

Vergleichsarbeiten
10. Klasse
Schuljahr 2002/03

Auswertungsbericht

DEUTSCH

Impressum

Herausgeber

Senatsverwaltung für Bildung, Jugend und Sport
Beuthstraße 6 - 8, 10117 Berlin-Mitte

www.senbjs.berlin.de

Verantwortlich

Tom Stryck
Referat I D, Schul- und Qualitätsentwicklung,
Schulforschung, Fort- und Weiterbildung

Ansprechpartner für diesen Bericht

Dr. Wolfgang Wendt
I D 12
Telefon 030 9026 6508
wendt@zedat.fu-berlin.de

Fritz Tangermann
I D 8
Telefon 030 9026 5773
fritz.tangermann@senbjs.verwalt-berlin.de

INHALTSVERZEICHNIS

A	Vergleichsarbeiten Klasse 10 im Schuljahr 2002/03 - Der zweite Probelauf in den Fächern Deutsch, Englisch, Französisch und Mathematik.....	3
B	Der Bericht zu den Vergleichsarbeiten im Fach Deutsch.....	5
1.	Fragestellung, Beteiligung, Ergebnisse	5
1.1	<i>Zielsetzung, Vorgehensweise und Beteiligungsquoten.....</i>	5
1.2	<i>Wahl der Aufgabenvorschläge und der Aufgabenvarianten.....</i>	8
1.3	<i>Unterschiede in den Ergebnissen der einzelnen Schularten</i>	10
1.4	<i>Gemeinsamkeiten in den Ergebnisse der einzelnen Schularten.....</i>	13
1.5	<i>Einheitlichkeit beim Bewerten der Arbeiten.....</i>	16
2.	Resümee aus dem Probelauf der Vergleichsarbeiten Deutsch	19
2.1	<i>Zusammenfassung</i>	19
2.2	<i>Schlussfolgerungen und Empfehlungen.....</i>	22
3.	Anhang: Materialien, Tabellen, Kommentare.....	25
3.1	<i>Dokumentation der beiden Aufgabenvorschläge</i>	25
3.2	<i>Dokumentation der Erhebungsbögen samt einer Grundauszählung und einigen Erläuterungen</i>	29
3.3	<i>Aufgabenvorschlag 2: Schulartspezifische Ergebnisse bei den einzelnen Bewertungsaspekten.....</i>	40
3.4	<i>Wie tauglich sind die Kategorien des Bewertungsbogens?</i>	44
C	Zusammenfassung der wichtigsten Ergebnisse in den Fächern Englisch, Französisch und Mathematik	47
C 1	<i>Englisch</i>	47
C 2	<i>Französisch</i>	50
C 3	<i>Mathematik</i>	52
D	Glossar.....	57

A VERGLEICH SARBEITEN KLASSE 10 IM SCHULJAHR 2002/03 - DER ZWEITE PROBELAUF IN DEN FÄCHERN DEUTSCH, ENGLISCH, FRANZÖSISCH UND MATHEMATIK

Im Januar 2003 wurde durch den Senator für Bildung, Jugend und Sport, Klaus Böger, die Entscheidung getroffen, am Ende des Schuljahrs 2002/03 erneut Vergleichsarbeiten am Ende der 10. Klasse auf freiwilliger Basis in den Fächern Mathematik, Englisch, Französisch und Deutsch schreiben zu lassen.

Im ersten Probelauf für Vergleichsarbeiten in der Klasse 10 (Schuljahr 2001/02) waren den Schulen von der Senatsverwaltung Aufgabenbeispiele geliefert worden, die sie als Aufgaben oder als Muster für schulinterne Vergleichsarbeiten (Parallelarbeiten) nutzen konnten. Im Zentrum sollten die Erfahrungen der Lehrerinnen und Lehrer sowie der Schülerinnen und Schüler mit Vergleichsarbeiten stehen. Dementsprechend konzentrierte sich die Auswertung dieses Durchlaufs auf Fragen zur Planung, Durchführung und Bewertung der Arbeit.

Die Schulen bekamen in Form einer Broschüre und als Präsentation im Internet Aufgabemuster für die Fächer. Es war ihnen frei gestellt, die vorgegebenen Aufgaben unverändert zu übernehmen, sie an ihre Bedingungen anzupassen oder eigene Aufgaben zu entwickeln und ihre Arbeiten selber zusammenzustellen. Außerdem gab es seinerzeit eine schriftliche Befragung an den beteiligten Schulen, in der die Lehrkräfte ihre Schwierigkeiten berichten und Vorschläge zum weiteren Vorgehen machen konnten.

Auf der Grundlage der Auswertung des ersten Probelaufs wurde der zweite Probelauf für Vergleichsarbeiten am Ende der Klasse 10 für das Schuljahr 2002/03 konzipiert. Der erneute Probelauf richtete sich an alle Oberschulen und war wiederum freiwillig. Dieses Mal wurde die Aufgabenstellung jedoch zentral vorgegeben.

Der Fokus richtete sich nunmehr auf die Leistungen der Schüler/innen und die empirisch-statistisch zu überprüfende Güte der Aufgaben und der Aufgabenformate, die sich an gesetzten Bildungsstandards für das jeweilige Fach und in den Fremdsprachen auch am „Gemeinsamen europäischen Referenzrahmen“ (GeR) orientierten. Das Ergebnis dieser Vergleichsarbeiten sollte darüber Auskunft geben, inwieweit die Schülerinnen und Schüler mit solchen Formaten umgehen können und wo sie, gemessen an den bundesweiten bzw. europaweiten Standards (Anforderungen), stehen.

Hierzu mussten nicht nur geeignete Aufgaben gefunden oder eigens entwickelt werden, es war auch sicherzustellen, dass an den Schulen die Arbeiten nach einheitlichen Kriterien bewertet wurden, um die Ergebnisse vergleichbar und statistisch analysierbar zu machen. In Englisch und Französisch konnte auf europäisches und empirisch erprobtes Material zurückgegriffen werden. Im Fach Mathematik sicherte der gute fachdidaktische Diskussionsstand im Anschluss an TIMSS und PISA - in höherem Maße als dies bei anderen Fächern der Fall ist - geeignete Aufgaben und vor allem über die Schulen hinweg vergleichbare Bewertungsverfahren. Das Fach Deutsch weist hinsichtlich beider Aspekte Besonderheiten und fachdidaktischen Nachholbedarf auf.

Die folgende Tabelle zeigt im Vergleich die Anzahl Berliner Schülerinnen und Schüler und wie viele Arbeiten in die Auswertung gingen. Keine Berücksichtigung finden dort die Arbeiten der Berufsfachschulen, da keine genauen Angaben zur Grundgesamtheit der

Berliner Schülerschaft zur Verfügung standen. Wegen der besonderen Bedingungen befindet sich der analoge Vergleich für das Fach Französisch im entsprechenden Fachbericht.

Table: *Vergleich der Verteilungen von Schüler/innen in der Grundgesamtheit und in der Stichprobe differenziert nach Fächern*

	Grundgesamtheit	Stichprobe		
	Schüler/innen	Deutsch	Englisch	Mathematik
OH	3 438 (10%)	178 (13%)	769 (9%)	359 (7%)
OR	7 364 (23%)	311 (23%)	1 763 (20%)	1 104 (22%)
OG	11 606 (36%)	671 (50%)	3 363 (38%)	2 230 (44%)
O/OG	10 113 (31%)	189 (14%)	2 895 (33%)	1 377 (27%)
Gesamt	32 521 (100%)	1 349 (100%)	8 790 (100%)	5 117 (100%)

Für das Fach Englisch liegt eine weitgehende Übereinstimmung der Verteilungen auf die Schularten in der Grundgesamtheit der Stichprobe vor. Für das Fach Mathematik und stärker noch für das Fach Deutsch gilt eine Überrepräsentanz der Arbeiten aus den Gymnasien und eine unterproportionale Beteiligung der Gesamtschule.

Der vorliegende Bericht speist sich aus verschiedenen Quellen. Das federführende Referat I D der Senatsverwaltung für Bildung, Jugend und Sport hat Frau Roumiana Nikolova, Mitarbeiterin am Institut für Erziehungswissenschaften der Humboldt-Universität Berlin - Abteilung Empirische Bildungsforschung, Prof. Dr. Dr. Rainer Lehmann - beauftragt, eine Auswertung der Vergleichsarbeiten in den Fächern Mathematik, Englisch und Französisch vorzunehmen. Ihr umfangreiches Datenmaterial war in einem zweiten Schritt Grundlage für weitere Auswertungen, insbesondere zu den Lösungsverteilungen bei den Aufgaben der Vergleichsarbeiten in Mathematik, Englisch und Französisch. Diese weiteren Auswertungen, die fachdidaktischen Einordnungen der jeweiligen Tests und die fachliche Bewertung der Ergebnisse insgesamt wurden von den Mitarbeiterinnen und Mitarbeitern des Referats I D vorgenommen. Im Einzelnen waren dies:

Christian Bänsch (I D 7) für Mathematik,

Elke Dragendorf (I D 6) für Englisch,

Marita Hebisch-Niemsch (I D 9) für Französisch.

Im Fach Deutsch ist die Datenaufbereitung, -analyse und -interpretation von Dr. Wolfgang Wendt (I D 12) vorgenommen worden; er wurde fachlich begleitet von Fritz Tangermann (I D 8).

Koordination, Redaktion und Erstellung der Endfassungen erfolgte durch Tom Stryck (Referatsleiter I D).

B DER BERICHT ZU DEN VERGLEICH SARBEITEN IM FACH DEUTSCH

1. FRAGESTELLUNG, BETEILIGUNG, ERGEBNISSE

1.1 Zielsetzung, Vorgehensweise und Beteiligungsquoten

Im Mittelpunkt stand die Frage, ob es möglich sei, Vergleichsarbeiten im Fach Deutsch schulartübergreifend zu entwickeln, nicht zuletzt weil ihnen eine wichtige Funktion für einen mittleren Schulabschluss zukommen sollte. Schulartübergreifende Vergleichsarbeiten müssen das gesamte Leistungsspektrum der Klasse 10 dergestalt abdecken, dass sie an jeder Stelle dieses Spektrums mit einer näher festzulegenden Feinheit zwischen den Schülerinnen und Schülern differenzieren.

Hierzu mussten nicht nur geeignete Aufgaben gefunden werden, sondern es war auch sicherzustellen, dass an den Schulen die Arbeiten nach einheitlichen Kriterien bewertet wurden, damit die Ergebnisse vergleichbar sind und statistisch auszuwerten. Das Fach Deutsch weist hinsichtlich beider Aspekte Besonderheiten auf. In Englisch und Französisch konnte auf europäisches und erprobtes Material zurückgegriffen werden. In Mathematik sichert der wohl strukturierte Stoff selber - in höherem Maße als dies bei anderen Fächern der Fall ist - geeignete Aufgaben und vor allem über die Schulen hinweg vergleichbare Bewertungsverfahren. Bei diesen drei Fächern lagen also entweder bereits standardisierte Arten und Weisen der Bewertung vor oder konnten relativ einfach entwickelt werden.

Im Fach Deutsch konnte in Berlin an keine Tradition angeknüpft werden, die es erlaubt hätte, rasch und ohne allzu großen Aufwand einen ähnlichen Entwicklungsstand zu erreichen. Dies führte zu einigen Abweichungen im Vergleich mit den anderen Fächern. So wurde nicht eine, sondern es wurden zwei Arbeiten von einer Fachgruppe entwickelt:

- Textgrundlage 1: "Ein dicker Sack", Gedicht von Wilhelm BUSCH.
- Textgrundlage 2: "Schuldenberg per SMS. Mit der Handyrechnung kommt der Schock. Viele Jugendliche können ihre Schulden nicht zahlen.", Artikel aus der Berliner Zeitung vom 5./6. Mai 2001.

Die beiden Texte und die Aufgabenstellungen sind im Anhang dokumentiert. Die Schulen waren gehalten, beide Aufgabenvorschläge ihren Schülern/innen vorzulegen; diese sollten die Auswahl treffen.

Von ihrer Anlage her wurde im Fach Deutsch eine Vergleichsarbeit entwickelt, die weder signifikant vom Lernniveau der 10. Klassen abweicht, noch typologisch aus dem Rahmen fällt. Sie entspricht einem für Berliner Verhältnisse charakteristischen Stand in der Konstruktion von Aufgaben für das Ende der Sekundarstufe I, denn hinsichtlich der Art der Konstruktion, des Anforderungsniveaus und des Schwierigkeitsgrad liegen fachdidaktisch beiden Aufgaben die traditionellen Prinzipien des Rahmenplans der Sekundarstufe I, hier der 10. Klasse, zugrunde. Die Verbindung zu den Kompetenzbereichen ist gewahrt, wenngleich in den er-

warteten Leistungen nicht explizit gemacht. Zugleich weisen die Aufgaben intentional auf die Einheitlichen Prüfungsanforderungen (EPA) Deutsch hin und lassen sich von ihnen ableiten. Auch zu den Bildungsstandards für den Mittleren Schulabschluss werden Bezüge hergestellt.

Von den gewählten Themen ausgehend versuchte die Fachgruppe ferner, möglichst viele der einer Deutscharbeit üblicherweise zugrunde liegenden Bewertungsmaßstäbe explizit zu machen und in einzelne Bewertungsaspekte auszudifferenzieren. Die korrigierenden Lehrkräfte sollten - so das Ziel - bei jeder Arbeit für jeden dieser Aspekte überprüfen, in welchem Ausmaß die dadurch thematisierten Anforderungen erfüllt sind.

Anhand dieser Vorgaben der Fachgruppe ließ sich für jeden der beiden Aufgabenvorschläge ein Bewertungsbogen erstellen, der als eine Art Checkliste es den Lehrkräften ermöglichen sollte, möglichst einfach und unter möglichst geringen inhaltlichen Einbußen die Arbeiten zu bewerten und dies nach möglichst einheitlichen Kriterien, um die Vergleichbarkeit der Ergebnisse und die Zulässigkeit ihrer Auswertung zu sichern. Wir dokumentieren die beiden Bewertungsbögen samt einer Grundauszählung im Anhang. Sie enthalten, wie jedes empirische Erhebungsinstrument, eine qualitative und eine quantitative Komponente, nämlich zum einen die Bewertungsdimensionen und zum anderen mögliche Positionen auf diesen Dimensionen, wo sich die jeweilige Arbeit (zumindest näherungsweise und durch einfaches Ankreuzen) einordnen lässt. Damit können auch die Bewertungen von Deutscharbeiten empirisch-statistisch ausgewertet werden.

Im zweiten Probelauf konkretisiert sich für das Fach Deutsch eine zwifache Fragestellung:

1. Ist der hier eingeschlagene Weg, ein standardisiertes Bewertungsverfahren zu entwickeln, praktikabel? Führt das Bewerten zu nachvollziehbaren, plausiblen Ergebnissen?
2. Ist es überhaupt möglich, schulartübergreifend identische Deutscharbeiten schreiben zu lassen, die relevante spezifische Stärken und Schwächen aller Schüler/innen erfassen und - zunächst unabhängig von einer Zensur - eben diese Schüler/innen in einem einheitlichen Bewertungsraster verorten?

Nach einer Reihe von Informationsveranstaltungen zu den Ergebnissen des ersten und zum Verfahren des zweiten Probelaufes wurden Anfang Mai an alle Oberschulen Berlins die beiden Aufgabenvorschläge samt Erläuterungen und der Erhebungsbögen verschickt. Im Begleitschreiben wurde als Termin für die Vergleichsarbeiten der 27. Mai 2002 genannt. Die Beteiligung stand - mit Ausnahme der Sonderschulen - allen öffentlichen und privaten allgemein bildenden Oberschulen sowie einem Teil der berufsbildenden Schulen frei, nämlich jenen Klassen aus den Berufsfachschulen, in denen die Schüler/innen eine dem Realschulabschluss gleichwertige Ausbildung vermittelt bekommen. Die bis Ende August eingehenden Bewertungsbögen wurden per EDV erfasst und ausgewertet. Auf diese Grundlage stützt sich die nachstehende Darstellung der zentralen Ergebnisse. Differenziert nach Schulart zeigt *Tabelle 1* den Umfang der Beteiligung.

Table 1: Anzahl beteiligter Schulen, Klassen und Schüler/innen differenziert nach Schulart.

Schulart	Schulen	Klassen	Schüler/innen
OH	6 (19%)	14 (19%)	178 (12%)
OR	10 (31%)	18 (24%)	311 (22%)
OG	11 (34%)	27 (36%)	671 (47%)
O	3 (9%)	10 (14%)	189 (13%)
B	2 (6%)	5 (7%)	68 (5%)
Σ	32 (100%)	74 (100%)	1417 (100%)

Um einen Hinweis darauf zu erhalten, wie repräsentativ diese Stichprobe für die Berliner Schullandschaft ist, sind diese Zahlen mit den Werten für Gesamtberlin zu vergleichen. Zu unterscheiden ist hierbei der allgemein und der berufsbildende Bereich.

Für die (öffentlichen und privaten) berufsbildenden Schulen lassen sich aus den Statistiken nur Schätzungen ableiten. Vergleichsarbeiten sind möglich an den Berufsfachschulen, die insgesamt 14.731 Schüler/innen hatten. Hiervon durchliefen 5.563 Schülerinnen und Schüler den einjährigen Bildungsgang und gehören damit zum Potenzial. Ferner kommen etwa 4.500 weitere Schüler/innen für Vergleichsarbeiten in Frage, so dass sich die Grundgesamtheit auf etwa 8.000 Schüler/innen belaufen dürfte. Die Beteiligung an den berufsbildenden Schulen war mit nicht einmal 1% äußerst gering, so dass deren Ergebnisse nicht repräsentativ sind.

Von den allgemein bildenden Schulen liegen 1.349 Deutscharbeiten vor. Damit wurden etwas über 4% des Potenzials ausgeschöpft, wie *Table 2* zeigt, in der die Verteilung der Arbeiten auf die Schularten der Verteilung in der Grundgesamtheit gegenübergestellt wird. Ferner zeigt sich, dass Haupt- und Realschulen in der Stichprobe zu (etwa) denselben Anteilen wie in der Grundgesamtheit vertreten sind. Das Gymnasium hingegen ist über-, die Gesamtschule unterrepräsentiert.

Table 2: Vergleich der Verteilung von Schulen und Schüler/innen in der Grundgesamtheit und in der Stichprobe.

Schulart	Grundgesamtheit		Stichprobe	
	Schulen	Schüler/innen	Schulen	Schüler/innen
OH	62 20%	3 438 10%	6 20%	178 13%
OR	86 27%	7 364 23%	10 33%	311 23%
OG	121 38%	11 606 36%	11 37%	671 50%
O	49 15%	10 113 31%	3 10%	189 14%
	318 100%	32 521 100%	30 100%	1 349 100%

Table 3: Anzahl beteiligter Schulen, Klassen und Schüler/innen differenziert nach Stadthälften. (Ohne die beiden berufsbildenden Schulen, da deren Einzugsbereiche sich nicht auf bestimmte Bezirke beschränken.)

Schulart	Schulen	Klassen	Schüler/innen
Ost	16 (53%)	38 (55%)	641 (48%)
West	14 (47%)	31 (45%)	708 (52%)
Σ	30 (100%)	69 (100%)	1349 (100%)

Im Gegensatz zur Grundgesamtheit befinden sich in der Stichprobe mehr Schulen, Klassen und Schüler/innen aus der östlichen als aus der westlichen Stadthälfte. *Tabellen 2 und 3* machen deutlich, dass die Stichprobe keineswegs repräsentativ für die gesamte Berliner Schullandschaft ist. Dennoch lassen sich verlässliche Schlüsse aus den erfassten Daten ziehen, wenngleich mit gewissen Einschränkungen. So dürfen Gesamtwerte, in die alle Arbeiten eingehen, nicht als Berliner Werte schlechthin interpretiert werden, denn die einzelnen Komponenten (Schularten, Stadthälften) gehen nicht entsprechend den Anteilen in der Grundgesamtheit in deren Berechnung ein. Da zugleich aber die Fallzahlen für die einzelnen Schularten relativ hoch sind, können die schulartspezifischen Werte als gute Schätzungen für die Verhältnisse in der Grundgesamtheit genommen werden, auch wenn sicherlich noch Verzerrungen in unbekannter Weise in den Teilstichproben vorhanden sind. Dies sind z.B. regionale Ungleichgewichte - vgl. *Table 3* - und - was sich nur vermuten lässt, aber plausibel erscheint - eine positiv selektierte Stichprobe hinsichtlich der Leistungsfähigkeit der Schüler/innen.

1.2 Wahl der Aufgabenvorschläge und der Aufgabenvarianten

Von den zwei vorgegebenen Aufgaben, unter denen die Schüler/innen frei wählen konnten, wurde die erste deutlich bevorzugt. Eine Ausnahme bildet die Hauptschule, bei der sich die Wahl in etwa gleichmäßig auf beide Aufgaben verteilte.

Table 4: Gewählte Aufgaben differenziert nach Schulart

Aufgabe	OH	OR	OG	O	B	Σ
1	82 46%	61 20%	95 14%	28 15%	9 13%	275 19%
2	96 54%	250 80%	576 86%	161 85%	59 87%	1142 81%
Σ	178 100%	311 100%	671 100%	189 100%	68 100%	1417 100%

Die beiden Aufgaben existieren in mehreren Varianten, von denen jeweils die erste (Brief bzw. Problemerkörterung) am weitaus häufigsten gewählt wurde, wie *Table 5* zeigt.

Table 5: Gewählte Aufgaben differenziert nach Aufgabenvarianten und nach Schulart. (1-1: Inhaltsangabe/Brief, 1-2: Inhaltsangabe/appellative Rede, 1-3: Charakteristik/Brief, 1-4: Charakteristik/appellative Rede; 2-1: Problemerkörterung, 2-2: Leserbrief, 2-3: Appellative Rede.)

Aufgabe	OH	OR	OG	O	B	Σ
1-1	65 37%	46 15%	65 10%	17 9%	7 10%	200 14%
1-2	6 3%	8 3%	13 2%	4 2%	2 3%	33 2%
1-3	5 3%	4 1%	8 1%	6 3%	0 0%	23 2%
1-4	4 2%	1 0,3%	8 1%	1 0,5%	0 0%	14 1%
1-1, 1-2 #	2 1%	1 0,3%	1 0,1%	-- ---	- ---	4 0,3%
1-3, 1-4 #	-- ---	1 0,3%	-- ---	-- ---	- ---	1 0,1%
2-1	59 33%	192 62%	447 67%	123 65%	34 50%	854 60%
2-2	26 15%	46 15%	101 15%	32 17%	19 28%	224 16%
2-3	7 4%	12 4%	26 4%	4 2%	5 7%	54 4%
2-2, 2-3 #	4 2%	--- ---	2 0,3%	3 2%	1 1%	10 0,7%
Σ	178 100%	311 100%	671 100%	189 100%	68 100%	1417 100%

Bei einigen Bögen ließ sich nicht genau feststellen, auf welche der Aufgabenvariante die Bewertung sich bezieht.

Eine Erklärung der auffällig ungleichen Wahlhäufigkeiten könnte die Antwort auf die Frage liefern, worauf die beiden Aufgaben im Einzelnen zielen und welche impliziten didaktischen Tendenzen in ihnen enthalten sind.

Zu Aufgabenvorschlag 1

Das Leseverstehen eines Textes in gebundener/lyrischer Sprache wird im Gedicht von Wilhelm Busch mit einem semantischen Bereich verbunden, der jenseits der üblichen Erfahrungswelt liegt, zumindest ist er für viele Schüler/innen nicht offensichtlich ("Mühle", "Ährenfeld", "Federvieh"). Ein sprachlicher Ausdruck ("... der hoch von Nöten ist ..."), der ohne Anmerkung nur Verständnisschwierigkeiten bietet und damit enigmatisch wirkt, motiviert nicht dazu, die Aufgabe zu wählen.

Inhaltsangabe und Charakteristik prüfen neben dem Textverstehen die grundlegende Kompetenz, zentrale Aussage und spezifische Orientierung auf ein Item (den Sack) in eigene Worte zu fassen und sind Teil einer Transferleistung. Eine Inhaltsangabe bietet zusätzliche Probleme und kann nur sehr knapp ausfallen, da kaum Handlungsschritte im Text zu entdecken sind.

Auf der Ebene des begrifflichen Kennens und Verwendens sind Leistungen mit mittlerem Schwierigkeitsgrad zu erbringen (Ironie, Pointe, Autorenabsicht, Textsorte). Der Anforderungsbereich der Urteilsfähigkeit wird mit den zwei Varianten der produktiven Aufgabe erfüllt, denn Argumentation und Transfer von Kenntnissen aus Geschichte und Lebenswelt sind genuine Lernziele dieser Jahrgangsstufe. Damit wäre auch das diskursive Denken gefordert.

Hier wird der Text durch seine archaisierende Sprachstruktur ebenfalls nicht motivieren, sondern eher abschrecken. Auch wenn die Fabel schon längst Lern- und Lehrstoff in den Klassenstufen zuvor gewesen ist, erscheint das Wagnis ihn zu wählen als zu groß.

Die Frage nach den Fähigkeitsmustern, die den Lernenden mit dieser Aufgabe abverlangt werden, zielt weniger auf die Konstruktionsfähigkeit, mit dem Angebot der Aufgabe umzugehen, als auf die Verstehensleistung, ihre Grundkonstellation erst einmal zu begreifen. Der Aufgabenvorschlag leidet unter der didaktisch problematischen Textauswahl und versperrt so den Lernenden einen unter Testbedingungen besonders wichtigen motivationsfördernden Auswahlimpuls.

Zu Aufgabenvorschlag 2

Der Informationsgehalt der Aufgabe ist leicht zu erfassen, sie hat einen erkennbaren lebensnahen Bezug und ist vom Schwierigkeitsgrad des Textes her leicht. Eine konstruktive Leistung der gedanklichen und sprachlichen Gliederung ist unzweideutig gefordert und zu erbringen, insofern sind grundlegende Kompetenzbereiche angesprochen. Möglichkeiten zu Abstraktion und Konkretisierung sind gleichermaßen gegeben und können auf unterschiedlichen Niveaustufen ausgeführt werden. Daher ist die Aufgabe für eine Vergleichsarbeit geeignet.

Bei den Analysen wird wegen der weitaus größeren Fallzahlen und der angemesseneren Verteilung auf die Schularten die zweite Aufgabe im Vordergrund stehen. Auf die erste wird nur dann eingegangen, wenn zusätzlicher Erkenntnisgewinn zu erwarten ist.

1.3 Unterschiede in den Ergebnissen der einzelnen Schularten

In der Darstellung der Ergebnisse

- werden zunächst die schulartspezifischen Leistungsunterschiede aufgezeigt (Abschnitt 1.3),
- die aber zu relativieren sind angesichts erheblicher Überlappungen der Leistungsstände (Abschnitt 1.4), und
- schließlich wird der für Vergleichsarbeiten zentralen Frage nachgegangen, wie weit ein einheitliches Vorgehen beim Bewerten erreicht werden konnte (Abschnitt 1.5).

Nachstehend werden ausschließlich die Daten aus den allgemein bildenden Schulen herangezogen. Da nur zwei berufsbildende Schulen mit 68 Schülern/innen teilgenommen haben, sind Aussagen über diese Schulart nicht zulässig und verzerren in nicht kontrollierbarer Weise die Gesamtwerte. Die weiteren Darstellungen beziehen sich demnach auf (maximal) 266 Arbeiten für die Aufgabe 1 und auf 1.083 Arbeiten für die Aufgabe 2.

Am Ende des Bewertungsbogens wurden die Lehrkräfte gebeten, für die Arbeit einen zusammenfassenden Wert von 0 (sehr schlecht) bis 9 (sehr gut) zu vergeben. Mit Bedacht wurde bei dieser Spanne vom üblichen Zensurenpektrum abgesehen. Damit sollte unterstrichen werden, dass es bei der Punktevergabe nicht um die legitimerweise auch pädagogisch motivierte Notenvergabe, sondern um eine nur die Arbeit selber zu berücksichtigende Bewertung handelt. Die gegenüber Zensuren größere Abstufung sollte zudem die Möglichkeit fei-

nerer Differenzierung eröffnen.

Tabelle 6: Zusammenfassende Bewertung der Aufgabe 1. (Von 7 Arbeiten fehlt die zusammenfassende Bewertung.) Mittelwert = 4,4, Streuung = 2,3

	Häufigkeit	Anteil
0: sehr schlecht	14	5%
1: 1 von 9 Punkten	13	5%
2: 2 ...	36	14%
3: 3 ...	22	8%
4: 4 ...	49	19%
5: 5 ...	39	15%
6: 6 ...	32	12%
7: 7 ...	25	10%
8: 8 von 9 Punkten	20	8%
9: sehr gut	9	4%
	259	100%

Die in *Tabelle 6* dokumentierte Verteilung stellt keine Normalverteilung dar, auch wenn die Häufigkeiten an den Rändern geringer werden, die Verteilung also einer Glockenkurve ähnelt. *Tabelle 7* gibt die Mittelwerte und Streuungen pro Schulart wieder.

Tabelle 7: Zusammenfassende Bewertungen der Aufgabe 1 differenziert nach Schularten (Mittelwerte und Streuungen).

	OH	OR	OG	O	Σ
N	80	60	91	28	259
Mittelwert	2,9	5,2	5,4	4,3	4,4
Streuung	2,0	2,0	2,2	1,9	2,3

Wie erwartet befindet sich im Leistungsspektrum die Hauptschule am unteren und das Gymnasium am oberen Ende - bei einer allerdings überraschend geringen Differenz zur Realschule. Es lässt sich nicht mit Sicherheit sagen, was die Schüler/innen bei ihrer Auswahl bewogen hat. Betrachten wir die Verhältnisse bei der Hauptschule, bei der sich als einziger Schulart die Wahl auf beide Aufgaben in etwa gleichmäßig verteilt, dann spricht einiges dafür, dass die Aufgabe 1 eine höhere Attraktivität für schwächere Schüler/innen besitzt als Aufgabe 2. Daher ist es möglich, dass der gymnasiale Mittelwert höher gelegen hätte, hätten sich mehr Gymnasiasten/innen für die Aufgabe 1 entschieden.

Für die Aufgabe 2 liefert *Tabelle 8* die Verteilung der Punktwerte, *Tabelle D9* wiederum die schulartspezifischen Mittelwerte und Streuungen.

Table 8: Zusammenfassende Bewertung der Aufgabe 2. (Von 24 Arbeiten fehlt die zusammenfassende Bewertung.) Mittelwert = 5,1, Streuung = 1,8

	Häufigkeit	Anteil
0: sehr schlecht	5	0,5%
1: 1 von 9 Punkten	21	2%
2	64	6%
3: 3 von 9 Punkten	119	11%
4	179	17%
5: 5 von 9 Punkten	231	22%
6	202	19%
7: 7 von 9 Punkten	148	14%
8	73	7%
9: sehr gut	17	2%
	1059	100%

Table 9: Zusammenfassende Bewertungen der Aufgabe 2 differenziert nach Schularten: Mittelwerte und Streuungen

	OH	OR	OG	O	Σ
N	92	240	568	159	1059
Mittelwert	3,9	4,8	5,4	4,9	5,1
Streuung	2,0	1,6	1,8	1,6	1,8

Die Punktwerte folgen bei der zweiten Aufgabe einer Normalverteilung. Insgesamt gilt für die Mittelwerte

Aufgabe 1:	Mittelwert = 4,4	Streuung = 2,3
Aufgabe 2:	Mittelwert = 5,1	Streuung = 1,8

Im ersten Fall ist das Gesamtmittel nahezu identisch mit dem theoretischen Mittel von 4,5 (der Mitte zwischen 0 und 9), im zweiten Fall etwas darüber, ein Hinweis, dass die Aufgabe etwas zu leicht gewesen sein könnte, wobei allerdings zu berücksichtigen ist, dass die Arbeit für einen schulartübergreifenden Einsatz konzipiert wurde.

Beide Tabellen liefern vergleichbare Resultate: Die Mittelwerte sind so angeordnet, dass wir die Hauptschule am unteren Ende, das Gymnasium am oberen finden. Zu bedenken ist, dass die Lehrkräfte in der Regel ihre eigenen Schüler/innen bewertet haben, was bedeutet, dass die Hauptschullehrer/innen – ohne die Arbeiten von Schüler/innen anderer Schularten zu kennen – unterdurchschnittlich eingeschätzt haben; Entsprechendes gilt für die Lehrkräfte der anderen Schularten. Die hier erzielten Ergebnisse decken sich mit allem, was ansonsten über den Leistungsstand der verschiedenen Schularten bekannt ist. Daraus lassen sich als gut begründete Hypothesen zwei bedeutsame Ergebnisse festhalten:

1. Es ist offensichtlich gelungen, bei zentral vorgegebenen Aufgaben und dezentraler Bewertung ein weitgehend einheitliches Vorgehen zu sichern.
2. Die gewiss vorhandenen Unterschiede in der Leistungsfähigkeit der Schüler/innen werden durch die Vergleichsarbeit abgebildet - wie adäquat und wie genau können wir nicht sagen, aber die Ergebnisse der Vergleichsarbeit sind in hohem Maße realitätshaltig.

1.4 Gemeinsamkeiten in den Ergebnisse der einzelnen Schularten

Die *Tabelle 10* listet für die einzelnen Schularten die Verteilung der Punktwerte aus der zusammenfassenden Bewertung auf. Die Tabelle zeigt Unterschiede, wie sie sich bereits im vorhergehenden Abschnitt aus dem Vergleich der Mittelwerte ergeben haben, aber auch Gemeinsamkeiten.

Tabelle 10: Zusammenfassende Bewertung der Aufgabe 2 differenziert nach Schularten: Verteilung der Punktwerte

(Angabegeben werden jeweils die Häufigkeiten und die prozentualen Anteile.)

	Insgesamt		OH		OR		OG		O	
<i>sehr schlecht</i>										
0 Punkte	5	0,1%	5	5%	0	0%	0	0%	0	0%
1 Punkt	21	2%	7	8%	4	2%	8	1%	2	1%
2 Punkte	64	6%	13	14%	12	5%	27	5%	12	8%
3 Punkte	119	11%	10	11%	33	14%	57	10%	19	12%
4 Punkte	179	17%	22	24%	44	18%	81	14%	32	20%
5 Punkte	231	22%	18	20%	69	29%	109	19%	35	22%
6 Punkte	202	19%	10	11%	46	19%	117	21%	29	18%
7 Punkte	148	14%	3	3%	20	8%	100	18%	25	16%
8 Punkte	73	7%	4	4%	11	5%	54	9%	4	2%
9 Punkte	17	2%	0	0%	10	4%	15	3%	1	1%
<i>sehr gut</i>										
	1059	100%	92	100%	240	100%	568	100%	159	100%
Mittelwert	5,1		3,9		4,8		5,4		4,9	
Streuung	1,8		2,0		1,6		1,8		1,6	

Betrachtet werden hierzu die Prozentwerte, denn aufgrund der unterschiedlichen Repräsentanz der Schularten in der Grundgesamtheit und aufgrund der unterschiedlichen Beteiligung am zweiten Probelauf Vergleichsarbeiten variieren die Fallzahlen stark, ohne dass dies darüber hinausgehende inhaltliche Gründe hätte, die sich auf die Bewertungsergebnisse auswirkten. Die Prozentwerte hingegen geben uns Auskunft darüber, wie sich innerhalb der jeweiligen Schülerschaft die Punktwerte verteilen. So sehen wir beispielsweise, dass von den

(92) Hauptschülern/innen 20% (nämlich 18) den Punktwert 5 erreichten, 29% (69) der Realschüler/innen, 19% (109) der Gymnasiasten/innen und 22% (35) der Gesamtschüler/innen.

Tabelle 10 zeigt zum einen, dass 0 Punkte nur in Hauptschulen vergeben wurden und - verallgemeinert - bis zum Punktwert 2 die größten Prozentwerte bei den Hauptschulen auftreten, zum anderen das Gymnasium die größten Prozentwerte bei den hohen Punktwerten ab 6 aufweist. Die sich hier andeutenden Unterschiede sollen anhand der so genannten kumulierten Prozentwerte verdeutlicht werden. Es sind dies die Prozentwerte in der jeweils dritten Zeile der Darstellung, die sich durch Aufaddieren der Einzelprozentwerte ergeben (Beispiel OH: 5%; 5%+8%=13%; 13%+14%=27% etc.), daher kumulierte Prozentwerte.

Tabelle 11: Zusammenfassende Bewertung der Aufgabe 2 differenziert nach Schularten: Verteilung der Punktwerte und Verteilungsfunktion.

(Angabegeben werden jeweils die Häufigkeiten, die prozentualen Anteile und die kumulierten Prozentwerte. Von 24 Arbeiten fehlt die zusammenfassende Bewertung.)

Punkte	0	1	2	3	4	5	6	7	8	9
OH	5	7	13	10	22	18	10	3	4	0
	5%	8%	14%	11%	24%	20%	11%	3%	4%	0%
	5%	13%	27%	38%	62%	82%	93%	96%	100%	100%
OR	0	4	12	33	44	69	46	20	11	1
	0%	2%	5%	14%	18%	29%	19%	8%	5%	0%
	0%	2%	7%	21%	39%	68%	87%	95%	100%	100%
OG	0	8	27	57	81	109	117	100	54	15
	0%	1%	5%	10%	14%	19%	21%	18%	9%	3%
	0%	1%	6%	16%	30%	49%	70%	88%	97%	100%
O	0	2	12	19	32	35	29	25	4	1
	0%	1%	8%	12%	20%	22%	18%	16%	2%	1%
	0%	1%	9%	21%	41%	63%	81%	97%	99%	100%

Dem Punktwert 3 ist bei der Hauptschule der Prozentwert 38 zugeordnet; das bedeutet, dass 38% der Hauptschüler/innen einen Punktwert von 3 oder kleiner, d.h. einen Punktwert von höchstens 3 erreicht haben; andersherum: 62% der Hauptschularbeiten wurden mit 4 Punkten und mehr bewertet. Vier Punkte und mehr erzielten bei den Real- und Gesamtschulen 79% und bei den Gymnasien 84% der Schüler/innen. Die besseren Ergebnisse bei diesen Schularten lassen sich also daran ablesen, dass die kumulierten Prozentwerte langsamer ansteigen als bei der Hauptschule. Der Anstieg ist bei den Gymnasiasten/innen am langsamsten, d.h. hier ist der Prozentsatz der Schüler/innen im oberen Leistungsbereich am größten.

Die kumulierten Prozentwerte erlauben es, auf einfache Art und Weise den Anteil der Schüler/innen zu bestimmen, der einen bestimmten Schwellenwert überschreitet (oder eben nicht

erreicht). Beispiel: Das theoretische Mittel der 9-Punkteskala liegt bei 4,5. Unterhalb dieses Wertes liegen bei

der Hauptschule	62%	oberhalb	38%
der Realschule	39%		61%
dem Gymnasium	30%		70%
der Gesamtschule	41%		59%.

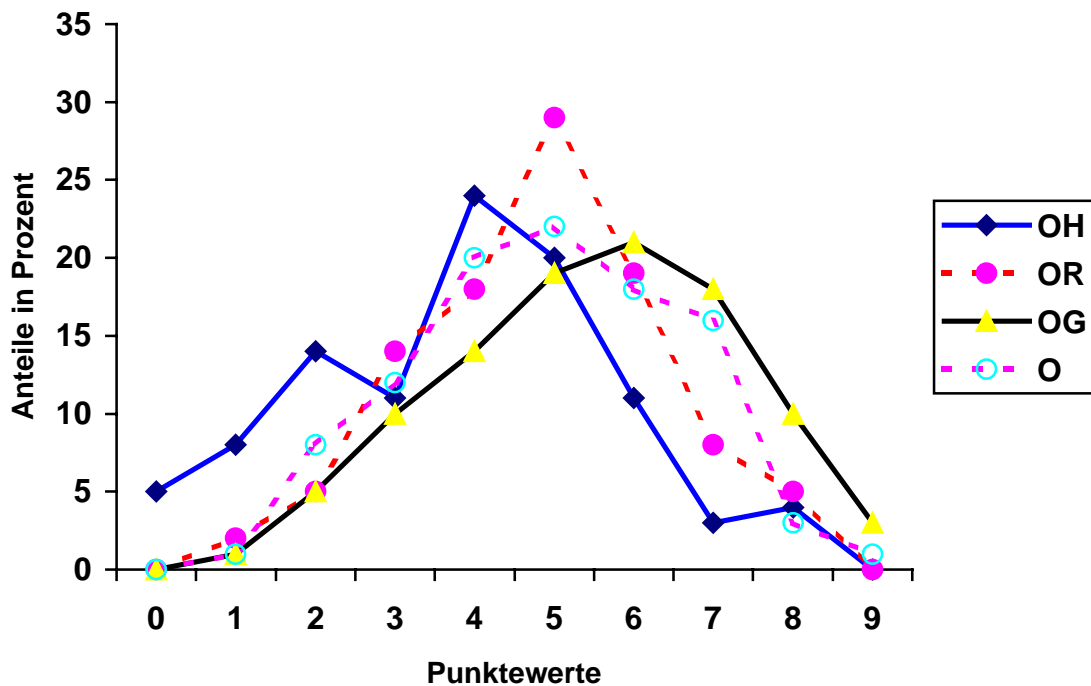
Bemerkenswert sind aber nicht allein die zwischen den Schularten auftretenden und in ihrer Richtung erwartbaren Unterschiede. Bemerkenswert sind auch und vor allem die Gemeinsamkeiten: Die Unterschiede sind in ihrem Ausmaß nicht so groß wie vorab zu vermuten gewesen wäre. Auch dies machen die kumulierten Prozentwerte deutlich, die - abgesehen von der Hauptschule - bei den anderen drei Schularten erst ab dem Punktwert 4 um mehr als 10 Prozentpunkt auseinanderklaffen (OG: 30%, O: 41%).

Deutlich werden die Gemeinsamkeiten auch daran, dass es nur zwei Punktwerte gibt, bei denen nicht alle vier Schularten vertreten sind, nämlich die beiden extremen Punktwerte 0 (OR, OG, O nicht vertreten) und 9 (OH nicht vertreten). Es gibt also große Überlappungsbereiche bei den vier schulartspezifischen Verteilungen.

Abbildung 1 zeigt die großen Gebiete, in denen sich die Werte der einzelnen Schularten überschneiden. Definieren wir mit LEHMANN et al.¹ (2002, 31) als typischen gymnasialen Leistungsbereich jenen, der sich oberhalb des Schnittpunktes der Verteilungen OR/OG befindet, also mit dem Punktwert 6 beginnt, dann befinden sich noch substantielle Anteile der Schülerschaften OH (7%), OR (13%) und O (19%) in diesem Bereich; vgl. *Tabelle 11*. Zu beachten sind allerdings die teilweise recht niedrigen Fallzahlen, die hinter den Prozentwerten stehen - dies gilt vor allem für die Hauptschule -, was die Verallgemeinerbarkeit stark einschränkt. Dass sich jedoch auch in unserer kleinen Erhebung das andernorts ebenfalls beobachtete Überlappungsphänomen zeigt, belegt die Behauptung, dass hier ein Stück Realität abgebildet wurde.

¹ R. H. LEHMANN, R. PEEK, R. GÄNSFÜß, V. HUSFELDT: Aspekte der Lernausgangslage und der Lernentwicklung - Klassenstufe 9. Ergebnisse einer Längsschnittuntersuchung in Hamburg. Hamburg: Amt für Schule 2002

Abbildung 1: Schulartspezifische Punkteverteilung der zusammenfassenden Bewertung.



1.5 Einheitlichkeit beim Bewerten der Arbeiten

Die schulartspezifischen Leistungsergebnisse, die sich mit denen anderer Untersuchungen decken, und der Befund, dass sich dieses Resultat mittels eines Auswertungsmodus ergeben hat, bei dem die Lehrkräfte jeweils die Arbeiten ihrer eigenen Schüler/innen bewertet hätten, ohne die Arbeiten anderer Schularten zu kennen, legen den Schluss nahe, dass es offensichtlich gelungen ist, bei zentral vorgegebenen Aufgaben und dezentraler Bewertung ein weitgehend einheitliches Vorgehen zu sichern. Daraus kann gefolgert werden, dass es einen verinnerlichteten Bewertungsmaßstab gibt, an dem sich schulartübergreifend alle orientieren.

Unterstützt worden ist dieses offensichtlich einheitliche Vorgehen durch einige Faktoren. Zu ihnen zählen neben der allerdings in der Regel weit zurückliegenden Ausbildung die aktuellen Erläuterungen hinsichtlich der Anforderungen und des Erwartungshorizontes, die die Themenstellungen begleiteten, auch wenn diese einen Interpretationsspielraum ließen, der schulartspezifisch hätte genutzt werden können.

Ferner sicherte der standardisierte Bewertungsbogen bis zu einem gewissen Maße eine einheitliche Bewertung, da er für alle korrigierenden Lehrkräfte wesentliche Aspekte explizit aufführte.

Und schließlich wird auch das Wissen, an einem schulartübergreifenden Vergleich teilzunehmen, die ansonsten auf Spezifika der jeweiligen Schüler/innen oder der jeweiligen Schule beruhende Relativierung der Bewertung zurückgedrängt haben.

In diesem Abschnitt wird die eingangs getroffene Feststellung anhand des vorliegenden Datensatzes weitergehend überprüft und differenziert. Dies geschieht anhand des Aufgabenvorschlags 2 und der Daten aus den allgemein bildenden Schulen. Ausgangspunkt ist die Frage, was die Lehrkräfte zu ihrer zusammenfassenden Bewertung am Ende des Bewertungsbogens bewogen hat. Welche der Bewertungsaspekte tragen am meisten zur globalen Einschätzung bei?

Zur Beantwortung werden alle Einzelaspekte mit dem Gesamturteil korreliert. (Die Korrelationskoeffizienten werden im vorliegenden Bericht nicht dokumentiert.) Wenn die schulart-spezifischen Muster der Korrelationskoeffizienten miteinander verglichen werden, ergeben sich Indizien dafür, wie einheitlich die Lehrkräfte in den verschiedenen Schularten ihre Bewertung vornehmen - der zentrale Aspekt bei schul(art)übergreifenden Vergleichsarbeiten. Eine derartige einheitliche Bewertung zu sichern, ist im Fach Deutsch mit größeren Problemen verbunden als in den anderen Fächern.

Der Bewertungsbogen zum Aufgabenvorschlag 2 besteht aus drei Teilen. Der erste betrifft allein die Varianten "Leserbrief" und "appellative Rede". Der zweite Teil ist inhaltsbezogen und themenspezifisch, während der dritte Teil sich der Darstellung (Gliederung, Sprache) widmet.

Über die Gesamtstichprobe hinweg ergeben sich die größten Korrelationskoeffizienten im dritten Teil. Allerdings trifft dies nicht auf die *Verstöße gegen die sprachliche Richtigkeit* zu, die nur einen mäßigen Zusammenhang zur Gesamtbewertung aufweisen; Fehler der Orthographie, Interpunktion und Grammatik besitzen für den Gesamteindruck einer Arbeit in der Regel nur untergeordnete Bedeutung. Der Umfang der Arbeit, quantifiziert durch die Anzahl der Wörter, ist etwas bedeutsamer, ragt aber nicht aus dem Feld der Bewertungsaspekte hervor. Als ebenfalls nachrangig müssen die Aspekte gelten, die nach umgangssprachlichen oder verkürzten Formulierungen fragen.

Bemerkenswert ist aber - von der Ausnahme *Verstöße gegen die sprachliche Richtigkeit* abgesehen -, dass die themenspezifischen Bewertungsteile 1 und 2, die sich auf die Inhalte der Themenstellung beziehen, im Allgemeinen zu kleineren Korrelationen führen als die Bewertungskriterien, die wir bei jedem Aufsatz anwenden könnten. *Gliederung* und *Sprachverwendung* (*Begriffe, Ausdruck*) dominieren z.B. gegenüber *Argumentation* und *Textwiedergabe*. Zwar gibt es die Aspekte *Argumentation* und *Textwiedergabe* bei allen Aufsätzen, die sich mit einer Vorlage auseinandersetzen, aber bedeutsam ist eben, dass die hieraus resultierende Anforderung an die korrigierende Lehrkraft, themenspezifische Inhalte der Arbeit bewerten zu müssen, für ihr Gesamturteil eine geringere Rolle als die Darstellungsleistung spielt.

Insbesondere die Aspekte *Gründe dafür/dagegen* (im Hinblick auf den Handybesitz) und *Konzentration auf das Wesentliche* sowie *Eigene Ratschläge* (Rubrik *Textwiedergabe*) weisen die diesbezüglich niedrigsten Korrelationen auf. Am wichtigsten für das Gesamturteil ist die Ausdrucksweise: rhetorisch gekonnt, flüssig und geschickt.

Wie ähnlich wird an den einzelnen Schularten bewertet? Gibt es vergleichbare Korrelationsmuster?

Es gibt nirgends wirklich gravierende Unterschiede; die Korrelationskoeffizienten gleichen sich in den Größenordnungen. Dies gilt insbesondere für die nachrangige Bedeutung der *Verstöße gegen die sprachliche Richtigkeit* und für die wichtige Rolle von *Gliederung* und *Sprachverwendung*.

Dennoch finden sich einige durchgehende und bemerkenswerte Unterschiede, etwa wenn bei der Gesamt- und vor allem bei der Hauptschule im Gegensatz zur Realschule und Gymnasium der zweite Bewertungsteil eine mindestens gleichwertige oder gar größere Bedeu-

tung für das Gesamturteil spielt. *Gliederung* und *Sprachverwendung* sind auch bei der Hauptschule wichtige Faktoren für die Bewertung der Deutscharbeit, aber im Gegensatz zu den anderen Schularten treten gleichrangig die inhaltsbezogenen und themenspezifischen Bewertungsaspekte wie Argumentation im Hauptteil und Schluss und die Textwiedergabe hinzu.

2. RESÜMEE AUS DEM PROBELAUF DER VERGLEICH SARBEITEN DEUTSCH

2.1 Zusammenfassung

Entwicklung der Aufgaben und des Bewertungsbogens

Der zweite Probelauf Vergleichsarbeiten gab allen freiwillig teilnehmenden Oberschulen zentral die Aufgaben vor, denn das Interesse galt nunmehr den Leistungen der Schüler/innen und der empirisch-statistisch zu überprüfenden Güte der Aufgaben. Für das Fach Deutsch konnte bei der Entwicklung der Vergleichsarbeit und des Bewertungsmodus nicht auf europäisches und erprobtes Material wie in den Fremdsprachen zurückgegriffen werden. Zudem ist die Materie selber nicht von sich aus im selben Ausmaß strukturiert wie z.B. in Mathematik.

Der Notwendigkeit, mehr als in den anderen beteiligten Fächern erproben zu müssen, wurde dadurch Rechnung getragen, dass eine Fachgruppe Deutsch zwei Aufgabenvorschläge (und diese in vier bzw. drei Varianten) entwickelte. Textgrundlage des ersten Vorschlages war ein Gedicht von Wilhelm Busch, des zweiten ein Zeitungsartikel zum Handygebrauch und zum Risiko der Verschuldung. Die Fachgruppe versuchte ferner, möglichst viele der einer Deutscharbeit üblicherweise zugrundeliegenden Bewertungsmaßstäbe explizit zu machen und in einzelne Bewertungsaspekte auszudifferenzieren. Anhand dieser Vorgaben ließ sich für jeden der beiden Aufgabenvorschläge ein standardisierter Bewertungsbogen erstellen.

Fragestellung und Ziel des zweiten Probelaufs

Zwei Fragen stehen somit im Mittelpunkt des Probelaufes Vergleichsarbeiten für das Fach Deutsch:

1. Ist der hier eingeschlagene Weg, ein standardisiertes Bewertungsverfahren zu entwickeln, praktikabel? Führt das Bewerten zu nachvollziehbaren, zu plausiblen Ergebnissen?
2. Ist es überhaupt möglich, schulartübergreifend identische Deutscharbeiten schreiben zu lassen, die relevante spezifische Stärken und Schwächen aller Schüler/innen erfassen und - zunächst unabhängig von einer Zensur - eben diese Schüler/innen in einem einheitlichen Bewertungsraster verorten?

Beteiligung, Aufgabenwahl und Repräsentativität der Stichprobe

Die Auswertung beruht auf insgesamt 1.417 Bewertungsbogen, die sich sehr ungleich auf Aufgabenvorschläge und Schularten verteilen: 80% der Schüler/innen wählten den Aufgabenvorschlag 2 (und von diesen wiederum drei Viertel die Variante 1 *Problemerkörterung*); Arbeiten aus Gymnasien sind über-, aus Gesamtschulen unterrepräsentiert, während die bei-

den anderen Schularten Haupt- und Realschule entsprechend ihrem Anteil an der gesamten Berliner Schülerschaft vertreten sind.² Diese Stichprobe ist zwar nicht repräsentativ für die gesamte Berliner Schülerschaft, aber hinreichend groß, um zu verlässlichen schulartspezifischen Aussagen auch im Vergleich zu kommen.

Angesichts der Fallzahlen beschränkt sich die Darstellung der Ergebnisse auf die Daten zum Aufgabenvorschlag 2. Im Abschnitt 1.2 wird versucht, anhand fachdidaktischer Überlegungen die so überaus eindeutige Vorliebe für den zweiten Vorschlag zu erklären.

Zusammenfassende Bewertung: Punktevergabe von 0 bis 9

Am Ende des Bewertungsbogens wurden die Lehrkräfte gebeten, für die Arbeit einen zusammenfassenden Wert von 0 (sehr schlecht) bis 9 (sehr gut) zu vergeben. Mit Bedacht wurde bei dieser Spanne vom üblichen Zensurenpektrum abgesehen. Damit sollte unterstrichen werden, dass es bei der Punktevergabe nicht um die legitimerweise auch pädagogisch motivierte Notenvergabe, sondern um eine nur die Arbeit selber zu berücksichtigende Bewertung handelt. Die gegenüber Zensuren größere Abstufung sollte zudem die Möglichkeit feinerer Differenzierung eröffnen.

Gesamtergebnis

Der Gesamtmittelwert des Globalurteils beträgt (beim Aufgabenvorschlag 2) 5,1, die Streuung 1,8. Damit liegt der Mittelwert etwas über der (zwischen 0 und 9) theoretischen Mitte von 4,5. Es gibt etwas mehr Arbeiten mit einer zusammenfassenden Bewertung oberhalb des Mittelwertes als darunter, also eine leicht asymmetrische Verteilung, aber der Normalverteilung stark angenähert; vgl. *Tabelle 8*. Zwei Folgerungen ergeben sich:

- Gehen wir von dem Ziel aus, dass eine alle Schularten umfassende Vergleichsarbeit das "mittlere Niveau" aller Berliner Schüler/innen treffen sollte, so ist dies weitgehend gelungen. Dass die Arbeit auf die Gesamtheit bezogen etwas zu leicht war (Mittelwert 5,1 ist größer als das theoretische Mittel 4,5) liegt an der Überrepräsentanz des Gymnasiums. (Zur strikten Ausrichtung einer Vergleichsarbeit an rein inhaltlichen Kriterien (Standards) siehe weiter unten.)
- Die Streuung von 1,8 verweist auf die Heterogenität der Ergebnisse, die Ausdruck der tatsächlich vorhandenen Heterogenität in der Berliner Schülerschaft sind.

Das gute Gesamtergebnis darf aber nicht vergessen machen, dass bei zahlreichen der einzelnen Bewertungsaspekte die maximale Punktzahl von bestenfalls der Hälfte der Schüler/innen (auch jenen des Gymnasiums) erreicht wurde. Dies gilt insbesondere für die Kriterien der inhaltlichen Bewältigung des Themas (vgl. Anhang 3.2 und 3.3).

Unterschiede zwischen den Schularten

Das Ergebnisspektrum des Globalurteils bietet bei beiden Aufgabenvorschlägen dasselbe Bild: Die Mittelwerte sind so angeordnet, dass wir die Hauptschule am unteren Ende, das

² Es gab nur 68 Arbeiten aus berufsbildenden Schulen, die bei der Auswertung i.d.R. außen vor blieben.

Gymnasium am oberen finden. Die hier erzielten Ergebnisse decken sich mit allem, was ansonsten über den Leistungsstand der verschiedenen Schularten bekannt ist. Das bedeutet:

Die Vergleichsarbeit bildet die tatsächlich vorhandenen Unterschiede in der Leistungsfähigkeit der Schüler/innen ab. Zwar lassen sich nur schwer Aussagen darüber treffen, wie adäquat und wie genau dies Abbild ist, aber die Ergebnisse sind in hohem Maße realitätshaltig.

Gemeinsamkeiten der Schularten

Treten einerseits zwischen den Schularten die in ihrer Richtung erwarteten Unterschiede auf, so gibt es andererseits - wie in anderen Untersuchungen auch - Gemeinsamkeiten dergestalt, dass die schulartspezifischen Verteilungen große Überlappungsbereiche aufweisen; vgl. *Abbildung 1*. Als typischer gymnasialer Leistungsbereich lässt sich jener bezeichnen, der sich oberhalb des Schnittpunktes der Verteilungen OR/OG befindet, also mit dem Punktwert 6 beginnt. In diesem Bereich befinden sich noch substantielle Anteile der Schülerschaften OH (7%), OR (13%) und O (19%). Dieses Ergebnis wird von den Werten des Abschnittes 3.3 im Anhang unterstrichen, die zeigen, dass es eine ganze Reihe von Bewertungskriterien gibt, die die Schüler/innen der Hauptschule ebenso gut erfüllen wie jene der Realschule.

Einheitliches Bewerten

Vergleichsarbeiten erfüllen nur dann ihre Funktion, den Leistungsstand von Klassen und Schulen im Vergleich untereinander zu bestimmen, wenn alle Arbeiten nach (in etwa) denselben Kriterien beurteilt werden. Dass diese zentrale Forderung weitgehend im zweiten Probelauf erfüllt werden konnte, zeigen zwei wichtige Belege:

- Die Lehrkräfte haben i.d.R. ihre eigenen Schüler/innen bewertet. Die schulartspezifischen Ergebnisse zum Globalurteil sind nicht nur plausibel, sondern decken sich mit den Resultaten aus der Schulforschung. Das bedeutet, dass die Hauptschullehrer/innen – ohne die Arbeiten von Schüler/innen anderer Schularten zu kennen – die Arbeiten ihrer Schüler/innen unterdurchschnittlich eingeschätzt haben; Entsprechendes gilt für die Lehrkräfte der anderen Schularten.
- Der (standardisierte) Bewertungsbogen erlaubt es (statistisch) zu überprüfen, wie wichtig die einzelnen der dort aufgeführten Bewertungsaspekte für die zusammenfassende Bewertung der korrigierenden Lehrkräfte sind. Es lassen sich somit Konstellationen von Bewertungskriterien ableiten. Und diese Konstellationen ähneln sich weitgehend über die Schularten hinweg; Genauer nachstehend.

Es ist somit gelungen, bei zentral vorgegebenen Aufgaben und dezentraler Bewertung ein weitgehend einheitliches Vorgehen zu sichern.

Konstellation der Bewertungskriterien

Der Bewertungsbogen besteht i.W. aus zwei Blöcken: Der eine enthält Bewertungsaspekte, die auf die inhaltliche Auseinandersetzung mit dem Thema zielen (Begründungen, Bewertungen), der andere konzentriert sich auf die Darstellung. Die themenspezifischen Bewertungsteile sind für das Globalurteil weniger bedeutsam als die Bewertungskriterien, die bei jedem Aufsatz angewandt werden könnten. *Gliederung* und *Sprachverwendung* (*Begriffe, Aus-*

druck) dominieren z.B. gegenüber *Argumentation* und *Textwiedergabe*. Am wichtigsten für das Gesamturteil ist die Ausdrucksweise: rhetorisch gekonnt, flüssig und geschickt.

Hier dürfte die Erklärung liegen, warum die Leistungen im zweiten Bewertungsblock höher liegen als die des ersten (im Schnitt werden hier mehr Punkte vergeben als bei den Bewertungsaspekten des ersten Blockes). Wenn die Lehrkräfte mehr Wert auf die Darstellungsleistungen legen, dann werden sie im Unterricht verstärkt darauf eingehen, was zu besseren Leistungen führt. Zudem lässt sich "inhaltliche Qualität" schwerer üben als Darstellungsfähigkeiten, zu denen sich in gewissem Umfang Regeln aufstellen und befolgen lassen.

Auch wenn sich bei allen Schularten ähnliche Bewertungsmuster zeigen, so finden sich einige durchgehende und bemerkenswerte Unterschiede. So spielt im Gegensatz zur Realschule und zum Gymnasium der zweite Bewertungsteil bei der Gesamt- und vor allem bei der Hauptschule eine mindestens gleichwertige oder gar größere Bedeutung für das Gesamturteil. *Gliederung* und *Sprachverwendung* sind auch bei der Hauptschule wichtige Faktoren für die Bewertung der Deutscharbeit, aber im Gegensatz zu den anderen Schularten treten gleichrangig die inhaltsbezogenen und themenspezifischen Bewertungsaspekte wie *Argumentation* im Hauptteil und *Schluss* und die *Textwiedergabe* hinzu.

2.2 Schlussfolgerungen und Empfehlungen

Zentraler Befund

Eine schulartübergreifende Vergleichsarbeit ist auch im Fach Deutsch möglich. Es ist gelungen bei zentral vorgegebenen Aufgaben und dezentraler Bewertung ein weitgehend einheitliches Vorgehen zu sichern. Die Unterschiede in der Leistungsfähigkeit der Schüler/innen werden durch die Vergleichsarbeit abgebildet.

Weiterentwicklungen der Aufgaben/Standards und Kompetenzen

Zwar hat die jetzige Vergleichsarbeit offensichtlich ein mittleres Schwierigkeitsniveau getroffen, so dass das gesamte Leistungsspektrum der Berliner Schülerschaft auf der zehnten Klassenstufe abgebildet werden konnte, aber ungeklärt bleibt der Bezug zu Standards und Kompetenzen: Was genau wird von den Schülern/innen erwartet? Welche kognitiven und metakognitiven Kompetenzen spielen eine Rolle? Wo liegen die Kriterien dafür, warum welche Kompetenzen überprüft werden sollen? Bei der Konstruktion der neuen Aufgabenformate sollten die Stärken und Mängel vor der Folie des Könnens sichtbar werden. Hierzu sind die zu fördernden und zu erwartenden Kompetenzen genau zu bestimmen. Über die bereits angeführten Fragen hinaus, ist auch zu klären, ob die Anforderungsbereiche in den neuen Standards für den Mittleren Schulabschluss den Aufgabentypen und Anforderungsbereichen in den EPA entsprechen.

Die Analyse des Leseverstehens in den Aufgaben kann nicht getrennt von der Darstellungsleistung erfasst werden. Hier wird die integrative Funktion des Deutschunterrichts deutlich. Zu überlegen wäre, wie weit eine deutliche Trennung für die Zwecke der Testkonstruktion anzustreben ist. Auch hinsichtlich der Frage eines einfachen Textverstehens und wie es gelöst wurde, können mit Hilfe der Auswertung bislang keine Schlüsse gezogen werden.

Anforderungen an die Vergleichsarbeit als Test- und Diagnoseinstrument

Die vorstehend erhobene Forderung, die einzelnen Leistungs- und Bewertungsaspekte deutlich herauszuarbeiten und in der Auswertung getrennt zu berücksichtigen, hat nicht zuletzt das Ziel, Vergleichsarbeiten in ihrer Funktion als Diagnoseinstrument zu schärfen. Damit gewinnen Vergleichsarbeiten an Bedeutung zugleich für die teilnehmenden Lehrkräfte und für das Systemmonitoring sowie die künftigen Bildungsberichte.

Sicherzustellen ist hierbei, dass weiterhin das gesamte Leistungsspektrum gleichmäßig abgebildet wird. Dies ist insbesondere deswegen wichtig, weil im Fach Deutsch die Vergleichsarbeiten für alle Schularten und alle Kursniveaus verbindlich vorgeschrieben sind.

Vergleichsarbeiten haben also zwei Bezugspunkte: Die an Kompetenzen orientierten Standards für das Fach Deutsch und das spezifische Leistungsspektrum der Berliner Schülerschaft in den zehnten Klassen.

Bewertung der einzelnen Vergleichsarbeiten

Offensichtlich ist es gelungen, ein weitgehend einheitliches Vorgehen der Lehrkräfte beim Bewerten der Vergleichsarbeiten zu sichern. Entscheidend hierfür dürften zwei Maßnahmen gewesen sein:

- Eine genaue Beschreibung des Erwartungshorizontes,
- der standardisierte Bewertungsbogen.

Beides lässt sich sicherlich noch verbessern; vgl. hierzu u.a. Abschnitt 3.4 im Anhang. Hierzu gehört beispielsweise die Prüfung, ob mit mehr Abstufungen eine differenziertere und angemessenere Bewertung ermöglicht werden sollte. Die Abkoppelung Punktebewertung (Globalurteil am Ende des Bewertungsbogens) - Zensurenvergabe sollte beibehalten werden. Das Erste ist von großer Bedeutung für die schulübergreifenden Vergleiche, das Zweite berücksichtigt schul- und klassenspezifische Belange.

Landesweite Auswertung der Vergleichsarbeiten

Die Ergebnisse aus den Vergleichsarbeiten müssen landesweit ausgewertet werden, um Bezugspunkte sowohl für die Schulen als auch für das System Monitoring zu bekommen. Es müssen Aussagen über das allgemeine Leistungsniveau, spezifische inhaltliche Stärken und Schwächen sowie Unterschiede zwischen den Schulen und Schularten möglich sein. Hierfür können und müssen die einzelnen Arbeiten "vor Ort" korrigiert und bewertet werden. Darüber hinaus ist eine repräsentative Stichprobe aus allen Arbeiten zu ziehen und zentral auszuwerten.

Bewertungsverhalten der korrigierenden Lehrkräfte

Zu überprüfen ist in weiteren Durchgängen, ob sich der Befund des zweiten Probelaufs bestätigt, dass die Lehrkräfte bei ihrer Einschätzung mehr Wert auf die Darstellungsleistung als die inhaltliche Auseinandersetzung legen. (Die Darstellungsleistungen der Schüler/innen werden ebenfalls als besser eingestuft als die der inhaltlichen Auseinandersetzung.)

Zweitkorrektur

Bislang konnten die Deutscharbeiten noch keinen Zweitkorrekturen unterzogen werden. Dies ist deshalb bedauerlich, weil erst durch eine Zweitkorrektur der endgültige Nachweis erbracht werden kann, dass tatsächlich die Lehrkräfte nach einheitlichen Kriterien bewertet haben.

Für die künftigen Durchgänge ist sicherzustellen, dass an einer Stichprobe der Arbeiten eine Zweitkorrektur durchgeführt wird. Dabei sollten die zweitkorrigierenden Lehrkräfte Arbeiten sowohl aus ihrer als auch aus anderen Schularten bekommen, um den schulartübergreifenden Charakter der Bewertung untersuchen zu können.

3. ANHANG: MATERIALIEN, TABELLEN, KOMMENTARE

3.1 Dokumentation der beiden Aufgabenvorschläge

AUFGABENVORSCHLAG 1

Untersuchendes und gestaltendes Erschließen eines literarischen Textes

Textgrundlage: Fabel

Wilhelm Busch (1832 – 1908)

Ein dicker Sack³

Ein dicker Sack – den Bauer Bolter,
der ihn zur Mühle tragen wollte,
um auszuruhen, mal hingestellt
dicht an ein reifes Ährenfeld –
5 legt sich in würdevolle Falten
und fängt'ne Rede an zu halten.

„Ich“, sprach er, „bin der volle Sack.
Ihr Ähren seid nur dünnes Pack.
Ich bins, der euch auf dieser Welt
10 in Einigkeit zusammenhält.
Ich bins, der hoch von Nöten ist,
dass euch das Federvieh nicht frisst;
ich, dessen hohe Fassungskraft
euch schließlich in die Mühle schafft.
15 Verneigt euch tief, denn ich bin der!
Was wäret ihr, wenn ich nicht wär?“
Sanft rauschen die Ähren:
„Du wärest ein leerer Schlauch,
wenn wir nicht wären.“

³ Zitiert nach: Fabeln, Parabeln, Gleichnisse. Hrsg. von R. Dithmar, dtv-Taschenbuch 404, München 1970, S. 267

Aufgabenvorschlag 1: Aufgabenstellung

Du kannst zwischen vier Möglichkeiten wählen:

1. Inhaltsangabe + Brief
- oder 2. Inhaltsangabe + appellative Rede
- oder 3. Charakteristik + Brief
- oder 4. Charakteristik + appellative Rede

Gib in der Überschrift deine Wahl an: ... (Nummer) Aufgabenstellung

Inhaltsangabe:

Aufgabe: Gib den Inhalt der Fabel kurz wieder! (keine Nacherzählung, keine Erlebniserzählung).

Arbeitshinweis: Erkläre auch die Autorenabsicht und wie dies verdeutlicht wird!

Charakteristik:

Aufgabe: Charakterisiere den dicken Sack und beurteile, wer damit gemeint sein könnte!

Arbeitshinweis: Berücksichtige dabei, wer alles den dicken Sack charakterisiert und in welcher Absicht!

Brief o d e r appellative Rede

Aufgabe: Schreibe **entweder** einen Brief an diesen dicken Sack

oder Verfasse eine appellative Rede an ihn. Nimm dessen Argumentation zu seiner Bedeutung kritisch unter die Lupe und mach ihm klar, was er erkennen bzw. lernen sollte!

Arbeitshinweise (für den Brief und die appellative Rede):

- Denke an eine Anrede. Beziehe bei deiner Argumentation ein, in welchen Rollen sich der dicke Sack in seiner Rede sieht und was es für die Ähren bedeutet, von ihm zur Mühle getragen zu werden!
- Du kannst bei deiner Argumentation auch Beispiele aus Geschichte und Gegenwart zur Verdeutlichung anführen.

AUFGABENVORSCHLAG 2**Erörterndes oder gestaltendes Erschließen eines Gebrauchstextes**

Vorbemerkung: Die Textgrundlage des Aufgabenvorschlages 2 ist ein Zeitungsartikel, der den Schüler/innen in Kopie aus der Zeitung vorgelegt wurde. Das Original erschien in der Berliner Zeitung, Nummer 104 vom 5./6. Mai 2001. Wir geben den Artikel an dieser Stelle als Abschrift wider und müssen daher auf die Zeilennummerierung verzichten. Auf diese Nummerierung bezog sich eine Anmerkung: Hinweis zu Zeile 19: 150 DM = ca. 75 Euro.

Schuldenberg per SMS

Mit der Handy-Rechnung kommt oft der Schock: Viele Jugendliche können ihre Schulden nicht zahlen

von Jochen Schneider

Handys gehören mittlerweile zur Grundausstattung. Auf dem Schulhof klingelt es in allen Tonlagen und geplagt Mitmenschen verdrehen die Augen. Fast die Hälfte aller Jugendlichen besitzt inzwischen ein Mobiltelefon. Neben dem Telefonieren fungiert das Handy aber vor allem als Sendestation für SMS-Nachrichten. Beim freudigen Hin- und Hertippen werden allerdings die Kosten oft außer Acht gelassen. Auch Anrufe wie "Mal eben hören wie in Kreuzberg das Wetter ist" sind auf Dauer ziemlich teuer. Die Berliner Schuldnerberatung stellte einen Richtwert auf: Ein 8-Minuten-Gespräch und zwei SMS pro Tag kosten monatlich rund 150 Mark. Viele Schüler können die Rechnungen nicht zahlen. Was sie nicht daran hindert, weiter zu telefonieren. Denn wenn sie bei einem Handyanbieter gesperrt werden, schließen sie bei einem anderen einfach einen neuen Vertrag ab. Handyrechnungen sind jedoch nicht der einzige Grund für frühe "Verschuldungskarrieren". "Oft werden Bestellungen aufgegeben, die nicht bezahlt werden könnten", sagt Bettina Heine, Schuldnerberaterin in Charlottenburg. Um die frühen "Verschuldungskarrieren" zu verhindern, hat die Berliner Schuldnerberatung jetzt ein "Handy-Booklet" herausgebracht, das Tipps zum vernünftigen Telefonieren vermitteln will.

Für Heine ist der erste Schritt getan, wenn sich ihre Klienten eingestehen, dass sie ein Problem haben. Gemeinsam mit den Jugendlichen wird dann ein Weg gesucht, wie die Schulden wieder abgearbeitet werden können. Wer auf sein Handy nicht verzichten will, sollte vom Festvertrag auf ein Telefon mit Prepaid-Karte umsteigen.

Infos zum "Handy-Booklet" unter: www.schuldnerberatung-berlin.de

Aufgabenvorschlag 2: Aufgabenstellung

3 Aufgabenarten zur Auswahl

Thema

Ein Handy – notwendig für jeden Jugendlichen heutzutage?

Textgrundlage: Artikel "Schuldenberg per SMS" der Berliner Zeitung vom 5./6. Mai 2001

1. Aufgabenart: Problemerkörterung

Aufgabe: Erörtere die Themafrage, indem du die Textgrundlage einbeziehst, und gib eine gut begründete Antwort!

2. Aufgabenart: Leserbrief

Aufgabe: Schreibe, indem du die Textgrundlage einbeziehst, einen Leserbrief an die Schülerzeitung einer bzw. deiner Schule zur Themafrage und nimm Stellung!

3. Aufgabenart: Appellative Rede

Aufgabe: Verfasse, indem du die Textgrundlage einbeziehst, zur Themafrage eine appellative Rede, in der du den Adressaten von deiner Auffassung überzeugen willst! Den Adressaten, z.B. Eltern, Lehrer oder Jugendliche, die anderer Meinung als du sind, kannst du selbst bestimmen.

Arbeitshinweise (für alle drei Aufgabenarten):

Bei der Darstellung des Problems kannst du einen aktuellen Anlass als Grund für deine Auseinandersetzung mit der Themafrage anführen.

Beziehe auch eigene Erfahrungen in die Argumentation ein!

Je nach deiner Auffassung musst du Tipps zum vernünftigen Gebrauch des Handys geben oder von dessen Anschaffung begründet abraten!

Gib die von dir gewählte Aufgabenart in der Überschrift an!

3.2 Dokumentation der Erhebungsbögen samt einer Grundauszählung und einigen Erläuterungen

Um die Ergebnisse der Grundauszählung in die Erhebungsbögen eintragen zu können, wurde ihr Layout leicht verändert. Im Original gab es für jeden der beiden Aufgabenvorschläge einen Erhebungsbogen, der aus zwei Seiten bestand. Wie die Antworten der beurteilenden Lehrkräfte numerisch codiert wurden, geht aus den Erhebungsbögen selber hervor; die entsprechenden Zahlen finden sich oberhalb der verbalen Abstufungen der einzelnen Antworten.

Beispiel: Aufgabenvorschlag 1, Bewertungsaspekt 01 "Angabe der Textsorte". Die mögliche Antwort "nicht vorhanden" wurde mit 0 codiert, "vorhanden, aber eher unpassend/ungenau" mit 1 etc.

Aufgrund dieser numerischen Codierungen ließ sich für jeden Bewertungsaspekt ein dazugehöriger *Mittelwert* (\emptyset) berechnen, in den eingeht, wie häufig die 0, die 1 etc. angekreuzt wurden: je größer der Mittelwert ist, desto besser die Bewertung.

Wegen der besseren Lesbarkeit ist in den folgenden Auswertungen auf die Mittelwertdarstellung verzichtet worden. Stattdessen werden pro Bewertungsaspekt die Anteile in Prozent angegeben, die auf die einzelnen Antwortabstufungen entfallen, also die Antwortverteilung.

Zudem steht am Anfang eines jeden Teiles die Anzahl aller Arbeiten (N), die ihm zugrunde liegen. Längst nicht bei allen Arbeiten jedoch sind auch alle Aspekte bei der Bewertung berücksichtigt worden, d.h. N gibt die Zahl maximal möglicher Bewertungen an. Der Frage, welche Bewertungsaspekte häufiger ausgelassen wurden und was das bedeuten könnte, wird im Anhang 3.4 nachgegangen.

In die hier dokumentierten Werte flossen auch die Daten der berufsbildenden Schulen ein.

Die Darstellung geht an einigen Stellen über eine Dokumentation des Erhebungsbogens und der Grundauszählung hinaus, um zum einen auf einige besonders interessante Ergebnisse hinzuweisen, und um zum anderen beispielhaft vorzuführen, wie sich die aufgeführten Werte interpretieren lassen. Dabei bleiben die Ergebnisse des Aufgabenvorschlages 1 aufgrund der geringen Fallzahlen unkommentiert.

AUFGABENVORSCHLAG 1: Erhebungsbogen samt Grundauszählung

Schulnr.		Klassennr.		Schüler/in-nr.
30 Schulen		60 Klassen		275 Sch

teilgenommen

Der Arbeit liegt die Aufgabenstellung

1 Inhaltsangabe	-	Brief	200	73%
2 Inhaltsangabe	-	appellative Rede	33	12%
3 Charakteristik	-	Brief	23	8%
4 Charakteristik	-	appellative Rede	<u>14</u>	5%
			270 ⁴	zugrunde.

⁴ Von den restlichen fünf Arbeiten ließ sich nur noch feststellen, dass vier der Arbeiten einer der beiden Aufgabenstellungen 1 oder 2 und die fünfte zur Aufgabenstellung 3 oder 4 gehörten.

Aufgabenstellung 1 und 2 (Aspekt: Inhaltsangabe)

ZUM ERSTEN SATZ	0 nicht vorhanden	1 vorhanden, aber eher unpassend/ ungenau	2 vorhanden, aber eher passend/ genau	3 vorhanden und passend/ treffend
N=237	Anteil	Anteil	Anteil	Anteil
01 Angabe der Textsorte	19%	5%	10%	56%
02 des Titels	26%	6%	6%	61%
03 des Autors	21%	5%	6%	68%
04 des Themas	37%	17%	18%	28%
	niedrig	eher niedrig	eher hoch	hoch
05 Abstraktionsgrad	29%	35%	24%	12%
WIEDERGABE DES GESCHEHENSABLAUFES	unzutreffend	eher unzutreffend	eher zutreffend	zutreffend
06 Darstellung	9%	18%	45%	29%
	niedrig	eher niedrig	eher hoch	hoch
07 Abstraktionsgrad	27%	38%	23%	12%
ZU DEN GESTALTUNGSMITTELN VON W. BUSCH	nicht erkannt	ansatzweise erkannt	wesentliche Aspekte erkannt	(nahezu) voll- ständig erkannt
08 Pointe	55%	19%	19%	7%
09 Ironie	66%	18%	12%	4%
10 Autorenabsicht	30%	24%	31%	15%
STIL DES/DER SCHÜLERS/IN	keiner	eher wenig	eher häufig	überwiegend
11 Nacherzählung: Anteil am Text	41%	32%	18%	8%
12 Erlebniserzählung: Anteil am Text	64%	25%	9%	3%
ZITIERWEISE	völlig falsch	eher falsch	eher korrekt	völlig korrekt
13 Korrektheit	19%	21%	43%	17%

AUFGABENVORSCHLAG 1 (Fortsetzung)**Aufgabenstellung 3 und 4 (Aspekt: Charakteristik)**

Zum ersten Satz	0 nicht vorhanden	1 vorhanden, aber eher unpassend/ ungenau	2 vorhanden, aber eher passend/ genau	3 vorhanden und passend/ treffend
N=38	Anteil	Anteil	Anteil	Anteil
14 Angabe der Textsorte	46%	11%	3%	40%
15 des Titels	46%	6%	6%	43%
16 des Autors	40%	6%	6%	49%
Thematik/Problematik	nicht erfasst	erfasst, aber ungenau	erfasst und eher treffend	erfasst und treffend
17 Einleitung/Schluss: Typus	14%	14%	40%	31%
	niedrig	eher niedrig	eher hoch	hoch
18 Abstraktionsgrad	11%	37%	31%	20%
Art der Charakteristik	nicht angegeben	angegeben, aber ungenau	angegeben und eher treffend	angegeben und treffend
19 indirekt (Erzähler, Ähren)	29%	29%	17%	26%
20 direkt (Rede)	29%	16%	29%	26%
Art der Charakteristik	nicht erfasst	erfasst, aber ungenau	erfasst und eher treffend	erfasst und treffend
21 Selbstverständnis des dicken Sackes	6%	20%	46%	29%
22 Verhalten den Ähren gegenüber	9%	20%	43%	29%
23 Wirkung/Absicht	14%	29%	34%	23%
Zitierweise	völlig falsch	eher falsch	eher korrekt	völlig korrekt
24 Korrektheit	9%	5%	57%	29%

AUFGABENVORSCHLAG 1 (Fortsetzung)**Aufgabenstellung 1, 2, 3 und 4 (Aspekt: Brief bzw. appellative Rede)**

Anrede	0 nicht vorhanden	1 vorhanden, aber eher unangemesse n	2 vorhanden, aber eher angemessen	3 vorhanden und angemessen
N=275	Anteil	Anteil	Anteil	Anteil
25 Qualität der Anrede	10%	13%	26%	51%
Zur Einleitung: Hinführung/ Darstellung der Problematik	nicht vorhan- den/ sehr kurz	eher kurz	eher ausführlich	ausführlich
26 Umfang der Darstellung	24%	38%	25%	12%
	Darstellung nicht vorhanden	erfasst, aber ungenau	erfasst und eher treffend	erfasst und treffend
27 Art der Darstellung	18%	26%	36%	20%
Zum Hauptteil: Deutung der ...	nicht vorhanden	vorhanden, aber eher unzutreffend	vorhanden und eher zutreffend	vorhanden und zutreffend
28 eigenen Rollenzuweisung des Sackes	17%	15%	46%	22%
29 Funktion des Sackes	18%	18%	41%	23%
30 Beispiele zur Verdeutlichung	45%	12%	26%	17%
Der Schluss	nicht vorhanden	vorhanden, aber eher unzutreffend	vorhanden und eher zutreffend	vorhanden und zutreffend
31 Kern-/Zielsatz	25%	22%	35%	18%
32 Appell	20%	18%	40%	22%
Argumentation	nicht vorhanden/ gering	vorhanden, aber eher unzureichend	vorhanden, und eher angemessen	vorhanden und angemessen
33 Orientierung auf die Sache	18%	28%	32%	15%
34 Überzeugungskraft	22%	31%	33%	14%
Zitierweise	völlig falsch	eher falsch	eher korrekt	völlig korrekt
35 Korrektheit	20%	12%	47%	20%

AUFGABENVORSCHLAG 1 (Fortsetzung)**Aufgabenstellung 1, 2, 3 und 4: Bewertung der Darstellungsleistungen**

GLIEDERUNG	0 gar nicht	1 nur bedingt	2 gegeben
N=275	Anteil	Anteil	Anteil
36 Klarer Aufbau	10%	43%	47%
37 Funktionale Einleitung	16%	42%	42%
38 Unterteilung in Absätze	30%	39%	31%
39 Zusammenhang in der Gedankenführung	9%	40%	51%
40 Überleitungen	23%	52%	25%
41 Abrundung am Schluss	20%	37%	43%
SPRACHVERWENDUNG: BEGRIFFE	nicht verwendet	nur bedingt verwendet	verwendet
42 Fachbegriffe	45%	41%	14%
43 abstrahierende Begrifflichkeit	39%	41%	20%
SPRACHVERWENDUNG: AUSDRUCK	nie, kaum	zeitweilig	überwiegend
44 rhetorisch gekonnt	28%	51%	21%
45 angemessen/flüssig	17%	47%	37%
46 geschickt	29%	50%	21%
47 umgangssprachlich	51%	38%	11%
48 verkürzt	39%	38%	23%

VERSTÖßE GEGEN DIE SPRACHLICHE RICHTIGKEIT	ANZAHL DER FEHLER		\bar{x}	s
49 Anzahl der Wörter: \bar{x} = 280	50 Rechtschreibung:	7,2	7,5	
s = 140	51 Zeichensetzung:	6,7	5,7	
	52 Grammatik:	3,7	4,0	
	Gesamtfehlerzahl:	18	14	

53 Bitte vergeben Sie abschließend für die Arbeit einen zusammenfassenden Wert zwischen 0 und 9:

sehr schlecht	0	1	2	3	4	5	6	7	8	9	sehr gut
Verteilung:	6%	5%	13%	10%	18%	15%	12%	10%	8%	3%	

\bar{x} = 4,4

s = 2,3

N = 275

AUFGABENVORSCHLAG 2: Erhebungsbogen samt Grundauszählung

Schulnr.		Klassennr.		Schüler/in-
32 Schulen		72 Klassen		1142 SCH

teilgenommen

Der Arbeit liegt die Aufgabenstellung

1 Problemerkörterung	854	75%
2 Leserbrief	224	20%
3 appellative Rede	<u>54</u>	5%
	1.132 ⁵	zugrunde.

⁵ Von den restlichen zehn Arbeiten ließ sich nur noch feststellen, dass sie zur Aufgabenstellung 2 oder 3 gehörten.

Aufgabenstellung 2 und 3 (Leserbrief - appellative Rede)

		0 nicht vorhanden	1 vorhanden, aber unpassend/ ungenau	2 vorhanden und passend/ treffend
N=288		Anteil	Anteil	Anteil
01	Anrede	22%	25%	53%
02	Adressatenbezug	24%	32%	44%
03	Appellative Rede: Kern-/Zielsatz	24%	35%	41%
04	Schlussappell	21%	35%	44%

Erläuternder Kommentar:

Zum Aufgabenvorschlag 2 liegen Arbeiten von 1142 Schüler/innen aus 72 Klassen an 32 Schulen vor. Drei Viertel der Schüler/innen haben sich für die erste Aufgabenstellung *Problemerörterung* entschieden, die restlichen (N=288) für die zweite und dritte, auf die sich der Beginn des Bewertungsbogens bezieht.

Letzte Spalte der Bewertungsaspekte 01 bis 04: Zwischen 41% und 53% der Arbeiten verdienen nach Ansicht der korrigierenden Lehrkräfte die volle Punktzahl. Bei einem Fünftel bis einem Viertel der 288 Arbeiten fehlen (bei einer Aufgabenstellung vom Typus *Leserbrief* bzw. *Appellative Rede*) die Anrede und/oder der Adressatenbezug oder zentrale Elemente wie ein Kern-/Zielsatz und Schlussappell.

Aufgabenstellung 1, 2 und 3 (Problemerörterung - Leserbrief - appellative Rede)

EINLEITUNG		0 nicht vorhanden	1 vorhanden, aber unpassend/ ungenau	2 vorhanden und passend/ treffend
N=1142		Anteil	Anteil	Anteil
05	Darstellung des Problemzusammenhanges	12%	38%	50%
06	Eigenes Beispiel	52%	19%	29%
07	These der Textgrundlage	27%	32%	41%
08	Themafrage: Art	32%	35%	33%
		kurz	mittel	ausführlich
09	Themafrage: Umfang	32%	42%	19%

HAUPTTEIL: Argumentation		0 gar nicht	1 zeitweilig	2 durchgehend
10	Verbindungen von Textwiedergabe, Beispielen, Begründungen, Bewertungen	12%	59%	29%

Handybesitz		keine, wenige	einige	zahlreiche
11	Gründe dafür	18%	53%	28%
12	Gründe dagegen	22%	52%	26%

Textwiedergabe		gar nicht	zeitweilig	durchgehend
13	Konzentration auf Wesentliches	16%	50%	34%
14	Begründungen	16%	56%	28%
15	Bewertungen	20%	56%	24%

SCHLUSS: Mögliche Maßnahmen		gar nicht	zeitweilig	durchgehend
16	Textwiedergabe: Konzentration auf Wesentliches	25%	47%	29%
17	Begründungen	22%	52%	26%
18	Bewertungen	25%	51%	24%
19	Eigene Ratschläge	15%	53%	33%
		nicht vorhanden	vorhanden, aber nicht hergeleitet/ begründet	vorhanden und hergeleitet/ begründet
20	Antwort auf Themafrage.	22%	38%	40%

Erläuternder Kommentar:

Der zweite Teil des Bewertungsbogens (Aspekte 05 bis 20) konzentriert sich auf die inhaltliche Bewältigung der Aufgabe. Eine Betrachtung der letzten Spalte, jener, die das vollständige Erfüllen des Bewertungskriteriums bescheinigt, zeigt, dass bei allen Bewertungsaspekten die maximale Punktzahl von bestenfalls der Hälfte der Schüler/innen erreicht wird (Aspekt 05) und dieser Anteil auf 19% absinken kann (Aspekt 09).

Insgesamt ergibt sich ein Eindruck des "Mittelmaßes" in den Leistungen. Dies allerdings darf nicht überraschen, da an dieser Stelle die Ergebnisse über alle Schularten hinweg betrachtet werden. Es zeigt sich somit auch an dieser Stelle, dass es möglich ist, eine Aufgabe bzw. eine Aufgabenstellung über alle Schularten hinweg den Schüler/innen in Berlin vorzugeben und zu einer angemessenen Abbildung des Leistungsspektrums zu kommen. Wäre dies nämlich nicht gelungen, dann hätten sich asymmetrische Ballungen der Bewertungen entwe-

der im oberen oder im unteren Bereich ergeben müssen.

Aufgabenstellung 1, 2 und 3: Bewertung der Darstellungsleistungen

GLIEDERUNG		0 gar nicht	1 nur bedingt	2 gegeben
N=1142		Anteil	Anteil	Anteil
21	Klarer Aufbau	6%	45%	49%
22	Funktionale Einleitung	12%	41%	47%
23	Unterteilung in Absätze	22%	36%	43%
24	Zusammenhang in der Gedankenführung	4%	42%	54%
25	Überleitungen	12%	54%	34%
26	Abrundung am Schluss	9%	42%	49%

SPRACHVERWENDUNG: BEGRIFFE		nicht verwendet	nur bedingt verwendet	verwendet
27	Fachbegriffe	11%	48%	41%
28	abstrahierende Begrifflichkeit	21%	52%	26%

SPRACHVERWENDUNG: AUSDRUCK		nie, kaum	zeitweilig	überwiegend
29	rhetorisch gekonnt	21%	55%	24%
30	angemessen/flüssig	7%	45%	48%
31	geschickt	21%	51%	29%
32	umgangssprachlich [#]	56%	34%	9%
33	verkürzt [#]	60%	34%	6%

Erläuternder Kommentar:

Verschiedene Bewertungsaspekte der Darstellung bilden den dritten Teil des Bogens. Zunächst fällt auf, dass die Lösungsprozente nahezu durchgängig höher liegen als im zweiten Teil. Dieser dritte Teil wiegt für das Gesamturteil schwerer als der zweite; vgl. Abschnitt 1.5. Eine Betrachtung der Aspekte im Einzelnen zeigt, dass der Block *Gliederung* durchgängig höhere Lösungsprozente aufweist als die beiden anderen zur *Sprachverwendung* (*Begriffe*, *Ausdruck*). Ob in diesen Bereichen die Leistungen der Schüler/innen tatsächlich schlechter oder die Anforderungen der Lehrkräfte höher sind, lässt sich nicht entscheiden.

[#] Zu beachten: Die Aspekte 32 und 33 sind anders gepolt als die übrigen, d.h. hier bedeuten weniger Punkte eine bessere Leistung.

VERSTÖßE GEGEN DIE SPRACHLICHE RICHTIGKEIT		
34 Anzahl der Wörter	:	419 170
ANZAHL DER FEHLER		
35 Rechtschreibung	:	7,4 7,4
36 Zeichensetzung	:	7,4 6,1
37 Grammatik	:	4,5 4,4
Gesamtfehlerzahl	:	19 14

38 Bitte vergeben Sie abschließend für die Arbeit einen zusammenfassenden Wert zwischen 0 und 9:

sehr schlecht	0	1	2	3	4	5	6	7	8	9	sehr gut
Verteilung:	0,4%	2%	7%	11%	17%	22%	19%	14%	7%	2%	

$$\bar{x} = 4,4 \quad s = 2,3 \quad N = 1.142$$

Erläuternder Kommentar:

Der letzte Teil des Bewertungsbogens umfasst den quantitativen Aspekt des Gesamtumfanges der Arbeit und der Fehlerzahlen unterschieden nach Fehlerarten sowie die zusammenfassende Bewertung auf der Skala von 0 bis 9.

Die Angaben der Lehrkräfte zu den Fehlern waren - wie bei anderen Bewertungsaspekten auch; vgl. Abschnitt 5.6.4 des Anhangs - nicht immer vollständig. Auffällig ist die ziemlich hohe Streuung ($s=170$) bei einem Mittelwert von knapp 420 Wörtern für den Umfang der Arbeiten. Hier schlägt sich die außerordentliche Heterogenität der Schülerschaft nieder. (Laut Angaben auf den Bewertungsbögen besteht die kürzeste Arbeit aus 36 Wörtern, die längste aus 1800.) Orthographie und Interpunktion werden häufiger bemängelt als grammatikalische Fehler.

3.3 Aufgabenvorschlag 2: Schulartspezifische Ergebnisse bei den einzelnen Bewertungsaspekten

Nachstehend finden sich für jeden der Bewertungsaspekte des Aufgabenvorschlags 2 schulartspezifische Ergebnisse⁶. Dabei fand eine Beschränkung und Zuspitzung auf die Frage statt, wie viele der Arbeiten die maximal mögliche Punktzahl erhalten. Aufgeführt sind also die prozentualen Anteile der Arbeiten, denen von den korrigierenden Lehrkräften das vollständige Erfüllen des jeweiligen Bewertungskriteriums bescheinigt wurde.

Aufgabenstellung 2 und 3 (Leserbrief - appellative Rede)

N=263		gesamt	OH	OR	OG	O
01	Anrede	53%	35%	54%	62%	49%
02	Adressatenbezug	44%	31%	40%	54%	39%
03	Appellative Rede: Kern-/Zielsatz	41%	43%	44%	43%	25%
04	Schlussappell	44%	35%	59%	41%	50%

Erläuternder Kommentar zu den Bewertungsaspekten 01 bis 04:

Die Aufgabenvarianten 2 und 3 wurden nur von N=263 Schüler/innen gewählt. Der Anforderung passender Anrede und treffenden Adressatenbezuges konnten in keiner Schulart alle Schüler/innen genügen. Die Anteile liegen überall unter 100%, auch jene im Gymnasium. Es ergibt sich über alle vier Bewertungsaspekte hinweg kein einheitliches Bild. Die zu erwartenden Anordnung vom Gymnasium bis zur Hauptschule liegt in diesem Bewertungsblock nicht vor.

⁶ Ohne die nicht repräsentativen 59 Arbeiten aus dem berufsbildenden Bereich.

Aufgabenstellung 1, 2 und 3 (Problemerkörterung - Leserbrief - appellative Rede)

N=1083	gesamt	OH	OR	OG	O	
EINLEITUNG						
05	Darstellung des Problemzusammenhanges	50%	30%	39%	58%	59%
06	Eigenes Beispiel	29%	12%	20%	36%	35%
07	These der Textgrundlage	41%	20%	34%	47%	50%
08	Themafrage: Art	33%	10%	20%	43%	34%
09	Themafrage: Umfang	19%	11%	10%	25%	22%
HAUPTTEIL:						
Argumentation						
10	Verbindungen von Textwiedergabe, Beispielen, Begründungen, Bewertungen	29%	16%	21%	39%	23%
Handybesitz						
11	Gründe dafür	28%	10%	20%	38%	25%
12	Gründe dagegen	26%	13%	14%	36%	25%
Textwiedergabe						
13	Konzentration auf Wesentliches	34%	28%	28%	42%	26%
14	Begründungen	28%	16%	16%	40%	23%
15	Bewertungen	24%	18%	14%	34%	16%
SCHLUSS: Mögliche Maßnahmen						
16	Textwiedergabe: Konzentration auf Wesentliches	29%	18%	17%	37%	28%
17	Begründungen	26%	12%	15%	36%	21%
18	Bewertungen	24%	12%	14%	35%	14%
19	Eigene Ratschläge	33%	18%	26%	39%	28%
20	Antwort auf Themafrage.	40%	24%	35%	47%	40%

Erläuternder Kommentar zu den Bewertungsaspekten 05 bis 20:

In diesem Block, der die inhaltliche Auseinandersetzung mit der Aufgabenstellung thematisiert, findet sich als Grundmuster die Anordnung OG > O > OR > OH. Dabei können die Werte der Gesamtschule zuweilen die des Gymnasiums erreichen (05, 06, 07, 09 - diese Bewertungsaspekte beziehen sich ausschließlich auf die Einleitung), die der Realschule (10, 13, 15, 18, 19) oder der Hauptschule (15, 18), was bereits andeutet, dass es - zumindest in dieser Arbeit - auf unterschiedlich hohen Niveaus schulartspezifische Stärken und Schwächen gibt, die genauer zu bestimmen Aufgabe künftiger Vergleichsarbeiten sein wird, die besser als die jetzige eine Funktion als Diagnoseinstrument erfüllen können.

Es gibt Bewertungsaspekte, bei denen relativ große Anteile der Hauptschüler/innen maximale Punktzahlen erreichen (05: 30%, 13:28%, 20: 24%). Elementare Anforderungen inhaltlicher Auseinandersetzung (z.B.: *Konzentration auf Wesentliches* (13)), heißt dies, werden

auch von einem substantziellen Teil der Hauptschüler/innen bewältigt. Zugleich werden die Grenzen deutlich, wenn ein zentraler Aspekt, Argumente für und wider zu finden (hier hinsichtlich des Handybesitzes, Punkte 11 und 12), nur ein Zehntel der Hauptschüler/innen das Kriterium vollständig erfüllen.

An dieser Stelle werden bedeutsame Schwächen in allen Schularten deutlich. Hinsichtlich einer argumentativen Auseinandersetzung erzielen nicht einmal im Gymnasium 40% der Schüler/Innen die maximale Punktzahl. Geringe Prozentsätze finden wir ebenfalls für die inhaltliche Bewältigung des Schlussteiles (Aspekte 16 bis 20).

Aufgabenstellung 1, 2 und 3: Bewertung der Darstellungsleistungen

N=1083		gesamt	OH	OR	OG	O
GLIEDERUNG						
21	Klarer Aufbau	49%	22%	41%	59%	52%
22	Funktionale Einleitung	47%	34%	30%	54%	56%
23	Unterteilung in Absätze	43%	25%	32%	55%	33%
24	Zusammenhang in der Gedankenführung	54%	41%	44%	64%	52%
25	Überleitungen	34%	20%	33%	40%	32%
26	Abrundung am Schluss	49%	42%	45%	55%	46%
SPRACHVERWENDUNG: BEGRIFFE						
27	Fachbegriffe	41%	32%	29%	47%	46%
28	abstrahierende Begrifflichkeit	26%	9%	12%	39%	18%
SPRACHVERWENDUNG: AUSDRUCK						
29	rhetorisch gekonnt	24%	9%	17%	33%	18%
30	angemessen/flüssig	48%	34%	29%	63%	43%
31	geschickt	29%	15%	14%	41%	18%
32	umgangssprachlich [#]	56%	24%	49%	69%	50%
33	verkürzt [#]	60%	41%	52%	72%	56%

Erläuternder Kommentar zu den Bewertungsaspekten 21 bis 33:

Zunächst fällt auf, dass als Schulartkonstellation wiederum dasselbe Muster wie eben vorherrscht: OG>O>OR>OH. Ferner liegen die Anteile der Arbeiten, die die maximale Punktzahl erreicht haben, in diesem Block, der der Darstellung gewidmet ist, generell höher liegen als im zweiten, dem inhaltlichen Block. Eine Erklärung hierfür lässt sich aus dem ableiten, was im Abschnitt 1.3 dargestellt wurde: Die Kriterien der Darstellungsleistung spielen tendenziell eine größere Rolle für die Lehrkräfte, wenn diese sich einen Gesamteindruck von den Arbeiten bilden. Wenn dem so ist, dann werden sie auch ihren Unterricht verstärkt daran ausricht-

[#] Da die Bewertungsaspekte 32 und 33 so gepolt sind, dass hohe Werte schlechte Leistungen bedeuten, werden die Prozentanteile der niedrigsten Ausprägung *nie, kaum* angegeben.

ten, also z.B. der Darstellung mehr Zeit einräumen und mehr üben, was sich wiederum in besseren Leistungen niederschlägt.

Bemerkenswert sind jene Bewertungsaspekte, bei denen ein anderer Schularten vergleichbarer Anteil an Hauptschülern/innen die maximale Punktzahl erreicht: 22 (*Funktionale Einleitung*, 34% - OR: 30%), 24 (*Zusammenhang in der Gedankenführung*, 41% - OR: 44%), 26 (*Abrundung am Schluss*, 42% - OR: 45%), 27 (*Verwendung von Fachbegriffen*, 32% - OR: 29%), 30 (*Ausdruck angemessen/flüssig*, 34% - OR: 29%). Die im Gegensatz hierzu weit stärkere Verwendung umgangssprachlicher Ausdrücke (Bewertungsaspekt 32) in den Arbeiten aus der Hauptschule als aus den anderen Schularten ist eine sehr plausible Abrundung des Bildes, das dieser Bewertungsblock liefert.

gesamt OH OR OG O

VERSTÖßE GEGEN DIE SPRACHLICHE RICHTIGKEIT						
34	Anzahl der Wörter:	428	241	384	481	414
ANZAHL DER FEHLER						
35	Rechtschreibung:	7,3	12	9	5	9
36	Zeichensetzung:	7,4	9	9	6	9
37	Grammatik:	4,5	6	5	4	6
Gesamtfehlerzahl:		19	27	24	15	23
Fehlerquotient:		5%	12%	6%	3%	6%

38 Zusammenfassende Bewertung (Globalurteil)

sehr schlecht

sehr gut.

0 1 2 3 4 5 6 7 8 9

gesamt	∅ =	4,4	s =	2,3	N =	1059
OH		3,9		2,0		92
OR		4,8		1,6		240
OG		5,4		1,8		568
O		4,9		1,6		159

Erläuternder Kommentar zu den Bewertungsaspekten 34 bis 38:

Umfang der Arbeiten sowie Fehlerzahl und Fehlerquotient weisen die zu erwartenden schulartspezifischen und teilweise erheblichen Unterschiede auf. Erkennbar wird, dass es auch

noch in der zehnten Klasse Schwierigkeiten in der formalen Bewältigung des Schreibens gibt. Hier nicht dokumentiert, aber zu betonen ist ein zweifacher Umstand: Zum einen bestehen erhebliche Unterschiede zwischen den Schülern/innen - so schwankt die Gesamtzahl der Fehler zwischen 0 und 101 - zum anderen sind die Verteilungen der Fehlerzahlen glücklicherweise nicht symmetrisch, sondern linkssteil, d.h. es gibt mehr Arbeiten mit weniger Fehlerzahlen als umgekehrt.

3.4 Wie tauglich sind die Kategorien des Bewertungsbogens?

Von besonderem Interesse sind die Bewertungsaspekte, zu denen viele der Lehrkräfte keine Angaben gemacht haben. Die Gründe sind nicht bekannt, es darf aber vermutet werden, dass eine hohe Anzahl *missing data* auf Defizite verweist: Ein solcher Bewertungsaspekt ist entweder von der Sache her nicht angemessen oder gar irrelevant oder er wurde unscharf formuliert, so dass eine bewertende Lehrkraft im Unklaren sein muss, was eigentlich gemeint ist.

In der folgenden *Tabelle 12* werden derartige Items zusammengestellt, um daraus Hinweise abzuleiten, wie ein Bewertungsbogen verbessert werden kann. Es werden alle Bewertungsaspekte mit einem Anteil fehlender Angaben berücksichtigt, der über 5% der jeweils maximal möglichen gültigen Angaben liegt. Das sind für den Aufgabenvorschlag 2 im ersten Bewertungsblock N=288 und ansonsten N=1.142.

Tabelle 12: *Bewertungsaspekte mit einem Anteil fehlender Angaben von über 5% der maximal möglichen gültigen Angaben.*

Alle drei Aufgabenstellungen		
N=1142		Anzahl missing data
EINLEITUNG		
08	Themafrage: Art	145
09	Themafrage: Umfang	186
SCHLUSS: Mögliche Maßnahmen		
	Textwiedergabe: ⁷	48
16	Konzentration auf Wesentliches	
17	Begründungen	47
18	Bewertungen	53
19	Eigene Ratschläge	46
20	Antwort auf Themafrage.	79
SPRACHVERWENDUNG: AUSDRUCK		
33	verkürzt	88

⁷ Für die Bewertungsaspekte 16 bis 19 wird das 5%-Kriterium zwar verfehlt, denn 5% von 1.142 sind etwa 57, so dass alle missing data-Anzahlen von 46 bis 53 darunter liegen, aber die Durchgängigkeit, mit der in diesem Teil relativ viele Angaben fehlen, lässt eine Dokumentation als angezeigt erscheinen.

Im Einzelfall ist schwer zu entscheiden, warum bei einem Bewertungsaspekt hohe Ausfälle auftreten. Ist es die ungenaue, die verkürzte und diffuse Formulierung oder wurde zwar das Bewertungskriterium erkannt, aber für die Aufgabenstellung der Arbeit als irrelevant oder nicht adäquat empfunden? Oder - dritte Möglichkeit - liegt es an den Antwortabstufungen, an deren Inhalt und an der zu geringen oder zu großen Anzahl? Wir können nur Vermutungen anstellen, die sich allerdings auf eine Analyse der Aufgabenstellung gründen.

Irritierend ist eine grundsätzliche inhaltliche Ungenauigkeit; die Aufgabenstellung gibt einen anderen Schwerpunkt als der Text vor: so stehen die Notwendigkeit des Handys (Thema) dem Verschuldungsproblem bei Jugendlichen (Zeitungsartikel) gegenüber. Die Bearbeitung der Aufgabe und Bewertung der Arbeit spiegeln diese Unsicherheit.

Beispiel: Im Erhebungsbogen (08) ist von der *Art der Themafrage* die Rede; die Formulierung erzeugt bei der Bearbeitung Unklarheiten: Ist die "Art" des Themas oder die "Art" der Frage gemeint? Von daher dürfte auch die Bewertung des Aspektes 20 *Antwort auf die Themafrage* schwierig gewesen sein. Ebenso ist auch die Funktion der *Textwiedergabe* am Schluss (16, 17, 18) unklar, da der Begriff doppelt auftritt (vgl. die Bewertungsaspekte 13, 14, 15) und somit an Trennschärfe für die Bearbeitung verliert.

Weniger klar allerdings ist, warum der Bewertungsaspekt 19 *Eigene Ratschläge* relativ häufig ausgelassen wurde. Vielleicht bestanden hier Unsicherheiten, nach welchen Kriterien Ratschläge der Schüler/innen zu bewerten seien. Ähnliches dürfte auch für den Aspekt 33 *Verwendung verkürzter Ausdrücke* gelten.

Diese kurze Übersicht belegt, dass bei dem nächsten Durchgang der Vergleichsarbeiten noch Anstrengungen unternommen werden müssen, um die Eindeutigkeit und Verständlichkeit der Aufgabenstellung und der Bewertungskriterien zu sichern. Zu prüfen ist in diesem Zusammenhang, wie weit drei oder vier Abstufungen für die Bewertung ausreichend sind.

C ZUSAMMENFASSUNG DER WICHTIGSTEN ERGEBNISSE IN DEN FÄCHERN ENGLISCH, FRANZÖSISCH UND MATHEMATIK

C 1 Englisch

Teilnehmerzahlen

Von den 142 Schulen, die sich für die Teilnahme am Probelauf angemeldet hatten und denen das Material für die Vergleichsarbeiten zur Verfügung gestellt wurde, übersandte die folgende Anzahl an Schulen ihre Schülertestbögen zur Auswertung zurück:

Schulform	Schulen	Klassen	Schülerinnen und Schüler
Gymnasien	41	150	3.363
Realschulen	28	87	1.763
Hauptschulen	18	52	769
Gesamtschulen	30	163	2.895
Berufsfachschulen	10	34	410
Gesamt	124	485	9.200

Aufgaben

Die für den Test ausgewählten Aufgaben wurden von Cambridge-ESOL, dem Prüfungssyndikat der Universität Cambridge zur Verfügung gestellt. Diese Aufgaben entsprechen internationalen Standards für einen Mittleren Schulabschluss. Sie erfüllen die Anforderungen des Gemeinsamen europäischen Referenzrahmens für Sprachen, auf dessen Grundlage die Kultusministerkonferenz Bildungsstandards für den Mittleren Schulabschluss entwickelt hat. Zur Verdeutlichung der Standards wurden auch auf KMK-Ebene Aufgabenbeispiele aus den Cambridge-Tests ausgewählt.

Der Test im Schuljahr 2002/2003 umfasste die Bereiche Hörverstehen, Leseverstehen und Schreiben. Ein Probelauf "Mündliche Überprüfung des Kompetenzbereiches Sprechen" ist für die kommende Vergleichsarbeit vorgesehen.

Die Testteilnehmerinnen und -teilnehmer wurden mit Texten konfrontiert, die Menschen in der Regel bei einem Auslandsaufenthalt begegnen und die eine Basis für die Bewältigung kommunikativer und lebenswichtiger Situationen bieten: Hinweisschilder, Zeitungsannoncen, Klappentexte, mündliche und schriftliche Interessensbekundungen, mündliche Ausschnitte aus Museums- bzw. Stadtführungen, Werbebroschüren, Briefe, auf Anrufbeantworter gesprochene Texte, Telefonanrufe, Anmeldeformulare, Fragebogen, etc.

Auswertung

Allen Schulen war bekannt, dass Cambridge ESOL, das die Aufgaben zur Verfügung gestellt hat, einen Test dann als bestanden wertet, wenn mindestens 70% der Aufgaben korrekt gelöst sind. Darüber hinaus wussten die Schulen, dass die Grenze bei geschlossenen Aufgabenformaten (Multiple Choice, Matching, Lückentext etc.) mindestens bei 66% der erbrachten Leistungen liegt. In den KMK-Bildungsstandards gilt ein Standard bei weitgehend geschlossenen Aufgaben dann als erreicht, wenn 2/3 der Aufgaben korrekt gelöst werden. Auch zu der produktiven Schreibaufgabe gab es Kriterien für die inhaltliche und sprachliche Leistungsfeststellung. In sechs Stufen wurden jeweils Kriterien für die Bewertung der inhaltlichen und der sprachlichen Leistung der Lernenden benannt.

Alle Aufgaben außer der produktiven Schreibaufgabe (Verfassen eines Briefes) wurden komplett maschinell ausgewertet. Bei letzterer wurde die Bewertung der Lehrerinnen und Lehrer zusammengefasst und der Mittelwert errechnet.

Folgende Ergebnisse liegen nach Auswertung der Tests vor:

Schulen insgesamt (alle Aufgaben außer der produktiven Schreibaufgabe):

Prozentzahl richtig gelöster Aufgaben Mindestwerte in %	Schüler					
	Gymn. %	Realsch. %	Hauptsch. %	Gesamtsch. %	Fachobersch. %	Insgesamt %
Mehr als 50	8,1	76,4	32,3	56,4	42,5	73,8
Mehr als 60	94,2	52,8	15,6	36,3	20,6	59,2
Mehr als 70	83,0	31,8	7,4	20,2	10,9	44,7

Schulen differenziert nach Niveaustufen (alle Aufgaben außer der produktiven Schreibaufgabe):

Prozentzahl richtig gelöster Aufgaben Mindestwerte in %	Schüler						
	Hauptschule			Gesamtschule			
	Niveau	Niveau	Niveau	Niveau	Niveau	Niveau	Niveau
	A %	B %	C %	F %	E %	G %	A %
Mehr als 50	37,1	7,4	0,0	96,6	78,2	34,6	10,5
Mehr als 60	16,1	0,6	0,0	93,1	54,7	14,0	0,0
Mehr als 70	5,6	0,6	0,0	81,6	31,4	4,4	0,0

Das Ergebnis belegt, dass nur im Bereich der Gymnasien und der F-Kurse der Gesamtschulen sowohl die 60%- als auch die 70%-Hürde von einer hohen Anzahl an Schülerinnen und Schülern bewältigt wurde. Es gibt eindeutige Parallelen zwischen den einzelnen Gesamtschulniveaus und den Schulformen im gegliederten Schulsystem:

Die Ergebnisse der Gymnasien sind vergleichbar mit denen der F-Kurse der Gesamtschulen, die Realschulen liegen sehr dicht bei denen der E-Kurse der Gesamtschulen. Die Schüler der A-Kurse der Hauptschulen und der G-Kurse der Gesamtschulen erbrachten zwar keine mit den eben aufgeführten Teilnehmergruppen vergleichbaren Leistungen; sie zeigten sich

aber deutlich leistungsstärker als die Lerner in den B- und C-Kursen der Hauptschulen und in den A-Kursen der Gesamtschulen. Parallelen zeigen sich auch zwischen den Ergebnissen der B-Kurse der Hauptschulen und denen der A-Kurse an Gesamtschulen. Die größten Schwierigkeiten mit dem Test hatten die Hauptschüler der C-Kurse.

Bei der Auswertung der produktiven Schreibaufgabe ergaben sich folgende Mittelwerte:

Inhalt:		0 - 5 Punkte	Sprachverwendung:	
Schulform	Mittelwert		Schulform	Mittelwert
Gymnasien	4,8		Gymnasien	4,3
Realschulen	4,0		Realschulen	3,4
Hauptschulen	2,5		Hauptschulen	2,3
Gesamtschulen	3,6		Gesamtschulen	3,1
Berufsfachschulen	2,9		Berufsfachschulen	2,6

Besonders im Bereich des Inhalts zeigten auch die Realschulen und Gesamtschulen bessere Durchschnittswerte (hier kann vom Erreichen des Mindeststandard bei 50% ausgegangen werden) als bei den oben angeführten Aufgaben.

Schlussfolgerungen

Die zwischen den einzelnen Schulformen und den Niveaustufen innerhalb der Haupt- und Gesamtschulen stark divergierenden Ergebnisse der Vergleichsarbeiten belegen die derzeit bestehende Problematik bei der Durchführung schulformübergreifender Tests am Ende eines Bildungsganges.

Die Aufgaben entsprechen internationalen Standards für einen Mittleren Schulabschluss. Sie erfüllen die Anforderungen des Gemeinsamen europäischen Referenzrahmens für Sprachen, auf dessen Grundlage die Kultusministerkonferenz Bildungsstandards für den Mittleren Schulabschluss entwickelt hat.

Will man nicht hinnehmen, dass in einzelnen Schulformen die Hälfte bzw. mehr als die Hälfte der Schülerinnen und Schüler unter den erwarteten Leistungen bleibt, so müssen dringend Maßnahmen zur Unterrichtsentwicklung ergriffen, Schülerinnen und Schüler mit den neuen Aufgabenformaten vertraut gemacht und Lernprozesse mit nachhaltiger Wirkung initiiert werden.

Erfolgreich lernen können ganz besonders schwache Schülerinnen und Schüler nur, wenn Lehrkräfte, aber auch Schülerinnen und Schüler selbst befähigt werden, Lernschwächen zu erkennen und zu beheben. Hierzu sind neben standardorientierten Rahmenlehrplänen auch neue Unterrichtsmethoden und gezielte Fortbildungsmaßnahmen für Lehrerinnen und Lehrer vonnöten.

Für die kommenden Vergleichsarbeiten, die vom LISUM entwickelt werden, erhalten die Schulen Übungsmaterialien und Hinweise, wie sie ihrer Schülerinnen und Schüler auf das Erreichen der Standards für den Mittleren Schulabschluss vorbereiten können.

C 2 Französisch

Aufgaben

Die für den Text ausgewählten Aufgaben wurden dem Testprogramm für Sprachenzertifikate DELF (Diplôme d'études en langue française) entnommen. Die DELF-Zertifikate sind international anerkannte Sprachdiplome für Französisch als Fremdsprache, die vom staatlichen Institut CiEP entwickelt werden. Die Aufgaben erfüllen die Anforderungen des Gemeinsamen europäischen Referenzrahmens für Sprachen, auf dessen Grundlage die Bildungsstandards für den Mittleren Schulabschluss der KMK festgelegt wurden.

Die erstmals zentral gestellten Aufgaben in den Vergleichsarbeiten überprüften die Fertigkeiten im Hörverstehen, Leseverstehen und Schreiben. Eine mündliche Überprüfung des Bereichs Sprechen wird probeweise für die nächste Vergleichsarbeit vorgesehen. Bei der Auswahl der Textvorlagen war die Funktionalität in Bezug auf die Bewältigung von Herausforderungen im Alltag, im Beruf und in der Schule ein wesentliches Kriterium.

Teilnehmer

An dem zweiten Probelauf haben 15 Oberschulen teilgenommen, von denen 11 Gymnasien waren. Diese fast ausschließlich auf den Gymnasialbereich beschränkte Teilnahme hat zur Folge, dass die Auswertung ebenfalls nur Schlussfolgerungen für diese Schulform zulässt. Darüber hinaus sind aufgrund der geringen Beteiligung insgesamt auch die Ergebnisse nicht als repräsentativ anzusehen.

Auswertung

Die Auswertung berücksichtigte eine Differenzierung in die Lerngruppen erste und zweite Fremdsprache.

Im Hörverstehen wiesen die Schülerinnen und Schüler die besten Ergebnisse auf. Die geforderte Punktzahl wurde im Durchschnitt zu 68 %, in Teilbereichen zu 90 % erreicht. Insbesondere konnten die Schülerinnen und Schüler gut mit den Zuordnungs- und Auswahlaufgaben umgehen. Im Leseverstehen hatten die Schülerinnen und Schüler Schwierigkeiten mit dem verwendeten Vokabular, so dass der Textzusammenhang nicht immer verstanden wurde. Die Punktzahl wurde nur zu 45 % erreicht. Die produktive Schreibaufgabe ist, gemessen an der zu erreichenden Punktzahl, von der Hälfte (53 %) der Schülerinnen und Schüler erfolgreich bewältigt worden. Die konkrete Schreibeinzelleistung ist bei der Auswertung allerdings nicht erfasst worden und ist durch Stichproben der Arbeiten nachzuholen. Eine auffällige Unterscheidung nach 1. und 2. Fremdsprache konnte bei der Schreibleistung nicht festgestellt werden.

Schlussfolgerungen und Empfehlungen

Das Anspruchsniveau und die Formate der Aufgaben entsprechen dem Standard der KMK-Vorgaben und damit auch dem Gemeinsamen europäischen Referenzrahmen. Die

gewählten Aufgaben von DELF hätten allerdings abwechslungsreicher und vielfältiger in der Gestaltung sein können und thematisch eine stärkere Schülerorientierung haben müssen. Den Schülerinnen und Schülern hätte dadurch ein größerer Spielraum im sprachlichen Handeln gegeben werden können. Das ist im Hinblick auf die bisher nicht berücksichtigten Real- und Gesamtschulen relevant.

Bei einer Gegenüberstellung der rezeptiven und produktiven Leistungen der Schülerinnen und Schüler zeichnen sich bessere Leistungen im rezeptiven Bereich ab. Allerdings ist die geforderte Schreibleistung nicht umfassend genug. Darüber hinaus müsste die Vergabe der Punkte auf die konkrete Schreibleistung (Ausdrucksfähigkeit, Textproduktion, Wortschatz, Grammatik) durch Stichproben der Arbeiten analysiert werden.

Ferner wäre empfehlenswert, bei der Auswahl der Aufgaben eine enge Abstimmung mit den Aufgabenformaten der englischen Vergleichsarbeiten vorzunehmen, um vergleichbare Anforderungsniveaus zu sichern.

C 3 Mathematik

Teilnehmerzahlen

Die Teilnahme an den Vergleichsarbeiten war freiwillig. Die Anzahl der Schulen, die ihre Ergebnisse bzw. die Schülerarbeiten zur weiteren Auswertung versandten, deckte sich nicht genau mit der Zahl der Anmeldungen. Insgesamt waren die Gymnasien über- und die Haupt- und Gesamtschulen unterrepräsentiert, die Teilnehmerquote ergibt jedoch - Ausnahme: die Berufsfachschulen - eine solide Datengrundlage.

Schulform	Schüler/innen im 10. Jahrgang in Berlin 2002/03		ausgewertete Teilnahme an der Vergleichsarbeit Mathematik				
	absolut	Anteil in %	Schulen	Klassen	Schüler/innen	Anteil an der Grundgesamtheit in %	Anteil der Teilnehmer in %
OH	3 438	10	14	26	359	10,6	7,0
OR	7 364	23	21	45	1 104	15,0	21,6
OG	11 606	36	28	88	2 230	19,2	43,6
O/OG	10 113	31	16	77	1 377	13,6	27,0
OBF	--	--	2	4	49	--	1,0
Gesamt	32 521	100	81	240	5 119	15,7	100

Aufgaben

Die Arbeit bestand aus 19 Aufgaben mit einer Bearbeitungszeit von 90 Min. und war in zwei Teile eingeteilt: Teil 1 (Aufgaben 1 bis 11) ohne Hilfsmittel mit 20 Min. Bearbeitungszeit und Teil 2 (Aufgaben 12 bis 19) mit Taschenrechner für 70 Minuten Bearbeitungszeit. Eine Formelsammlung war nicht zugelassen.

Die Themenfelder der Aufgaben bezogen sich nicht nur auf Inhalte des Rahmenplans der Klassenstufe 10, sondern deckten auch Kernbereiche der Inhalte der Klassenstufen 7 bis 9 ab. Der Schwerpunkt lag rahmenplangemäß in Arithmetik und Algebra (13 Aufgaben).

Es lag eine breite Streuung der Schwierigkeiten vor. Zur besseren Validierung der Kompetenzstufen der Aufgaben enthielt der Test fünf freie PISA-Aufgaben. Die gewählten Aufgabenformate waren ansonsten eher traditionell. Der weit überwiegende Teil der Aufgaben bezog sich auf das Niveau I des Rahmenplans. Bei Niveau-II-Aufgaben fand sich durchgängig ein höherer Modellierungsanteil. Alle Aufgaben wiesen einen nur geringen Textanteil auf. Als Kriterium für eine kompetenzorientierte Einordnung der Aufgaben und die Zusammenstellung der Arbeit wurde das Kompetenzmodell nach PISA-Framework 2000 benutzt.

Die beigelegte Musterlösung folgte bei der Bepunktung der Vorgehensweise bei Klassenarbeiten, d. h. die Bewertungseinheiten wurden etwa proportional zum angenommenen Zeitaufwand verteilt. Die Korrektur führten gemäß der Musterlösung die

jeweiligen Fachlehrer durch. Auf den Auswertungsbögen wurde nur die Gesamtpunktzahl pro Aufgabe erfasst, so dass Detailfragen zur Vergabe einzelner Punkte, also Fragen bzgl. Teilkompetenzen innerhalb von Aufgaben, unbeantwortet blieben. Die zum Zeitpunkt der endgültigen Zusammenstellung der Arbeit vorliegende Entwurfsfassung der KMK-Standards, insb. das Konzept der Leitideen, wurde berücksichtigt.

Auswertung

Das federführende Referat I D der Senatsverwaltung für Bildung, Jugend und Sport hat Frau R. Nikolova, Mitarbeiterin am Institut für Erziehungswissenschaften der Humboldt-Universität Berlin - Abteilung Empirische Bildungsforschung, Prof. Dr. Dr. Rainer Lehmann - beauftragt, eine Auswertung der anonym erhobenen Daten aus den Vergleichsarbeiten vorzunehmen. Ihr Bericht ist Grundlage der Ergebnispräsentation. Die anderen Berichtsteile für Mathematik wurden von Herrn C. Bänsch, I D 7 verfasst.

Die Zweitkorrektur einer Stichprobe von ca. 250 Arbeiten durch das Entwicklerteam ergab, dass die Bepunktung der Schülerarbeiten durch die Fachlehrer eine verlässliche Datengrundlage darstellt.

Auf Grund der statistischen Auswertung ergibt sich eine Fähigkeitsskala, auf der sich die 19 Aufgaben gut in fünf abstrakte Kompetenzstufen einteilen lassen. Diese korrelieren in der Abfolge hinreichend gut mit den PISA-Kompetenzstufen. Das zeigt, dass die theoretische Einschätzung der Aufgaben und ihre Zusammenstellung im Test geeignet sind, Schülerkompetenzen adäquat zu überprüfen.

Die Zusammenstellung der Aufgaben in der Arbeit bildet das Kompetenzspektrum der Teilnehmer gut ab. Alle Kompetenzstufen sind vertreten, auch die Leistungsspitze wird erfasst. Die meisten Aufgaben repräsentieren die Kompetenzstufen III und IV, d. h. die Arbeit war für einen mittleren Bildungsabschluss, bezogen auf den Leistungsstand der beteiligten Schüler/innen, etwas zu schwer. Mindeststandards wurden etwas zu wenig berücksichtigt.

Das statistische Ergebnis ist annähernd eine Normalverteilung mit einem Mittelwert, der ca. 37 von den 70 maximal möglichen Bewertungseinheiten der Arbeit entspricht. 5 % der Teilnehmer erreichten bis zu 10 % der möglichen Punkte, insgesamt 25 % erreichten bis zu 30 %, weitere 25 % bis zu 50 %, die dritten 25 % bis zu 70 % der erreichbaren Punkte. Die obersten 25 % der Teilnehmer erreichten mehr als 70 % der erreichbaren Punkte.

Die Risikogruppe der Zehntklässler/innen, die mit ihren mathematischen Fähigkeiten unter dem Niveau der Kompetenzstufe I liegen, umfasst 8 % der Teilnehmer. Diese sind nicht mit hinreichender Wahrscheinlichkeit in der Lage, Kompetenzstufe I zu erreichen. Die Zahl der Schülerinnen und Schüler, die mit einiger Sicherheit Aufgaben der Kompetenzstufe I, nicht aber schwierigere Aufgaben bewältigen können, beträgt 11 %.

Die Gruppe der Schüler/innen, die Kompetenzstufe II erreichen, umfasst 23 % der Stichprobe. Die Leistungen von 29 % der Teilnehmer/innen entsprechen der Kompetenzstufe III. 16 % sind in der Lage, Kompetenzstufe IV zu erreichen, und 7 % der Stichprobe sind Schülerinnen und Schüler, die sich auf Kompetenzstufe V befinden. Diese besitzen besonderes Wissen im Fach Mathematik.

Setzt man Kompetenzstufe III als Regelstandard, so erreichen bzw. überschreiten insgesamt 52 % der Zehntklässlerinnen und Zehntklässler dieses Anforderungsniveau.

Bei den Gesamtschulen unterscheiden sich getrennte G- und A-Kurse nur unwesentlich von GA-Kursen, ebenso wenig F- oder E-Kurse von FE-Kursen. Bei den Hauptschulen gibt es große Überschneidungen zwischen A-, B- und C-Kursen. Die Gymnasialleistungen liegen sehr deutlich über denen der anderen Schulformen. Signifikante geschlechtsspezifische Unterschiede traten nicht zu Tage.

In der nachfolgenden Tabelle wird schulform- bzw. leistungsniveauspezifisch angegeben, wie viele Schülerinnen und Schüler die mit ausgewählten Skalenwerten aus den fünf abstrakten Kompetenzstufen markierten Fähigkeitsniveaus mit 50-prozentiger Wahrscheinlichkeit erreichen bzw. übertreffen. Weiterhin sind die schulformbezogenen Mittelwerte auf der abstrakten Fähigkeitskala angegeben.

Fähigkeitsniveau bzw. abstrakte Kompetenzstufe (KS)						Mittelwert auf der Gesamt-Fähigkeitsskala
Schulform	KS I (untere Grenze)	KS II (oberer Grenze)	KS III (Mitte)	KS IV (Mitte)	KS V (untere Grenze)	
Hauptschulen, A	79 %	23 %	7 %	0 %	0 %	84,4
B	50 %	13 %	9 %	2 %	0 %	72,7
Realschulen	97 %	59 %	33 %	7 %	4 %	99,8
Gymnasien	99 %	82 %	63 %	24 %	14 %	110,4
Gesamtschulen, FE	98 %	60 %	34 %	7 %	3 %	100,2
GA	75 %	17 %	9 %	2 %	0 %	82,0
gesamt	92 %	60 %	40 %	13 %	7 %	100

Wegen der geringen Beteiligung sind die C-Kurse der Hauptschulen weggelassen. Wegen der geringen Abweichungen sind die Gesamtschulkurse zu GA- bzw. FE-Niveau zusammengefasst.

Schlussfolgerungen und Empfehlungen

Wir wissen jetzt, dass eine schulformübergreifende Arbeit in Mathematik auf angemessenem Niveau für eine Prüfung zu einem Mittleren Schulabschluss machbar ist. Die Arbeit muss, wenn man die Grenze für das Bestehen bei dieser Komponente eines Mittleren Schulabschlusses etwa in dem Bereich ansetzen möchte, der auch bei einer Klassenarbeit gilt (45 – 50 %), zumindest im Moment noch etwas leichter und etwas kürzer sein als die geschriebene. Ein größerer Anteil der Arbeit sollte sich auf Mindest- und nicht auf Regelstandards nach KMK beziehen.

Die eher geringen Lösungshäufigkeiten bei Aufgaben zu Basiswissen wie m1, m2, m3 und m5 legen den Schluss nahe, dass im Unterricht noch mehr darauf geachtet werden könnte, Basiskompetenzen auf Grund weiter zurück liegender Inhaltsbereiche stärker in neue Sachzusammenhänge wiederholend einzubinden.

Von den drei Hauptdomänen Algebra, Geometrie, Berufsvorbereitung war die Algebra überrepräsentiert, so wie im derzeitigen Rahmenplan. Eine weiter gehende Beschränkung auf eine Hauptdomäne oder eine Leitidee der KMK-Standards könnte die Aussagekraft über Stärken und Schwächen der geprüften Schülerpopulation verbessern. Bei einer frühzeitigen

Bekanntgabe von solchen Schwerpunkten kann eine genauere Überprüfung von Detailkompetenzen statt finden.

Sämtliche Erkenntnisse, insb. auch die Empfehlung zu noch stärkerer Berücksichtigung von Mindeststandards und für eine geringfügige Reduzierung des Umfangs, wurden in die Vorgaben an das LISUM für die Entwicklung der kommenden Vergleichsarbeit aufgenommen.

D GLOSSAR

Vorbemerkung: Das Glossar wurde im Hinblick auf die Berichte aller vier Fächer erstellt. Nicht alle der nachstehenden Begriffserläuterungen sind daher für den vorliegenden Deutsch-Bericht von Bedeutung.

ITEM

Der kleinste Baustein in einem empirischen Erhebungsinstrument.

Beispiele: Frage in einem Fragebogen, Aufgabe in einem Mathematiktest.

ITEM-RESPONSE-THEORIE (IRT)

Mit den Vergleichsarbeiten sollen Leistungen der Schüler/innen festgestellt werden. Die Schüler/innen, so der Ausgangspunkt, verfügen über bestimmte Fähigkeiten, die nicht direkt beobachtbar sind (latent traits). Was sich beobachten (messen) lässt, ist, wie die Schüler/innen auf die vorgegebenen Items reagieren, wie sie also z.B. die Aufgaben in einem Mathematiktest lösen.

Ob ein Item bewältigt wird oder nicht, hängt von zwei Größen ab: Der Kompetenz (den Fähigkeiten) der Person und der Schwierigkeit (den Anforderungen) des Items. Lassen sich beide Aspekte (Personenkompetenz, Itemschwierigkeit) auf derselben Dimension verorten, gilt im Prinzip, dass eine Person dann ein Item löst, wenn sein Kompetenzniveau über dem Schwierigkeitsniveau liegt.

Die klassische Annahme der Testtheorie lautet: Es gibt einen deterministischen Zusammenhang dergestalt, dass eine Person alle Items löst, die unterhalb seiner Fähigkeitsniveaus liegen und keines oberhalb. In probabilistischen Modellen (z.B. bei der → RASCH-SKALIERUNG) wird von Lösungswahrscheinlichkeiten ausgegangen. Mit diesen Modellen verträglich ist der beobachtbare Umstand, dass zuweilen sehr fähige Personen auch manches einfache Item nicht bewältigen und manch "blindes Huhn auch ein Korn findet". Der strikte Zusammenhang zwischen Fähigkeit und Schwierigkeit wird zugunsten eines tendenziellen aufgegeben: Je fähiger eine Person ist, desto größer ist die Wahrscheinlichkeit, Items verschiedener Schwierigkeitsstufen zu lösen. Diese Wahrscheinlichkeit nimmt mit zunehmender Schwierigkeit ab.

KORRELATION

Korrelation meint den statistischen Zusammenhang zwischen zwei Merkmalen. Die Enge des Zusammenhanges wird mittels des Korrelationskoeffizienten ausgedrückt. I.d.R. (aber nicht zwangsläufig) misst der Korrelationskoeffizient lineare Zusammenhänge, also solche der Proportionalität (je - desto; Beispiel: Körpergröße und Körpergewicht).

Korrelationskoeffizienten sind Zusammenhangsmaße, die zwischen -1 und +1 variieren können. Das Vorzeichen zeigt an, in welcher Richtung ein Zusammenhang besteht (gleichsinnig/proportional oder ungleichsinnig/umgekehrt proportional), der Betrag des Koeffizienten quantifiziert das Ausmaß des (linearen) Zusammenhanges. Der Koeffizient darf aber nicht als Prozentanteil des maximal möglichen Zusammenhanges interpretiert werden. Sein Quadrat macht eine Aussage über den Anteil gemeinsamer Varianz, d.h. in welchem Ausmaße die beiden miteinander korrelierten Merkmale sich (gleichsinnig oder ungleichsinnig) verändern. (Und das Quadrat z.B. eines Koeffizienten von 0,5 beträgt nur 0,25.)

Korrelative Zusammenhänge dürfen nicht automatisch als kausale interpretiert werden. Wenn zwei Merkmale miteinander korrelieren, kann dies vielfältige Ursachen haben. Erforderlich ist in jedem Falle eine eingehende inhaltliche Überprüfung.

RASCH-SKALIERUNG

Eine Skalierungsmethode, die auf der probabilistischen → **Item-Response-Theorie** beruht.

Ihr Vorteil beruht darauf, dass sich die Fähigkeiten von Personen und die Schwierigkeiten von Items auf derselben Dimension abbilden lassen, wie dies im z.B. im Bericht zu den Ergebnissen der Mathematikvergleichsarbeiten vorgeführt wird.

Durch die Zuordnung von Items zu bestimmten RASCH-Werten, die i.W. im Bereich von 60 bis 140 liegen, können die unterschiedlichen Niveaus inhaltlich interpretiert werden. Hierzu sind die jeweiligen Items daraufhin zu untersuchen, welche Anforderungen sie beinhalten (Beherrschen der Grundrechenarten, Anwenden einfacher Kalküle etc. Können auf diese Art und Weise die unterschiedlichen RASCH-Niveaus im Hinblick auf die Items inhaltlich bestimmt werden, so lässt sich dies auf die Kompetenzen der Personen (Schüler/innen) übertragen: Die herausgearbeiteten Anforderungen, die in den Items stecken, sind die Kompetenzen, die die Schüler/innen aufweisen müssen, um die Items (mit einer gewissen Wahrscheinlichkeit) zu bewältigen.

Zu beachten ist, dass die konkrete Größe von Mittelwert und Streuung der RASCH-Werte (100 und 20 im vorliegenden Fall) sich nicht zwingend aus dem mathematischen Teil der RASCH-Skalierung ergeben. Dort ergeben sich durch das Rechenverfahren bedingt i.d.R. "krumme" Werte, die erst durch anschließende Normierung auf übliche und anschauliche Größenordnungen gebracht werden. Das stellt keine Manipulation dar und verfälscht nicht die Verhältnisse, denn die Anordnung der Personen und der Items auf der kombinierten Fähigkeits-/Schwierigkeitsdimension ändert sich hierdurch nicht.

SIGNIFIKANZ

Bedeutsamkeit, im vorliegenden Zusammenhang **statistische Bedeutsamkeit** im Sinne von Überzufälligkeit.

Empirische Untersuchungen beruhen im Allgemeinen auf Stichproben. Gewollt sind aber Aussagen über die zugrunde liegende Grundgesamtheit. Die (sog. Inferenz-)Statistik hilft dabei, die Verallgemeinerung von der Stichprobe auf die Grundgesamtheit abzusichern.

Beispiel: Zeigen Jungen oder Mädchen bessere Leistungen in Mathematik? Je einer Gruppe von Jungen und Mädchen wird derselbe Mathematiktest vorgegeben, der zu zwei Mittelwerten führt, einen für die Jungen, einen für die Mädchen, etwa $MW(J)=11,4$ und $MW(M)=10,5$.

Ist diese Differenz ein zufälliges Ergebnis, das nur für die Stichprobe gilt, oder gilt es für alle Jungen und Mädchen?

Die Statistik stellt Verfahren bereit, anhand derer diese Frage (innerhalb gewisser Grenzen) beantwortet werden kann. Als signifikant gilt das Ergebnis dann, wenn der statistische Test zum Ergebnis führt, dass die gefundene Differenz überzufällig, also unabhängig von der konkreten Stichprobe sei. Verschiedene Größen sind beim statistischen Testen zu berücksichtigen, u. a. die Größe der Stichproben; es gilt: Je größer die Stichproben, desto eher ergeben sich Signifikanzen.

Statistische Signifikanz und inhaltliche Bedeutsamkeit gehen nicht zwangsläufig miteinander einher. Bei großen Stichproben sind selbst kleinste Differenzen statistisch signifikant, aber inhaltlich bedeutungslos. Statistik ersetzt also nicht die inhaltliche Auseinandersetzung.

SKALIERUNG

Eine Skala bezeichnet das einem Messvorgang zugrunde liegende Bezugssystem, das, worauf z.B. eine zu testende Person verortet werden soll, etwa eine Skala zur Lesekompetenz oder zur Rechenfertigkeit.

Skalierung ist der Vorgang, eine Skala zu erstellen. Ausgangspunkt sind in der Regel eine größere oder kleinere Menge von Items, von denen angenommen wird, sie trügen etwas zur zu messenden Eigenschaft bei. Gesammelt werden könnten beispielsweise verschiedene Rechenaufgaben, die einer Stichprobe vorgelegt werden. Das Problem besteht nun (wie in einer Klassenarbeit) darin, die Einzellösungen so zusammenzufassen, dass jeder Person der Stichprobe ein möglichst aussagekräftiges Gesamtergebnis auf der Dimension Rechenfähigkeit zugeordnet werden kann.

Ein häufig gewähltes Verfahren besteht darin, jedem Item eine bestimmte maximal mögliche Punktzahl zuzuordnen. Summenbildung ergibt dann die einzelnen Ausprägungen der Skala. Dies Vorgehen geht u. a. davon aus, dass alle Items von ihren Anforderungen her einen Beitrag genau zu der zu messenden Eigenschaft leisten, dessen Ausmaß angemessen durch die Punkte repräsentiert wird.

Die Verfahren der Testkonstruktion systematisieren den Vorgang der Skalierung. Hierbei gibt es verschiedene Ansätze; → **ITEM-RESPONSE-THEORIE**.

STREUUNG (STANDARDABWEICHUNG, *standard deviation* (s, sd oder SD))

Die Streuung ist ein Maß für die Heterogenität. Die Streuung ist kein absolutes Maß, d.h. einer Streuung von z.B. 4,4 lässt sich nicht ansehen, ob dahinter eine große oder kleine Heterogenität steht. Streuungen lassen sich in der Regel erst dann sinnvoll interpretieren, wenn zwei Streuungen miteinander verglichen werden können.

Beispiel: Erhoben werden die Mathematikleistungen in zwei Parallelklassen. Mittelwert und Streuung betragen in der ersten Klasse $MW=22,3$ - $s=2,9$ und in der zweiten $MW=22,5$ - $s=4,4$. Daraus lässt sich ablesen, dass bei demselben mittleren Leistungsniveau (nahezu identische Mittelwerte) die Schüler/innen der ersten Klasse in ihren Mathematikleistungen homogener als jene der zweiten Klasse sind.

Verdeutlichen lässt sich die Bedeutung der Streuung als Homogenitätsmaß anhand der in diesem Bericht vorgelegten Ergebnisse, Beispiel Deutsch: Aus Tabelle D10 gehen die spezi-

fischen Streuungen hervor (OH: 2,0; OR: 1,6; OG: 1,8; O:1,6). Die Unterschiede zeigt die Abbildung D1: Die OH-Verteilung ist am breitesten, gefolgt von der OG-Verteilung, während die beiden anderen Verteilungen im Vergleich dazu homogener sind.

In die Berechnung der Streuung geht der Mittelwert ein. Dessen Kenntnis erleichtert neben dem eben exemplarischen Vergleich zweier Streuungen das Einschätzen, ob eine Streuung groß oder klein ist. Hierzu muss die Höhe der Streuung auf die Größe des Mittelwertes bezogen werden.

Beispiel: Eine Streuung von $s=4,4$ ist eher klein, beträgt der Mittelwert $MW=35,7$, und groß bei einem Mittelwert von $MW=5,1$.

Ist die Art der Verteilung bekannt, aus der die Daten stammen, für die Mittelwert und Streuung berechnet wurden, dann lassen sich weitergehende Aussagen treffen. Für die Normalverteilung gilt, dass innerhalb des Bereiches $MW-s$ und $MW+s$ rund zwei Drittel aller Werte liegen.

Beispiel: Die für PISA 2000 entwickelten Skalen aus den nationalen Ergänzungstests besitzen einen Mittelwert von 100 und eine Streuung von 30. Damit liegen die Leistungen von zwei Drittel aller in Deutschland untersuchten Schüler/innen im Bereich von $100-30=70$ und $100+30=130$.

Das Quadrat der Streuung heißt Varianz.

VARIANZ

→ **STREUUNG**