



Institut für Schulqualität der Länder
Berlin und Brandenburg e.V.

Expertise MSA / P10 Deutsch

Bezug der Prüfungsaufgaben in Berlin und Brandenburg
2006/07 zu den kompetenzorientierten Anforderungen
der länderübergreifenden Bildungsstandards für das
Fach Deutsch



Institut für Schulqualität der Länder
Berlin und Brandenburg

Impressum

Alle Rechte vorbehalten

© Institut für Schulqualität der Länder Berlin und Brandenburg e. V.

Otto-von-Simson-Str. 15

14195 Berlin

030 84416680

www.isq-bb.de

info@isq-bb.de

Berlin 2008

Vorbemerkung

Das Brandenburger Ministerium für Bildung, Jugend und Sport (MBS) hat gemeinsam mit der Senatsverwaltung für Bildung, Wissenschaft und Forschung (SenBWF) des Landes Berlin im Oktober 2007 in einem Perspektivpapier beschlossen, Möglichkeiten der Angleichung der Prüfung 10 (P10) in Brandenburg und des Mittleren Schulabschlusses (MSA) in Berlin zu prüfen.

Die zu diesem Zweck eingerichtete Arbeitsgruppe auf Ebene der Bildungsverwaltungen umfasst neben Vertretern des MBS und der SenBWF auch Vertreter des ISQ und des Landesinstituts für Schule und Medien Berlin-Brandenburg (LISUM).

In diesem Zusammenhang wurde das ISQ gebeten, eine Expertise in Auftrag zu geben. Sie sollte auf der Basis der Prüfungsaufgaben in Mathematik und Deutsch des Schuljahres 2006/07 für beide Länder vergleichend darstellen:

- ➔ ob und ggf. wie die jeweiligen Prüfungsaufgaben den Bezug zu den kompetenzorientierten Anforderungen der länderübergreifenden Bildungsstandards inhaltlich und fachdidaktisch herstellen;
- ➔ wie gut die in den Prüfungsarbeiten verwendeten Aufgaben- und Antwortformate mit den Qualitätsstandards der Aufgabenentwicklung im Rahmen der Bildungsstandards übereinstimmen;
- ➔ welche Veränderungen empfohlen werden, um die Leistungsüberprüfungen am Ende der Jahrgangsstufe 10 im Sinne der KMK-Bildungsstandards aufgabenseitig zu optimieren.

Für die Fächer Mathematik und Deutsch wurden vom ISQ getrennte Expertisen beauftragt. Die Gutachtergruppen setzten sich für beide Fächer aus Vertretern des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin und ausgewiesenen Expertinnen und Experten der Fachdidaktik des jeweiligen Faches zusammen.

Die Expertise für das Fach Deutsch wird hiermit vorgelegt.

Berlin, im Januar 2008

Dr. Hans Anand Pant

Wissenschaftliche Leitung und Geschäftsführung des ISQ



Institut zur Qualitätsentwicklung
im Bildungswesen



Kommentierung der schriftlichen Prüfungsarbeit zum
Mittleren Schulabschluss 2007 im Fach Deutsch (Berlin)
bzw. der Prüfungen am Ende der Jahrgangsstufe 10
Deutsch (Brandenburg)
hinsichtlich ihrer Orientierung
an den länderübergreifenden Bildungsstandards
aus didaktischer sowie testdiagnostischer Perspektive

Katrin Böhme

(IQB, Humboldt-Universität zu Berlin)

Albert Bremerich-Vos

(Universität Duisburg-Essen)

Inhaltsverzeichnis

1.	Zum Gegenstand dieses Gutachtens	3
2.	Bildungspolitische Einordnung.....	3
2.1	Die Einführung länderübergreifender Bildungsstandards.....	3
2.2	Charakterisierung der länderübergreifenden Bildungsstandards für den Mittleren Schulabschluss im Fach Deutsch	5
3.	Testdiagnostische Grundlegung.....	6
3.1	Kennzeichnung eines standardisierten Leistungstests	6
3.2	Prinzipien bildungsstandardorientierter Testentwicklung am IQB	8
3.3	Fundierung diagnostischer Entscheidungen	11
4.	Beschreibung der Prüfungsaufgaben.....	12
4.1	Berlin	12
4.2	Brandenburg.....	12
5.	Anmerkungen zu den Prüfungsaufgaben des Landes Berlin.....	14
5.1	Übergreifende Anmerkungen zu Items und Aufgaben	14
5.2	Anmerkungen zur Instruktion.....	14
5.3	Aufgaben- und itemspezifische Anmerkungen.....	15
5.4	Anmerkungen zur Bewertung der Prüfungsaufgaben.....	19
5.5	Anmerkungen zu den in der Handreichung verwendeten Kompetenzmodellen.....	19
6.	Anmerkungen zu den Prüfungsaufgaben des Landes Brandenburg.....	21
6.1	Übergreifende Anmerkungen zu Items und Aufgaben	21
6.2	Aufgaben- und itemspezifische Anmerkungen.....	23
6.3	Anmerkungen zur Bewertung der Prüfungsaufgaben.....	27
7.	Vergleich der Prüfungsaufgaben in Berlin und Brandenburg unter dem Gesichtspunkt der Orientierung an den Bildungsstandards	27
8.	Vergleich der Prüfungsaufgaben mit Testaufgaben, welche zur Evaluierung der Bildungsstandards eingesetzt werden	28
9.	Fazit.....	30
	Literatur.....	31

1. Zum Gegenstand dieses Gutachtens

Die nachfolgende Kommentierung der Prüfungsaufgaben der Länder Berlin und Brandenburg für den Mittleren Schulabschluss im Fach Deutsch des Jahres 2007 wird von der Frage geleitet, inwieweit die vorliegenden Prüfungsaufgaben einerseits aus didaktischer und andererseits aus testdiagnostischer Perspektive den länderübergreifenden Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss (vgl. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004) gerecht werden.

Im Rahmen der Expertise wird für beide Länder vergleichend dargestellt,

- ob und ggf. wie die jeweiligen Prüfungsaufgaben mit den kompetenzorientierten Anforderungen der länderübergreifenden Bildungsstandards im Fach Deutsch (Sekundarstufe I) kompatibel sind;
- inwieweit, die Prüfungsaufgaben fachdidaktisch und testdiagnostisch plausibel sind;
- wie die Leistungsüberprüfungen am Ende der Jahrgangsstufe 10 im Sinne der KMK-Bildungsstandards aufgabenseitig optimiert werden können.

2. Bildungspolitische Einordnung

2.1 Die Einführung länderübergreifender Bildungsstandards

Das lediglich mittelmäßige Abschneiden deutscher Schülerinnen und Schüler in internationalen Vergleichsstudien wie PISA, TIMSS oder IGLU hat wiederholt deutlich gemacht, dass in Deutschland für einen Anschluss an die internationale Leistungsspitze umfassende Maßnahmen für eine Sicherung und Erhöhung des Bildungsertrages erforderlich sind.

Eine differenzierte Analyse der Bildungs- und Schulsysteme und somit auch der Erfolgsfaktoren jener Staaten, die in internationalen Vergleichsstudien wiederholt besonders gute Ergebnisse erzielt haben, konnte die folgenden Reformansätze und Steuerungsmodelle identifizieren (vgl. van Ackeren, 2003):

- eine eher spät einsetzenden Differenzierung in verschiedene Bildungsgänge,
- eine intensive individuelle Förderung der Schülerinnen und Schüler,
- die Etablierung nationaler Bildungsstandards,
- die regelmäßige, professionelle Durchführung von zentralen Vergleichsstudien und
- eine an die Ergebnisse dieser Evaluation geknüpfte Ressourcenzuweisung.

In den Jahren 2003 und 2004 hat die Kultusministerkonferenz unter anderem auf Grundlage dieser Erkenntnisse länderübergreifende Bildungsstandards für verschiedene Fächer und unterschiedliche Schulabschlüsse verabschiedet.

Im Sinne der von Klieme et al. (2003) vorgelegten Expertise wurden die Bildungsstandards als schulische bzw. fachliche Kompetenzen verstanden, die sich in gezeigtem Verhalten abbilden und daher als Can-Do-Statements („Die Schülerinnen und Schüler können ...“) formuliert werden. Dabei ist zu beachten, dass aus beobachtbarem Verhalten nicht unmittelbar und fehlerfrei auf den Grad der Ausprägung einer zugrundeliegenden Kompetenz geschlossen werden kann. Kompetenzen werden somit als nicht direkt beobachtbare Konstrukte verstanden, die sich mittels Indikatoren, bspw. Testaufgaben, operationalisieren lassen.

Die länderübergreifenden Bildungsstandards beschreiben fachbezogene Kompetenzen, die Schülerinnen und Schüler bis zu einem bestimmten Zeitpunkt ihrer Bildungsbiographie in der Regel erreicht haben sollen. Sie definieren verbindliche Zielvorgaben. Es werden also Leistungserwartungen im Sinne eines zu erreichenden Zieles formuliert, wobei der Weg zur Erreichung dieses Zieles nicht vorgegeben wird, sondern unter Berücksichtigung der spezifischen Unterrichtssituationen frei gewählt werden kann. Das Hauptaugenmerk liegt somit auf dem Erfolg, welchen der Unterricht in Hinblick auf die Förderung und Entwicklung der Schülerleistungen hat. Diese Form der Output-Orientierung ergänzt die bislang leitende Input-Orientierung des deutschen Bildungssystems, welche sich in erster Linie auf die Entwicklung didaktischer Modelle konzentrierte. Diese Grundprinzipien eines Faches werden in den Bildungsstandards aufgegriffen und um die Perspektive zentraler, langfristig aufgebauter Lernergebnisse im Sinne von Basisqualifikationen erweitert. Somit beschreiben Bildungsstandards erwünschte Lernergebnisse und zugrunde liegende Wissensbestände, welche die Schüler bis zu einem bestimmten Zeitpunkt erworben haben sollen. Dies beinhaltet, dass die Bildungsstandards nicht schulformbezogen formuliert sind, sondern einer abschlussbezogenen Konzeption folgen.

Zunächst beziehen sich die Formulierungen der Bildungsstandards auf die Definition eines mittleren Anforderungsniveaus. Im Zuge der Entwicklung von Kompetenzstufenmodellen wird konkretisiert, was unter dem jeweiligen Regelstandard zu verstehen ist, und somit verdeutlicht, welche Leistungserwartung an Schülerinnen und Schüler, die einen bestimmten Abschluss anstreben, gestellt werden sollte.

Die Übersetzung und Konkretisierung der Bildungsstandards in Aufgaben führt zur Gegenüberstellung von Unterrichts- und Testaufgaben, wobei eine Kontrastierung in dieser Absolutheit zumeist wenig sinnvoll ist. Im Sinne einer Betrachtung von Prototypen dienen erstere zur Unterstützung der Implementierung der Bildungsstandards und ihrer Verankerung im Unterrichtsalltag. Letztere ermöglichen in Form standardisierter Leistungstests die Messung und

somit die Überprüfbarkeit bereits erworbener Kompetenzen sowie der klassenbezogenen Ermittlung von Förderbedarf.

2.2 Charakterisierung der länderübergreifenden Bildungsstandards für den Mittleren Schulabschluss im Fach Deutsch

Zur Herstellung begrifflicher Klarheit soll zunächst eine zentrale Annahme vorangestellt werden. Die länderübergreifenden Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss, kurz die „Bildungsstandards“, bezeichnen im Folgenden die verbal dargelegten Inhalte des gleichnamigen Dokuments der Kultusministerkonferenz (Sekretariat der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland, 2004). Dieses Dokument umfasst inhaltlich-didaktische Beschreibungen der angezielten Kompetenzen, ist jedoch nicht im Sinne eines klar umgrenzten und präzise gefassten Leistungsstandards zu verstehen, an dessen Erreichung die Vergabe des Mittleren Schulabschlusses im Fach Deutsch geknüpft werden kann.

In inhaltlich-didaktischer Hinsicht greifen die Bildungsstandards im Fach Deutsch für den Mittleren Abschluss einen pragmatisch orientierten literacy-Begriff auf, welcher danach fragt, inwieweit die erworbenen Kompetenzen zu einem selbst bestimmten Leben und einer aktiven Teilhabe an der Gesellschaft beitragen können. Die Relevanz einer umfassenden Sprachkompetenz in produktiver und rezeptiver Form erschließt sich hierbei unmittelbar und erstreckt sich nicht nur auf den schulischen Wissenserwerb, sondern auch auf den beruflichen Erfolg und die soziale Integration. Nicht nur während der Schulzeit, sondern auch im späteren Berufsleben und in Bezug auf lebenslanges Lernen bedeuten bspw. unzureichende Lesekompetenzen eine entscheidende Einschränkung im Wissenserwerb und damit erhebliche Nachteile in Bezug auf die stetig wachsenden Anforderungen der heutigen Wissensgesellschaft.

Konkret beinhalten die 2004 von der Kultusministerkonferenz veröffentlichten und seit dem Schuljahr 2005/2006 bundesweit verbindlich geltenden Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss Beschreibungen angestrebter Kompetenzen in den Bereichen

- (1) Sprechen und Zuhören,
- (2) Schreiben,
- (3) Lesen – mit Texten und Medien umgehen sowie
- (4) Sprache und Sprachgebrauch untersuchen.

Neben der Definition von Kompetenzen und Standards werden drei Anforderungsbereiche ausgewiesen, die verschiedene Grade kognitiver Komplexität repräsentieren.

Anforderungsbereich I thematisiert die Verfügbarkeit der für die Bearbeitung der Aufgaben notwendigen inhaltlichen und methodischen Kenntnisse.

Anforderungsbereich II umfasst das selbstständige Erfassen, Einordnen, Strukturieren und Verarbeiten der aus der Thematik, dem Material und der Aufgabenstellung erwachsenden Fragen bzw. Probleme und deren entsprechende gedankliche und sprachliche Bearbeitung.

Im Anforderungsbereich III werden eine eigenständige Reflexion, Bewertung bzw. Beurteilung einer komplexen Problemstellung bzw. Thematik oder entsprechenden Materials und ggf. die Entwicklung eigener Lösungsansätze erwartet.

3. Testdiagnostische Grundlegung

3.1 Kennzeichnung eines standardisierten Leistungstests¹

Standardisierte Leistungstests müssen verschiedenen psychometrischen Anforderungen genügen, da sichergestellt werden soll, dass das jeweils interessierende Kompetenzkonstrukt mit Hilfe des entwickelten Tests bestmöglich operationalisiert wird. Mit anderen Worten muss der Test ein geeignetes Instrument darstellen, um aus den vorliegenden Antworten tatsächlich ein quantitatives Maß für den Grad der Kompetenzausprägung zu ermitteln und somit die fragliche Kompetenz messbar zu machen (Mislevy et al., 2002). Die hierfür relevanten psychometrischen Anforderungen können mit Hilfe der so genannten Hauptgütekriterien eines Tests (Lienert & Raatz, 1998) veranschaulicht werden.

Die *Objektivität* eines Tests beschreibt den Grad, in dem die Testergebnisse unabhängig von der Person sind, die den Test einsetzt und auswertet. Somit hängt die Objektivität insbesondere von einer objektiven Testdurchführung und einer objektiven Bewertung der Schülerantworten ab. Eine objektive Durchführung bedeutet, dass die Art der Durchführung von Testung zu Testung nicht variieren darf und somit die Rahmenbedingungen der Testung, bspw. das Verhalten des Testleiters während der Testung, für alle am Test teilnehmenden Schülerinnen und Schüler annähernd gleich sind. Weiterhin muss die Auswertung und Beurteilung der Schülerantworten in objektiver Weise erfolgen. Eine gegebene Schülerlösung muss nach immer gleich bleibenden Kriterien bewertet werden, unabhängig davon, wer die Schülerantwort beurteilt. Hierfür sind genaue Instruktionen erforderlich, wie die Antworten auf eine Aufgabe einzuschätzen sind. Die bei der Auswertung erreichbare Objektivität hängt in starkem Maße vom gewählten Aufgabenformat ab. Dies kann im Fall von geschlossenen Aufgabenformaten zu einer nahezu perfekten Auswertungsobjektivität

¹ Dieser Abschnitt ist in enger Anlehnung an den Beitrag von Granzer, Böhme & Köller (2008) entstanden.

führen, stellt aber bei offenen Aufgabenformaten, die freie Antworten der Schülerinnen und Schüler erfordern, mitunter ein Problem dar. Hier sind detaillierte Erklärungen und prototypische Antworten erforderlich, damit die Beurteiler die freie Schülerantwort nicht nach ihrem subjektivem Gefühl beurteilen, sondern hinreichend genau und objektiv arbeiten können. Beispielsweise muss erläutert werden, welche Antwort als richtig, welche als teilweise richtig und welche als falsch zu bewerten ist (Mullis et al., 2004). Eine Möglichkeit zur Überprüfung der Auswertungsobjektivität besteht darin, eine gegebene Schülerantwort von mehreren Beurteilern einschätzen zu lassen, um dann die Übereinstimmung zwischen diesen zu ermitteln (Wirtz & Casper, 2002). Nur wenn die Durchführungs- und Auswertungsobjektivität gesichert sind, können die Testleistungen der Schülerinnen und Schüler und damit ihre Kompetenz fair beurteilt werden.

Durch die *Reliabilität* (Zuverlässigkeit) wird das Ausmaß der Genauigkeit beschrieben, mit der ein Test die zu messende Kompetenz abzubilden vermag, zunächst unabhängig davon, ob es sich bei der gemessenen Kompetenz auch wirklich um die fragliche Kompetenz handelt. Die Reliabilität kann empirisch mit Hilfe statistischer Verfahren bestimmt werden, nachdem der Test einer Schülerstichprobe vorgelegt wurde. Quantifizierbar ist die Reliabilität mittels eines Reliabilitätskoeffizienten, welcher angibt, inwieweit das Testergebnis reproduzierbar ist.

Es sind verschiedene Faktoren bekannt, welche die Reliabilität beeinflussen. So ist es zum Beispiel hilfreich, in einem Test möglichst viele Aufgaben zu verwenden, die dieselbe Kompetenz erfassen. Werden nur wenige Aufgaben für die Erfassung ein und derselben Kompetenz verwendet, so ist die Messung im Allgemeinen weniger zuverlässig.

Die *Validität* (Gültigkeit) ist ein weiteres wesentliches Güte Merkmal eines Tests und beschreibt, inwieweit ein Test dasjenige Kompetenzkonstrukt misst, welches er messen soll bzw. zu messen vorgibt. Ein Test gilt dann als valide, wenn er wirklich das Merkmal erfasst, welches er erfassen soll, und nicht irgendein anderes. Ein gegebener Test kann sich durchaus als objektiv und reliabel erweisen und dennoch ein anderes als das ursprünglich intendierte Kompetenzkonstrukt messen, was dann ein Ausdruck mangelnder Validität ist.

Enthält bspw. eine Schreibaufgabe eine komplizierte und stark textlastige Instruktion, so kommt bei ihrer Bearbeitung nicht nur die Schreib-, sondern auch die Lesekompetenz der Schülerinnen und Schüler zum Tragen. Kinder mit geringerer Lesekompetenz verstehen möglicherweise die Instruktion falsch oder nicht vollständig und bewältigen daher die Schreibaufgabe schlechter, als es ihre tatsächliche Schreibkompetenz erlauben würde. Somit wird nicht nur die Schreibkompetenz erfasst, sondern zusätzlich auch die Lesekompetenz, obwohl die Lesekompetenz hier nicht Bestandteil der fraglichen Kompetenz und ihre Messung nicht intendiert ist.

Inhaltsvalidität ist dann gewährleistet, wenn der Test die Bandbreite verschiedener Anforderungen, die die angezielte Kompetenz beinhaltet, möglichst vollständig und entsprechend ihrer Relevanz abfragt.

Zwischen den Testgütekriterien Objektivität, Validität und Reliabilität bestehen wechselseitige Abhängigkeiten. Grundvoraussetzung für eine zufrieden stellende Reliabilität ist eine hinreichende Objektivität, denn nur ein objektiver Test kann auch reliabel sein. Weiterhin kann ein Test (kriterienbezogen) nicht valider sein, als er reliabel ist (Lienert & Raatz, 1998).

Zusammenfassend lässt sich somit festhalten, dass standardisierte Tests bestimmten Gütekriterien entsprechen müssen, damit sie eine zuverlässige Messung der intendierten Kompetenzen ermöglichen. Objektiv ist ein Test, wenn seine Ergebnisse unabhängig vom Untersucher sind. Reliabel ist er, wenn er das, was er erfasst, konsequent und zuverlässig erfasst. Valide ist ein Test, wenn er das misst, was er messen soll.

3.2 Prinzipien bildungsstandardorientierter Testentwicklung am IQB

Damit ein Test den soeben erläuterten Gütekriterien genügen kann, ist es von ausschlaggebender Bedeutung, dass das Ziel eines objektiven, reliablen und validen Tests bereits während der Aufgabenentwicklung stets präsent ist. Um dies gewährleisten zu können, sollten alle an der Aufgabengenerierung Beteiligten zunächst eingehend geschult und in die entsprechenden Konzepte eingeführt werden.

Da im Kontext groß angelegter Schulleistungsstudien zumeist erfahrene Lehrkräfte für die Entwicklung von Aufgaben gewonnen werden, ist darauf zu achten, die relevanten testdiagnostischen Hintergründe verständlich darzustellen und ausführliche Hinweise und Erläuterungen zur Testkonstruktion zur Verfügung zu stellen. Hierbei sollte u. a. darauf Wert gelegt werden, dass die Aufgabenentwickler bereits vor der Formulierung eines Items eine klare Entscheidung treffen, welche (Teil-)Kompetenz sie innerhalb welchen Anforderungsbereichs mit Hilfe dieses Items messbar machen wollen. Im gesamten Entwicklungsprozess sollte das primäre Ziel, eine vorab eindeutig definierte Kompetenz zu messen, stets präsent und handlungsleitend sein (vgl. AERA, APA & NCME, 1999; Osterlind, 1998). Alle Einfälle für Aufgaben, die Auswahl von Stimulusmaterial oder die Einbeziehung von Abbildungen müssen daher immer vor dem Hintergrund bewertet werden, ob sie dazu beitragen, dass das Item die intendierte Kompetenz misst. In gleicher Weise sollte stets berücksichtigt werden, welche nicht intendierten Kompetenzbereiche durch ein Item möglicherweise ebenfalls angesprochen werden könnten. Natürlich können bestimmte Kompetenzen oft nicht isoliert von anderen Kompetenzen überprüft werden. In diesem Falle ist es für die spätere Interpretation der Testergebnisse jedoch unerlässlich

zu wissen, welche Kompetenzen zusätzlich zu den primär überprüften für die Lösung des Items relevant sind. Auch die Auswahl des Itemformats sollte immer überlegt erfolgen. So bilden bspw. Aufgaben mit standardisierten Antwortalternativen wie Multiple-Choice-Items für Tests verschiedene Vorteile, da sie eine hohe Auswertungsobjektivität versprechen. Bezogen auf die Bildungsstandards im Fach Deutsch besteht die Herausforderung der Operationalisierung unter anderem darin, dass sie teils wenig präzise formuliert sind, so dass es schwierig ist, sie in Testaufgaben zu transformieren.

Die am IQB entwickelten Testaufgaben zur Überprüfung der Erreichung der Bildungsstandards dienen vorrangig der Diagnostik auf Schul- oder Klassenebene. Die Aufgaben sollen daher den Leistungsstand von Schülergruppen hinreichend objektiv, zuverlässig und valide erfassen. Individualdiagnostische Aussagen über einzelne Schülerinnen und Schüler werden derzeit nicht angestrebt, da es nicht ohne weiteres möglich ist, den hohen Anforderungen zu genügen, welche gute Tests für Zwecke der individuellen Leistungsdiagnostik erfüllen müssen. Will man genaue Aussagen über einzelne Schülerinnen und Schüler treffen, so müsste jeder Einzelne sehr viele Aufgaben bearbeiten, um eine hinreichend genaue Schätzung der Personenfähigkeit zu ermöglichen. Zuverlässige (reliable) Aussagen über die Fähigkeiten einer Testperson sind nicht möglich, wenn diese ihre Fähigkeiten im Test nur anhand von zwei oder drei Aufgaben zeigen konnte. Verfolgt man zusätzlich die Zielstellung, kompetenzbereichspezifische Aussagen zu ermöglichen, so müsste jeder Schüler und jede Schülerin viele Stunden lang getestet werden, um für jeden Kompetenzbereich ausreichend Aufgaben für eine sichere Schätzung vorlegen zu können. Im Zuge der Evaluierung der Erreichung der Bildungsstandards bearbeitet jede einzelne Testperson jedoch nur eine kleine Zahl von Aufgaben pro Kompetenzbereich, weshalb lediglich Aussagen über die Fähigkeiten von Schülergruppen, nicht jedoch für Individuen möglich sind.

Durch die Tests soll eine möglichst akkurate Erfassung der Kompetenzen der Schüler gewährleistet werden. Das bedeutet, dass die größte Anzahl an Items im mittleren Schwierigkeitsbereich angesiedelt sein sollte, da dies dem Bereich der mittleren Fähigkeiten entspricht, in dem auch die Kompetenzen der meisten Jugendlichen liegen. Gleichzeitig muss sichergestellt werden, dass auch Aussagen über die Gruppen der besonders schwachen und besonders starken Schülerinnen und Schüler getroffen werden können. Deshalb gehören zu einer umfassenden Beurteilung der Schülersamtheit auch Items, die selbst leistungsstarke Jugendliche herausfordern, sowie solche, die auch Schwächeren noch eine Chance geben.

Die diagnostische Information, welche für die Entscheidungsfindung herangezogen wird, kann sich auf eine Quelle beschränken, also univariat vorliegen, oder aus mehreren Informationsquellen gespeist werden und somit multivariat beschaffen sein. Um die Validität und damit auch die

Entscheidungssicherheit zu erhöhen, sollten stets mehrere Prädiktoren, also Informationen verschiedener Quellen herangezogen werden (vgl. Amelang & Zielinski, 2004).

Bezogen auf die bildungsstandardorientierte Testentwicklung wird daher das Ziel verfolgt, verschiedene Stimuli und variierende Aufgabentypen aus allen Kompetenzbereichen und über alle Anforderungsbereiche hinweg zu konstruieren.

Zusammenfassend wird am IQB darauf hin gearbeitet, die Bildungsstandards zunächst theoretisch zu präzisieren, um das jeweils zu untersuchende Konstrukt möglichst konkret definieren zu können und es anschließend mit Hilfe von Aufgaben bzw. Tests zu operationalisieren. Auf der Basis theoretisch fundierter und empirisch validierter Testinstrumente werden anschließend Kompetenzstufenmodelle mit den zugehörigen Niveaustufen entwickelt, um die von den Schülerinnen und Schülern erreichten Leistungsstände festhalten und Entwicklungsfortschritte illustrieren zu können.

Die bildungsstandardorientierten Aufgaben des IQB dienen auf Grundlage der Normierungsergebnisse dazu, ein Kriterium zu definieren, welches prospektiv als erwartbarer Leistungsstandard gelten kann. Es handelt sich also um kriteriumsorientiertes Testen. Anders als bei norm- bzw. bezugsgruppenorientierten Tests, bei denen die individuellen Ergebnisse relativ zu den Ergebnissen einer Referenzgruppe interpretiert werden, will man mit kriteriums- bzw. lernzielorientierten Tests ermitteln, ob und in welchem Ausmaß ein Lernender ein im Detail beschriebenes Kriterium erreicht hat (vgl. Ingenkamp, 2005).

Die bildungsstandardorientierte Testentwicklung am IQB orientiert sich an folgendem Prozessablauf (vgl. bspw. Humboldt Universität zu Berlin. IQB, 2007), der einerseits eine Ausrichtung an den Bildungsstandards darstellt und andererseits dem state of the art der Testentwicklung entspricht:

1. fachdidaktisch und lernpsychologisch fundierte Konkretisierung der Kompetenzen
2. auf diesen Vorüberlegungen basierende Erarbeitung von Richtlinien zur Konstruktion von Testaufgaben (Item- und Testspezifikationen)
3. Itementwicklung durch erfahrene Lehrkräfte
4. an empirischen Hinweisen in Abstimmung zwischen Fachdidaktik und Psychometrie stattfindende Optimierung der generierten Testaufgaben
5. empirische Erprobung der entwickelten Aufgaben in großen Schülerstichproben
6. Normierung der Aufgaben auf der Basis national repräsentativer Stichproben
7. Entwicklung empirisch fundierter Kompetenzstufenmodelle, auf deren Grundlage bestimmt werden kann, welche konkreten Kompetenzerwartungen an die Erreichung von

Mindest-, Regel- und Exzellenzstandards gestellt werden können und welcher prozentuale Anteil an Schülerinnen und Schülern den Leistungsanforderungen gerecht werden konnte. Die in den Schritten 6 und 7 angesprochenen Arbeiten sind für den Mittleren Schulabschluss im Fach Deutsch derzeit noch nicht beendet, allerdings liegen hierfür Erfahrungen aus anderen Projekten vor.

3.3 Fundierung diagnostischer Entscheidungen

Cizek (2005) betont, dass Schulleistungsdiagnostik, welche bildungsbiographische Konsequenzen nach sich zieht, dem unausweichlichen Erfordernis entspringt, Entscheidungen zu treffen. Auch die Vergabe des Mittleren Schulabschlusses ist eine solche kategoriale Entscheidung: Entweder die Schülerin oder der Schüler erhält den angestrebten Abschluss oder nicht. Diese Entscheidungsfindung sollte möglichst sorgsam und in einer Weise erfolgen, die auch tatsächlich all denjenigen Schülerinnen und Schülern den Abschluss zubilligt, welche die gestellten Anforderungen erfüllen können.

Dies legt die Vermutung nahe, dass tatsächlich ein Kriterium vorliegt, anhand dessen überprüft werden kann, ob die mit der Abschlussvergabe verbundenen Leistungserwartungen vorliegen oder nicht. Wie bereits oben ausgeführt ist dies u. E. derzeit jedoch nicht der Fall, da die Formulierungen der Bildungsstandards keinen präzisen, mindestens zu erfüllenden Standard vorgeben. Mit den Dokumenten der Kultusministerkonferenz liegen also keine quantitativen Aussagen in Hinblick auf die erwarteten Kompetenzen der Schülerinnen und Schüler vor, sondern eher qualitative Aussagen über das, was Absolventen des Mittleren Schulabschlusses im Regelfall können sollten.

Umso wichtiger ist es aber, sich tatsächlich an den inhaltlich-didaktischen Vorgaben der Bildungsstandards einerseits und an den Prinzipien der Testentwicklung andererseits zu orientieren, um trotz des Fehlens eines konkreten Leistungsmaßstabs einen länderübergreifend fairen und von den Leistungsanforderungen her vergleichbaren Maßstab für die Vergabe des Mittleren Schulabschlusses zu wählen.

4. Beschreibung der Prüfungsaufgaben

4.1 Berlin

Die schriftliche Prüfungsarbeit zum mittleren Schulabschluss 2007 im Fach Deutsch des Landes Berlin umfasst in der Terminologie der Bildungsstandards Aufgaben zur Überprüfung der Lese- und Schreibkompetenz sowie zur Diagnostik im Kompetenzbereich „Sprache und Sprachgebrauch untersuchen“. Hierbei liegt der Prüfungsschwerpunkt auf der Überprüfung der Lesekompetenz (48%), gefolgt von der Überprüfung der Schreibkompetenz (32%) und schließlich der Überprüfung des Sprachwissens (20%). Die Aufgaben zu allen drei Kompetenzbereichen sind in den thematischen Rahmen „Geruchsinn und Gerüche“ eingebettet.

Als Aufgabenstämme werden verschiedene kontinuierliche Informations- sowie literarische Texte ebenso wie ein diskontinuierlicher Text eingesetzt. Zu diesen Stimuli müssen insgesamt 66 Aufgaben bearbeitet werden, wobei diese mitunter in Teilaufgaben gegliedert sind und die freie Schreibaufgabe² von uns als nur eine zu bearbeitende Aufgabe gezählt wird. Das Spektrum der verwendeten Aufgabentypen reicht von geschlossenen Formaten, wie Multiple-Choice-Items und Zuordnungsaufgaben, über Kurzantworten bis hin zu einer umfangreichen Essay-Aufgabe. Als Hilfsmittel ist ein Wörterbuch zugelassen. Die verfügbare Bearbeitungszeit liegt bei 180 Minuten. Die Gewichtung einzelner Prüfungsleistungen ist für die Schüler bereits während der Bearbeitung der Aufgaben einsichtig.

Das Aufgabenmaterial und die Aufgabenstellungen gelten ebenso wie die Auswertungsanweisungen ohne Differenzierung nach besuchter Schulform für alle teilnehmenden Schülerinnen und Schüler.

Für die Bewertung der Prüfungsarbeit durch die jeweilige Lehrkraft liegen für jedes Item Hinweise zur Punktvergabe vor. Sämtliche Prüfungsarbeiten werden schulintern zweifach korrigiert (Neumann, 2006).

Für die Benotung der Prüfungsarbeit erfolgt nach ungewichteter Aufsummierung der vergebenen Punkte eine Umrechnung in Noten gemäß einer tabellarischen Vorlage.

4.2 Brandenburg

Die Prüfungen am Ende der Jahrgangsstufe 10 des Landes Brandenburg im Fach Deutsch umfassen in der Terminologie der Bildungsstandards Aufgaben zur Leistungsüberprüfung in den Kompetenzbereichen Lesen und Schreiben, wobei es sich bei der Schreibaufgabe um eine „komplexe Aufgabe mit offenen Teilaufgabenstellungen“ handelt, „deren Ergebnisse in einem

² Für die Beurteilung der Schreibaufgabe werden vier Teilkriterien (681-684) definiert.

zusammenhängenden Text dargestellt werden müssen“ (Landesinstitut für Schule und Medien Brandenburg, Schuljahr 2006/2007).

Da die Gesamtnote der Prüfung im Wesentlichen der Bewertung der Schreibaufgabe entspricht (siehe unten), liegt die Gewichtung auf der Bewertung der Schreibkompetenz bzw. der in Essayform demonstrierten Kompetenz des Leseverstehens der Schülerinnen und Schüler. Den Prüflingen werden Aufgaben aus vier Themenfeldern (Analyse und Interpretation eines fiktionalen Textes, Erörtern von Problemen mithilfe von Materialgrundlagen, Produktiver Umgang mit Texten: Veränderung des Erzählsystems eines epischen Textes, Analyse eines nicht fiktionalen Textes) zur Auswahl vorgelegt, von denen sie eine bearbeiten müssen. Es besteht eine Abstufung der Anforderungen in den Aufgabenstellungen je nach besuchter Schulform bzw. nach Kursniveau. Die Stimulustexte sind jedoch für alle Schulformen identisch. Jede Wahlaufgabe besteht aus fünf Multiple-Choice-Items zur Überprüfung der Lesekompetenz sowie einer freien Schreibaufgabe, für die nach Schulform differenzierte Bearbeitungshinweise gegeben werden. Entsprechend den differenzierten Aufgabenstellungen und Anforderungen liegen für die unterschiedlichen Schulformen bzw. für verschiedene Kursniveaus leicht variierende Auswertungshinweise vor.

Als Hilfsmittel ist ein Wörterbuch zugelassen. Die verfügbare Bearbeitungszeit liegt bei 180 Minuten.

Die Bewertung der Prüfungsarbeit erfolgt durch die jeweilige Lehrkraft, wobei für den geschlossenen Aufgabenteil ein Lösungsschlüssel und für die Beurteilung des Prüfungsaufsatzes ein Erwartungshorizont bereitgestellt werden. Die Note des Prüfungsaufsatzes setzt sich aus einer Bewertung der Aspekte Inhalt (50%), Sprachliches Können (25%) sowie Sprachrichtigkeit (25%) zusammen. Für jeweils vier zufällig ausgewählte Arbeiten pro Klasse wird eine Korrektur durch eine weitere Lehrkraft vorgenommen.

Die Kriterien der Bewertung der Prüfungsleistung sind für die Schülerinnen und Schüler während der Bearbeitung der Aufgaben nicht transparent.

Die Prüfungsnote entspricht der Note des Aufsatzteils, es sei denn, dass die Note für den geschlossenen Teil um mindestens drei Noten abweicht. Nur in diesem Fall ändert sich die Gesamtnote um eine Note nach oben oder nach unten (Landesinstitut für Schule und Medien Brandenburg, 2006).

5. Anmerkungen zu den Prüfungsaufgaben des Landes Berlin

5.1 Übergreifende Anmerkungen zu Items und Aufgaben

- Zur Anzahl der Auswahloptionen bei Multiple-Choice-Items (MC-Items):
Als Standardvorgehen werden heutzutage für MC-Items zumeist vier oder fünf Antwortoptionen entwickelt, von denen genau eine die richtige Antwort ist. Diese Verwendung von vier bzw. fünf Optionen hat sich bewährt und stellt den weit verbreiteten Normalfall dar. Zwar wird in der einschlägigen Literatur diskutiert, dass bereits drei Antwortoptionen eine hinreichende Messgenauigkeit gewährleisten können. Dennoch besteht Einigkeit, dass die Hinzunahme weiterer plausibler Distraktoren die Reliabilität erhöht. (vgl. Impara & Foster, 2006).
- Um die Lesezeit zu minimieren und unnötige Redundanzen zu vermeiden, sollten identische Eingangsformulierungen der Antwortoptionen (vgl. bspw. den identischen Beginn der Antwortoptionen in Item 109: „weil man mit“⁶) in den Stamm des MC-Items übernommen werden (vgl. Haladyna, 2004; Nico, 2004).
- Es sollte nach Möglichkeit vermieden werden, in MC-Items Antwortoptionen zu verwenden, die sich gegenseitig ausschließen (vgl. bspw. Item 109: Optionen (a) und (c); 202: Optionen (a) und (c)) (vgl. Nitko, 2004).
- Der Einsatz von Complex Multiple-Choice-Items (CMCs) sollte insgesamt überdacht werden, da in der einschlägigen Literatur seit Anfang der 90er Jahre von der Verwendung dieses Aufgabentyps aufgrund empirischer Befunde eher abgeraten wird. CMCs sind zwar zumeist schwerer als vergleichbare single-best-answer MCs, unterliegen aber in stärkerem Maße dem Einfluss von Testbearbeitungsstrategien, bspw. wenn ein Prüfling einzelne Optionen klar als richtig oder falsch erkennt und Distraktoren auf diese Weise ausschließen kann. Schwerwiegender ist jedoch der Befund, dass CMcs eine geringere Trennschärfe aufweisen, was die Reliabilität des gesamten Tests negativ beeinflusst (vgl. zusammenfassend Haladyna, 2004).
- Bei der Itemkonstruktion sollte darauf geachtet werden, dass die Items untereinander keine Abhängigkeiten aufweisen.

5.2 Anmerkungen zur Instruktion

- Es sollten keine Aufgabentypen instruiert werden, die nicht vorkommen. Dies betrifft hier die so genannten Richtig-Falsch-Abfragen.

- Alle auftretenden Aufgabenformate sollten aber instruiert werden. So fehlen bspw. Instruktionen für Kurzantworten (bspw. Aufgabe 104) sowie Zuordnungsaufgaben (bspw. Aufgabe 102).
- Complex Multiple-Choice-Items (bspw. 105) sollten separat instruiert werden, da aus der Instruktion nicht hervorgeht, dass es solche Aufgaben geben wird und wie diese zu bearbeiten sind. In der Instruktion ist von *der zutreffenden Lösung* die Rede, eigentlich geht es aber um zwei zutreffende Lösungen.
- Die Instruktion „Fragen beantworten Sie im ganzen Satz oder im verständlichen Stichpunkt.“ erscheint insofern unglücklich, als die meisten Arbeitsaufträge, die mit einer Kurzantwort zu beantworten sind, gar nicht als Frage formuliert werden, siehe bspw. Aufgabe 103, 106, 107 etc.
- Die Schreibaufgabe wird in ihrem Aufbau im Instruktionsteil erläutert. Dies ist ungewöhnlich, da Aufbau und Arbeitsaufträge der Lese- und Sprachgebrauchsaufgaben hier nicht erläutert werden. Dieser Punkt könnte entweder ersatzlos gestrichen oder der eigentlichen Schreibaufgabe vorangestellt werden.

5.3 Aufgaben- und itemspezifische Anmerkungen

Leseaufgaben zu Text 1 „Ein Riecher für den Richtigen“

102: Hier könnte das Layout optimiert werden (siehe Vorschlag). Wozu dient die in der Tabelle vorgenommene Nummerierung (a) bis (f), wenn sie gar nicht verwendet werden darf? Wenn diese Nummerierung ohnehin angeboten wird, könnten die Schülerinnen und Schüler doch gleich diese Buchstaben den Absatzüberschriften zuordnen, ohne alles abschreiben zu müssen. Das Eintragen der Absatzüberschriften in die unten stehende Tabelle kostet unnötig Testzeit und hat keinen eigenen diagnostischen Wert.

Optimierungsvorschlag:

102: Ordnen Sie die folgenden Überschriften den Absätzen (*siehe* Zeilenangaben) zu.

- (A) Bedeutung der Ergebnisse
- (B) Beispiele aus der Geschichte
- (C) Bezug zu älteren Versuchen
- (D) Ergebnisse des Versuchs
- (E) Forschungsvorhaben und Versuchsbeschreibung
- (F) Versuche aus der Tierwelt

Notieren Sie jeweils vor den Zeilenangaben den Buchstaben der passenden Überschrift.

_____	Zeile 1 bis 22
_____	Zeile 23 bis 26
_____	Zeile 27 bis 36
_____	Zeile 37 bis 43
_____	Zeile 44 bis 48
_____	Zeile 49 bis 53

- 104: Die Variablen (b) und (c), welche nach Beginn und Ende des Versuchs fragen, erschließen sich den Schülerinnen und Schüler evtl. nicht sofort, da die Wortwahl als gesuchte Information ein Datum oder eine Uhrzeit nahe legt. Möglicherweise könnte man hier nach dem „Tag des Versuchsbeginns“ bzw. dem „Tag des Versuchsendes“ fragen.
- 104: Die Formulierungen der Variablen (e) und (f) sind etwas unglücklich („Zahl der Teilnehmer (Frauen)“ bzw. („Zahl der Teilnehmer (Männer)“). Als alternative Formulierungen kämen hier bspw. „Anzahl der teilnehmenden Frauen“ sowie „Anzahl der teilnehmenden Männer“ oder „Anzahl weiblicher Teilnehmer“ bzw. „Anzahl männlicher Teilnehmer“ in Frage.
- 105: Ungünstig erscheint hier die Formulierung der beiden falschen Antwortoptionen, die beide nach dem Muster „über den ... zu informieren.“ gebildet wurden. Ein fiktiver Roman dient im Allgemeinen nicht der Information über Sachthemen. Dies ist Weltwissen, weshalb nicht aus dem Text geschlossen werden muss, dass diese Antwortoptionen nicht korrekt sein können.
- 109/110: Diese beiden Items sind voneinander nicht logisch unabhängig. Das in 109 als gegeben Formulierte ist zugleich eine der möglichen Lösungen von 110 (s. Lösungshinweise). Das ist sowohl testtheoretisch als auch fachdidaktisch ungünstig und sollte vermieden werden.
- 113: Hier könnte problematisch sein, dass der Satzanfang „Der Artikel informiert über...“ die Angabe des *Textthemas* nahe legt, während die Instruktion auf die Formulierung der *Kernaussage* zielt.
Bei diesem Item erscheinen die Lösungshinweise als nicht präzise genug. Würde formuliert „Es gibt einen Zusammenhang zwischen dem Körpergeruch und der Partnerwahl“, dann würde nur reproduziert, was in 109 als Teil der Instruktion bereits gegeben ist.
- In Bezug auf die übrigen halboffenen sowie offenen Items dieser Aufgabe kann die Punktvergabe im Allgemeinen als plausibel gelten: Ist die gesuchte Information der in der Aufgabenstellung gegebenen benachbart und/oder sind nur einfache textbasierte Schlüsse nötig, wird eine „1“ vergeben, in anderen, etwas komplexeren Fällen eine „2“.

Sprachgebrauchsaufgaben zu Text 1 „Ein Riecher für den Richtigen“

Hierzu ist generell Folgendes zu sagen:

In den Bildungsstandards (2004, 7) wird postuliert, der Bereich „Sprache und Sprachgebrauch untersuchen“ habe mit jedem der anderen drei Bereiche in Beziehung zu stehen. Gefordert ist demnach hier wie auch überwiegend im fachdidaktischen Diskurs ein „integratives“ Verständnis dieses Kompetenzbereichs.

Im Rahmen der Items 151 bis 159 wird versucht, diesem Ansatz Rechnung zu tragen. Alle Items lassen sich problemlos einschlägigen Standards zuordnen:

- 151 und 152: differenzierter Wortschatz
- 153: Wortart
- 154: Satzglied
- 155: genus verbi
- 156 und 157: Satzstrukturen
- 158 und 159: orthografische Regeln.

In dieser Hinsicht sind die Items also nicht zu beanstanden. In anderer Hinsicht ist aber ein zentrales Problem unübersehbar: Zur Erhellung der Struktur des Lesetextes oder auch zur Explikation von Aspekten des Leseverstehens tragen diese Items eigentlich nichts bei. Die Wahl der Exempel für grammatische Begriffe mutet willkürlich an. Der Text wird sozusagen als Steinbruch genutzt, es hätten auch beliebige andere Exempel herangezogen werden können. Es ist letztlich nicht einsichtig, inwiefern dieses Verfahren in didaktischer Hinsicht plausibler ist als die Nutzung von isoliert dargebotenen Beispielsätzen.

Wie das integrative Programm überzeugend eingelöst werden kann, ist derzeit allerdings auch noch alles andere als klar. Insofern wäre es unbillig, den Eindruck zu erwecken, die Berliner Prüfungsarbeit könnte in dieser Hinsicht anhand eines eindeutigen und etablierten Maßstabs beurteilt werden. Immerhin liegen einige „unterrichtstaugliche“, recht detailliert ausgearbeitete Konzepte vor (z.B. Einecke 1999).

Es wäre z.B. denkbar gewesen, den argumentativen Duktus des Textes „Ein Riecher für den Richtigen“ stärker zu fokussieren. Einige Beispiele:

- Mit welcher Konjunktion könnte der Satz in Zeile 27 beginnen (vgl. das Leseitem 108)?
- Wie kann man den Gebrauch des Adverbs „immer“ in Zeile 37 beurteilen? Inwiefern stützen die folgenden Beispiele die Behauptung, in der „immer“ enthalten ist?
- Man könnte auch nach alternativen Formulierungen fragen: Wie müsste man z.B. schreiben, wenn man im 3. Abschnitt nur mit dem Präteritum, ohne ein Plusquamperfekt auskommen wollte? (Hier geht es um das Verständnis des Ablaufs des Experiments.)

- Man könnte auch einen schwer verständlichen Satz wie den in Zeile 49 ff. in eine Abfolge von Hauptsätzen oder dergleichen transformieren lassen. Es müsste jedenfalls darum gehen, die Funktionalität der Sprachreflexionsaufgaben für das Leseverständnis zu plausibilisieren.

Leseaufgaben zu Text 2 „Der Duft der Verführung“

- 204: Hier kann man darüber streiten, ob nicht 2 Punkte angemessener sind. Das Beispiel ist ja nicht im Text zu lokalisieren, sondern selbständig zu konstruieren.
- 202/205: Bei diesen beiden Items liegt der begründete Verdacht nahe, dass sie logisch-semantic nicht unabhängig voneinander sind.
- 207: Das in Item 207 (b) als gegeben Unterstellte ist u. E. problematisch. Dass sich das Geruchsgedächtnis in ca. 50 Jahren kaum noch nutzen lässt, folgt nicht aus dem im Text explizit Gesagten. In den Zeilen 38f. wird konjunktivisch formuliert und in Zeile 39f. ist von noch immer nicht zu unterschätzenden Fähigkeiten des Geruchsgedächtnisses die Rede. Möglicherweise wird das Geruchsgedächtnis in 50 Jahren also durchaus noch nutzbar sein, die Geruchsbeispiele werden sich aber womöglich geändert haben (statt Gras usw. Diesel oder Pommes Frites).

Sprachgebrauchsaufgaben zu Text 2 „Der Duft der Verführung“

Sieht man von Item 251 ab, bei dem es um eine für das Verständnis des Textes, speziell die lokale Kohäsion, wichtige Stelle geht, treffen die zum Text 1 formulierten Einwände auch hier zu.

Leseaufgaben zum diskontinuierlichen Text 3 „Verführerische Düfte“

Zu den Items zur Grafik „Verführerische Düfte“ ist anzumerken, dass 302 und 303 nicht unabhängig voneinander sind.

Leseaufgaben zu Text 4 „Das Parfum“

Auch hier sind einige Items nicht unabhängig voneinander, so 405 und 408, aber wohl auch 407 und 412. Hinzu kommt, dass zumindest bei 409 die Punktvergabe als nicht einsichtig erscheint. Hier geht es ja nicht um eine textuelle (bzw. aktuelle), sondern um die lexikalische Bedeutung von „traumwandlerisch“, und die Distraktoren sind unplausibel. Insofern wäre wohl eine „1“ angemessen.

Sprachgebrauchsaufgaben zu Text 4 „Das Parfum“

Hier haben die Items 452 und 455 eher nicht funktional-integrativen Charakter. Man könnte sich vorstellen, dass auch schwierigere Items zu konstruieren sind, z.B.

- „Einmal heißt es, der Duft sei „unbeschreiblich“ (Z. 39), und doch wird er zweimal mit Hilfe desselben Adjektivpaars beschrieben. Wie lautet dieses Paar?“ (Zeilen 13 und 35)
- Denkbar wäre auch, lokalisieren zu lassen, womit der Geruch im vorletzten Absatz explizit verglichen wird (Z. 33f) und wie dieser Vergleich – zweiter Teil des Items - beurteilt wird.

Schreibaufgabe zu Text 5 „Filmkritik“

Diese Aufgaben, die man zu einem guten Teil auch dem Bereich „Sprachgebrauch und Sprachgebrauch untersuchen“ zuschlagen könnte, sind unserer Auffassung nach plausibel und standardkompatibel. Denkbar ist allerdings, dass es Deckeneffekte gibt, dass also in der Gruppe der leistungsstärksten Schülerinnen und Schüler nicht mehr hinreichend differenziert werden kann.

Schreibaufgabe 6

Die Instruktion ist u. E. nicht ganz eindeutig. Es sollen drei Argumente für oder gegen den Film ausgeführt werden. Dass als Pro-Argument für die gewählte Position auch die Widerlegung eines Contra-Arguments zählen kann, wird in der Instruktion nicht deutlich, ist aber – mit Recht – in den Hinweisen zur Lösung vermerkt. Was wiederum diese Hinweise angeht, so ist die Unterscheidung von These und Schlussfolgerung argumentationstheoretisch nicht haltbar. Die These *ist* die Konklusion bzw. die Schlussfolgerung. Dass sie sozusagen zweimal vorkommen soll, einmal eher zu Beginn, einmal eher am Schluss, im Rahmen einer Zusammenfassung etwa, gibt die Instruktion nicht her. Problematisch ist darüber hinaus die Vergabe eines Punktes für Einfallsreichtum auf der inhaltlichen Ebene. Wenn wir recht sehen, dann kann man die Instruktion so lesen, dass man sich *nur* auf die gelesenen Texte beziehen solle. (Es fehlt ja z.B. ein „auch“, oder ein „unter anderem“ oder dergleichen.) Dann wäre Einfallsreichtum auf der Inhaltsebene eigentlich ausgeschlossen.

5.4 Anmerkungen zur Bewertung der Prüfungsaufgaben

Die Anweisungen zur Bewertung von halboffenen und offenen Antworten erscheinen im Wesentlichen als plausibel. Es wäre jedoch sinnvoll, einen expliziten Hinweis zu geben, dass *inhaltlich äquivalente* Antworten auch äquivalent zu handhaben sind.

5.5 Anmerkungen zu den in der Handreichung verwendeten Kompetenzmodellen

Anmerkung zur Klassifikation der Leseleistung

Die Berufung auf das PISA-Modell ist einsichtig, auch wenn es sich nur um PISA 2000 (und nicht PISA 2003) und noch dazu um eine kreative Bezugnahme handelt, insofern nicht fünf, sondern

nur drei „Stufen“ unterschieden werden. Verknüpft man das auf S. 8 der Handreichung formulierte Modell mit dem Aufgabenset, dann wird u. E. allerdings Folgendes deutlich:

- Leseaufgaben in der Dimension „Informationen ermitteln“ dominieren eindeutig.
- Aufgaben auf Level 3 dieser Dimension sind – vorsichtig formuliert – kaum vertreten.
- Wenn es überhaupt zu textbezogenem Interpretieren kommt, dann sind z.B. Aufgaben kaum vertreten, bei denen es um das text- und vor allem vorwissengesteuerte Verknüpfen von über den Text verstreuten Informationen geht.
- Leseaufgaben zum Reflektieren und Bewerten (sensu PISA bzw. Anforderungsbereich III) kommen nach unserem Eindruck gar nicht vor.

Tendenzen der damit angedeuteten Art sind auch in anderen Bundesländern und auch im Rahmen der Arbeit des IQB zu beobachten. Sie dürften u. a. darauf zurückzuführen sein, dass man in Kenntnis der Schwierigkeiten, freie Antworten in Textform objektiv auszuwerten, solchen Formaten gegenüber skeptisch geworden ist. Wenn es z.B. um Reflektieren und Bewerten geht, sind solche Formate aber weitgehend unersetzbar.

Anmerkung zur Bewertung der Leistung im Bereich „Sprache und Sprachgebrauch untersuchen“

Die Ausführungen zu „Stufen“ des *Sprachwissens und Sprachbewusstseins* sind u. E. (allzu) spärlich. Hier wird allein auf der Basis der Verben „kennen“, „anwenden“ und „beurteilen“ unterschieden. Differenzierungen wie die zwischen prozeduralem und deklarativem Wissen bzw. knowing how und knowing that spielen gar keine Rolle, obwohl sie im fachdidaktischen Diskurs seit längerem erörtert werden.

Anmerkung zu den „Kompetenzrastern“ für die Schreibaufgabe

Das auf S. 7 der Handreichung der Senatsverwaltung für Bildung, Wissenschaft und Forschung mitgeteilte Raster zur Beurteilung von *Schreibprodukten* ist eine leicht modifizierte Version eines in NRW zunächst im Kontext der Lernstandserhebung 9 im Jahr 2005 entwickelten Modells, das seinerseits in wesentlichen Teilen auf dem Züricher Textanalyseraster (Nußbaumer, 1991) beruht. Insofern der männliche Autor dieses Gutachtens an der Entwicklung des Rasters für NRW beteiligt war, sind ihm auch erhebliche Probleme im Hinblick auf die Auswertungsobjektivität bzw. die Beurteilerübereinstimmung und -reliabilität bekannt. Alternative, *empirisch getestete* Instrumente standen nach unserer Kenntnis in Deutschland Ende 2006, Anfang 2007, d.h. in dem Zeitraum, in dem die Berliner Aufgaben konstruiert wurden, allerdings gar nicht zur Verfügung. (Inwiefern man die erst 2007 publizierte Arbeit von Harsch u. a., die als Teil der „DESI-Studie“ erschienen ist, für die Konstruktion eines Modells der Dimensionen und Niveaus der Schreibkompetenz hätte nutzen können, braucht hier nicht erörtert zu werden.)

Das Modell der *prozessbezogenen Schreibkompetenzen* ist nach unserem Eindruck im hier gegebenen Kontext eigentlich funktionslos. Es wird ja z.B. nicht erhoben, inwieweit ein Schüler in der Lage ist, Schreibideen zu generieren und die Planungsideen beim Formulieren zu berücksichtigen. Dass dies *nicht* getestet wird, ist aus unserer Sicht übrigens sehr zu begrüßen. Denn Schreiben lässt sich – jedenfalls in vielen Fällen – als ein *rekursiver* Problemlöseprozess begreifen. Dass jemand Ideen, die er zunächst generiert hat, dann nicht ausformuliert, mag demnach daraus resultieren, dass er sie im Laufe des Schreibprozesses verwirft. Daraus zu schließen, dass er weniger kompetent ist als jemand, der die Einträge in einer mind map durchgängig abarbeitet, ist irrig. Hinzu kommt, dass in dem Modell keine Niveaus unterschieden werden. Insofern trägt es allenfalls – was aber auch nicht wenig sein mag – dazu bei, für die Lehrpersonen vor Ort die Wahl einiger Items zu plausibilisieren.

6. Anmerkungen zu den Prüfungsaufgaben des Landes Brandenburg

6.1 Übergreifende Anmerkungen zu Items und Aufgaben

- Im Prinzip positiv hervorzuheben ist das Bemühen, den Schülerinnen und Schülern eine Wahlmöglichkeit einzuräumen und verschiedene Textsorten (literarische Texte, diskontinuierliche Texte sowie kontinuierliche Informationstexte) und Aufgabenstellungen (Analyse und Interpretation fiktionaler Texte, produktiver Umgang mit Texten, Erörtern von Problemen sowie die Analyse nichtfiktionaler Texte) zu berücksichtigen. Allerdings ist zu bedenken, dass hierdurch ein großer Anteil der verfügbaren Testzeit gebunden ist.
- Die Schülerinnen und Schüler erhalten keine allgemeinen Instruktionen zur Aufgabenbearbeitung, bspw. wie viel Zeit auf die Entscheidung für eine Aufgabe verwendet werden sollte. Dies ist u. E. ungünstig.
- Dass in den Prüfungsaufgaben fast ausschließlich bestimmte Aspekte der Schreibkompetenz fokussiert werden, halten wir angesichts der Breite der in den Standards genannten Kompetenzen für problematisch.
- Die Aufgaben zur Kompetenzfeststellung müssen einen validen und reliablen Weg der Diagnostik darstellen. Dies beinhaltet bspw., dass zweifelsfrei bestimmt werden kann, welche (Teil-)Kompetenzen Gegenstand der Diagnostik sind, dies scheint im Falle der Gestaltung der Prüfungsaufgaben des Landes Brandenburg mitunter schwierig.
- Hinzu kommt, dass die MC-Items zur Überprüfung des Leseverständnisses insgesamt als zu leicht erscheinen, sodass es im oberen Teil des Leistungsspektrums zu Deckeneffekten kommen könnte. Insbesondere die Dimension des Reflektierens und Bewertens kommt nicht zur Sprache.

- Optimierungsbedürftig erscheint auch die Plausibilität der MC-Items.
In der aktuellen Literatur (vgl. Nitko, 2004) wird diskutiert, dass die Distraktoren (also die falschen Antwortoptionen) für Schülerinnen und Schüler, welche die richtige Antwort nicht kennen, plausibel sein sollten. Dies trifft für zumindest einige Items nicht zu.
- Weiterhin wird in der Literatur empfohlen, für die korrekte Antwortoption weder optische noch sprachliche Hinweisreize zu setzen (vgl. Haladyna, 2004; Nitko, 2004). Dies geschieht hier aber bspw., insofern die verschiedenen Optionen unterschiedlich lang sind. Testteilnehmer tendieren dazu, deutlich längere Optionen zu wählen, da sie annehmen, hier sei das Richtige präziser formuliert. Dies trifft auf die hier vorliegenden MC-Items auch verschiedentlich zu und sollte künftig vermieden werden. Beispiel: Wahlaufgabe 4, Item 1.1: „Was ist die Späthstraße heute?“ a) eine Autobahn, b) eine Sachgasse c) eine viel befahrene Ost-West-Achse.
- Unnötige Redundanzen in den Antwortoptionen der MC-Items können vermieden werden, indem identische Eingangsformulierungen in den Stamm des MC-Items übernommen werden (vgl. bspw. den identischen Beginn der Antwortoptionen in Wahlaufgabe 1, Item 1.3 „Er denkt über ...“ (vgl. Haladyna, 2004; Nico, 2004).
- Wie in Berlin sind bei MC-Aufgaben jeweils nur drei Optionen vorgesehen, was dem in der Methodenliteratur in der Regel empfohlenen Standard (4 oder auch 5 Optionen) nicht entspricht (siehe oben).
- Auch hier sollte nach Möglichkeit vermieden werden in MC-Items Antwortoptionen zu verwenden, die sich gegenseitig ausschließen (vgl. bspw. Wahlaufgabe 2, Item 1.5) (vgl. Nitko, 2004).
- Üblicherweise werden Textstimuli den Items vorangestellt, dies sollte auch hier berücksichtigt werden.
- Die verständnisprüfenden Multiple-Choice-Items haben u. E. zu wenig Gewicht, da die hier erbrachte Leistung nur dann in die Gesamtnote einfließt, wenn die Bewertung dieses Prüfungsteils um drei oder mehr Noten von der Bewertung der komplexen Aufgabe abweicht.
- Es ist nicht unmittelbar einsichtig, warum eine Abstufung nach verschiedenen Bildungsgängen intendiert ist. Die Bildungsstandards für den mittleren Schulabschluss sind explizit abschlussbezogen und *nicht* schulformbezogen formuliert.

6.2 Aufgaben- und itemspezifische Anmerkungen

Zu Wahlaufgabe 1 „Analyse und Interpretation fiktionaler Texte“

Zu den MC-Aufgaben

- Bei 1.1 ist eine Information zu lokalisieren und es ist ein einfacher Schluss („wenn Fünf nach eins, dann mittags“) zu ziehen.
- Bei 1.2 fällt der Schluss komplexer aus. Die Informationen, die hier zu verknüpfen sind, sind allerdings benachbart.
- Bei 1.3 braucht man für eine Lösung eigentlich nur das Signalwort „Punkband“ (vs. „Popband“ und „Rockband“).
- Bei 1.4 sind mehrere Detailinformationen, die sich in einem Abschnitt befinden, zu integrieren.
- Will man 1.5 lösen, muss man Informationen in zwei benachbarten Sätzen verknüpfen.

Komplexere Schlüsse sind also kaum verlangt. Es wäre z.B. denkbar gewesen, nach der Referenz des Pronomens „sie“ zu fragen. Auf welche Personen wird hier wahrscheinlich Bezug genommen? Ist ein solcher Bezug auf Figuren zu Beginn eines Textes üblich? Man hätte auch fragen können, ob es im Text Indizien dafür gibt, warum sich Achim zum Aussteigen genötigt sieht (er glaubt sich womöglich verplant, man sieht „immer dasselbe“).

Zu der Schreibaufgabe für die Gesamtschule (Grundkurs)

Hier interessiert vor allem die Frage, ob die Aufgabenstellung eine hinreichende Beurteilerübereinstimmung wahrscheinlich macht.

Die Frage nach den sprachlichen Mitteln bezieht sich primär auf die Zeilen 1 bis 8. Die Instruktion ist aber wohl so zu verstehen, dass auch andere Textteile berücksichtigt werden sollen („insbesondere“).

Im „Erwartungshorizont“ werden genannt Aufzählung bzw. Wiederholung, Verben („robbte“ allerdings in Z. 11) und negierende Attribute. In der Terminologie der Rhetorik könnte man von Wiederholungsfiguren, von Klimax und Oxymora sprechen. Auffällig ist hier aber auch Syntaktisches. Man könnte auf unvollständige Sätze verweisen (Ellipsen). Diese und weitere, in den Folgezeilen laut Instruktion ja auch noch aufzuspürende „sprachliche Mittel“ werden nicht weiter namhaft gemacht.

Im Hinblick auf die Teilaufgabe, die „Handlungen“ zu erfassen, dürfte es nicht zu gravierenden Diskrepanzen zwischen Raterurteilen kommen.

Anders verhält es sich mit der Interpretationsaufgabe. Hier handelt es sich um die schwierigste Teilaufgabe. Die im „Erwartungshorizont“ angedeuteten Versionen sind unterschiedlich plausibel.

Dass Achim mit sich unzufrieden und auf der Suche nach seiner Identität ist, dürfte zutreffen, ist aber als Motiv für das Zerschlagen des Spiegels ersichtlich zu unspezifisch. Auch das Stichwort „Identifikationsmöglichkeit“ ist im hier gegebenen Kontext u. E. verwirrend. Würde sich Achim mit dem bunten Spiegelbild identifizieren, wäre die Zerschlagung unplausibel. Will er – zweite Lesart – zu seiner realen „Farblosigkeit“ stehen und sich damit identifizieren, dann ist das gemeint, was jetzt als viertes Motiv angegeben ist. [„Achim möchte das (eigentlich gewünschte, imaginierte) ‚Bunte‘ zerstören und zur realen ‚Blässe‘ stehen.“]

Die Formulierung, dass weitere Varianten denkbar sind (Erwartungshorizont, S. 2), ist sicherlich zutreffend, offenbart aber auch exemplarisch ein Grundproblem im Umgang mit literarischen Texten: Wie will man das Spektrum der Deutungsmöglichkeiten plausibel begrenzen?

Zu der Schreibaufgabe für die Gesamtschule (Erweiterungskurs)

Die Teilaufgaben zum ersten und zum letzten Spiegelstrich sind mit denen für den Grundkurs identisch. Auffällig ist, dass im „Erwartungshorizont“ nun zum ersten Spiegelstrich (Analyse der Situation der Hauptfigur) u. a. Aspekte formuliert werden, die im Grundkurskontext den erzählerischen Mitteln zugeschlagen wurden. Was zum dritten Spiegelstrich ausgeführt wird, ist wiederum weitgehend anders geartet als das, was unter dieser Überschrift zu den Grundkursaufgaben gesagt wird. Jetzt geht es z.B. um die Erzählperspektive, um das Figurenensemble usw. Wir haben insofern den Eindruck, dass der komplexe (dazu noch singularische!) Ausdruck „Funktion erzählerischer Mittel“ nicht konsistent gebraucht wird.

Zu der Schreibaufgabe für die Realschule

Der Vergleich mit der Schreibaufgabe für den Grundkurs der Gesamtschule ergibt, dass nur eine Teilaufgabe partiell komplexer ist. Nun heißt es „Untersuchen Sie, wie Achim auf diese Situation reagiert.“ Früher war gefragt „Welche verschiedenen Handlungen vollzieht Achim am Spiegel?“ Die Differenz, um die es hier geht, ist sehr gering.

Zu der Schreibaufgabe für das Gymnasium

Hier wird ein „Erwartungshorizont“ aufgespannt, der dem für den Erweiterungskurs in der Gesamtschule analog ist. Was hier im Einzelnen zur symbolischen Funktion des Spiegels als Medium der Auseinandersetzung des Protagonisten mit sich selbst ausgeführt wird, ist u. E. allerdings partiell zu apodiktisch formuliert. Dass z.B. das Zerschlagen des Spiegels „das Zerstören der Bilder von sich und damit den Schritt zu sich selbst (Neubeginn)“ symbolisiert, ist ja keineswegs ausgemacht. (Diese Behauptung widerspricht im Übrigen auch der sonst in den „Horizonten“ mit Recht vertretenen Ansicht, es seien mehrere Lesarten plausibel.)

Zu Wahlaufgabe 2 „Produktiver Umgang mit Texten“

Im „Erwartungshorizont“, der sich auf die Schreibaufgabe für den Grundkurs bezieht, wird nicht nur in plausibler Weise erläutert, welche Elemente in inhaltlicher und formaler Hinsicht erwartet werden, sondern es wird auch auf Aspekte wie Eigenständigkeit, Originalität und ästhetische Absicht abgehoben. Wieder ist zu fragen, in welchem Maß die Beurteiler in diesen Hinsichten übereinstimmen. (Diese Frage bezieht sich auch auf die Schreibaufgaben für die anderen Schulformen.)

Bei der Schreibaufgabe für den Erweiterungskurs halten wir das Wort „beispielsweise“ für irritierend. Hier geht es ja nicht um ein Beispiel für „Gut-Sein“, dem andere an die Seite zu stellen wären.

Die Aufgabe für die Realschule dürfte wie intendiert strukturell einfacher sein als die für den Erweiterungskurs, die für das Gymnasium wiederum ist vermutlich schwieriger als die für die Realschule. (Ob die Exemplifizierung des Fachbegriffs „personales Erzählverhalten“ hilfreich ist, kann im „Lehnstuhl“ nicht beurteilt werden.)

Zu Wahlaufgabe 3 „Erörtern von Problemen mithilfe von Materialgrundlagen“

Für die Beantwortung der 1. MC-Aufgabe müssen Informationen verknüpft werden, die sich in zwei aufeinander folgenden Abschnitten eines Dokuments finden lassen. Bei der 2. MC-Aufgabe muss nur eine Information lokalisiert werden, wobei Formulierungen in der Aufgabe und im Text partiell identisch sind. Bei 3. sind Informationen aus mehreren Dokumenten zu verbinden. Sie sind allerdings jeweils an prominenter Stelle lokalisiert. Bei 4. müssen benachbarte Informationen zusammengeführt werden. Bei 5. schließlich ist wieder nur ein kurzes Dokument im Spiel, wobei überdies die Distraktoren recht unplausibel sind. Diese MC-Aufgabe halten wir darüber hinaus nicht für glücklich formuliert. Streng genommen müsste es heißen: „Was bewegt die meisten Sch. nach Auffassung von Frau Hungerland ...?“ Wieder werden also u. E. komplexere Schlüsse, Interpretationen und Reflexionen/Bewertungen im Kontext der MC-Aufgaben nicht verlangt. Die Schreibaufgabe für Sch. aus Grundkursen der Gesamtschulen ist als „klassische“ Ja/Nein-Quaestio gestellt. Präziser müsste die Aufgabe wohl lauten, ob sie in ihrer freien Zeit jobben *können* sollten. Dass auf Informationen aus mindestens zwei Materialien zurückgegriffen werden soll, dürfte nicht präzise genug sein. Material M3 müsste doch in jedem Fall bedacht werden, soll nicht im rechtlich „luftleeren“ Raum argumentiert werden.

Die Quaestio ist in der Schreibaufgabe für den Erweiterungskurs so formuliert wie für den Grundkurs. Hier halten wir die Aufforderung, wesentliche Argumente (noch einmal) *zusammenzufassen*, für problematisch. Wird hier nicht nur Redundanz gefordert? Eigentlich geht es doch darum, nach der Gegenüberstellung von Pro und Contra Urteile zur *Relevanz* von Argumenten abzugeben und dann zu einer Konklusion zu kommen.

Vergleicht man die Schreibaufgabe für Realschüler insbesondere mit der für Schülerinnen und Schüler aus Grundkursen, so erscheint das Bemühen um eine Differenzierung zwischen den Schulformen doch als etwas krampfhaft. Die Differenz ist ja nur quantitativer Natur. Es ist unwahrscheinlich, dass auf diese Weise (per definitionem qualitativ) unterschiedliche Niveaus der Schreibkompetenz erfasst werden können. Derselbe Einwand lässt sich im Kontext des Vergleichs von Realschul- und gymnasialen Aufgaben formulieren. Gymnasiasten haben, so könnte man resümieren, alle Kontra-Argumente zu bedenken, Realschüler sollen nur eines dieser Argumente entkräften.

Zu Wahlaufgabe 4 „Analyse nicht fiktionaler Texte“

Die MC-Aufgaben werden hier nicht mehr im Einzelnen kommentiert. Bei Item 1.5 ist a) zwar deutlich plausibler als die beiden Alternativen, u. E. aber auch nicht gänzlich zutreffend. (Man könnte das Wasserschutzgebiet ins Feld führen, vor allem aber Z. 66. Es geht wohl um ein nicht näher charakterisiertes, aber eigentlich leicht nachvollziehbares Gefühl der Zugehörigkeit.) Vergleicht man die Schreibaufgaben, dann springt zunächst die qualitative Differenz zwischen der Grund- und der Erweiterungskursaufgabe ins Auge. Schülerinnen und Schüler der Erweiterungskurse sollen – anders als die Grundkursschülerinnen und -schüler – den Symbolgehalt der Weide erschließen. Sie sollen aber auch untersuchen, „was am Beispiel von Herrn Schulz veranschaulicht wird.“ Diese Instruktion, die auf ein Spezifikum der Textsorte „Reportage“ zielt, halten wir für problematisch. Sie könnte u. E. sehr viel präziser formuliert werden. Hinzu kommt, dass die Frage, was veranschaulicht wird, eigentlich schon im Rahmen der ersten Teilaufgabe beantwortet werden sollte. Eine alternative Formulierung könnte insofern so (oder ähnlich) lauten: „Warum hat der Autor für die Darstellung der Veränderungen gerade die Perspektive eines Bürgers wie Schulz gewählt?“ Der Vergleich der Schreibaufgaben für die Grundkursschülerinnen und -schüler und die Realschülerinnen und -schüler wiederum ergibt, dass jenseits von Formulierungsunterschieden *inhaltlich* eigentlich überhaupt keine Differenzen bestehen. (Hier läuft die Bemühung, Differenzen zu „setzen“, sozusagen ins Leere.) Gymnasialschülerinnen und -schüler haben zusätzlich sprachliche Mittel der Veranschaulichung zu exemplifizieren und darüber hinaus die Mehrdeutigkeit des Titels der Reportage zu erläutern. Dass hier die Weide eine Rolle spielt, ist einsichtig. Für problematisch halten wir aber die Rede vom „Beispiel von Herrn (und Frau) Schulz als Veranschaulichung der Entwicklung im Neuköllner Wohngebiet [...]“. (Erwartungshorizont, 9) Herr Schulz ist keine Veranschaulichung, es geht, wie gesagt, darum, welche Funktion die Wahl seiner Perspektive hat.

6.3 Anmerkungen zur Bewertung der Prüfungsaufgaben

Für die Bewertung der Prüfungsaufsätze liegen nach Bildungsgang differenzierte Erwartungshorizonte vor. Es wird darauf hingewiesen, dass der Prüfungsaufsatz eine komplexe Leistung darstellt und deshalb als Ganzes zu bewerten ist, wobei sich „die Note [...] aus dem Vergleich zwischen den möglichen erwarteten Leistungen [...] und der tatsächlich erbrachten Leistung auf der Grundlage der Aufgabenstellung und der verbindlichen Vorgaben“ ergibt. „Alternative Lösungen, die der Aufgabenstellung entsprechen“, sollen als gleichwertig aufgefasst werden (Landesinstitut für Schule und Medien Brandenburg, Schuljahr 2006/2007).

Die Erwartungshorizonte scheinen trotz ihrer Ausführlichkeit an manchen Stellen recht vage und von daher in ihrer Umsetzung subjektiv gefärbt. Wie sich bspw. eine „angemessene und treffende Wortwahl“ im Einzelnen gestaltet oder wie ein Schüler dem Ziel der „Originalität in der Gestaltung“ gerecht werden kann, könnte anhand von Beispielen aus Schülertexten erläutert werden. Außerdem beinhalten die Erwartungshorizonte Aspekte, die über die Bildungsgänge hinweg nicht konsistent gehandhabt werden (bspw. „erzählerische Mittel“ im Bereich Inhalt). Inwieweit auf dieser Grundlage bei der Bewertung der Prüfungsaufsätze eine von der Person der Lehrkraft unabhängige, also objektive Beurteilung gewährleistet ist, bleibt offen. Es werden jedoch pro Klasse vier zufällig ausgewählte Prüfungsaufsätze von einer weiteren Lehrkraft korrigiert. Ob und, falls ja, inwieweit Abweichungen in der Benotung zu Konsequenzen in der Beurteilung der Prüfungsleistung führen, ist jedoch unklar.

7. Vergleich der Prüfungsaufgaben in Berlin und Brandenburg unter dem Gesichtspunkt der Orientierung an den Bildungsstandards

In beiden Bundesländern gibt es keine Aufgaben zum Kompetenzbereich Sprechen und Zuhören, was aus testökonomischen Gründen und Beschränkungen der technischen Machbarkeit nachvollziehbar ist.

Was den Kompetenzbereich Lesen angeht, so sind die Items des Landes Berlin aus unserer Sicht überzeugender, u. a. weil die Aufgabenformate vielfältiger sind und auch deutlicher zwischen verschiedenen Schwierigkeitsniveaus differenziert werden kann.

In Brandenburg liegt das Hauptaugenmerk auf dem Kompetenzbereich Schreiben, wobei wir im Hinblick auf die Auswertungsobjektivität größere Probleme erwarten. Hinzu kommt, dass die Differenzierung der Schreibaufgaben auf der Basis von Schulformen zum Teil als eher willkürlich und konstruiert erscheint. Im Vergleich hierzu wirkt der Stellenwert der Schreibaufgabe im Land

Berlin bei einiger Kritik im Einzelnen angemessener, zumal hier auch die Überarbeitungskompetenz als wesentliche Komponente im Schreibprozess teilweise erfasst wird. In Berlin lassen sich einige geschlossene und halboffene Items der Orthographie zuordnen, in Brandenburg nicht. Hier kommt die Orthographie nur als integraler Bestandteil der Beurteilung des Prüfungsaufsatzes zum Tragen.

Die in den länderübergreifenden Bildungsstandards postulierte Verknüpfung von Aufgaben im Kompetenzbereich „Sprache und Sprachgebrauch untersuchen“ mit den anderen Kompetenzbereichen wird u. E. in Berlin allenfalls ansatzweise geleistet. Die Zuordnungen der Items zu den Lesetexten erscheinen als eher willkürlich. Die einschlägigen Aufgaben tragen nicht zur Erhellung der Struktur der Lesetexte oder zur Explikation von Verstehensproblemen bei. In Brandenburg firmieren einschlägige Aufgaben im Wesentlichen unter dem Etikett „erzählerische Mittel“. Dieses Etikett wird aber u. E. in den verschiedenen, schulformspezifischen Auswertungshinweisen nicht konsistent verwendet.

In Hinblick auf die Verteilung der Aufgaben über die drei Anforderungsbereiche hinweg kann festgehalten werden, dass in beiden Bundesländern geeignete Items, die höhere kognitive Anforderungen im Bereich des Reflektierens und Bewertens messen, weitgehend fehlen, wobei das Bemühen um derartige Items in den Prüfungsaufgaben des Landes Berlin deutlicher zu erkennen ist.

8. Vergleich der Prüfungsaufgaben mit Testaufgaben, welche zur Evaluierung der Bildungsstandards eingesetzt werden

Ob Verbindungen zwischen verschiedenen Tests hergestellt werden können, hängt davon ab, inwieweit die zu vergleichenden Tests bestimmte Merkmale teilen. Um die Ähnlichkeit verschiedener Tests zu quantifizieren, benennen Kolen & Brennan (2004) als Mindestanforderung für die Vergleichbarkeit vier Kriterien, für welche eine Übereinstimmung geprüft werden muss. Dies betrifft zunächst die *Schlussfolgerungen*, welche aus den Ergebnissen der Tests gezogen werden sollen. Welche Art von Schlüssen lässt sich aufgrund der Testleistungen ziehen? Welche Messintention liegt den Tests zugrunde? Weiterhin stellt sich die Frage nach den operationalisierten *Konstrukten*: Zu welchem Grad messen die Tests tatsächlich ein und dasselbe Konstrukt? Ebenfalls relevant ist die Frage, für welche *Zielpopulation* die Tests konstruiert wurden. Abschließend thematisieren Kolen & Brennan (2004) die *Merkmale und Umstände der Messung*. Entscheidend sind hier bspw. vergleichbare Testlänge, Aufgabenformate sowie Durchführungsbedingungen. Nur wenn sich die zu vergleichenden Tests in Hinblick auf alle vier Kriterien hinreichend ähnlich sind, ist eine Gegenüberstellung der Tests überhaupt sinnvoll und nutzbringend.

Sehr schnell erschließt sich aufgrund dieser Kriterien, dass eine Vergleichbarkeit zwischen den hier betrachteten Prüfungsaufgaben und den für die Überprüfung der Bildungsstandards entwickelten Tests nur eingeschränkt möglich ist, da deutliche Unterschiede bestehen:

- *Schlussfolgerungen:* Die Tests für die Evaluierung der Bildungsstandards sollen eine Monitoringfunktion erfüllen und sind auf eine Diagnostik ausgelegt, welche auf der Ebene von Klassen und Schulen hinreichend genaue Aussagen zulässt. Die Prüfungsaufgaben hingegen dienen der Individualdiagnostik, da entschieden werden soll, ob der Mittlere Schulabschluss vergeben werden kann oder nicht. Für die einzelnen Schülerinnen und Schüler verknüpfen sich mit den Prüfungsaufgaben also weit reichende bildungsbiographische Konsequenzen.
- *Konstrukte:* In Bezug auf die zu überprüfenden Konstrukte strebt das IQB eine vollständige und umfassende Überprüfung aller sprachlicher Kompetenzen an, also sowohl der rezeptiven (Hören und Lesen) als auch der produktiven (Sprechen und Schreiben) Bestandteile. Wie bereits oben dargelegt, stößt dies jedoch im Falle des Kompetenzbereichs „Sprechen und Zuhören“ an technische und testdiagnostische Grenzen, die auch in den Prüfungsaufgaben nicht überwunden werden können. Insbesondere in den Prüfungsaufgaben des Landes Brandenburg werden aber auch die übrigen Kompetenzbereiche nicht in der Breite überprüft, wie dies im Rahmen der IQB-Arbeit geschieht. Insofern kann auch in Hinblick auf die Konstrukte nicht von einer hinreichenden Ähnlichkeit ausgegangen werden.
- *Zielpopulation:* Die Zielpopulation ist weitgehend vergleichbar.
- *Merkmale und Umstände der Messung:* In Hinblick auf die Merkmale und Umstände der Testdurchführung bestehen wiederum deutliche Unterschiede. Gelten für die Testung der Bildungsstandards standardisierte Durchführungsbedingungen, so liegen diese bei den Prüfungsaufgaben vollständig in der Hand der Schulen. Insofern kann hier nicht mehr von einer Durchführungsobjektivität ausgegangen werden. Erhebliche Einschränkungen sind auch im Hinblick auf die Objektivität der Auswertung der offenen Antworten anzunehmen. Somit kann festgehalten werden, dass die Prüfungsaufgaben keine standardisierten Leistungstests sind.

Dennoch könnte die Qualität der Items und Aufgaben aus testdiagnostischer Perspektive optimiert werden.

Man könnte sich bei der der Entwicklung der Prüfungsaufgaben bspw. am oben dargestellten Prozessverlauf (vgl. Kapitel 3.2) orientieren und eine Erprobung der Aufgaben vor ihrem Einsatz durchführen.

Für die geschlossenen Formate bestehen, wie angemerkt, einige Verbesserungsmöglichkeiten hinsichtlich der Konstruktion (bspw. Verwendung von mehr als 3 Antwortoptionen, Entwicklung

plausibler Distraktoren, Vermeidung identischer Eingangsformulierungen der Antwortoptionen etc.).

Außerdem könnte in stärkerem Maße darauf geachtet werden, dass die Items voneinander unabhängig sind.

Von besonderer Bedeutung ist auch die Optimierung der Auswertungshinweise zur Steigerung der Auswertungsobjektivität, was insbesondere dort relevant ist, wo offene Formate in großem Umfang zum Einsatz kommen.

9. Fazit

Abschließend kann festgehalten werden, dass die Prüfungsaufgaben für den Mittleren Schulabschluss im Fach Deutsch der Länder Berlin und Brandenburg in unterschiedlichem Maße mit den kompetenzorientierten Anforderungen der länderübergreifenden Bildungsstandards kompatibel sind. So ist in Brandenburg die Bandbreite der überprüften Kompetenzen und der einbezogenen Anforderungsbereiche deutlich geringer als in Berlin. Auch in testdiagnostischer Hinsicht sind die Prüfungsaufgaben des Landes Brandenburg in der Summe weniger überzeugend als die des Landes Berlin.

Literatur

- Amelang, M. & Zielinski, W. (2004). *Psychologische Diagnostik und Intervention*. Berlin: Springer.
- American Educational Research Association, American Psychological Association & National Council on Measurement in Education (Eds.). (1999). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.
- Becker-Mrotzek, M. & Böttcher, I. (2006). *Schreibkompetenz entwickeln und beurteilen*. Berlin: Cornelsen.
- Cizek, G. J. (2005). High-Stakes Testing: Contexts, Characteristics, Critiques, and Consequences. In: R. P. Phelps (Ed.). *Defending Standardized Testing*. Mahwah, NJ: Lawrence Erlbaum, pp. 23-54.
- Deutsches PISA-Konsortium (2001). (Hrsg.). *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske und Budrich.
- Downing, S. M. & Haladyna, T. M. (2006). (Eds.). *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum.
- Einecke, G. (1999). Auf die sprachliche Ebene lenken. Gesprächssteuerung, Erkenntniswege und Übungen im integrierten Grammatikunterricht. In: A. Bremerich-Vos (Hrsg.). *Zur Praxis des Grammatikunterrichts*. Freiburg/Br.: Fillibach, S. 125-191.
- Fix, M. (2006). *Texte schreiben – Schreibprozesse im Deutschunterricht*. Paderborn: Schöningh.
- Granzer, D., Böhme, K. & Köller, O. (2008). Kompetenzmodelle und Aufgabenentwicklung für die standardisierte Leistungsmessung im Fach Deutsch. In: A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.). *Lernstandsbestimmung im Fach Deutsch*. Weinheim: Beltz, S. 10 - 28.
- Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items*. Mahwah, NJ: Lawrence Erlbaum.
- Harsch, C., Neumann, A., Lehmann, R. & Schröder, K. (2007): Schreibfähigkeit. In: B. Beck und E. Klieme (Hrsg.). *Sprachliche Kompetenzen. Konzepte und Messungen. DESI-Studie*. Weinheim: Beltz, S. 42-62.
- Humboldt Universität zu Berlin. Institut zur Qualitätsentwicklung im Bildungswesen, IQB. (Hrsg.). (2007). *Perspektiven und Visionen. Tätigkeitsbericht 2005/06*.
- Impara, J. C. & Foster, D. (2006). Item and Test Development Strategies to Minimize Test Fraud. In: S. M. Downing & T. M. Haladyna (Eds.). *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum, pp. 91-114.
- Ingenkamp, K. & Lissmann, U. (2005). *Lehrbuch der Pädagogischen Diagnostik*. Weinheim: Beltz.
- Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Berlin: Bundesministerium für Bildung und Forschung.
- Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking*. New York: Springer.

- Landesinstitut für Schule und Medien Brandenburg (Hrsg.). (2006). *Ergänzende Informationen und Aufgabenbeispiele zu den schriftlichen Prüfungen am Ende der Jahrgangsstufe 10 im Fach Deutsch*.
- Landesinstitut für Schule und Medien Brandenburg (Hrsg.). (Schuljahr 2006/2007). *Prüfungsschwerpunkte und Hinweise zu den schriftlichen Prüfungen am Ende der Jahrgangsstufe 10 im Fach Deutsch*.
- Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz.
- Mislevy, R. J., Steinberg, L. S., Almond, R. G. (2002). On the Roles of Task Model Variables in Assessment Design. In: S. H. Irvine & P. C. Kyllonen (Eds). *Item Generation for test Development*. Mahwah, NJ: Lawrence Erlbaum, pp. 97-128.
- Mullis, I. V. S., Martin, M. O. & Kennedy, A. M. (2004). Item-writing guidelines for the PIRLS 2006 field test. Paper presented at the 2nd PIRLS 2006 NRC Meeting Bratislava, Slovak Republic.
- Neumann, A. (2006). *Vergleich der Konzeptionen und Inhalte der Prüfungen zum Mittleren Schulabschluss in Berlin und der Prüfungen Jahrgangsstufe 10 in Brandenburg Schuljahr 2005/06*. Institut für Schulqualität der Länder Berlin und Brandenburg e. V.
- Nitko, A. J. (2004). *Educational Assessment of Students*. Upper Saddle River, NJ: Merrill Prentice Hall.
- Nussbaumer, M. (1991). *Was Texte sind und wie sie sein sollen*. Tübingen: Niemeyer.
- Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats*. New York, NY: Kluwer.
- PISA-Konsortium Deutschland (2004). (Hrsg.). *PISA 2003 – Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs*. Münster: Waxmann.
- Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004). (Hrsg.). *Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss. Beschluss vom 4.12.2003*. München: Kluwe.
- van Ackeren, I. (2003). *Nutzung großflächiger Tests für die Schulentwicklung. Exemplarische Analyse der Erfahrungen aus England, Frankreich und den Niederlanden*. Berlin: Bundesministerium für Bildung und Forschung.
- Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen: Hogrefe.

Institut für Schulqualität der Länder Berlin und Brandenburg e. V.

www.isq-bb.de

