivität und Interpretationsobjektivität (Lienert & Raatz 1998[6]).

Bei Delfin 4 ist die Durchführungsobjektivität durch ein hohes Maß an Standardi-sierung gewährleistet. Die Testdurchführung ist klar und strikt strukturiert, die Anleitungen für die Testleiter/innen sind völlig eindeutig und sehr genau. Außerdem wurde ein Video ins Internet gestellt, das jeder Erzieher/in bzw. Lehrer/in ein Modell bietet, wie der Test richtig durchgeführt wird.

Die Auswertungsobjektivität
Die Auswertung beider Stufen erfolgt mit Hilfe eines Protokollbogens. Die dort festgehaltenen Punktwerte werden dann auf einen Ergebnisbogen übertragen. Wir haben die Auswertungsobjektivität zweifach empirisch geprüft.
Einmal wurden neun Kinder mit Delfin 4 Stufe 2 (BiP) getestet. Die Testdurchführung erfolgte durch eine Wissenschaftliche Mitarbeiterin des Projekts. Dabei anwesend waren insgesamt fünf im Umgang mit Delfin 4 vertraute Erzieher/innen und Lehrer/innen. Ihre Aufgabe bestand darin, die Testdurchführung völlig unabhängig voneinander mit zu protokollieren und anschließend die Punktwerte zu bestimmen. Ein Vergleich der Resultate der fünf Personen ergab eine mittlere Übereinstimmung von 92 Prozent. Die geringste Übereinstimmung bestand beim Untertest „Bilderzählung BE“ (82 Prozent), die größte beim Untertest „Wortproduktion WP“ (96 Prozent).
Darüber hinaus wurden 15 Erzieher/innen gebeten, verschriftete Bilderzählungen von zehn Kindern anhand des Auswertungsrasters einzuordnen. Diese Studie ergab eine mittlere Übereinstimmung von 84 Prozent.
Alles in allem ermöglichen also die Testmaterialien objektive Auswertungen.

Auch die Interpretationsobjektivität ist gegeben. Umrechnungstabellen ermöglichen es, die Leistungen eines Kindes, im Vergleich zu gleichaltrigen Kindern einzuordnen. Des weiteren liegen Prozentrang- und Standardnormen (T-Werte) vor, die es den Erzieher/innen und Lehrer/innen erlauben, die Sprachleistungen eines Kindes anhand von Vergleichswerten der Alterskohorte eindeutig einzuordnen bzw. zu bewerten. Außerdem wird in den Normen klar markiert, in welchen Fällen die gezeigte Leistung eine Risikoschwelle (cut-off-point) überschreitet, so dass eine zusätzliche Sprachförderung anzuraten ist. Bei der Berechnung dieser „cut-off-points“ wurden Konfidenzintervalle berücksichtigt.

Reliabilität

Mit Reliabilität oder Zuverlässigkeit bezeichnet man den Grad der Genauigkeit, mit dem ein Test die Leistung einer Person zu erfassen vermag. Dieser Grad der Genauigkeit wird durch Reliabilitätskoeffizienten näher bestimmt (Lienert & Raatz 1998[6]). Bei Delfin 4 wurden Interne Konsistenzschätzungen berechnet.

Wie die folgende Übersicht dokumentiert, unterstreichen die zusammengefassten Befunde der Reliabilitätsanalysen, dass Delfin 4 sowohl in Stufe 1, als auch in Stufe 2 zuverlässig misst.

N		Reliabilitätskoeffizienten	
		Screening	**Untertest**
BiZ	678 Pilotierung 2007	r_{GES} = .85	r_{HA} = .76; r_{KN} = .81; r_{SN} = .87; r_{BB} = .78
	1.552 Normierung 2007	r_{GES} = .86	r_{HA} = .75; r_{KN} = .81; r_{SN} = .89; r_{BB} = .78
BiP	542 Pilotierung 2007	r_{GES} = .95	r_{WV} =.88; r_{BK} = .70; r_{KN} =.77; r_{SN} = .93;r_{PB} = .83; r_{WP} = .91; r_{BE} = .92;
	1.703 Normierung 2007	r_{GES} = .95	r_{WV} =.85; r_{BK} = .75; r_{KN} =.80; r_{SN} = .94; r_{PB} = .79; r_{WP} = .91; r_{BE} = .92;
	2.278 Normierung 2008	r_{GES} = .93	r_{WV} = .86; r_{BK} = .73; r_{KN} = .71; r_{SN} = .94; r_{PB} = .78; r_{WP} = .89; r_{BE} = .85.

Tabelle 58: Reliabilitätskoeffizienten Delfin 4

Validität

Mit Validität oder Gültigkeit bezeichnet man das Ausmaß, in dem ein Test diejenige Fähigkeit zu messen vermag, die er zu messen vorgibt. Dabei unterscheidet man verschiedene Arten von Validität (Lienert & Raatz 1998[6]). Bei Delfin 4 wurden die Konstruktvalidität sowie die Konkurrente Validität oder Übereinstimmungsvalidität empirisch geprüft.

Konstruktvalidität

Bei der empirischen Prüfung der Konstruktvalidität wird die Faktorenanalyse (FA) genutzt um zu prüfen, ob bzw. wieweit die damit extrahierte Faktorenstruktur die Struktur des Konstruktionsrationals nachzubilden vermag. Das ist möglich, weil das Prinzip der FA darin besteht, die in den Daten enthaltenen Informationen auf eine möglichst geringe Zahl von hypothetischen Dimensionen oder Faktoren zu reduzieren.

Nachfolgend werden die Ergebnisse je einer FA auf Basis der Delfin 4-Testwerte berichtet. Als erstes werden die Befunde zu BiZ dargestellt. Diese Analysen wurden schon sehr früh im Verlauf des Projekts berechnet, basieren somit auf den Daten der Pilotierungsstudie.

Das BiZ zugrunde liegende Sprachkompetenzmodell ließ sich mit der FA ganz klar nachbilden. Es besteht also Konstruktvalidität.

Die extrahierten vier Faktoren klären 58,3% der Varianz auf. Die Kennzahl für die Stichprobeneignung (KMO: 0.90) macht deutlich, dass eine Stichprobenpassung der Faktorenstruktur gegeben ist.

KMO: .900; Bartlett: .000; Varianzaufklärung: 58,33%		Faktorladungen[59]			
Untertest	Aufgabe	1	2	3	4
KN	KN2	.684			
	KN3	.667			
	KN5	.664			
	KN4	.647			
	KN1	.637			
	KN8	.576			
	KN6	.569			
	KN7	.545			
HA	HA2		.779		
	HA3		.775		
	HA1		.761		
	HA4		.752		
	HA5		.665		
SN	SN4			.805	
	SN3			.789	
	SN2			.774	
	SN1			.720	
BB[60]	BB1				.870
	BB2				.812

Tabelle 59: Ergebnisse der Faktorenanalyse zu BiZ

Als nächstes werden die Befunde zu BiP ergänzt. Diese beruhen auf der Normierungsstichprobe.

[59] Koeffizienten unter .300 werden nicht angeführt.
[60] BB besteht aus einer Bildaufgabe. Die Reaktionen der Kinder auf das Bild werden anhand einer Kriterienliste charakterisiert. Dabei beziehen sich die meisten Kriterien auf „übersatzmäßge“ Leistungen. Die Liste beinhaltet aber auch einige Kriterien zu „satzmäßigen“ Leistungen. Wir haben das berücksichtigt, indem wir die Gesamtleistung der Kinder in „übersatzmäßige Leistungen“ (BB1) und „satzmäßige Leistungen“ (BB2) unterteilt haben.

KMO: .945; Bartlett: .000; Varianzaufklärung: 58,96%		Faktorladungen[61]					
Untertest	Aufgaben	1	2	3	4	5	6
SN	SN6	.757					
	SN2	.752					
	SN3	.752					
	SN4	.732					
	SN1	.728					
	SN5	.707					
	SN9	.701					
	SN7	.689					
	SN8	.645					
WS[62]	WVN		.625				
	WVA		.606				
	WVV		.589				
	WPA		.576				
	WPN	.394	.570				
	WPP	.436	.566				
	WPV	.433	.557				
	WVP		.553				
BE	Bild1			.883			
	Bild2			.844			
	Bild3			.819			
	Bild4			.812			
PB	PBU1				.704		
	PBB3				.635		
	PBB4				.625		
	PBB2				.552		
	PBU2				.553		
	PBB1				.505		
	PBU5				.498		
	PBB5				.422		
	PBU3				.414		
	PBU4				.411		
	PBB6				.399		
KN	KN4					.665	
	KN1					.626	
	KN6					.562	
	KN5					.559	
	KN7					.554	
	KN2					.541	
	KN3					.430	
	KN9					.427	
	KN8					.401	
BK	BK Obst						.743
	BK Obst						.731
	BK Obst						.718

Tabelle 60: Ergebnisse der Faktorenanalyse zu BiP

[61] Koeffizienten unter .300 werden nicht angeführt.
[62] Die Untertests WV und WP laden auf einem Faktor, der mit WS (= Wortschatz) überschrieben wird.

Wie die tabellarische Zusammenstellung zeigt, ließ sich auch das BiP zugrunde liegende Sprachkompetenzmodell eindeutig nachbilden[63]. Insofern ist auch bei BiP die Konstruktvalidität nachgewiesen.

Bei BiP klären die extrahierten sechs Faktoren insgesamt 59,0% der Varianz auf. Die Kennzahl für die Stichprobeneignung (KMO: 0.95) macht deutlich, dass eine Stichprobenpassung der Faktorenstruktur gegeben ist.

Übereinstimmungsvalidität

Die Übereinstimmungsvalidität kann durch das Ermitteln von empirischen Zusammenhängen mit relevanten Außenkriterien belegt werden (Moosbrugger & Kelava 2007, S. 156). Dabei gilt: Je enger der Zusammenhang ausfällt, desto besser ist die Validität belegt. Bei Delfin 4 wurde die Übereinstimmungsvalidität in zwei Richtungen geprüft:

Korrelationen mit anderen Sprachtests:

Wie bereits in den Ausführungen zur Pilotierung dargestellt, wurden parallel zu Delfin 4 andere Sprachtests eingesetzt, deren Messgüte – laut vorliegenden Expertisen (vgl. Fried 2008) – gewährleistet ist. Da es keinen allgemeinen Sprachtest gibt, der die ganze Breite der durch Delfin4 BiZ und BiP erfassten Bereiche akademischer Sprachkompetenz abdeckt, mussten die Analysen unter Verwendung unterschiedlicher Verfahren auf Untertestebene durchgeführt werden. Dabei konnten für die meisten Delfin 4-Untertests passende Untertests aus anderen Verfahren ausgespürt und verwendet werden. Das ist allerdings nicht der Fall, wo es um die Untertests BB und BE zur Erfassung der Erzählkompetenz geht. Hierzu gibt es noch kein passendes Verfahren.

Wie bereits in vorhergehenden Kapiteln berichtet, wurden Korrelationen zwischen Untertests von BiZ und Untertests des „SSV- Sprachscreenings für das Vorschulalter“ (Grimm 2003) berechnet. In einer weiteren Validierungsstudie wurden Untertests von BiZ mit einem Untertest von „HASE – Heidelberger Auditives Screening in der Einschulungsuntersuchung“ (Brunner & Schöler 2002) in Beziehung gesetzt. Diese Ergänzung

[63] Dass einzelne Wortproduktionsaufgaben (Nomen, Verben, Adjektive) nicht nur auf dem Faktor „Wortschatz“, sondern parallel dazu in geringerem Maße auch auf dem Faktor „Sätze nachsprechen“ (Wortproduktion: Nomen, Verben, Präpositionen) laden, ist mit dem Sprachkompetenzmodell vereinbar.

wurde gewählt, weil die Aufgaben der betreffenden Untertests von SSV und HASE nach differenten Regeln konstruiert worden sind.
SSV ist ein Sprachtest mit der Funktion, potentielle Risikokinder zu identifizieren. Er hat die Untertests Phonologisches Arbeitsgedächtnis für Nichtwörter (PGN), Morphologische Regelbildung (MR) und Satzgedächtnis (SG). Durchgeführt wurden – wie von Grimm vorgeschrieben - je nach Alter des Kindes jeweils zwei Untertests. Bei Kindern von 3;0 bis 3;11 Jahren fanden die Untertests PGN und MR Anwendung; bei den Kindern von 4;0 bis 5;11 Jahren kamen die Untertests PGN und SG zum Einsatz.

HASE zielt darauf, Kinder aufzuspüren, die aufgrund von auditiven Informationsverarbeitungsstörungen Sprach- und Schriftspracherwerbsprobleme und -störungen entwickeln und daher gezielte Förderungen benötigen. Das Verfahren besteht aus vier Untertests. Für die Validitätsstudie wurde nur der Untertest „Nachsprechen von Kunstwörtern (KNG)“ verwendet, denn die anderen Untertests weisen nur bedingt Nähe zu den BiZ-Untertests auf. Bei der Testdurchführung haben wir – wie von Brunner und Schöler empfohlen - die CD als Vorgabe eingesetzt.

BiP wurde zusammen mit dem „Aktiven Wortschatztest für 3- bis 5-jährige Kinder – Revision – (AWST-R) „ (Kiese-Himmel 2005) durchgeführt. Dieser Wortschatztest erfasst ausschließlich Sprachentwicklungsrückstände, die sich im Bereich des expressiven Wortschatzes niedergeschlagen haben.

Der nachfolgenden Übersicht zu allen „Übereinstimmungsstudien“ können die jeweiligen Stichprobengrößen und Korrelationskoeffizienten entnommen werden.

		Validitätsskoeffizienten	
Stufe 1 (BiZ)	388	Delfin 4–KN x SSV–PGN	.50**
	235	Delfin 4–HA x SSV–MR	.51**
	225	Delfin 4–SN x SSV–SG	.67**
	31	Delfin 4–KN x HASE–KNG	.70**
Stufe 2 (BiP)	368	Delfin 4–WV x AWST–R	.77**
	357	Delfin 4–WP x AWST–R	.80**

N = Stichprobengröße
** = Die Korrelation (nach Pearson) ist auf dem Niveau von 0,001 (2-seitig) signifikant

Tabelle 61: Validitätskoeffizienten von Delfin 4 Stufe 1 und Stufe 2 (bezogen auf andere Sprachtests)

Was BiZ betrifft, so wurden zwischen dem SSV-PGN und Delfin 4-KN, sowie zwischen dem SSV-MR und Delfin 4-HA mittlere Übereinstimmungen, zwischen dem SSV-SG und Delfin 4-SN sogar eine hohe Übereinstimmung ermittelt. Dass zwischen Delfin 4-KN und HASE-KNG eine höhere Übereinstimmung resultierte als zwischen Delfin 4-KN und SSV-PGN erklären wir uns damit, dass die Delfin 4-Kunstwörter vom Konstruktionsprinzip her eher den HASE-Kunstwörtern gleichen (Konsonsant-Vokal-Folgen) als denen des SSV (die nach dem Prinzip der Wort(un)ähnlichkeit gebildet worden sind, so dass dort auch komplizierte Konsonantencluster und zahlreiche Zischlaute auftauchen).

Im Hinblick auf BiP zeigte sich, dass zwischen dem AWST-R und Delfin 4-WV sowie Delfin 4-WP eine hohe Übereinstimmung besteht. Erwartungsgemäß ist dabei der Zusammenhang zwischen den AWST-R-Ergebnissen und der mit Delfin 4 gemessenen Wortproduktion stärker ausgeprägt, da der AWST-R ja den expressiven Wortschatz erfasst.

Bezogen auf andere Sprachtests ließ sich also die Übereinstimmungsvalidität von Delfin 4 für die Bereiche Wortschatz, Morpho-Syntax und Phonembewusstheit belegen.

Linguistische Analysekriterien:

Um auch die Validität der Untertests „Bildbeschreibung (BB)“ von BiZ und „Bilderzählung (BE)“ von BiP belegen zu können, mussten wir – aus den bereits genannten Gründen - andere Wege gehen. Im Bereich der (psycho-)linguistischen Forschung gibt es aber eine lange Tradition, die Qualität von Texten mit Hilfe linguistischer Analysekriterien zu qualifizieren. So bedient sich z.B. Bishop (2004) bei ihrem Test ERRNI zur Erfassung der Erzählkompetenz von Kindern ab vier Jahren solcher Kriterien.

Angesichts dessen erfolgte die Prüfung der Validität des Subtests BE von Delfin 4 – Stufe 2 anhand externer linguistischer Analysekriterien durch linguistisch geschulte Rater/innen.

Zu diesem Zweck wurden die im Verlauf der Durchführung der Untertests BE von Delfin 4 – Stufe 2 produzierten Erzählungen der Kinder audiografiert und anschließend nach vorgegebenen Transitionsregeln transkribiert. Diese Texte wurden parallel analysiert: a) Mit dem gröberen und dadurch in der Kita- und Schulpraxis praktikablen Auswertungsraster von Delfin 4; b) mittels Inhaltsanalysen, bei denen gängige linguistische Kriterien eingesetzt wurden. Erstere Analysen wurden durch diejenigen Erzieher/innen bzw. Lehrer/innen vorgenommen, welche das Verfahren durchgeführt haben; letztere durch je zwei linguistisch ausgebildete Personen.

Für die Erfassung quantitativer als auch qualitativer Aspekte des kindlichen Lexikons und der syntaktischen Komplexität kindlicher Äußerungen stehen verschiedene Messmethoden zur Verfügung, die sich im Hinblick auf ihre Aussagekraft stark unterscheiden.

Eine Reihe von Maßen verwendet die Anzahl aller vom Kind geäußerten Wörter, die Anzahl der verschiedenen Wörter sowie das Verhältnis dieser beiden Größen. Da sich jedoch die spontansprachlichen Samples stärker in ihrer Größe unterscheiden als Kindäußerungen bei anderen (Test)aufgaben, werden verschiedene Begrenzungen definiert (z.B. Zeitvorgaben, fixe Anzahl von Äußerungen, die in die Untersuchung einfließen), um die Vergleichbarkeit der Samples sicher zu stellen.

Die Studien von z.B. Hess u.a. (1984) und Richards (1987) zeigten nämlich, dass die Größe des untersuchten Samples erheblichen Einfluss auf das traditionelle Maß der TTR (type-token-ratio: „number of different words“ geteilt durch „total number of words“) hat: Je weniger Äußerungen, desto günstiger erscheint der Quotient. Auch Richards & Malvern (1997) wiesen darauf hin, dass inter- und intraindividuelle Vergleiche nur bei vergleichbarer Samplegröße zulässig sind. Zur Abschwächung des Effektes der Samplegröße wurden verschiedene Vorschläge gemacht (konstante Äußerungsanzahl wie z.B. 50 oder 100 Äußerungen, eine Zeitbegrenzungen auf z.B. 10 Min.). Kritisiert wird hier die Gefahr der Vermischung von lexikalischer Vielfalt und Äußerungslänge oder Sprechfreude bzw. Motivation (vgl. Kauschke 2000: 82).

Ein empirisches Problem stellt die verhältnismäßig kurze Sample-Länge unserer Stichprobe dar (teilweise nur vier Äußerungen); auch eine zeitliche Begrenzung bei der Bearbeitung der Aufgabe erscheint nicht sinnvoll.

Günstiger für unsere Zwecke ist der Vorschlag, die Anzahl der verschiedenen Wörter an einer festen Tokenzahl zu berechnen (z.B. 100, 200, …) (vgl. dazu Richard & Malvern 1997; sie halten allerdings eine Anzahl von 400 Wörtern für notwendig, damit die TTR stabil bleibt und als reliables Maß gelten kann. Damit können wir bei unserer Stichprobe nicht dienen - die Kinder produzieren durchschnittlich 46 Wörter)

Problematisch bleibt, dass beim interindividuellen Vergleich nur das Kind mit der geringsten Sample-Länge die Bezugsgröße vorgeben kann, so dass entweder viele Kinder aus der Stichprobe herausfallen oder von einer sehr kleinen Sample-Länge ausgegangen werden muss, wie die folgenden Angaben der TTR für Samples mit 40, 50, 60 und 70 Wörtern zeigen:

Was die diagnostische Relevanz der verschiedenen Maße betrifft, so dienen nach Klee (1992) dienen zur Unterscheidung von Kindern mit ungestörter und gestörter Sprachentwicklung u. a. folgende Maße:

- number of different words
- total number oft words und
- MLU + MSL (mean length of utterance und mean syntactic length)

Bei konstant gehaltener Sample-Länge zeigt die TTR (100 oder 200) zwar Korrelationen zum Alter, kann jedoch keine Entwicklungsprobleme erfassen.
Watkins u. a. (1995) zeigten jedoch, dass die TTR bei konstanter Tokenzahl diagnostisch signifikant war (50 oder 100 Tokens) – die SLI-Kinder fielen in dieser Studie deutlich auf.

Als günstigste Maße zur Erfassung der Vielfalt des kindlichen Lexikons gelten demnach für unsere Zwecke:

- Types-Zählung (number of different words innerhalb einer bestimmten Sample-Länge, also die TTR bei konstanter Tokenzahl)
- Anzahl der Tokens (total number of words), wenn auch Einflüssen wie Sprechfreudigkeit unterworfen;
- Vergleich von Anzahl von Nomen, Verben, etc. (steht bei uns noch aus).

(vgl. Watkins u.a. 1995; Gopnik & Choi 1995, zusammenfassend dazu Kauschke 2000: 85)

Diese sollen noch kurz charakterisiert werden.

<u>Anzahl Tokens:</u>
Gesamtzahl der von einem Kind geäußerten Wörter bei der Aufgabe BE.

<u>Anzahl Sätze:</u>
Gesamtzahl der Äußerungen („Sätze“) eines Kindes bei der Aufgabe BE
Dieses Maß gilt allenfalls als grober Indikator für Sprechfreudigkeit, jedoch nicht für lexikalische oder grammatische Kompetenz (jedoch dies bei unserer Studie auch nru bedingt). Wir haben es dennoch mit dem Gesamtpunktwert von BE korreliert, um zu zeigen, dass der Zusammenhang nicht besonders hoch ist. Ein Kind, das „mehr“ erzählt, ist also nicht auch „kompetenter“ im Erzählen – dies entspricht durchaus unserem theoretischen Modell, bei dem eine Erzählung beispielsweise aus vier inhaltlich und gram-

matisch komplexen Äußerungen „gehaltvoller“ sein kann als eine, die zwar aus vielen, aber unterkomplexen Äußerungen besteht.

MLU (words):

Eingeführt von Brown (1973); mittlere Äußerungslänge (in Wörtern, nicht in Morphemen) – gilt als grober Anhaltspunkt zur Einschätzung grammatischer Fähigkeiten bzw. des allgemeinen Standes der Sprachentwicklung; nach Klee 1992 diagnostisch relevantes Maß zur Unterscheidung von Kindern mit und ohne Sprachauffälligkeiten. Zahlreiche Korrelationen zwischen der MLU und anderen syntaktischen und morphologischen Leistungen bestätigen dieses Maß als geeigneten Indikator für die grammatische Kompetenz eines Kindes (vgl. Rondal u. a. 1987; Bates u. a. 1988). Allerdings gilt dies nur für die ersten Phasen des Spracherwerbs bis zu einem alter von etwa 3,6 Jahren. Danach können aus der MLU zwar keine klaren Rückschlüsse auf die grammatische Komplexität der Äußerungen gezogen werden, jedoch gilt sie dennoch als allgemeiner Hinweis zum Sprachentwicklungsstand (vgl. Rosenthal Rollins u.a. 1996).

TTR 40 und 50:

Type-Token-Ratio (TTR): $\frac{\text{number of different words (NDW)}}{\text{total number of words (TNW)}}$

Quotient aus der Anzahl der unterschiedlichen Wörter geteilt durch die Gesamtzahl der Wörter (verglichen wird stets die gleiche Textlänge, hier für die ersten 40 und 50 Wörter).

Satzlänge MAX:

Anzahl der Wörter im längsten Satz – (upper bound nach Brown 1973); dieses Maß gibt einen Hinweis auf größtmögliche Kapazität des Kindes, die Reichweite der syntaktischen Kompetenz

Profilanalyse (0-4):

Da die MLU bei Kindern über 3,6 Jahren als nicht mehr so zuverlässiger Indikator für den grammatischen Entwicklungsstand gilt, haben wir ein weiteres Analysekriterium hinzugezogen, bei dem die kindlichen Äußerungen direkt auf das Vorkommen von bestimmten, komplexeren syntaktischen Formen durchsucht werden und dann einer bestimmten Entwicklungsstufe zugeordnet werden können. Im Rahmen von Profilanalysen

werden ähnliche Kriterien angelegt, wir lehnen uns an die Stufen von 0 - 4 von Grießhaber 2002 an (adaptierte, radikal vereinfachte Fassung von Clahsen 1985). Die Stellung des finiten Verbs gilt hier als Hauptmerkmal der Spracherwerbsstufen, ähnlich der Analyse von Reich/Roth in HAVAS 5. Die Stufen reichen von 0: bruchstückhafte Äußerung bis hin zu 4: Nebensatzbildungen mit Verb in Endstellung.

Nach Marjanovic-Umek, Kranjc & Fekonja (2002) lassen sich Erzählungen von Kindern im Vorschulalter im wesentlichen durch zwei Textualitätskriterien kennzeichnen: Kohärenz und Kohäsion. Die Kohärenz basiert hierbei auf der konzeptionell zu erschließenden Struktur einer Geschichte, d.h. wie die verschiedenen Teile einer Geschichte inhaltlich-logisch verbunden werden; die Kohäsion hingegen bezieht sich auf die durch Sprachmittel verknüpften Teile der Geschichte. Die von uns herangezogenen linguistischen Analysekriterien (quantitative und qualitative Aspekte des kindlichen Lexikons und der syntaktischen Komplexität der kindlichen Äußerungen) können als Indikatoren dieser Textualitätskriterien angesehen werden. Sie äußern sich u. E. konkret darin, dass sich Kinder in der Lage zeigen, mehrere Akteure logisch miteinander zu verknüpfen, mehrere Ereignisse zeitlich zu verknüpfen, den Höhepunkt eines Ereignisablaufs zu markieren usw. Das gelingt ihnen am ehesten, wenn sie schon fähig sind, in Mehrwortsätzen bzw. Satzreihen oder –gefügen zu sprechen.

Im einzelnen haben wir folgende Analysekriterien angewandt: Anzahl unterschiedlicher Wörter, Mittlere Äußerungslänge auf Basis der Wörter (MLU words), Type-token-ratio (TTR: Quotient aus der Anzahl der unterschiedlichen Wörter geteilt durch die Gesamtzahl der Wörter) und Satzlänge MAX (Anzahl der Wörter im längsten Satz). Außerdem haben wir auf ein Verfahren zurückgegriffen, bei dem die kindlichen Äußerungen direkt auf das Vorkommen von bestimmten, komplexeren syntaktischen Formen durchsucht und anschließend einer bestimmten Entwicklungsstufe zugeordnet werden. Diese sogenannten Profilanalysen lehnen sich an die Stufen 0 - 4 von Grießhaber (2004) an.

Die Aufbereitung der mündlichen Daten erfolgte gemäß den Angaben in Bishop, D.V.M. (2004): Expression, Reception and Recall of Narrative Instrument. ERRNI Manual. Oxford; allerdings für das Deutsche an einigen Stellen leicht verändert. Die Analyse der Texte erfolgte durch speziell geschulte Personen mit linguistischer Vorbildung (einschlägiges Studium). Die „interscorer reliability“ wurde anhand von parallel und dabei

unabhängig durchgeführten Analysen geprüft. Die berechnete durchschnittliche Raterübereinstimmung (> 92) dokumentiert das hohe Ausmaß an interpersonaler Übereinstimmung.

Die Korrelationsberechnung beruhen auf transkribierten Erzählungen, welche der der Durchführuing des Untertests BE von BiP evoziert worden waren. Diese wurden einerseits mit Hilfe des Analyserasters eingeordnet, das der BE-Aufgabe zugrunde liegt, andererseits mittels der oben genannten linguistischen Textanalysekriterien charakterisiert. Der jeweilige Grad der Übereinstimmung kann der folgenden Übersicht entnommen werden.

	Validitätskoeffizienten		N
Linguistische Analysekriterien	Anzahl Wörter	.61**	132
	MLU (words)	.61**	132
	TTR	.49**	48
	Satzlänge-MAX	.50**	132
	Profilstufe (0 – 4)	.57**	132

Tabelle 62: Validitätskoeffizienten zum Untertests BE von BiP (bezogen auf linguistische Analysekriterien)

N = Stichprobengröße

** = Die Korrelation (nach Pearson) ist auf dem Niveau von 0,001 (2-seitig) signifikant

Aus diesen Befunden folgern wir, dass es mit Delfin 4 nicht nur möglich ist, den Wortschatz, die Morphosyntax und das Phonemgedächtnis, sondern auch die Erzählkompetenz (insbesondere im Hinblick auf Kohäsion und Kohärenz) valide zu erfassen.

Sensitivität und Spezifität

Delfin 4 ist ein Screening, welches Kinder anhand ihrer im Test gezeigten Sprachleistungen in zwei Gruppen einteilt: in Kinder mit und ohne angezeigtes Sprachentwicklungsrisiko. Bei Klassifikationen dieser Art sind zwei Arten von Fehlern möglich: Ein Kind wird als risikogefährdet angesehen, obwohl es das nicht ist, oder es wird umgekehrt als nicht in seiner Sprachentwicklung gefährdet angesehen, obwohl es der Fall ist.

In der Medizin sowie der Sprachheilpädagogik wendet man spezifische Maße an, um die Zuverlässigkeit derartiger binären Klassifikation (z.B. Sprachentwicklungsrisiko ja/nein) zu kennzeichnen. Damit kann man quantifizieren, um wie viel besser ein Screening im Vergleich zu einer zufälligen Zuordnung abschneidet. Die Werte variieren dabei zwischen 0 und 1. Ein Wert ab 0,6 wird bereits als deutliche Verbesserung gegenüber einer zufälligen Zuordnung eingestuft.

Delfin 4 ist zu dem Zweck entwickelt worden, potentielle Sprachentwicklungsrisiken anzuzeigen, ohne gleich so weit zu gehen, die Kinder als „sprachentwicklungsverzögert“ oder gar als „sprachentwicklungsgestört“ zu klassifizieren. Solche Festlegungen können höchstens durch speziell geschulte Personen und das auch nur im Verlauf einer längerfristigen Diagnostik erfolgen.

Im Rahmen der Entwicklung von Delfin 4 haben wir deshalb bewusst auf solche Klassifizierungen verzichtet. Trotz dieser grundsätzlichen Bedenken haben wir beschlossen, die in unseren Daten steckenden Möglichkeiten zu nutzen, um die Klassifizierungsstärke von Delfin 4 zu bestimmen. Dabei kam uns zugute, dass die Testleiter/innen bei der Durchführung bzw. Auswertung von BiZ aufgefordert waren, die Sprachäußerungen der Kinder hinsichtlich dabei auftretender Auffälligkeiten zu charakterisieren. Konkret wurde gefragt: „Was fällt an den Sprachäußerungen auf?“. Die Antworten sollten dabei bezüglich folgender Kategorien (Laut-/Sprach-Artikulation, Wortschatz, Morpho-Syntax) näher bestimmt werden. Damit war es möglich, das Delfin 4-Ergebnis mit der Einschätzung der Sprachäußerungen durch die Testleiter/innen in Beziehung zu setzen.

Zu diesem Zweck haben wir die Normierungsstichprobe BiZ 2007 genutzt, um zu überprüfen, ob die Kinder anhand der Biz- und der Beobachtungsergebnisse der Testlei-

ter/innen (BEOB) übereinstimmend als „risikogefährdet“ („rote Ampelfarbe“), „möglicherweise risikogefährdet“ („gelbe Ampelfarbe“ oder „nicht erkennbar risikogefährdet“ („grüne Ampelfarbe“) eingeordnet wurden. Die Verteilung dieser Klassifizierung kann der folgenden Übersicht entnommen werden. Bei der Variable BiZ ist die Einteilung nach „Ampelfarben“ Teil der Testauswertung; bei der Variable „BEOB“ wurde anhand der Anmerkungen der Testleiter/innen im Protokollheft (Frage: „Was fällt an den Sprachäußerungen auf?“) als „rot“ eingeordnet, wenn übergreifende Auffälligkeiten im Bereich Artikulation und/oder im Bereich Wortschatz sowie Morphosyntax berichtet worden waren; als „gelb“, wenn „Artikulationsprobleme“ genannt worden waren und als „grün“, wenn keinerlei Hinweise auf Sprachauffälligkeiten erfolgten. Die vollständigen Aufgabenlösungen BiZ nebst Angaben zu den beobachteten Sprachauffälligkeiten lagen von 834 Testleiter/innen vor.

		BiZ			
		grün	gelb	rot	Gesamt
BEOB	grün	455	113	1	68,2%
	gelb	1	93	67	19,3%
	rot	0	0	104	12,5%
Gesamt		54,6%	24,8%	20,6%	100,0%

Tabelle 63: Klassifizierungen gemäß Biz und BEOB (N = 834)

Die Sensitivität indiziert den relativen Anteil derjenigen Kinder, die – gemessen am Kriterium „Beobachtungen der Praktiker/innen“ – durch das Screening korrekt als „risikobehaftet“ im Sinne von „starke Sprachauffälligkeiten“ (rote Ampelfarbe)bzw. „geringe Sprachauffälligkeiten“ (gelbe Ampelfarbe)aufweisend. Unter der Voraussetzung, dass man alle rot-rot-, gelb-gelb- sowie gelb-rot-Übereinstimmungen als „korrekte Klassifikation“ wertet, beträgt der Sensitivitätskoeffizient 0,996. BiZ vermag somit Risiken mit sehr hoher Treffsicherheit zu identifizieren.

Die Spezifität zeigt den relativen Anteil derjenigen Kinder an, die – gemessen am Kriterium „Beobachtungen der Praktiker/innen“ – durch das Screening korrekt als „nicht er-

kennbar risikobehaftet“ klassifiziert wurden. Unter der Voraussetzung, dass nur die grün-grün-Übereinstimmung zählt, beträgt der berechnete Spezivitäts-koeffizient 0,800. BiZ weist also auch beim Herausfiltern von Kindern ohne erkennbare Sprachentwicklungsrisiken eine zwar geringere, aber immer noch hohe Treffsicherheit auf.

8. Normierung

Die Normierung (Eichung) eines Tests erfordert die Festlegung eines Bezugssystems, das es ermöglicht, die Merkmalsausprägungen eines Kindes im Vergleich zu den Merkmalsausprägungen anderer Kinder eindeutig einzuordnen und zu interpretieren (vgl. z.B. Lienert & Raatz 1998). Um solche Vergleichswerte zu erhalten, muss ein einheitlicher Maßstab (soziale oder kriterienbezogene Norm) bestimmt werden. Wie bei den bislang im Früh- und Elementarbereich vorgelegten deutschen Sprachtests der Fall, wurden der Delfin 4-Sprachtest sozial normiert; was u.a. darin begründet liegt, dass der wissenschaftliche Erkenntnisstand zur akademischen Sprachkompetenz von etwa vierjährigen Kindern noch nicht präzise und detailliert genug kriterial bestimmt werden kann.

Delfin 4 Stufe 1 „Besuch im Zoo (BiZ)“

Stichprobe

Bei der Normierung kommt es entscheidend darauf an, möglichst aussagekräftige Vergleichswerte von solchen Kindern zu erhalten, die den mit BiZ zu untersuchenden Kindern hinsichtlich relevanter Merkmale (z.B. Alter, Geschlecht, Sozialstatus, Muttersprache) ähnlich sind (Moosbrugger & Kelava 2007, S. 19). Die Normierung muss deshalb auf der Basis einer möglichst großen Stichprobe erfolgen, welche die Grundgesamtheit (hier alle Kinder in Nordrhein-Westfalen zwei Jahre vor Schulbeginn) in Bezug auf relevante Merkmale möglichst weitgehend repräsentiert.

Dem wurde in zwei aufeinanderfolgenden Jahren durch folgende Bemühungen Rechnung getragen:

Im Jahr 2007 wurde eine erste Normierungsstichprobe gewonnen, die sich aus Kindern mit deutscher Muttersprache[64] zwei Jahre vor Schulbeginn zusammensetzt. Sie umfasst insgesamt 14.859 Kinder. Im Durchschnitt sind diese 4 Jahre, 1 Monat alt (Stan-

[64] Die Beschränkung auf Kinder mit deutscher Muttersprache ergibt sich aus einer der beiden Funktionen des zweistufigen Screeningverfahrens „Delfin 4 – Stufe 1 und 2“, nämlich zu überprüfen, ob ein Kind zwei Jahre vor der Schule hinreichend Deutsch beherrscht. Das kann aber nur angemessen entschieden werden, wenn man sich am Sprachkompetenzlevel von Kindern mit deutscher Muttersprache orientiert.

dard-abweichung: 4 Monate). Die jüngsten Kinder sind 3 Jahre, 6 Monate, die ältesten 4 Jahre, 10 Monate alt.

Die Repräsentativität einer Normierungsstichprobe wird gemeinhin dadurch gewährleistet, dass sie gezielt nach relevanten Faktoren geschichtet wird (Schichtungs- oder Quotierungsstichprobe). Die Normierungsstichprobe von 2007 ist in Bezug auf zwei potentiell sprachentwicklungsbedeutsame Faktoren repräsentativ geschichtet: Geschlecht der Kinder sowie Sozialindex des Schulaufsichtsbezirks, in dem die Kinder getestet werden.

Was das Geschlecht betrifft, so ist man sich in der wissenschaftlichen Literatur nicht einig, ob und wieweit es bedeutsame Differenzen in der Sprachentwicklung von Jungen und Mädchen gibt, d.h. es liegen widersprüchliche Befunde vor; „..., wobei nicht klar ist, ob das daran liegt, dass keine Analysen durchgeführt wurden, oder daran, dass die Ergebnisse keine Unterschiede aufzeigten." (Hoff-Ginsberg 2000, S. 479). Dort, wo in Forschungen Unterschiede zutage traten, wie z.B. bei Grimm, Aktas und Frevert (2000) sowie Schöler und Schäfer (2004) der Fall, zeigten sich die Mädchen den Jungen überlegen. Vor diesem Hintergrund haben wir Jungen und Mädchen in der Normierungsstichprobe von 2007 so ins Verhältnis gesetzt, wie in der Statistik der Kinder- und Jugendhilfe Nordrhein-Westfalen angegeben (Landesamt für Datenverarbeitung und Statistik NRW 2007). Letztendlich setzt sich die Normierungsstichprobe aus 51,2 Prozent Jungen und 48,9 Prozent Mädchen zusammen.

Dass die Sprachentwicklung von Kindern bedeutsam durch sozioökonomische Faktoren bestimmt wird, ist in der Wissenschaft unumstritten (Barnett 2001; Bradley & Corwyn 2002; Fried 2003; Gonzales 2001; Halle et al. 2003; Hoff 2003; Hoff-Ginsberg 2000). Deshalb haben wir die Normierungsstichprobe von 2007 auch nach Sozialindex geschichtet. Es handelt sich dabei um ein Maß, das die soziale Belastung von Schulaufsichtsbezirken (d.h. Kreise bzw. kreisfreie Städte) anzeigt. Es basiert auf den vier sozialografischen Merkmalen: Arbeitslosenquote, Sozialhilfequote, Migrantenquote, Sozialhilfequote (Ausländer und Aussiedler) und Quote der Wohnungen in Einfamilienhäusern. Die einzelnen Merkmale wurden in einem statistischen Verfahren (Faktorenanalyse) zur Skala „Sozialindex" zusammengefasst. Der Konstruktion des Sozialindex liegt die Annahme zugrunde, dass Kreise mit vielen Arbeitslosen, Sozialhilfeempfängern und Mig-

ranten sowie einem geringen Anteil von Wohnungen in Einfamilienhäusern sozial stärker belastet sind als Kreise mit einer einheimischen, erwerbstätigen, den Lebensunterhalt selbständig bestreitenden und Einfamilienhäuser bewohnenden Bevölkerung (vgl. Frein, Möller, Petermann & Wilpricht 2006).

Da sich der Sozialindex jeweils auf den ganzen Kreis bzw. die kreisfreie Stadt bezieht, in dem bzw. in der das Kind getestet wird, handelt es sich dabei um kein individuelles Merkmal. Dennoch darf der Faktor als aussagekräftig gelten, weil belegt werden konnte, dass bedeutsame Zusammenhänge zwischen Sozialindex und Bildungschancen bzw. -erfolgen der Kinder eines Schulaufsichtsbezirks bestehen (ebd., S. 188).

Wie die folgende Tabelle zeigt, bildet die Normierungsstichprobe von 2007 fast perfekt die Sozialindexverteilung der Kinder beim ersten Testdurchgang 2007 ab.

Sozial-index	Nordrhein-Westfalen	Prozentanteil	Normierungs-stichprobe	Prozentanteil
6	32.716	21,3	3.129	21,1
5	55.755	36,2	5.412	36,4
4	10.240	6,7	983	6,6
3	17.234	11,2	1.639	11,0
2	25.350	16,5	2.459	16,5
1	12.625	8,2	1.237	8,3
Summe	153.920	100,1*	14.859	99,9*

Tabelle 64: Stichprobenverteilung nach Sozialindex
* = Rundungsfehler;

Im Jahr 2008 haben wir die Normierungsbemühungen mittels einer – auf den Kreis Recklinghausen beschränkten – weiteren Normierungsstichprobe erweitert. Diese setzt sich ebenfalls aus Kindern mit deutscher Muttersprache zwei Jahre vor Schulbeginn zusammen. Sie umfasst insgesamt 3.636 Kinder. Im Durchschnitt sind diese 4 Jahre, 3 Monate alt (Standardabweichung: 2 Monate). Die jüngsten Kinder sind 3 Jahre, 3 Monate, die ältesten 5 Jahre, 1 Monat alt. Das Verhältnis Jungen: Mädchen beträgt 50,6% : 49,4%.

Die Normentabellen basieren auf beiden Normierungsstichproben.

Mit „Normierung“ bezeichnet man das Berechnen von Kennzahlen, die das Verhältnis jedes einzelnen Rohwertes zu den Ergebnissen der Normierungsstichprobe zum Ausdruck bringen (Wottawa 1980, S.102). Dazu wird für jedes Kind der Normierungsstichprobe die Zahl der Punkte ermittelt, die es für seine „Lösungen“ erhält. Diese „Rohwerte“ sind nur bedingt aussagekräftig, denn sie ermöglichen nur Aussagen im Hinblick auf diese eine Stichprobe. Somit können sie nicht ohne weiteres verglichen werden. So markiert z.B. der Rohwert 6 im Untertest „Kunstwörter nachsprechen“ des BiZ ein ganz anderes Leistungsniveau, als der Rohwert 6 im Untertest „Sätze nachsprechen“ des BiZ.

Deshalb werden die Rohwerte in Standardwerte transformiert, d.h. auf einer einheitlichen Skala abgebildet und so vergleichbar gemacht. Es gibt verschiedene Möglichkeiten, Rohwerte statistisch in Standardwerte zu transformieren, wie z.B. Prozentränge, T-Werte, IQ-Werte, C-Werte, z-Werte und Stanine-Werte (vgl. z.B. Tent & Stelzl 1993, S. 58). Welche dieser Möglichkeiten in Frage kommt, hängt nicht zuletzt davon ab, ob die Normierungsstichprobe normalverteilt ist[65]. Da dies (wie bei den meisten anderen Sprachtests auch) bei der BiZ-Normierungsstichprobe nicht gegeben ist, wurden nur die für diesen Fall vorgesehenen Standardnorm-Äquivalente berechnet.

Die dafür erforderlichen statistischen Transformationen gehen von der verteilungsfreien Prozentrangskala aus. Es handelt sich dabei um eine Rang- oder Ordinalskala. Die Prozentrangplätze werden über die kumulativen Häufigkeiten errechnet. Sie markieren die relative Position eines Kindes im Vergleich mit der Rangordnung der Kinder der Normierungsstudie. Wenn z.B. ein Kind bei BiZ einen Prozentrang (PR) von 75 hat, weiß man, dass seine Sprachkompetenz gleich gut oder besser ist als die von 75 Prozent der Normierungsstichprobe; oder anders formuliert: nur 25 Prozent der Kinder der Normierungsstichprobe besitzen eine noch höhere Sprachkompetenz (vgl. Moosbrugger & Kelava 2007, S. 19). Der Nachteil von Prozentrangnormen ist, dass sie im mittleren Bereich erheblich feiner differenzieren als in den Extrembereichen.

[65] Nichtnormalverteilte Merkmale können durch eine „Flächentransformation“ normalisiert, das meint: in sogenannte Standardnormäquivalente transformiert werden (vgl. Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007).

Deshalb haben wir die Prozentrangnormen in Standardnorm-Äquivalente transferiert[66]. Diese geben jeweils an, wie weit ein Rohwert vom arithmetischen Mittelwert der Normierungsstichprobe entfernt ist. Dabei wird diese Differenz in Form von Standardabweichungen[67] ausgedrückt.

Das bei Sprachtests ohne Normalverteilung gebräuchlichste (und auch bei BiZ verwendete) Standardnorm-Äquivalent ist die T-Wert-Skala (TW). Sie hat einen definierten Mittelwert von TW 50 und eine definierte Standardabweichung von TW 10. Per Definition markieren 40 bis 60 TW eine durchschnittliche und TW über 60 TW eine überdurchschnittliche Leistung. Ein TW unter 40 indiziert eine Risikoent-wicklung.

Gruppennormen

In der Fachliteratur wird zwischen Gesamtnormen und Gruppennormen unterschieden (Lienert & Raatz 1998[6]). Von Gesamtnormen spricht man, wenn die Normierung über die Gesamtpopulation der Normierungsstichprobe erfolgt, von Gruppennormen, wenn sie über eine oder mehrere Teilpopulationen erfolgt. Gruppennormen sind dann angebracht, wenn die Gesamtpopulation der Normierungsstichprobe aus Teilpopulationen zusammengesetzt ist, die sich bezüglich des relevanten Merkmals systematisch unterscheiden. Ob das der Fall ist, lässt sich anhand von Varianzanalysen ermitteln. Mit diesem Verfahren wird getestet, ob die Varianz zwischen bestimmten Gruppen größer ist als die Varianz innerhalb der Gruppen. Dadurch kann ermittelt werden, ob sich bestimmte Gruppen statistisch signifikant unterscheiden oder ob das nicht der Fall ist.

Varianzanalytische Berechnungen anhand der BiZ-Normierungsstichprobe ergaben, dass es sowohl in Bezug auf Geschlecht, als auch in Bezug auf das Alter statistisch signifikante Unterschiede gibt. Die Belege dazu finden sich im Anhang.

Den statistisch bedeutsamen Differenzen zwischen den Kindern, die unter vier Jahre alt und denen, die vier Jahre und älter sind, wird durch dementsprechende Altersnormen entsprochen (vgl. Normentabellen im Anhang). Dagegen werden – ungeachtet der ebenfalls statistisch bedeutsamen Unterschiede zwischen Mädchen und Jungen (vgl.

[66] Die Transformation von Prozentrangwerten in T-Werte wurde anhand statistischer Umrechnungstabellen vorgenommen (vgl. z.B. Bühner, 2006[2], S. 106).
[67] Maß für die Streuung um den mittleren Wert.

ebenfalls Anhang) – keine Geschlechtsnormen angeboten. Das hat zwei Gründe: Zum einen variieren die Leistungsunterschiede zwischen den Geschlechtern nur in einem solchen Ausmaß,, dass eine getrennte Normierung nicht zwangsläufig erforderlich ist. Zum anderen ist der praktische Nutzen von Geschlechtsnormen fraglich, weil eine nach Geschlechtern getrennte Sprachförderung pädagogisch keinen Sinn macht.

Risikoniveau

BiZ hat eine Screeningfunktion[68], d.h. es soll dazu dienen, belastbare Informationen für die Entscheidung zu liefern, welche Kinder zwei Jahre vor der Schule eine zusätzliche Sprachförderung erhalten sollten, damit ihre Bildungschancen zum Schulstart möglichst gewahrt bleiben. Lienert und Raatz (1998[6]) warnen davor, bei derart weitreichenden Konsequenzen allzu feine Normen, zu verwenden, weil diese dazu verführen, die Messgenauigkeit von Tests zu überschätzen; sie empfehlen statt dessen, gröbere Normen zugrundezulegen. Deswegen nutzen wir die feinen T-Wert-Normen nur für die Umrechnung der Rohwerte bzw. die Berechnung von Untertest- und Gesamttestwerten; orientieren jedoch die Entscheidung, ob zusätzlicher Sprachförderbedarf besteht, an den wesentlich gröberen Stanine-Normen (SN)[69].

Die SN-Skala[70] (Stanine oder SN steht für „standard nine“) umfasst die Werte 1 bis 9. Sie hat einen Mittelwert von 5 und eine Standardabweichung von 2. Dabei kennzeichnet SN 1 eine extrem unterdurchschnittliche Leistung. SN 2 bis 4 markieren unterdurchschnittliche Leistungen. Ab SN 5 können Leistungen als durchschnittlich bzw. überdurchschnittlich gelten. Diese grobe Einteilung reicht aus, um Risikoniveaus zu definieren bzw. Grenzwerte festzulegen, die den Übergang von einer Niveaustufe zur nächsten markieren.

Anhand der „Tabelle zur Ermittlung des Risikoniveaus“ (vgl. Anhang) kann abgelesen werden, in welche von drei Risikoniveaus die Leistung eines Kindes zwei Jahre vor der Schule einzuordnen ist: in den Rot-Bereich (SN 1), der signalisiert, dass ein extremes Risiko erkennbar wird, so dass eine zusätzliche Sprachförderung vonnöten ist; in den

[68] Verfahren, mit dem Kinder „herausgefiltert“ werden sollen, die bestimmte Merkmale aufweisen.
[69] Es handelt sich dabei, wie auch bei den T-Wert-Normen der Fall, um Normen, die keine Normalverteilung der Normierungsstichprobe voraussetzen.
[70] Die Transformation der Prozentrangnormen in Stanine-Werte wurde anhand statistischer Umrechnungstabellen vorgenommen (vgl. z.B. Bühner, 2006[2], S. 106).

Gelb-Bereich (SN 2-4), der besagt, dass ein Risiko nicht auszuschließen ist, so dass nochmals (mit Delfin 4 – Stufe 2) genauer geprüft werden muss, ob eine zusätzliche Sprachförderung angebracht ist; und in den Grün-Bereich (SN 5 und größer), der ausdrückt, dass sich kein Risiko abzeichnet, so dass keine über die in Kindertageseinrichtungen übliche Sprachförderung hinausgehende Maßnahmen erforderlich ist.

Um die Zuordnung der Leistungen einzelner Kinder so treffsicher wie möglich zu gestalten, wurde folgendermaßen vorgegangen: Bei der Definition der Grenzwerte, die den Übergang von einem Risikoniveau zum nächsten markieren, wurden „Vertrauensintervalle“ (auch „Normbänder“ genannt) berechnet, welche den Bereich angeben, in dem sich der „wahre Wert“, also die tatsächliche Leistung eines Kindes mit 95%iger Wahrscheinlichkeit bewegt. Die Vertrauensintervalle haben wir berechnet, indem wir den Standardmessfehler, den ein Test immer hat, zum Schutz vor Überinterpretation kleiner Wertdifferenzen, gleich bei der Grenzwertbestimmung eingearbeitet haben. Bei BiZ wurden folgende Grenzwertintervalle (Vertrauensintervalle eines Grenzwerts) ermittelt: Rot-/Gelb-Bereich: TW 32,3 bis 33,7; Gelb-/Grün-Bereich: TW 46,3 bis 47,7.

Da man bei einem Sprachtest mit Screeningfunktion „... möglichst nicht den Fehler begehen möchte, Risikokinder für Sprach- und Schriftspracherwerbsprobleme zu übersehen...“, nimmt man „... eher in Kauf, dass Kinder als Risikokinder herausgesiebt werden, die tatsächlich aber keiner Intervention, also beispielsweise keiner weiteren Sprachförderung oder einer Therapie bedürfen.“ (Schöler & Schäfer 2004, S. 18); als den Fehler zu begehen, Risikokinder zu übersehen. Demzufolge wurde als Grenzwert immer derjenige TW bestimmt, der das geringste Risiko birgt, Risikokinder zu übersehen; also beim Übergang vom Rot- zum Gelb-Bereich der TW 32,3 und beim Übergang vom Gelb- zum Grün-Bereich der TW 47,7.

Wie in vorhergehenden Kapiteln bereits berichtet, haben wir BiP bereits 2007 an einer ausreichend großen Stichprobe normiert. Allerdings haben die Praxiserprobungen sowie die empirischen Aufgabenanalysen eine Überarbeitung des Materials nach sich gezogen. Diese „verschlankte“ Version bedurfte einer erneuten Normierung, um einen einheitlichen Bezugsmaßstab für alle Testergebnisse zu ermitteln.

Um auch für BiP Altersnormen berechnen zu können, wurde eine möglichst große repräsentative Stichprobe von Kindern benötigt, die an dem Verfahren teilnehmen. Wichtig war dabei, dass sich diese Stichprobe von Kindern wie ein „normaler“ Jahrgang zusammensetzt. Für die Normierung des Verfahrens war es daher wichtig, dass Kinder mit unterschiedlichem Sprachentwicklungsstand eingeschätzt wurden. Daher verbot es sich, nur diejenigen Kinder mit einzubeziehen, bei denen die Testung mit BiZ einen zusätzlichen Sprachförderbedarf anzeigte.

Bei der Stichprobengewinnung haben wir diesem Tatbestand dadurch Rechnung getragen, dass wir auch Eltern um Erlaubnis gebeten haben, ihr Kind in die Normierungsstudie einbeziehen zu dürfen, deren Kinder schon in der ersten Stufe des Verfahrens eine altersgemäße Sprachentwicklung gezeigt haben, so dass eine weitere Testung mit Delfin 4 eigentlich nicht nötig war. Durch die tatrkräftige Unterstützung der pädagogischen Fach- und Lehrkräfte konnten wir dann auch viele Eltern dafür gewinnen. Auf diese Weise ist sichergestellt, dass beide Stufen des Verfahrens auf der Basis von Normierungsstichproben „geeicht“ worden sind, welche die Verteilung der Sprachfähigkeit von vierjährigen Kindern in Nordrhein-Westfalen valide abzubilden vermögen.

Der Ablauf der Stichprobengewinnung kann dem folgenden Algorithmus entnommen werden. Im Verlauf dieser Schritte gelang es, eine nach Sozialindex, Alter, Geschlecht und Migrationshintergrund für Nordrhein-Westfalen näherungsweise repräsentativ zusammengesetzten Normierungsstichprobe zu gewinnen. Die Stichprobengröße betrug 2.278 Kinder. Die Zusammensetzung repräsentiert die Situation in Nordrhein-Westfalen gut: 49 Prozent Mädchen und 51 Prozent Jungen. 70,1 Prozent von ihnen wachsen monolingual in ihren Familien auf, davon sprechen 56,6 Prozent deutsch und 13,5 Prozent eine andere Muttersprache. 29,8 Prozent der Kinder wachsen bilingual auf. Von ihnen sprechen 28,6 Prozent deutsch und eine weitere Sprache und 1,2 Prozent andere Sprachen ohne deutsch. Gemäß dem Ergebnis von Delfin 4 - Stufe 1 nahmen 50,2 Prozent der Kinder freiwillig an den Erhebungen zur Stufe 2 teil[71]. Bei 40,6 von ihnen war das Ergebnis im grünen und bei 9,6 Prozent im roten Bereich. Für 49,8 Prozent der Kinder war die Teilnahme an der zweiten Stufe obligatorisch. Ihr Ergebnis lag im gelben

[71] Bei 40,6 von ihnen war das Ergebnis im grünen und bei 9,6 Prozent im roten Bereich. Für 49,8 Prozent der Kinder war die Teilnahme an der zweiten Stufe obligatorisch. Ihr Ergebnis lag im gelben Bereich.

Bereich. Die Verteilung der Kinder auf die Schulamtsbezirke ist so, dass die Verteilung des Sozialindex gut abgebildet wird.

Zeitleiste

Normierungsstudie „BiP“ (2008)

am 7. Febr. 2008

Informationsveranstaltung

TU Dortmund: VertreterInnen des Delfin4 Teams, des MSW, des MGFFI, der beteiligten Schul- und Jugendämter

bis 14. Febr. 2008

Rückmeldungen der Schul- und Jugendämter

(per E-Mail an heidrun.besler@msw.nrw.de)

bis 18. Febr. 2008

Auswahl der beteiligten Lehrkräfte und KiTas durch Schul- und Jugendamt

(Weiterleitung der Informationsschreiben und Elternbriefe)

ab 18. Febr. 2008

- **Schulungen der TestleiterInnen**
  Verteilung der Test-Materialien)
- **Einholung des Einverständnisses der Erziehungsberechtigten**
  (durch die ErzieherInnen)

Durchführung

3. März 2008 bis 14. März 2008

Durchführung der Stufe 1 (Biz)

Achtung: verkürzter Zeitraum!

bis 11. April 2008

Durchführung der Stufe 2 (BiP)

Achtung: vorgezogener Zeitraum!

TestleiterInnen: ErzieherInnen oder Studentische Hilfskräfte + LehrerInnen:

(direkt im Anschluss an die Stufe 1)

- Kinder, deren Ergebnis der Stufe 1 im „gelben“ Bereich lag: **LehrerInnen**
- Kinder, deren Ergebnis der Stufe 1 im „roten“ oder „grünen“ Bereich lag: **ErzieherInnen oder Studentische Hilfskräfte**

spätestens am 11. April 2008

Abholung der Protokollbogen
(durch das Delfin4 Team - zentrale Sammelstellen)

|

bis Ende April 2008

Dateneineingabe

|

bis Anfang Mai 2008

Berechnung der Ergebnismatrix

Nützlichkeit

Früher als „Nebengütekriterium“ erachtet, wird der pädagogische Nutzen eines Sprachtests inzwischen als zentrales Qualitätsmerkmal angesehen. Wie „nützlich“ ein Sprachtest ist, wird u.a. daran gemessen, ob bzw. wieweit er eine praktische pädagogische Funktion erfüllt, die über bereits vorhandene Angebote hinausgeht. Beim Delfin 4-Sprachtest kommt die „Nützlichkeit“ u.a. darin zum Ausdruck, dass dieses Verfahren nicht isoliert dasteht, sondern Teil eines umfassenden Professionalisierungssystems ist, das Erzieher/innen und Lehrkräften - z.B. in Form von Handreichungen zur Sprachförderung, Elternarbeit und Selbst- bzw. Teamqualifizierung (im weiteren „Handreichung“ (vgl. Fried 2009) – Orientierung vermittelt, wie sich die Testergebnisse rational in adäquate pädagogische Maßnahmen übersetzen lassen (vgl. Fried, Briedigkeit & Schunder 2008). Damit ist unseres Wissens erstmals in Deutschland ein kombiniertes professionelles Instrumentarium zur Diagnose und Förderung allgemeiner akademischer Sprachkompetenz von Kindergartenkindern vorgelegt worden.

Die Handreichung soll dazu dienen, die Sprachkompetenz von Kindern im Rahmen des umfassenden Bildungsangebots sowohl integriert, als auch spezifisch und geplant zu fördern. Sie basiert – ebenso wie der zweistufige Delfin 4-Sprachtest - auf dem Konstruktionsrational. Dementsprechend beinhaltet sie folgende Basismodule zur Sprachförderung: Wortschatz, Morphosyntax, Erzählen und Phonembewusstheit. Dazu kommen noch die Zusatzmodule: Artikulation, Elternarbeit und Selbst- und Teamqualifizierung. Die Handreichung ist nicht als Rezeptbuch oder Trainingsprogramm angelegt, sondern als professionelle Orientierung, die einschlägiges Fachwissen, gestaltungsrelevante Hinweise und prototypische Beispiele beinhaltet, wie man von den Delfin 4-Testergebnissen zu Förderarrangements gelangt, die genau auf die Stärken und Schwächen des Kindes zielen und sowohl alltagsorientiert, als auch in spezifisch gestalteten Übungssituationen durchführbar sind. An der Entwicklung der Handreichung waren rund 50 Erzieherinnen, Fachberatungen und Lehrkräfte beteiligt. Deren Beobachtungen bzw. Anmerkungen bei der Praxiserprobung haben entscheidend zur Verbesserung der Handreichungen beigetragen.

Literatur

Barnett, W.S. (2001): Preschool education for economically disadvantaged children: Effects on reading achievement and related outcomes. In: S.B. Neumann/D.K. Dickinson (Eds.): Handbook of early literacy research (pp. 421-443). New York: Guilford.

Bishop, D.V.M. (2004): Expression, Reception and Recall of Narrative Instrument. ERRNI Manual. Oxford

Bradley, R.H. & Corwyn, R.F. (2002): Socioeconomic status and child development. In: Annual Review of Psychology 53, pp. 371-399.

Brown, R. (1973): A First Language. The Early Stages. Cambridge, Massachusetts, London: Harvard University Press.

Brunner, M. & Schöler, H. (2002): HASE – Heidelberger Auditives Screening in der Einschulungsuntersuchung (auch zum Einsatz in der Untersuchung U9). Wertingen, Westra.

Bühner, M. (2006[2]). Einführung in die Test- und Fragebogenkonstruktion. München: Pearson.

Frein, T., Möller, G., Petermann, A. & Wilpricht, M. (2006): Die empirische Seite: Bedarfsgerechte Stellen-zuweisung: Das neue Instrument Sozialindex. In: SchulVerwaltung Nr. 6, S. 188.

Fried, L. (2003): (Schrift-)Sprachfähigkeit als kulturelle Basiskompetenz von Kindergartenkindern? In: R. Arnold/H. Günther (Hrsg.): Innovative Bildungs- und Erziehungsprozesse. Kaiserslautern: Fachgebiet Pädagogik der Universität Kaiserslautern.

Fried, L. (2007): „Sprachstand 4“ misst Sprachentwicklungsstand. In: Kita aktuell 16 (3), S. 53-55.

Fried, L. (2008): Delfin 4: Diagnostik, Elternarbeit und Sprachförderung bei Vierjährigen in NRW. In: SchulVerwaltung 19 (11), S. 300 – 302.

Fried, L. (2008): Pädagogische Sprachdiagnostik für Vorschulkinder – Dynamik, Stand und Ausblick. In: Zeitschrift für Erziehungswissenschaft 10, Sonderheft 11/2008, 63-78.

Fried, L., Briedigkeit, E. & Schunder, R. (2008): Delfin 4 - Sprachförderorientierungen. Eine Handreichung. Düsseldorf, MGFFI NRW.

Fried, L., unter Mitarbeit von Briedigkeit, E., Isele, P. & Schunder, R. (2009): Testmanual Delfin 4 - Teil 1: Sprachkompetenzmodell. Dortmund, Technische Universität Dortmund, Lehrstuhl Pädagogik der frühen Kindheit.

Fried, L., Briedigkeit, E., Isele, P. & Schunder, R. (2009): Delfin 4 – Sprachkompetenzmodell und Messgüte eines Instrumentariums zur Diagnose, Förderung und Elternarbeit in Bezug auf die Sprachkompetenz vierjähriger Kinder. Zeitschrift für Grundschulforschung, 2 (2), 13–26.

Fried, L. & Briedigkeit, E. (2007): Delfin4 – Hintergründe und Einblicke zum neuen System der Sprachstandsfeststellung und –förderung. In: Kompakt Spezial (5), S. 10-11.

Fried, L., Briedigkeit, E., Isele, P. &Schunder, R. (2007): Notwendigkeit und Praxis von Sprachtests bei Kindergartenkindern. In: Städte- und Gemeinderat 61 (9), S. 16 - 19.

Fried, L., Briedigkeit, E. & Schunder, R. (2009): Sprachförderung gemäß Delfin 4. In: Kita aktuell NRW 18 (1), S. 8-11.

Gonzales, V. (2001): The role of socioeconomic and sociocultural factors in language minority children´s development: An ecological research view.

Gopnik, A. & Choi, S. (1995): Names, Relational Words, and Cognitive Development in English and Korean Speakers: Nouns are not always learned before verbs. In: M. Tomasello & W. Merriman (eds.): Beyond names for things: young children's acquisition oft verb (pp. 63-80). Hillsdale, N.J.: Erlbaum.

Grießhaber, W. (2002, modifiziert 2004 und 2009) ursprünglich entwickelt im Förderprojekt Deutsch und PC, ausgehend von: Clahsen, H. (1985) Profiling second language development: A procedure for assessing L2 proficiency. In: Hyltenstam, K. & Pienemann, M. (eds.) Modelling and Assessing Second Language Acquisition. Clevedon: Multilingual Matters, 283-331.

Grießhaber, W. (2004): Zwischenbericht der wissenschaftlichen Begleitung. März 2004. Projekt "Deutsch und PC". Münster, WWU Sprachenzentrum.

Grimm, H. (2003): SSV - Sprachscreening für das Vorschulalter. Kurzform des SETK 3-5. Göttingen: Hogrefe.

Grimm, H., unter Mitarbeit von Aktas, M. & Frevert, S. (2000): SETK-2. Sprachentwicklungstest für zweijährige Kinder. Manual. Göttingen: Hogrefe.

Halle, T., Calkins, J., Berry, D. & Johnson, R. (2003): Promoting language and literacy in early childhood care and education settings. In: Child Care & Early Education Research Connections (CCEERC), September 2003, pp. 2-17..

Hartung, J. & Elpelt, B. (2007). Multivariate Statistik. Lehr- und Handbuch der angewandten Statistik. München: Oldenbourg.

Hess, C./Landry, R. & Ritchie, K. (1984): The type-toke ratio and vocabulary performance. In: Psychological Reports, 55, pp. 51-57.

Hoff, E. (2003): The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. In: Child Development, 74,, pp. 1368-1378.

Hoff-Ginsberg, E. (2000): Soziale Umwelt und Sprachlernen. In: Grimm, H. (Hrsg.): Sprachentwicklung (S. 463-494). Göttingen: Hogrefe (= Enzyklopädie der Psychologie, C III, Bd. 3).

Kauschke, C. (2000): Der Erwerb des frühkindlichen Lexikons. Tübingen: Narr.

Kiese-Himmel, Ch. (2005): Aktiver Wortschatztest für 3- bis 5-jährige Kinder – Revision – (AWST-R). Göttingen, Hogrefe.

Klee, T. (1992): Developmental and diagnostic characteristics of quantitative measures of children's language production. In: Topics of Language Disorders, 12/2; pp. 28-41.

Landesamt für Datenverarbeitung und Statistik NRW (2007): Statistiken der Kinder- und Jugendhilfe. Teil III.1. Kinder und tätige Personen in Tageseinrichtungen am 15.03.2006. Dortmund: Technische Universität, Dortmunder Arbeitsstelle Kinder- und Jugendhilfestatistik.

Lienert, G.A. & Raatz, U. (1998[6]): Testaufbau und Testanalyse. Weinheim: Beltz.

Marjanovic-Umek, L, Kranjc, S. & Fekonja, U. (2002): Developmental levels of the child's storytelling. Paper presented at the Annual Meeting of the European Early Childhood Education Research Association EECERA (12th, Lefkosia, Cyprus, August 28-31, 2002).

Moosbrugger, H. & Kelava, A. (2007): Qualitätsanforderungen an einen psychologischen Test (Testgüte-kriterien). In: H. Moosbrugger & A. Klelava (Hrsg.), Testtheorie und Fragebogenkonstruktion (S. 8-26). Berlin: Springer.

Moosbrugger, H. & Kelava, A. (Hrsg.). (2007): Testtheorie und Fragebogenkonstruktion. Berlin: Springer

Richards (1987): Type/Token Rations: what do they really tell us? In: Journal of Child Language, 14, pp. 201-209.

Richards, B.J. & Malvern, D.D. (1997): Quantifying Lexical Diversity in the Study of Language Development. Reading: The New Bulmershe Papers.

Rosenthal Rollins, P., Snow, C. & Willett, J.B. (1996): Predictors of MLU: semantic and morphological developments. In: First Language, 16, pp. 243-259.

Rost, J. (1996): Lehrbuch Testtheorie, Testkonstruktion. Bern: Huber.

Schöler, H. & Schäfer, P. (2004): HASE. Heidelberger Auditives Screening in der Einschulungsuntersuchung – Itemanalysen und Normen. Heidelberg: Pädagogische Hochschule Heidelberg, Arbeitsbericht Nr. 17 aus dem Forschungsprojekt „Differenzialdiagnostik“.

Schöler, H. (2001): Sprachleistungsmessung im Schulalter. Ein Überblick. Heidelberg, Pädagogische Hochschule Heidelberg, Fakultät I – Psychologie in der Fachrichtung Lernbehindertenpädagogik, Arbeitsbericht Nr. 11 aus dem Forschungsprojekt „Differentialpädagogik“.

Tent, L. & Stelzl, I. (1993). Pädagogisch-psychologische Diagnostik. Band 1: Theoretische und methodische Grundlagen. Göttingen: Hogrefe.

Frein, T, Möller, G. Petermann, A. & Wilpricht, M. (2006):Die empirische Seite: Bedarfsgerechte Stellenzuweisung. Das neue Instrument Sizialindex. In: SchulVerwaltung, &, S. 188.

Watkins, R.V., Harbers, H., Hollis, W. & Kelly, D. (1995): Measuring Children’s Lexical Diversity: Differentiating Typical and Impaired Language Learners. In: Journal of Speech and Hearing Research, 38, pp. 1349-1355.

Wottawa, H. (1991[12]): Grundriß der Testtheorie. München: Juventa.

Anhang

Pilotierungsstudie zu Delfin 4 – Stufe 1: Besuch im Zoo (BiZ)

Evaluationsbogen

Liebe/r Begleiter/in, liebe/r Protokollant/in,

es ist uns wichtig, das Verfahren mit Ihrer Hilfe zu optimieren. Deshalb bitten wir Sie, uns Ihre Erfahrungen anhand der folgenden Fragen kurz zu vermitteln. Wenn Sie feststellen, dass die Fragen nicht alle (aus Ihrer Sicht wesentlichen) Ereignisse ansprechen, so sollten Sie den Raum bei „Sonstiges“ nutzen, um all das kurz festzuhalten.

Wir danken Ihnen jetzt schon für Ihren wichtigen Beitrag zur Optimierung des Verfahrens.

Zum thematischen Rahmen:

Hatten die Kinder Probleme mit dem thematischen Rahmen „Besuch im Zoo“?
Wenn ja, welche?

Zum Material:

Sind die Kinder mit dem Material klar gekommen?
Wenn nein, warum nicht?

Sind Sie selbst mit dem Material klar gekommen?
Wenn nein, warum nicht?

Haben Sie Ideen, wie man das Material noch verbessern könnte?

Zur Situation:

Haben Sie die Situation als natürlich empfunden?
Wenn nein, warum nicht?

Haben Sie die Situation als kindgemäß empfunden?
Wenn nein, warum nicht?

Haben Sie die Situation als praxisgerecht empfunden? Wenn nein, warum nicht?

Traten in der Situation besondere Probleme auf?
Wenn ja, welche?

Haben Sie Ideen, wie man die Situation noch verbessern könnte?

Zur Durchführungsanleitung:

Haben Sie sich durch die Durchführungsanleitung gut auf die Durchführung vorbereitet gefühlt?

Wenn nein, warum nicht?

Fanden Sie die Anleitung verständlich?
Wenn nein, warum nicht?

Haben Sie in der Anleitung Fehler entdeckt?
Wenn ja, welche?

Hat Ihnen in der Anleitung etwas gefehlt?
Wenn ja, was?

Haben Sie Ideen, wie man die Durchführungsanleitung noch verbessern könnte?

Sonstiges:

Pilotierungsstudie zu Delfin – Stufe 2: Besuch im Pfiffikus-Haus (BiP)

Evaluationsbogen

Liebe Testleiterinnen und Testleiter,

wir bedanken uns ganz herzlich für Ihr Engagement, das Screening der zweiten Stufe von Delfin 4 für unsere Pilotierungsstudie durchzuführen. Sie haben damit einen ganz wesentlichen Beitrag zur empirischen Überprüfung der Messqualität des Instrumentes beigetragen.
Es ist uns wichtig, das Verfahren mit Ihrer Hilfe zu optimieren. Deshalb bitten wir Sie, uns Ihre Erfahrungen anhand der folgenden Fragen kurz zu vermitteln. Wenn Sie feststellen, dass die Fragen nicht all (aus Ihrer Sicht wesentlichen) Ereignisse ansprechen, so sollten Sie den Raum bei „Sonstiges“ nutzen, um weitere Punkte festzuhalten.
Wir bitten Sie, den Bogen bis zum 15. Mai 2007 an uns zurück zu schicken und danken Ihnen schon jetzt für Ihre Unterstützung!

Für das Delfin 4 Team
gez. Prof. Dr. Lilian Fried

Testleiter/in: ○ Erzieher/in

○ Lehrer/in

Zum thematischen Rahmen:

Hatten die Kinder Probleme mit dem thematischen Rahmen „Besuch im Pfiffikus-Haus?“

Wenn ja, welche?

Zum Material:

Sind die Kinder gut mit dem Material klar gekommen?
Wenn nein, warum nicht?

Haben Sie Ideen, wie man das Material noch verbessern könnte?

Zur Situation:

Haben Sie die Situation als natürlich empfunden? Wenn nein, warum nicht?

Haben Sie die Situation als kindgemäß empfunden?

Wenn nein, warum nicht?

Haben Sie die Situation als praxisgemäß empfunden?

Wenn nein, warum nicht?

~~Traten in der Situation besondere Probleme auf?~~

~~Wenn ja, welche?~~

Haben Sie Ideen, wie man die Situation noch verbessern könnte?

Zur Handanweisung:
Haben Sie sich mit der Handanweisung gut auf die Durchführung vorbereitet gefühlt?

Wenn nein, warum nicht?

Fanden Sie die Handanweisung verständlich?

Wenn nein, warum nicht?

Haben Sie in der Handanweisung Fehler entdeckt?

Wenn ja, welche?

Hat Ihnen in der Handanweisung etwas gefehlt?

Wenn ja, was?

Haben Sie Ideen, wie man die Handanweisung noch verbessern könnte?

Zur Protokollierung:

Gab es bei der Protokollierung Proble-

me? Wenn ja, welche?

Sonstiges:

Schichtung Normierungsstichprobe

1. Sozialindex: 0 bis unter 20 (Gruppe 6)

	Schulamt	Kinderzahl der Stichpro-be	Anteil in % der Stichpro-be	Sozialindex	Sozialindex Gruppe	Zahl der Kin-der, die an der ersten Stufe teilgenommen haben	Anteil der Kin-der in %, die an der ersten Stufe teilge-nommen ha-ben	Anzahl der zum 01.08.2009 schul-pflichtigen Kinder	Anteil in %, der zum 01.08.2009 schul-pflichtigen Kinder
1.	Kreis Kleve	1447	4,4	0,0	6	2783	1,8	3198	1,8
2.	Kreis Coesfeld	72	0,2	2,1	6	2217	1,4	2437	1,4
3.	Kreis Borken	3609	11,1	3,5	6	3875	2,5	4164	2,4
4.	Kreis Euskirchen	1680	5,2	7,9	6	1589	1,0	1956	1,1
5.	Kreis Steinfurt	6	0	9,4	6	4351	2,8	4990	2,8
6.	Kreis Viersen	0	0	9,9	6	2056	1,3	2897	1,6
7.	Kreis Olpe	1234	3,8	10,4	6	1222	0,8	1462	0,8
8.	Rhein-Sieg-Kreis	0	0	14,0	6	5576	3,6	6355	3,6
9.	Kreis Heinsberg	35	0,1	16,5	6	2288	1,5	2635	1,5
10.	Kreis Höxter	0	0	16,7	6	1389	0,9	1579	0,9
11.	Kreis Warendorf	60	0,2	17,2	6	2800	1,8	3076	1,7
12.	Rheinisch-Bergischer Kreis	122	0,4	19,0	6	2570	1,6	2974	1,7
			25,4				**21,0**		**21,3**

2. Sozialindex: 20 bis unter 40 (Gruppe 5)

	Schulamt	Kinderzahl der Stichpro-be	Anteil in % der Stichpro-be	Sozialindex	Sozialindex Gruppe	Zahl der Kin-der, die an der ersten Stufe teilgenommen haben	Anteil der Kinder in %, die an der ersten Stufe teilgenommen haben	Anzahl der zum 01.08.2009 schul- pflich-tigen Kinder	Anteil in %, der zum 01.08.2009 schul-pflichtigen Kinder
13.	Rhein-Kreis Neuss	162	0,5	20,3	5	4194	2,7	4474	2,5
14.	Kreis Düren	43	0,1	22,4	5	2367	1,5	2635	1,5
15.	Kreis Siegen-Wittgenstein	24	0,1	22,7	5	2356	1,5	2795	1,6
16.	Rhein-Erft-Kreis	1374	4,2	22,9	5	4218	2,7	4874	2,8
17.	Hochsauerlandkreis	19	0,1	23,2	5	2594	1,7	2861	1,6
18.	Kreis Wesel	134	0,4	23,3	5	3688	2,4	4236	2,4
19.	Kreis Soest	0	0	24,4	5	2946	1,9	3382	1,9
20.	Kreis Gütersloh	32	0,1	25,2	5	3378	2,2	4004	2,3
21.	Oberbergischen Kreis	0	0	26,1	5	2402	1,6	3023	1,7
22.	Kreis Mettmann	117	0,4	28,6	5	4566	3,0	4885	2,8
23.	Kreis Minden-Lübbecke	27	0,1	29,4	5	2877	1,9	3377	1,9
24.	Stadt Münster	0	0	31,3	5	2255	1,5	2569	1,5
25.	Kreis Paderborn	2825	8,7	31,6	5	2865	1,9	3314	1,9
26.	Kreis Aachen	0	0	33,8	5	2718	1,8	3268	1,8
27.	Kreis Herford	0	0	34,1	5	2392	1,6	2550	1,4
28.	Stadt Bonn	2112	6,5	36,4	5	2772	1,8	3161	1,8
29.	Ennepe-Ruhr-Kreis	77	0,2	37,0	5	2820	1,8	3101	1,8
30.	Stadt Leverkusen	17	0,1	38,3	5	1309	0,9	1599	0,9
31.	Kreis Lippe	135	0,4	39,3	5	3020	2,0	3767	2,1
			21,9				**36,4**		**36,2**

3. Sozialindex: 40 bis unter 50 (Gruppe 4)

	Schulamt	Kinderzahl der Stichprobe	Anteil in % der Stichprobe	Sozialindex	Sozialindex Gruppe	Zahl der Kinder, die an der ersten Stufe teilgenommen haben	Anteil der Kinder in %, die an der ersten Stufe teilgenommen haben	Anzahl der zum 01.08.2009 schulpflichtigen Kinder	Anteil in %, der zum 01.08.2009 schulpflichtigen Kinder
32.	Stadt Mülheim an der Ruhr	643	2,0	41,5	4	1450	0,9	1616	0,9
33.	Kreis Unna	69	0,2	45,1	4	3447	2,2	3901	2,2
34.	Märkischer Kreis	460	1,4	47,9	4	3717	2,4	4385	2,5
35.	Stadt Hamm	1044	3,2	49,4	4	1626	1,1	1838	1,0
			6,8				**6,6**		**6,6**

4. Sozialindex: 50 bis unter 60 (Gruppe 3)

	Schulamt	Kinderzahl der Stichprobe	Anteil in % der Stichprobe	Sozialindex	Sozialindex Gruppe	Zahl der Kinder, die an der ersten Stufe teilgenommen haben	Anteil der Kinder in %, die an der ersten Stufe teilgenommen haben	Anzahl der zum 01.08.2009 schulpflichtigen Kinder	Anteil in %, der zum 01.08.2009 schulpflichtigen Kinder
36.	Stadt Bottrop	944	2,9	50,3	3	985	0,6	1086	0,6
37.	Kreis Recklinghausen	202	0,6	50,5	3	5469	3,6	6016	3,4
38.	Stadt Krefeld	1355	4,2	51,4	3	1760	1,1	2177	1,2
39.	Stadt Solingen	10	0	51,7	3	1304	0,8	1562	0,9
40.	Stadt Aachen	0	0	53,0	3	1880	1,2	2200	1,2
41.	Stadt Remscheid	450	1,4	54,3	3	964	0,6	1124	0,6
42.	Stadt Bochum	1148	3,5	54,9	3	2836	1,8	3118	1,8
43.	Stadt Mönchengladbach	98	0,3	59,2	3	2036	1,3	2484	1,4
			12,9				**11,0**		**11,1**

5. Sozialindex: 60 bis unter 75 (Gruppe 2)

	Schulamt	Kinderzahl der Stichprobe	Anteil in % der Stichprobe	Sozialindex	Sozialindex Gruppe	Zahl der Kinder, die an der ersten Stufe teilgenommen haben	Anteil der Kinder in %, die an der ersten Stufe teilgenommen haben	Anzahl der zum 01.08.2009 schulpflichtigen Kinder	Anteil in %, der zum 01.08.2009 schulpflichtigen Kinder
44.	Stadt Düsseldorf	4243	13,0	62,8	2	4380	2,9	5341	3,0
45.	Stadt Oberhausen	884	2,7	64,7	2	1510	1,0	1929	1,1
46.	Stadt Essen	141	0,4	65,9	2	4436	2,9	5077	2,9
47.	Stadt Herne	1	0	70,6	2	1211	0,8	1500	0,8
48.	Stadt Wuppertal	25	0,1	71,3	2	2747	1,8	3205	1,8
49.	Stadt Köln	0	0	72,2	2	8357	5,4	9212	5,2
50.	Stadt Bielefeld	489	1,5	73,1	2	2709	1,8	3242	1,8
			17,7				**16,5**		**16,6**

6. Sozialindex: 75 und mehr (Gruppe 1)

	Schulamt	Kinderzahl der Stichprobe	Anteil in % der Stichprobe	Sozialindex	Sozialindex Gruppe	Zahl der Kinder, die an der ersten Stufe teilgenommen haben	Anteil der Kinder in %, die an der ersten Stufe teilgenommen haben	Anzahl der zum 01.08.2009 schulpflichtigen Kinder	Anteil in %, der zum 01.08.2009 schulpflichtigen Kinder
51.	Stadt Dortmund	3016	9,3	75,0	1	4715	3,1	5506	3,1
52.	Stadt Hagen	0	0	75,6	1	1656	1,1	1854	1,0
53.	Stadt Duisburg	28	0,1	77,7	1	3996	2,6	4439	2,5
54.	Stadt Gelsenkirchen	1966	6,0	100,0	1	2258	1,5	2530	1,4
			15,4				**8,3**		**8,0**

	alle Schulämter	**32579**	**100,1**	**NRW 38,8**	**5**	**159302**	**99,8**	**176940**	**99,8**

Nordrhein-Westfalen - Statistik der Kinder- und Jugendhilfe

1. Anzahl der Kinder im Alter von 3 bis 5 Jahren

Altersgruppe	Anzahl	Anteil
Anzahl 3- bis 4jährige	111.708	42,00%
Anzahl 4- bis 5jährige	154.271	58,00%
Gesamt:	**265.979**	100,00%

2. Geschlecht

Geschlecht	3- bis 4jährige	Anteil der 3- bis 4jährigen	Anteil der 3- bis 5jährigen
männlich	57.050	**51.07%**	21,45%
weiblich	54.658	**48,93%**	20,55%
Gesamt:	**111.708**	100,00%	42,00%

Geschlecht	4- bis 5jährige	Anteil der 4- bis 5jährigen	Anteil der 3- bis 5jährigen
männlich	78.946	**51,17%**	29,68%
weiblich	75.325	**48,83%**	28,32%
Gesamt:	**154.271**	100,00%	58,00%

Geschlecht	3- bis 5jährige	Anteil der 3- bis 5jährigen
männlich	135.996	**51,13%**
weiblich	129.983	**48,87%**
Gesamt:	**265.979**	100,00%

3. Vorrangige Familiensprache

Familiensprache	3- bis 4jährige	Anteil der 3- bis 4jährigen	4- bis 5jährige	Anteil der 4- bis 5jährigen	3- bis 5jährige	Anteil der 3- bis 5jährigen
deutsch	89.319	**79,96%**	122.447	**79,37%**	211.766	**79,62%**
nicht -	22.389	**20,04%**	31.824	**20,63%**	54.213	**20,38%**
Gesamt:	**111.708**	100%	**154.272**	100,00%	**265.979**	100.00%

**Normierungstabelle/Entscheidungsmatrix zu Delfin 4 – Stufe 1 –
Test „Besuch im Zoo“**

Anleitung zur Entscheidungsfindung

Der Test „Besuch im Zoo (BiZ)“ liefert die Grundlage für die Entscheidung, ob für ein Kind die Sprachstandsfeststellung nach § 36 Absatz 2 Schulgesetz abgeschlossen ist, oder ob das Kind ab Mai mit dem vertiefenden Einzeltest „Besuch im Pfiffikus-Haus“ erneut getestet wird, um festzustellen, ob die Sprachentwicklung altersgemäß ist und ob es die deutsche Sprache hinreichend beherrscht.

Bei der Auswertung von „Besuch im Zoo“ sind drei Fallkonstellationen – je nach Ergebnis – möglich:

- Das Kind benötigt keine zusätzliche pädagogische Sprachförderung („grün“). Die Sprachstandsfeststellung ist beendet.
- Das Testergebnis lässt noch keine Aussage über die Notwendigkeit einer zusätzlichen pädagogischen Sprachförderung zu („gelb“). Das Kind wird zu einem späteren Zeitpunkt mit „Besuch im Pfiffikus-Haus“ erneut getestet.
- Das Testergebnis legt eine zusätzliche pädagogische Sprachförderung nahe („rot“). Wird diese Einschätzung von den pädagogischen Fachkräften der Kindertageseinrichtung geteilt, so wird die Notwendigkeit dieser Förderung als Testergebnis bescheinigt. Die Sprachstandsfeststellung ist beendet. (Allerdings haben die Eltern das Recht, ihr Kind dennoch zum Test mit „Besuch im Pfiffikus-Haus“ anzumelden.)

Unterlagen

Zur Ermittlung eines dieser Ergebnisse sind folgende Unterlagen notwendig:

1. Protokollheft
2. Umrechnungstabellen
3. Entscheidungstabelle
4. Auswertungsraster
5. evtl. Ergebnisse systematischer Beobachtung im Alltag der Kindertageseinrichtung bei der Gruppe der „roten“ Kinder, sofern die Einverständniserklärung der Eltern über den Austausch zwischen ErzieherInnen und Lehrkräften vorliegt.

Schritte der Entscheidungsfindung

Die Entscheidungsfindung umfasst folgende Schritte:

1. Schritt	Rohpunktwerte der Untertests ermitteln (RW)
2. Schritt	Standardwerte der Untertests ermitteln (TW)
3. Schritt	Durchschnittsstandardwert des Gesamttests ermitteln (DTW)
4. Schritt	Einordnung des DTW in

1. Schritt:
Die Rohpunktwerte (RW) jedes der vier Untertests werden dem Protokollheft des einzelnen Kindes entnommen und in das Auswertungsraster übertragen. Hat ein Kind einen/mehrere Untertest/s verweigert, erhält es hierfür 0 Rohpunktwerte. In das Auswertungsraster wird ebenfalls das Alter des Kindes am Testtag notiert (jünger als vier Jahre bzw. vier Jahre und älter).

2. Schritt:
Die Rohpunktwerte werden in Standardwerte (TW) umgewandelt. Anhand der Umrechnungstabellen lässt sich ablesen, welche Standardwerte (TW) den vier Rohpunktwerten der jeweiligen Untertests entsprechen. Dabei ist je nach dem Alter des getesteten Kindes unbedingt zwischen der Tabelle mit den Altersnormen für jüngere Kinder unter vier Jahren und der Tabelle mit den Altersnormen für vierjährige Kinder zu unterscheiden.

Da die empirischen Berechnungen ergeben haben, dass sich die Sprachleistungen der jüngeren Kinder (unter vier Jahren) von denen der älteren Kinder (über vier Jahre) bedeutsam unterscheiden, wurden altersspezifische Umrechnungstabellen (Altersnormen) entwickelt. Es hängt vom Alter des Kindes am Testtag ab, aus welcher Umrechnungstabelle man jeweils den T-Wert entnimmt.

3. Schritt:
Die Standardwerte (TW) der vier Untertests (HA-TW, KN-TW, SN-TW, BB-TW) werden addiert. Die Summe dividiert durch 4 ergibt den durchschnittlichen Standardwert des Gesamttests (DTW).[1]

4. Schritt:
Der DTW wird mit den farbig markierten (rot, gelb, grün) Zeilen der Entscheidungstabelle verglichen. Damit ist die Entscheidung über das Ergebnis des Tests „Besuch im Zoo“ möglich.

1 Berechungsformel: (HA-TW + KN-TW + SN-TW + BB-TW) : 4

Umrechnungstabellen: Altersnormen für jüngere Kinder (unter vier Jahre)[2]

HA (jünger) RW	TW
0	30
1	32
2	34
3	36
4	37
5	39
6	42
7	45
8	47
9	52
10	55
11	75

KN (jünger) RW	TW
0	32
1	35
2	39
3	41
4	45
5	49
6	54
7	61
8	75

BB (jünger) RW	TW
0	33
1	40
2	43
3	45
4	47
5	50
6	53
7	56
8	59
9	62
10	65
11	67
12	75

[2] Basis: (1) Normierungsstichprobe von 2007 (N = 14 859); die in Bezug auf Geschlecht und Sozialindex von Schulaufsichtsbezirken die Verteilung in der Grundgesamtheit (alle Kinder, mit denen Delfin 4 Stufe 1 durchgeführt wurde) repräsentiert; (2) Normierungsstichprobe von 2008 (N = 1 451), die Kinder aus dem Kreis Recklinghausen umfasst.

SN (jün-ger) RW	**TW**
0	34
1	36
2	37
3	
4	38
5	39
6	40
7	42
8	43
9	
10	44
11	
12	45
13	46
14	
15	47
16	48
17	
18	49
19	50
20	51
21	52
22	
23	53
24	54
25	55
26	56
27	57
28	59
29	60
30	63
31	75

Umrechnungstabellen: Altersnormen für ältere Kinder (ab 4 Jahren)[3]

HA (älter) RW	TW
0	25
1	29
2	30
3	31
4	34
5	36
6	38
7	42
8	44
9	48
10	52
11	75

KN (älter) RW	TW
0	30
1	33
2	37
3	39
4	42
5	46
6	51
7	58
8	75

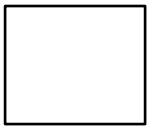
BB (älter) RW	TW
0	31
1	37
2	39
3	42
4	44
5	48
6	50
7	53
8	56
9	59
10	62
11	65
12	75

[3] Basis: (1) Normierungsstichprobe von 2007 (N = 14 859); die in Bezug auf Geschlecht und Sozialindex von Schulaufsichtsbezirken die Verteilung in der Grundgesamtheit (alle Kinder, mit denen Delfin 4 Stufe 1 durchgeführt wurde) repräsentiert; (2) Normierungsstichprobe von 2008 (N = 2 185), die Kinder aus dem Kreis Recklinghausen umfasst.

SN (älter) RW	**TW**
0	30
1	31
2	32
3	33
4	34
5	35
6	36
7	37
8	
9	38
10	39
11	
12	40
13	41
14	42
15	43
16	
17	44
18	
19	45
20	46
21	47
22	48
23	
24	49
25	50
26	52
27	53
28	55
29	57
30	61
31	75

Berechnung des „Durchschnittlichen TW“:

(HA-TW + KN-TW + BB-TW + SN-TW) : 4 = DTW



Entscheidungstabelle

Niveau	DTW		Entscheidung
Rot	**20,0 – 33,7**		zusätzliche SF[4]; keine Stufe 2
Gelb	**33,8 – 47,7**		Delfin 4 Stufe 2
Grün	**47,8 - 75,0**		keine zusätzliche Sprachförderung

[4] SF = Sprachförderung

Auswertungsraster

Name:		Alter:
Untertest	Rohpunktwert (RW) (Protokollheft)	Standardwert (TW)
Handlungsanweisungen ausführen(HA)		
Kunstwörter nachsprechen (KN)		
Bild beschreiben (BB)		
Sätze nachsprechen (SN)		
Summe der T-Werte:		
Ergebniswert (DTW) (Summe der T-Werte geteilt durch vier):		

Auswertungsraster - Beispiele

Beispiel 1: Kind im Alter über vier Jahre

Name: Max Mustermann		Alter: über vier Jahre
Untertest	Rohpunktwert (RW) (Protokollheft)	Standardwert (TW)
Handlungsanweisungen ausführen (HA)	7	42
Kunstwörter nachsprechen (KN)	5	46
Bild beschreiben (BB)	4	44
Sätze nachsprechen (SN)	20	46
Summe der T-Werte:		178
Ergebniswert (DTW) (Summe der T-Werte geteilt durch vier):		44,5

Nach der Entscheidungstabelle liegt der DTW 44,5 im „gelben“ Bereich. Max Mustermann wird zur 2. Stufe des Verfahrens eingeladen.

Beispiel 2: Kind im Alter unter vier Jahren

Name: Marie Musterfrau		Alter: unter vier Jahren
Untertest	Rohpunktwert (RW) (Protokollheft)	Standardwert (TW)
Handlungsanweisungen ausführen (HA)	7	45
Kunstwörter nachsprechen (KN)	5	49
Bild beschreiben (BB)	4	47
Sätze nachsprechen (SN)	20	51
Summe der T-Werte:		192
Ergebniswert (DTW) (Summe der T-Werte geteilt durch vier):		48

Nach der Entscheidungstabelle liegt der DTW 48 im „grünen“ Bereich. Marie Musterfrau benötigt keine zusätzliche pädagogische Sprachförderung

.

Fachliche Grundlagen

Nur wenn ein Verfahren normiert ist, kann fundiert entschieden werden, ob ein Kind zusätzlichen Sprachförderbedarf hat oder nicht. Wie die Normierung von „Delfin 4 – Stufe 1: Besuch im Zoo (BiZ)“ vollzogen wurde, wird nachfolgend kurz zusammengefasst.

Definition:
Die Normierung (Eichung) eines Tests erfordert die Festlegung eines Bezugssystems, das es ermöglicht, die Rohpunktwerte, die ein Kind bei BiZ erreicht hat, im Vergleich zu den Rohpunktwerten anderer Kinder einzuordnen und zu interpretieren. Normen sind also nichts anderes als Vergleichswerte. Um solche Vergleichswerte zu erhalten, muss ein einheitlicher Maßstab (Norm) bestimmt werden. Anhand solcher Normen lässt sich angeben, welche Position ein Kind bezüglich der Werte anderer Kinder einnimmt (vgl. z.B. Lienert & Raatz, 1998).

Stichprobe:
Bei der Normierung kommt es entscheidend darauf an, möglichst aussagekräftige Vergleichswerte von solchen Kindern zu erhalten, die den mit BiZ zu untersuchenden Kindern hinsichtlich relevanter Merkmale (z.B. Alter, Geschlecht, Sozialstatus, Muttersprache) ähnlich sind (Moosbrugger & Kelava, 2007, S. 19). Die Normierung muss deshalb auf der Basis einer möglichst großen Stichprobe erfolgen, welche die Grundgesamt (hier alle Kinder in Nordrhein-Westfalen zwei Jahre vor Schulbeginn) in Bezug auf relevante Merkmale möglichst weitgehend repräsentiert.

Dem wurde in zwei aufeinanderfolgenden Jahren durch folgende Bemühungen Rechnung getragen:

Im Jahr 2007 wurde eine Normierungsstichprobe gewonnen, die sich aus Kindern mit deutscher Muttersprache[5] zwei Jahre vor Schulbeginn zusammensetzt. Sie umfasst insgesamt 14.859 Kinder. Im Durchschnitt sind diese 4 Jahre, 1 Monat alt (Standardabweichung: 4 Monate). Die jüngsten Kinder sind 3 Jahre, 6 Monate, die ältesten 4 Jahre, 10 Monate alt.

Die Repräsentativität einer Normierungsstichprobe wird gemeinhin dadurch gewährleistet, dass sie gezielt nach relevanten Faktoren geschichtet wird (Schichtungs- oder Quotierungsstichprobe). Die Normierungsstichprobe von 2007 ist in Bezug auf zwei potentiell sprachentwicklungsbedeutsame Faktoren repräsentativ geschichtet: Geschlecht der Kinder sowie Sozialindex des Schulaufsichtsbezirks, in dem die Kinder getestet werden.

Was das Geschlecht betrifft, so ist sich die Wissenschaft nicht einig, ob und wieweit es bedeutsame Differenzen in der Sprachentwicklung von Jungen und Mädchen gibt, weil widersprüchliche Befunde vorliegen; „..., wobei nicht klar ist, ob das daran liegt, dass keine Analysen durchgeführt wurden, oder daran, dass die Ergebnisse keine Unterschiede

5 Die Beschränkung auf Kinder mit deutscher Muttersprache ergibt sich aus einer der beiden Funktionen des zweistufigen Screeningverfahrens „Delfin 4 – Stufe 1 und 2“, nämlich zu überprüfen, ob ein Kind zwei Jahre vor der Schule hinreichend Deutsch beherrscht. Das kann aber nur angemessen entschieden werden, wenn man sich am Sprachkompetenzlevel von Kindern mit deutscher Muttersprache orientiert.

aufzeigten.“ (Hoff-Ginsberg, 2000, S. 479). Dort, wo in Forschungen Unterschiede zutage traten, wie z.B. bei Grimm, Aktas und Frevert (2000) sowie Schöler und Schäfer (2004) der Fall, zeigten sich die Mädchen den Jungen überlegen. Vor diesem Hintergrund haben wir Jungen und Mädchen in der Normierungsstichprobe von 2007 so ins Verhältnis gesetzt, wie in der Statistik der Kinder- und Jugendhilfe Nordrhein-Westfalen für das Jahr 2006 angegeben (Landesamt für Datenverarbeitung und Statistik NRW, 2007) dargestellt. Dementsprechend setzt sich die Normierungsstichprobe aus 51,2 Prozent Jungen und 48,9 Prozent Mädchen zusammen.

Dass die Sprachentwicklung von Kindern bedeutsam durch sozioökonomische Faktoren bestimmt wird, ist in der Wissenschaft unumstritten (Barnett 2001; Bradley/Corwyn 2002; Fried 2003; Gonzales 2001; Halle et al. 2003; Hoff 2003; Hoff-Ginsberg 2000). Deshalb haben wir die Normierungsstichprobe von 2007 auch nach Sozialindex geschichtet. Es handelt sich dabei um ein Maß, das die soziale Belastung von Schulaufsichtsbezirken (d.h. Kreise bzw. kreisfreie Städte) anzeigt. Es basiert auf den vier soziografischen Merkmalen: Arbeitslosenquote, Sozialhilfequote, Migrantenquote, Sozialhilfequote (Ausländer und Aussiedler) und Quote der Wohnungen in Einfamilienhäusern. Die einzelnen Merkmale wurden in einem statistischen Verfahren (Faktorenanalyse) zur Skala „Sozialindex“ zusammengefasst. Der Konstruktion des Sozialindex liegt die Annahme zugrunde, dass Kreise mit vielen Arbeitslosen, Sozialhilfeempfängern und Migranten sowie einem geringen Anteil von Wohnungen in Einfamilienhäusern sozial stärker belastet sind als Kreise mit einer einheimischen, erwerbstätigen, den Lebensunterhalt selbständig bestreitenden und Einfamilienhäuser bewohnenden Bevölkerung (vgl. Frein/Möller/Petermann/Wilpricht 2006).

Da sich der Sozialindex auf den ganzen Kreis bzw. die kreisfreie Stadt bezieht, in dem bzw. in der das Kind getestet wird, handelt es sich dabei um kein individuelles Merkmal. Dennoch darf der Faktor als aussagekräftig gelten, weil belegt werden konnte, dass bedeutsame Zusammenhänge zwischen Sozialindex und Bildungschancen bzw. -erfolgen der Kinder eines Schulaufsichtsbezirks bestehen (ebd., S. 188).

Wie die folgende Tabelle zeigt, bildet die Normierungsstichprobe von 2007 fast perfekt die Sozialindexverteilung der Kinder beim ersten Testdurchgang 2007 ab.

Sozial-index	Nordrhein-Westfalen	Prozentanteil	Normierungs-stichprobe	Prozentanteil
6	32.716	21,3	3.129	21,1
5	55.755	36,2	5.412	36,4
4	10.240	6,7	983	6,6
3	17.234	11,2	1.639	11,0
2	25.350	16,5	2.459	16,5
1	12.625	8,2	1.237	8,3
Summe	153.920	100,1*	14.859	99,9*

* = Rundungsfehler;

Im Jahr 2008 haben wir die Normierungsbemühungen mittels einer – auf den Kreis Recklinghausen beschränkten – weiteren Normierungsstichprobe erweitert. Diese setzt sich ebenfalls aus Kindern mit deutscher Muttersprache zwei Jahre vor Schulbeginn zusammen. Sie umfasst insgesamt 3.654 Kinder. Im Durchschnitt sind diese 4 Jahre, 3 Monate alt (Standardabweichung: 2 Monate). Die jüngsten Kinder sind 3 Jahre, 3 Monate, die ältesten 5 Jahre, 1 Monat alt. Das Verhältnis Jungen: Mädchen beträgt 50,6% : 49,4%.

Die Normentabellen basieren auf beiden Normierungsstichproben.

Normierung:
Mit „Normierung“ bezeichnet man das Berechnen von Kennzahlen, die das Verhältnis jedes einzelnen Rohpunktwertes zu den Ergebnissen der Normierungsstichprobe zum Ausdruck bringen (Wottawa, 1991, S.102). Dazu wird für jedes Kind der Normierungsstichprobe die Zahl der Punkte ermittelt, die es für seine „Lösungen“ erhält. Diese „Rohwerte“ sind nur bedingt aussagekräftig, denn sie ermöglichen nur Aussagen im Hinblick auf diese eine Stichprobe. Auch können sie nicht ohne weiteres verglichen werden. So markiert z.B. der Rohwert 6 im Untertest „Kunstwörter nachsprechen“ des BiZ ein ganz anderes Leistungsniveau, als der Rohwert 6 im Untertest „Sätze nachsprechen“ des BiZ.

Deshalb werden die Rohpunktwerte in Standardwerte transformiert, d.h. auf einer einheitlichen Skala abgebildet und so vergleichbar gemacht. Es gibt verschiedene Möglichkeiten, Rohpunktwerte statistisch in Standardwerte zu transformieren, wie z.B. Prozenträngе, T-Werte, IQ-Werte, C-Werte, z-Werte und Stanine-Werte (vgl. z.B. Tent & Stelzl, 1993, S. 58). Welche dieser Möglichkeiten in Frage kommen, hängt nicht zuletzt davon ab, ob die Normierungsstichprobe normalverteilt ist[6]. Da dies (wie bei den meisten anderen Sprachtests auch) bei der BiZ-Normierungsstichprobe nicht gegeben ist, wurden nur die für diesen Fall vorgesehenen Standardnorm-Äquivalente berechnet.

Die dafür erforderlichen statistischen Transformationen gehen von der verteilungsfreien Prozentrangskala aus. Es handelt sich dabei um eine Rang- oder Originalskala. Die Prozentrangplätze werden über die kumulativen Häufigkeiten errechnet. Sie markieren die relative Position eines Kindes im Vergleich mit der Rangordnung der Kinder der Normierungsstudie. Wenn z.B. ein Kind bei BiZ einen Prozentrang (PR) von 75 hat, weiß man, dass seine Sprachkompetenz gleich gut oder besser ist als die von 75 Prozent der Normierungsstichprobe; oder anders formuliert: nur 25 Prozent der Kinder der Normierungsstichprobe besitzen eine noch höhere Sprachkompetenz (vgl. Moosbrugger & Kelava, 2007, S. 19). Der Nachteil von Prozentrangnormen ist, dass sie im mittleren Bereich erheblich feiner differenzieren als in den Extrembereichen. Da BiZ vor allem dazu dient,

[6] Nichtnormalverteilte Merkmale können durch eine „Flächentransformation“ normalisiert, das meint: in sogenannte Standardnormäquivalente transformiert werden (vgl. Lienert & Raatz, 1998; Moosbrugger & Kelava, 2007).

Deshalb wurden die Prozentrangnormen in Standardnorm-Äquivalente transferiert[7]. Diese geben jeweils an, wie weit ein Rohpunktwert vom arithmetischen Mittelwert der Normierungsstichprobe entfernt ist. Dabei wird diese Differenz in Form von Standardabweichungen[8] ausgedrückt.
Das bei Sprachtests ohne Normalverteilung gebräuchlichste (und auch bei BiZ verwendete) Standardnorm-Äquivalent ist die T-Wert-Skala (TW). Sie hat einen definierten Mittelwert von TW 50 und eine definierte Standardabweichung von TW 10. Per Definition markieren 40 bis 60 TW eine durchschnittliche und TW über 60 TW eine überdurchschnittliche Leistung. Ein TW unter 40 indiziert eine Risikoentwicklung.

Gruppennormen:
In der Fachliteratur wird zwischen Gesamtnormen und Gruppennormen unterschieden (Lienert & Raatz, 1994). Von Gesamtnormen spricht man, wenn die Normierung über die Gesamtpopulation der Normierungsstichprobe erfolgt, von Gruppennormen, wenn sie über eine oder mehrere Teilpopulationen erfolgt. Gruppennormen sind dann angebracht, wenn die Gesamtpopulation der Normierungsstichprobe aus Teilpopulationen zusammengesetzt ist, die sich bezüglich des relevanten Merkmals systematisch unterscheiden. Ob das der Fall ist, lässt sich anhand von Varianzanalysen ermitteln. Mit diesem Verfahren wird getestet, ob die Varianz zwischen bestimmten Gruppen größer ist als die Varianz innerhalb der Gruppen. Dadurch kann ermittelt werden, ob sich bestimmte Gruppen statistisch signifikant unterscheiden oder ob das nicht der Fall ist.

Varianzanalytische Berechnungen anhand der BiZ-Normierungsstichprobe ergaben, dass es sowohl in Bezug auf Geschlecht, als auch in Bezug auf das Alter statistisch signifikante Unterschiede gibt.

Den statistisch bedeutsamen Differenzen zwischen den Kindern, die unter vier Jahre und denen, die vier Jahre und älter sind (vgl. Anhang), wird durch dementsprechende Altersnormen entsprochen (vgl. Normentabellen). Dagegen werden – ungeachtet der ebenfalls statistisch bedeutsamen Unterschiede zwischen Mädchen und Jungen (vgl. Anhang) – keine Geschlechtsnormen angeboten. Das hat zwei Gründe: Zum einen variieren die Leistungsunterschiede zwischen den Geschlechtern nur in einem solchen Ausmaß, dass sich eine getrennte Normierung erübrigt. Zum anderen ist der praktische Nutzen von Geschlechtsnormen fraglich, weil eine nach Geschlechtern getrennte Sprachförderung pädagogisch keinen Sinn macht.

Risikoniveau:
BiZ hat eine Screeningfunktion[9], d.h. es soll dazu dienen, belastbare Informationen für die Entscheidung zu liefern, welche Kinder zwei Jahre vor der Schule eine zusätzliche Sprachförderung erhalten sollten, damit ihre Bildungschancen zum Schulstart möglichst gewahrt bleiben. Lienert und Raatz (1998) warnen davor, bei derart weitreichenden Konsequenzen allzu feine Normen, zu verwenden, weil diese dazu verführen, die Messgenauigkeit von Tests zu überschätzen; sie empfehlen statt dessen, sich an gröberen Normen auszurichten. Deswegen nutzen wir die feinen T-Wert-Normen nur für

[7] Die Transformation von Prozentrangwerten in T-Werte wurde anhand statistischer Umrechnungstabellen vorgenommen (vgl. z.B. Bühner 2006, S. 106).
[8] Maß für die Streuung um den mittleren Wert.
[9] Verfahren, mit dem Kinder „herausgefiltert“ werden sollen, die bestimmte Merkmale aufweisen.

die Umrechnung der Rohpunktwerte bzw. die Berechnung von Untertest- und Gesamttestwerten; orientieren jedoch die Entscheidung, ob zusätzlicher Sprachförder-bedarf besteht, an den wesentlich gröberen Stanine-Normen (SN)[10].

Die SN-Skala[11] (Stanine oder SN steht für „standard nine“) umfasst die Werte 1 bis 9. Sie hat einen Mittelwert von 5 und eine Standardabweichung von 2. Dabei kennzeichnet SN 1 eine extrem unterdurchschnittliche Leistung. SN 2 bis 4 markieren unterdurchschnittliche Leistungen. Ab SN 5 können Leistungen als durchschnittlich bzw. überdurchschnittlich gelten. Diese grobe Einteilung reicht aus, um Risikoniveaus zu definieren bzw. Grenzwerte festzulegen, die den Übergang von einer Niveaustufe zur nächsten markieren.

So kann der „Tabelle zur Ermittlung des Risikoniveaus“ entnommen werden, in welche von drei Risikoniveaus die Leistung eines Kindes zwei Jahre vor der Schule einzuordnen ist: in den Rot-Bereich (SN 1), der signalisiert, dass ein extremes Risiko erkennbar wird, so dass eine zusätzliche Sprachförderung vonnöten ist; in den Gelb-Bereich (SN 2-4), der besagt, dass ein Risiko nicht auszuschließen ist, so dass nochmals (mit Delfin 4 – Stufe 2) genauer geprüft werden muss, ob eine zusätzliche Sprachförderung angebracht ist; und in den Grün-Bereich (SN 5 und größer), der ausdrückt, dass sich kein Risiko abzeichnet, so dass keine über die in Kindertageseinrichtungen übliche Sprachförderung hinausgehende Maßnahmen erforderlich ist.

Um die Zuordnung der Leistungen einzelner Kinder so treffsicher wie möglich zu gestalten, wurde folgendermaßen vorgegangen: Bei der Definition der Grenzwerte, die den Übergang von einem Risikoniveau zum nächsten markieren, wurden „Vertrauensintervalle“ (auch „Normbänder“ genannt) berechnet, welche den Bereich angeben, in dem sich der „wahre Wert“, also die tatsächliche Leistung eines Kindes mit 95%iger Wahrscheinlichkeit bewegt. Die Vertrauensintervalle berechnet man, indem der Standardmessfehler, den ein Test immer hat, zum Schutz vor Überinterpretation kleiner Wertdifferenzen, gleich bei der Grenzwertbestimmung eingearbeitet wird. Bei BiZ wurden folgende Grenzwertintervalle (Vertrauensintervalle eines Grenzwerts) ermittelt: Rot-/Gelb-Bereich: TW 32,3 bis 33,7; Gelb-/Grün-Bereich: TW 46,3 bis 47,7.

Da man bei einem Sprachtest mit Screeningfunktion „... möglichst nicht den Fehler begehen möchte, Risikokinder für Sprach- und Schriftspracherwerbsprobleme zu übersehen...“, nimmt man „... eher in Kauf, dass Kinder als Risikokinder herausgesiebt werden, die tatsächlich aber keiner Intervention, also beispielsweise keiner weiteren Sprachförderung oder einer Therapie bedürfen.“ (Schöler & Schäfer, 2004, S. 18); als den Fehler zu begehen, Risikokinder zu übersehen. Demzufolge wurde als Grenzwert immer derjenige TW bestimmt, der das geringste Risiko birgt, Risikokinder zu übersehen; also beim Übergang vom Rot- zum Gelb-Bereich der TW 32,3 und beim Übergang vom Gelb-zum Grün-Bereich der TW 47,7.

[10] Es handelt sich dabei, wie auch bei den T-Wert-Normen der Fall, um Normen, die keine Normalverteilung der Normierungsstichprobe voraussetzen.

[11] Die Transformation der Prozentrangnormen in Stanine-Werte wurde anhand statistischer Umrechnungstabellen vorgenommen (vgl. z.B. Bühner 2006, S. 106).

Literatur

Barnett, W.S. (2001): Preschool education for economically disadvantaged children: Effects on reading achievement and related outcomes. In: Neumann, S.B./Dickinson, D.K. (Eds.): Handbook of early literacy research (pp. 421-443). New York: Guilford.

Bradley, R.H. & Corwyn, R.F. (2002): Socioeconomic status and child development. In: Annual Review of Psychology 53, pp. 371-399.

Bühner, M. (2006); Einführung in die Test- und Fragebogenkonstruktion. München: Pearson.

Frein, T./Möller, G./Petermann, A. & Wilpricht, M. (2006): Die empirische Seite: Bedarfsgerechte Stellenzuweisung: Das neue Instrument Sozialindex. In: SchulVerwaltung Nr. 6, S. 188.

Fried, L. (2003): (Schrift-)Sprachfähigkeit als kulturelle Basiskompetenz von Kindergartenkindern? In: Arnold, R./Günther, H: (Hrsg.): Innovative Bildungs- und Erziehungsprozesse. Kaiserslautern: Fachgebiet Pädagogik der Universität Kaiserslautern.

Grimm, H., unter Mitarbeit von Aktas, M. & Frevert, S. (2000): SETK-2. Sprachentwicklungstest für zweijährige Kinder. Manual. Göttingen: Hogrefe.

Gonzales, V. (2001): The role of socioeconomic and sociocultural factors in language minority children's development: An ecological research view. (http://www.findarticles.com/p/articles/mi qa3722/is 200101/ai n8943090/print)

Halle, T./Calkins, J./Berry, D. & Johnson, R. (2003): Promoting language and literacy in early childhood care and education settings. In: Child Care & Early Education Research Connections (CCEERC), September 2003, pp. 2-17.

Hoff, E. (2003): The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. In: Child Development 74, pp. 1368-1378.

Hoff-Ginsberg, E. (2000): Soziale Umwelt und Sprachlernen. In: Grimm, H. (Hrsg.): Sprachentwicklung (S. 463-494). Göttingen: Hogrefe (= Enzyklopädie der Psychologie, C III, Bd. 3).

Landesamt für Datenverarbeitung und Statistik NRW (2007): Statistiken der Kinder- und Jugendhilfe. Teil III.1. Kinder und tätige Personen in Tageseinrichtungen am 15.03.2006. Dortmund: Technische Universität, Dortmunder Arbeitsstelle Kinder- und Jugendhilfestatistik.

Lienert, G.A. & Raatz, U. (1998): Testaufbau und Testanalyse. Weinheim: Beltz.

Moosbrugger, H. & Kelava, A. (2007): Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In: Moosbrugger, H. & Klelava, A. (Hrsg.), Testtheorie und Fragebogenkonstruktion (S. 8-26). Berlin: Springer.

Rost, J. (1996): Lehrbuch Testtheorie, Testkonstruktion. Bern: Huber.

Schöler, H. & Schäfer, P. (2004): HASE. Heidelberger Auditives Screening in der Einschulungsuntersuchung – Itemanalysen und Normen. Heidelberg: Pädagogische Hochschule Heidelberg, Arbeitsbericht Nr. 17 aus dem Forschungsprojekt „Differenzialdiagnostik".

Tent, L. & Stelzl, I. (1993). Pädagogisch-psychologische Diagnostik. Band 1: Theoretische und methodische Grundlagen. Göttingen: Hogrefe.

Wottawa, H. (1991): Grundriß der Testtheorie. München: Juventa.

Anhang

Varianzanalyse Jungen - Mädchen

Deskriptive Ergebnisse: Jungen und Mädchen

Delfin 4 Untertests		N	Mittelwert	Stan-dardab-weichung	Standard-fehler	95%-Konfidenzintervall für den Mittelwert	
						Untergrenze	Obergrenze
Handlungsanweisungen ausführen	Mädchen	15 666	7,99	3,241	,026	7,94	8,05
	Junge	16 376	7,83	3,343	,026	7,78	7,88
Kunstwörter nachsprechen	Mädchen	15 452	5,10	2,479	,020	5,07	5,14
	Junge	16 214	4,97	2,408	,019	4,93	5,01
Sätze nachsprechen	Mädchen	15 440	17,02	10,900	,088	16,85	17,19
	Junge	16 169	16,18	10,711	,084	16,01	16,34
Bild beschreiben	Mädchen	15 440	4,88	3,492	,028	4,82	4,93
	Junge	16 192	4,88	3,452	,027	4,83	4,93

* = Stichprobengröße: Eigentlich umfasst die Gesamtstichprobe 14 859 Kinder; da jedoch nicht in allen Testprotokollen (eindeutige) Geschlechtsangaben gemacht wurden und nicht alle Kinder jeden Untertest durchgeführt bzw. zu Ende geführt haben, besteht ein – von Untertest zu Untertest – variierender „Stichprobenschwund“;

** = Das Konfidenzintervall (auch Vertrauensbereich genannt) sagt etwas über die Präzision der Schätzung des Mittelwerts aus. Es gibt mit einer 95%-Wahrscheinlichkeit an, in welchem Wertebereich der geschätzte Mittelwert variieren kann, kennzeichnet also die Präzision der Schätzung.

Prüfstatistische Ergebnisse: Unterschiede zwischen Jungen und Mädchen

Delfin 4 Untertests		Quadratsumme	df	Mittel der Quadrate	F	*p*
Handlungsanweisungen ausführen	Zwischen den Gruppen	224,989	1	224,989	20,736	,000
	Innerhalb der Gruppen	347634,092	32040	10,850		
	Gesamt	347859,081	32041			
Kunstwörter nachsprechen	Zwischen den Gruppen	142,480	1	142,480	23,881	,000
	Innerhalb der Gruppen	188914,335	31664	5,966		
	Gesamt	189056,815	31665			
Sätze nachsprechen	Zwischen den Gruppen	5659,024	1	5659,024	48,481	,000
	Innerhalb der Gruppen	3689414,459	31607	116,728		
	Gesamt	3695073,483	31608			
Bild beschreiben	Zwischen den Gruppen	,079	1	,079	,007	,936
	Innerhalb der Gruppen	381295,915	31630	12,055		
	Gesamt	381295,994	31631			

df = Freiheitsgrade;
F = Prüfgröße des F-Tests; bei signifikantem F-Wert (vgl. *p*) unterscheiden sich die Mittelwerte der Gruppen signifikant;
p = Signifikanzniveau (sollte nicht höher als 0,05 sein);

Varianzanalyse: „unter 4 Jahre“ - „4 Jahre und älter“

Deskriptive Ergebnisse: Kindergruppen „unter 4 Jahre“ sowie „4 Jahre und älter“

Delfin 4 Untertests		N*	Mittelwert	Standard-abweichung	Standard-fehler	95%-Konfidenzintervall für den Mittelwert**	
						Untergrenze	Obergrenze
Handlungsanweisungen ausführen	unter 4 Jahre	5 284	8,50	2,536	,035	8,43	8,57
	4 Jahre und älter	9 066	9,19	2,195	,023	9,15	9,24
	Gesamt	14 350	8,94	2,350	,020	8,90	8,98
Kunstwörter nachsprechen	unter 4 Jahre	4 638	5,37	1,963	,029	5,31	5,42
	4 Jahre und älter	8 523	5,85	1,825	,020	5,81	5,89
	Gesamt	13 161	5,68	1,889	,016	5,65	5,71
Sätze nachsprechen	unter 4 Jahre	4 510	18,33	9,201	,137	18,06	18,60
	4 Jahre und älter	8 438	22,33	8,236	,090	22,15	22,51
	Gesamt	12 948	20,94	8,794	,077	20,78	21,09
Bild beschreiben	unter 4 Jahre	4 725	5,68	2,976	,043	5,59	5,76
	4 Jahre und älter	8 491	6,40	3,035	,033	6,33	6,46
	Gesamt	13 216	6,14	3,034	,026	6,09	6,19

* = Stichprobengröße: Eigentlich umfasst die Gesamtstichprobe 14 859 Kinder; da jedoch nicht in allen Testprotokollen (eindeutige) Altersangaben gemacht wurden und nicht alle Kinder jeden Untertest durchgeführt bzw. zu Ende geführt haben, besteht ein – von Untertest zu Untertest – variierender „Stichprobenschwund“;

** = Das Konfidenzintervall (auch Vertrauensbereich genannt) sagt etwas über die Präzision der Schätzung des Mittelwerts aus. Es gibt mit einer 95%-Wahrscheinlichkeit an, in welchem Wertebereich der geschätzte Mittelwert variieren kann, kennzeichnet also die Präzision der Schätzung.

Prüfstatistische Ergebnisse: Unterschiede zwischen den Kindergruppen „unter 4 Jahre“ sowie „4 Jahre und älter“

Delfin 4 Untertests		Quadrat-summe	df	Mittel der Quadrate	F	*p*
Handlungsanweisungen ausführen	Zwischen den Gruppen	1 596,6	1	1 596,6	294,967	0,000
	Innerhalb der Gruppen	77 661,9	14 348	5,4		
	Gesamt	79 258,6	14 349			
Kunstwörter nachsprechen	Zwischen den Gruppen	705,1	1	705,1	200,608	0,000
	Innerhalb der Gruppen	46 250,1	13 159	3,5		
	Gesamt	46 955,1	13 160			
Sätze nachsprechen	Zwischen den Gruppen	47 067,5	1	47 067,5	638,649	0,000
	Innerhalb der Gruppen	954 100,0	12 946	73,7		
	Gesamt	1 001 167,5	12 947			
Bildbeschreibung	Zwischen den Gruppen	1 567,2	1	1 567,2	172,483	0,000
	Innerhalb der Gruppen	120 065,3	13 214	9,1		
	Gesamt	121 632,7	13 215			

df = Freiheitsgrade;
F = Prüfgröße des F-Tests; bei signifikantem F-Wert (vgl. *p*) unterscheiden sich die Mittelwerte der Gruppen signifikant;
p = Signifikanzniveau (sollte nicht höher als 0,05 sein);

Normierungstabelle/Entscheidungsmatrix zu Delfin 4 – Stufe 2 –
Test „Besuch im Pfiffikushaus (BiP)“

Anleitung zur Entscheidungsfindung

Der Test „Besuch im Pfiffikushaus (BiP)“ liefert die Grundlage für die Entscheidung, ob für ein Kind die Sprachstandsfeststellung nach § 36 Absatz 2 Schulgesetz ....

Bei der Auswertung von „Besuch im Pfiffikushaus“ sind zwei Fallkonstellationen – je nach Ergebnis– möglich:

- Das Kind benötigt keine zusätzliche pädagogische Sprachförderung („grün“). Die Sprachstandsfeststellung ist beendet.
- Das Testergebnis legt eine zusätzliche pädagogische Sprachförderung nahe („rot“). Die Sprachstandsfeststellung ist beendet.

Unterlagen

Zur Ermittlung eines dieser Ergebnisse sind folgende Unterlagen notwendig:

1. Protokollheft
2. Umrechnungstabellen
3. Entscheidungstabelle
4. Auswertungsraster

Schritte der Entscheidungsfindung

Die Entscheidungsfindung umfasst folgende Schritte:

1. Schritt	Rohpunktwerte der Untertests ermitteln (RW)
2. Schritt	Standardwerte der Untertests ermitteln (TW)
3. Schritt	Durchschnittsstandardwert des Gesamttests ermitteln (DTW)
4. Schritt	Einordnung des DTW in

1. Schritt:
Die Rohpunktwerte (RW) jedes der sieben Untertests werden dem Protokollheft des einzelnen Kindes entnommen und in das Auswertungsraster übertragen. In das Auswertungsraster wird ebenfalls das Alter des Kindes am Testtag notiert (jünger als vier Jahre bzw. vier Jahre und älter). Hat ein Kind bei einem Untertests nicht mitgearbeitet, erhält es hierfür „0“ Rohpunktwerte.

2. Schritt:
Die Rohpunktwerte werden in Standardwerte (TW) umgewandelt. Anhand der Umrechnungstabellen lässt sich ablesen, welche Standardwerte (TW) den sieben Rohpunktwerten der jeweiligen Untertests entsprechen. Dabei ist je nach dem Alter des getesteten Kindes unbedingt zwischen der Tabelle mit den Altersnormen für jüngere Kinder unter vier Jahren und der Tabelle mit den Altersnormen für vierjährige Kinder zu unterscheiden.

Da die empirischen Berechnungen ergeben haben, dass sich die Sprachleistungen der jüngeren Kinder (unter vier Jahren) von denen der älteren Kinder (über vier Jahre) bedeutsam unterscheiden, wurden altersspezifische Umrechnungstabellen (Altersnormen) entwickelt. Es hängt vom Alter des Kindes am Testtag ab, aus welcher Umrechnungstabelle man jeweils den T-Wert entnimmt.

3. Schritt:
Die Standardwerte (TW) der sieben Untertests $_{(WVTW}$ + $_{BKTW}$ + $_{KNTW}$ + $_{SNTW}$ + $_{PBTW}$ + WP_{TW} + $_{BETW)}$ werden addiert. Die Summe dividiert durch 7 ergibt den durchschnittlichen Standardwert des Gesamttests (DTW).[1]

4. Schritt:
Der DTW wird mit den farbig markierten (rot, grün) Zeilen der Entscheidungstabelle verglichen. Damit ist die Entscheidung über das Ergebnis des Tests „Besuch im Pfiffikushaus“ möglich.

---

1 Berechungsformel: $_{(WVTW}$ + $_{BKTW}$ + $_{KNTW}$ + $_{SNTW}$ + $_{PBTW}$ + $_{WPTW}$ + BE_{TW}) : 7

Umrechnungstabellen: Altersnormen für jüngere Kinder (unter vier Jahre)[2]

WV (jünger) RW	TW
0	20
1	25
2	29
3	31
4	34
5	36
6	38
7	39
8	42
9	44
10	45
11	47
12	49
13	51
14	54
15	57
16	62
17	66
18	75

BK (jünger) RW	TW
0	38
1	39
2	40
3	41
4	43
5	44
6	45
7	46
8	51
9	52
10	53
11	56
12	75

[2] Stichprobengröße insgesamt: N = 2.315; die Stichprobe umfasst die Testergebnisse von Kindern aus den Schulaufsichtsbezirken Düsseldorf, Hamm, Mühlheim und Münster, die unterschiedliche Sozialindices repräsentieren.

KN (jün-ger) RW	**TW**
0	34
1	37
2	40
3	44
4	50
5	53
6	59
7	75

SN (jünger) RW	TW
0	28
1	37
2	37
3	38
4	38
5	38
6	39
7	39
8	40
9	40
10	41
11	41
12	42
13	43
14	43
15	44
16	44
17	45
18	45
19	45
20	46
21	46
22	47
23	47
24	48
25	48
26	48
27	49
28	49
29	49
30	50
31	51
32	52
33	53
34	53
35	54
36	55
37	56
38	57
39	59
40	63
41	75

PB (jünger) PW	TW
0	36
1	37
2	38
3	45
4	47
5	48
6	51
7	54
8	57
9	60
10	65
11	72
12	75

WP (jünger) RW	TW
0	24
1	28
2	30
3	32
4	34
5	34
6	35
7	35
8	36
9	37
10	38
11	39
12	40
13	41
14	42
15	42
16	43
17	44
18	44
19	45
20	46
21	47
22	48
23	49
24	50
25	51
26	52
27	53
28	54
29	55
30	58
31	59
32	63
33	64
34	69
35	72
36	75

BE (jünger) RW	TW
0	34
1	35
2	35
3	36
4	37
5	39
6	39
7	40
8	41
9	43
10	44
11	44
12	45
13	46
14	47
15	48
16	49
17	50
18	51
19	52
20	53
21	53
22	54
23	55
24	56
25	56
26	57
27	58
28	59
29	60
30	61

BE (jünger) Fortsetzung	
RW	TW
31	62
32	62
33	63
34	64
35	65
36	66
37	66
38	67
39	68
40	69
41	69
42	70
43	71
44	72
45	73
46	73
47	74
48	74
49	75
50	75
51	75
52	75
53	75
54	75
55	75
56	75
57	75
58	75
59	75
60	75

Umrechnungstabellen: Altersnormen für ältere Kinder (ab vier Jahren)

WV (älter) RW	TW
0	24
1	26
2	28
3	31
4	34
5	36
6	38
7	40
8	42
9	43
10	45
11	47
12	48
13	50
14	52
15	55
16	58
17	63
18	75

BK (älter) RW	TW
0	38
1	39
2	39
3	40
4	42
5	43
6	44
7	45
8	49
9	50
10	51
11	53
12	75

KN (älter) RW	TW
0	36
1	39
2	41
3	44
4	48
5	51
6	57
7	75

SN (älter) RW	TW
0	34
1	35
2	35
3	36
4	36
5	37
6	37
7	38
8	38
9	39
10	39
11	40
12	40
13	41
14	41
15	41
16	42
17	43
18	43
19	43
20	44
21	44
22	45
23	45
24	46
25	46
26	47
27	47
28	48
29	48
30	49
31	49
32	49
33	50
34	51
35	52
36	53
37	54
38	55
39	57
40	61
41	75

PB (älter) RW	TW
0	39
1	40
2	44
3	45
4	46
5	48
6	50
7	53
8	56
9	60
10	65
11	70
12	75

WP (älter) RW	TW
0	32
1	33
2	34
3	34
4	35
5	35
6	36
7	37
8	38
9	38
10	39
11	39
12	40
13	40
14	41
15	42
16	42
17	43
18	44
19	44
20	45
21	45
22	46
23	47
24	48
25	48
26	49
27	50
28	52
29	53
30	55
31	56
32	60
33	62
34	66
35	71
36	75

BE (älter) RW	TW
0	29
1	31
2	32
3	34
4	35
5	36
6	37
7	38
8	39
9	40
10	41
11	42
12	43
13	44
14	45
15	46
16	47
17	48
18	49
19	50
20	51
21	52
22	53
23	54
24	54
25	55
26	56
27	57
28	57
29	58
30	59

BE (älter) Fortsetzung	
RW	TW
31	61
32	62
33	62
34	63
35	64
36	65
37	66
38	67
39	68
40	69
41	70
42	71
43	72
44	72
45	73
46	74
47	74
48	75
49	75
50	75
51	75
52	75
53	75
54	75
55	75
56	75
57	75
58	75
59	75
60	75

Berechnung des „Durchschnittlichen TW“:

$(WV_{TW} + BK_{TW} + KN_{TW} + SN_{TW} + PB_{TW} + WP_{TW} + BE_{TW}) : 7 = DTW$

Entscheidungstabelle

Niveau	**DTW**		**Entscheidung**
Rot	**20,0 – 42,4**		zusätzliche pädagogische Sprachförderung
Grün	**42,5 - 75,0**		keine zusätzliche Sprachförderung

Auswertungsraster

Name:		Alter:
Untertest	Rohpunktwert (RW) (Protokollheft)	Standardwert (TW)
Wortverständnis (WV)		
Begriffsklassifikation (BK)		
Kunstwörter nachsprechen (KN)		
Sätze nachsprechen (SN)		
Pluralbildung (PB)		
Wortproduktion (WP)		
Bilderzählung (BE)		
Summe der Standardwerte:		
Ergebniswert (DTW) (Summe der Standardwerte geteilt durch sieben):		

Auswertungsraster - Beispie-

le Beispiel 1: Kind im Alter über vier Jahre

Name: Max Mustermann		Alter: über vier Jahre
Untertest	Rohpunktwert (RW) (Protokollheft)	Standardwert (TW)
Wortverständnis (WV)	7	40
Begriffsklassifikation (BK)	3	40
Kunstwörter nachsprechen (KN)	2	41
Sätze nachsprechen (SN)	12	40
Pluralbildung (PB)	0	39
Wortproduktion (WP)	10	39
Bilderzählung (BE)	0	29
Summe der Standardwerte:		268
Ergebniswert (DTW) (Summe der Standardwerte geteilt durch sieben):		38,2

Nach der Entscheidungstabelle liegt der DTW 38,2 im „roten“ Bereich. Bereich. Max Mustermann benötigt zusätzliche pädagogische Sprachförderung.

Beispiel 2: Kind im Alter unter vier Jahren

Name: Marie Musterfrau		Alter: unter vier Jahren
Untertest	Rohpunktwert (RW) (Protokollheft)	Standardwert (TW)
Wortverständnis (WV)	7	39
Begriffsklassifikation (BK)	9	52
Kunstwörter nachsprechen (KN)	0	34
Sätze nachsprechen (SN)	19	45
Pluralbildung (PB)	5	48
Wortproduktion (WP)	22	49
Bilderzählung (BE)	16	315
Summe der Standardwerte:		45
Ergebniswert (DTW) (Summe der Standardwerte geteilt durch sieben):		

Nach der Entscheidungstabelle liegt der DTW 45 im „grünen“ Bereich. Marie Muster-frau benötigt keine zusätzliche pädagogische Sprachförderung.