KöBeS

Kölner Beiträge zur Sprachdidaktik

Reihe B I/2007

herausgegeben von
Michael Becker-Mrotzek, Ursula Bredel & Hartmut Günther

Markus Linnemann

Entwicklung und Validierung eines Tests zur Erfassung der Lesekompetenz von Berufsschülern und -schülerinnen

Kölner Beiträge zur Sprachdidaktik

herausgegeben von
Hartmut Günther, Ursula Bredel & Michael Becker-Mrotzek

Reihe B

Markus Linnemann

Entwicklung und Validierung eines Tests zur Erfassung der Lesekompetenz von Berufsschülern und -schülerinnen

Dieser Text wurde im Oktober 2006 als Diplomarbeit im Fach Erziehungswissenschaft an der Universität zu Köln eingereicht. Sie wurde angefertigt bei Prof. Dr. Hilde Haider. Die ursprünglich im Anhang abgedruckten Materialien sind in dieser Textversion nicht enthalten.

Anschrift des Verfassers:

Markus Linnemann
Bremer Str. 3
50670 Köln
0221 / 1392914
markus.linnemann@uni-koeln.de
www.jazznbass.de

Mein Dank geht an meine Betreuerin Hilde Haider, an Michael Becker-Mrotzek, an Christian Huber und Jürgen Wilbert für das Lesen und die Kritik. Ein besonderer Dank geht an Wolfgang Winkelmann, der mich vor einigen Jahren ins statistisch gesehen kalte Wasser geworfen hat...

Inhalt

Vorwort			7
Einleitung			9
		I. Theoretischer Teil	
1.	Lesekompetenz		11
	1.1	Lesekompetenz bei PISA	14
	1.2	Weitere Lesetests	19
	1.3	Zusammenfassung	20
2.	Berufskollegs		21
	2.1	Duales System der Berufsausbildung	21
	2.2	Forderungen an einen Lesetest für Berufsschüler	22
	2.3	Zusammenfassung	23
3.	„Lesetest für Berufsschüler/innen“ - Anforderungen an die Entwicklung		24
	3.1	Anforderungen	24
	3.2	Zusammenfassung	27
4.	Testtheorie		28
	4.1	„Klassische Testtheorie“	29
	4.2	Probabilistische Testmodelle	37
	4.3	Kriteriumsorientierte Testkonstruktion	40
	4.4	Begründung für das Modell der Klassischen Testtheorie als Modell für den Test	42
	4.5	Die theoretische Konstruktion der Kompetenzstufen	43
	4.6	Zusammenfassung	45
5.	Zusammenfassung und Fragestellung der Arbeit		46
		II. Empirischer Teil	
1.	Entwicklung des „Lesetests für Berufsschüler/innen“		49
	1.1	Phasen der Entwicklung des LTB^{-3}	49
	1.3	Pretests	52
2.	Validierung des „Lesetests für Berufsschüler/innen“		71
	2.1	Untersuchungen zur Validität	71
	2.2	Die Kompetenzstufen	89
	2.3	Normierung des Tests	95

	III. Diskussion	99
1.	Objektivität	99
2.	Reliabilität	100
3.	Validität	100
4.	Normierung	102
5.	Testökonomie	102
6.	Bildung der Kompetenzstufen	102
7.	Fazit und Ausblick	103
	IV. Zusammenfassung	105
Literatur		107
Anhang		111
A	Tabellen	111
B	Material	118

„[...] for reasons unknown that as a result of the public works of Puncher and Wattmann it is established beyond all doubt that in view of the labors of Fartov and Belcher left unfinished for reasons unknown of Testew and Cunard left unfinished it is established what many deny that man in Possy of Testew and Cunard that man in Essy that man in short that man in brief in spite of the strides of alimentation and defecation wastes and pines [...]"

Aus Lucky's Monolog in Samuel Becketts „Waiting for Godot", 1954.

Vorwort

Ein Test kann nicht von einer Person entwickelt werden und auch der „Lesetest für Berufsschüler/innen - LTB-3" ist in einer Arbeitsgruppe entstanden, die wiederum in ein Projekt zur „Leseförderung in der Berufsbildung" eingebettet war, das zwischen 2002 und 2006 an drei berufsbildenden Schulen im Kreis Düren und Jülich durchgeführt wurde.

Mein Auftrag bestand zunächst darin, die Auswahl der Texte und die Konstruktion der Fragen in testtheoretischer Hinsicht zu begleiten. Der hauptsächliche Beitrag lag zudem in der Auswertung der Pretests und der Validierung des Messinstruments. Dieser Beitrag ist in dieser Arbeit vorwiegend beschrieben. Nicht alle testtheoretischen Forderungen konnten vollkommen durch- und umgesetzt werden, an vielen Stellen mußte ein „pragmatischer" Weg gewählt werden.

Neben der wissenschaftlichen Arbeit kam mir die Aufgabe zu, das Testhandbuch zusammenzustellen sowie die Materialien soweit aufzuarbeiten, dass sie von einem potentiellen Verlag veröffentlich werden könnten. Dieser Verlag ist mit dem Gilles & Francke-Verlag in der Zwischenzeit gefunden.

Im Hauptteil dieser Arbeit werden, außer in wenigen Beispielen, keine Testfragen wortwörtlich genannt. Die Fragen werden nummeriert „Item x" genannt. Die genauen Wortlaute sind im Anhang B nachzuschlagen, sie sind zum Verständnis dieser Arbeit jedoch nicht nötig.

Der Aufbau dieser Arbeit weicht leicht vom Aufbau grundlagenwissenschaftlicher Arbeiten in der Psychologie und verwandten Fächern ab, da es hier um ein „technologisches Forschungsproblem", zu dem

eine Testentwicklung gehört (vgl. Herrmann, zitiert nach Hager, Spies und Heise, 2001, S. 13), und nicht um das theoriegeleitete Aufstellen und Testen von Hypothesen geht. Die übliche Unterteilung in einen theoretischen und einen empirischen Teil wird beibehalten. In der Diskussion wird jedoch Problemen bei der Testkonstruktion und Validitätsmängel breiten Raum gegeben.

In der gesamten Arbeit wird aus Gründen der besseren Lesbarkeit ausschließlich die männliche Form benutzt, es sei denn, der Test wird bei seinem Namen - „Lesetest für Berufsschüler/innen“ - genannt.

Einleitung

Lesekompetenz ist eine der wesentlichen Schlüsselqualifikationen für die Teilhabe an hoch literalisierten Gesellschaften. Lesekompetenz meint einerseits die basale Fähigkeit, Gedrucktes zu entziffern, darüber hinaus jedoch insbesondere *„die Fähigkeit, [...] das Begriffsvermögen und die Fähigkeit zur Synthesebildung auf all das anzuwenden, was man hört oder sieht"* (Mathews, 1974, S. 64). Denn Texte begegnen uns eben nicht nur in geschriebener, sondern auch in gesprochener Form, beispielsweise in Vorträgen oder Radiobeiträgen. Lesekompetenz meint im Kern die Fähigkeit, sprachliche Äußerungen außerhalb ihres ursprünglichen Handlungszusammenhangs zu verstehen.

Lesekompetenz ist jedoch nicht nur für die Teilhabe an gesellschaftlichen Prozessen erforderlich, sondern auch für die Bewältigung vieler anderer Aufgaben, wie der enge Zusammenhang von Lesekompetenz und anderen kognitiven Größen zeigt. So ist bei PISA die Leistung im Lesen eng verbunden mit der Leistung in Mathematik und ein *„Großteil des Einflusses kognitiver Hintergrundmerkmale [ist] über die Lesekompetenz vermittelt"* (PISA 2000, S. 185).

Mitte der 1980er Jahre hatte man die Befürchtung, der Stellenwert des Lesens würde durch die starke Zunahme audiovisueller Medien mit dem Aufkommen des Privatfernsehens zurückgedrängt. Es zeigte sich jedoch, dass das Lesen durch diese Medien gerade nicht abgelöst wurde. Die sogenannten neuen Medien, wie das Internet, E-Mail-Verkehr oder SMS, erfordern geradezu Kompetenzen im Lesen und Schreiben. Auch wenn E-Mails und vor allem SMS nicht gerade Beispiele für komplexe Texte sind, so fordert gerade ihre besondere Kürze, etwa einer SMS, spezifische Kompetenzen, u.a. die Fähigkeit zur begrifflichen Verdichtung.

Lesen ist jedoch nie nur einfache „Entnahme" von Bedeutung aus einem Text, sondern immer verbunden mit der aktiven Rekonstruktion des Sinns. Erst durch das Lesen wird aus einem „Text-auf-dem Papier" ein „Text-im-Kopf". Das setzt ein Nachdenken, eine Reflexion über Inhalt und Form des Textes voraus. Hinzu kommt die Fähigkeit, Bezüge zum Autor herzustellen, die im Falle einer SMS von großer Wichtigkeit sein kann. Zudem kommt mit den neuen Medien ein neuer Kommunikationstypus auf: Stellt die ursprüngliche schriftliche Kommunikation eine *„zerdehnte Kommunikation"* (Ehlich, 1983) dar, d.h. eine Verständigung

über Raum und Zeit hinweg, so ist es mit Hilfe von online-Medien wie E-Mail, SMS und Chatroom möglich, die schriftliche Kommunikation zwar über den Raum, jedoch nicht über die Zeit zu „dehnen".

Aus dieser Vorstellung heraus wird deutlich, dass eine Diagnose der Lesekompetenz nicht nur Ansätze zur Leseförderung im Sinne einer Alphabetisierung bieten kann, vielmehr bietet sie einen Anhaltspunkt zur Erklärung vielfältiger Defizite. Wenn es also gelänge, einen Test zur Erfassung der Lesekompetenz zu entwickeln, der nicht (nur) auf die Diagnostik einer Leseschwäche im Sinne einer psychischen Störung oder der Entdeckung von Analphabeten aus ist, könnte man schlechte Schulleistungen und sogar Schulversagen anders einordnen. Auch soziologische und sozialpädagogische Studien könnten davon profitieren, die Variable „Lesekompetenz" in den Blick zu nehmen, wenn sie die Veränderung der Teilhabe an einer hoch literalisierten Gesellschaft analysieren.

Aus diesem Grund sollte ein Lesetest entwickelt, der zwar, das ergab sich aus dem Projektzusammenhang, für Berufsschüler und -schülerinnen gedacht ist, jedoch auch für Schulabgänger der Sekundarstufe I und Erwachsene gleichermaßen geeignet sein sollte. Der zu entwickelnde Test soll es ermöglichen, auf ökonomische und objektive Weise die Lesekompetenz zu erfassen. Er soll einfach handzuhaben sein, Testdurchführung und -auswertung sollen nur wenig Zeit in Anspruch nehmen und eine problemlose Integration in den Schulalltag ermöglichen.

Ziel dieser Arbeit ist die Beschreibung der Entwicklung und die Untersuchung der Validität des Lesetests. Er orientiert sich stark am Lesekompetenzkonzept der PISA-Studie, deshalb wird im theoretischen Teil dieser Arbeit zunächst das angelsächsische Konzept der „Reading Literacy" und das testtheoretische Modell der PISA-Studie mit den besonderen Anforderungen der Berufskollegs verknüpft. Im Anschluss daran wird beschrieben, welche testtheoretischen Modelle zur Entwicklung des Tests in Frage kommen und wie sie mit der Lesekompetenz und den Ansprüchen der Berufskollegs zum „Lesetest für Berufsschüler/innen" führen.

Im empirischen Teil wird, ausgehend von diesen Überlegungen gezeigt, wie der Test konkret entwickelt wurde und wie sich die Befundlage zur Validität gestaltet. In diesen Kapiteln werden bereits Hinweise auf mögliche Schwachstellen des Tests gegeben. In der Diskussion wird der so entwickelte Test einer kritischen Prüfung unterzogen. Im abschließenden Kapitel „Forschungsausblick" soll ein weiterer Verlauf der Kompetenzforschung angedacht werden.

I. Theoretischer Teil

Der theoretische Teil dieser Arbeit gliedert sich in vier Bereiche. Im Abschnitt *„Lesekompetenz“* geht es zunächst um eine Definition der Lesekompetenz. Da sich die Testkonstruktionsgruppe am PISA-Test orientiert hat, wird der Lesefähigkeit nach der PISA-Studie ein größerer Raum gegeben. Im Anschluss daran wird gezeigt, wieso gerade *„Berufsschüler“* hier im Focus des Interesses stehen und welche besonderen Anforderung seitens der Berufskollegs gestellt werden. Die Definition der PISA-Studie und die Anforderungen des Berufskollegs werden dann in Bezug zum zu entwickelnden Lesetest gesetzt. Im Anschluss daran werden testtheoretische Modelle vorgestellt. Besonderes Augenmerk liegt hier auf der Klassischen Testtheorie, da sie für die Entwicklung des Lesetests nutzbar gemacht wurde.

1. Lesekompetenz

Sprache entwickelte sich zur Kommunikation zwischen zwei oder mehr Personen, die sich unmittelbar gegenüberstanden. Die Erfindung der Schrift, die eine „zerdehnte Kommunikation“ im Sinne Ehlichs (1983) erst ermöglichte, war ein wichtiger Wendepunkt der Zivilisation. Seit dem Aufkommen der ersten Schriftzeichen etwa 4000 v. Chr. bis heute, entwickelten sich die Schriftsprachen (mit der Ausnahme von Chinesisch und anderen asiatischen Sprachen, die auf dem Chinesischen beruhen) zu einem alphabetischen Schreibsystem, in dem eine geringe Anzahl Symbole Klänge repräsentieren, die die gesprochene Sprache bilden (vgl. Carlson, Buskist und Martin, 2000, S. 319 ff.).

Nach PISA ist *„Lesen [...] ein höchst komplexer Vorgang, der aus mehreren Teilprozessen besteht“* (PISA 2000, S. 71). Die Bedeutung eines Textes wird auf der Wort-, Satz- und Textebene unterschiedlich erfasst, d.h. es sind unterschiedliche Kompetenzen beteiligt, deren Prozesse *„wohl eher parallel als sequentiell“* (Becker-Mrotzek & Drommler, 2006, S. 31) ablaufen. Kognitionspsychologen, Psycholinguisten und Linguisten haben unterschiedlich versucht, diese Teilprozesse zu erforschen. So wurde z.B. von Just und Carpenter (1980) untersucht, wie einzelne Buchstaben und Wörter wahrgenommen werden, um daraus ein Modell des

Lesens zu entwerfen. Einen Text zu lesen bedeutet zunächst, diesen mit den Augen abzufahren, zu „scannen". Die weitergehende Verarbeitung visueller Reize geschieht jedoch nicht während der Bewegung der Augen, sondern ausschließlich während sogenannter „Fixationen", d.h. eine Betrachtung eines Wortes ohne Augenbewegung. Die durchschnittliche Dauer dieser Fixationen beträgt 250 ms, diese Dauer ist allerdings sehr variabel. Ein guter Leser benötigt weniger Zeit, um ein Wort zu erkennen, als ein ungeübter Leser. Ein schlechter Leser springt sogar zurück zu Wörtern, die er bereits gelesen hat und verweilt dort erneut. Auch gibt es Unterschiede zwischen den Wörtern: Nur 40% der Funktionswörter werden überhaupt fixiert, Inhaltswörter zu 80%. Wie werden aber nun Wörter erkannt? Dabei spielen zwei Arten eine Rolle: phonetisches Lesen und die Ganz-Wort-Erkennung. Phonetisches Lesen bedeutet, ein Wort lautsprachlich nach den Regeln, die über die Graphem-Phonem-Korrespondenz bekannt sind, zu Dekodieren. Das Ganz-Wort-Lesen bedeutet, das Wort als Ganzes erkennen zu können. Während des Lesens treten beide Arten der Worterkennung auf: je nachdem, ob ein Wort bekannt ist oder nicht. Phonetisches Lesen tritt immer dann auf, wenn ein Wort unbekannt ist. Bei einem ungeübten Leser sind das die meisten Wörter, während bei einem geübten Leser die Ganz-Wort-Erkennung den meisten Teil des Lesens ausmacht. Rubinstein et al. (1971) konnten zeigen, dass die visuelle Information mit Hilfe der Graphem-Phonem-Korrespondenz in einen phonologischen Code übersetzt wird. Nach Ellis (1992) ist dies jedoch nur für unbekannte Wörter nötig.

Die Bedeutung der Wörter wird durch Erfahrung erlernt. Die Erinnerung an die Bedeutung der Inhaltswörter umfasst das visuelle, auditive und somatosensorische Gedächtnis sowie das Gedächtnis für den Geschmacks- und Tastsinn. Das Verständnis der Adjektive ist stets mit den Objekten assoziiert, die die Träger der jeweiligen Eigenschaft sind. Werden Wörter in eine syntaktisch richtige Reihenfolge gebracht, spricht man von einem Satz. Um zu einer Bedeutung eines ganzen Satzes zu kommen, müssen die Bedeutungen der Kombinationen der einzelnen Wörter analysiert werden. Dies kann unter Verwendung syntaktischer wie semantischer Hinweise geschehen. *„Die syntaktischen Hinweise der Wortreihenfolge und der Flexionen unterstützen die Interpretation eines Satzes beim Verstehen"* (Anderson, 2001, S. 369) genauso wie die semantischen Strategien, die Menschen beim Sprachverstehen einsetzen.

Doch was bedeutet es, einen längeren Text zu lesen? Lesen ist nicht nur bloße Informationsentnahme. Ein Modell, dass vom einfachen Decodieren von Informationen ausgeht, greift zu kurz. Sprache ist ein *„Medium der Kommunikation und des Denkens, das immer einen verständigen, aktiven Nutzer verlangt"* (Becker-Mrotzek & Drommler, 2006, S. 31). Somit ist auch das Lesen von geschriebener Sprache ein aktiver Prozess der Bedeutungsre- und -konstruktion, bei dem der Leser versucht, mit Hilfe seines Sprach- und Weltwissens zu verstehen, was der Autor des Textes mitteilen wollte. Dies geschieht zwar nicht unbedingt vollkommen, jedoch *„sind die bleibenden Unterschiede zwischen Gemeinten und Verstandenem für die praktischen Zwecke unerheblich"* (ebd.). Kintsch und van Dijk (1978) entwickelten ein Informationsverarbeitungsmodell des Textverstehens. Das Modell setzt voraus, dass der Text bereits durch die syntaktischen und semantischen Verarbeitungsprozesse in Propositionen zerlegt wurde. Eine Proposition ist *„die kleinste Wissenseinheit, die eine unabhängige Aussage darstellen kann"* (Anderson, 2001, S. 472). Verstehen ist dabei das Verarbeiten dieser Propositionen, indem neue mit früheren Propositionen verknüpft werden. Dazu müssen sich die Propositionen überlappen. Tun sie dies nicht, und ist eine Inferenz, ein „Überbrückungsschluss" nötig, ist das Textverständnis erschwert. Nach Kintsch und van Dijk (1978) ist die Zahl der Propositionen im Arbeitsgedächtnis auf durchschnittlich vier beschränkt. Ist die frühere, sich mit der neueren Proposition überlappende, an die beim Lesen angeknüpft werden soll, bereits aus dem Arbeitsgedächtnis gelöscht, kann das dazu führen, dass der Leser den Text nicht versteht, es sei denn, er kann die frühere Proposition wieder aus dem Langzeitgedächtnis reaktivieren. Eine weitere Konsequenz aus der Beschränkung des Arbeitsgedächtnisses ist die Frage, welche der Propositionen und wie lange sie im Arbeitsgedächtnis aufrecht erhalten bleiben. Es wird angenommen *„dass eine Kombination aus zeitlicher Nähe und Wichtigkeit der Information verwendet wird, um diejenigen Propositionen auszuwählen, die aktiv gehalten werden"* (Anderson, 2001, S. 419). „Zeitliche Nähe" meint damit die Proposition, die gerade eben verarbeitet wurde.

Kintsch und van Dijk (1978) geben zwei Arten von mentalen Ausarbeitungen an: zum Einen den bereits erwähnten Überbrückungsschluss, um den Text mit Propositionen anzureichern und zum Anderen die Bildung von Makropropositionen, also Zusammenfassungen, „Überschriften", von wesentlichen Aussagen des Textes. Die Bildung dieser

Makropropositionen führten zu besseren Wiedergabeleistungen der wichtigsten Aussagen des Textes.

Diese höheren Prozesse sind es, die in der vorliegenden Arbeit ausschließlich betrachtet werden, wenngleich auch nicht mit Hilfe eines kognitionspsychologischen Zugangs. Es wird also im Folgenden davon ausgegangen, dass die Probanden „lesen können", also die grundlegenden Prozesse, z.B. das Erkennen der Wörter während der Fixationen, beherrschen.

1.1 Lesekompetenz bei PISA

Die PISA-Studie umfassend darzustellen, also ihre Methode, ihr Konzept, ihre Ergebnisse und die daraus sich ergebenden Schlussfolgerungen, würde den Rahmen dieser Arbeit sprengen (siehe dazu PISA 2000). Deshalb soll hier im Wesentlichen auf die Punkte eingegangen werden, die für eine Testentwicklung relevant sind, da diese Ausgangspunkt der Konzeption des „Lesetests für Berufsschüler/innen" waren.

Die meisten „großen" nationalen und internationalen Studien zur Lesekompetenz bedienen sich des aus dem angelsächsischen Sprachraums stammenden „Reading Literacy-Konzepts". Die *„Idee einer Universalisierung von Basisqualifikationen"* (PISA 2000, S. 78) steht hier im Mittelpunkt. Lesekompetenz bei PISA wird definiert als die Fähigkeit, *„geschriebene Texte zu verstehen, zu nutzen und über sie zu reflektieren, um eigene Ziele zu erreichen, das eigene Wissen und Potential weiterzuentwickeln und am gesellschaftlichen Leben teilzunehmen"* (PISA 2000, S. 78). Damit geht die PISA-Studie über die Fähigkeit, Buchstaben, Wörter und Sätze zu entschlüsseln, weit hinaus und sieht sie, neben der mathematischen und naturwissenschaftlichen Kompetenz, als ein „basales Kulturwerkzeug" (ebd.) gemäß des Reading-Literacy-Konzepts.

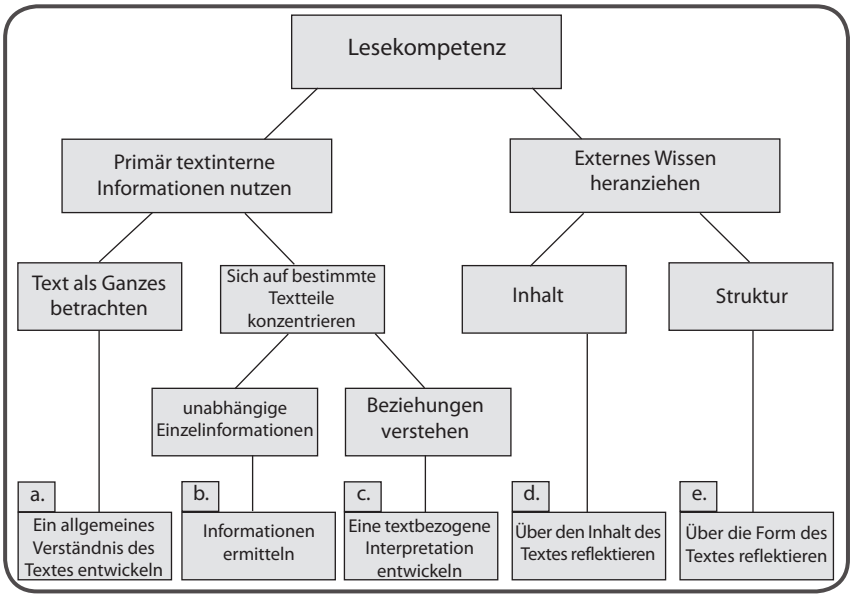
Der PISA-Studie liegt ein empirisches Kompetenzmodell zugrunde. *„Ziel ist es, empirisch nachprüfbare Modelle von Kompetenz zu entwickeln, die gerade auch Entwicklungsprozesse und individuelle Unterschiede zu erfassen in der Lage sind"* (Becker-Mrotzek & Drommler, 2006, S. 34). Kompetenz wird verstanden als die Fähigkeit, sich in authentischen Situationen zu bewähren, für die verschiedene Basisqualifikationen, die Lesekompetenz, mathematische und naturwissenschaftliche Kompetenzen, erforderlich sind.

1.1.1 Struktur der Lesekompetenz

Abbildung 1 zeigt die in PISA angenommene Struktur der Lesekompetenz. Das Modell unterscheidet zunächst die beiden Dimensionen „primär textinterne Informationen nutzen“ und „externes Wissen heranziehen“. Bei der ersten Dimension sind die textimmanenten Informationen ausreichend Grundlage zur richtigen Beantwortung der Testfragen, im zweiten Fall muss externes Wissen herangezogen werden. Diese Dimension der textimmanenten Leistung fächert sich letztlich auf in „ein allgemeines Verständnis des Textes entwickeln“, „Informationen ermitteln“ und „eine textbezogene Interpretation entwickeln“. Diese Dimension ist mit 70 Prozent der Aufgaben repräsentiert. Die Dimension „Externes Wissen heranziehen“ unterscheidet „über den Inhalt reflektieren“ und „über die Form des Textes reflektieren“ (vgl. PISA 2000, S. 82).

Im Unterschied zu kognitionspsychologischen Modellen des Textverstehens, wird durch die Tatsache, dass die Schüler mit dem Text arbeiten können, dass der Text also während der gesamten Testphase verfügbar bleibt, in erster Linie der Aspekt des Verstehens betont. Lesekompetenz ist jedoch mehr als nur Verstehen: Sie ist zusätzlich eine Gedächtnisleistung, wenn Inhalte von Texten behalten werden sollen. Dies ist jedoch Voraussetzung dafür, sich den Text zu eigen machen zu können. Im nationalen Ergänzungsteil der PISA-Studie wird deshalb die Bearbeitung der Lesetest-Aufgaben in eine Lese-, Lern- und Aufgabenphase zerlegt, um auch diese Aspekte untersuchen zu können. Diesem Ergänzungsteil werden kognitionspsychologische Modelle von Kintsch und von van Dijk und Kintsch (1983, vgl. PISA 2000, S. 24) zu Grunde gelegt, dessen Grundzüge weiter oben bereits vorgestellt wurden.

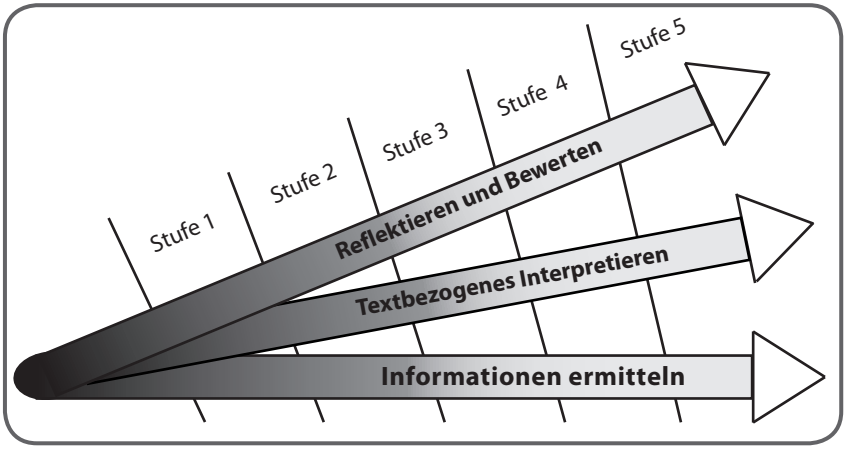
Abbildung 1: Theoretische Struktur der Lesekompetenz in PISA



1.1.2 Berichtskalen

Aufgrund von empirisch ermittelten Kriterien und inhaltlichen Gemeinsamkeiten werden in der PISA-Studie die fünf Aspekte der unteren Strukturebene (s. Abbildung 1) zu drei Subskalen zusammengefasst, die als Berichtskalen dienen: 1. „Informationen ermitteln", 2. „textbezogenes Interpretieren" (mit den Aspekten b. „ein allgemeines Verständnis des Textes entwickeln" und c. „eine textbezogene Interpretation entwickeln") und 3. „Reflektieren und bewerten" (mit den Aspekten d. „über den Inhalt des Textes reflektieren" und e. „über die Form des Textes reflektieren"). Diese Subskalen korrelieren so hoch miteinander (zwischen 0,90 und 0,95), *„dass von Subskalen einer zu Grunde liegenden Lesekompetenz statt von qualitativ verschiedenartigen Kompetenzen gesprochen wird"* (PISA 2000, S. 134). Abbildung 2 zeigt die drei Subskalen.

Abbildung 2: Die drei Subskalen des PISA-Tests



Diese drei Subskalen lassen sich hinsichtlich ihrer Anforderungen folgendermaßen charakterisieren: „Informationen ermitteln" fordert vom Leser, die passende Information oder Teilinformation in den Materialien zu finden (PISA 2000, S. 83). Die Aufgaben auf dieser Skala haben einen verschiedenen Komplexitätsgrad. Je nach Schwierigkeit der Aufgabe, ist das Verstehen größerer Textabschnitte erforderlich, es ist sogar möglich, dass Informationen, die erfragt werden nicht im Text enthalten ist und geschlussfolgert werden müssen. Die Aufgaben der Subskala „textbezogenes Interpretieren" verlangen vom Leser, dass er Bedeutung konstruiert und durch schlussfolgerndes Denken Schlüsse aus einem oder mehreren Textteilen zieht. Es wird auch verlangt, dass der Leser Schlüsse über die Absichten des Autors zieht (vgl. PISA 2000, S. 83). Die Aufgaben der Subskala „Reflektieren und Bewerten" *„verlangen vom Leser, den Text mit eigenen Erfahrungen, Wissensbeständen und Ideen in Beziehung zu setzen"* (ebd., S. 83).

Für jeden getesteten Schüler wird sowohl ein Testwert für jede dieser Subskalen als auch für die Gesamtskala, d.h. für alle Aufgaben des Tests, ermittelt. Diese Rohwerte werden linear in einen standardisierten Wert mit dem Mittelwert 500 und der Standardabweichung 100 transformiert (vgl. PISA 2000, S. 522). Um diese Werte inhaltlich interpretieren zu können *„besteht die Möglichkeit, die abstrakte Fähigkeitsskala in einzelne Stufen zu unterteilen, die anhand von typischen Aufgabenmerkmalen genauer gekennzeichnet werden können"* (ebd., S. 88). Es kann davon

ausgegangen werden, *„dass die auf einem bestimmten Schwierigkeitsniveau geforderte Lesekompetenz“* von einer Person des *„entsprechenden Fähigkeitsniveaus gerade eben noch erbracht werden kann“* (ebd., S. 90), sie aber mit hoher Wahrscheinlichkeit alle leichteren Aufgaben löst. Diese inhaltliche Interpretation geschieht mit Hilfe des auf dem Rasch-Modell aufbauenden Proficiency Scaling. Das Rasch-Modell wird weiter unten beschrieben.

1.1.3 Bildung der Kompetenzstufen

Auf den Voraussetzungen des gewählten Testmodells aufbauend, lassen sich aus der kontinuierlichen Fähigkeitsskala Kompetenzstufen herausarbeiten, indem sich Klassen von Aufgaben beschreiben lassen, die homogen sind, sich jedoch *„von Anforderungen auf einem höheren Schwierigkeitsniveau (höhere Kompetenzstufe) abgrenzen lassen“* (ebd., S. 90). Die Unterteilung des Kontinuums in Kompetenzstufen erfolgt für die drei Subskalen getrennt. *„Die Kompetenzstufen der Gesamtskala Lesekompetenz wurde über die Aufgaben der entsprechenden Kompetenzstufen aus den drei Subskalen bestimmt“* (ebd., S. 88).

Die Aufgaben der einzelnen Kompetenzstufen beinhalten bestimmte Merkmale und Anforderungen. Diese sind für Aufgaben der gleichen Kompetenzstufe homogen, für verschiedene Kompetenzstufen unterscheiden sie sich systematisch (vgl. PISA 2000, S. 88). Becker-Mrotzek & Drommler (2006) fassen die Merkmale und Anforderungen der Kompetenzstufen folgendermaßen zusammen:

- *„Kompetenzstufe I: Oberflächliches Verständnis einfacher Texte. Die notwendige Information muss im Text deutlich erkennbar sein und der Text darf nur wenig konkurrierende Elemente enthalten, die von der relevanten Information ablenken könnten.*
- *Kompetenzstufe II: Herstellen einfacher Verknüpfungen. Leser sind in der Lage, einfache Verknüpfungen zwischen verschiedenen Teilen eines Textes herzustellen, auf die Bedeutung einzelner Elemente durch einfache Schlussfolgerungen zu schließen und so den Hauptgedanken eines relativ vertrauten Textes zu identifizieren.*
- *Kompetenzstufe III: Integration von Textelementen und Schlussfolgerungen. Leser sind in der Lage, verschiedene Teile des Textes zu integrieren, auch wenn die einzubeziehende Information wenig offensichtlich ist, mehrere Kriterien zu erfüllen hat und ihre Bedeutung indirekt erschlossen werden muss.*

- *Kompetenzstufe IV: Detailliertes Verständnis komplexer Texte. Leser können mit unvertrauten Texten umgehen, eingebette Informationen nutzen und sie den Anforderungen entsprechend organisieren.*
- *Kompetenzstufe V: Flexible Nutzung unvertrauter, komplexer Texte. Expertenleser, die auch komplexe, unvertraute und lange Texte für verschiedene Zwecke nutzen können.“*

(Becker-Mrotzek & Drommler, 2006, S. 38)

Festzuhalten ist, dass sowohl die Subskalen als auch die Kompetenzstufen zum größten Teil mit Hilfe statistischer Verfahren ermittelt wurden. Kritik an diesem Vorgehen äußern die Autoren des PISA-Berichtes selber, wenn sie behaupten, dass durch die sehr pragmatischen Kriterien, die durch das Board of Participating Countries festgelegt wurden, die Nutzung des Potentials des Materials eingeschränkt ist (vgl. PISA 2000, S. 90).

1.1.4 Texte, Tabellen und Diagramme

Neben den Subskalen und Kompetenzstufen unterscheidet PISA zudem kontinuierliche von nicht-kontinuierlichen Texten. *„Kontinuierliche Texte bestehen normalerweise aus Sätzen, die in Absätzen organisiert sind. Sie können Teil von größeren Strukturen wie Abschnitten, Kapiteln oder Büchern sein“* (PISA 2000, S. 80), hingegen liegen nicht-kontinuierliche Texte *„häufig in Matrix-Format vor und beruhen auf Listen“* (ebd.). Unter nicht-kontinuierliche Texte fallen Tabellen, Grafiken und Abbildungen. Prinzipiell ließen sich für kontinuierliche und nicht-kontinuierliche Texte jeweils Subskalen und Kompetenzstufen bilden. Etwa 2/3 der Aufgaben in PISA werden zu kontinuierlichen Texten gestellt.

1.2 Weitere Lesetests

Handhabbare Lesetests liegen vor allem für Kinder vor. Sie sind meist Teil von Schulleistungs- oder Schulfähigkeitstests und dienen entweder der Früherkennung von Lese- und Rechtschreibschwierigkeiten (z.B. das „Bielefelder Screening zur Früherkennung von Lese- und Rechstschreibschwierigkeiten (BISC)“ von Jansen, Mannhaupt, Marx & Skowronek, 1999) oder zur Diagnose einer allgemeinen Schulfähigkeit. Lesetests sind dort meist kurze Subtests einer umfassenderen Testbatterie. Der Allgemeine Schulleistungstest für 4. Klassen (AST 4) von

Fippinger (1978) z.B. überprüft in seinem Untertest „Leseverständnis" vor allem das Sinnverständnis beim Lesen. Dargeboten werden zwei Geschichten mit jeweils 9 Mehrfach-Wahl-Aufgaben. Die Geschichten sind dem 4. Schuljahr angemessen. In seiner Neuauflage (Fippinger, 1992) ist der Lesetest nicht mehr enthalten.

Ein weiteres Einsatzgebiet von Lesetests beschäftigt sich mit dem Feststellen der Eignung zum Studium in Deutschland für Studierende mit Deutsch als Fremdsprache (TestDaF) (vgl. Arras & Grotjahn, 2002). Zudem gibt es einige Tests für Schüler mit Deutsch als Fremd- und Zweitsprache. Da die meisten der Berufsschüler jedoch Deutsch nicht als Fremdsprache sprechen (wenngleich einige von ihnen Migrationshintergrund haben), eignen sich diese Tests nur bedingt für Berufsschüler, in denen auch Schüler mit Abitur lernen.

Ein Lesetest speziell für Berufsschüler eines Alters, in dem die Entwicklung der Lesekompetenz bereits abgeschlossen sein sollte, fehlte bislang.

1.3 Zusammenfassung

Lesen setzt sich aus vielen Teilprozessen zusammen. Zunächst müssen Buchstaben und Wörter wahrgenommen und entschlüsselt, dann müssen die Wörter zu Sätzen und schließlich zu längeren Texten zusammengesetzt und verstanden werden. Lesekompetenz kann also auf all diesen Stufen gemessen werden. Der Lesetest der PISA-Studie geht jedoch von einem Leser aus, dem es möglich ist, Wörter und Sätze zu entschlüsseln. Das Literacy-Konzept sieht das Lesen als Basisqualifikation. Gemessen wird dort ein Umgang mit Fließtexten und anderen Textarten, der befähigt, das eigene Potential weiterzuentwickeln. Der PISA-Test zielt also gerade nicht auf die grundlegenden Fähigkeiten der Worterkennung ab. Dieses Konzept lässt sich für die Entwicklung eines Lesetests für Berufsschüler fruchtbar machen, denn ein Leseverständnistest für diese Altersklasse existiert nicht. In den Berufsschulen und Berufskollegs sollte es den Schülern möglich sein, Wörter zu erkennen und sie zu verstehen. Wichtiger ist dort jedoch ein tieferes Verständnis der Texte. Somit beschreibt das folgende Kapitel, welche spezifischen Anforderungen an einen Test zum Leseverständnis durch die besondere Situation der Berufsschule gestellt werden.

2. Berufskollegs

Im Focus des Lesetests stehen Berufsschüler - und schülerinnen. Dies ist durch das Projekt „Sprachförderung in der Berufsbildung“ so festgelegt. Die nächsten beiden Kapitel sollen jedoch zeigen, inwieweit diese Sichweise für eine Testentwicklung wichtig ist.

2.1 Duales System der Berufsausbildung

Unter dem dualen System der Berufsausbildung versteht man die gleichzeitige Ausbildung in Betrieb und in der (Berufs-) Schule. Damit ist zugleich eine Verbindung zwischen privater und staatlicher Verantwortung gegeben. Die berufspraktische Ausbildung findet im Betrieb statt, der theoretische sowie der allgemeinbildende Unterricht in der Berufsschule. Die Vorteile dieses Systems liegen in der Praxisnähe des Unterrichts, eines relativ geringen Bürokratisierungsgrades und damit verbunden eine Kostengünstigkeit bezüglich des staatlichen Budgets. (vgl. Kaiser und Pätzold, 1999)

Das Berufskolleg ist eine *„Bildungseinrichtung, in der die beruflichen Qualifikationen und Abschlüsse entweder doppeltqualifizierend zusammen mit den allgemeinbildenden Abschlüssen der Sekundarstufe I und II [...] oder für sich alleine erworben werden können“* (ebd., S. 121). Daraus wird deutlich, dass es nicht nur um die Vermittlung des „theoretischen Teils“ des Berufs geht, denn darüberhinaus sollen umfassende Handlungskompetenzen vermittelt werden, mit dem Ziel, dass die Schüler *„lernen, sich aktiv und verantwortungsbewußt in ihren Berufen sowie in der privaten und öffentlichen Lebenswelt einzubringen und die Entwicklungen dort mitgestalten“* können (Kusch, 2006, S. 14).

Bei Abschluss eines Ausbildungsvertrages ist der Schüler berechtigt bzw. verpflichtet eine Berufsschule zu besuchen. Da für den Abschluss eines Ausbildungsvertrages keine formalen Voraussetzungen festgelegt sind, ist es den ausbildenden Betrieben überlassen, wen sie ausbilden. Das führt in der Berufsschule bezüglich der Lernvoraussetzungen und -bedürfnissen zu sehr heterogenen Klassen. Diese Tatsache lässt sich für eine Testentwicklung positiv nutzen, da, wie weiter unten beschrieben wird, eine große Variabilität zu größerer Zuverlässigkeit des Tests führt. Außerdem läßt sich der zuvor erbrachte Schulabschluss zur Validierung nutzen.

2.2 Forderungen an einen Lesetest für Berufsschüler

Die Lesekompetenz ist von elementarer Bedeutung für den Erfolg der Ausbildung und für die sich daran anschließende berufliche Tätigkeit. Die angehenden Fachkräfte müssen sich in der Ausbildung und im Beruf fachliche Inhalte aus Fachbüchern, Fachberichten und Produktinformationen erschließen. Neben Fließtexten müssen auch Tabellen und Diagramme gelesen werden. Gerade an Berufskollegs nehmen diese Darstellungsarten eine zentrale Rolle ein, literarische Texte hingegen sind eher peripher wichtig. Lesekompetenzdiagnostik an Berufskollegs muss also auch das Verstehen und Deuten dieser Darstellungsarten in den Blick nehmen.

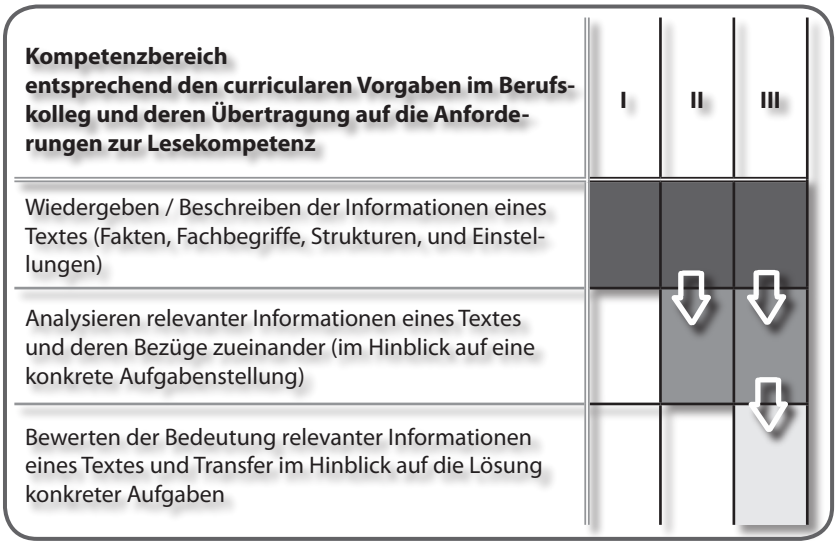
Die Lesekompetenz hat maßgeblichen Einfluss auf den Erfolg und auf die Effizienz der Lernprozesse im Betrieb und insbesondere in der Berufsschule. Wem es an der erforderlichen Lesekompetenz mangelt, hat geringere Chancen, die Gesellen- und Facharbeiterprüfung erfolgreich zu bestehen und damit den Grundstein für die eigene berufliche Karriere zu legen.

Ein zu entwickelnder Lesetest für die dualen Bildungsgänge muss die Ziele und die Rahmenbedingungen der dualen Ausbildung berücksichtigen. Die sich daraus ergebenden Anforderungen beschreiben Stevens und Wahlers (2006, S. 73) folgendermaßen:

- *„Konzentration auf wesentliche in der Berufswelt und in der Ausbildung relevante Textarten*
- *Offenlegung der darauf bezogenen individuellen Stärken und Förderbedarfe der einzelnen Schüler*
- *Bereitstellung von möglichst konkreten Diagnoseergebnissen, die Möglichkeiten für eine fächerübergreifende Leseförderung bieten.*
- *Praktikabilität im Hinblick auf den Umfang der Durchführung und Auswertung des Tests im Unterrichtsalltag“*

Weitere Anforderungen ergeben sich aus den curricularen Vorgaben im Berufskolleg. Ein Lesetest für Berufsschüler muss diese natürlich auch berücksichtigen. Abbildung 3 fasst die Kompetenzbereiche entsprechend den curricularen Vorgaben im Berufskolleg zusammen.

Abbildung 3: Kompetenzbereiche des Lesens im Berufskolleg, aus: Stevens & Wahlers, 2006, S. 74.



2.3 Zusammenfassung

Die Schülergruppe in Berufskollegs ist hinsichtlich ihrer Leistungen sehr heterogen. Dies ist eine besondere Herausforderung für die Berufskollegs, denn bezüglich des Lesens befinden sich die Schüler auf höchst unterschiedlichem Niveau. Ein Test zur Lesekompetenz, der zu einer gezielten Förderung herangezogen werden kann, liegt bisher nicht vor und muss nach den Anforderungen der Berufskollegs entwickelt werden. Das Kompetenzkonzept, das der PISA-Studie zu Grunde liegt, scheint geeignet, jedoch ist der dort konkret durchgeführte Test nicht vollkommen auf die Personengruppe der Berufsschüler zu übertragen. Der Lesetest der PISA-Studie ist zudem nicht öffentlich zugängig und nicht für diagnostische Zwecke konzipiert. Im nächsten Kapitel werden daher die besonderen Anforderungen der Berufskollegs mit dem PISA-Modell verknüpft.

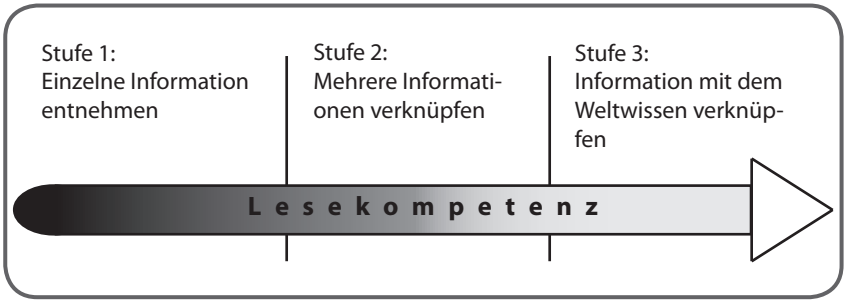
3. „Lesetest für Berufsschüler/innen“ - Anforderungen an die Entwicklung

Ziel ist es, einen Lesetest für Berufsschüler zu entwickeln, der sich am Konzept der Reading Literacy, das im Teil I, Kapitel 1 vorgestellt wurde, orientiert. Er hat jedoch eine andere Personengruppe im Blick, nämlich Berufsschüler und -schülerinnen.

3.1 Anforderungen

Die in PISA (2001) zu Grunde gelegte Definition der Lesekompetenz und die Konzeption des „Arbeitens mit dem Text“ soll für den zu entwickelnden Test übernommen werden. Dem dreiskaligen Modell in PISA wird ein Modell mit nur einer Skala „Lesekompetenz“ gegenübergesetzt. Neben einer normorientierten Auswertung soll die Skala in drei Kompetenzstufen eingeteilt werden (s. Abbildung 4). Die Fragen sollen bereits der inhaltlichen Konzeption der Kompetenzstufen entsprechen. Insofern liegt ein Unterschied zur PISA-Studie darin, dass die Kompetenzstufen nicht nachträglich aus den empirischen Daten „geschätzt“ werden sondern bereits vorher inhaltlich manifestiert sind.

Abbildung 4: Lesekompetenzmodell des LTB[-3]



Die Bedingung, dass für die Durchführung des Tests im Schulalltag nur eine geringe Zeit zur Verfügung stehen kann, führt dazu, dass der Test nur die für die Berufskollegs wichtigsten Aspekte des PISA-Strukturmodells (vgl. Abbildung 1) beinhalten kann.

Die drei Kompetenzstufen fordern vom Schüler folgende Fähigkeiten:

- Die erste Kompetenzstufe soll die Fähigkeit umfassen, explizite Informationen eines Textes zu erfassen und korrekt und umfassend wiederzugeben. Diese Kompetenz umfasst im Wesentlichen den Aspekt „Informationen ermitteln" des PISA-Strukturmodells.

- Die zweite Kompetenzstufe soll über diese einfache Fähigkeit hinaus gehen. Es wird gefordert, dass der Proband relevante Informationen und deren Bezüge zueinander im Hinblick auf eine konkrete Aufgabenstellung analysieren kann. Dies umfasst die selbstständige Auswahl und Herausarbeitung von Parallelen, Entwicklungen und Kontrasten, bis hin zur Strukturierung der wesentlichen Informationen und der Konstruktion von Transferbezügen. Diese Stufe bezieht sich am Ehesten auf den Aspekt „eine textbezogene Interpretation entwickeln".

- Die dritte Kompetenzstufe soll darauf ausgerichtet sein, dass der Proband relevante Kerninhalte der Texte, Tabellen und der Grafik bewerten und sie zu dem eigenen Weltwissen in Bezug setzen kann. Erst diese Fähigkeit ermöglicht ein kritisches Hinterfragen des Textes und der eigenen Position sowie einen Transfer zur Lösung einer konkreten Aufgabe. Diese Kompetenzstufe bezieht sich auf den Aspekt „über den Inhalt des Textes reflektieren" des PISA-Strukturmodells.

Diese Anforderungen implizieren, dass die Kompetenzstufen hierarchisch aufeinander aufbauen müssen: Ohne einzelne Informationen aus dem Text extrahieren zu können, ist es nicht möglich, mehrere Informationen zu verknüpfen. Ebensowenig ist es dann möglich, diese mit dem Weltwissen zu verbinden und so zu einem reflektierten Standpunkt oder einer geforderten Problemlösung zu gelangen. Eine Kompetenzstufe stellt also eine Voraussetzung zum Erreichen der nächst höheren Stufe dar.

Den spezifischen Anforderungen der Berufskollegs entsprechend sollen im „Lesetest für Berufsschüler/innen" ähnlich wie im PISA-Test Fragen zu Texten, Diagrammen und Tabellen gestellt werden. Die verschiedenen Anforderungen für die einzelnen Textarten verdeutlicht

Tabelle 1. Für jede dieser Textarten sollen getrennt Kompetenzstufen berechnet werden.

Es sollen ausschließlich nicht-fiktionale Texte dargeboten werden. Um die Motivation der Berufsschüler, zu erhalten, sollen Authentische Texte, Tabellen und Diagramme aus der Lebenswelt der Berufsschüler die nur geringfügig modifiziert werden, eignen sich dazu, ihre Motivation zu erhalten. Die Texte dürfen nicht zu schwierig und nicht zu lang, inhaltlich jedoch kohärent sein. Es muss möglich sein, genügend Fragen pro Kompetenzstufe zu erstellen.

Tabelle 1: Anforderungen der drei Kompetenzstufen an die Schüler für die verschiedenen Textarten

Kompetenzstufe	Texte (Fließtexte)	Diagramme	Tabellen
1 Einzelne Informationen entnehmen	Die Schüler können richtig und vollständig explizit gegebene Einzelinformationen wiedergeben	Die Schüler können Einzelwerte und Skalierungen benennen.	Die Schüler können explizite Einzelwerte oder Spalten- und Zeilenbezeichnungen benennen.
2 Mehrere Informationen verknüpfen	Die Schüler können implizite Informationen entnehmen.	Die Schüler können erkennbare Zusammenhänge (Entwicklungen, Kontraste, Parallelen) zwischen Einzelwerten ableiten.	Die Schüler können Entwicklungen und Zusammenhänge zwischen Einzelwerten ableiten. Die Schüler können Beziehungen zwischen komplexen Zusammenhängen ableiten, wie den Vergleich zweier Spalten oder Zeilen.

Kompetenz-stufe	Texte (Fließtexte)	Diagramme	Tabellen
	a) Die Schüler können im Text dargestellte kausale Zusammenhänge, z.B. Argumentationsstrukturen, erkennen. b) Die Schüler können implizite Informationen verstehen, die durch sprachliche Mittel wie Diminution, Pejoration, Hyperbel usw. ausgedrückt werden.	a) Die Schüler können Beziehungen zwischen mehreren Zusammenhängen ableiten, z.B. einen Vergleich zweier Kurven. b) Die Schüler können die Zuverlässigkeit einer Quelle ableiten, die u.a. durch die grafische Gestaltung, das Layout oder Hervorhebungen und den damit verbundenen Effekten auf den Leser ausgedrückt werden.	
3 Information mit dem Weltwissen verknüpfen	Die Schüler können das aus dem Text abgeleitete Wissen bewerten und auf neue Kontexte außerhalb des Textes anwenden.	Die Schüler können das aus dem Text abgeleitete Wissen bewerten und auf neue Kontexte außerhalb des Textes anwenden.	Die Schüler können das aus dem Text abgeleitete Wissen bewerten und auf neue Kontexte außerhalb des Textes anwenden.

3.2 Zusammenfassung

Der zu entwickelnde Lesetest lehnt sich stark an den Lesekompetenzbegriff der PISA-Studie an, geht jedoch auf die Bedürfnisse der Berufskollegs ein. Das Strukturmodell der PISA-Lesekompetenz, das fünf Aspekte beinhaltet, wird auf drei Kompetenzen reduziert. Diese werden nicht empirisch ermittelt sondern sollen bereits bei der Itemkonstruktion berücksichtigt werden. Die Kompetenzstufen bauen, im Unterschied zur PISA-Studie, hierarchisch aufeinander auf. Im nächsten Kapitel werden drei grundlegend verschiedene testtheoretische

Modelle vorgestellt und im Hinblick auf ihre Tauglichkeit für die Konstruktion des zu entwickelnden Lesetests diskutiert.

4. Testtheorie

Auf die Frage, was ein Test ist, gibt Lienert (1989, S. 7) folgende Definition:

> *„Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung.“*

Rost (1996) erweitert diesen Begriff dahingehend, dass die Aussage nicht ausschließlich quantitativer Natur sein muss. *„Vielmehr können auch **qualitative, d.h. kategoriale** Aussagen [...] Ziel des Testens sein“* (vgl. dort, S. 18, Hervorhebung im Original). Aus dem Antwortverhalten im Test wird auf ein psychisches Merkmal geschlossen. Dies macht eine Theorie über den Zusammenhang zwischen psychischem Merkmal und Testverhalten nötig. Eine einheitliche Theorie darüber gibt es nicht, jedoch stehen verschiedene formale Modelle zur Verfügung. *„Ein formales Modell, dass auf einen Inhalt angewendet wird, wird zu einer Theorie“* (Rost, 1996, S. 25). Eins der formalen Modelle ist das bereits erwähnte Rasch-Modell, die Item-Response-Theorie (IRT) (vgl. dort, S. 120 ff.), ein anderes das der „klassischen“ Testtheorie. Beide Modelle arbeiten normorientiert. Das Modell einer durchgehend kriteriumsorientierten Messung ist dem entgegengesetzt. Es orientiert sich nicht an einer Norm, die verstanden wird, als der Durchschnitt einer Menge von Probanden, sondern an einem festgelegten Kriterium . Zunächst werden die drei Messmodelle - die Klassische Testtheorie, das Rasch-Modell und das kriteriumsorientierte Modell - ausführlicher beschrieben. Nachdem die wesentlichen Unterschiede erläutert wurden, wird erklärt, aus welchem Grund für den zu entwickelnden Test das Modell der Klassischen Testtheorie gewählt wurde.

4.1 Klassische Testtheorie

*„Die Bezeichnung **Klassische Testtheorie** [...] hat sich eingebürgert, weil diese den ersten Ansatz darstellt, der Interpretation psychologischer Messung eine mathematische Fundierung zu geben“* (Hervorhebung im Original, Perleth und Sierwald, 1991, S. 46). Im Folgenden werden kurz die Grundannahmen der Klassischen Testtheorie beschrieben, daran anschließend werden die wichtigsten Gütekriterien diskutiert.

4.1.1 Axiome der Klassischen Testtheorie

Die Klassische Testtheorie wird wegen ihres besonderen Umgangs mit dem Messfehler auch „Messfehlertheorie“ genannt. Sie geht zunächst davon aus, dass es eine „wahre“ Eigenschaftsausprägung gibt. Diese lässt sich zwar messen, dabei tritt jedoch immer ein Messfehler auf. Dieser Messfehler lässt sich beschreiben.

Der „wahre Wert“ einer Person wird definiert als der Mittelwert unendlich häufiger Messungen (Erwartungswert ε) einer Variable. Die Abkürzung für den wahren Wert ist τ (für „true“). Bei einer einzelnen Messung setzt sich der Beobachtungswert x aus dem wahren Wert τ und dem Messfehler e zusammen. Die Grundgleichung der Klassischen Testtheorie lautet also $x_i = \tau_i + e_i$ (vgl. Huber, 1973, S. 49). Das Konzept des wahren Wertes wird wichtig bei der Konstruktion eines Vertrauensintervalls um den wahren Wert. Misst man nun (hypothetisch) unendlich viele Personen jeweils einmal, werden einige unterschätzt, andere überschätzt. Der Erwartungswert ε der Messfehler E ist 0, $\varepsilon(E) = 0$. Dies ist das erste Axiom der Klassischen Testtheorie. Das zweite Axiom besagt, dass der Messfehler nicht mit dem wahren Wert korreliert, also $r(E,\tau) = 0$. Auch darf der Messfehler nicht mit irgendeiner anderen Variable Y korrelieren $(r(E_X,\tau_Y)) = 0$ (drittes Axiom). Das vierte Axiom besagt, dass der Messfehler nicht mit dem Messfehler der Variable Y korrelieren darf, also $r(E_X,E_Y) = 0$ (vgl. Rost, 1996, S. 35).

Aus diesem formalen Modell wird eine Testtheorie, wenn man sie auf konkrete Messwerte anwendet. *„In diesem Fall macht die Messfehlertheorie die Aussage, dass ein Messwert X **nur eine latente Variable** τ_X repräsentiert und die Abweichungen der Messwerte von den wahren Variablenausprägungen den **Erwartungswert 0 haben und mit nichts anderem korrelieren**“* (ebd., S. 39, Hervorhebungen im Original).

4.1.2 Gütekriterien

In allen Testtheorien stellt sich die Frage, was einen guten Test auszeichnet. Dies ist auch in der Klassischen Testtheorie nicht anders. Hier werden vorrangig drei Testgütekriterien beschrieben (vgl. Lienert 1989, S. 12): Objektivität, Reliabilität und Validität. Als weitere Nebenkriterien werden genannt, dass der Test normiert, vergleichbar, ökonomisch und nützlich sein soll.

Objektivität

Die Objektivität ist ein Maß für die Unabhängigkeit der Testung vom Versuchsleiter. Man unterscheidet zwischen Durchführungs-, Auswertungs- und Interpretationsobjektivität je nachdem auf welche Phase sich das Gütekriterium bezieht. Die Durchführungsobjektivität lässt sich durch Standardisierung der Anweisung erhöhen. Eine Steigerung der Auswertungsobjektivität lässt sich erreichen, indem die Auswertung des Tests nach genau vorgegebenen Regeln stattfindet, also etwa mit Hilfe einer Schablone oder maschineller Verfahren. Eine hohe Interpretationsobjektivität ist dann gegeben, wenn aus den gleichen Auswertungsergebnissen die gleichen Schlüsse gezogen werden. Bei Leistungstests, in denen das Ergebnis ein Skalenwert ist, der die zu messenden Fähigkeit abbildet, ist diese Art der Objektivität trivial.

Reliabilität

Die Reliabilität ist ein Maß für die Genauigkeit der Messung, also inwieweit der Messwert (Beobachtungswert) mit dem wahren Wert übereinstimmt. Testtheoretisch ist die Reliabilität definiert als der Anteil der Varianz der wahren Werte an der Varianz der Beobachtungswerte einer Grundgesamtheit. Sie liegt zwischen 0, was fehlende, und 1, was vollständige Reliabilität bedeutet. Da man die Varianz der wahren Werte nicht kennt, lässt sich die Reliabilität nur schätzen. Dazu kommen mehrere Verfahren zur Ermittlung des Reliabilitätskoeffizienten in Betracht.

Durch die Korrelation zweier Messwerte an derselben Person in einem zeitlichen Abstand erhält man die Wiederholungs- oder Retestreliabilität. Sie ist ein Maß für die zeitliche Stabilität eines Tests. Die Wiederholungsreliabilität ist nicht geeignet, stark veränderliche Merk-

male zu erfassen. Entwicklungs- und Erinnerungseffekte werden dem Messfehler zugerechnet.

Ist die Gefahr eines Lerneffektes zu groß, bedient man sich eines Paralleltests. Von Paralleltest spricht man, wenn die beiden Testformen in Inhalt und Aufbau vergleichbar sind. Streng genommen müssen sie die gleichen Mittelwerte und eine übereinstimmende Verteilung besitzen. (vgl. Huber, 1973, S. 53) In dieser Stringenz sind sie in der Praxis jedoch kaum zu konstruieren. Erinnerungs- und Lerneffekte sind kleiner als bei der Retestreliabilität, jedoch können Entwicklungseinflüsse zum Verlust der Parallelität führen. (vgl. Perleth und Sierwald, 1991, S. 56)

Zwei weitere Verfahren werden der Reliabilität zugeordnet, jedoch zumeist Verfahren der Konsistenzanalyse genannt (vgl. Perleth und Sierwald, 1991, S. 56 f.). Nach der Testhalbierungsmethode wird der Test in zwei gleichgroße Teile geteilt und für jede dieser Testhälften ein Punktwert ermittelt. Die Korrelation dieser Hälften ist dann ein Schätzer für die innere Konsistenz des Tests (Testhalbierungs- oder Splithalf-Reliabilität). Sinnvollerweise halbiert man den Test so, dass in der einen Hälfte die Items mit geraden Nummern, in der anderen Hälfte die Items mit ungeraden Nummern sind (Odd-even-Methode), denn dies schließt weitgehend eine Einteilung in zwei unterschiedlich schwierige Teile aus, wenn die Items nach steigender Schwierigkeit geordnet sind. Da ein längerer Test zuverlässiger ist als ein kurzer, lässt sich der Koeffizient mit Hilfe einer von Spearman und Brown entwickelten Formel korrigieren. Anstatt den Test in zwei Hälften einzuteilen, ist es auch möglich ihn in so viele Teile zu teilen, wie er Items hat. Mit Hilfe einer von Cronbach entwickelten Formel wird die innere Konsistenz berechnet.

Obwohl die Reliabilität ein Maß für die Genauigkeit eines Messinstrumentes ist, hängt sie neben der Güte des Messverfahrens von verschiedenen Einflüssen ab (vgl. Perleth und Sierwald, 1991, S. 58 f.). Homogenität bzw. Heterogenität der Personenstichprobe beeinflussen sie genauso wie die Homogenität bzw. Heterogenität der Items. Desweiteren ist die Methode der Reliabilitätsbestimmung bedeutsam. Bei einer homogenen Stichprobe ist die ermittelte Reliabilität kleiner als bei einer heterogenen. Wenn alle Probanden den gleichen wahren Wert hätten, würde sie sogar auf Null sinken ($r_{tt} = 0$). Auf diesen Tatbestand wird im Zusammenhang mit kriteriumsorientierten Tests noch einzugehen sein. *„Die Reliabilität ist umso größer, je heterogener die Stichprobe*

ist, an der sie bestimmt wird" (Perleth und Sierwald, 1991, S. 59). Auch ist die Reliabilität für heterogene Items geringer als für homogene. „*Die Reliabilität ist umso größer, je homogener das Messverfahren (also das Aufgabenset) ist*" (ebd.). Wie bereits aus der Beschreibung der verschiedenen Methoden zur Reliabilitätsbestimmung hervorgeht, beeinflusst auch das Verfahren die Höhe der Reliabilität, denn „*jeder Reliabilitätskoeffizient schlägt demnach andere Komponenten des Messwertes dem Fehlerwert bzw. dem wahren Wert zu*" (ebd.). Perleth und Sierwald (1991) geben für praktische Zwecke konkrete Anforderungen an die Höhe der Reliabilität an. Für Leistungstests sollte die Paralleltest- und Retestreliabilität zwischen 0,80 und 0,90 liegen, für die Konsistenzreliabilität zwischen 0,85 und 0,95.

Validität

Unter Validität versteht man den Grad, in dem ein Test mißt, was zu er messen vorgibt. Auch für die Validität gibt es mehrere Ansätze, jedoch ist das Konzept nicht so einheitlich zu fassen, wie das der Reliabilität, „*es existiert kein einheitlicher Katalog von Validitätsmerkmalen*" (Perleth und Sierwald, 1991, S. 63). Die meisten Autoren unterscheiden jedoch drei Arten der Validität, die zum Teil noch weiter differenziert werden: die inhaltliche Validität, die Kriteriumsvalidität und die Konstruktvalidität.

„*Unter inhaltlicher Validität eines Messverfahrens ist das Ausmaß zu verstehen, in dem das Gemessene das zu Messende unmittelbar repräsentiert*" (Perleth und Sierwald, 1991, S. 64), mit anderen Worten: Der Beobachtungswert steht nicht für eine latente Variable sondern für sich selbst. Perleth und Sierwald (1991, S. 64) nennen als Beispiel einen Test zum Schreibmaschineschreiben. Die gemessene Variable steht direkt für die Fertigkeit des Maschineschreibens, nicht für eine latente Variable. Ein Verfahren für die inhaltliche Validität gibt es nicht. Deshalb wird sie im Rahmen der Klassischen Testtheorie selten gefordert, im Rahmen der kriteriumsorientierten Tests ist diese Art der Validität einer der Kernpunkte.

Die Kriteriumsvalidität ist die Korrelation der Messwerte im Test mit einem Wert des Kriteriums. Das Kriterium ist eine Variable, die das Konstrukt, das dem eigentlichen Test zugrunde liegt, wenigstens zum Teil repräsentiert. Bei der Übereinstimmungsvalidität werden Testwert und Kriteriumswert gleichzeitig erhoben. Nützlich kann diese Art der

Validität sein, wenn der neu erstellte Test hoch mit einem älteren Test korreliert, jedoch wesentlich ökonomischer ist. Bei der Vorhersagevalidität lassen sich Prognosen aus dem Testergebnis erstellen. So kann mit Hilfe des Testergebnisses z.B. Schul- oder Berufserfolg vorhergesagt werden.

„Unter Konstruktvalidität versteht man das Ausmaß, in dem das Messverfahren das theoretische Konstrukt, das es messen soll, erfasst" (Perleth und Sierwald, 1991, S. 67). Ein Konstrukt ist eine nicht beobachtbare latente Eigenschaft, die mittels Operationalisierung „sichtbar" gemacht wird. Diese Operationalisierung kann mehrere Verhaltensweisen oder Merkmale als Indikatoren des Konstruktes umfassen, die empirisch gemessen werden können. Wie die inhaltliche Validität gibt es auch zur Konstrukvalidität keine Kennwerte. Dieterich (1977, zitiert nach Perleth und Sierwald, 1991, S. 68) gibt jedoch einige Verfahren zur Überprüfung der Konstruktvalidität an. Unter anderen nennt er folgende.

- Neben der inhaltlichen Validierung (s.o.) spielt die Überprüfung der Homogenität eines Tests eine große Rolle. Sie kann mit Hilfe einer Faktorenanalyse und ähnlichen Verfahren kontrolliert werden. Eine hohe Homogenität, die sich zum Beispiel in einem Faktor widerspiegelt, spricht dafür, dass es sich um einen Test handelt, der ein eng umgrenztes Konstrukt beinhaltet. Bei einem sehr heterogenen Test mit sehr vielen Faktoren muss theoretisch fundiert dargelegt werden, wieso sich gerade diese Faktoren zu einem Konstrukt zusammenfassen lassen.

- Die konvergente Validität zielt darauf ab, dass sich zu anderen empirischen Indikatoren positive, zu theoretisch nicht verwandten Konstrukten negative oder keine Korrelationen finden lassen. Eine Faktorenanalyse mit den Testitems und Items aus anderen Tests, die das gleiche und andere Konstrukte messen, muss zeigen, dass die Items des gleichen Konstuktes auf dem gleichen Faktor laden, während andere Items auf anderen Faktoren laden.

- Es muss zu zeigen sein, dass es Korrelationen zu anderen empirischen Indikatoren gibt, die mit dem Konstrukt in Zusammenhang stehen. Korrelationen zu Indikatoren, die für Konstrukte stehen, die theoretisch nicht mit dem Konstrukt in Beziehung stehen, sollten nur geringe Werte aufweisen.

Nebengütekriterien

Neben den Hauptgütekriterien spielen vier weitere Kriterien eine Rolle in der Testentwickung.

Die Normierung eines Tests ist zu diagnostischen Zwecken nötig. Ein nicht normierter Test hat eine geringe diagnostische Brauchbarkeit, auch wenn er als Forschungsinstrument durchaus brauchbar ist.

Mit „Vergleichbarkeit" meint Lienert (1989, S. 18) einen Vergleich des Tests mit einer Parallelform oder mit validitätsähnlichen Tests zur Kontrolle der intraindividuellen Reliabilität - oder Validität.

Ein Test muss ökonomisch, d.h. kurz und einfach durchzuführen, einfach und bequem auszuwerten sein und wenig Material beanspruchen.

Darüberhinaus sollte der Test nützlich sein, d.h. es muss ein praktisches Bedürfnis bestehen und das Personenmerkmal sollte nicht mit anderen Tests ebensogut untersucht werden können.

Für die Nebenkriterien lassen sich keine Kennwerte berechnen.

4.1.3 Itemanalyse

Ziel der Itemanalyse ist es, die Kennwerte der genannten Gütekriterien zu verbessern. Das geschieht in der Regel, indem die Items auf ihre Brauchbarkeit im Hinblick auf das zu messende Konstrukt untersucht werden. Die wesentlichen Kennwerte dazu sind der Trennschärfekoeffizient und die Aufgabenschwierigkeit. Auch eine Analyse der Distraktoren, also der falschen Antwortmöglichkeiten bei einem Multiple-choice Test, sind von Bedeutung.

Trennschärfe

Die Trennschärfe ist ein Maß dafür, inwieweit Items zwischen guten und schlechten Probanden trennen. Sie ist die Korrelation zwischen Item und Skalenwert bei einer Stichprobe von Probanden. Der Skalenwert ist die Summe der richtigen Lösungen aller Items der betreffenden Skala. Diese Skala kann auch alle Items des Tests beinhalten. Da in diesem Skalenwert das betreffende Item enthalten ist, korrigiert man diese Korrelation, indem man das Item nicht in die Berechnung des Skalenwertes einbezieht. Man erhält so die korrigierte Trennschärfe. Der Wert liegt zwischen -1 und 1. Eine hohe Trennschärfe sagt aus,

dass das Item tatsächlich Teilaspekte des Konstruktes misst. Je höher die Trennschärfe, desto besser differenziert es zwischen guten und schlechten Probanden hinsichtlich der Gesamtpunktzahl des Tests. Im Sinne einer hohen Reliabilität und Differenzierungsfähigkeit sollten die Trennschärfen möglichst hoch sein. Items mit sehr geringer oder gar negativer Trennschärfe müssen aus dem Test ausgeschlossen werden, da sie nichts zum Testkonstrukt beitragen und den Test nur unnötig verlängern.

Schwierigkeitsindizes

Der Schwierigkeitsindex eines Items ist der Anteil der richtigen Antworten an der Gesamtzahl der Antworten auf ein Item die an einer Stichprobe ermittelt wurden. Eine zu hohe oder zu niedrige Schwierigkeit trennt nicht zwischen den Probanden und das Item sollte eliminiert werden. Eine Frage, die zu 100% richtig beantwortet wird, führt nur zu einer Addition einer Konstanten und verlängert den Test unnötig. Während eine hohe Trennschärfe angibt, dass ein Item gut zwischen Probanden trennt, gibt die Schwierigkeit an, wo das Item trennt. Im Sinne einer hohen Differenzierungsfähigkeit sollten deshalb Items den gesamten Bereich von 0 bis 1 abdecken. Soll der Test besonders im unteren Bereich differenzieren, kann es sinnvoll sein, gehäuft Fragen geringer Schwierigkeit (also hoher Schwierigkeitsindizes) in den Test aufzunehmen. Die korrigierte Schwierigkeit, die die Ratewahrscheinlichkeit berücksichtigt, ist bei der Itemselektion zu bevorzugen. Die Korrektur errechnet sich nach Lienert (1989, S. 89) nach folgender Formel:

$$P = 100 * \frac{N_R - \frac{N_F}{m-1}}{N}$$

N_R = Anzahl der VP, die die Aufgabe richtig beantwortet haben
N_F = Anzahl der VP, die die Aufgabe falsch beantwortet haben
m = Anzahl der Antwortalternativen
N = Gesamtzahl der VP

Zusammenhang zwischen Schwierigkeit und Trennschärfe

Trennschärfe und Schwierigkeit sind nicht unabhängig voneinander. Bei sehr hoher und sehr niedriger Schwierigkeit ist die Trennschärfe gering. Den höchsten Wert kann die Trennschärfe bei einer Schwierigkeit von 0,5 erreichen. Da aber der Test differenzieren soll, müssen auch andere Schwierigkeiten aufgenommen werden. Das vermindert jedoch die Reliabilität. Perleth und Sierwald (1991, S. 74) kommen zu folgendem Fazit: Die Itemauswahl ist *„so zu treffen, dass...*

- *die verbleibenden Items möglichst trennscharf (z.B. Trennschärfe >0,5) sind*

- *und die Schwierigkeitsindizes der verbleibenden Items in etwa Normalverteilung aufweisen, wobei Items mit extrem hohen (p > 0,8) und niedrigen (p < 0,2) Schwierigkeiten ausgeschlossen werden"* (Perleth und Sierwald, 1991, S. 74).

Langfeldt (1984, S. 98) gibt folgende Beurteilungskriterien für die Schwierigkeit und Trennschärfe an:

	niedrig	mittel	hoch
Schwierigkeit	≥ 0,80	0,80 bis 0,20	≤ 0,20
Trennschärfe	≤ 0,30	0,30 bis 0,50	≥ 0,50

„Ein brauchbarer Schulleistungstest sollte (mindestens) den unter „mittel" genannten Anforderung entsprechen" (ebd.)

Distraktorenanalyse

Bei Multiple-choice Aufgaben mit einer richtigen Antwort *„erweist sich jedoch sehr oft eine systematische Analyse der Nicht-R-Antworten, der sog. Alternativ-Antworten, als fruchtbar"* (Lienert, 1989, S. 120). Jeder Distraktor sollte, nach Abzug der richtigen Antworten, gleich häufig gewählt werden, und sie sollten *„nach psychologischer Erwartung von der schlechten Gruppe häufiger gegeben werden als von der guten Gruppe"* (ebd., S. 121).

4.1.4 Zusammenfassung

Die Klassische Testtheorie geht davon aus, dass sich psychische Merkmale, latente Variablen, messen lassen. Jede Messung ist jedoch fehlerhaft. Um eine möglichst fehlerfreie Messung zu erhalten, ist es nötig, mit Hilfe der Itemanalyse und -selektion diejenigen Fragen auszuwählen, die den gesetzten Ansprüchen der „Messfehlerfreiheit" und der „Gültigkeit" standhalten. Die Klassische Testtheorie bietet dazu Kennwerte wie Reliabilität respektive Validität. Die Reliabilität kann zur Schätzung des wahren Wertes herangezogen werden. Somit bietet die Klassische Testtheorie alle Voraussetzungen zur Entwicklung eines Lesetests.

4.2 Probabilistische Testmodelle

Probabilistische Testmodelle, eng mit dem Namen des dänischen Mathematikers Georg Rasch verbunden, existieren schon seit seinen ab etwa 1960 veröffentlichten Publikationen. Zur weiteren Ausarbeitung des Modells und zur praktischen Anwendung kommt es aber erst in neuerer Zeit, vor allem mit den großen internationalen Studien TIMMS (vgl. Martin & Kelly, 1998), PISA 2001 und anderen. Dies mag an einer zunehmenden Kritik an der Klassischen Testtheorie liegen, aber auch an den Anforderungen, die bei der praktischen Umsetzung des Modells an die Rechenkapazität von Computern gestellt wird, die in der 60er- und 70er Jahren nicht gegeben war. Die probabilistischen Testmodelle, die auch Item-Response-Modelle heißen, sind vielfältig (vgl. Perleth und Sierwald, 1991, S. 83) . Hier sollen jedoch nur die Grundannahmen beschrieben werden, die dem Rasch-Modell zugrunde liegen.

Das testtheoretische Konzept des Rasch-Modells ist ein Latent Trait Modell, d.h. die latente Variable „Fähigkeit" wird durch die gemessene Leistung erschlossen (vgl. Perleth und Sierwald, 1991, S. 78 ff.). Das Rasch-Modell erlaubt es, im Unterschied z.B. zur Klassischen Testtheorie, Personenparameter (Fähigkeit) Θ von Itemparametern (Schwierigkeit) ξ zu trennen. Da beide Parameter Wahrscheinlichkeitswerte sind, lassen sie sich auf der gleichen Skala abbilden. Die Beziehung zwischen dem latenten Fähigkeitsmerkmal und der Antwortwahrscheinlichkeit kann für jedes Item grafisch mithilfe der „Item Characteristic Curve" (ICC, s. Abbildung 5, S. 39) dargestellt werden. Die Lösungswahrscheinlichkeit eines Items bei einem Probanden wird

aus dem Fähigkeitsparameter f_i und der Schwierigkeit s_i geschätzt. Die Grundannahme ist also:

$$P(\text{Person v löst Aufgabe i}) = \frac{f_v}{f_v + s_i}$$

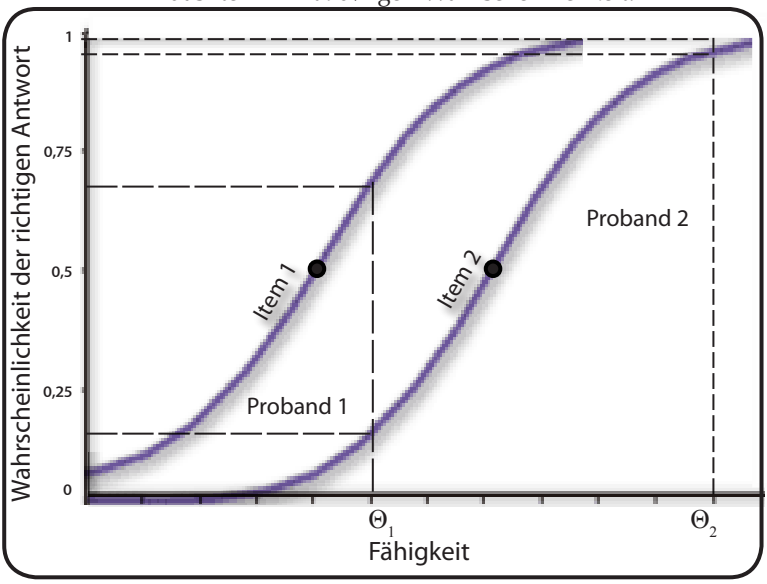
Diese Gleichung wird in ein logistisches Modell überführt. $\ln(f_v)$ wird durch Θ_v und $\ln(s_i)$ wird durch ξ_i ersetzt:

$$P(\text{Person v löst Aufgabe i}) = \frac{e^{(\Theta_v - \xi_i)}}{1+e^{(\Theta_v - \xi_i)}}$$

Die Wahrscheinlichkeit einer Aufgabenlösung geht gegen 1, wenn die Fähigkeit größer wird. Der Wendepunkt, an dem die Lösungswahrscheinlichkeit genau 0,5 ist, definiert die Schwierigkeit eines Items. Entsprechen sich Fähigkeit der Person Θ_v und Schwierigkeit der Aufgabe ξ_i, dann löst diese Person das Item mit einer Wahrscheinlichkeit von 50%. (Im PISA-Test wurde von diesem üblichen Vorgehen insofern abgewichen, dass die Lösungswahrscheinlichkeit auf 62 Prozent festgelegt wurde, d.h. ein Schüler mit einer bestimmten Fähigkeit löst ein Item mit der entsprechenden Schwierigkeit mit einer Wahrscheinlichkeit von 62%. *„Mit dieser pragmatischen Entscheidung wurde beabsichtigt, dass Personen, deren Fähigkeit der Itemschwierigkeit entspricht, dieses Item mit hinreichender Sicherheit (eben $p(x_{vi} = 0{,}62)$ statt nur $p(x_{vi} = 0{,}50)$ lösen“* (PISA 2000, S. 519).)

Eine zentrale Grundannahme des Rasch-Modells ist, dass der Test „spezifisch objektiv“ ist. Das bedeutet einerseits Itemunabhängigkeit (unabhängig von den Items lassen sich Differenzen zwischen den Personen feststellen) und Stichprobenunabhängigkeit (unabhängig von der Stichprobe lassen sich Differenzen zwischen den Itemschwierigkeiten feststellen). Die Itemunabhängigkeit spielte beim Multi-Matrix-Sampling der PISA-Studie eine wichtige Rolle (vgl. PISA 2000, S. 49 f.), da nicht jedem Schüler jedes Item vorgelegt werden musste und die Ergebnisse trotzdem auf einer gemeinsamen Skala abgebildet werden konnten.

Abbildung 5: Item-Characteristik-Kurven (ICC). Item 1 ist leichter als Item 2 (die schwarzen Punkte zeigen die Schwierigkeit an). Der Proband 1 mit der Fähigkeit Θ_1 löst das Item 1 mit einer Wahrscheinlichkeit von ca. 69%, das Item 2 mit ca. 13%. Der Proband 2 mit der Fähigkeit Θ_2 löst das Item 1 mit fast 100%, das Item 2 mit 90%iger Wahrscheinlichkeit.



Die Parameter des Modells werden aufgrund der Messwerte geschätzt. Dies geschieht mit Hilfe des Maximum-Likelihood-Kriteriums, d.h. die Werte der Parameter werden so geschätzt, dass *„die Wahrscheinlichkeit, dass sie den vorliegenden Datensatz nach dem gegebenen Modell ‚produziert' haben, maximieren"* (Perleth und Sierwald, 1991, S. 80). Nicht alle Itempools können Rasch-skaliert werden. Anforderungen an die Daten können aber mit Hilfe spezieller Verfahren überprüft werden. Während die Klassische Testtheorie davon ausgeht, dass die Summe der richtigen Lösungen eine erschöpfende Statistik darstellt, also das latente Merkmal notwendig und hinreichend beschreibt, lässt sich diese Annahme mit dem Rasch-Modell überprüfen. Lassen sich die Daten nicht mit einem bestimmten probabilistischen Modell (wie dem Rasch-Modell) darstellen, stehen weitere Modelle zur Verfügung, in denen sich auch Ratewahrscheinlichkeiten und Trennschärfen berücksich-

tigen lassen (vgl. Perleth und Sierwald, 1991, S. 83). Dies geht jedoch in einigen Modellen auf Kosten der Trennbarkeit von Fähigkeits- und Schwierigkeitsparametern.

Konstruiert man die Items des Rasch-Modell so, dass sich die Item-Charakteristik-Kurven nicht überschneiden (dies ist die Aufgabe der Itemanalyse), führt dies dazu, dass die Probanden nur mit geringer Wahrscheinlichkeit Aufgaben überspringen. Abbildung 5 zeigt, dass Proband 1 das Item 1 mit großer Wahrscheinlichkeit lösen kann (ca. 69%) , Item 2 jedoch wahrscheinlich nicht (ca. 13%). Alle schwierigeren Items, die noch weiter rechts im Schaubild angeordnet wären, wären von Proband 1 mit sehr hoher Wahrscheinlichkeit nicht mehr zu lösen gewesen. Dies ist natürlich nur durch eine hohe Trennschärfe der Items (dargestellt als Steigung der Kurve) zu erreichen. Dieses Verfahren führt zu einer eindeutigen Reihenfolge der Schwierigkeiten der Items, die es ermöglicht, mit Hilfe des sogenannten „Proficiency Scalings", Items nach speziellen Anforderungen zu gruppieren. Dies ist in der PISA-Studie durch die Etablierung der Kompetenzstufen geschehen.

4.3 Kriteriumsorientierte Testkonstruktion

Während die normorientierten Tests sich an einer Bezugs- oder Normgruppe orientieren, kann man die kriteriumsorientierten Tests auch als „lehrzielorientiert" bezeichnen (vgl. Klauer, 1987, S. 5). Sie nutzen als Einordnung eines Individuums nicht die Kennwerte einer Gruppe (Mittelwert, Standardabweichung) sondern bedienen sich eines sogenannten „Kriteriums". Als Ergebnis erhält man keinen Wert, mit dessen Hilfe man eine Person mit einer Gruppe vergleichen kann, sondern eine Aussage, ob diese Person den Test (bzw. den Lernleistungsstandard) „bestanden" oder „nicht bestanden" hat. Das Kriterium ist ein Punktwert (oder Prozentwert), der vom Testentwickler nach bestimmten Regeln festgelegt wird. Diesen Wert gilt es zu erreichen (oder, je nach Test, darunter zu bleiben).

Die Unterschiede in der Testkonstruktion und auch der Anwendungsbereiche sind erheblich. Dies lässt sich leicht an der Art der Reliabilitätsberechnung der Klassischen Testtheorie demonstrieren. Dort ist die Reliabilität definiert als der Anteil der Varianz der wahren Werte an der Varianz der Beobachtungswerte. Wird die Varianz der wahren Werte 0, ist die Reliabilität $r_{tt} = 0$. Ziel eines Unterrichts ist es jedoch,

alle Schüler so erfolgreich zu unterrichten, dass sie alle Aufgaben richtig lösen. *„In dem Maße, in dem eine Lehrperson erfolgreich ist, streben die Reliabilität und die Validität eines entsprechenden Leistungstests gegen Null“* (Klauer, 1987, S. 7). Die Häufigkeitsverteilung sollte demnach auch nicht normal sondern rechtsgipflig sein. Daher fordert Klauer eine neue Art von Tests, die „kriteriumsorientierten Tests“. Die Unterscheidung geht jedoch bereits auf Glaser (1963, zit. nach Büscher, 1984, S. 99) zurück.

Die Kennwerte der Klassischen Testtheorie werden fast alle unter Verwendung der Varianz berechnet. Reliabilität und Validität beruhen auf Korrelationskoeffizienten, die nicht definiert sind, wenn eine Variable keine Streuung aufweist. Für die Reliabilität werden eine Reihe anderer Gütekriterien vorgeschlagen, die sich mit der Konsistenz der Klassifikation beschäftigen. Wie bei der Klassischen Testtheorie lassen sich auch hier Retest-, Paralleltest und Testhalbierungsmethode nutzen. Ausgangspunkt der Berechnungen sind jedoch nicht die Messwerte der Testdarbietungen, sondern die Klassifikationen. Frickes Übereinstimmunskoeffizient Ü *„gibt das Verhältnis zwischen der empirischen (beobachteten) und der maximal möglichen Übereinstimmung [...] wieder“* (Büschel, 1984, S. 108). Die interne Optimalität setzt also den wahren Wert in Beziehung zum beobachteten Wert. Liegen bei allen Probanden wahrer Wert und beobachteter Wert unter der Zielmarke oder beide Werte oberhalb der Zielmarke, ist eine hohe interne Optimalität gewährleistet, jede Fehlklassifikation vermindern sie. Da auch hier der wahre Wert nicht bekannt ist, gibt es Schätzformeln, wie der Koeffizient δ von Linden & Mellenbergh (1978, zit. nach Klauer 1987, S. 88). Die externe Optimalität setzt den Beobachtungswert zu Beobachtungswerten anderer Tests nach dem selben Schema in Beziehung.

Als Kernstück der kriteriumsorientierten Tests dient die Konstruktion der Items. Hierbei muss sichergestellt werden, dass diese aus einem genau definierten, endlichen oder unendlichen Aufgabenpool stammen. Sie müssen aus diesem zufällig „gezogen“ werden. Nur bei einem genau festgelegt Aufgabenpool, lassen sich kontentvalide Items konstruieren. Hiermit ergibt sich eine wesentliche Anforderung an eine Itemanalyse. Die Aufgaben können nicht mit Hilfe der Itemanalyse der Klassischen Testkonstruktion ausgewählt werden. Dazu Büscher (1984, S. 104):

> *„Wenn eine Aufgabe einen relevanten Aspekt des Lehrziels erfasst, dann hat sie ihre Berechtigung in einem lehrzielorientierten Test, gleichgültig ob sie nun leicht oder schwer ist. Aufgaben, die von allen oder von den meisten Schülern richtig beantwortet werden, entsprechen der Intention des Lehrers.“*

An Stelle der Normierung steht die Definition einer „Zielmarke“, d.h. eines Kriteriums, mit dessen Hilfe ein Proband als Könner oder Nicht-Könner klassifiziert wird. Statt einer Dichotomisierung kann auch eine Einteilung in mehrere Kompetenzgrade geschehen (z.B. bei Schulnoten).

Kriteriumsorientierte Tests eignen sich besonders zur Evaluation des Unterrichts und zur Lehr- bzw. Lernerfolgsmessung. Für einen zu entwickelnden Lesetest eignen sich besonders die ausgearbeiteten Methoden der Ermittlung einer Kompetenz anhand einer Zielmarke.

4.4 Begründung für das Modell der Klassischen Testtheorie als Modell für den Test

Alle hier vorgestellten Modelle zeigen Vor- und Nachteile für eine Testkonstruktion zum Textverständnis.

Das Modell der kriteriumsorientierten Messung nach Klauer eignet sich eher zur Evaluation von Unterricht und Lernerfolgen. Mit Hilfe einer Zufallsauswahl Fragen aus einem genau umrissenen Itemuniversum, werden kontentvalide Tests konstruiert. Somit ist das Ergebnis des Tests kein Wert einer latenten Variable, sondern direkt erfahrbares Wissen. Für die Kompetenzstufe 1 des zu entwickelnden Lesetests ist die Konstruktion eines kriteriumsorientierten Tests noch vergleichbar unproblematisch, da das Universum möglicher Fragen nach einzelnen Informationen im Text begrenzt ist. Doch schon die Kompetenzstufe 2 und erst recht die dritte Stufe lassen eine Testkonstruktion nach kriteriumsorientierten Richtlinien nicht zu, da das Itemuniversum aus dem zufällig eine Stichprobe gezogen werden muss, nicht mehr exakt beschrieben werden kann. Die Fragen würden nicht kontentvalide werden. Nach der im Test zu Grunde gelegten Definition, umfasst Lesen nicht ausschließlich das Extrahieren von Informationen aus dem Text.

Insofern ist Lesekompetenz eine latente Variable, die sich nur anhand bestimmter empirischer Indikatoren messen lässt. Eine vollständig kriteriumsorientierte Testkonstruktion ist also ausgeschlossen.
Das Rasch-Modell erlaubt eine Trennung von Itemschwierigkeiten und Personenfähigkeit. Mit Hilfe einer angemessenen logistischen Transformation lässt sich für jedes Item eine Item-Characteristic-Kurve berechnen, auf der sich eine Lösungswahrscheinlichkeit für eine bestimmte Fähigkeit abbildet. Wie bereits gezeigt, lässt sich eine klare Reihenfolge der Items nach Schwierigkeit bilden, und mit Hilfe eines bestimmten Verfahrens lassen sich nun aus der kontinuierlichen Fähigkeitsskala Kompetenzstufen ableiten. Die Berechnung des Modells, also die gesamte Testerstellung nach dem Rasch-Modell ist jedoch sehr aufwendig. Sie benötigt nicht nur eine hohe Computerleistung, die sicherlich heutzutage vorhanden ist, sondern sie verursacht auch einen hohen Personaleinsatz. Dies ist im Rahmen dieser Testentwicklung jedoch nicht zu leisten. Die Vorteile des Modells bieten keine wesentliche Verbesserung des zu entwickelnden Tests, so ist z.B. das Multi-Matrix-Sampling dieser Testentwicklung nicht nötig, da jeder Proband alle Fragen dargeboten bekommt. Auch die Konstruktion der Stufen mit Hilfe des Proficiency Scaling ist durch die theoretische Ableitung der Items nicht nötig.

Aus theoretischen und pragmatischen Überlegungen ist die Klassische Testtheorie das Modell, das den Erfordernissen und Gegebenheiten am Ehesten entspricht. Es ist einfach und vergleichsweise schnell zu verwenden, außerdem ist es unter Testanwendern eingehend bekannt.

4.5 Die theoretische Konstruktion der Kompetenzstufen

Die Zuweisung einer Kompetenz, also das Bestehen einer Kompetenzstufe, geschieht mit Hilfe der Berechnung von Vertrauensintervallen um den wahren Wert des Schülers. Der Klassischen Testtheorie folgend ist der Beobachtungswert (Testrohwert) immer fehlerbehaftet. Das Vertrauensintervall berücksichtigt die Reliabilität und somit den Messfehler. Mittels der Reliabilität und der Streuung kann um einen Beobachtungswert ein Bereich angegeben werden, der es ermöglicht, Aussagen darüber zu treffen, mit welcher Wahrscheinlichkeit sich der wahre Wert in einem Bereich befindet. Grundlage der Berechnung

ist das 90%-Vertrauensintervall $CI_{90}\tau$. Die Größe dieses Intervalls ist je nach Kompetenzstufe verschieden, da die jeweiligen Reliabilitäten sowie die Streuungen unterschiedlich sind.

Neben der Größe des Vertrauensintervalls ist es nötig, eine „Zielmarke" für den wahren Wert anzugeben, d.h. wieviel Prozent der Aufgaben richtig gelöst werden müssen, um eine Testperson als Könner einzustufen. Dazu gibt es verschiedene Methoden, die in den nächsten Kapiteln kurz beschrieben werden.

4.5.1 Festsetzung mit Hilfe von Expertenratings

Expertenurteile können auf mehrere Arten eingesetzt werden: Die Experten können die Items oder die Probanden beurteilen. Einige der Verfahren, in denen Expertenratings verwendet werden um Items zu schätzen, gehen von einem „Grenzfallschüler" aus. Das ist ein Schüler, der eine 50 prozentige Chance hat, das geforderte Ziel zu erreichen.

Nedelsky (1954, zitiert nach Klauer, 1987, S. 63) schlägt vor, dass ein Item daraufhin geprüft werden soll, welchen Distraktor ein Grenzfallschüler sicher erkennen wird. Die Zahl der möglichen Alternativen reduziert sich dann um diese Zahl. Zwischen den verbleibenden Antworten rät der Schüler. Der erwartete Wert des Grenzfallschülers für diese Aufgabe ist dann 1/m, wobei m die Zahl der verbliebenen Antwortmöglichkeiten ist. Die Summe dieser Werte ist dann der kritische Punktwert. Angoff (1971, zitiert nach Klauer, 1987) ändert dieses Verfahren insoweit ab, dass die Experten für einen Grenzfallschüler direkt die Wahrscheinlichkeit der Lösung für ein Item abschätzen.

Livingston & Zieky (1982, zitiert nach Klauer, 1987) gehen davon aus, dass Lehrer die Fähigkeit ihrer Schüler besser einschätzen können als die Schwierigkeit von Aufgaben. Deshalb verwenden sie das Grenzfallschülerkonzept in abgewandelter Form. Die Experten schätzen die Schüler, die „auf der Kippe stehen".

Neben diesen Methoden gibt es noch eine Reihe anderer Verfahren. Allen ist aber gemein, dass sich die Experten, sollten mehrere am Rating beteiligt sein, einig sind. Befunde, wonach bessere Experten höhere Anforderungen stellen, zeigen mehrere Untersuchungen (z.B. Saunders, Ryan und Hyunh, 1981, zitiert nach Klauer, 1987, S. 68)

4.5.2 Festsetzung mit Hilfe von Außenkriterien

Zur Festlegung des Kompetenzgrades mittels Außenkriterien gibt es die Möglichkeit, den früheren oder späteren Lernerfolg heranzuziehen.

Nach der Methode von Berk (1976, zitiert nach Klauer, 1987, S. 69) bildet man zwei Gruppen. Nur eine Gruppe erhält den Unterricht, beide Gruppen führen am Ende der Unterrichtseinheit den Test durch. Mit Hilfe der Unterschiede zwischen beiden Testergebnissen und einer entscheidungstheoretischen Prozedur kann der Punktwert ermittelt werden.

Baut eine Lerneinheit auf eine andere auf, ist es möglich den Erfolg einer Lerneinheit auf die in der früheren Lerneinheit erlangte Kompetenz zu beziehen. Späteres Lernen als Außenkriterium bedient sich dieser Methode. Es lässt sich nachträglich der Kompetenzgrad feststellen, der nötig ist, um die nächste Lerneinheit zu bestehen.

4.5.3 Festsetzung mit Hilfe von Fehlertoleranzen

Der Grundgedanke der Methode mit Hilfe der Fehlertoleranzen ist, dass eine perfekte Beherrschung des Lehrziels angestrebt wird, d.h. die eigentliche Zielmarke ist $\pi_z = 100$. Klauer (1987, S. 72) fordert daher:

> *„Da dieser Grenzwert aber nie erreicht werden kann, sondern auch bei Könnern gelegentlich mit Fehlern gerechnet werden muss, wird eine Fehlertoleranz γ eingeführt, so dass der Kompetenzgrad $\pi_z = 1 - \gamma$ nunmehr als Norm gilt."*

Klauer (ebd.) schlägt in Anlehnung an Irrtumswahrscheinlichkeiten als sinnvolle Fehlertoleranzen $\gamma = 0{,}05$ und $\gamma = 0{,}10$ vor. Hohe Werte, so Klauer (ebd. S. 73), *„haben den Nachteil, dass man dann nur noch schwerlich von wirklichen Könnern sprechen kann"*.

4.6 Zusammenfassung

Ausgehend der Vorgaben, auf denen der Lesetest für Berufsschüler entwickelt werden soll, werden drei grundsätzlich unterschiedliche

Testmodelle beschrieben: das der Klassischen Testtheorie, der probabilistische Ansatz und das Modell der kriteriumsorientierten Messung. Klassische und probabilistische Testtheorie gehen von einem zu messenden latenten Merkmal aus, die kriteriumsorientierten Tests sehen das Ergebnis des Tests als direkt sichtbares. Testkonstruktion, Itemselektion und die Berechnung von Kennwerten, die die Güte des Tests beschreiben sind grundsätzlich anders. Demnach ist auch der Einsatzbereich der Testarten leicht anders. Während „klassisch" und probabilistisch konstruierte Tests Lernvoraussetzungen messen, lassen sich mit kriteriumsorientierten Tests eher Lernergebnisse messen. Für den zu entwickelnden „Lesetest für Berufsschüler/innen" wurde aus theoretischen und pragmatischen Gründen das Modell der Klassichen Testtheorie gewählt. Mit diesem Modell ist es mit Hilfe von Vertrauensintervallen um den wahren Wert möglich, messfehlerkorrigierte Entscheidungen darüber zu treffen, ob eine Person eine Kompetenzstufe bestanden hat. Zusätzlich ist es nötig, festzulegen, wann ein Kompetenzgrad erreicht ist. Dazu wurden drei Verfahren näher beschrieben.

5. Zusammenfassung und Fragestellung der Arbeit

Zunächst wurde beschrieben, dass Lesen ein komplexer Vorgang ist, der aus vielen Teilprozessen besteht. Das Lesen und Verstehen von ganzen Sätzen und längeren Texten sind dabei höhere Prozesse. In der PISA-Studie werden ausschließlich diese höheren Prozesse betrachtet. Der Lesetest der PISA-Studie ist sehr umfangreich und zeitaufwendig, zudem nicht zu diagnostischen Zwecken konstruiert. Außerdem ist der Test für 15-jährige Schüler bestimmt. Neben Fließtexten werden auch Tabellen und Diagramme dargeboten. Andere Tests, die die Lesekompetenz messen, sehen als Zielgruppe entweder Kinder oder Personen mit Migrationshintergrund, oder sie messen Leseschwäche als psychische Störung. Ein Lesetest für Berufsschüler der auf das Leseverständnis abhebt, existiert nicht und muss daher konstruiert werden, will man die Lesekompetenz der Berufsschüler im Hinblick auf eine adäquate Förderung erheben.
Der „Lesetest für Berufsschüler/innen" soll das Leseverständnis der Berufsschüler anhand von authentischen Fließtexten, Tabellen und Diagrammen auf drei Kompetenzstufen ermitteln. Diese Kompetenzstufen

sind das „Entnehmen von Informationen“, das „Verknüpfen mehrerer Informationen“ und das „Verknüpfen der Informationen mit dem Weltwissen“. Der Test soll nach der Klassischen Testtheorie konstruiert werden und den üblichen Standards der Gütekriterien standhalten. Hohe Objektivität, Reliabilität und Validität müssen ebenso gegeben sein wie eine hohe Ökonomie. Ausreichend hohe Gütekriterien lassen sich nur durch sorgfältige Konstruktion der Items, der Distraktoren und der Itemanalyse erreichen. Deshalb muss zunächst durch Pretests erprobt werden, ob die konstruierten Items geeignet sind, das zu messende Konstrukt in ausreichend sicherem Maße, also hoch reliabel, zu messen. Der fertige Test wird dann in verschiedener Hinsicht validiert, d.h. es muss geprüft werden, ob der Test tatsächlich die Variable „Lesekompetenz“ misst. Hinzu kommt eine Normierung des Tests, damit dieser als diagnostisches Instrument seine Wirkung entfalten kann.

Der folgende empirische Teil folgt dieser Einteilung, indem zunächst beschrieben wird, wie die Pretestversionen entwickelt und die Pretests durchgeführt wurden. Zu jedem Pretest wird daher auch die Itemanalyse und die zugehörigen Kennwerte beschrieben. In einem zweiten Schritt wird dann die Validität des Testes analysiert und die Normierung durchgeführt.

II. Empirischer Teil

Der empirische Teil dieser Arbeit ist in zwei Abschnitte unterteilt. Der erste Abschnitt beschreibt die Entwicklung des Tests, indem die Phasen der Entwicklung, die Durchführung der Pretests und die Itemanalysen besprochen werden.
Der zweite Teil beschäftigt sich mit der Validierung des Tests. Untersuchung zur Validität, die statistische Konstruktion der Kompetenzstufen, die Auswertung und die Normierung des Tests stehen hier im Focus.

1. Entwicklung des „Lesetests für Berufsschüler/innen“

Im Folgenden wird beschrieben, wie aus den Vorgaben des theoretischen Teils ein Lesetest für Berufsschüler/innen entwickelt wurde.

1.1 Phasen der Entwicklung des LTB^{-3}

Die Entwicklung des Lesetests für Berufsschüler/innen, im Folgenden auch LTB^{-3} genannt, lässt sich in drei Phasen beschreiben (s. Abbildung 6). In der theoretischen Phase (Phase I), die den November 2004 umfasste, wurden Zielgruppe und Rahmenkonzeption festgelegt sowie die Konzeption der Stufen erarbeitet. Als Zielgruppe wurden Jugendliche und junge Erwachsene in Berufsschulen festgelegt. Daraufhin wurde als testtheoretisches Rahmenkonzept ein Verfahren vorgeschlagen, das eine kriteriumsorientiere Auswertung mit Hilfe von Kompetenzstufen zulässt. Die Fragen sollten ausschließlich im geschlossenen oder halboffenen Format vorliegen, um eine hohe Auswertungsobjektivität zu erreichen. Auch die Auswertungsdauer sollte durch geschlossene Fragen wesentlich verkürzt werden. Die Anzahl der Antwortmöglichkeiten wurde auf drei festgelegt. Die Bearbeitung des Tests sollte dabei zwei Schulstunden nicht überschreiten. Neben kontinuierlichen Texten sollten auch nicht-kontinuierliche Texte, also Tabellen und Diagramme, aufgenommen werden, auf denen jeweils die drei Kompetenzstufen, „Informationen entnehmen“, „mehrere

Informationen verknüpfen“ und „eine Information mit dem Weltwissen verknüpfen“, ermittelt werden sollten.

Abbildung 6: Ablauf der Entwicklung des LTB-3

Phase	Zeitraum	Arbeit
I	November 2004	Definition der Zielgruppe
	November 2004	Rahmenkonzeption
	November 2004	Konzeption der Stufen
	↓	
II	März 2005	Sammlung von Texten
	März - Mai 2005	Erstellung der Items
	Mai / Juni 2005	Besprechung des Materials
	Juni 2005	Erster Pretest, Besprechung / Änderung des Materials
	September 2005	Zweiter Pretest, Besprechung / Änderung des Materials
	September - November 2005	Dritter Pretest, Besprechung / Änderung des Materials
	↓	
III	Januar 2006	Festlegung der Zielmarken für die kriteriumsorientierte Auswertung der Kompetenzstufen
	Februar 2006	Testdarbietung „Fragen ohne Texte“
	Februar - Juni 2006	Testnormierung
	April - Juli 2006	Erstellen der Materialien für die endgültige Testversion
	Juni 2006	Nachuntersuchung „Intelligenztests“
	Juli 2006	Abgabe an den Verlag, Erstellen einer verkaufsfertigen Version

Die zweite Phase (II) ist die empirische Phase. Zunächst wurden aus einer Vielzahl möglicher Materialien Texte, Diagramme und Tabellen ausgewählt. Diese Texte durften nicht zu lang und nicht zu kompliziert sein, um die Motivation der Schüler nicht zu gefährden. Die Textmenge wurde auf eine Seite beschränkt und die Texte daraufhin zum Teil gekürzt, inhaltlich jedoch weitgehend unverändert gelassen. Der Inhalt der Texte sollte dem Erfahrungshorizont der Berufsschüler entnommen sein. Zu den Texten, die in die nähere Auswahl kamen, wurden zunächst von einem Mitglied der Konstruktionsgruppe Items erstellt, die dann in einem Diskussionsprozess erweitert und verbessert wurden. Für den ersten Pretest kristallisierten sich vier Texte sowie ein Diagramm und zwei Tabellen heraus. Für die weiteren Pretests wurden zwei der Texte sowie das Diagramm und die Tabellen benutzt. Nach jedem Pretest wurden die Ergebnisse der Itemanalysen besprochen und die Fragen und Antwortalternativen verbessert. Zu Anfang der Phase II wurde die Idee verworfen, halboffene Fragen zu stellen. Das Format des Tests stand damit fest: Es wurden Fragen mit jeweils drei Antwortmöglichkeiten gebildet, aus denen jeweils nur eine Antwort richtig ist. Auf jeder Kompetenzstufe wurden so viele Fragen wie möglich gebildet, um später auswählen zu können, welche in die Endfassung des Tests gelangen. Als Minimum für eine Kompetenzstufe wurden pro Testabschnitt (Text, Diagramm, Tabelle) drei Fragen festgelegt.

Zusammenfassend wurden vier inhaltliche Anforderungen an die Items gestellt:

1. Misst die Frage die zu erfassende Kompetenz?
2. Lässt sich die Frage genau einer Kompetenzstufe zuordnen?
3. Sind die Kompetenzstufen durch die Fragen ausreichend abgebildet?
4. Lässt sich die Frage nur mit Hilfe des Textes beantworten?

Die dritte Phase (III) umfasste die Normierung des Tests für die normorientierte sowie die Festlegung der Zielmarken für die kriteriumsorientierte Auswertung. Des Weiteren wurden alle Materialen entwickelt, die für den Test zusätzlich nötig oder nützlich sind. Die Endfassung des Tests besteht somit aus fünf Materialien: dem Testhandbuch „Lesetest für Berufsschüler/innen - LTB^{-3}“, dem Testheft, dem Antwortbogen, der Antwortschablone und dem Ergebnisbogen (Drommler, Linnemann, Becker-Mrotzek, Haider, Stevens, Wahlers, 2006).

Das Testhandbuch (1) umfasst Angaben zum Testkonstrukt in Abgrenzung zum Lesekonstrukt der PISA-Studie, statistische Kennwerte und die Normentabellen. Darüberhinaus gibt es Durchführungs-, Auswertungs- und Interpretationsanweisungen. Das Testheft (2) beinhaltet die Texte, das Diagramm und die Tabellen sowie die Fragen. Das Testheft ist als Kopiervorlage konzipiert, so dass sich die Lehrer selbst einen Klassensatz zusammenstellen können. Auf dem Antwortbogen (3) werden von den Schülern die Antworten gegeben. Er ist maschinell auswertbar. Mit der Antwortschablone (4) lässt sich auch per Hand leicht ablesen, welche Fragen vom Schüler richtig beantwortet wurden. Die Werte können leicht im Kopf addiert und in den Ergebnisbogen (5) eingetragen werden. Dieser ist für die Lehrer bestimmt. Hierauf können alle Werte einer Schulklasse übersichtlich zusammengestellt werden. Auf diesem Bogen wird auch angekreuzt, ob ein Schüler die jeweilige Kompetenzstufe erreicht hat. Die Spalten stimmen farblich mit der Auswertungsschablone überein. Die Materialien 2 bis 5 werden mit einer Spiralbindung zusammengehalten und dienen als Kopiervorlage.

1.2 Pretests

Der fertigen Version des Lesetests für Berufsschüler/innen, gingen drei Prestestversionen voraus. Der erste Pretest diente dazu, ausgewählte Texte, Diagramme und Tabellen auf ihre Eignung zu prüfen. Die geringe Lese- und Bearbeitungszeit des ersten Pretests machten es möglich, im zweiten Pretest zwei Texte sowie ein Diagramm und zwei Tabellen mit den entsprechenden Fragen vorzugeben. Der zweite Pretest wurde an 74 Probanden erprobt. Der dritte Pretest, der fast der Endversion entsprach, wurde mit 576 Berufsschülern durchgeführt. Die Endversion unterscheidet sich lediglich in einer leichten Änderung zweier Antwortmöglichkeiten. Eine Frage wurde ersatzlos gestrichen, so dass der LTB^{-3} nun 46 Fragen beinhaltet. An weiteren 55 Probanden wurden die Fragen ohne die Texte, das Diagramm und die Tabellen dargeboten, um prüfen zu können, inwieweit die Fragen ohne Informationen aus dem Text lösbar sind (s. Kapitel 2.1.5). Im Folgenden werden Stichprobe, Material, Durchführung, Ergebnisse sowie die Itemanalysen für jeden Pretest beschrieben.

1.2.1 Erster Pretest

Der erste Pretest diente im Wesentlichen dazu, die ausgewählten Texte, das Diagramm und die Tabellen auf ihre prinzipielle Eignung zu überprüfen, sowie einen Schätzwert dafür zu bekommen, wie lang die Schüler zum Lesen der Texte und zum Beantworten der Fragen benötigen. Der Test sollte über eine doppelte Schulstunde nicht hinausgehen, um eine problemlose Implementierung in den Unterricht möglich zu machen. Des Weiteren sollte die Motivation der Schüler beobachtet werden.

Stichprobe

Die Stichprobe des ersten Pretests bestand aus 168 Berufsschülern verschiedener Schulklassen, die sich aus insgesamt 11 verschiedenen Ausbildungsgängen zusammensetzten (darunter Verpackungsmittelmechaniker, Bäckereifachverkäuferinnen, Technischer Zeichner, Metallbauer, etc., für eine detaillierte Übersicht siehe Anhang A-1). Die Daten des ersten Pretests wurden zwischen dem 14. und 17.06.2005 erhoben.

Material

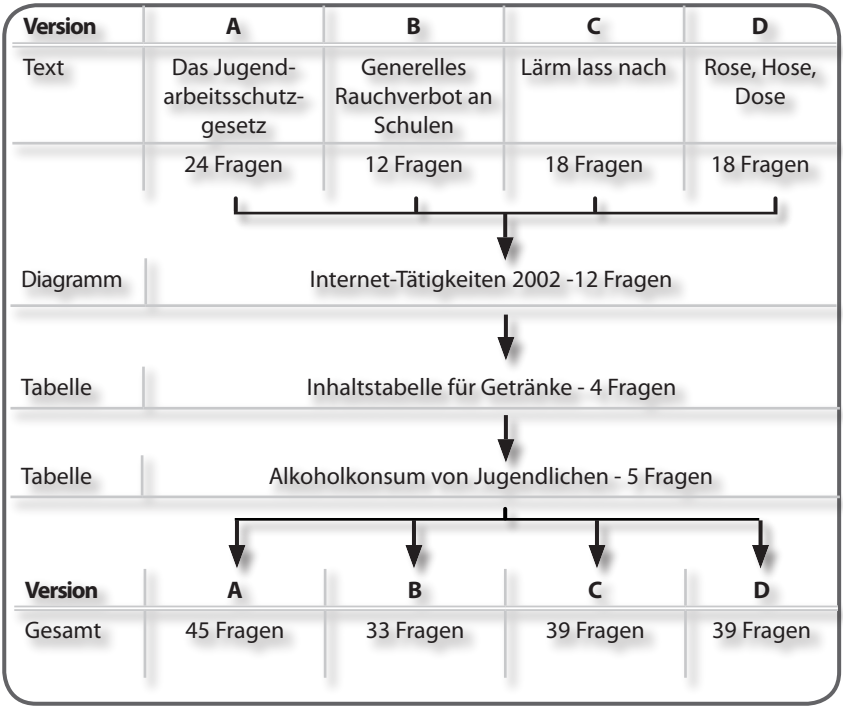
Das Material bestand insgesamt aus vier Texten, einem Diagramm und zwei Tabellen sowie den dazugehörenden Fragen. Jeweils ein Text sowie das Diagramm und die Tabellen bildeten die Testversion für einen Schüler. Die Anzahl der Fragen pro Testversion, pro Kompetenzstufe und der Ablauf des ersten Pretests zeigt Abbildung 7.

An den vier Testversionen haben unterschiedlich viele Probanden teilgenommen. Tabelle 2 zeigt die Verteilung.

Tabelle 2: Verteilung der Probanden auf die Testversion.

Testversion	Anzahl der Probanden	Prozent der Probanden	Anzahl der männlichen Probanden	Anzahl der weiblichen Probanden
A	58	34,5	50	8
B	36	21,4	29	6
C	38	22,6	36	2
D	36	21,4	17	19
Gesamt	**168**	**100**	**132**	**35**

Abbildung 7: Abfolge der Teile der vier Testversionen des ersten Pretests



Das Material bestand aus einem Testheft („Testheft 01“), das die Texte, das Diagramm und die Tabellen sowie die Fragen enthielt, und einem

Antwortbogen („Antwortbogen 01“). Bei den meisten Fragen konnte eine Antwort aus drei Möglichkeiten ausgewählt werden. Lediglich beim Text „Das Jugendarbeitsschutzgesetz“ gab es Fragen mit fünf bzw. zwei Antwortmöglichkeiten („Ja-Nein-Fragen“). Jede Frage wurde gleich gewichtet, d.h. jede richtige Antwort gab einen Punkt. Die einzelnen Fragen gehörten entweder der ersten, zweiten oder dritten Kompetenzstufe an. Zusätzlich zum Gesamtpunktwert wurde für jede dieser Stufen ein gesonderter Punktwert gebildet. Am Anfang des Test, zwischen den einzelnen Teilen sowie am Ende des Tests musste die Uhrzeit eingetragen werden, aus denen sich dann die Dauer des jeweiligen Testabschnitts berechnen ließ.

Da Zweifel bestanden, ob die Schüler mit dem Antwortbogen umgehen können, durften sie im Testheft ankreuzen und mussten dann die Antworten in den Antwortbogen übertragen, wofür ihnen 15 Minuten zur Verfügung standen. Ausgewertet wurden die Antworten aus dem Testheft. Hiermit sollte geprüft werden, ob die Schüler mit dem Antwortbogen umgehen können, um diesen in weiteren Pretests und der Endfassung ausschließlich einsetzen zu können.

Durchführung

Die Testdarbietung fand im Klassenverband im jeweiligen Klassenraum statt. Als Testleiter fungierte der jeweilige Lehrer. Die Schüler wurden angewiesen, zunächst die persönlichen Daten auf der ersten Seite des Testheftes zu notieren. Da der Test anonym durchgeführt wurde, konnte der Name frei gewählt werden.

Am Ende einer doppelten Schulstunde wurden die Testhefte und die Antwortbögen durch die Lehrer eingesammelt.

Hilfsmittel, wie z.B. Wörterbücher waren nicht erlaubt. Die Fragen mussten nicht aus dem Gedächtnis beantwortet werden. Die Schüler konnten jederzeit in den Text schauen.

Die Lehrer erstellten Durchführungsprotokolle, auf denen sie Auffälligkeiten während der Testdurchführung notierten.

Ergebnisse

Nur ein geringer Teil der Versuchspersonen kam mit dem Test nicht zurecht. Eine Person gab auf mehrere Fragen mehr als eine Antwort und füllte zudem die letzte Seite des Testheftes gar nicht aus. Zwei

weitere Probanden ließen viele Fragen unbeantwortet. Somit lässt sich sagen, dass der Test durch die Schüler positiv angenommen und nicht boykottiert wurde. Eine ausreichende Motivation schien also gegeben. Während des Testdurchgangs beklagte sich keiner der Schüler über den Test und auch die Durchsicht der Protokolle zeigte keine strukturellen Schwierigkeiten. Die Zeit reichte aus, um alle Fragen zu beantworten. Dies ist für einen reinen Niveautest von elementarer Wichtigkeit.

Nur eine Person nutzte die Möglichkeit in den Texten Unterstreichungen und Markierungen vorzunehmen. Das Übertragen der Antworten vom Testheft auf den Antwortbogen gelang fast ausnahmslos fehlerfrei.

Durch die klassenweise Vergabe der Testversionen kam es zu einem Ungleichgewicht bezüglich der Geschlechter. Die Version D wurde von etwa gleich vielen Männern wie Frauen bearbeitet, die anderen Versionen wurden wesentlich häufiger von Männern bearbeitet (s. Tabelle 2, Verteilung der Probanden auf die Testversionen)).

Das Durchschnittsalter aller Probanden betrug m = 20,7 Jahre (s = 3,1). Ca. 80% der Probanden gab Deutsch als Muttersprache an. Ein Ungleichgewicht bezüglich der Muttersprache konnte nicht festgestellt werden.

Die von den Schülern bereits absolvierten Schulabschlüsse wurden in drei Kategorien eingeteilt: niedriger, mittlerer und hoher Abschluss (s. Anhang A-2). Es ergab sich eine ungleiche Verteilung der Abschlüsse auf die Testversionen. Auch dies ist der klassenweisen Verteilung der verschiedenen Testversionen geschuldet. Tabelle 3 zeigt die Verteilung auf die Testversionen.

Tabelle 3: Verteilung der Schulabschlüsse auf die vier Testversionen in Prozent

	Schulabschluss		
Testversion	**niedrig**	**mittel**	**hoch**
A	36,2	39,7	24,1
B	47,1	47,1	5,9
C	8,3	72,2	19,4
D	22,9	45,7	31,4

Zusätzlich zu den bereits genannten Variablen wurden die letzten Schulnoten der Fächer Deutsch und Mathematik erhoben. Da diese Noten von verschiedenen Schulformen stammen (Hauptschule bis Gymnasium), lassen sie sich jedoch nicht vergleichen. Eine weitere Analyse der Noten wird hier und auch im Folgenden nicht vorgenommen.

Tabelle 4 zeigt die Dauer für die Bearbeitung der vier verschiedenen Texte sowie die Gesamtdauer des Tests. Das Bearbeiten der Fragen zum Diagramm dauerte m = 7,8 min (s = 6,1), das Bearbeiten der Tabellen m = 7,8 min (s = 5,3). Da der Test mit knapp 30 min Gesamtdauer im Mittel kürzer ausfiel als erwartet, konnte der zweite Pretest mit zwei Texten durchgeführt werden. Nicht alle Probanden trugen alle Uhrzeiten ein.

Tabelle 4: Dauer der Bearbeitung der verschiedenen Texte und die Gesamtdauer der jeweiligen Testversion (Text, Diagramm, Tabelle). Die „Dauer der Bearbeitung des Textes“ beinhaltet sowohl das Lesen des jeweiligen Textes als auch die Bearbeitung der zugehörenden Fragen.

		Mittelwert *(Standardabweichung)*	
Testversion	**n**	**der Dauer der Bearbeitung des Textes**	**der Gesamtdauer der Testversion**
A	35	16,3 *(5,9)*	34,9 *(17,7)*
B	34	10,4 *(3,0)*	26,9 *(9,6)*
C	29	12,0 *(3,9)*	24,0 *(5,0)*
D	35	14,0 *(4,9)*	28,4 *(7,2)*
Gesamt	**144**		**29,2 *(12,3)***

Itemanalyse

Für jede der vier Testversionen wurde eine eigene Itemanalyse durchgeführt, um zu entscheiden, welcher Text und welche Fragen zu diesem und dem Diagramm und den Tabellen in die Endversion aufgenommen werden können.

Zur Beurteilung der Güte eines Items kommen die in Teil I, Kapitel 4.1.3 genannten Faktoren in Betracht. Da die Itemanalyse nicht nur zu einer hohen inneren Konsistenz führen soll, sondern auch zu einer Vali-

ditätsverbesserung, wurden neben der Schwierigkeit und Trennschärfe auch die Korrelation zwischen Schulabschluss und Gesamtpunktwert (Validitätskoeffizient der externen Validität) sowie die Differenzierungsfähigkeit der einzelnen Items zwischen den Schulabschlüssen betrachtet. Die Stichprobe der Schüler ließ es zu, die Schwierigkeiten differenziert nach bereits absolvierten Schulabschlüssen zu betrachten. Das Testkonstrukt mit seinen drei aufeinander aufbauenden Kompetenzstufen legt nahe, dass die Fragen zur zweiten Stufe schwieriger sind als die der ersten, die der dritten schwieriger als die der ersten beiden Kompetenzstufen sind. Schließlich sollten die Alternativantworten - also die falschen Antworten - möglichst gleichverteilt sein.

Die durchgeführte Itemanalyse ergab die in Tabelle 5 (S. 59) dargestellten Schwierigkeits- und Trennschärfeindizes für die Testversion A und die in Tabelle 6 (S. 60) dargestellten Werte für die Testversion B.

Tabelle 5: Schwierigkeiten und Trennschärfen für die Items der ersten Pretestversion, Version A: Text „Jugendarbeitsschutzgesetz“ + Diagramm + Tabelle.

Item Nr.	Schwierigkeit unkorrigiert	Schwierigkeit korrigiert	Trennschärfe	Abschnitt
1	,44	,30	-	Das Jugendarbeitsschutzgesetz
2	,89	,87	,43	
3	,58	,47	,32	
4	,82	,78	,46	
5	,79	,58	,12	
6	,84	,68	,49	
7	,65	,30	,34	
8	,81	,61	,52	
9	,93	,86	,43	
10	,77	,54	,10	
11	,82	,65	,27	
12	,81	,61	,00	
13	,89	,79	-,15	
14	,84	,68	,47	
15	,72	,44	-,13	
16	,89	,79	,22	
17	,89	,84	,33	
18	,89	,84	,27	
19	,72	,58	,54	
20	,56	,34	,47	
21	,75	,63	,60	
22	,93	,89	,34	
23	,72	,58	,32	
24	,96	,95	,44	
25	,89	,84	,46	Diagramm
26	,96	,95	,56	
27	,60	,39	,10	
28	,91	,87	,22	
29	,79	,68	,50	
30	,19	-,21	,17	
31	,60	,39	,07	
32	,95	,92	,21	
33	,74	,61	,55	
34	,74	,61	,59	
35	,88	,82	,49	
36	,16	-,26	,23	
37	,86	,79	,20	Tabelle
38	,82	,74	,25	
39	,91	,87	,12	
40	,82	,74	,51	
41	,42	,13	,28	
42	,91	,87	,24	
43	,70	,55	,50	
44	,63	0,45	,34	
45	,70	0,55	,11	

Die Fragen zum Text „Jugendarbeitsschutzgesetz“ stellten sich als zu einfach heraus. In der Stufe 2 und 3 (Items 17 bis 24) betrug die geringste korrigierte Schwierigkeit 0,34, die höchste 0,95. Die Fragen zu Stufe 3 (Items 22 bis 24) wurden daraufhin gestrichen.

Tabelle 6: Schwierigkeiten und Trennschärfen für die Items der ersten Pretestversion, Version B: Text „Rauchverbot an Schulen“ + Diagramm + Tabelle.

Item Nr.	Schwierigkeit unkorrigiert	Schwierigkeit korrigiert	Trennschärfe	Abschnitt
1	,97	,96	,31	Rauchverbot an Schulen
2	,09	-,37	-,00	
3	1,00	1,00		
4	,97	,96	,08	
5	,34	,01	-,10	
6	1,00	1,00		
7	,74	,61	,30	
8	,71	,57	,24	
9	1,00	1,00		
10	,91	,87	,40	
11	,20	-,20	,32	
12	,39	,09	,12	
13	,77	,66	,19	Diagramm
14	,94	,91	,16	
15	,63	,44	,19	
16	,94	,91	,04	
17	,94	,91	,20	
18	,23	-,16	,23	
19	,69	,53	,17	
20	1,00	1,00		
21	,94	,91	-,08	
22	,51	,27	,25	
23	,89	,83	,07	
24	,09	-,37	,34	
25	,83	,74	,30	Tabellen
26	,83	,74	,04	
27	,94	,91	,08	
28	,83	,74	,12	
29	,46	,19	,40	
30	,86	,79	,11	
31	,60	,40	,16	
32	,63	,44	,22	
33	,57	,36	,06	

Im Text „Rauchverbot an Schulen“ stellten sich die Fragen zu Kompetenzstufe 1 als zu einfach heraus und wurden gestrichen, so dass der zweite Pretest aus den Kompetenzstufen 1 und 2 beim Text „Jugendarbeitsschutzgesetz“ und aus den Stufen 2 und 3 beim Text „Rauchverbot an Schulen“ bestand, zu dem noch drei neue Fragen konstruiert wurden. Tabelle 7 zeigt die Änderung der Items nach der Itemanalyse.

In die Berechnung der Reliabilität der Testversion A ging die Frage 01 nicht mit ein, da sie falsch gestellt wurde („In welchem Abschnitt wird etwas gesagt über am Arbeitsplatz?“). Die Items 24 und 26 konnten wegen der fehlenden Varianz nicht in die Analyse aufgenommen werden. In der Testversion B mussten die Items 03, 06 und 09 weggelassen werden, da auch sie eine Varianz von $s^2 = 0$ hatten. Tabelle 8 (S. 62) gibt die Reliabilitätskoeffizienten (Cronbachs α) und Validitätskoeffizienten (Spearman-Korrelation zwischen Gesamttestwert und Schulabschluss) für die einzelnen Testversionen vor der Itemselektion wieder. Eine zusätzliche Berechnung der Reliabilitäten nach der Itemselektion erschien vor dem Hintergrund der weitreichenden Änderungen nicht sinnvoll.

Tabelle 7: Änderungen der Items nach dem ersten Pretest

Test-version	Item Nr.	Grund der Änderung	Änderung
A	01 bis 04, 06, 08, 09, 18, 19, 22 bis 24	zu geringe Schwierigkeit	Frage geändert
	10, 11, 14, 17, 21	zu geringe Schwierigkeit, differenziert nicht zwischen Schulabschlüssen	Frage geändert
	12, 13, 15, 16	zu geringe Schwierigkeit, negative Trennschärfe	Frage geändert
B	01 bis 07	zu geringe oder zu hohe Schwierigkeit, zu geringe Trennschärfen	Fragen gestrichen
	08	Antwort b zu selten gewählt	Antwort b geändert
	09 bis 10	zu geringe Schwierigkeit	Frage gestrichen
	12	zu geringe Trennschärfe	Antworten geändert

Test-version	Item Nr.	Grund der Änderung	Änderung
Diagramm A und B	25, 27, 33	Antwort c zu selten gewählt	Antwort c geändert
	26, 32	zu geringe Schwierigkeit	Frage gestrichen
	28, 29, 35	zu geringe Schwierigkeit	Antworten geändert
	36	zu hohe Schwierigkeit	Frage geändert
Tabellen A und B	38, 43	Antwort b zu selten gewählt	Antwort b geändert
	39, 42	zu geringe Schwierigkeit	Antworten geändert
	45	Antwort c zu selten gewählt	Antwort c geändert

Anm.: Die Nummern beim Diagramm und bei den Tabellen beziehen sich auf die Nummerierung der Version A.

Trotz der vergleichsweise hohen Reliabilität der Version C (mit dem Text: „Lärm lass nach“) war die Validität gering. Aus diesem Grund wurde diese Versionen verworfen. Version D (mit dem Text „Rose, Hose, Dose“) schien bei mäßig hoher Reliabilität und Validität zu einfach. Der Mittelwert der Schwierigkeiten betrug m = 0,78, die Streuung s = 0,19. Er wurde deshalb nicht weiter verwendet.

Tabelle 8: Reliabilitäts- und Validitätskoeffizienten der Testversionen. Die Validität ist die Rangkorrelation des Gesamttestwertes mit dem Schulabschluss.

Testversion	n	Cronbachs α	Validität
A	48	0,713	0,672
B	35	0,586	0,547
C	38	0,888	0,382
D	36	0,681	0,597

1.2.2 Zweiter Pretest

Stichprobe

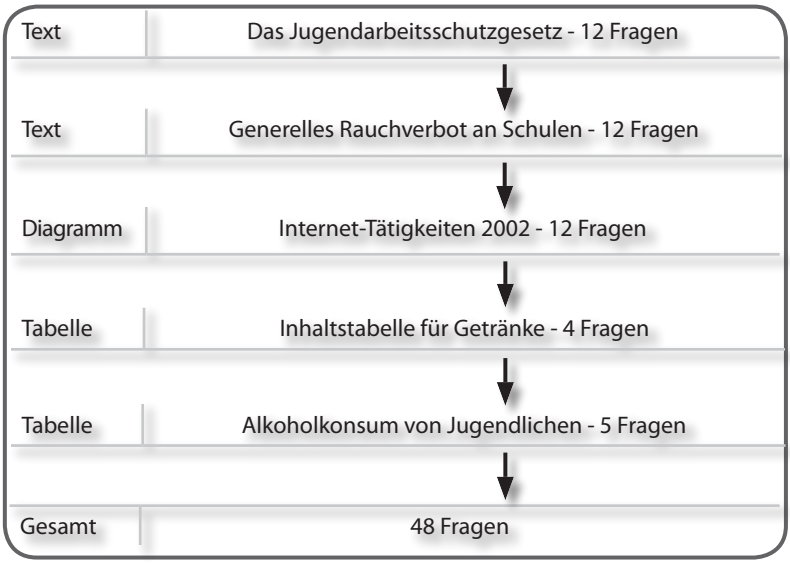
Die Stichprobe des am 05.09.2005 durchgeführten zweiten Pretests setzte sich zusammen aus Schülern der an der Testkonstruktion teil-

nehmenden Berufskollegs, die noch nicht am ersten Pretest teilgenommen hatten. Die 74 Schüler stammten aus vier Schulklassen. Das Durchschnittsalter der Probanden betrug m = 17,7 Jahre (s = 1,76), die überwiegende Zahl war männlich (86,3%), 4 von 5 Schülern sprachen Deutsch als Muttersprache. 1/3 der Probanden hatte einen niedrigen Schulabschluss, 2/3 einen mittleren oder hohen Abschluss. (Eine Tabelle mit den Ausbildungsberufen befindet sich in Anhang A-3.)

Material

Jedem Schüler wurde ein Testheft („Testheft 02“) und ein Antwortbogen („Antwortbogen 02“) vorgelegt. Das Testheft bestand aus zwei Texten, „Das Jugendarbeitsschutzgesetz“ und „Rauchverbot an Schulen“, dem Diagramm „Internet-Tätigkeiten 2002“ und den beiden Tabellen „Inhaltstabelle für Getränke“ und „Alkoholkonsum von Jugendlichen“. Die Fragen und Antwortalternativen wurden von der Testgruppe aus den Ergebnissen des ersten Pretests entwickelt. Abbildung 8 zeigt die Abfolge der Testteile.

Abbildung 8: Abfolge der Teile des zweiten Pretests.



Durchführung

Die Durchführung fand erneut im Klassenverband statt. Die Testanweisung entsprach der des ersten Pretests. Auf der ersten Seite des Testheftes mussten zunächst die gleichen persönlichen Daten wie im ersten Pretest angegeben werden. Der Test wurde anonym durchgeführt, d.h. jeder Schüler konnte einen Code-Namen benutzen. Auf dem Antwortbogen musste der Name wiederholt notiert werden. Die Antworten konnten im Testheft gegeben werden, mussten aber am Ende auf den Antwortbogen übertragen werden. Dazu standen wieder 15 Minuten zur Verfügung. Im Gegensatz zum ersten Pretest wurden nur die Antworten auf dem Antwortbogen ausgewertet.

Ergebnisse

Die durchschnittliche Bearbeitungsdauer lag zwischen 20 und 60 Minuten ($m = 37,5$; $s = 10,5$) und lag somit im Rahmen dessen, was durch den ersten Pretest erwartet und angestrebt wurde.

Itemanalyse

Die Itemanalyse ergab die in Tabelle 9 (S.65) dargestellten Schwierigkeits- und Trennschärfeindizes. Wie bereits im ersten Pretest sind die geringen Trennschärfeindizes auffällig, keine der Trennschärfen liegt über 0,58, die meisten im Bereich um 0,4.

Tabelle 9: Schwierigkeiten und Trennschärfen für die Items der zweiten Pretestversion. Grau unterlegt sind die Items, die aufgrund der nach diesem Pretest durchgeführten Itemanalyse geändert wurden.

Item Nr.	Schwierigkeit unkorrigiert	Schwierigkeit korrigiert	Trennschärfe	Text
1	,73	,66	,42	Das Jugendarbeitsschutzgesetz
2	,66	,58	,44	
3	,70	,63	,50	
4	,80	,75	,45	
5	,61	,22	,43	
6	,89	,78	,59	
7	,64	,28	,35	
8	,57	,14	,30	
9	,86	,72	,40	
10	,84	,68	,44	
11	,89	,78	,46	
12	,66	,32	,41	
13	,68	,36	,26	
14	,62	,43	,53	
15	,36	,04	,34	
16	,78	,67	,46	
17	,11	-,34	,01	
18	,23	-,16	,19	
19	,54	,31	,40	Rauchverbot
20	,32	-,02	,29	
21	,57	,36	,19	
22	,54	,31	,25	
23	,54	,31	,36	
24	,14	-,29	,14	
25	,46	,19	,33	an Schulen
26	,39	,09	,27	
27	,70	,55	,39	
28	,72	,58	,39	
29	,72	,58	,29	Diagramm
30	,39	,09	,14	
31	,81	,72	,48	
32	,84	,76	,58	
33	,11	-,34	,07	
34	,27	-,09	-,10	
35	,73	,60	,42	
36	,55	,31	,32	
37	,51	,27	,31	
38	,58	,37	,37	
39	,14	-,29	,23	
40	,72	,58	,47	Tabelle
41	,76	,64	,34	
42	,82	,73	,49	
43	,76	,64	,39	
44	,66	,49	,39	
45	,36	,04	,33	
46	,59	,38	,48	
47	,43	,15	,14	
48	,61	,42	,37	

Aufgrund der im Teil I, Kapitel 4.1.3 beschriebenen Kriterien zur Itemselektion, wurden insgesamt neun Items geändert. Drei Items, die aufgrund der Kriterien hätten geändert werden müssen, wurden erst nach dem dritten Pretest geändert. Tabelle 10 fasst die Änderungen zusammen.

Tabelle 10: Änderungen der Items nach dem zweiten Pretest

Item Nr.	Grund der Änderung	Änderung
17	geringe Schwierigkeit, geringe Trennschärfe, keine ausreichende Differenzierung zwischen den Schulabschlüssen	Frage geändert
22, 25	geringe Trennschärfe, Antwort b zu selten gewählt	Antwort b geändert
27, 32	keine ausreichende Differenzierung zwischen den Schulabschlüssen	Antwort a geändert
28, 36	keine ausreichende Differenzierung zwischen den Schulabschlüssen, Antwort a zu selten gewählt	Antwort a geändert
29, 48	keine ausreichende Differenzierung zwischen den Schulabschlüssen	Antwort c geändert
33	zu geringe Trennschärfe, zu hohe Schwierigkeit	keine
34	negative Trennschärfe	keine
44	Antwort b zu selten gewählt	keine

Die Reliabilität für die Gesamtzahl der Items des zweiten Pretests betrug Cronbachs $\alpha = 0{,}88$, die nach der Spearman-Brown-Formel korrigierte split half-Reliabilität ist $r_{tt} = 0{,}90$. Sie liegen im von Perleth und Sierwald (1991, s. Teil I, Kapitel 4.1.2) angegebenen Bereich für Leistungstests. Cronbachs α‘s für die einzelnen Kompetenzstufen waren jedoch geringer (Stufe 1: $\alpha = 0{,}82$; Stufe 2: $\alpha = 0{,}64$; Stufe 3: $\alpha = 0{,}53$). Die externe Validität, gemessen als Spearman-r zwischen dem Schulabschluss und dem Gesamttestwert ist $r = 0{,}48$ ($n = 73$). Dieser Wert ist ausreichend hoch, da der Schulabschluss kein erschöpfendes Kriterium für die Leseleistung ist. Die Reliabilitäten für die Gesamtpunkte sind ausreichend hoch, die Werte für die Kompetenzstufen, besonders für die zweite und dritte Stufe sind für einen diagnostischen Test als gering zu bewerten.

1.2.3 Dritter Pretest

Stichprobe

Die Stichprobe umfasste 576 Berufsschüler der teilnehmenden Berufskollegs. Die Probanden stammten aus 21 Ausbildungsgängen der drei teilnehmenden Schulen. Die Bezeichnungen der Ausbildungsgänge sind im Anhang A-4 einzusehen.

Material

Das Material entsprach dem des zweiten Pretests mit den durch die Itemanalyse geänderten Fragen und Antwortalternativen. Die Probanden bekamen nunmehr „Testheft 03“ und „Antwortbogen 03“. Die Items im Wortlaut sind den Materialien im Anhang B zu entnehmen.

Durchführung

Die Durchführung der Tests fand wie bei den ersten beiden Pretests in den Klassenräumen als Gruppentest statt.

Ergebnisse

Das Alter der Probanden lag zwischen 16 und 39 Jahren (m = 19,7 Jahre; s = 2,76; n = 567). Ca. 61% der Probanden war männlich, 39% weiblich. Deutsch als Muttersprache gaben 83% an, eine andere Muttersprache 15,5%, 1,5% machten keine Angabe. Anders als in den ersten beiden Pretests, konnte der Bildungsstand anhand des bereits absolvierten Schulabschlusses in vier Kategorien eingeteilt werden (in Klammern die benutzte Abkürzung): kein oder niedriger Abschluss (HS), mittlere Reife (MR), Fachhochschulabschluss (FH) und Abitur oder Hochschulabschluss (AB). Tabelle 11 zeigt die Anteile der Schulabschlüsse an der Stichprobe. Etwa 2/3 der Stichprobe waren dem unteren und mittleren Bildungsniveau zuzuordnen (für eine genau Einteilung der Schülabschlüsse s. Anhang 5.

Die benötigte Zeit betrug im Mittel m = 35,6 Minuten (s = 10,3). Die Unterschiede zwischen den Schulabschlüssen waren signifikant ($F(3, 483) = 3,57$; $p < 0,05$; partielles $\eta^2 = 0,022$), jedoch gering. Da es sich

um einen reinen Niveau-Test handelt, muss dieser Befund nicht näher betrachtet werden.

Tabelle 11: Anteil der Schulabschlüsse an der Stichprobe der Berufsschüler und die für den Test benötigte Zeit.

Schulabschluss	n	Prozent	benötigte Zeit m *(SD)*
kein oder niedriger Schulabschluss	88	15,7	36,6 *(9,99)*
mittlere Reife	314	56,2	36,6 *(11,31)*
Fachhochschulabschluss	102	18,2	33,7 *(8,65)*
Abitur oder Hochschulabschluss	55	9,8	32,6 *(6,98)*
fehlend	*17*		
Gesamt	**576**	**100**	**35,6** *(10,30)*

Itemanalyse

Tabelle 12 (S. 69) zeigt die Schwierigkeiten und Trennschärfen der Items.

Tabelle 12: Schwierigkeiten und Trennschärfen für die Items der dritten Pretestversion. Grau unterlegt sind die Items, die aufgrund der nach diesem Pretest durchgeführten Itemanalyse geändert wurden.

Item Nr.	Schwierigkeit unkorrigiert	Schwierigkeit korrigiert	Trennschärfe	Text
01	,88	,85	,43	Das Jugendarbeitsschutzgesetz
02	,78	,73	,35	
03	,74	,68	,37	
04	,65	,56	,38	
05	,76	,52	,45	
06	,91	,82	,32	
07	,72	,44	,35	
08	,40	-,20	,24	
09	,89	,78	,29	
10	,85	,70	,32	
11	,92	,84	,39	
12	,78	,56	,31	
13	,74	,48	,34	
14	,69	,54	,34	
15	,60	,40	,42	
16	,89	,84	,17	
17	,63	,45	,33	
18	,63	,45	,28	
19	,78	,67	,38	Rauchverbot
20	,44	,16	,25	
21	,59	,39	,10	
22	,58	,37	,16	
23	,68	,52	,31	
24	,24	-,14	,18	
25	,65	,48	,36	an Schulen
26	,56	,34	,32	
27	,80	,70	,42	
28	,80	,70	,30	
29	,81	,72	,42	Diagramm
30	,39	,09	,10	
31	,85	,78	,23	
32	,86	,79	,27	
33	,15	-,28	,00	
34	,30	-,05	,10	
35	,84	,76	,28	
36	,69	,54	,32	
37	,75	,63	,33	
38	,17	-,25	,24	
39	,84	,76	,32	Tabelle
40	,78	,67	,29	
41	,70	,55	,14	
42	,79	,69	,34	
43	,69	,54	,14	
44	,47	,21	,28	
45	,73	,60	,29	
46	,61	,42	,33	
47	,28	-,08	,15	

Insgesamt wurden vier Fragen bzw. Antwortmöglichkeiten geändert. Tabelle 13 zeigt eine Übersicht der Änderungen.

Tabelle 13: Änderungen der Items nach dem dritten Pretest

Item Nr.	Grund der Änderung	Änderung
21	Antwort a zu häufig genannt	Antwort a geändert
33	Antwortmuster bei den Schulabschlüssen zu willkürlich, falsche Antwort a steigt mit zunehmendem Schulabschluss, Trennschärfe zu gering, Schwierigkeit zu hoch	Frage und Antwortmöglichkeiten a,b und c geändert
34	Antwort b zu selten gewählt, Trennschärfe sehr gering, Antwortmuster differenziert bei den Schulabschlüssen zu gering	Frage gestrichen
43	Antwort c zu selten genannt	Frage und Antwortmöglichkeiten a,b und c geändert

Tabelle 14 zeigt die Reliabilitäten sowie die Validität für die Gesamtpunktzahl und die drei Kompetenzstufen des dritten Pretests vor der Itemselektion.

Tabelle 14: Reliabilitäts- und Validitätskoeffizienten der Punktzahl der Kompetenzstufen sowie der Gesamtpunktzahl. Die Validität ist die Rangkorrelation des Gesamtpunktzahl mit dem Schulabschluss.

Kompetenzstufe	Cronbachs α (n=576)	split half (n=576)	Validität (n=559)
Gesamttest	0,831	0,836	0,497
1	0,766	0,796	0,432
2	0,574	0,580	0,388
3	0,531	0,472	0,422

2. Validierung des „Lesetests für Berufsschüler/innen“

Dieses Kapitel beschäftigt sich mit der Untersuchung zur Validität und den Analysen der Kompetenzstufen. Es geht ferner auf die Unterschiede der normorientierten und kriteriumsorientierten Auswertung und auf die Normierung des Tests ein. Den Berechnungen dieses Kapitels wurde die Stichprobe aus dem dritten Pretest zu Grunde gelegt.

2.1 Untersuchungen zur Validität

Die folgenden Kapitel beschäftigen sich mit der Validität des Tests. Zur Validität gibt es kein einheitliches Konzept und so lassen sich nur Indikatoren sammeln, inwieweit der Test das Konstrukt „Lesefähigkeit“ trifft. Möglichkeiten dazu bestehen darin, die faktorielle Struktur und die Homogenität zu überprüfen. Des Weiteren ist es möglich, zu bestimmen, inwieweit der Test mit anderen Größen zusammenhängt, die einen Hinweis auf Lesefähigkeit geben. Dies geschieht im Kapitel „Korrelationsanalysen“. Außerdem sollte der Test Personengruppen unterscheiden, von denen angenommen werden kann, dass sie über verschieden hohe Lesefähigkeiten verfügen. Dies wird im Kapitel „Unterschiedshypothesen“ geprüft. Im Kapitel „Pretest ohne Darbietung der Texte“ wird schließlich geprüft, ob der Test neben Lesefähigkeit auch andere Eigenschaften misst.

2.1.1 Faktorenanalyse

Eine Faktorenanlyse dient u.a. *„zur Klassifikation von Fragebogen- und Testaufgaben nach ihrer Dimensionalität mittels der Analyse der Aufgabeninterkorrelationen“* (Lienert, 1989, S. 490 f.). Diese Klassifikation kann wiederum dazu dienen, einen faktoriell homogenen oder heterogenen Test zu konstruieren. Neben dem Einsatz der Faktorenanalyse während der Testkonstruktion kann sie aber auch nachträglich eingesetzt werden, um eine innere Struktur des Tests zu ermitteln. Im Folgenden werden mehrere Faktorenanalysen durchgeführt: zunächst eine Analyse mit allen 46 Items, um zu sehen, welche Dimensionen sich abbilden lassen. Im Anschluss werden drei Einzelanalysen durchgeführt, um zu sehen, welche Dimensionen sich auf den jeweiligen Textarten abbilden lassen.

Sollten die Kompetenzen streng aufeinander aufbauen, dürften sich die Kompetenzen nicht als unabhängige Faktoren herausbilden, denn die Kompetenzen sind voneinander abhängig: Ohne die Fragen der ersten Stufe (mehrheitlich) richtig zu beantworten, dürfte es den Probanden nicht möglich sein, die Fragen der dritten Kompetenzstufe (mehrheitlich) richtig zu beantworten. Andersherum gilt das jedoch nicht.

Faktorenanalyse mit allen Items

Zur Überprüfung der faktoriellen Validität wurde eine Hauptkomponentenanalyse mit Varimax-Rotation über alle Items durchgeführt. Die Zahl der Komponenten wurde auf drei festgelegt. Auch der Scree-Plot zeigte eine Dreifaktorenlösung als geeignet an. Insgesamt wurden 21,33% der Gesamtvarianz aufgeklärt. Die Korrelationen zwischen den Items sind durchweg gering (s. auch Teil II, Kapitel 2.1.2).

Tabelle 15 (S. 73) zeigt die rotierte Faktormatrix. Betrachtet man die Tabelle, so fällt auf, dass die Komponente 1 die ersten 13 Fragen des Textes „Das Jugendschutzgesetz“ abbildet. Dies sind die Fragen zur Kompetenzstufe 1. Die Komponente 2 umfasst Teile des Textes „Jugendschutzgesetzes“ (die der Kompetenzstufe 2) sowie nahezu den kompletten Text „Rauchverbot an Schulen“, der die Stufen 2 und 3 enthält. Die dritte Komponente umfasst die Fragen zum Diagramm und zu den Tabellen. Die große Übereinstimmung bei den ersten beiden Kompetenzstufen könnte am unterschiedlichen Fragemodus liegen: Die Items 01 bis 13 sind wenig schwierige Multiple choice-Fragen mit fünf, bzw. zwei Antwortmöglichkeiten, während alle anderen Items drei Antwortmöglichkeiten zulassen.

Tabelle 15: Faktorenladungen für die drei Komponenten. Die fettgedruckten Kästchen geben die pro Item höchste Faktorladung an. Ladungen unter 0,10 sind nicht aufgeführt. Die schwarz umrandeten Kästchen geben die Kompetenzstufe an. Überlagern sich Unterlegung und Umrandung, stimmt die Kompetenzstufe mit dem Faktor überein.

Item Nr.	Faktorladung (Umrandung = Kompetenzstufe) 1	2	3	Text
01	,62	-,14	,35	Das Jugendarbeitsschutzgesetz
02	,48		,33	
03	,47		,36	
04	,34		,43	
05	,37	,35	,20	
06	,72			
07	,45	,20	,11	
08	,18	,30		
09	,33	,12	,17	
10	,63	,14		
11	,74	,13		
12	,55	,11		
13	,46	,25		
14	,14	,29	,25	
15		,42	,34	
16		,25		
17	,12	,38	,21	
18	,21	,34		
19	,23	,36	,20	Rauchverbot
20		,26	,19	
21		,29		
22			,21	
23		,37	,22	
24		,30		

Item Nr.	Faktorladung (Umrandung = Kompetenzstufe) 1	2	3	Text
25	,14	,41	,19	an Schulen
26	,10	,34	,22	
27	,19	,24	,42	
28	,19	,28	,17	
29		,28	,47	Diagramm
30			,10	
31		,11	,33	
32			,46	
33		,10	-,11	
34		,34		
35		,19	,32	
36		,24	,36	
37	,13	,14	,37	
38		,35		
39		,17	,47	Tabelle
40			,53	
41	,12	-,16	,32	
42	,15	,26	,28	
43			,32	
44	,12	,22	,23	
45		,11	,41	
46		,20	,36	
47		,46		

Item Nr.	Faktorladung (Umrandung = Kompetenzstufe) 1	2	3	Text

	Item Nr.	Faktorladung (Umrandung = Kompetenzstufe) 1	2	3	Text
Eigenwert		3,77	3,41	2,85	
erklärte Varianz		8,03%	7,25%	6,05%	

Insgesamt bleibt festzuhalten, dass sich die Kompetenzstufen faktorenanalytisch nicht abbilden ließen. Dies zeigt, dass die Kompetenzstufen keine orthogonalen Skalen sind. Da diese im Testkonstrukt nicht so angelegt sind, ist die Faktorenanalyse ein Indiz für die Abhängigkeit der Skalen. Die Korrelationen zwischen den Kompetenzstufen liegen zwischen $r_{13} = 0{,}52$ und $r_{23} = 0{,}57$ (s. Anhang A-7).

Einzelne Faktorenanalyse für die Texte, das Diagramm und die Tabellen

Um zu überprüfen, ob sich die Kompetenzstufen für die einzelnen Textteile faktorenanalytisch zeigen, wurden getrennte Faktorenanalysen für jede dieser Textarten durchgeführt. Die Varianzaufklärung für die Texte war mit insgesamt 29% gering. Wenn man die rotierten Faktorenladungen für die Textarten einzeln betrachtet (s. Tabelle 16, S. 76 f.), wird deutlich, dass, wie auch bei der Analyse des gesamten Itempools, eher der Fragemodus eine Rolle spielt und die Kompetenzstufen sich dort nicht abbilden lassen. Die Items 01 bis 13 sind leichte Multiple choice-Fragen mit fünf, bzw. zwei Antwortmöglichkeiten, alle anderen Items lassen drei Antwortmöglichkeiten zu.

Ein ähnlicher Befund zeigte sich beim Diagramm und den Tabellen. Die Varianzaufklärung war hier mit 41% bzw. 45,8% zwar besser, jedoch immer noch nicht zufriedenstellend. Außerdem lässt sich festhalten, dass die empirisch gefundenen Komponenten mit den Kompetenzstufen beim Diagramm und den Tabellen (s. Tabelle 16, S. 76) noch weniger übereinstimmen als bei den Texten.

Weitere Analysen zeigten, dass sich auf den einzelnen Kompetenzstufen der Faktor „Texte“ von einem Faktor „Diagramm und Tabellen“ unterscheiden lässt. Dies wird besonders deutlich, wenn man die Ergebnisse einer Zwei-Faktorenlösung für Kompetenzstufe 1 betrachtet: Von 13 Fragen zum Text luden 12 hoch auf Faktor 1, von 7 Fragen zum Diagramm und zu den Tabellen luden alle sehr hoch auf Faktor 2. Die Varianzaufklärung betrug ca. 30%. Bei den anderen Kompetenzstufen war das Bild nicht so eindeutig.

Zusammenfassung

Die durchgeführten Faktorenanalysen lassen keinen eindeutigen Schluss über die Homogenität des Testes zu. Die Analysen sowohl mit allen Items als auch die getrennten Analysen für Texte, Diagramm und Tabellen weisen darauf hin, dass die Faktoren entweder die Textart oder den Fragemodus abbilden jedoch nicht die Kompetenzstufen.

Tabelle 16: **Faktorenladungen der drei Komponenten für** die Texte (ob**en)** und die Diagramme und Tabellen (unten). **Die fettgedruckten Kästchen geben die jeweils höchste Faktorladung an. Ladungen unter 0,10 sind nicht aufgeführt. Die schwarz umrandeten Kästchen geben die Kompetenzstufe an. Überlagern sich Unterlegung und Umrandung, stimmt die Kompetenzstufe mit dem Faktor überein.**

Texte:

Item Nr.	Faktorladung (Umrandung = Kompetenzstufe) 1	2	3	Text
01	,26	,04	**,78**	Das Jugendarbeitsschutzgesetz
02	,16	,07	**,67**	
03	,15	,12	**,66**	
04	,09	,17	**,59**	
05	**,38**	**,39**	,13	
06	**,70**	-,04	,25	
07	**,48**	,25	,07	
08	,17	**,24**	,07	
09	,22	,19	**,30**	
10	**,61**	,03	,21	
11	**,75**	,07	,19	
12	**,62**	,05	,09	
13	**,58**	,21	-,03	
14	,17	**,41**	,04	
15	,07	**,59**	,08	
16	-,04	**,34**	,08	
17	,02	**,53**	,12	
18	,09	**,41**	,14	
19	,17	**,49**	,11	
20	,08	**,32**	,02	Rauchverbot an Schulen
21	-,02	**,27**	-,13	
22	-,10	**,25**	,19	
23	,02	**,40**	,11	
24	,05	**,21**	-,01	
25	,17	**,47**	,00	
26	,07	**,47**	,08	
27	,14	**,42**	,23	
28	,07	**,35**	,23	
Eigenwert	2,87	2,91	2,348	
erklärte Varianz	10,26%	10,38%	8,39%	

Diagramm:

Item Nr.	Faktorladung (Umrandung = Kompetenzstufe) 1	2	3	Text
29	,61	,14	-,03	Internet-Tätigkeit
30	,18	,48	-,34	
31	,50	-,15	,44	
32	,52	-,19	,10	
33	-,07	,20	,83	
34	-,11	,66	,07	
35	,58	,03	,02	
36	,42	,15	-,01	
37	,57	,12	-,17	
38	,20	,62	,17	
Eigenwert	**1,81**	**1,21**	**1,07**	
erklärte Varianz	**18,1%**	**12,1%**	**10,7%**	

Tabellen:

Item Nr.	Faktorladung (Umrandung = Kompetenzstufe) 1	2	3	Text
39	,50	,37	-,03	Alkoholkonsum
40	,70	,13	,10	
41	,67	-,10	-,01	
42	,08	,53	,11	
43	-,06	,32	,81	
44	,01	,62	-,29	
45	,41	,38	-,01	
46	,15	,63	,10	
47	-,14	,38	-,56	
Eigenwert	1,4	1,6	1,1	
erklärte Varianz	15,6%	18,1%	12,1%	

2.1.2 Homogenität

Die Korrelationen zwischen den Items sind durchgängig sehr gering. Der Mittelwert der Korrelationen beträgt für alle Items $m_r = 0{,}098$ ($s = 0{,}07$). Die minimale Korrelation beträgt dabei $r_{min} = -0{,}093$, die höchste $r_{max} = 0{,}523$. Die Verteilung ist normal. Die Mittelwerte der Korrelationen für die einzelnen Kompetenzstufen sind ähnlich gering.

Ein sehr homogener Test müsste hohe Korrelationen zwischen den Items aufweisen. Bereits im vorangegangenen Kapitel über die Faktorenanalyse wurde deutlich, dass die Interkorrelationen gering sind, denn sonst wären die Ergebnisse der Faktorenanalyse eindeutiger

gewesen. Eine Faktorenanalyse analysiert Cluster von Korrelationen, die es hier nicht zu geben scheint. Nach Lienert bedeutet Homogenität *„(faktorielle) Einheitlichkeit der zu messenden psychischen Funktion“* (1989, S. 251). Des Weiteren geht er von folgenden Annahmen aus:

> *„In einem vollkommen homogenen Test müsste jede Aufgabe genau dieselbe Funktion, dasselbe Persönlichkeitsmerkmal prüfen wie jede andere, wie auch der ganze Test. Wenn nun die einzelnen Aufgaben streng nach ansteigender Schwierigkeit geordnet sind, so wird es - außer durch Zufall - nicht vorkommen dürfen, dass Aufgaben übersprungen [...] werden.“ (Lienert, 1989, S. 252)*

Von diesen Annahmen ausgehend hat Loevinger (vgl. Lienert, 1989, S. 252) einen Homogenitätsindex H_t entwickelt, der die beobachtete Varianz an der Varianz eines vollständig homogenen Tests mit gleicher Schwierigkeitsverteilung der Aufgaben relativiert. Die Formel, die Loevinger angibt, sieht folgendermaßen aus:

$$H_t = \frac{V_x - V_{het}}{V_{hom} - V_{het}}$$

V_x = tatsächlich beobachtete Testvarianz
V_{het} = Varianz eines vollkommen heterogenen Tests
V_{hom} = Varianz eines vollkommen homogenen Tests

„Der Quotient selbst kann im Sinne eines Korrelationskoeffizienten interpretiert werden“ (ebd., S. 252). Berechnet man mit Hilfe der Formel (Loevinger gibt dazu eine Schätzformel an, vgl. Lienert, 1989, S. 252) Loevingers H_t für den gesamten Test sowie die drei Kompetenzstufen getrennt, so ergibt sich das in Tabelle 17 (S. 82) dargestellt Bild von sehr geringen Homogenitätswerten. Hat dieser Homogenitätswert auch für die Testkonstruktion *„keine unmittelbare Bedeutung“* (ebd., S. 254), so ist die fehlende Einheitlichkeit des Tests doch bemerkenswert. Das Konstrukt „Lesefähigkeit“ scheint somit sehr homogen zu sein.

Tabelle 17: Loevingers H_t für den Gesamttest und die Kompetenzstufen

Stufe	Loevingers H_t
Gesamtpunktzahl	0,07
Kompetenzstufe 1	0,09
Kompetenzstufe 2	0,07
Kompetenzstufe 3	0,06

2.1.3 Korrelationsanalysen

Zur Untersuchung der externen Validität standen mehrere Außenkriterien zur Verfügung: der bereits erreichte Schulabschluss, der Intelligenzquotient, die erhaltenen Schulnoten, das Geschlecht und die Muttersprache. Im Folgenden werden die Korrelationen zu den drei erstgenannten Variablen beschrieben und diskutiert. Die Analysen der übrigen Variablen befinden sich weiter unten im Kapitel „Unterschiedshypothesen".

Korrelation mit bereits erreichtem Schulabschluss

Die Schüler mussten angeben, welchen Schulabschluss sie bereits erreicht haben. Da die Klassen des Berufskollegs dem dualen Ausbildungssystem angehören, werden dort Schüler aller Schulformen unterrichtet. Tabelle 11 (S. 53) zeigt die Verteilung der Probanden auf einzelnen Schulabschlüsse. Die Rangkorrelation nach Spearman beträgt dabei $r = 0{,}50$ ($n = 559$; $p < 0{,}001$). Eine zu hohe Korrelation würde den Test obsolet machen, ein zu niedriger Wert wäre ein Indikator für eine geringe externe Validität. Insofern ist dieser Wert als befriedigend zu bewerten.

Intelligenzquotient

Zur Messung des Intelligenzquotienten wurde der Mehrfachwahl-Wortschatz-Intelligenztest nach Lehrl (1999) (MWT-B) benutzt. Der Vorteil liegt vor allem in einer schnellen und einfachen Durchführung. Der MWT-B wurde am 28.06.2006 an 69 Schülern durchgeführt, die am dritten Pretest des LTB[-3] bereits teilgenommen hatten.

Der MWT-B misst die kristalline Intelligenz, er will *„die Rückstände messen, die Spuren der geistigen Auseinandersetzung eines Menschen mit seiner*

Umwelt, die sich als Fertigkeiten und Wissen niedergeschlagen haben“ (Lehrl, 1999, S. 16). Seine Reliabilität ist mit r_{tt} = 0,95 (Retestreliabilität nach 30 min) ausreichend hoch. Die Korrelationen mit anderen Intelligenztests (HAWIE bzw. WAIS, ZVT, Raven SPM) liegen zwischen r = 0,64 und r = 0,78 (ebd. S. 22). Seine Aufgaben bestehen ausschließlich aus sprachlichem Material, insofern gibt ein hohes Ergebnis auch einen Hinweis auf hohe sprachliche Kompetenz, zumal die Korrelationen des MWT-B mit den Verbalteilen der jeweiligen verglichenen Intelligenztests durchweg höher sind als mit den Handlungsteilen. Da sprachliche Kompetenz aus mehr Komponenten als der Lesekompetenz besteht und der MWT-B keine kompletten Texte sondern nur Wörter darbietet, erreicht die Korrelation mit dem LTB^{-3} mit r = 0,335 (n = 69; Spearman-rho) nur ein geringes Niveau. Der LTB^{-3} erfasst die Intelligenz somit nur in geringem Maße.

Schulnoten

Die Probleme bei der Erfassung der Schulnoten wurden bereits erwähnt. Die Frage an die Schüler nach den letzten Schulnoten war zu unpräzise gestellt. So ist nicht klar, ob sich die angegebene Schulnote auf die letzte Note im Berufskolleg oder auf die der abgebenden Schule bezieht. Die Schulnoten müssten dann an den jeweiligen Schulabschlüssen relativiert werden, denn die Note „sehr gut“ an einer Hauptschule im Fach Deutsch entspricht nicht der Note „sehr gut“ beim Abitur an einem Gymnasium. Die Note „gut“ beim Schulabschluss Abitur kann aber sehr wohl einem „gut“ beim Fachhochschulabschluss oder der Mittleren Reife entsprechen, da diese Abschlüsse auf dem Gymnasium erworben werden (können). Forschungsergebnisse sprechen sogar davon, dass nicht einmal die Noten der gleichen Schulform vergleichbar sind. Die IGLU-Studie zeigt, dass die Zensurengebung stark von der Klassenzusammensetzung abhängt (vgl. IGLU-Studie, Bos et al., 2003). Aus diesen Gründen wird auf eine Analyse dieser Noten verzichtet.

Auf einen zweiten Aspekt sei hier noch hingewiesen. Es war nicht möglich, die aktuellen Schulnoten der Schüler von den Lehrern zu erhalten, um diese mit dem Testergebnis zu korrelieren.

2.1.4 Unterschiedshypothesen

Die Unterschiedshypothesen zielen auf eine konkurrente Validität ab. *„Die konkurrente Validität sollte unter anderem dazu führen, dass verschie-*

dene Personengruppen unterschieden werden können“ (Klauer, 1987, S. 46). Die Variablen, die erhoben wurden, waren Schulabschluss, Geschlecht und Muttersprache. Die Muttersprache wurde dichotomisiert in „Muttersprache Deutsch“ („L_1“) und „andere Muttersprache“ („L_2“). Für die Unterschiedshypothesen wurde zunächst eine Messwiederholungs-Varianzanalyse mit den Prozentwerten der richtigen Antworten der drei Kompetenzstufen als Messwiederholungsfaktoren und dem Geschlecht (zweistufig), dem Schulabschluss (vierstufig) und der Muttersprache (zweistufig) als unabhängige Variablen durchgeführt. Das Modelldesign der ANOVA wurde dabei auf die Haupteffekte und die Interaktionen zweiten Grades beschränkt. Das Signifikanzniveau wurde auf 5% festgelegt, signifikante Unterschiede werden mit $p < 0{,}05$ kenntlich gemacht, hoch signifikante mit $p < 0{,}01$ und höchst signifikante Unterschiede mit $p < 0{,}001$. Bei den Post hoc-Tests wurde das α-Niveau, wenn nötig, korrigiert.

Unterschiede zwischen den Schulabschlüssen in den Kompetenzstufen

Schon bei der Itemselektion wurde darauf geachtet, dass die Items zwischen den bereits vorhandenen Schulabschlüssen differenzieren. Dieser Unterschied müsste sich auch in den Ergebnissen der Varianzanalyse widerspiegeln. In der Tat zeigte die durchgeführte ANOVA einen höchst signifikanten Haupteffekt „Schulabschluss“ ($F(3, 544) = 55{,}07$; $p < 0{,}001$; partielles $\eta^2 = 0{,}23$) über alle Kompetenzstufen hinweg. Bis auf den Unterschied „Fachhochschulabschluss-Abitur“ waren alle Unterschiede höchst signifikant. Tabelle 18 zeigt die Mittelwerte der richtigen Antworten. Die Interaktion Kompetenzstufe × Schulabschluss war nicht signifikant.

Tabelle 18: Mittelwert und Standardabweichung der richtigen Antworten in Prozent für die Schulabschlüsse

Schulabschluss	**Mittelwert** ***(Standardabweichung)***
kein oder niedriger Abschluss	54,9 *(13,92)*
mittlere Reife	65,3 *(13,43)*
Fachhochschulreife	74,8 *(10,27)*
Abitur oder Hochschulabschluss	78,8 *(9,93)*
Insgesamt	**66,7** *(14,46)*

Das Augenmerk auf die Unterschiede während der Testkonstruktion führte also durchaus zu signifikanten Unterschieden.

Geschlechterunterschiede in den Kompetenzstufen

Laut der PISA-Studie (vgl. Stranat & Kunter, 2001) sind *„die größten und konsistentesten Geschlechterunterschiede [...] im Bereich Lesen zu beobachten“*, wobei die Mädchen in allen PISA-Teilnehmerstaaten *„im Lesen signifikant höhere Testwerte als die Jungen“* haben. Dieser Unterschied beträgt etwa 1/3 Standardabweichungen.

Auch im LTB-3 war dieser Unterschied signifikant (Haupteffekt Geschlecht: $F(1, 544) = 6{,}14$; $p < 0{,}05$; partielles $\eta^2 = 0{,}01$). Der Effekt war jedoch mit $d = 0{,}14$ nicht nur klein (vgl. Cohen, 1977), sondern auch dem in PISA gefundenen Unterschied entgegengesetzt. Die Unterschiede bezogen sich auch nicht auf alle Kompetenzstufen gleich, wie die signifikante Interaktion zwischen den beiden Variablen zeigte ($F(2, 1088) = 6{,}33$; $p < 0{,}01$; partielles $\eta^2 = 0{,}01$). Es konnte lediglich auf der Kompetenzstufe 3 ein signifikanter Kontrast gefunden werden ($p < 0{,}016$; $d = 0{,}27$). Auch dieser Befund ist dem aus der PISA-Studie entgegengesetzt. Dort schneiden die Mädchen auf den höheren Skalen besser ab als die Jungen, wogegen es bei den unteren Skalen („Informationen heraussuchen“) keine signifikanten Unterschiede gibt (ebd., S. 260). Der Lesetest für Berufsschüler/innen differenziert also nicht (oder im Falle der Kompetenzstufe 3 nicht gut) zwischen den Geschlechtern.

Unterschiede zwischen Muttersprachlern und Nicht-Muttersprachlern

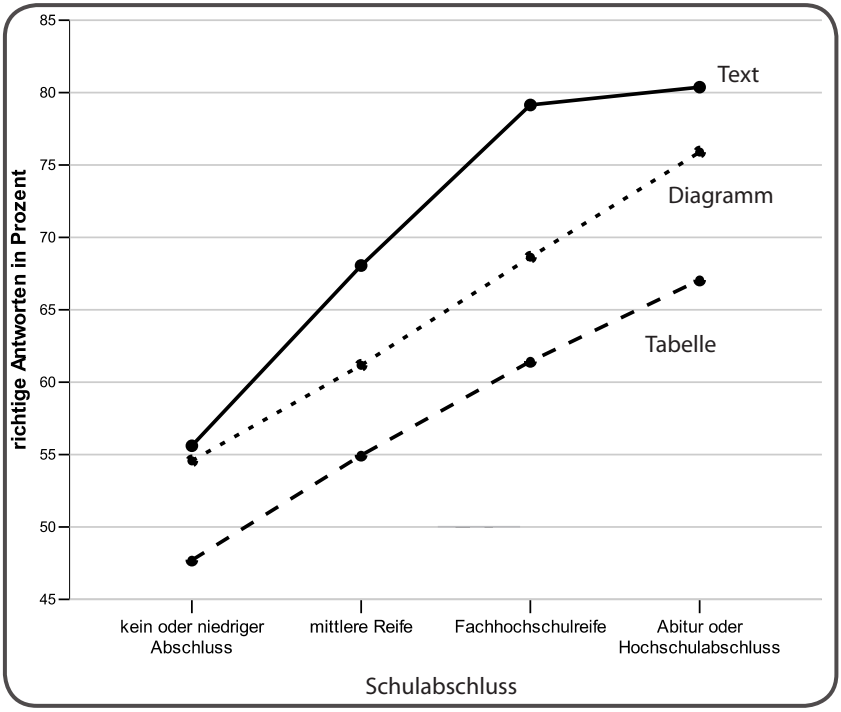
In der PISA-Studie gefundenen Unterschiede legen nahe, dass es auch im LTB-3 signifikante Unterschiede bezüglich der Muttersprache gibt, legt man die Muttersprache als Kriterium für einen Migrationshintergrund an. Die Ergebnisse der Varianzanalyse unterstützten diese Hypothese. Schüler mit Deutsch als Muttersprache hatten höhere Werte über alle Kompetenzstufen hinweg als Schüler, die eine andere Muttersprache angegeben hatten. Dieser Effekt war signifikant ($F(1, 544) = 5{,}87$; $p < 0{,}05$; partielles $\eta^2 = 0{,}01$), und mit $d = 0{,}48$ klein bis mittelgroß. Die Interaktion Kompetenzstufe × Muttersprache war nicht signifikant.

Die Befunde der Faktorenanalyse legten nahe, den Test in die Teile Texte, Diagramm und Tabellen zu unterteilen. Die mit den Kompetenzstufen als Messwiederholung berechnete ANOVA wurde mit der Messwiederholung Texte, Diagramm und Tabellen („Textart“) erneut berechnet.

Unterschiede zwischen den Schulabschlüssen in den verschiedenen Textarten

Wie bereits bei den Kompetenzstufen ist der Haupteffekt Schulabschluss höchst signifikant ($F(3, 544) = 45{,}55$; $p < 0{,}001$; partielles $\eta^2 = 0{,}20$). Dies ist nicht verwunderlich, da der Text so konstruiert wurde. Interessant ist jedoch die signifikante Interaktion zwischen dem Messwiederholungsfaktor „Textart“ und dem Schulabschluss ($F(3, 544) = 2{,}86$; $p < 0{,}05$; partielles $\eta^2 = 0{,}02$). Post hoc-Tests zeigten keinen signifikanten Unterschied zwischen den Textarten bei den Hauptschülern, signifikante Unterschiede zwischen allen Textarten bei den Schülern mit mittlerer Reife und Fachhochschulreife und einen signifikanten Unterschied zwischen Text-Diagramm und Diagramm-Tabelle bei den Abiturienten. Abbildung 9 (S. 84) zeigt die Interaktion.

Abbildung 9: Interaktion zwischen dem Schulabschluss und der Textart

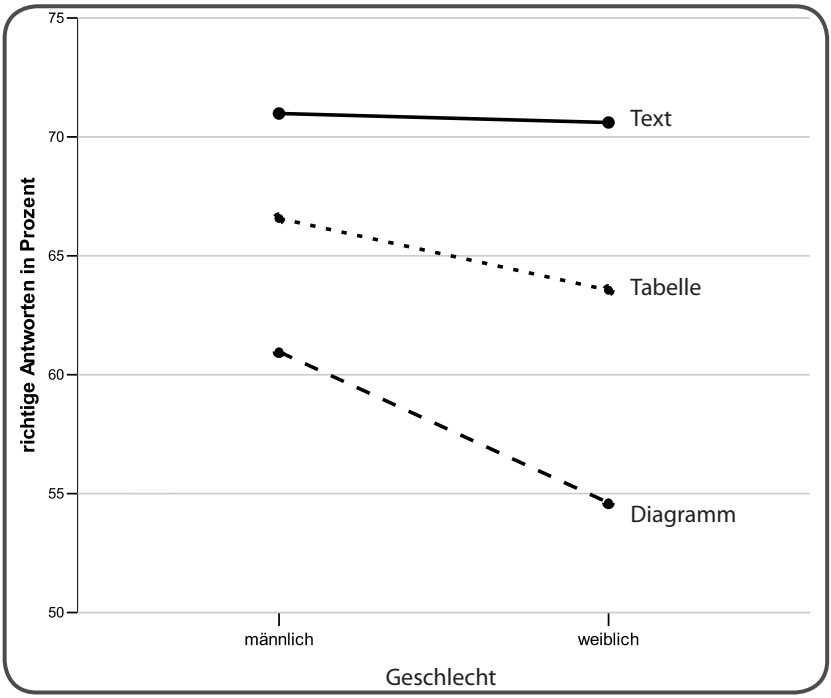


Geschlechterunterschiede in den verschiedenen Textarten

Die PISA-Studie legt nahe, dass es innerhalb der verschiedenen Textarten Unterschiede zwischen den Geschlechtern gibt (vgl. PISA 2000, S. 260). Signifikante Unterschiede ergaben sich dort bei den kontinuierlichen Texten (Mädchen sind besser als Jungen), wogegen bei den nicht-kontinuierlichen Texten (Diagramme und Tabellen) keine signifikanten Unterschiede gezeigt werden konnten. Die Unterschiede bei den kontinuierlichen Texten betrug ca. 0,2 Standardabweichungen und ist etwa für alle Bildungsgänge gleich. Die für den LTB^{-3} durchgeführte ANOVA ergab eine hoch signifikante Interaktion Textart × Geschlecht mit ($F(1,9; 1032,3) = 6,53$; $p < 0,01$; partielles $\eta^2 = 0,01$). Post hoc durchgeführte t-Tests mit α-Korrektur ergaben ausschließlich einen Unterschied bezüglich der Diagramme ($t(566) = 4,55$; $p < 0,016$). Diese werden von den weiblichen Probanden schlechter bearbeitet, der Effekt

ist nach Cohen mit d = 0,40 klein bis mittelgroß. Abbildung 10 veranschaulicht die Interaktion grafisch. Auch der Haupteffekt Geschlecht ist signifikant ($F(1, 544) = 8{,}68$; $p < 0{,}01$; partielles $\eta^2 = 0{,}02$), jedoch hier nicht weiter von Interesse.

Abbildung 10: Interaktion zwischen dem Geschlecht und der Textart



Unterschiede zwischen Muttersprachlern und Nicht-Muttersprachlern in verschiedenen Textarten

Bezüglich der Muttersprache zeigten sich keine signifikanten Unterschiede zwischen den Textarten ($F(1{,}9; 1032{,}2) = 1{,}33$; $p > 0{,}05$).

2.1.5 Pretest ohne Darbietung der Texte

Um zu sehen, ob bestimmte Fragen alleine durch Wissen oder Raten richtig zu beantworten sind, wurde eine Testversion erstellt, bei der

die Probanden die Antworten ohne die Texte, das Diagramm oder die Tabellen dargeboten bekamen. Die Probanden konnten die Antworten also wissen, ohne die Texte zu kennen, oder sie mussten sie raten. Bei reinem Raten ist die korrigierte Schwierigkeit $\Pi = 0$. Ein Wert, der größer ist, bedeutet, dass die Frage auch ohne die Kenntnis der Texte richtig zu beantworten ist. Es kann also sein, dass die richtige Antwort gewusst wurde oder die richtige Antwort gegenüber den Distraktoren zu logisch erscheint. Dies wäre natürlich nicht im Sinne des Tests. Ein negativer Wert zeigt, dass eine Antwort bei Unkenntnis der Texte eher falsch gegeben wird, eine falsche Antwortmöglichkeit lenkt hier also von der richtigen Möglichkeit ab. Dies, sowie der Fall, dass sich die korrigierte Schwierigkeit nicht signifikant von $\Pi = 0$ unterscheidet, ist für die Testkonstruktion gewünscht. Der Test wäre also gut, wenn sich zeigen ließe, dass die Punktzahl ohne Lesen der Texte nicht über die Ratewahrscheinlichkeit hinaus geht oder sogar darunter liegt. Dies gilt entsprechend für die einzelnen Items.

Stichprobe

Die Stichprobe des am 09.02.2006 durchgeführten Tests bestand aus 42 Schülern eines Berufskollegs. Das Alter der Probanden lag zwischen 16 und 21 Jahren ($m = 18{,}4$; $s = 1{,}53$). Die Geschlechter waren nahezu gleich verteilt. 24 Probanden gaben an, Deutsch als Muttersprache gelernt zu haben, 17 gaben eine andere Sprache an. 60 % der Schüler hatten Abitur, 33 % mittlere Reife und 2 Schüler gaben einen Fachhochschulabschluss an.

Material

Die Fragen entsprachen den Fragen aus dem dritten Pretest. Die Texte, das Diagramm und die Tabellen wurden weggelassen. Zusätzlich wurde der Intelligenztest MWT-B durchgeführt.

Durchführung

Den Probanden wurde mitgeteilt, dass sie einen Lesetest ohne Texte bearbeiten sollen, um herauszufinden, welche Fragen man ohne sie beantworten kann. Im Gegensatz zu den Pretests zuvor wurden die Probanden dazu angehalten, die Antwort anzukreuzen, die ihnen am logischsten erscheint. Sie sollten die Antwort raten, wenn sie sie nicht

wissen. Es wurde ihnen gesagt, dass sie bei einigen Fragen die Lösung nicht wissen können.

Ergebnisse

Es zeigte sich, dass einige Fragen auch ohne Lesen der Texte, des Diagramms und der Tabellen richtig zu beantworten waren. Tabelle 19 zeigt, wieviele Fragen auf den jeweiligen Kompetenzstufen durch Wissen richtig zu beantworten sind.

Tabelle 19: Anzahl der Fragen, die durch Wissen richtig zu beantworten sind. Grau unterlegt sind die für die Testkonstruktion unerwünschten Fragen.

Kompetenzstufe	schlechter als geraten $\Pi < 0$	geraten $\Pi = 0$	besser als geraten $\Pi > 0$	Gesamt
1	6	7	8	21
2	5	7	4	16
3	4	3	3	10
Gesamt	15	17	15	47

Die Unterschiede zwischen den Stufen sind nicht signifikant (Exakter Test nach Fisher; $p > 0{,}898$), d.h. es ist für die Testkonstrukteure für jede Stufe gleich schwer gewesen, die Fragen so zu konstruieren, dass sie ausschließlich mit Informationen aus dem Text zu beantworten sind. Gerade Tests, die den Erfahrungshorizont der Jugendlichen betreffen, sind gefährdet, mit Fragen zu operieren, deren Antworten die Schüler kennen. Hierzu seien zwei Beispiele aus dem Test genannt.

Beispiel 1.	**Prozent**	
Frage 11: **An Tagen mit viel Arbeit ist es Jugendlichen ausnahmsweise erlaubt, ohne Pause durchzuarbeiten.**	**ohne Text**	**mit Text**
richtige Antwort	93	96,6
falsche Antwort	7	3,4

Beispiel 1 zeigt, dass die Unterschiede zwischen den Probanden, die die Texte bekommen haben und denen, die sie nicht bekommen haben, bei dieser Frage nicht wesentlich sind. Die korrigierte Schwierigkeit der Frage 11 weicht in der Gruppe „ohne Text“ signifikant positiv von $\Pi = 0$ ab ($t(41) = 21{,}31$; $p < 0{,}001$), d.h. die Probanden raten nicht.

Beispiel 2.	**Prozent**	
Frage 21: **Wozu führt regelmäßiger Zigarettenkonsum?**	**ohne Text**	**mit Text**
a. Zum Konsum illegaler Drogen	2,4	31,5
b. Zu Organschäden	80,5	61,8
c. Zu sofortigen Gesundheitsstörungen	17,1	6,7

Beispiel 2 zeigt, dass die meisten Probanden der „ohne Text“-Version Antwort b („zu Organschäden“) genannt haben. Dies ist die richtige Antwort, die die Probanden der „mit Text“-Version nur zu 61,8 % genannt haben. Das Lesen des Textes hat fast 1/3 dazu gebracht, Antwort a anzukreuzen. Hier zeigt sich, dass die Darbietung eines in diesem Punkt wohl verwirrenden Textes zu einer falschen Antwort führen kann und dies über alle Schulabschlüsse hinweg, denn vom niegrigen bis hohen Abschluss erhält diese Antwort fast durchgängig ca. 30 %.

Die Daten zeigen noch einen weiteren interessanten Aspekt. Bei der Analyse der gewußten Antworten gibt es bei zwei Fragen Unterschiede zwischen den Schülern unterschiedlicher Schulabschlüsse. Bei Item 32 entscheiden sich die Schüler mit mittlerem Schulabschluss auf die Frage „Wie viele Jugendliche wurden für die Studie befragt“ für

Antwort b („51%“), während sich die Schüler mit Abitur mehrheitlich und überzufällig für Antwort a („904“) entscheiden (s. Tabelle 20). Für sie ist die Antwort „51%“ anscheinend keine adäquate Antwort auf die Frage. Auch bei Item 45 gibt es signifikante Unterschiede im Antwortverhalten von Schülern mittlerer und hoher Schulabschlüsse, jedoch ist es hier fraglich, dass die Abiturienten die Zahlen kennen.

Abschließend kann man sagen, dass es Antworten gibt, die „gewußt“ werden können. Unterschiede zwischen den verschiedenen Schulabschlüssen lassen sich post hoc nicht erklären. Die Reliabilität des „Pretests ohne Texte“ ist mit Cronbachs $\alpha = 0{,}16$ erwartungsgemäß sehr klein.

Tabelle 20: Verteilung der Schüler in Prozent auf die Antworten der Items 32 und 45 unterteilt nach Schulabschluss (nur Probanden der „ohne Text“-Version; n = 42)

	Schulabschluss	
Frage 32: Wie viele Jugendliche wurden für die Studie befragt?	**mittel**	**hoch**
Antwort		
904	21,4	66,7
51 %	57,1	7,4
Geht aus dem Diagramm nicht hervor	21,4	25,9
Frage 45: Wie viele der 16- bis 17-Jährigen trinken insgesamt ein - bis mehrmals pro Monat Bier?		
Antwort		
23 %	7,1	0,0
25 %	57,1	23,1
26 %	35,7	76,9

2.2 Die Kompetenzstufen

Bei der kriteriumsorientierten Auswertung wird der Schüler mit einem feststehenden Punktwert, dem Kriterium, verglichen. Liegt er über dem Kriteriumswert, hat er die Kompetenzstufe erreicht (und gilt als

Könner). Liegt er darunter, hat er die Stufe nicht erreicht (und muss als Nicht-Könner gelten). Zunächst wird beschreiben, wie diese Kompetenzstufen gebildet wurden bevor sie dann einer kritischen Analyse unterzogen werden.

2.2.1 Die empirische Konstruktion der Kompetenzstufen

Die Klassifikation der Schüler in Könner und Nicht-Könner wurde mit Hilfe des Vertrauensintervalls um den wahren Wert vorgenommen. Die Größe dieses Intervalls ist je nach Kompetenzstufe verschieden, da die jeweiligen Reliabilitäten sowie die Streuungen unterschiedlich sind. Tabelle 21 zeigt die Reliabilitäten, die Streuung und die Größe des Vertrauensintervalls.

Tabelle 21: Reliabilitäten, Standardabweichungen und Konfidenzintervalle für die drei Kompetenzstufen der Endfassung des Tests

Kompetenzstufe	split half	SD	$CI_{90}\tau$
1	0,80	3,62	± 2,63
2	0,60	2,57	± 2,69
3	0,45	1,88	± 2,28

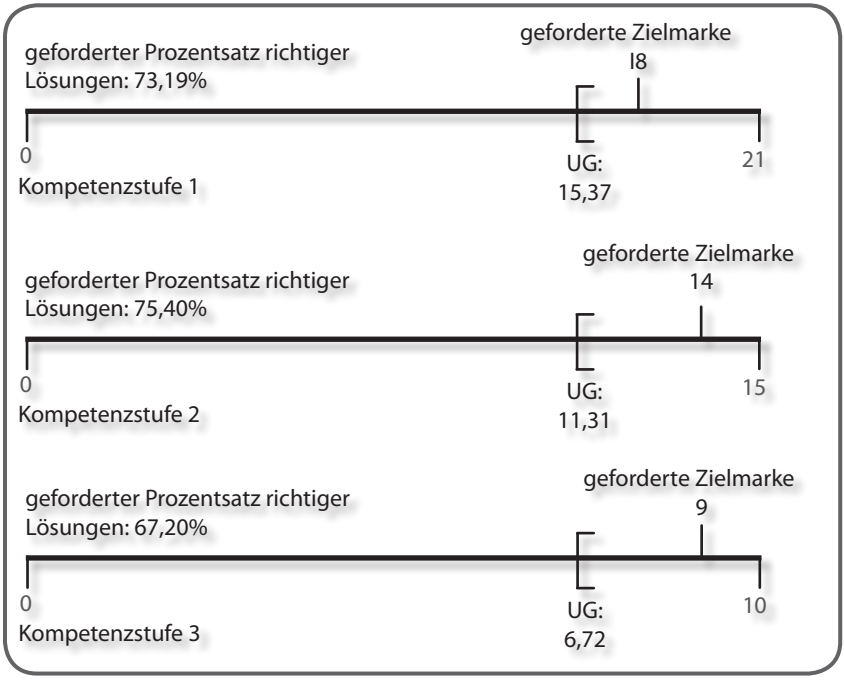
Anm.: $CI_{90}\tau$ = 90%-Vertrauensintervall für den wahren Wert

Der Prozentwert des Intervalls ist willkürlich gesetzt. Da es neben der Größe des Vertrauensintervalls nötig ist, eine „Zielmarke“ für den wahren Wert anzugeben, d.h. wieviel Prozent der Aufgaben richtig gelöst werden müssen, um eine Testperson als Könner einzustufen, wurden diese Werte durch die Autoren mit Hilfe des Modells der Fehlertoleranzen von Klauer (1987) festgelegt. Für die Kompetenzstufen 1 und 2 wurde 75% richtig gelöste Aufgaben gefordert, für die Kompetenzstufe 3 70%. Diese Werte schienen, auch im Hinblick auf die Größe der Vertrauensintervalle, als vertretbar. Abbildung 11 (S. 91) zeigt den Zusammenhang zwischen geforderter Zielmarke, Untergrenze des Vertrauensintervalls und dem daraus resultierenden Prozentsatz. Die theoretisch geforderten 75% sind wegen der diskreten Punktzahl nur näherungsweise erreicht.

Die höchstmögliche Punktzahl der Kompetenzstufe 1 ist 21. Bei einer festgelegten Zielmarke von 75% muss eine Person also

21 × 0,75 = 15,75 Punkte erreichen, um als Könner zu gelten. Da dieses Ergebnis mit einem Messfehler behaftet ist, kann es also sein, dass eine Person einen Punktwert von z.B. 16 Punkten erlangt, unter Einbezug des Messfehlers jedoch unter 15,75 Punkten liegt. Um sicher zu gehen, dass der wahre Punktwert mindestens 15,75 Punkte beträgt, muss man den real geforderten Wert auf 18,38 erhöhen, denn dann liegt die untere Grenze des Vertrauensintervalls beim 15,75. Die geforderte Zielmarke muss auf 18 Punkte abgerundet werden, da es nur ganzzahlige Punktwerte gibt. 18 Punkte entsprechen einer Vertrauensintervalluntergrenze von 15,37 Punkten. Diese entsprechen wiederum 73,19%. Durch das Vertrauensintervall wurde also mit einem 95%igen Vertrauensgrad sichergestellt, dass der Proband die erforderliche Zielmarke erreicht hat. Mit den übrigen beiden Kompetenzstufen wurde analog verfahren.

Abbildung 11: Untere Grenze der Vertrauensintervalle, geforderte Zielmarke und damit einhergehender real geforderter Prozentsatz richtiger Lösungen



Dieses Verfahren wertet den α-Fehler (hier: Unterschätzung des Probanden, vgl. Klauer, 1987, S. 61) und den β-Fehler (hier: Überschätzung eines Probanden) unterschiedlich. Durch die strenge Festlegung wird eher ein α-Fehler begangen, d.h. das Modell weist eher einen Könner zurück, als es einen Nicht-Könner durchlässt.

2.2.2 Analysen der Kompetenzstufen

Die folgenden Analysen betreffen die Kompetenzstufen. Es ist nicht möglich mit einer Testkonstruktion Hypothesen zu prüfen. Deshalb soll hier auch nicht getestet werden, ob die Kompetenzstufen theoretisch aufeinander aufbauen oder gar, welche Module (z.B. kognitiver Art) für die einzelnen Kompetenzstufen herangezogen werden können. Es soll lediglich gezeigt werden, ob die nach testtheoretischen Maßgaben konstruierten Kompetenzstufen im Test sich mit dem theoretischen Konstrukt vereinbaren lassen. Dazu werden zunächst die Schwierigkeiten der Kompetenzstufen miteinander verglichen, um dann zu testen, ob sich für die Probanden des Tests zeigen lässt, dass die Kompetenzstufen aufeinander aufbauen.

Vergleich der Schwierigkeiten der Kompetenzstufen

Kompetenzstufe 3 baut auf Stufe 1 auf, d.h., um Weltwissen mit Informationen aus dem Text zu verknüpfen, ist es natürlich unabdingbar, dass es dem Probanden möglich ist, Informationen aus dem Text zu ermitteln. Der Prozess „Weltwissen mit Informationen zu verknüpfen" müsste also schwieriger sein, als nur die bloße Information zu ermitteln, die Schwierigkeit einer Frage auf Stufe 3 müsste also größer sein, als eine Frage auf Stufe 1. Im Durchschnitt der Fragen ist dies auch der Fall. Schließlich wurde darauf schon während der Itemselektion ein Augenmerk darauf gelegt. Die Schwierigkeit nimmt mit zunehmender Kompetenzstufe zu (s. Tabelle 22, vgl. dazu auch Teil II, Kapitel 2.1.5 , Haupteffekt Kompetenzstufen).

Tabelle 22: Mittelwerte der Schwierigkeitsindizes für die Kompetenzstufen und die Schulabschlüsse

Kennwert	Skala	Gesamt	Schulabschluss			
			HS	MR	FH	AB
n	alle	576	79	273	93	52
m der Schwierigkeit	Gesamt	0,68	0,54	0,67	0,76	0,79
	1	0,76	0,62	0,75	0,84	0,86
	2	0,64	0,50	0,63	0,71	0,74
	3	0,57	0,45	0,55	0,66	0,73

Die Reihenfolge der Schwierigkeiten der Fragen ist jedoch nicht so eindeutig. Dies sei an zwei Beispielen erläutert. Beispiel 1 zeigt eine Frage zu Kompetenzstufe 1. Offensichtlich ist diese Frage schwierig zu beantworten. Bei den Hauptschülern zeigt sich eine Gleichverteilung der Antwortmöglichkeiten, d.h. die Verteilung weicht nicht vom „Raten“ ab. Bei den Abiturienten haben immerhin 50 % die richtige Lösung gefunden. Die Frage bezieht sich nur auf eine einfache Information, allerdings ist Weltwissen in bestimmtem Umfang nötig. Unter Umständen wissen die Schüler nicht, welche Achse die X-Achse ist. Allerdings sind sämtliche Antworten auch für die Y-Achse falsch.

Beispiel 1.

Frage 30: **Was wird auf der X-Achse dargestellt?**	**Prozent**
a. Die Anzahl der Nutzer	27,4
b. Der Anteil der Internettätigkeiten in Prozent	40,0
c. Der Anteil von Jungen und Mädchen	32,7

Beispiel 2 zeigt eine Frage zu Kompetenzstufe 3. Sie müsste eigentlich verhältnismäßig schwer sein, da eine Information mit dem Weltwissen verknüpft werden muss.

Beispiel 2.

Frage 28: **Ein 25-Jähriger raucht seit seinem 15. Lebensjahr. Trotz der Bitten seiner Freundin hört er nicht auf zu rauchen. Warum?**	**Prozent**
a. Weil er weiß, dass er jederzeit aufhören könnte.	3,3
b. Weil er noch keine gesundheitlichen Beeinträchtigungen bemerkt.	82,9
c. Weil er sich von Rauchverboten nicht bevormunden lassen will.	13,7

Es kann also durchaus sein, dass es schwieriger ist, eine Information aus einem Text zu extrahieren, die verborgen ist, als eine leicht zu entdeckende Information mit Weltwissen zu verknüpfen. Der Test diente nicht dazu, dies zu testen. Letztlich ließe sich der Test so konstruieren, dass die Stufen aufeinander aufbauen. Eine Testkonstruktion kann nicht dem Hypothesentesten dienen. Es müssten weitere Analysen zum Zusammenhang der Kompetenzstufen folgen, um abschließend zu klären, ob sie aufeinander aufbauen.

Ist das empirische Kompetenzstufenmuster theoriekonform?

Legt man die im Teil II, Kapitel 2.2.1 erstellten Kriterien an die Stichprobe an, sieht man, dass der Anteil der Schüler, der die jeweilige Kompetenzstufe erreicht hat, sehr unterschiedlich ist. Kompetenzstufe 1 haben 40,5% der Probanden erreicht, Stufe 2 nur 3,3% und Stufe 3 6,3%. Man sieht, dass die Stufe 2 etwas zu streng bewertet wurde. Durch die Tatsache, dass das Vertrauensintervall komplett über der theoretisch geforderten Zielmarke liegt, ist die Zuordnung in eine Richtung verschoben: Das unberechtigte Nichtbestehen einer Kompetenzstufe wiegt schwerer als das ungerechtfertigte Bestehen.
Tabelle 23 zeigt die Anteile der Probanden, die die jeweilige Kompetenzstufe erreicht haben aufgeteilt nach Schulabschluss. Wären die Kompetenzen streng aufeinander aufbauend, dann dürfte es nicht vorkommen, dass Probanden eine Kompetenzstufe erreicht haben, ohne die darunter liegende erreicht zu haben. In der Gruppe der Abiturienten sieht man jedoch, dass 27,3% die Kompetenzstufe 3 erreicht haben, jedoch nur 7,3% die Stufe 2. Das heißt, die Stufen bauen nicht aufeinan-

der auf. Ermittelt man den Anteil der Personen, die mindestens eine Kompetenzstufe übersprungen haben, ergibt sich ein Anteil von 5,9%. Dies ist ein guter Wert, der aber auch zum Teil daran liegt, dass die Schwierigkeit für die Probanden mit niedrigem oder mittlerem Schulabschluss sehr hoch ist. Wenn alle Probanden keine Kompetenzstufe bestehen, wäre der Anteil 0%. Bei den Abiturienten ist der Anteil somit auch erheblich höher: 23,6% überspringen mindestens eine Stufe.

Tabelle 23: Anteil der Probanden, die die jeweilige Kompetenzstufe bestanden haben, getrennt nach Schulabschluss

	Schulabschluss			
Kompetenzstufe	niedriger Abschluss	mittlere Reife	Fachhochschulreife	Abitur
1	13,8	36,5	55,9	80,0
2	0,0	1,6	9,8	7,3
3	1,1	3,2	9,8	27,3

2.3 Normierung des Tests

Neben der kriteriumsorientierten Auswertung kann der Test auch normorientiert ausgewertet werden, d.h. jeder Proband kann mit einer Gruppe verglichen werden, um ihn ins Verhältnis zu dieser zu setzen. Dazu ist es nötig, dass Normen vorliegen, die dies ermöglichen. Die Normen sind im Handbuch des Tests erschienen (Drommler, et al., 2006). Im Folgenden werden Normierungsstichprobe und Normierung beschrieben.

2.3.1 Normierungsstichprobe

Als Normierungsstichprobe diente die Stichprobe des dritten Pretests (s. Teil II, Kapitel 1.3.3). Die Stichprobe zeigte zwar eine gute Repräsentativität bezüglich der Gruppe der Berufsschüler des dualen Systems, problematisch ist jedoch, dass nach der Erhebung der Stichprobe und Durchführung der Itemselektion drei Fragen leicht geändert und eine gestrichen wurde, denn *„nach der Erstellung der Test-Endform darf diese nicht mehr verändert werden, wenn die Normen Gültigkeit behalten sollen“*

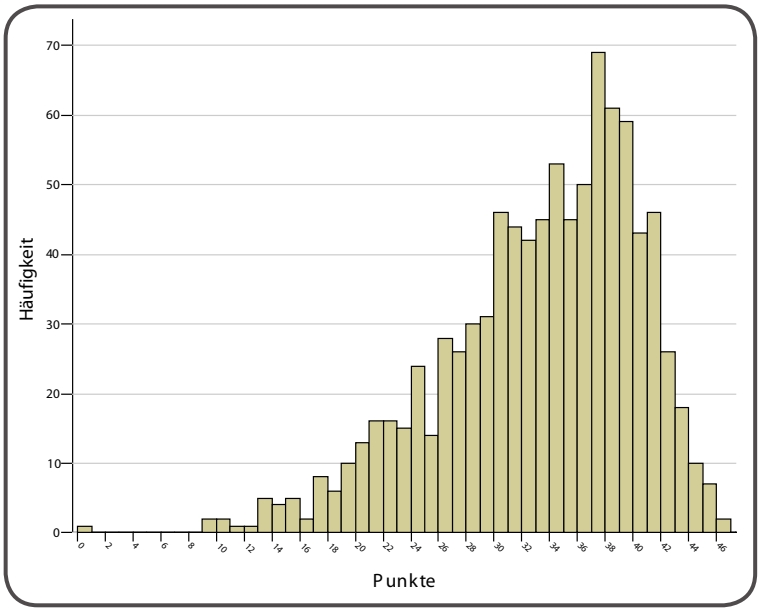
(Perleth und Sierwald, 1991, S. 72). Demnach kann die Normierung nur vorläufig sein.

2.3.2 *Verfahren der Normierung*

Für den LTB^{-3} wurden Normen ausschließlich für die Gesamtpunkte errechnet. Die Normierungsstichprobe wurde aufgeteilt in die Schulabschlüsse „kein oder niedriger Abschluss“, „mittlere Reife“, „Fachhochschluabschluss“ und „Abitur oder Hochschulabschluss“, da sich hier die größten Unterschiede in der Varianzanalyse ergeben haben, und in Normen für Schüler mit Deutsch als Muttersprache (L_1-Schüler) oder als Fremdsprache (L_2-Schüler), da auch dieser Unterschied signifikant war. Letzteres ist zudem wichtig, weil das prozentuale Verhältnis von L_1 und L_2-Schülern in den Schulen sehr verschieden sein kann und somit die Möglichkeit gegeben wird, die Schüler auch in Schulen mit sehr verschiedener Verteilung bezüglich dieser Variable zu vergleichen. Auf getrennte Normen für die Geschlechter wurde verzichtet.

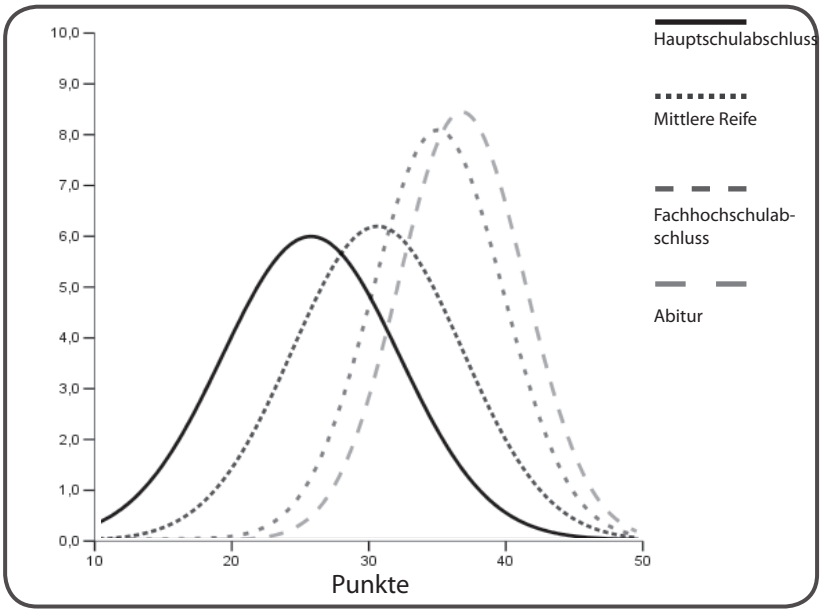
Abbildung 12 (S. 97) zeigt die Verteilung der Rohwerte des Tests. Die Verteilung hat die Form einer rechtsgipfligen Normalverteilung. Zur Berechnung der Normwerte ist daher eine Umrechung in z-Werte unangebracht. Daher wurde auf z-Wert-basierte Normen verzichtet und die genaue kumulative Häufigkeit der Rohwerte zugrunde gelegt. Diese Prozentränge werden im Testhandbuch direkt angegeben, sie wurden aber zusätzlich in T-Werte umgerechnet, um die Möglichkeit zu geben, die Testleistung mit Leistungen in anderen Tests zu vergleichen. Der Mittelwert der T-Skala beträgt 50, die Standardabweichung 10 Punkte. Für die Umrechnung wurden die Prozentränge in z-Äquivalente umgewandelt und dann linear in T-Werte transformiert.

Abbildung 12: Verteilung der Gesamtpunktzahl



Die Verteilungen der einzelnen Schulabschlüsse sind annähernd normal, deshalb konnten hier die Normen mittels z-Transformation in Prozenträngе und T-Werte errechnet werden. Abbildung 13 (S. 98) zeigt die Verteilung der Gesamtpunkte nach Schulabschluss getrennt. Je größer die Überlappung in einem Test ist, desto weniger trennt er die betreffenden Gruppen.

Abbildung 13: Verteilung der Gesamtpunkte nach Schulabschluss



III. Diskussion

Die vorliegende Arbeit hatte zum Ziel, die Entwicklung und Validierung des „Lesetest für Berufsschüler/innen - LTB-3“ zu beschreiben. Ausgehend vom Leseverständnis der PISA-Studie wurde beschrieben, wie die besonderen Anforderungen der Berufskollegs zu einem Lesetest führten, der nicht nur für Berufsschüler, sondern auch für Abgänger der Sekundarstufe I und für Erwachsenen geeignet ist.

Die Phasen der Testentwicklung, theoretische, empirische und Normierungsphase, dauerten insgesamt 1,5 Jahre.

Der Test wurde auf der Basis der Klassischen Testtheorie entwickelt. Hierfür spielten theoretische und pragmatische Gründe eine Rolle. Trotz einer geforderten kriteriumsorientierten Auswertung konnte der Test nicht kriteriumsorientiert konstruiert werden, da kein lehrzielorientierter, inhaltsvalider Test entwickelt werden sollte. Eine kriteriumsorientierte Auswertung lässt sich auf der Basis des probabilistischen Rasch-Modells mit Hilfe des Proficiency Scalings erstellen, jedoch ist der Aufwand dafür sehr hoch und für das Projekt nicht zu leisten. Die Klassische Testtheorie lässt mit Hilfe der Reliabilität und Vertrauensintervallen um den wahren Wert jedoch auch eine kriteriumsorientierte Auswertung zu.

In der Klassischen Testtheorie werden im Wesentlichen drei Gütekriterien beschrieben: Objektivität, Reliabilität und Validität. Hinzu kommen Normierung und eine Evaluation der Testökonomie. Diese für eine Testentwicklung wesentlichen Punkte werden in den folgenden Kapiteln kritisch beleuchtet. Im Fall der Konstruktion von Kompetenzstufen, die in dieser Testentwicklung versucht wurde, kann auch dieses Vorgehen kritisch betrachtet werden. Im Anschluss daran wird ein Fazit gezogen und ein Ausblick auf die weitere Forschung gegeben.

1. Objektivität

Die Objektivität wird im Test durch eine sehr standardisierte Anweisung und eine klare Auswertung und Interpretation gewährleistet. Fehler, die zu weniger objektiven Ergebnissen führen könnten, liegen im Bereich des Ablesens von der Schablone und dem damit verbundenen Addieren der einzelnen Punkte.

2. Reliabilität

Die Reliabilität erreicht im dritten Pretest eine für Leistungstests ausreichende Höhe für den Gesamtpunktwert. Somit ist eine normbezogene Auswertung des Gesamtpunktwertes mit geringen Fehlern versehen. Die Reliabilität für die einzelnen Kompetenzstufen ist jedoch geringer. Für die zweite und dritte Kompetenzstufe ist die Reliabilität weit unterhalb der für Leistungstests geforderten Reliabilität. Dies hat zur Folge, dass die Vertrauensintervalle um den wahren Wert bei diesen beiden Kompetenzstufen wesentlich größer sind, als für die erste Stufe. Die real geforderte Zielmarke liegt daher weiter von der wahren Zielmarke entfernt. Anzumerken ist auch, dass die Reliabilitäten im dritten Pretest geringfügig kleiner sind als für den zweiten Pretest, die Itemselektion hat somit nicht den gewünschten Effekt gebracht. Durch die größere Anzahl der Versuchspersonen beim dritten Pretest, sind die Werte der Reliabilität jedoch mit geringeren Fehlern versehen.

Die Reliabilität wurde ausschließlich mit Hilfe der Split-half Reliabilität und mit Cronbachs α berechnet, also mit Maßen der inneren Konsistenz. Eine weitere Analyse der Reliabilität mit Hilfe einer Testwiederholung wäre wünschenswert. Ein Lerneffekt des Lesens, also eine Kompetenzsteigerung alleine durch das Bearbeiten des Tests, wäre zwar auszuschließen, nicht aber ein Effekt der Testkenntnis. Da der Test jedoch kein Speed-Test ist und keine Gedächtnisleistung erbracht werden muss, wäre dies hinzunehmen.

3. Validität

Die durchgeführten Faktorenanlysen sowie die Unterschiedsuntersuchungen bezüglich der Kompetenzstufen und der Textarten zeigten, dass die Textarten sich eher zur Erklärung der Unterschiede zwischen den Geschlechtern eignen. Offensichtlich sind die gewählten Textarten sehr verschieden. Die größten Unterschiede zeigten sich jeweils beim Diagramm. Es trennte nicht nur die beiden Geschlechter, sondern auch die Schulabschlüsse am besten. Das eigentliche Vorhaben, die drei Kompetenzstufen für jede der Textarten zu verwirklichen, musste aufgegeben werden, da für eine reliable und valide Prüfung der Lesefähigkeit mehr als drei Fragen pro Kompetenzstufe erforderlich wären. Ein Test mit jeweils z.B. acht Fragen pro Textart und Kompetenzstufe hätte jedoch einen Test zur Folge gehabt, der den Rahmen von zwei Schul-

stunden sprengen würde und somit nicht ökonomisch geworden wäre. Außerdem ist es schwierig, zu den relativ kurzen Texten ausreichend viele Fragen zu stellen.

Die Heterogenität, gemessen mit Loevingers H, aber auch die Korrelationen zwischen den Items zeigen, dass das zu Grunde gelegte Konstrukt „Lesen" sehr heterogen ist. Es ist also zu vermuten, dass am Lesen sehr viele unterschiedliche Prozesse beteiligt sind. Jede Testfrage misst zwar in einem bestimmten Umfang „Lesekompetenz", jedoch ist die spezifische Varianz, die jedes Items nicht mit den anderen teilt, sehr groß. Vor diesem Hintergrund ist es fraglich, ob Skalenwerte, d.h. eine Zusammenfassung von Punkten, bei derartig heterogenen Tests überhaupt sinnvoll sind. Das heterogene Testergebnis kommt sicherlich auch deshalb zustande, weil die Lesefähigkeit entweder nicht klar definiert ist, oder sie tatsächlich so komplex ist, dass es eine Lesefähigkeit im engeren Sinne nicht gibt, da zuviele Teilprozesse daran beteiligt sind. Schlussfolgerndes Denken und Gedächtnis sind nur zwei von ihnen, die in diesem Test nicht einmal abgefragt wurden. Weltwissen jedoch, das in der dritten Kompetenzstufe vorausgesetzt wird, ist kein hinreichender Teil der Lesekompetenz.

Die Ergebnissen der Faktorenanalyse und zur Homogenität deuten auf eine nur geringe Konstruktvalidität hin. Der „Pretest ohne Darbietung der Texte" zeigte ein weiteres Argument für die mangelnde Konstruktvalidität: Es ist durchaus möglich, einige Fragen ohne das Lesen der Texte, des Diagramms oder der Tabelle richtig zu beantworten. Da der Lesetest jedoch kein Test für schlussfolgerndes Denken ist, lässt sich hier in Bezug auf die Validität ein Problem erkennen. Die Konsequenz daraus könnte lauten, dass die Texte nicht nach der Lebenswelt der Berufsschüler auszusuchen sind, da die Schüler zuviel Vorwissen besitzen. In diesem Fall müsste das Argument der Motivation neu überdacht werden, jedoch schienen die Berufsschüler bereits vor dem Bearbeiten der Tests eine genügend hohe Motivation zu seiner Bearbeitung zu besitzen. Dieses Problem stellt sich natürlich nicht, wenn der Test als „Selektionsinstrument", z. B. als Einstellungskriterium, genutzt wird.

Die Korrelationen mit den Außenkriterien als externes Validitätsmerkmal sind ausreichend hoch. Es fehlen jedoch wichtige Außenkriterien wie die Schulnote oder ein Rating der Lehrer. Die Frage nach den letzten Noten der Schüler im Fach Deutsch und Mathematik konnte nicht mit der Testpunktzahl korreliert werden, da die Frage nach den Schulnoten zu unpräzise formuliert war. Außerdem ließen sich die

Noten nicht vergleichen, da sie an verschiedenen Schulformen gewonnen wurden. Somit fehlen wichtige Zusammenhänge mit ähnlichen Konstrukten.

Für die Berufsschulen wäre es darüber hinaus wichtig zu wissen, wie hoch der Zusammenhang zwischen Lesekompetenz und Ausbildungserfolg bzw. Berufserfolg ist. Eine prognostische Validität fehlt jedoch bislang völlig.

4. Normierung

Die Normierung des Tests erfolgte nach der letzten Änderung der Items. Die Reliabilität sowie die Mittelwerte sind jedoch für den dritten Pretest und die Endversion vergleichbar. Trotzdem wäre es wünschenswert die Normierung während des praktischen Einsatzes des Tests zu korrigieren. Auf eine Normierung der einzelnen Kompetenzstufen wurde aus Gründen mangelnder Reliabilität verzichtet.

5. Testökonomie

Die Ökonomie des Tests kann abschließend als zufriedenstellend bewertet werden. Der Test lässt sich innerhalb einer doppleten Schulstunde einsetzen. Die Auswertung ist äußerst praktikabel und auch für einen Klassensatz in geringer Zeit zu leisten.

6. Bildung der Kompetenzstufen

Anders als im PISA-Test (PISA 2000) wurden die Kompetenzstufen bereits vor der Testkonstruktion festgelegt. Die Fragen wurden theoriegeleitet konstruiert. Die Fragen hätten während der Testentwicklung so konstruiert werden können, dass die Kompetenzstufen aufeinander aufbauen. Darauf wurde hier verzichtet. So ist eine Analyse der Kompetenzstufen erst möglich geworden und hat gezeigt, dass die Stufen nicht „automatisch“ aufeinander aufbauen. Ist dies kein Beweis, so ist es jedoch ein Hinweis darauf, dass die Kompetenzstufen keine hierarchisch aufeinander aufbauenden Stufen sind.

Die besondere Art der Bildung der Kompetenzstufen nach dem Modell der Fehlertoleranzen nach Klauer führte zu einer stärkeren

Gewichtung des α-Fehlers. Der Test weist also eher einen Könner zurück, als er einen Nicht-Könner durchlässt. Dies ist insofern akzeptabel, dass eher ein Schüler gefördert würde, der es nicht nötig hat. Dies kann zwar auch negative Konsequenzen haben (Unterforderung), jedoch ist jeder Schüler, der trotz Hilfebedarf nicht gefördert wird, davon bedroht, einen schlechteren Abschluss zu erreichen. Ein schlechter Schulabschluss führt aber letztlich zu geringeren Chancen in der Berufswelt Fuß zu fassen.

Diese Vorgehensweise könnte aber in Zukunft durch ein anderes Modell „entschärft" werden. Durch die Bildung eines zweiten Konfidenzintervalls, das komplett unterhalb der Zielmarke liegt, könnten Schüler, die auf keinen Fall als Könner gelten, besser aufgedeckt werden, wogegen das Konfidenzintervall, das komplett oberhalb der Zielmarke liegt, dazu dient, Könner aufzudecken. Dazwischen wird es Schüler geben, die in einem Indifferenzbereich liegen. Es müsste entschieden werden, wie diese Schüler zu klassifizieren sind. Solange jedoch die praktischen Konsequenzen für diese Schüler diffus bleiben (Leseförderung), kann auch eine sinnvolle Gewichtung des α- und β-Fehlers nicht vorgenommen werden.

7. Fazit und Ausblick

Mit Hilfe der Normierung lässt sich der Test gut nutzen, um einen Schüler hinsichtlich der Lesekompetenz einzuschätzen. Das heterogene (gemessene) Konstrukt muss dabei kein Nachteil sein, wenn am Lesen tatsächlich viele Prozesse beteiligt sind. Fraglich ist jedoch, ob die Kompetenzstufen tatsächlich den Anforderungen der Berufskollegs entsprechend, denn das Testergebnis sollte individuelle Stärken und Förderbedarfe offenlegen (vgl. Teil I, Kapitel 2.2) und möglichst konkrete Diagnoseergebnisse bereitstellen. Dies ist jedoch nur zum Teil verwirklicht. Sicherlich kann für einen Schüler entschieden werden, ob er basale Probleme mit dem Textverständnis hat, ob ein Nichtbestehen der dritten Kompetenzstufe an der mangelnden Lesefähigkeit, an fehlendem Weltwissen oder an einer mangelnden Verknüpfungsleistung liegt, kann nicht entschieden werden.

Die Definition der PISA-Studie hebt auf Lesen als eine Basisqualifikation ab. Lesen dient nach PISA dazu, *„Ziele zu erreichen, das eigene Wissen und Potential weiterzuentwickeln und am gesellschaftlichen Leben teilzunehmen"* (PISA 2000, S. 78). Die gestellten Fragen sollten authentische Situationen darstellen. Es ist allerdings, im Hinblick auf Validi-

tät, zu erkennen, dass ein Test (besonders ein Paper-and-pencil Test) keine besonders authentische Situation darstellt. Texte zu lesen, sie zu verstehen und sich zu eigen zu machen hilft sicherlich, Wissen und Potential weiterzuentwickeln. Ob anhand von empirisch ermittelter Leseleistung jedoch ein Rückschluss darauf gezogen werden kann, inwieweit jemand am gesellschaftlichen Leben teilnimmt, ist eine noch unbeantwortete Frage. Soziologische und sozialpsychologische und -pädagogische Studien müssten dies zu klären versuchen.

Wie an anderer Stelle bereits erwähnt, eignet sich eine Testentwicklung nicht zur Hypothesentestung. Als Ansatzpunkte für weitere Untersuchungen kann er aber sehr wohl geeignet sein.

Noch ungeklärt ist, warum fast gleichartig lautenden Fragen so gering miteinander korrelieren. Möglich wäre, dass es am (linguistischen) Umfeld der Information im Text liegt. Es könnte auch sein, dass leichte Unterschiede in der Fragestellung oder in den Antwortmöglichkeiten einen großen Effekt haben. Letztlich könnte es daran liegen, ob das nötige Vorwissen unterschiedlich hoch ist.

Der „Lesetest für Berufsschüler/innen" ist ein Test, der sich nicht an kognitionspsychologischen Vorstellungen orientiert, d.h. es soll weder schlussfolgerndes Denken noch eine Gedächtnisleistung getestet werden. Mögen diese beiden kognitiven Aspekte auch unweigerlich Prozesse sein, die beim Lesen eine große Rolle spielen, so sind sie jedoch nicht damit gleichzusetzen. Ohne diese beiden Prozesse zu betrachten, bleibt die Erforschung der Leseleistung bzw. das Testen dieser jedoch unvollständig. Legt man das Reading-Literacy-Konzept zu Grunde, muss auch untersucht werden, ob die Schüler sich die Texte, das Diagramm und die Tabellen „zu eigen" gemacht haben. Dies zu untersuchen wäre ein weiterer Schritt zur Erkundung der Validität.

Im „Lesetest für Berufsschüler/innen" wird das Bestehen der Kompetenzstufen mit Hilfe von Vertrauensintervallen bestimmt. In weiteren Untersuchungen müsste geklärt werden, ob die theoretisch hergeleiteten Kompetenzstufen, wie gefordert, aufeinander aufbauen. Dazu muss sichergestellt werden, dass die Fragen valide und reliabel genau eine Kompetenz messen. Das Ergebnis einer Faktorenanalyse wären dann drei orthogonale Faktoren.

Zuletzt lässt sich noch die Frage stellen, ob und wie man Lesen tatsächlich „fördern" kann. Dies ist jedoch erst möglich, wenn die Prozesse erkannt sind, die einen Leser zu einem Experten machen.

IV. Zusammenfassung

Die vorliegende Arbeit zeigt die Entwicklung und Validierung eines Tests zur Erfassung der Lesekompetenz von Berufsschülern- und schülerinnen. Den Anstoß dazu gab ein Projekt zur Leseförderung in der Berufsbildung an drei berufsbildenden Schulen in Düren und Jülich. Leseförderung kann nur gezielt wirksam werden, wenn die Lesestärken- und schwächen der Schüler hinreichend genau diagnostiziert wurden. Der entwickelte Test leistet einen nützlichen Beitrag zur Einschätzung der Lesekompetenz besonders der leseschwachen Schüler.

Lesekompetenz im „Lesetest für Berufsschüler und -schülerinnen - LTB^{-3}" wird analog zur PISA-Studie verstanden als die Fähigkeit *„geschriebene Texte zu verstehen, zu nutzen und über sie zu reflektieren, um eigene Ziele zu erreichen, das eigene Wissen und Potential weiterzuentwickeln und am gesellschaftlichen Leben teilzunehmen"* (PISA 2000, S. 78). Damit geht der Lesetest über die Fähigkeit, Buchstaben, Wörter und Sätze zu entschlüsseln hinaus und stellt das Textverständnis in den Vordergrund. Dieses wird in drei Kompetenzstufen gemessen: „Informationen ermitteln", „mehrere Informationen verknüpfen" und „Informationen mit dem Weltwissen verknüpfen". Neben authentischen, nicht-fiktiven Fließtexten werden auch Tabellen und ein Diagramm dargeboten, denn diese Textarten sind besonders in der Berufsausbildung von großem Interesse. Der Test ist als Gruppentest konzipiert. Die Schüler arbeiten mit dem Text, d.h. zur Beantwortung der Fragen dürfen sie diesen benutzen.

Die Entwicklung des Tests von der ersten Konzeption bis zur Veröffentlichung dauerte etwa 1,5 Jahre. Mehrere Pretests gingen der endgültigen Version voraus, die an über 570 Berufsschülern normiert wurde. Die Konstruktion des Tests geschah auf Basis der Klassischen Testtheorie. Die Objektivität des Tests ist durch eine standardisierte Instruktion, Durchführungs- und Auswertungsanleitung gewährleistet. Die Auswertung geschieht mit Hilfe einer Schablone. Die Reliabilität ist mit 0,831 (Cronbachs α) für die Gesamtpunkte für einen Leistungstest ausreichend hoch. Die Reliabilitäten für die einzelnen Kompetenzstufen sind geringer und für die Kompetenzstufen 2 und 3 mit 0,57 bzw. 0,53 für einen Leistungstest zu niedrig. Dies hat ein großes Vertrauensintervall um den wahren Wert zur Folge, was wiederum bedeutet, dass

die Zielmarke des beobachteten Wertes für die kriteriumsorientierte Auswertung weit über der geforderten Zielmarke liegen muss, um dem Messfehler gerecht zu werden. Die Retestreliabilität konnten nicht berechnet werden. Die Kriteriumsvalidität mit dem Kriterium „bereits erreichter Schulabschluss“ beträgt für den Gesamttestwert 0,50, was als ausreichend angesehen wird. Eine höhere Korrelation würde den Test obsolet machen, denn der Schulabschluss ist den Berufsschulen bekannt.

Faktoren- und Varianzanalysen konnten zeigten, dass nicht die Kompetenzstufen getrennte Skalen sind, sondern die Textarten Fließtexte, Tabellen und Diagramme einen weitaus höheren Effekt auf die Unterschiede der Leseleistung hatten: Diagramme werden schlechter gelesen als Texte. Die Kompetenzstufen ließen sich faktoriell nicht abbilden.

Die Untersuchungen zur Validität zeigten insgesamt, dass weitere Untersuchungen zur Lesekompetenz nötig sind, denn besonders die große Heterogenität ist auffällig. Die Korrelationen zwischen den Items ist durchweg sehr gering, und selbst gleichartig gestellte Fragen korrelieren nur mäßig.

Mit Hilfe der Normierung lässt sich der Test, besonders um schlechte Schüler hinsichtlich ihrer Lesekompetenz einzuschätzen, gut nutzen. Ob jedoch die Kompetenzstufen einen wesentlichen Beitrag zu dieser Einschätzung leisten, ist angesicht der geringen Reliabilität und Validität der zweiten und dritten Kompetenzstufe fraglich.

Literatur

Anderson, J. R. (2001). Kognitive Psychologie. Heidelberg: Spektrum Akademischer Verlag.

Arras, U./Grotjahn, R. (2002): TestDaF: Aktuelle Entwicklungen. In: Fremdsprachen und Hochschule, 66, S. 65-68.

Baumert, J./Klieme, E./Neubrand, M. u.a. (=Deutsches PISA-Konsortium) (Hg.) (2001) PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich.

Becker-Mrotzek, M./Drommler, R. (2006) Texte lesen. In: Becker-Mrotzek, M./Kusch, E./Wehnert, B. (Hg.) Leseförderung in der Berufsbildung. Köln: KöBeS Heft 2 (Kölner Beiträge zur Sprachdidaktik), Gilles & Francke, S. 17-44.

Beckett, Samuel (1954). Waiting for Godot. Frankfurt a. M.: Suhrkamp, 32. Auflage.

Bos, W./Lankes E.M./Prenzel, M./Schwippert, K./Walther, G./Valtin, R. (Hg.) (2003). Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. Münster: Waxmann.

Büscher, P. (1984). Theoretische Ansätze der kriteriumsorientierten Leistungsmessung. In: Heller, K. A. (Hg.) (1984) Leistungsdiagnostik in der Schule. Stuttgart: Verlag Hans Huber. 4. Auflage, S. 99-116.

Carlson, N. R./Buskist W./Martin G.N. (2000). Psychology. The science of behaviour. Harlow, UK: Pearson Education.

Cohen, J. (1977). Statistical Power Analysis for the Behavioral Sciences. New Jersey: Lawrence Erlbaum Associates.

Ehlich, K. (1983). Text und sprachliches Handeln. Die Entstehung von Texten aus dem Bedürfnis nach Überlieferung. In: Assmann, A./ Assmann, J./Hardmeier, Ch. (Hg.) (1983) Schrift und Gedächtnis. München: Fink, S. 24-43.

Ellis, A.W. (1992). Reading, Writing and Dyslexia: A cognitive analysis. Hove, UK: The Psychology Press, 2. Auflage.

Fippinger, F., (1978). Allgemeiner Schulleistungstest für 4. Klassen (AST 4). Göttingen :Hogrefe, 1. Auflage.

Fippinger (1992). Allgemeiner Schulleistungstest für 4. Klassen (AST 4). Hrsg. von Ingenkamp, K. Göttingen: Hogrefe, 3. Auflage.

Fritz, A./Suess, A. (1986). Lesen. Die Bedeutung der Kulturtechnik Lesen für den gesellschaftlichen Kommunikationsprozeß. Konstanz: Universitätsverlag Konstanz GmbH

Hager, W./Spies, K./Heise, E. (2001). Versuchsdurchführung und Versuchsbericht. Ein Leitfaden. Göttingen: Hogrefe.

Huber, H. P. (1973). Psychometrische Einzelfalldiagnostik. Weinheim: Beltz.

Jansen, H./Mannhaupt G./Marx H./Skowronek, H.: Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten (BISC) (1999). Göttingen: Hogrefe.

Just, M.A./Carpenter, P.A. (1980). A theory of reading: From eye fixation to comprehension. In: Psychological Review, 87, S. 329-354.

Kaiser, F.-J./Pätzold, G. (1999). Wörterbuch Berufs- und Wirtschaftspädagogik. Bad Heilbrunn: Klinkhardt.

Kintsch, W./van Dijk T.A. (1978). Toward a model of text comprehension and reproduction. In: Psychological Review, 85, S. 363-394.

Klauer, K. J. (1987). Kriteriumsorientierte Tests. Göttingen: Verlag für Psychologie.

Kusch, E. (2006). Problembeschreibung. In: Becker-Mrotzek, M./Kusch, E./Wehnert, B. (Hg.) Leseförderung in der Berufsbildung. Köln: KöBeS Heft 2 (Kölner Beiträge zur Sprachdidaktik), Gilles & Francke, S. 14-16.

Lehrl, S. (1999) Mehrfachwahl-Wortschatz-Intelligenztest: Manual mit Block MWT-B. Balingen: Spitta Verlag GmbH.

Langfeldt, H.-P. (1984). Die klassische Testtheorie als Grundlage normorientierter (standardisierter) Schulleistungstests. In: Heller, K. A. (Hg) (1984) Leistungsdiagnostik in der Schule. Stuttgart: Verlag Hans Huber. 4. Auflage, S. 65-98.

Lienert, G. A. (1989). Testaufbau und Testanalyse. München: Psychologie Verlags Union

Mathews, V. (1974). Zur Situation in Amerika. In: Bismarck, Klaus (Hg.) Köln: Die Rolle des Buches im audiovisuellen Zeitalter.

Martin, M.O./Kelly, D.L. (1998). Third International Mathematics and Science Study (TIMSS). Technical Report (Volume I): Design and Development. Chestnut Hill, MA: Boston College.

Perleth, C./Sierwald, W. (1991). Testtheoretische Konzepte der Begabungsdiagnostik. In: Heller, K. A.: Begabungsdiagnostik in der Schul- und Erziehungsberatung. Göttingen: Hans Huber.

Rost, J. (1996). Lehrbuch Testtheorie, Testkonstruktion. Göttingen: Verlag Hans Huber.

Stevens, T./Wahlers, J. (2006). Entwicklung des Diagnoseinstruments zur Lesekompetenz. In: Becker-Mrotzek, M./Kusch, E./Wehnert, B. (Hg.) Leseförderung in der Berufsbildung. Köln: KöBeS Heft 2 (Kölner Beiträge zur Sprachdidaktik), Gilles & Francke, S. 72-75.

Benutzte Texte / Tabellen / Grafiken für den Test:

„Das Jugendarbeitsschutzgesetz". In: autoFachmann, Ausgabe 1, August 2002. Würzburg: Vogel Auto Medien GmbH & Co. KG, S. 7-8.

„Generelles Rauchverbot an Schulen?". www.kinderaerzteimnetz.de/bvkj/aktuelles1/show.php3?id=1434&nodeid=26.

„Internet-Tätigkeit 2002". JIM2002, Stuttgart: mpfs-Medienpädagogischer Forschungsverbund Südwest. Weitere Informationen, sowie die neusten Daten unter www.mpfs.de/index.php?id=11.

Anhang

Anhang A: Tabellen

Tabelle A-1: Erster Pretest: Ausbildungsgänge

genannter Ausbildungsgang	Anzahl der Probanden	Prozent
Verpackungsmittelmechaniker	18	10,7
Sanitär-Heizung-Klima	14	8,3
Bäckereifachverkäuferin	11	6,5
Mechatronik	26	15,5
Bau-1	11	6,5
Technischer Zeichner	15	8,9
Fleischereifachverkäuferin	4	2,4
Kaufmann im Groß-und Außenhandel	24	14,3
Automobilkauffrau	15	8,9
Metallbauer	22	13,1
Konstruktionstechniker	1	0,6
fehlend	***7***	***4,2***
Gesamt	**168**	**100**

Tabelle A-2: Zuordnung der Schulabschlüsse auf die Kategorien für den ersten und zweiten Pretest

genannter Schulabschluss	Kategorie
kein Abschluss Sonderschulabschluss Hauptschulabschluss	niedriger Abschluss

genannter Schulabschluss	Kategorie
Realschulabschluss Fachoberschulabschluss Fachoberschulreife mit Qualifikation Hauptschulabschluss 10b FORQ höhere Handelsschule	mittlerer Abschluss
Fachhochschulreife Abitur Hochschulabschluss	hoher Abschluss

Tabelle A-3: Zweiter Pretest: Ausbildungsgänge

genannter Ausbildungsgang	Anzahl der Probanden	Prozent
Elektroniker für Betriebstechnik	14	18,9
Zerspannungsmechaniker	30	40,5
ELU1	2	2,7
Bauzeichner	6	8,1
Energieelektroniker	1	1,4
Textilstopfer	3	4,1
Textilmaschinenführer	18	24,3
Gesamt	**74**	**100**

Tabelle A-4: Dritter Pretest: Ausbildungsgänge

genannter Ausbildungsgang	Anzahl der Probanden	Prozent
Bäckereifachverkäuferin	20	3,9
Landmaschinenmechaniker	30	5,9
Fleischereifachverkäuferin	3	0,6
Automobilkaufmann	41	8,0
Tichler	17	3,3
Kfm. f. Bürokommunikation	38	7,4
Verpackungsmitteltechniker	21	4,1
Industriekaufmann	17	3,3

genannter Ausbildungsgang	Anzahl der Probanden	Prozent
Maler und Lackierer	14	2,7
Elektrotechniker	9	1,8
Industriemechaniker	16	3,1
Elektroniker f. Betriebstechnik	15	2,9
Mechatroniker	25	4,9
Rechtsanwaltsfachangestellte	19	3,7
Informatikkaufmann	12	2,3
Einzelhandelskauffrau	76	14,8
Zerspannungsmechaniker	22	4,3
Fachinformatiker	49	9,6
Elektroniker f. Geräte u. Systeme	19	3,7
Technischer Zeichner	13	2,5
Verwaltungsfachangestellte	10	2,0
fehlend	***26***	***5,1***
Gesamt	**512**	**100**

Tabelle A-5: Zuordnung der Schulabschlüsse auf die Kategorien für den dritten Pretest und die Normierung

genannter Schulabschluss	Kategorie	Abkürzung
kein Abschluss Sonderschulabschluss Hauptschulabschluss	kein oder niedriger Schulabschluss	HS
Realschulabschluss Fachoberschulabschluss Fachoberschulreife mit Qualifikation Hauptschulabschluss 10b FORQ höhere Handelsschule	mittlere Reife	MR
Fachhochschulabschluss	Fachhoschulab-schluss	FH
Abitur Hochschulabschluss	Abitur oder Hoch-schulabschluss	AB

Item Nr.		Schwierigkeit		Trennschärfe		
LL	RHD	unkorrigiert	korrigiert	LL	RHD	
01		,89	,84	,51		Lärm lass nach
02		,76	,64	,24		
03		,82	,72	,42		
04		,87	,80	,49		
05		,87	,80	,62		
06		,97	,96	,71		
07		,82	,72	,45		
08		,84	,76	,36		
09		,76	,64	,33		
10		,87	,80	,45		
11		,87	,80	,54		
12		,71	,57	,37		
13		,47	,21	,05		
14		,95	,92	,80		
15		,74	,61	,45		
16		,71	,57	,35		
17		,79	,68	,36		
18		,13	-,30	,07		
	01	,64	,46		-,03	Rose, Hose, Dose
	02	,94	,91		,11	
	03	,81	,72		-,14	
	04	,89	,84		,21	
	05	1,00	1,00			
	06	,64	,46		,20	
	07	,61	,42		,34	
	08	,89	,84		,48	
	09	,86	,79		,24	
	10	,86	,79		,15	
	11	,94	,91		,27	

Tabelle A-6:

Schwierigkeiten und Trennschärfen für die Items des ersten Pretests für die Testversionen „Lärm lass nach" (LL) (n = 38) und „Rose, Hose, Dose" (RHD) (n = 36). Für die Schwierigkeiten des Diagramms und der Tabellen wurden die beiden Stichproben zusammengefasst (n = 74). Signifikante Unterschiede lagen bei keinem Item vor (Messwiederholungsdesign ANOVA Interaktion Item * Testversion $F(20,1440)=1,33$; $p>0,150$; partielles $\eta^2=0,018$). Die Trennschärfen sind von ihrer „Umgebung" abhängig und unterscheiden sich z.T. erheblich.

Tabelle A-6 *fortgesetzt*

Item Nr.		Schwierigkeit		Trennschärfe		
JS	RHD	unkorrigiert	korrigiert	LL	RHD	
	12	,86	,79		,05	Rose, Hose, Dose
	13	,89	,84		,14	
	14	1,00	1,00			
	15	,83	,75		,28	
	16	,72	,58		,40	
	17	,94	,91		,30	
	18	,86	,79		,39	
19	19	,81	,72	,48	,66	Diagramm
20	20	,93	,90	,73	,40	
21	21	,69	,53	,29	,36	
22	22	,92	,88	,59	,03	
23	23	,89	,84	,58	,27	
24	24	,28	-,07	,24	-,21	
25	25	,64	,45	,10	,06	
26	26	,96	,94	,80	,03	
27	27	,88	,82	,76	,12	
28	28	,78	,68	,61	,00	
29	29	,80	,70	,45	,23	
30	30	,16	-,26	,04	,22	
31	31	,84	,76	,43	,11	Tabellen
32	32	,81	,72	,42	,42	
33	33	,92	,88	,39	,02	
34	34	,85	,78	,56	,44	
35	35	,50	,25	,22	-,06	
36	36	,84	,76	,58	,13	
37	37	,58	,37	,27	-,06	
38	38	,61	,41	,33	,53	
39	39	,74	,61	,28	,36	

Tabelle A-7: Korrelationen zwischen den Kompetenzstufen sowie mit der Gesamtpunktzahl des Tests

			Kompetenzstufe		
			1	2	3
Kompetenzstufe	1	r		0,539	0,520
		p		<0,001	<0,001
	2	r			0,566
		p			<0,001
	korrigierte* Gesamtpunkte	r	0,597	0,620	0,613
		p	<0,001	<0,001	<0,001

*** Gesamtpunkte minus jeweilige Punkte der Kompetenzstufe**

Tabelle A-8: Kennwerte des „Pretest ohne Texte". Die Fragen, deren Schwierigkeit signifikant positiv sind (Fragen, die überzufällig richtig beantwortet wurden), sind grau unterlegt. Dabei bedeuten *** = p<0,001; ** = p<0,01; * = p<0,05. n = 42.

Item-Nr.	korrigierte Schwierigkeit	Vertrauensintervall		(nicht α-justiertes) p
		Untergrenze	Obergrenze	
1	0,05	-0,09	0,18	0,478
2	-0,07	-0,18	0,04	0,198
3	0,02	-0,11	0,15	0,782
4	-0,19***	-0,26	-0,12	0,000
5	-0,57***	-0,70	-0,44	0,000
6	0,43***	0,29	0,57	0,000
7	0,29***	0,13	0,44	0,000
8	-0,20*	-0,35	-0,05	0,013
9	0,57***	0,44	0,70	0,000
10	0,33***	0,18	0,48	0,000
11	0,86***	0,78	0,94	0,000
12	0,24**	0,08	0,39	0,003
13	-0,14	-0,30	0,01	0,072
14	-0,07	-0,21	0,07	0,317

Item-Nr.	korrigierte Schwierigkeit	Vertrauensintervall		(nicht α-justiertes) p
		Untergrenze	Obergrenze	
15	-0,07	-0,21	0,07	0,317
16	0,82***	0,72	0,92	0,000
17	-0,36***	-0,45	-0,26	0,000
18	0,11	-0,05	0,26	0,170
19	-0,18**	-0,31	-0,05	0,008
20	-0,14	-0,28	-0,01	0,038
21	0,68***	0,55	0,81	0,000
22	0,36***	0,20	0,51	0,000
23	0,14	-0,01	0,30	0,072
24	-0,21**	-0,34	-0,09	0,001
25	-0,11	-0,25	0,03	0,126
26	-0,14*	-0,28	-0,01	0,038
27	-0,14*	-0,28	-0,01	0,038
28	0,39***	0,24	0,55	0,000
29	0,00	-0,15	0,15	1,000
30	-0,43***	-0,50	-0,36	0,000
31	-0,29***	-0,40	-0,18	0,000
32	0,25**	0,09	0,41	0,003
33	-0,32***	-0,42	-0,22	0,000
34	0,04	-0,12	0,19	0,636
35	-0,04	-0,18	0,11	0,623
36	-0,07	-0,21	0,07	0,317
37	0,43***	0,28	0,58	0,000
38	-0,25***	-0,37	-0,13	0,000
39	-0,07	-0,21	0,07	0,317
40	-0,04	-0,18	0,11	0,623
41	0,57***	0,43	0,71	0,000
42	-0,25***	-0,37	-0,13	0,000
43	0,39***	0,24	0,55	0,000
44	0,00	-0,15	0,15	1,000

Item-Nr.	korrigierte Schwierigkeit	Vertrauensintervall		(nicht α-justiertes) p
		Untergrenze	Obergrenze	
45	0,43***	0,28	0,58	0,000
46	-0,18**	-0,31	-0,05	0,008
47	0,07	-0,08	0,22	0,352

Anhang B: Material

Die Seitenzahlen auf den Materialien sind die Originalseitenzahlen.

Tabelle B-1: Übersicht über die verwendeten Materialien

Nr.	Material	Seite
B-01	Pretestversion 01. Testheft: „Das Jugendarbeitsschutzgesetz“	98
B-02	Pretestversion 01. Antwortbogen (beispielhaft für die weiteren Antwortbögen): „Das Jugendarbeitsschutzgesetz“	106
B-03	Pretestversion 01. Testheft: „Rauchen“	108
B-04	Pretestversion 01. Testheft: „Rose, Hose, Dose“	115
B-05	Pretestversion 01. Testheft: „Lärm lass nach“	123
B-06	Pretestversion 02. Testheft.	131
B-07	Pretestversion 03. Testheft.	141
B-08	Testversion „Pretest ohne Darbietung der Texte“ (Fragen entsprechen der Pretestversion 3)	nicht enthalten
B-09	Zusatztest „MWT-B“. Antwortbogen 06.	151
B-10	Endfassung. Testheft.	152
B-11	Endfassung. Antwortbogen.	164
B-12	Endfassung. Ergebnisbogen.	166

In den Kölner Beiträgen zur Sprachdidaktik (KöBeS) werden in loser Folge Forschungsergebnisse der Forschungsstelle Sprachentwicklung und Sprachdidaktik an der Universität zu Köln veröffentlicht. Die Herausgeber Prof. Dr. Michael Becker-Mrotzek, Prof. Dr. Ursula Bredel und Prof. Dr. Hartmut Günther sind die Leiter der Forschungsstelle.