

**Vergleichsarbeiten Klasse 10/Frühjahr 2004.
Zweiter Bericht:**

**Ergebnisse differenziert nach
Migrationshintergrund
und nach
inhaltlichen Teilbereichen der Fächer**

Überblick

Die wichtigsten Ergebnisse auf einen Blick

Vorbemerkung zum zweiten Bericht

- A Zum Migrationshintergrund: Leistung bei unterschiedlicher Konstellation von
Herkunfts- und Verkehrssprache**
- B Vertiefender Blick in Teilbereiche der Fächer**
 - B1 Deutsch**
 - B2 Englisch**
 - B3 Mathematik**
- C Fazit**

ANHANG

- Migrationshintergrund (Sprachkombination) und Schulart
- Leistung und Geschlecht
- Herausfinden bemerkenswerter Items am Beispiel Mathematik: Itemschwierigkeiten in Untergruppen
- Zusammenhang zwischen Zensuren und den Ergebnissen in den Vergleichsarbeiten
- Zusammenhänge zwischen den Ergebnissen in den drei Fächern

Quellenangaben
Abkürzungen

Ergebnisblätter aus dem Frühjahr 2004 entsprechend den neuen Einteilungen abgewandelt

Die wichtigsten Ergebnisse auf einen Blick

Der vorliegende Bericht führt den ersten Bericht aus dem September 2004 fort. Er stellt in zwei Themenblöcken die Ergebnisse weiterer Auswertungen der schulischen Rückmeldungen vor. Ausdifferenziert werden zum einen die Ergebnisse nach Migrationshintergrund (Kapitel A), zum anderen die Werte in den einzelnen Fächern nach inhaltlichen Teilbereichen, die die mit den Aufgaben verbundenen Anforderungen spezifizieren (Kapitel B).

Zentrale Ergebnisse aus dem Kapitel A:

Im Zuge der Rückmeldungen wurden die Schulen gebeten, Angaben zu zwei Merkmalen des Migrationshintergrundes zu machen, nämlich zur Herkunftssprache und zur Verkehrssprache zu Hause. Während das erste Merkmal zweifach gestuft war (Dichotomie: deutsche/nicht-deutsche Herkunftssprache; dH - ndH), wurden beim zweiten vier Kategorien unterschieden (deutsch (d), türkisch (t), russisch (r), andere (a) Sprachen). Ein Vergleich der Ergebnisse zeigt:

- Nur im Fach Deutsch erreicht die Gruppe dH/dV höhere Werte als die Gruppe ndH/dV, die ansonsten die besten Ergebnisse erzielt (wenn auch ohne statistische Signifikanz).
- Die Sprachkombination ndH/tV weist durchgängig die niedrigsten Leistungswerte auf.
- Die Unterschiede zwischen den Gruppen sind sinnfälligerweise in Deutsch am größten, in Mathematik am kleinsten.
- Ein Vergleich der Fächer zeigt, dass die besten Ergebnisse in Englisch, die schlechtesten in Mathematik erzielt wurden.

In der Gruppe ndH/dV dürften die integrationswilligen und aufstiegsorientierten Familienhäuser vertreten sein. Daher lässt sich das zentrale Ergebnis zur Formel zuspitzen: Integration lohnt sich.

Zentrale Ergebnisse aus dem Kapitel B:

Die Ausdifferenzierung der Gesamtergebnisse erfolgt nach Teilbereichen, die sich nicht mehr aus der äußeren Struktur, sondern aus den Konzeptionen ergeben, die den Vergleichsarbeiten zugrundeliegen:

- In Deutsch nach den drei Verstehensdimensionen V1 *Informationen ermitteln*, V2 *Textverständnis entwickeln*, *Informationen verknüpfen* und V3 *Texte reflektieren und bewerten*.
- In Englisch nach den drei Niveaustufen aus dem Gemeinsamen Europäischen Referenzrahmen A2 *Elementare Sprachverwendung*, B1 *Selbstständige Sprachverwendung* und B1+.
- In Mathematik nach den zwei am häufigsten mit Aufgaben vertretenen der fünf Leitideen, nämlich nach L1 *Zahl* und L4 *Funktionale Zusammenhänge*.

Diese konzeptorientierte Einteilung der Aufgaben wird im Hinblick auf die Struktur der Stichprobe nach Schularten/Kursen untersucht wird. Dabei ergibt sich u.a.:

- Die Schulart OH in Deutsch bzw. der Kurs OH/B,C in Englisch und Mathematik weist die niedrigsten Leistungswerte auf.
- In Englisch und Mathematik liegen die Ergebnisse des Kurses OH/A über jenen des Kurses O/GA.
- Die Unterschiede zwischen den Schularten/Kursen sind für die Leitideen in Mathematik am größten, in Deutsch für die Verstehensdimensionen am kleinsten.
- Ein Vergleich der Fächer zeigt wiederum, dass die besten Ergebnisse in Englisch, die schlechtesten in Mathematik erzielt wurden.

Von Schule zu Schule, von Klasse zu Klasse können die Konstellationen völlig unterschiedlich von den hier vorgestellten aussehen, die auf der Mittelung über die gesamte große Stichprobe beruhen. Der Wert der dokumentierten Ergebnisse liegt in dem, was mit dem Bericht selber nicht leistbar ist, wofür er aber die Voraussetzung schafft: Zu prüfen, wie es um die eigene Klasse bestellt ist (war).

Aus dem Anhang:

Korrelation zwischen den Zensuren und Ergebnissen der Vergleichsarbeiten: Die Korrelationskoeffizienten sind zwar alle statistisch signifikant von Null verschieden, aber nur mittelhoch. Die Einschätzung des Leistungsniveaus anhand der Zensuren ist ähnlich, aber bei weitem nicht identisch mit der der Vergleichsarbeiten. Dies ist durchaus plausibel, denn Vergleichsarbeiten haben einen anderen Zugriff auf die Kenntnisse und Fähigkeiten der Schüler/innen als Klassenarbeiten oder der Prozess der Zensurengebung. In Deutsch und in Englisch ist der Zusammenhang zwischen Zensur und Gesamtergebnis in etwa gleich stark ausgeprägt, in Mathematik deutlich schwächer. Weiter als in den anderen Fächern scheinen in Mathematik die Aufgaben der Vergleichsarbeit von jenen im Unterricht behandelten oder in Klassenarbeit vorkommenden entfernt zu sein.

Fächerübergreifende Korrelationen: Erwartungsgemäß sind die Korrelationen der Zensuren zwischen den beiden sprachlichen Fächern erheblich höher als mit der Zensur in Mathematik. Dies deckt sich mit der weit verbreiteten Annahme, dass sprachliche Fähigkeiten zu einem großen Teil unabhängig von mathematischen vorhanden sein können.

Die Korrelationen zwischen den Gesamtpunktzahlen der drei Vergleichsarbeiten liegen höher, teilweise deutlich höher, als die Korrelationen zwischen den Zensuren, und alle drei in etwa derselben Größenordnung. Offensichtlich enthalten die Aufgaben aus allen drei Vergleichsarbeiten Anforderungen, die weniger fachabhängig sind als jene, die sich in den Bewertungen der Lehrkräfte als Zensuren niederschlagen.

Folgerungen:

Die Analysen des vorliegenden Berichtes stießen an Grenzen, die durch die Aufgaben und die bislang vorliegenden Strukturkriterien, nach denen die Aufgaben gebündelt werden, gekennzeichnet sind.

Um die praktische Relevanz der Vergleichsarbeiten zu erhöhen, ist es erforderlich, möglichst trennscharfe Items zu finden, "analytische" Aufgaben, aus denen sich eine gewissermaßen atomare Struktur der Vergleichsarbeit ergibt. Durch eine weitgehend eindeutige Zuordnung von Aufgabe zur damit verbundenen Anforderung lässt sich dann im Idealfall unmittelbar aus dem über- oder unterdurchschnittlichen Abschneiden folgern, wo die Stärken und Schwächen der eigenen Klasse liegen.

Die diagnostische Funktion der Vergleichsarbeiten zu stärken verlangt Anstrengungen von beiden Seiten, von den "Produzenten" und von den "Abnehmern". Die Entwicklerteams sind gefordert, bei der Entwicklung der Aufgaben dem spezifischen Charakter von Vergleichsarbeiten Rechnung zu tragen und möglichst analytische Aufgaben zu finden. Die Lehrkräfte müssen in bislang wenig geübter Praxis eine diagnostisch orientierte Auswertung der Arbeit vornehmen, indem nicht allein die Gesamtergebnisse in Bezug zu den Vergleichswerten gesetzt werden, sondern auch von wohldefinierten Teilbereichen oder - falls erforderlich - von Einzelaufgaben.

Vorbemerkung zum zweiten Bericht

Der Bericht aus dem September 2004 dokumentiert die Lösungshäufigkeiten für die Gesamtergebnisse und die Einzelitems der Vergleichsarbeiten. In einer Ergänzung Ende Februar 2004 wurden diese Angaben weiter differenziert nach den Kursniveaus in den Haupt- und den Gesamtschulen. Der vorliegende Bericht stellt in zwei Themenblöcken die Ergebnisse weiterer Auswertungen der schulischen Rückmeldungen vor. Ausdifferenziert werden zum einen die Ergebnisse nach Migrationshintergrund, zum anderen die Werte in den einzelnen Fächern nach inhaltlichen Teilbereichen, die die mit den Aufgaben verbundenen Anforderungen spezifizieren.

Damit sollen nicht nur weitere wichtige Ergebnisse mitgeteilt, sondern auch Teilanalysen vorgestellt werden, die beispielhaft zeigen, wie auf Schul- und Klassenebene mit den Daten umgegangen werden kann, um Hinweise für die eigene Arbeit zu erhalten.

Hinweis zum Verständnis der Werte in den Tabellen:

Bei einem Vergleich von Tabellenangaben innerhalb dieses Berichts und über die verschiedenen Berichte hinweg können sich kleinere Abweichungen bei den Mittelwerten, Prozentangaben oder Fallzahlen ergeben. Ein Datensatz ist nie vollständig: Bei einer Reihe von Fällen fehlen die Angaben zu dem einen oder anderen oder zu mehreren Merkmalen. In einer Tabelle aber, die beispielsweise Kombinationen aus Herkunftssprache, Schulart und Mathematikleistung enthält, können nur Fälle (Vergleichsarbeiten) berücksichtigt werden, bei denen zu allen drei Merkmalen gültige Werte vorliegen. Stellt die Tabelle nur die Verteilung der Jugendlichen nach Herkunftssprache auf die Schularten dar, so reichen die Angaben zu diesen beiden Merkmalen, so dass die Fälle mit fehlenden Angaben zur Mathematikleistung ebenfalls Eingang in die Tabelle finden können.

Zu beachten ist auch, dass Anteile aufgrund von Auf- und Abrundungen sich nicht unbedingt auf 100% addieren müssen.

A

**Zum Migrationshintergrund:
Leistung bei unterschiedlicher Konstellation von
Herkunfts- und Verkehrssprache**

Im Zuge der Rückmeldungen wurden die Schulen gebeten, Angaben zu zwei Merkmalen des Migrationshintergrundes zu machen, nämlich zur Herkunftssprache und zur Verkehrssprache zu Hause. Während das erste zweifach gestuft war (Dichotomie: deutsche/nichtdeutsche Herkunftssprache; dH - ndH), wurden beim zweiten vier Kategorien unterschieden (deutsch, türkisch, russisch, andere Sprachen). Die Verteilung auf die sechs möglichen Kombinationen der beiden Merkmale zeigt Tabelle A1 am Beispiel des Datensatzes Deutsch.¹

A1 Tabelle: Verteilung auf die unterschiedlichen Kombinationen von Herkunfts- und Verkehrssprache. (Datensatz Deutsch)

Angegeben werden absolute Häufigkeiten und die prozentualen Anteile innerhalb der beiden Gruppen der Herkunftssprachen (Zeilenprozente).

Verkehrssprache Herkunftssprache	deutsch	türkisch	russisch	andere	gesamt
dH	1 459 100%	4 0%	1 0%	2 0%	1 466 100%
ndH	51 12%	170 42%	52 13%	136 33%	409 100%
gesamt	1 510 81%	174 9%	53 3%	138 7%	1 875 100%

Die in Tabelle A1 gezeigten Werte gelten grosso modo ebenfalls für die beiden anderen Fächer Englisch und Mathematik, insbesondere die Proportionen für die einzelnen Kombinationen. Die Kombinationen deutscher Herkunftssprache mit anderen Verkehrssprachen als dem Deutschen treten kaum auf. Für die Analysen bleiben diese daher außen vor. Grundlage für die Auswertung bilden demnach die Sprachkonstellationen, die Tabelle A2 wiedergibt.

Für alle drei Fächer sind die Verteilungen identisch bis auf geringe Schwankungen der absoluten Häufigkeiten. Das liegt daran, dass es sich im Prinzip um die Vergleichsarbeiten immer derselben Schüler/innen handelt, dass allerdings zugleich nicht für jeden Schüler/in vollständige Rückmeldungen für alle drei Fächer erfolgten, z.B. weil jemand nicht mitgeschrieben oder Französisch statt Englisch geschrieben hatte. Tabelle A2 jedenfalls zeigt, dass wir in allen drei Fächern von denselben Gegebenheiten ausgehen können, die wiederum stellvertretend für die Konstellationen der gesamten Berliner Schülerschaft auf der zehnten Klassenstufe stehen; vgl. Kapitel 1 des ersten Berichts.

¹ Die Datensätze für Deutsch, Englisch und Mathematik sind unterschiedlich groß, wenn auch in etwa vom selben Umfang von rund 2000 Arbeiten.

A2 Tabelle: Datengrundlage der Analyse in den drei Fächern: Die zu berücksichtigenden Kombinationen von Herkunfts- und Verkehrssprache.

Angegeben werden absolute Häufigkeiten und die prozentualen Anteile innerhalb der Fächer (Zeilenprozente). H/V: Herkunfts-/Verkehrssprache; d: deutsch, nd: nichtdeutsch, t: türkisch, r: russisch, a: andere Sprachen.

Kombination Fach	dH/dV	ndH/dV	ndH/tV	ndH/rV	ndH/aV	gesamt
Deutsch	1 459 78%	51 3%	170 9%	52 3%	136 7%	1 868 100%
Englisch	1 585 78%	57 3%	180 9%	56 3%	148 7%	2 026 100%
Mathematik	1 408 78%	49 3%	166 9%	49 3%	130 7%	1 802 100%

Es sei an die Vorbemerkung erinnert: In den nachfolgenden Tabellen werden nicht dieselben Fallzahlen auftreten, da nun über die beiden Sprachmerkmale hinaus auch die Leistungsdaten aus den Vergleichsarbeiten benötigt werden und nicht für jeden Fall Angaben zu allen Merkmalen vorliegen.

In den drei Fächern war die maximal erreichbare Punktzahl unterschiedlich; sie betrug in Deutsch 50 Punkte, in Englisch 75 und in Mathematik 45. Daher bedeutet der Gesamtmittelwert von 32,9 in Deutsch etwas anderes als die 54,2 in Englisch und die 25,3 in Mathematik. Diese Gesamtmittelwerte sind wichtige Hinweise auf das mittlere Leistungsniveau in den einzelnen Fächern. Um vergleichbare Größen zu gewinnen, können die Mittelwerte auf die maximal erreichbare Punktzahl bezogen werden, also in

Deutsch $32,9/50 = 0,68$ (68%)
 Englisch $54,2/75 = 0,72$ (72%)
 Mathematik $25,3/45 = 0,56$ (56%)

Es bietet sich an, den Wert in Prozenten anzugeben, denn er gibt wieder, zu welchem Anteil die Berliner Schüler/innen (im Schnitt) die maximal mögliche Punktmenge ausgeschöpft haben. Es ist also die Lösungshäufigkeit für die Gesamtarbeit, d.h. der durchschnittliche Anteil erreichter Punkte.

Diese fächerübergreifende quantifizierende Größe nennen wir **(Mittlerer) Lösungsanteil**.

Sie lässt sich auch für Teilbereiche angeben: Zunächst ist die maximal erreichbare Punktzahl eines Teilbereiches festzustellen und dann der dazugehörige Mittelwert darauf zu beziehen, also der Quotient zu bilden. Beispiel: In der Tabelle A3 wurden über alle Schüler/innen hinweg 24,7 von 35 maximal möglichen Punkte erreicht, also $24,7/35 = 71\%$.

Eine weitere fächerübergreifende Vergleichsgröße stellt der Anteil derjenigen dar, die die vorgegebene Erfolgsschwelle überschritten haben (**Bestehensquote**). Die Vereinfachung durch die Dichotomie *bestanden - nicht bestanden* liefert einen sicherlich vereinfachenden, aber pointierten Eindruck vom Gesamtergebnis einer Vergleichsarbeit. Diese Größe ist durch etwas charakterisiert, das zugleich ein Vorzug und ein Nachteil ist: Die Schwelle ist in den

drei Fächern an unterschiedlicher Stelle festgelegt worden: 50% (25) der maximal 50 Punkte in Deutsch, 60% (45) von 75 Punkten in Englisch und 45% (21) von 45 Punkten in Mathematik. In absoluten Zahlen sind die drei Schwellen damit unterschiedlich. Aber das bedeutet nicht zwangsläufig, dass es unterschiedlich schwierig ist, die Schwelle zu erreichen, denn die drei Arbeiten dürften nicht denselben Schwierigkeitsgrad besitzen. Der normative Eingriff, eine Bestehensschwelle fächerspezifisch festzulegen, stellt den Versuch dar, aus fachdidaktischer Sicht die geschätzten Unterschiede auszugleichen.

Die Bestehensquote besitzt demnach eine normative Komponente, der Lösungsanteil aggregiert die aufgetretenen Itemschwierigkeiten.

A3 Tabelle: Deutsch.
Leistungsergebnisse differenziert nach Sprachkonstellationen.

	dH/dV	ndH/dV	ndH/tV	ndH/rV	ndH/aV	gesamt
DEUTSCH						
Fallzahl N	1 417	50	163	51	125	1 806
Quote "bestanden"	87%	86%	52%	80%	81%	83%
Teil I: Mittelwert	25,5	23,7	20,7	23,7	23,9	24,7
Inhalt: Mittelwert	5,6	5,0	3,2	5,0	4,9	5,3
Darstellung: Mittelwert	3,0	3,0	1,7	2,5	2,4	2,9
Teil II: Mittelwert	8,9	8,0	5,0	7,5	7,4	8,2
Insgesamt: Mittelwert	34,0	31,7	25,6	31,2	31,2	32,9
Mittlerer Lösungsanteil	68%	63%	51%	62%	62%	66%

Betrachten wir die Zeile, die angibt, wie hoch der Anteil derjenigen ist, deren Leistung oberhalb der kritischen Schwelle liegt, so zeichnet sich eine Dreiteilung der fünf Sprachgruppen ab: Am oberen Ende des Leistungsspektrums diejenigen, die deutsch zu Hause sprechen und dies unabhängig davon, welches ihre Herkunftssprache ist. Die niedrigsten Leistungswerte weisen die Schüler/innen auf, deren Verkehrssprache zu Hause türkisch ist. Die beiden anderen Gruppen mit russisch oder einer anderen Verkehrssprache liegen dazwischen, aber weitaus dichter am oberen als am unteren Ende.

Dieser erste Eindruck bestätigt sich, ziehen wir die Gesamtpunktzahlen (unterste Zeile) heran,² allerdings mit kleinen Modifikationen. Die Spitze besteht nur noch aus den Schüler/innen mit der Kombination dH/dV, während die Kombination ndH/dV in das unmittelbar darunter liegende Feld rutscht. Der Unterschied zwischen den beiden Sprachkombinationen fällt in diesem Merkmal größer aus als bei der Bestehensquote. Die Ursache hierfür ist die größere Homogenität der Gruppe ndH/dV im Vergleich mit dH/dV, wie die hier nicht dokumentierten Streuungen zeigen. Der Mittelwert der Gruppe ndH/dV liegt relativ hoch und zugleich scharen sich die Schüler/innen dieser Gruppe dicht um den Mittelwert, so dass ein hoher Anteil über die kritische Bestehensschwelle kommt.

² Die Mittelwerte sind statistisch signifikant voneinander verschieden.

Die Gesamtpunktzahlen mit ihrer möglichen Spannweite von 0 bis 50 Punkten bilden die Leistungsergebnisse besser ab als das zweistufige (dichotome) Merkmal *bestanden/nicht bestanden*. Von besonderem Interesse ist es daher zu prüfen, wie es zu den Unterschieden im Gesamtwert kommt, indem die Komponenten der Vergleichsarbeit Deutsch betrachtet werden. Dabei zeigt sich, dass die Gruppe dH/dV in beiden Teilen der Arbeit gut abschneidet, wobei dies für den Teil II, der Schreibaufgabe, auf den guten Werten im Modul *Inhalt* beruht. Beim Modul *Darstellung* vermag die Gruppe ndH/dV aufzuschließen.

Was könnten die Ursachen für diese Konstellation sein?

Dass die Gruppe dH/dV jedem der Bewertungskriterien als beste genügt, ist erwartungsgemäß. Die Schüler/innen mit der Kombination ndH/dV dürften wahrscheinlich aus integrationswilligen und aufstiegsorientierten Elternhäusern stammen, was i.a. einen fördernden Effekt auf die Leistungsbereitschaft der Kinder hat.³ Zugleich gilt, dass es einfacher ist, sich mit formalen Anforderungen vertraut zu machen, eben mit der sprachlichen Gestaltung, als mit soziokulturellen Inhalten einer Gesellschaft, mit denen man später als die gleichaltrigen Mitschüler/innen konfrontiert wurde. Die Bewertung des *Inhalts* einer Schreibaufgabe⁴ ist wahrscheinlich stärker kulturell geprägt als die der *Darstellung*, so dass beim *Inhalt* deutlicher als bei der *Darstellung* Differenzen zwischen der beurteilenden Lehrkraft, die nahezu ausschließlich des Typs dH/dV ist, und den Schüler/innen ndH/dV zu Tage treten, ein Vorgang, wenn es ihn denn wie hier postuliert gibt, der i.d.R. unbewusst ablaufen dürfte.

Tabelle A4 gibt Antwort auf die Frage, wie weit die Konstellationen für Deutsch im Fach Englisch sich wiederholen.

Die Konstellationen in den Fächern Deutsch und Englisch zeigen bedeutsame Gemeinsamkeiten und Unterschiede; vgl. Tabelle A3 und A4. Am unteren Ende des Leistungsspektrums befindet sich in jedem Falle die Gruppe ndH/tV. Und besser als die dH/dV-Schüler/innen sind in Englisch die ndH/dV-Jugendlichen, auch wenn der Unterschied weder statistisch signifikant, noch inhaltlich bedeutsam ist. Diese Gleichheit der Mittelwerte gilt für alle Gruppen außer ndH/tV, die statistisch signifikant schlechter sind.

Zugleich lässt sich dasselbe Phänomen wie im Fach Deutsch beobachten: Die Gruppe ndH/dV ist homogener als die Gruppe dH/dV,⁵ so dass wiederum ein höherer Anteil über die kritische Bestehensschwelle kommt, als es nach den Mittelwerten zunächst den Anschein hätte; vgl. die beiden Zeilen "Quote "bestanden"" und "Insgesamt".

Bei allen fünf Gruppen gibt es einen starken Abfall von den Testteilen *Hören* und *Lesen* hin zu *Schreiben*. Besonders groß ist der Abfall in den Gruppen ndH/tV und ndH/aV. Letztere erzielt im Teil *Hören* die besten Ergebnisse. In dieser Gruppe könnten sowohl englischsprachige Schüler/innen sein als auch solche, die als ndH-Jugendliche bereits mit dem Englischen als (mündlichem) Verständigungsmittel vertraut waren, ohne aber über den Typ von Sprachkenntnissen zu verfügen, wie er in der Schule vermittelt und abgeprüft wird.

³ Vgl. hierzu die Tabelle Z1 aus dem Anhang.

⁴ Gefordert war, zur Aussage "Wichtigstes Ziel bei der Berufswahl sollte es sein, den Wunschberuf anzustreben" Stellung zu beziehen.

⁵ $s(dH/dV) = 13,4$; $s(ndH/dV) = 11,1$.

A4 Tabelle: Englisch.
Leistungsergebnisse differenziert nach Sprachkonstellationen.⁶

	dH/dV	ndH/dV	ndH/tV	ndH/rV	ndH/aV	gesamt
ENGLISCH						
Fallzahl N	1 497	56	168	56	142	1 518
Quote "bestanden"	81%	88%	59%	80%	84%	79%
Hören-1: Mittelwert	5,2	5,1	4,7	5,0	5,2	5,1
Hören-2: Mittelwert	8,0	8,0	7,3	8,2	8,3	8,0
Hören-3: Mittelwert	6,4	6,2	5,3	6,6	6,6	6,3
Hören: Mittelwert	19,5	19,3	17,3	19,8	20,0	19,4
Lesen-1: Mittelwert	4,1	4,1	3,8	4,1	4,1	4,1
Lesen-2: Mittelwert	6,5	6,5	6,2	6,5	6,6	6,4
Lesen-3: Mittelwert	8,6	8,8	7,8	8,4	8,8	8,5
Lesen: Mittelwert	19,1	19,4	17,8	19,0	19,4	19,0
Schreiben-1: Mittelwert	7,0	7,5	5,1	6,9	6,6	6,8
Schreiben-2: Mittelwert	3,2	3,1	2,3	3,2	3,2	3,1
Letter/TASK: Mittelwert	3,3	3,3	2,6	3,1	3,1	3,2
Letter/LANG: Mittelwert	2,9	2,9	2,2	2,8	2,7	2,8
Schreiben-3 (Letter): MW	6,1	6,1	4,8	5,9	5,8	6,0
Schreiben: Mittelwert	16,3	16,7	12,2	16,0	15,6	15,9
Insgesamt: Mittelwert	54,9	55,4	47,2	54,7	55,1	54,3
Mittlerer Lösungsanteil	73%	74%	63%	73%	72%	72%

Im Fach Mathematik gab es die Besonderheit, dass es von den letzten beiden Aufgaben (14 und 15) drei Varianten gab, die den Bereichen Körperberechnung (K), Trigonometrie (T) und Sachrechnen (S) entstammten. Nicht die Schüler/innen konnten zwischen diesen drei Bereichen wählen, aber die Lehrkräfte. Die Wahlentscheidung ist also ein Klassen-, kein Individualmerkmal. Wie die Verteilung der fünf Sprachgruppen auf die Wahlbereiche ausfiel, zeigt Tabelle A5.

⁶ MW: Mittelwert.

A5 Tabelle: Mathematik.

Aufteilung der Vergleichsarbeiten Mathematik auf die drei Wahlbereiche Körperberechnung (K), Trigonometrie (T) und Sachrechnen (S) differenziert nach Sprachkonstellation.

	dH/dV	ndH/dV	ndH/tV	ndH/rV	ndH/aV	gesamt
K	527 41%	11 23%	66 47%	22 49%	59 51%	685 42%
T	594 46%	33 69%	57 40%	16 36%	44 38%	744 45%
S	174 13%	4 8%	18 13%	7 16%	12 10%	215 13%
gesamt	1 295 100%	48 100%	141 100%	45 100%	115 100%	1 644 100%

Sachrechnen ist in allen Gruppen am schwächsten vertreten. Dies deckt sich mit Ergebnissen, die die Tabelle 2C.4 im ersten Bericht wiedergibt: Von allen Schularten wurde nur an den Berufsfachschulen und bereits wesentlich seltener an den Hauptschulen die Sachrechenaufgaben in substanziellem Maße gewählt. Tabelle A5 zeigt, dass die Gruppe ndH/dV sich überdurchschnittlich häufig mit Aufgaben aus der Trigonometrie auseinandersetzen musste und unterdurchschnittlich häufig mit der Körperberechnung. Dieser Umstand darf im Hinblick auf die gezeigten Leistungen nicht überbewertet werden, da - wie erwähnt - die Entscheidung für die Wahlbereiche von den Lehrkräften und nicht von den Schülern/innen getroffen wurde. Und wie Ergebnisse zu den Leistungen ausfallen, zeigt die Tabelle A6.

A6 Tabelle: Mathematik.

Leistungsergebnisse differenziert nach Sprachkonstellationen.⁷

	dH/dV	ndH/dV	ndH/tV	ndH/rV	ndH/aV	gesamt
MATHEMATIK						
Fallzahl N	1 302	48	141	45	116	1 652
Quote "bestanden"	68%	69%	60%	64%	60%	67%
Teil I (Aufg. 1 - 13): MW	19,8	19,9	19,8	19,6	19,0	19,8
Wahlbereich K: Mittelwert	5,9	4,1	3,8	6,7	4,7	5,6
Wahlbereich T: Mittelwert	6,5	6,8	6,6	6,6	6,0	6,5
Wahlbereich S: Mittelwert	2,4	1,8	0,6	1,9	1,0	2,1
Insgesamt: Mittelwert	25,5	25,7	24,3	25,5	23,8	25,3
Mittlerer Lösungsanteil	57%	57%	54%	57%	53%	56%

⁷ Fortsetzung von Tabelle A3 für das Fach Mathematik.

Bei den Mittelwerten für die Wahlbereiche sind die teilweise sehr niedrigen Fallzahlen zu beachten; vgl. Tabelle A5.

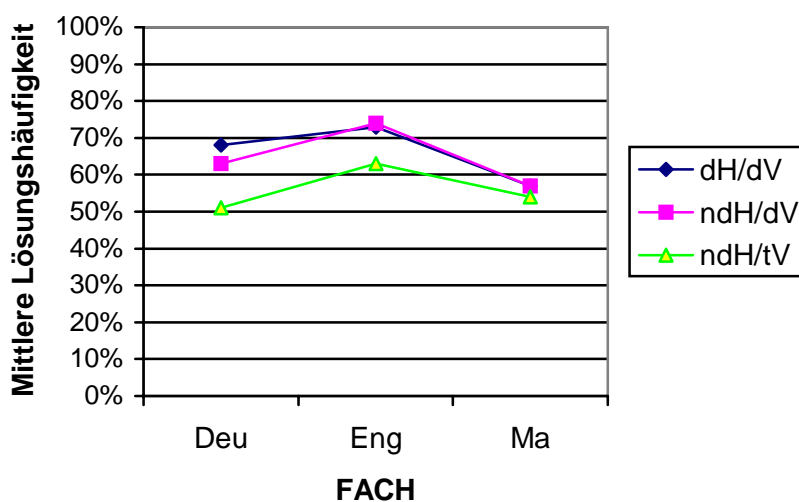
Insgesamt liegen die Mittelwerte aller Sprachenkombinationen weit dichter beisammen als in Deutsch oder in Englisch.⁸ Wie in Englisch liegt ndH/dV an der Spitze, auch wenn der Unterschied zu anderen Gruppen weder statistisch signifikant noch inhaltlich bedeutsam ist. Deutlich besser als in den anderen Fächern schneidet nun die Gruppe ndH/tV ab.

Die Mittelwertunterschiede, die nicht im Teil I, in dem Kern an Aufgaben, der von allen bearbeitet wurde, und nicht in der Gesamtpunktzahl auftreten, sollten in den Wahlbereichen nicht allein aufgrund der geringen Aufgabenzahl und nicht nur wg. der stark schwankenden und teilweise sehr niedrigen Fallzahlen überbewertet werden, sondern auch weil nicht bekannt ist, was die Lehrkräfte zu ihrer Entscheidung bewog, in ihren Klassen den einen Typ von Aufgaben bearbeiten zu lassen.

Das zentrale Ergebnis, dass der (statistische) Zusammenhang zwischen mittlerem Leistungsniveau und Sprachkombination in Mathematik weit weniger ausgeprägt ist als in den anderen Fächern, belegt, dass in Mathematik - zumindest bei dem in der Vergleichsarbeit verwendeten Typ von Aufgaben - die sprachlichen Fertigkeiten und Fähigkeiten eine relativ geringe Rolle spielen.

Dieses Kapitel abschließend seien die Gesamtergebnisse der drei Fächern in der Abbildung A7 nebeneinander gestellt, wobei der Übersichtlichkeit halber eine Beschränkung auf die drei Sprachkombinationen dH/dV, ndH/dV und ndH/tV erfolgt, die das Leistungsspektrum in seiner Breite abstecken.

A7 Abbildung: Mittlerer Lösungsanteil⁹ in den Fächern Deutsch, Englisch und Mathematik differenziert nach drei Sprachkombinationen.



⁸ Eine Varianzanalyse bestätigt diesen Eindruck: Statistisch signifikante Unterschiede treten nur in den Wahlbereichen auf und dort bei der Körperberechnung und beim Sachrechnen.

⁹ Vgl. die Erläuterung zur Tabelle A3. Die Mittelwerte der Gesamtpunktzahlen sind nicht miteinander vergleichbar, da die maximal erreichbaren Punkte von Fach zu Fach unterschiedlich sind.

Die Abbildung A7 fasst zusammen, was die Betrachtung der einzelnen Fächer ergab, und verdeutlicht einige Aspekte:

- Nur im Fach Deutsch erreicht die Gruppe dH/dV höhere Werte als die Gruppe ndH/dV, die ansonsten die besten Ergebnisse erzielt (wenn auch ohne statistische Signifikanz).
- Die Sprachkombination ndH/tV weist durchgängig die niedrigsten Leistungswerte auf.
- Die Unterschiede zwischen den Gruppen sind sinnfälligerweise in Deutsch am größten, in Mathematik am kleinsten.
- Ein Vergleich der Fächer zeigt, dass die besten Ergebnisse in Englisch, die schlechtesten in Mathematik erzielt wurden.

B**Vertiefender Blick in Teilbereiche der Fächer**

Um zu erkennen, wo Stärken und Schwächen der eigenen Klasse liegen, können Vergleichsarbeiten wertvolle Hinweise liefern, wenn die Ergebnisse differenziert nach Anforderungen vorliegen, wenn also die einzelnen Aufgaben in abgrenzbare und wohldefinierte inhaltliche Teilbereiche zusammengefasst und hieraus in Form spezifischer Teilergebnisse Indikatoren für Fähigkeitsprofile gewonnen werden können.

Der erste Bericht bot hierfür bereits Einiges an Material, denn er lieferte die Lösungshäufigkeiten für jede Einzelaufgabe und für einzelne Teile, aus denen die Vergleichsarbeiten bestanden, so dass - unterstützt durch eingefügte Tabellen - die Möglichkeit bestand, die Werte der eigenen Klasse unter unterschiedlichen Aspekten mit schulartspezifischen Berliner Gesamtwerten in Beziehung zu setzen. Die Mitteilung der Teilergebnisse beschränkte sich dabei auf die Unterteilung, die durch die Struktur der Arbeiten vorgegeben war.

Nachstehend soll in drei Abschnitten, die jeweils einem der Fächer gewidmet sind, ein Schritt darüber hinausgegangen werden. Es erfolgt eine Ausdifferenzierung der Gesamtergebnisse nach Teilbereichen, die sich aus den Konzeptionen ergeben, die den Vergleichsarbeiten zugrundeliegen. Diese Konzeptionen speisen sich aus unterschiedlichen Quellen, die teils eingeführt sind (z.B. Erwartungshorizont), teils erst seit wenigen Jahren die Diskussion bereichern (z.B. Kompetenzstufenmodelle). Die Sache selbst mit ihren je fachspezifischen Problemen bringt es mit sich, dass die Konzeptionen unterschiedlich weit elaboriert sind. Dementsprechend unterschiedlich weit reichen die nachstehend dokumentierten Versuche, zu weiteren aussagekräftigen Ergebnissen zu kommen. In jedem Fall aber ist dies hoffentlich ausreichend, interessierten Lehrkräften Ideen zu liefern, wie sie die Ergebnisse ihrer Klasse für sich und ihre Arbeit weiter auswerten können.

B1 Deutsch

Die Entwicklung der Deutschaufgaben fand vor dem Hintergrund dreier Kategorien von Leistungsaspekten statt (vgl. <http://www.kmk.org/schul/bildungsstandards>):

Vier Kompetenzbereiche:

- K1:** Sprechen und Zuhören
- K2:** Schreiben
- K3:** Lesen - mit Texten und Medien umgehen
- K4:** Sprache und Sprachgebrauch untersuchen.

Drei Anforderungsbereiche:

- A1:** Verfügbarkeit der für die Bearbeitung der Aufgaben notwendigen inhaltlichen und methodischen Kenntnisse
- A2:** Selbstständiges Erfassen, Einordnen, Strukturieren und Verarbeiten der aus der Thematik, dem Material und der Aufgabenstellung erwachsenden Fragen/Probleme und deren entsprechende gedankliche und sprachliche Bearbeitung
- A3:** Eigenständige Reflexion, Bewertung bzw. Beurteilung einer komplexen Problemstellung/Thematik oder entsprechenden Materials und ggf. die Entwicklung eigener Lösungsansätze.

Drei Verstehensdimensionen:**V1:** Informationen ermitteln**V2:** Textverständnis entwickeln, Informationen verknüpfen**V3:** Texte reflektieren und bewerten.

Die nachstehende exemplarische Auswertung zieht die drei Verstehensdimensionen heran. Die einzelnen Aufgaben lassen sich anhand der Informationen des Entwicklerteams wie folgt zuordnen (in Klammern jeweils die maximal zu vergebende Punktzahl):

V1: Informationen ermitteln.

Aufgabe 01 (2), 02 (1), 03 (1), 05(2), 08 (2).

Maximal mögliche Punktzahl: 8.

V2: Textverständnis entwickeln, Informationen verknüpfen.Aufgabe 04 (2), 06 (2), 07 (2), 09 (2), 10 (2), 13 (1), 14 (2), 15 (4), 16 (2),
18-01 (2), 18-06 (1), 18-08 (1), 18-09 (1).¹⁰

Maximal mögliche Punktzahl: 24.

V3: Texte reflektieren und bewerten.

Aufgabe 11 (1), 12 (3), 17 (4),

18-02 (2), 18-03 (2), 18-04 (2), 18-05 (2), 18-07 (1), 18-10 (1).

Maximal mögliche Punktzahl: 18.

Die Zuordnung ist teilweise sicherlich nicht eindeutig, ist aber unter pragmatischen Gesichtspunkten zufriedenstellend und genügt dem hier verfolgten Zweck, exemplarisch zu zeigen, wie sich inhaltlich unterrichtsrelevante Einteilungen abweichend von der äußeren Struktur der Vergleichsarbeit als Grundlage weiterer Auswertungen bestimmen lassen.¹¹ Die Tabelle B1.1-a hält die Ergebnisse für die drei Verstehensdimensionen fest. Sie erweitert die Tabelle 2A.1 aus dem ersten Bericht.

Tabelle B1.1-a liefert das zu erwartende Bild, sowohl wenn wir die Schularten als auch wenn wir die dH- und die ndH-Ergebnisse miteinander vergleichen. Dass sich hinter den niedrigen ndH-Werten ein breites Spektrum verbirgt, das der Differenzierung bedarf, wurde im Kapitel A bereits gezeigt.

¹⁰ Unter den Aufgaben (Items) 18-01 ff. werden die zehn Bewertungskriterien verstanden, wie sie im Begleittext zur Vergleichsarbeit formuliert werden. Für die ersten fünf, für die auf den Inhalt bezogenen, konnten bis zu 2 Punkte vergeben werden, für die Darstellungsitens 18-06 bis 18-10 nur 1 Punkt.

¹¹ Aus Sicht der Testkonstruktion werden hier drei Skalen gebildet, deren jeweiliger Wert sich durch Aufaddieren der Einzelwerte, also durch Summenbildung ergibt. Da hier ein exemplarischer Versuch aus inhaltliche3n Gründen interessiert, soll die statistische Sinnhaftigkeit nicht diskutiert werden. Wir kommen im Kapitel C darauf zurück. Mitgeteilt seien die α -Werte, die die interne Konsistenz quantifizieren:

$$\alpha(V1)=0,55, \alpha(V2)=0,67, \alpha(V3)=0,76.$$

Sie sind nicht besonders hoch, was teilweise an der geringen Itemanzahl liegt (V1), aber hoch genug, um mit der Auswertung fortzufahren.

B1.1-a Tabelle: Deutsch.**Punktwerte differenziert nach Schulart und Herkunftssprache.**

Angegeben wird jeweils der Mittelwert. V1 bis V3: Die drei Verständensdimensionen; siehe Text.

	O/FE	O/GA	OH	OR	OG	OBF	gesamt
gesamt	34,5	26,5	22,9	33,0	38,9	27,6	32,9
dH	34,7	27,6	23,8	34,0	39,6	28,9	34,0
ndH	33,8	23,2	20,9	28,8	35,6	24,1	28,9
V1							
gesamt	7,3	6,3	5,5	7,0	7,7	6,6	7,0
dH	7,4	6,5	5,8	7,1	7,7	6,8	7,1
ndH	7,1	5,7	4,9	6,7	7,5	6,3	6,6
V2							
gesamt	17,2	13,7	11,7	16,2	19,1	14,3	16,4
dH	17,2	14,1	12,0	16,7	19,4	15,0	16,8
ndH	17,1	12,6	10,9	14,2	17,8	12,3	14,6
V3							
gesamt	10,0	6,5	5,7	9,8	12,1	6,7	9,5
dH	10,1	7,0	6,0	10,3	12,5	7,1	10,0
ndH	9,5	5,0	5,0	8,0	10,4	5,6	7,7

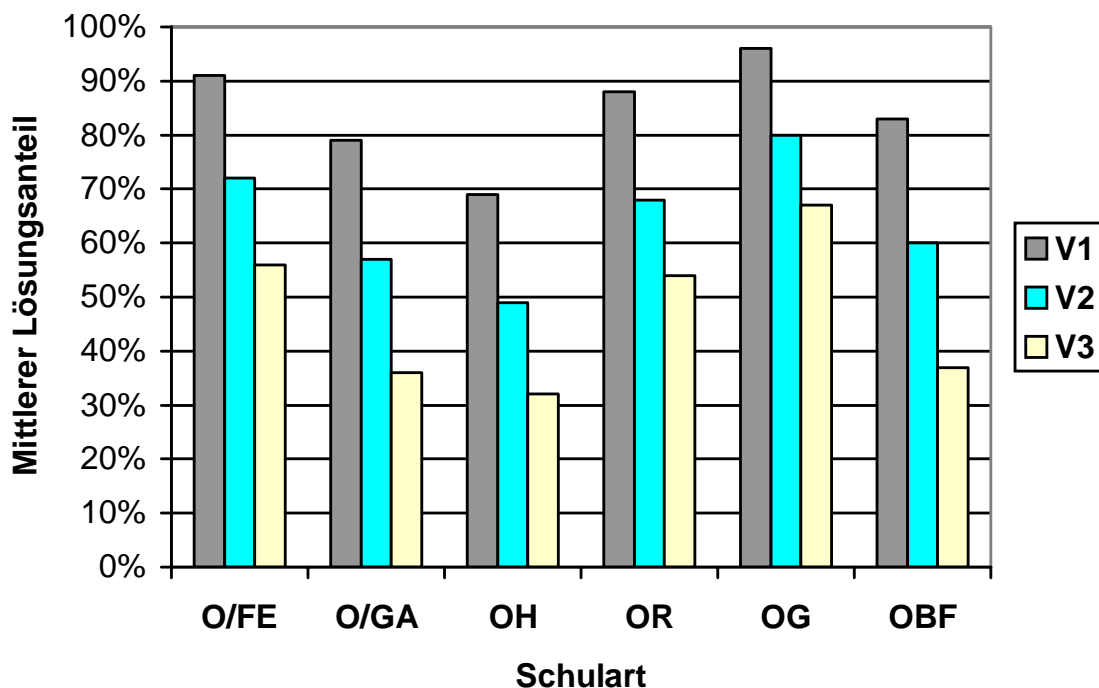
Bei den einzelnen Verständensdimensionen gibt es unterschiedliche maximale Punktwerte, was die Untersuchung erschwert, in welchen der drei Dimensionen Stärken oder Schwächen vorliegen. Die jeweils konkret erreichten Punktzahlen werden daher auf die entsprechenden Maximalwerte bezogen; wir gehen also über zu den mittleren Lösungsanteilen; vgl. Text zwischen den Tabellen A2 und A3. Beispiel: Der Kurs O/FE erreichte im Schnitt 34,5 Punkte von maximal möglichen 50 Punkten; das entspricht einem Anteil von $34,5/50 = 0,69$ oder 69%. Die entsprechenden Umrechnungen enthält die Tabelle B1.1-b, wobei aus Platzgründen die Differenzierung nach dH/ndH entfällt.

B1.1-b Tabelle: Deutsch.
Punktwerte und mittlere Lösungsanteile differenziert nach Schulart.

	O/FE	O/GA	OH	OR	OG	OBF	gesamt
gesamt	34,5 69%	26,5 53%	22,9 46%	33,0 66%	38,9 78%	27,6 55%	32,9 66%
V1	7,3 91%	6,3 79%	5,5 69%	7,0 88%	7,7 96%	6,6 83%	7,0 88%
V2	17,2 72%	13,7 57%	11,7 49%	16,2 68%	19,1 80%	14,3 60%	16,4 68%
V3	10,0 56%	6,5 36%	5,7 32%	9,8 54%	12,1 67%	6,7 37%	9,5 53%

Die unterschiedlich großen Anforderungen, die mit den Verstehensdimensionen verbunden sind, drücken sich in den abfallenden Werten von V1 zu V3 aus. Wie über die Schularten hinweg gleichmäßig oder ungleichmäßig dieser Abfall ist, zeigt die Abbildung B1.2.

B1.2 Abbildung: Deutsch.
Mittlerer Lösungsanteil differenziert nach Verstehensdimensionen V1, V2 und V3 und nach Schularten.



Der Rückgang der Werte ist dort besonders gering, wo insgesamt das Leistungsniveau hoch ist (OG insbesondere), und besonders stark, wo das Leistungsniveau relativ niedrig ist (O/GA, OBF). Eine Ausnahme bildet die Hauptschule, bei der der Rückgang nicht so drastisch ausfällt, so dass bei V3 die Unterschiede zu den anderen Schularten geringer werden.

An dieser Stelle wollen wir die Analyse der Gesamtberliner Werte abrechnen, indem in Analogie zur Tabelle 2A.5 aus dem ersten Bericht mit der Tabelle B1.3 den Lehrkräften ein Schema an die Hand geben, das die Analyse der Ergebnisse der eigenen Klasse erleichtern soll; vgl. auch die abgewandelten Ergebnisblätter am Ende des Anhangs.

B1.3 Tabelle: Deutsch.

Vergleich von Leistungsprofilen: Vorschlag eines Musters, anhand dessen die Werte der eigenen Klasse in den Berliner Zusammenhang gestellt werden können. Angegeben sind für jede Verstehensdimension der Mittelwert und der mittlere Lösungsanteil, d.h. der Mittelwert bezogen auf die maximal mögliche Punktzahl.

	Berlin: Insgesamt	Berlin: Meine Schulart	Meine Klasse
Verstehensdimension 1	7,0 88%		
Verstehensdimension 2	16,4 68%		
Verstehensdimension 3	9,5 53%		
Insgesamt	32,9 66%		

B2 Englisch

Die Entwicklung der Englischaufgaben fand vor dem Hintergrund zweier Kategorien von Leistungsaspekten statt (vgl. <http://www.kmk.org/schul/bildungsstandards>):

Vier Kompetenzbereiche:

- K1:** Kommunikative Fertigkeiten
- K2:** Verfügung über die sprachlichen Mittel
- K3:** Interkulturelle Kompetenzen
- K4:** Methodische Kompetenzen

Drei Niveaustufen aus dem Gemeinsamen Europäischen Referenzrahmen:

- A2:** Elementare Sprachverwendung.
Kann Sätze und häufig gebrauchte Ausdrücke verstehen, die mit Bereichen von ganz unmittelbarer Bedeutung zusammenhängen (z.B. Informationen zur Person und zur Familie, Einkaufen, Arbeit, nähere Umgebung). Kann sich in einfachen, routinemäßigen Situationen verständigen, in denen es um einen einfachen und direkten Austausch von Informationen über vertraute und geläufige Dinge geht. Kann mit einfachen Mitteln die eigene Herkunft und Ausbildung, die direkte Umgebung und Dinge im Zusammenhang mit unmittelbaren Bedürfnissen beschreiben.
- B1:** Selbstständige Sprachverwendung
Kann die Hauptpunkte verstehen, wenn klare Standardsprache verwendet wird und wenn es um vertraute Dinge aus Arbeit, Schule, Freizeit etc. geht. Kann die meisten Situationen bewältigen, denen man auf Reisen im Sprachgebiet begegnet. Kann sich einfach und zusammenhängend über vertraute Themen und persönliche Interessengebiete äußern. Kann über Erfahrungen und Ereignisse berichten, Träume, Hoffnungen und Ziele beschreiben und zu Plänen und Ansichten kurze Begründungen oder Erklärungen geben.
- B1+:** Selbstständige Sprachverwendung +.
[Pragmatisch eingeführte Zwischenstufe auf dem Weg zu B2.¹²]

Die nachstehende exemplarische Auswertung zieht die drei Niveaustufen heran. Die einzelnen Aufgaben lassen sich anhand der Informationen des Entwicklerteams wie folgt zuordnen (in Klammern die maximal zu vergebende Punktzahl, falls sie nicht der Regel eines zu vergebenden Punktes *falsch/richtig* folgt):

- A2: Elementare Sprachverwendung.**
Leseaufgabe 1 (Item L01 bis L05)
Maximal mögliche Punktzahl: 5.

¹² B2: Kann die Hauptinhalte komplexer Texte zu konkreten und abstrakten Themen verstehen; versteht im eigenen Spezialgebiet auch Fachdiskussionen. Kann sich so spontan und fließend verständigen, dass ein normales Gespräch mit Muttersprachlern ohne größere Anstrengung auf beiden Seiten gut möglich ist. Kann sich zu einem breiten Themenspektrum klar und detailliert ausdrücken, einen Standpunkt zu einer aktuellen Frage erläutern und die Vor- und Nachteile verschiedener Möglichkeiten angeben.

B1: Selbststaendige Sprachverwendung.

Höraufgabe 1 und 2	(Item H01 bis H16)
Leseaufgabe 2	(Item L06 bis L15)
Schreibaufgabe 1	(Item S01 bis S10)
Schreibaufgabe 2	(SMS (5))
Schreibaufgabe 3	(letter/task (5), letter/language (5))
Maximal mögliche Punktzahl: 51.	

B1+: Selbststaendige Sprachverwendung Plus.

Höraufgabe 3	(Item H17 bis H25)
Leseaufgabe 3	(Item L16 bis L25)
Maximal mögliche Punktzahl: 19.	

Die Zuordnung auf die drei Niveaustufen ist sehr ungleichmäßig und mag teilweise nicht eindeutig sein, ist aber unter pragmatischen Gesichtspunkten zufriedenstellend und genügt dem hier verfolgten Zweck, exemplarisch zu zeigen, wie sich inhaltlich unterrichtsrelevante Einteilungen abweichend von der äußeren Struktur der Vergleichsarbeit als Grundlage weiterer Auswertungen bestimmen lassen.¹³

Die Tabelle B2.1-a hält die Ergebnisse für die drei Verstehensdimensionen fest. Sie erweitert die Tabelle 2B.1 aus dem ersten Bericht.

Tabelle B2.1-a liefert hinsichtlich der Schularten das zu erwartende Bild, wenn auch der Hauptschulkurs A bessere Werte im Schnitt erzielt als der Gesamtschulkurs GA. Kapitel A hatte bereits gezeigt, was für ein breites Spektrum sich hinter den ndH-Werten verbirgt. Für Englisch gilt nun - in Abweichung der Deutschergebnisse -, dass die ndH-Jugendlichen nicht durchweg schlechtere Resultate erzielen als ihre dH-Mitschüler/innen, teils sind die Unterschiede nicht besonders groß, teils liegen die ndH-Werte über den dH-Werten; vgl. insbesondere den Kurs O/GA.

¹³ Aus Sicht der Testkonstruktion werden hier drei Skalen gebildet, deren jeweiliger Wert sich durch Aufaddieren der Einzelwerte, also durch Summenbildung ergibt. Da hier ein exemplarischer Versuch aus inhaltliche3n Gründen interessiert, soll die statistische Sinnhaftigkeit nicht diskutiert werden. Wir kommen im Kapitel C darauf zurück. Mitgeteilt seien die α -Werte, die die interne Konsistenz quantifizieren:

$$\alpha(A2)=0,71, \alpha(V2)=0,87, \alpha(V3)=0,84.$$

Die Koeffezienten sind deutlich höher als für die Verstehensdimensionen in Deutsch. Sie sind ausreichend hoch, um auch unter statistischen Gesichtspunkten die Auswertung weiter voranzutreiben.

B2.1-a Tabelle: Englisch.**Punktwerte differenziert nach Schulart¹⁴ und Herkunftssprache.**

Angegeben wird jeweils der Mittelwert. A2, B1, B1+: Niveaustufen aus dem *Gemeinsamen Europäischen Referenzrahmen*; siehe Text.

	O/FE	O/GA	OH/A	OH/B,C	OR	OG	OBF	gesamt
gesamt	55,2	41,1	43,7	35,0	54,4	63,1	45,3	54,3
dH	56,1	40,5	43,9	33,2	54,8	63,1	46,4	54,9
ndH	51,7	43,4	43,1	38,8	53,1	63,1	42,8	52,4
A2								
gesamt	4,1	3,1	3,3	2,5	4,1	4,7	3,6	4,1
dH	4,2	3,0	3,3	2,5	4,1	4,6	3,6	4,1
ndH	3,9	3,4	3,1	2,5	4,2	4,7	3,5	4,0
B1								
gesamt	36,4	26,7	29,0	22,7	35,7	41,3	28,6	35,5
dH	36,9	26,6	29,0	21,7	35,9	41,4	29,3	35,9
ndH	34,0	27,2	29,0	24,6	35,0	40,9	26,9	33,9
B1+								
gesamt	14,7	11,4	11,4	9,9	14,6	17,2	13,1	14,8
dH	15,0	10,9	11,5	9,0	14,8	17,1	13,4	14,9
ndH	13,8	12,8	11,0	11,8	13,9	17,4	12,4	14,5

Bei den einzelnen Niveaustufen gibt es unterschiedliche maximale Punktwerte, was den Vergleich erschwert. Die jeweils konkret erreichten Punktzahlen werden daher auf die entsprechenden Maximalwerte bezogen; wir gehen also über zu den mittleren Lösungsanteilen; vgl. Text zwischen den Tabellen A2 und A3. Beispiel: Der Kurs O/FE erreichte im Schnitt 55,2 Punkte von maximal möglichen 75 Punkten; das entspricht einem Anteil von $55,2/75 = 0,74$ oder 74%.¹⁵ Die entsprechenden Umrechnungen enthält die Tabelle B2.1-b, wobei aus Platzgründen die Differenzierung nach dH/ndH entfällt.

¹⁴ Aufgrund der niedrigen Fallzahlen fassen wir die Arbeiten aus den Kursstufen B und C der Hauptschule zusammen.

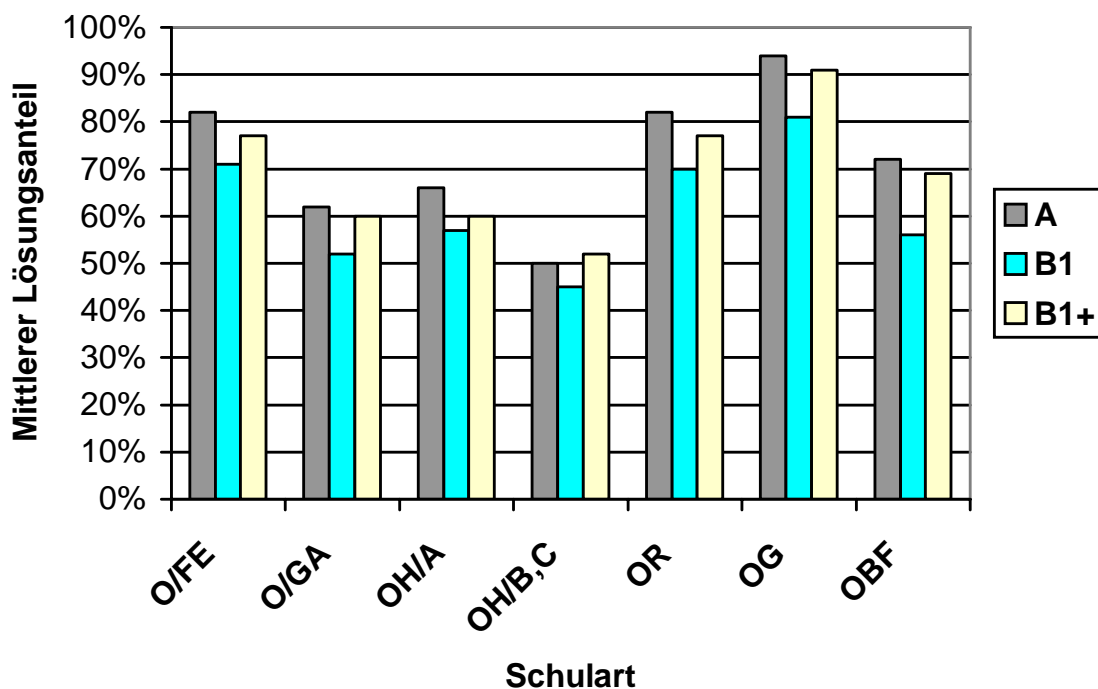
¹⁵ Die Lösungsanteile erlauben auch vorsichtige Vergleiche zwischen den Fächern, vorsichtig, weil wir nicht wissen, ob die Arbeiten "gleich schwer" waren und wie das zu definieren wäre.

B2.1-b Tabelle: Englisch.
Punktwerte und mittlere Lösungsanteile differenziert nach Schulart.

	O/FE	O/GA	OH/A	OH/B,C	OR	OG	OBF	gesamt
gesamt	55,2 74%	41,1 55%	43,7 58%	35,0 47%	54,4 73%	63,1 84%	45,3 61%	54,3 72%
A	4,1 82%	3,1 62%	3,3 66%	2,5 50%	4,1 82%	4,7 94%	3,6 72%	4,1 82%
B1	36,4 71%	26,7 52%	29,0 57%	22,7 45%	35,7 70%	41,3 81%	28,6 56%	35,5 70%
B1+	14,7 77%	11,4 60%	11,4 60%	9,9 52%	14,6 77%	17,2 91%	13,1 69%	14,8 78%

Erwartungsgemäß gibt es bei allen Schularten/Kursen einen Abfall der Werte von den Niveaustufe A zu B, dann allerdings wiederum einen Anstieg von B1 zu B1+. Die Zuordnung der Aufgabenteile zu den Niveaustufen erfolgte vor dem Einsatz der Vergleichsarbeiten im Zuge der Entwicklung, also a priori. Wie die Ergebnisse zeigen, lässt die Zuordnung empirisch sich nicht halten, ein bedeutsames Resultat für die Entwicklerteams, das zu einer genaueren Prüfung der Anforderungen, die mit den Items verbunden sind, nötig. Vor dem nächsten Durchgang sollten die Aufgaben möglichst pilotiert werden, um erste empirische Hinweise auf die Schwierigkeit der Aufgaben zu gewinnen und Fehleinschätzungen zu vermeiden. Wie über die Schularten/Kurse hinweg die Verläufe gleichmäßig oder ungleichmäßig sind, zeigt die Abbildung B2.2.

B2.2 Abbildung: Englisch.
Mittlerer Lösungsanteil differenziert nach den Niveaustufen A, B1 und B1+ und nach Schularten.



B3 Mathematik

Die Entwicklung der Mathematikaufgaben fand vor dem Hintergrund dreier Kategorien von Leistungsaspekten statt (vgl. auch <http://www.kmk.org/schul/bildungsstandards>):

Fünf Leitideen für inhaltsbezogene mathematische Kompetenzen:

- L1:** Zahl
- L2:** Messen
- L3:** Raum und Form
- L4:** Funktionaler Zusammenhang
- L5:** Daten und Zufall.

Sechs allgemeine mathematische Kompetenzen:

- K1:** Mathematisch argumentieren
- K2:** Probleme mathematisch lösen
- K3:** Mathematisch modellieren
- K4:** Mathematische Darstellungen verwenden
- K5:** Mit symbolischen, formalen und technischen Elementen der Mathematik umgehen
- K6:** Kommunizieren.

Drei Anforderungsbereiche:

- A1:** Reproduzieren
- A2:** Zusammenhänge herstellen
- A3:** Verallgemeinern und Reflektieren.

Im Gegensatz zu den Fächern Deutsch und Englisch liegen mit den *Leitideen* Kategorien vor, die nicht Anforderungsniveaus zum Ausdruck bringen, sondern sich auf Unterrichtsthemen beziehen. Für die weitere Analyse ziehen wir daher die *Leitideen* heran.

Bei diesen (wie bei den anderen beiden möglichen Unterteilungen nach *Kompetenzen* und *Anforderungsbereichen*) ist die Besonderheit der Mathematikvergleichsarbeit zu berücksichtigen, dass zwar die Aufgaben 1 bis 13 für alle Schüler/innen verbindlich waren, bei den Aufgaben 14 und 15 jedoch für jede Klasse die Mathematiklehrkraft zwischen drei Versionen wählen konnte: Körperberechnung (K), Trigonometrie (T) und Sachrechnen (S). Die Mathematikvergleichsarbeit liegt demnach in drei Versionen vor. Wie die Mathematikaufgaben - und damit die erreichbaren Punkte - den Einteilungskategorien *Leitidee*, *Kompetenzen*, *Anforderungsbereiche* zuzuordnen sind, geht aus dem Begleittext zur Vergleichsarbeit hervor. Wir listen nachstehend auf, wie viele Punkte maximal pro Kategorie bei jeder der drei Versionen der Arbeit zu erzielen sind, wie also sich die maximal möglichen 45 Punkte Gesamtwert je nach Version der Arbeit aufteilen (Gewichtung der Aspekte in der Vergleichsarbeit durch das Entwicklerteam):

	L1	L2	L3	L4	L5	Σ
Version K	13	13	4	14	1	45
Version T	13	13	4	14	1	45
Version S	22	2	4	16	1	45

	K1	K2	K3	K4	K5	K6	Σ
Version K	1	12	12	6	8	6	45
Version T	1	15	12	6	6	5	45
Version S	2	13	12	6	6	6	45

	A1	A2	A3	Σ
Version K	16	24	5	45
Version T	16	24	5	45
Version S	17	22	6	45

Die einzelnen Aspekte sind sehr unterschiedlich in der Arbeit vertreten. Unsere Analyse wird sich daher auf die beiden Leitideen **L1 Zahl** und **L4 Funktionale Zusammenhänge** beschränken, da nur für diese über alle drei Versionen der Arbeit hinweg eine ausreichende Anzahl von Items (Aufgaben, Teilaufgaben) vorliegt. Die Zuordnung der Items zu den beiden Leitideen ergibt folgende Aufstellung:

L1: Zahl.

Items für alle: 3F, 3H, 3B, 4, 6a, 6b, 8a, 8b, 8c-1, 8c-2, 12-2, 12-3, 12-4
 Zusätzlich für Version K: ---
 Zusätzlich für Version T: ---
 Zusätzlich für Version S: S14-1 bis S14-5, S15-1 bis S15-4.

L4: Funktionale Zusammenhänge.

Items für alle: 1-1a, 1-1b, 1-1c, 1-2, 5a, 5b, 5c, 5d, 7, 11-1 bis 11-5
 Zusätzlich für Version K: ---
 Zusätzlich für Version T: ---
 Zusätzlich für Version S: S15-5, S15-6.

Die Zuordnung zu den beiden Leitideen mag teilweise nicht eindeutig sein, ist aber unter pragmatischen Gesichtspunkten zufriedenstellend und genügt dem hier verfolgten Zweck, exemplarisch zu zeigen, wie sich inhaltlich unterrichtsrelevante Einteilungen abweichend von der äußeren Struktur der Vergleichsarbeit als Grundlage weiterer Auswertungen bestimmen lassen.¹⁶

Es zeigt sich, dass über die allen gemeinsamen Aufgaben aus dem Bereich der ersten dreizehn hinaus nur noch in der Version S weitere Items berücksichtigt werden können, einmal 9, das andere Mal 2 Items. Die Einteilung führte somit zu identischen Versionen K und T,

¹⁶ Aus Sicht der Testkonstruktion werden hier drei Skalen gebildet, deren jeweiliger Wert sich durch Aufaddieren der Einzelwerte, also durch Summenbildung ergibt. Da hier ein exemplarischer Versuch aus inhaltliche3n Gründen interessiert, soll die statistische Sinnhaftigkeit nicht diskutiert werden. Wir kommen im Kapitel C darauf zurück. Mitgeteilt seien die α -Werte, die die interne Konsistenz quantifizieren:

$$\alpha(L1)=0,84, \alpha(L4)=0,84.$$

Sie sind zufriedenstellend hoch, so dass auch unter statistischen Gesichtspunkten die Auswertung fortgeführt werden kann.

während die Version S davon abweicht. Wir entscheiden uns für eine einheitliche Version, die nur die Items aus dem allen gemeinsamen Kern berücksichtigt. Das hat den Vorteil, mit den beiden so entstehenden Skalen L1 und L4 vergleichbare Werte und dies für die gesamte Stichprobe zu erhalten. Der Auswertung liegt demnach folgende Zuordnung zugrunde:

L1: Zahl.

Items für alle: 3F, 3H, 3B, 4, 6a, 6b, 8a, 8b, 8c-1, 8c-2, 12-2, 12-3, 12-4
Maximal mögliche Punkte: 13.

L4: Funktionale Zusammenhänge.

Items für alle: 1-1a, 1-1b, 1-1c, 1-2, 5a, 5b, 5c, 5d, 7, 11-1 bis 11-5
Maximal mögliche Punkte: 14.

Die Zuordnung der Items zu den beiden Leitideen führt zu nahezu gleich langen Skalen. Die Tabelle B3.1-a hält die Ergebnisse für die beiden Leitideen fest. Sie erweitert die Tabelle 2C.1 aus dem ersten Bericht.

B3.1-a Tabelle: Mathematik.**Punktwerte differenziert nach Schulart¹⁷ und Herkunftssprache.**

Angegeben wird jeweils der Mittelwert und in Klammern die Fallzahlen.

	O/FE	O/GA	OH/A	OH/B,C	OR	OG	OBF	gesamt
gesamt	25,0 (N=199)	14,8 (N=207)	19,5 (N=52)	10,9 (N=39)	27,2 (N=390)	33,3 (N=569)	16,2 (N=199)	25,5 (N=1655)
dH	25,1 (N=166)	14,0 (N=154)	18,8 (N=38)	11,2 (N=35)	27,5 (N=312)	33,5 (N=463)	15,4 (N=142)	25,7 (N=1310)
ndH	24,6 (N=33)	17,0 (N=53)	21,4 (N=14)	8,0 (N=4)	25,9 (N=78)	32,6 (N=106)	18,2 (N=57)	24,8 (N=345)
L1: Zahl								
gesamt	7,4	4,3	5,9	2,8	7,6	9,9	5,1	7,5
dH	7,5	3,9	5,5	2,9	7,5	9,9	4,5	7,4
ndH	7,3	5,3	6,7	2,3	8,0	10,0	6,5	7,8
L4: Funktionale Zusammenhänge								
gesamt	7,7	5,1	7,1	4,1	8,5	10,2	6,0	8,1
dH	7,7	4,6	6,7	4,3	8,7	10,3	5,8	8,2
ndH	7,7	6,5	8,2	2,3	7,7	9,8	6,6	7,9

Tabelle B3.1-a liefert hinsichtlich der Schularten das zu erwartende Bild, wobei wiederum - wie in Englisch - der Hauptschulkurs A im Schnitt bessere Werte erzielt als der Gesamtschulkurs GA. Kapitel A hatte bereits gezeigt, was für ein breites Spektrum sich hinter den

¹⁷ Aufgrund der niedrigen Fallzahlen fassen wir die Arbeiten aus den Kursstufen B und C der Hauptschule zusammen.

ndH-Werten verbirgt. In Mathematik zeigt sich nun - in Abweichung der Deutschergebnisse und ausgeprägter als in Englisch -, dass die die ndH-Jugendlichen nicht durchweg schlechtere Resultate erzielen als ihre dH-Mitschüler/innen, teils sind die Unterschiede gering, teils liegen die ndH-Werte über den dH-Werten; vgl. insbesondere die Kurse O/GA und OH/A.

Bei den einzelnen Leitideen gibt es leicht unterschiedliche maximale Punktwerte, was den Vergleich beider Ergebnisse etwas erschwert. Die jeweils konkret erreichten Punktzahlen werden daher auf die entsprechenden Maximalwerte bezogen; wir gehen also über zu den mittleren Lösungsanteilen; vgl. Text zwischen den Tabellen A2 und A3. Beispiel: Der Kurs O/FE erreichte im Schnitt 25,0 Punkte von maximal möglichen 45 Punkten; das entspricht einem Anteil von $25,0/45 = 0,56$ oder 56%.¹⁸ Die entsprechenden Umrechnungen enthält die Tabelle B3.1-b, wobei aus Platzgründen die Differenzierung nach dH/ndH entfällt. Eine Illustration liefert die Abbildung B3.2.

**B3.1-b Tabelle: Mathematik.
Punktwerte und mittlere Lösungsanteile differenziert nach Schulart.**

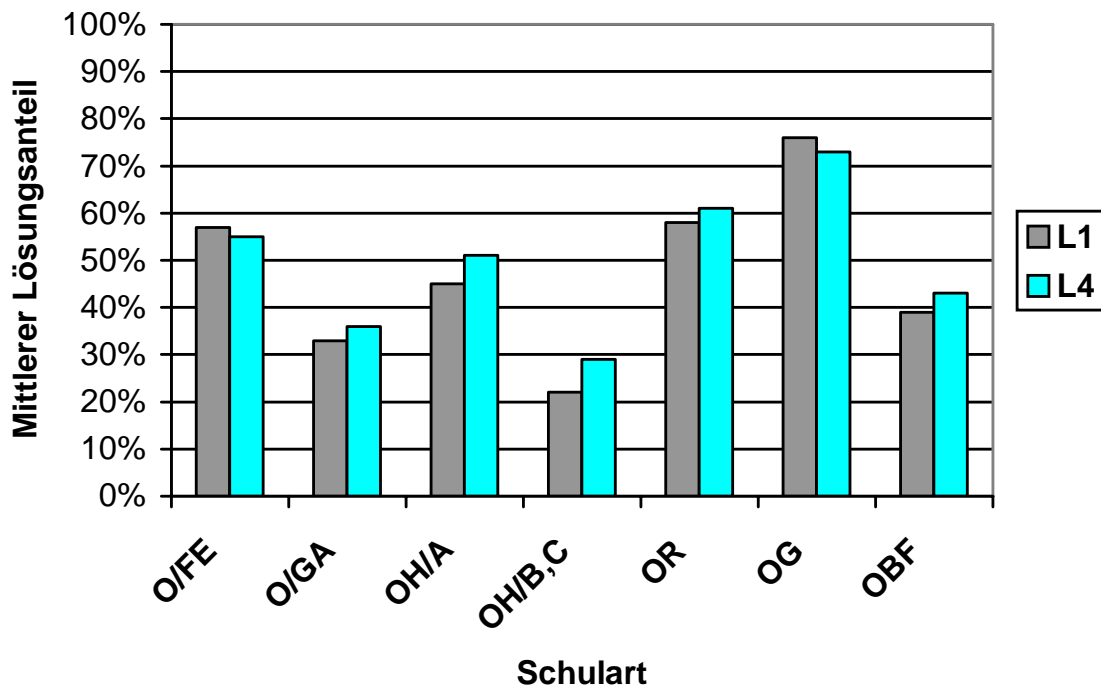
	O/FE	O/GA	OH/A	OH/B,C	OR	OG	OBF	gesamt
gesamt	25,0 56%	14,8 33%	19,5 43%	10,9 24%	27,2 60%	33,3 74%	16,2 36%	25,5 57%
L1	7,4 57%	4,3 33%	5,9 45%	2,8 22%	7,6 58%	9,9 76%	5,1 39%	7,5 58%
L4	7,7 55%	5,1 36%	7,1 51%	4,1 29%	8,5 61%	10,2 73%	6,0 43%	8,1 58%

Die Unterschiede in den Ergebnisse beider Leitideenskalen sind nicht bedeutsam.¹⁹ Allein bei den Hauptschulkursen deutet sich an, dass die Schüler/innen mit den Aufgaben zu den funktionalen Zusammenhängen besser zurecht kamen als mit den Items zum Umgang mit den Zahlen. Für die Hauptschule lässt sich aus diesen Ergebnissen also die Empfehlung ableiten, im Unterricht ein stärkeres Augenmerk auf den Bereich der Leitidee L1 zu legen. Für die anderen Schularten können keine derartige Schlussfolgerungen gezogen werden, jedenfalls nicht auf Landesebene, denn von Schule zu Schule, von Klasse zu Klasse können die Konstellationen völlig unterschiedlich von der hier vorgestellten aussehen, die auf der Mittelung über die gesamte große Stichprobe beruht. An dieser Stelle wollen wir daher die Analyse der Gesamtberliner Werte abbrechen, indem wir in Analogie zur Tabelle 2C.5 aus dem ersten Bericht mit der Tabelle B3.3 den Lehrkräften ein Schema an die Hand geben, das die Analyse der Ergebnisse der eigenen Klasse erleichtern soll; vgl. auch das abgewandelte Ergebnisblatt am Ende des Anhangs.

¹⁸ Die Lösungsanteile erlauben auch vorsichtige Vergleiche zwischen den Fächern, vorsichtig, weil wir nicht wissen, ob die Arbeiten "gleich schwer" waren und wie das zu definieren wäre.

¹⁹ Die Konstellationen verdeutlicht die Abbildung B3.2.

B3.2 Abbildung: Mathematik.
Mittlerer Lösungsanteil differenziert nach den Leitideen L1 und L4 und nach Schularten.



B3.3 Tabelle: Mathematik.
Vergleich von Leistungsprofilen: Vorschlag eines Musters, anhand dessen die Werte der eigenen Klasse in den Berliner Zusammenhang gestellt werden können. Angegeben sind Mittelwerte und mittlere Lösungsanteile.

	Berlin: Insgesamt	Berlin: Meine Schulart	Meine Klasse
gesamt	25,5 57%		
Leitidee L1	7,5 58%		
Leitidee L4	8,1 58%		

C

Fazit

Der vorliegende Bericht führt den ersten Bericht aus dem September 2004 fort. Er stellt in zwei Themenblöcken die Ergebnisse weiterer Auswertungen der schulischen Rückmeldungen vor. Ausdifferenziert werden zum einen die Ergebnisse nach Migrationshintergrund (Kapitel A), zum anderen die Werte in den einzelnen Fächern nach inhaltlichen Teilbereichen, die die mit den Aufgaben verbundenen Anforderungen spezifizieren (Kapitel B).

Zentrale Ergebnisse aus dem Kapitel A:

Kombination Herkunfts-/Verkehrssprache

Im Zuge der Rückmeldungen wurden die Schulen gebeten, Angaben zu zwei Merkmalen des Migrationshintergrundes zu machen, nämlich zur Herkunftssprache (H) und zur Verkehrssprache zu Hause (V). Während das erste Merkmal zweifach gestuft war (Dichotomie: deutsche/nichtdeutsche Herkunftssprache; dH - ndH), wurden beim zweiten vier Kategorien unterschieden (deutsch (d), türkisch (t), russisch (r), andere (a) Sprachen).

Von den acht möglichen Kombinationen wurden für die Auswertung aufgrund ihrer Auftrenshäufigkeiten nur fünf berücksichtigt: dH/dV, ndH/dV, ndH/tV, ndH/rV und ndH/aV.

Vergleichsgrößen

Bei der Analyse sind nicht nur Vergleiche der Gruppen innerhalb eines Faches von Belang, sondern auch zwischen den Fächern. Bei den Vergleichsarbeiten waren aber in den drei Fächern unterschiedlich viele Punkte maximal zu erreichen, so dass gleich hohe Mittelwerte Unterschiedliches bedeuten. Um einen fächerübergreifenden Vergleich der Leistungen zu ermöglichen, wurden zwei Vergleichsgrößen herangezogen:

Zum Einen der Anteil derjenigen, die die vorgegebene Erfolgsschwelle überschritten haben (**Bestehensquote**); zum Anderen wurde der jeweils erzielte Mittelwert auf die Maximalzahl erreichbarer Punkte bezogen, also gefragt, wie hoch die Quote ist, mit der die Maximalzahl ausgeschöpft wird (**Lösungsanteil**).

Fächerübergreifende Ergebnisse

Tabelle C1 liefert einen Überblick über zentrale Ergebnisse in den drei Vergleichsarbeiten; vgl. auch Abbildung A7:

- Nur im Fach Deutsch erreicht die Gruppe dH/dV höhere Werte als die Gruppe ndH/dV, die ansonsten die besten Ergebnisse erzielt (wenn auch ohne statistische Signifikanz).
- Die Sprachkombination ndH/tV weist durchgängig die niedrigsten Leistungswerte auf.
- Die Unterschiede zwischen den Gruppen sind sinnfälligerweise in Deutsch am größten, in Mathematik am kleinsten.
- Ein Vergleich der Fächer zeigt, dass die besten Ergebnisse in Englisch, die schlechtesten in Mathematik erzielt wurden.

C1 Tabelle: Überblick über die Gesamtergebnisse in den drei Fächern

	dH/dV	ndH/dV	ndH/tV	ndH/rV	ndH/aV	gesamt
DEUTSCH						
Fallzahl N	1 417	50	163	51	125	1 806
Quote "bestanden"	87%	86%	52%	80%	81%	83%
Insgesamt: Mittelwert	34,0	31,7	25,6	31,2	31,2	32,9
Mittlerer Lösungsanteil	68%	63%	51%	62%	62%	66%
ENGLISCH						
Fallzahl N	1 497	56	168	56	142	1 518
Quote "bestanden"	81%	88%	59%	80%	84%	79%
Insgesamt: Mittelwert	54,9	55,4	47,2	54,7	55,1	54,3
Mittlerer Lösungsanteil	73%	74%	63%	73%	72%	72%
MATHEMATIK						
Fallzahl N	1 302	48	141	45	116	1 652
Quote "bestanden"	68%	69%	60%	64%	60%	67%
Insgesamt: Mittelwert	25,5	25,7	24,3	25,5	23,8	25,3
Mittlerer Lösungsanteil	57%	57%	54%	57%	53%	56%

In der Gruppe ndH/dV dürften die integrationswilligen und aufstiegsorientierten Familienhäuser vertreten sein. Daher lässt sich das zentrale Ergebnis zur Formel zuspitzen: **Integration lohnt sich.**

Fächerspezifische Ergebnisse

Deutsch: Die Gruppe dH/dV schneidet in beiden Teilen der Arbeit gut ab, wobei dies für den Teil II, der Schreibaufgabe, auf den guten Werten im Modul *Inhalt* beruht. Beim Modul *Darstellung* vermag die Gruppe ndH/dV aufzuschließen. Was könnten die Ursachen für diese Konstellation sein?

Dass die Gruppe dH/dV jedem der Bewertungskriterien als beste genügt, ist erwartungsgemäß. Die Schüler/innen mit der Kombination ndH/dV dürften wahrscheinlich aus integrationswilligen und aufstiegsorientierten Elternhäusern stammen (s.o.), was i.a. einen fördernden Effekt auf die Leistungsbereitschaft der Kinder hat. Zugleich gilt, dass es einfacher ist, sich mit formalen Anforderungen vertraut zu machen, eben mit der sprachlichen Gestaltung, als mit soziokulturellen Inhalten einer Gesellschaft, mit denen man später als die gleichaltrigen Mitschüler/innen konfrontiert wurde. Die Bewertung des *Inhalts* einer Schreibaufgabe²⁰ ist wahrscheinlich stärker kulturell geprägt als die der *Darstellung*, so dass beim *Inhalt* deutlicher als bei der *Darstellung* Differenzen zwischen der beurteilenden Lehrkraft, die nahezu ausschließlich des Typs dH/dV ist, und den Schüler/innen ndH/dV zu Tage treten, ein Vorgang, wenn es ihn denn wie hier postuliert gibt, der i.d.R. unbewusst ablaufen dürfte.

²⁰ Gefordert war, zur Aussage "Wichtigstes Ziel bei der Berufswahl sollte es sein, den Wunschberuf anzustreben" Stellung zu beziehen. Zu den Bewertungskriterien des *Inhalts* und der *Darstellung* vgl. den ersten Bericht oder das Begleitmaterial zu den Vergleichsarbeiten.

Englisch: Für die Vergleichsarbeit Englisch gilt dasselbe Phänomen wie im Fach Deutsch: Die Gruppe ndH/dV ist homogener als die Gruppe dH/dV, so dass wiederum ein höherer Anteil über die kritische Bestehensschwelle kommt, als es nach den Mittelwerten zunächst den Anschein hätte; vgl. die beiden Zeilen "Quote 'bestanden' " und "Insgesamt".

Bei allen fünf Gruppen gibt es einen starken Abfall von den Testteilen *Hören* und *Lesen* hin zu *Schreiben*. Besonders groß ist der Abfall in den Gruppen ndH/tV und ndH/aV.

Mathematik: Der (statistische) Zusammenhang zwischen mittlerem Leistungsniveau und Sprachkombination in Mathematik ist weit weniger ausgeprägt als in den anderen Fächern, belegt, dass in Mathematik - zumindest bei dem in der Vergleichsarbeit verwendeten Typ von Aufgaben - die sprachlichen Fertigkeiten und Fähigkeiten eine relativ geringe Rolle spielen.

Zentrale Ergebnisse aus dem Kapitel B:

Eine diagnostische Funktion für den eigenen Unterricht können Vergleichsarbeiten nur dann erfüllen, wenn sich aus dem Vergleich der Berliner (schulartspezifischen) Landeswerte mit den Werten der eigenen Klasse Hinweise gewinnen lassen, auf welche thematischen Bereiche verstärkt Bemühungen zu richten sind. Im ersten Bericht konzentrierte sich die Mitteilung von Teilergebnissen auf die Unterteilung, die durch die Struktur der Arbeiten vorgegeben war. Neben den Gesamtergebnissen wurden in allen Fächern die Einzelergebnisse für jedes Item (Aufgabe, Teilaufgabe) ausgewiesen. Darüber hinaus wurden z.B. in Deutsch bei der Schreibaufgabe getrennt für die Aspekte *Inhalt* und *Darstellung* die Mittelwerte dokumentiert oder in Englisch für die drei Teiltests *Hören*, *Lesen* und *Schreiben*.

Konzeptorientierte Einteilung der Aufgaben

Im vorliegenden Bericht erfolgt eine Ausdifferenzierung der Gesamtergebnisse nach Teilbereichen, die sich nicht mehr aus der äußeren Struktur, sondern aus der Konzeption ergibt, die den Vergleichsarbeiten zugrundeliegt, um beispielhaft zu zeigen, wie Lehrkräfte die Ergebnisse ihrer Klasse für sich und ihre Arbeit weiter auswerten können. Unter mehreren möglichen Einteilungskriterien wurden für die drei Fächer die folgenden ausgewählt:

Deutsch: *Drei Verstehensdimensionen:*
V1: Informationen ermitteln
V2: Textverständnis entwickeln, Informationen verknüpfen
V3: Texte reflektieren und bewerten.

Englisch: *Drei Niveaustufen aus dem Gemeinsamen Europäischen Referenzrahmen*
A2: Elementare Sprachverwendung
B1: Selbstständige Sprachverwendung
B1+: Selbstständige Sprachverwendung Plus.

Mathematik: *Fünf Leitideen für inhaltsbezogene mathematische Kompetenzen*
L1: Zahl
L2: Messen
L3: Raum und Form
L4: Funktionaler Zusammenhang
L5: Daten und Zufall.

Da nicht alle Leitideen unter statistischen Gesichtspunkten ausreichend durch Items repräsentiert sind, beschränkt sich die Auswertung auf die beiden Leitideen L1 und L4.

Fächerspezifische Ergebnisse

Tabelle C2 liefert einen Überblick über die Ergebnisse. Mehr Details finden sich in den Abschnitten B1 bis B3.

C2 Tabelle: Überblick über die Gesamtergebnisse in den drei Fächern. Angabe Mittelwerte und Lösungsanteile.

Deutsch								
	O/FE	O/GA	OH ²¹	OR	OG	OBF	gesamt	
gesamt	34,5 69%	26,5 53%	22,9 46%	33,0 66%	38,9 78%	27,6 55%	32,9 66%	
V1	7,3 91%	6,3 79%	5,5 69%	7,0 88%	7,7 96%	6,6 83%	7,0 88%	
V2	17,2 72%	13,7 57%	11,7 49%	16,2 68%	19,1 80%	14,3 60%	16,4 68%	
V3	10,0 56%	6,5 36%	5,7 32%	9,8 54%	12,1 67%	6,7 37%	9,5 53%	
Englisch								
	O/FE	O/GA	OH/A	OH/B,C	OR	OG	OBF	gesamt
gesamt	55,2 74%	41,1 55%	43,7 58%	35,0 47%	54,4 73%	63,1 84%	45,3 61%	54,3 72%
A	4,1 82%	3,1 62%	3,3 66%	2,5 50%	4,1 82%	4,7 94%	3,6 72%	4,1 82%
B1	36,4 71%	26,7 52%	29,0 57%	22,7 45%	35,7 70%	41,3 81%	28,6 56%	35,5 70%
B1+	14,7 77%	11,4 60%	11,4 60%	9,9 52%	14,6 77%	17,2 91%	13,1 69%	14,8 78%
Mathematik								
	O/FE	O/GA	OH/A	OH/B,C	OR	OG	OBF	gesamt
gesamt	25,0 56%	14,8 33%	19,5 43%	10,9 24%	27,2 60%	33,3 74%	16,2 36%	25,5 57%
L1	7,4 57%	4,3 33%	5,9 45%	2,8 22%	7,6 58%	9,9 76%	5,1 39%	7,5 58%
L4	7,7 55%	5,1 36%	7,1 51%	4,1 29%	8,5 61%	10,2 73%	6,0 43%	8,1 58%

²¹ In Deutsch gab es im Schuljahr 2003/2004 keine äußere Leistungsdifferenzierung.

Deutsch: Die schulartspezifischen Werte entsprechen den Erwartungen. Ebenso plausibel ist der Rückgang der Lösungsanteile mit den steigenden Anforderungen der Verstehensdimensionen von V1 zu V3. Dieser Rückgang ist aber nicht gleichmäßig bei allen Schularten, sondern dort besonders gering, wo insgesamt das Leistungsniveau hoch ist (OG insbesondere), und besonders stark, wo das Leistungsniveau relativ niedrig ist (O/GA, OBF).²² Von V1 zu V3 nehmen also die Unterschiede zu.

Englisch: Auch hier entsprechen die schulartspezifischen Werte weitgehend den Erwartungen, wenn auch der Hauptschulkurs A im Schnitt bessere Werte erzielt als der Gesamtschulkurs GA. Auf der Niveaustufe A differieren die Leistungswerte zwischen den Schularten/Kursen deutlich höher als auf den beiden anderen Stufen - im Unterschied zu den Ergebnissen in Deutsch, bei denen auf der Ebene der höchsten Anforderungen auch die größten Unterschiede auftreten.

Tabelle B2.1-a zeigt darüber hinaus, dass - in Abweichung von den Deutschergebnissen - die ndH-Jugendlichen nicht durchweg schlechtere Resultate erzielen als ihre dH-Mitschüler/innen, teils sind die Unterschiede nicht besonders groß, teils liegen die nDh-Werte über den dH-Werten; vgl. insbesondere den Kurs O/GA.

Mathematik: Es ergibt sich ein ähnliches Bild wie im Englischen: Erwartungsgemäße schulartspezifische Unterschiede bei im Schnitt besseren Leistungen von OH/A gegenüber O/GA. Die Differenzen zwischen den Schularten/Kursen treten bei der Leitidee L1 stärker hervor als bei L4.

Betrachten wir pro Schulart/Kurs die Unterschiede zwischen den Ergebnissen beider Leitideenskalen, so sind diese nicht bedeutsam. Allein bei den Hauptschulkursen deutet sich an, dass die Schüler/innen mit den Aufgaben zu den funktionalen Zusammenhängen besser zu recht kamen als mit den Items zum Umgang mit den Zahlen. Für die Hauptschule lässt sich aus diesen Ergebnissen also die Empfehlung ableiten, im Unterricht ein stärkeres Augenmerk auf den Bereich der Leitidee L1 zu legen. Für die anderen Schularten können keine derartige Schlussfolgerungen gezogen werden, jedenfalls nicht auf Landesebene.

Fächerübergreifende Ergebnisse

Im Kapitel A wird die Stichprobe anhand ihrer Sprachkonstellation aufgeteilt und die Aufgaben entsprechend der äußeren Struktur der Arbeiten zusammengefasst. Im Kapitel B steht die konzeptorientierte Einteilung der Aufgaben im Vordergrund, die im Hinblick auf die Struktur der Stichprobe nach Schularten/Kursen untersucht wird. Die unterschiedlichen Ansätze führen dennoch zu teilweise ähnlichen Ergebnissen, die entsprechend dem o.a. Muster zusammengefasst werden:

- Die Schulart OH in Deutsch bzw. der Kurs OH/B,C in Englisch und Mathematik weist die niedrigsten Leistungswerte auf.
- In Englisch und Mathematik liegen die Ergebnisse des Kurses OH/A über jenen des Kurses O/GA.

²² Eine Ausnahme bildet die Hauptschule, bei der der Rückgang nicht so drastisch ausfällt, so dass bei V3 die Unterschiede zu den anderen Schularten geringer werden. Das dürfte damit zusammenhängen, dass es in Deutsch im Schuljahr 2003/2004 keine äußere Leistungsdifferenzierung gab.

- Die Unterschiede zwischen den Schularten/Kursen sind für die Leitideen in Mathematik am größten, in Deutsch für die Verstehensdimensionen am kleinsten.
- Ein Vergleich der Fächer zeigt wiederum, dass die besten Ergebnisse in Englisch, die schlechtesten in Mathematik erzielt wurden.

Von Schule zu Schule, von Klasse zu Klasse können die Konstellationen ganz anders als die hier vorgestellten aussehen, die auf der Mittelung über die gesamte große Stichprobe beruhen. Der Wert der dokumentierten Ergebnisse liegt in dem, was mit dem Bericht selber nicht leistbar ist, wofür er aber die Voraussetzung schafft: Zu prüfen, wie es um die eigene Klasse bestellt ist (war).

Aus dem Anhang:

Der Anhang enthält einige Ergänzungen und öffnet Perspektiven für weitere Auswertungen, die an künftigen Datensätzen vorgenommen werden können. Es wird auf den Anhang verwiesen, aus dem an dieser Stelle nur zwei Ergebnisse zusammengefasst werden sollen.

Korrelation zwischen den Zensuren und Ergebnissen der Vergleichsarbeiten: Die Korrelationskoeffizienten sind zwar alle statistisch signifikant von Null verschieden, aber nur mittelhoch, d.h. es liegt ein substanzieller Zusammenhang vor, der aber nicht sehr stark ist. Die Einschätzung des Leistungsniveaus anhand der Zensuren ist ähnlich, aber bei weitem nicht identisch mit der der Vergleichsarbeiten. Dies ist durchaus plausibel, denn Vergleichsarbeiten haben einen anderen Zugriff auf die Kenntnisse und Fähigkeiten der Schüler/innen als Klassenarbeiten oder der Prozess der Zensurengebung. In Deutsch und in Englisch ist der Zusammenhang zwischen Zensur und Gesamtergebnis in etwa gleich stark ausgeprägt, in Mathematik deutlich schwächer. Weiter als in den anderen Fächern scheinen in Mathematik die Aufgaben der Vergleichsarbeit von jenen im Unterricht behandelten oder in Klassenarbeit vorkommenden entfernt zu sein.

Fächerübergreifende Korrelationen: Über die Schülernummer können die drei Datensätze Deutsch, Englisch und Mathematik miteinander verknüpft werden, so dass die fachspezifischen Zensuren und Ergebnisse in den Vergleichsarbeiten miteinander verglichen werden können.

Erwartungsgemäß sind die Korrelationen der Zensuren zwischen den beiden sprachlichen Fächern erheblich höher als mit der Zensur in Mathematik. Dies deckt sich mit der weit verbreiteten Annahme, dass sprachliche Fähigkeiten zu einem großen Teil unabhängig von mathematischen vorhanden sein können.

Die Korrelationen zwischen den Gesamtpunktzahlen der drei Vergleichsarbeiten liegen höher, teilweise deutlich höher, als die Korrelationen zwischen den Zensuren, und alle drei in etwa derselben Größenordnung. Offensichtlich enthalten die Aufgaben aus allen drei Vergleichsarbeiten Anforderungen, die weniger fachabhängig sind als jene, die sich in den Bewertungen der Lehrkräfte als Zensuren niederschlagen.

Folgerungen:

Die Auswertung stieß auf zwei Grenzen, die eine ist quantitativer, die andere qualitativer Natur.

Ad 1: Beschränkung durch die Stichprobengröße

Es war zwar möglich, die Gesamtgruppe der Schüler/innen mit Migrationshintergrund anhand der beiden Merkmale Herkunfts- und Verkehrssprache zu unterteilen, aber einige der Untergruppen waren so klein, dass sie bei der Auswertung nicht berücksichtigt werden konnten oder keine statistisch zuverlässigen Aussagen ermöglichten. Bei knapp 40000 Schülern/innen insgesamt und rund 2000 Rückmeldungen²³ umfasst die Stichprobe etwas über 5% der Grundgesamtheit. Dies ist zwar ausreichend, um zuverlässige Landesmittelwerte - auch differenziert nach Schularten - zu bestimmen, unterschreitet aber rasch die Mindestgröße von Teilgruppen, wenn mehrere Aspekte zugleich - wie bei der Ausdifferenzierung des Migrationshintergrundes - untersucht werden sollen.

Da für die Vergleichsarbeiten im Frühjahr 2005 flächendeckende Rückmeldungen vorgesehen sind, also eine Gesamterhebung, verringert sich das Problem, wenn es auch nicht ganz verschwindet, nämlich dann nicht, wenn auch in der Grundgesamtheit bestimmte Konstellationen sehr selten auftreten. Dann aber ist die damit verknüpfte Fragestellung zumindest quantitativ nicht relevant. Möglich wird es dann sein, die Herkunftssprache aufzufächern. Hinter der Kategorie *nichtdeutscher Herkunftssprache* verbirgt sich ein heterogenes Spektrum, das nach differenzierter Betrachtung verlangt.

Ad 2: Beschränkung durch die Aufgaben

Im ersten Bericht wurden inhaltliche Teilergebnisse in den einzelnen Fächern so weit mitgeteilt, wie sie sich aus der Struktur der jeweiligen Arbeit in Deutsch, Englisch oder Mathematik ergaben. Das waren Lösungshäufigkeiten zum einen für jedes Einzelitem und zum zweiten für Itemgruppen, z.B. für die Schreibaufgabe in Deutsch jeweils nach Inhalt und Darstellung unterschieden oder für die Teiltests Hören, Lesen, Schreiben in Englisch.

Eine derartige Unterteilung kann bereits wichtige diagnostische Hinweise liefern, um Ansatzpunkte für die Weiterentwicklung des eigenen Unterrichts zu finden. Aber im Falle aggregierter Werte sind die Hinweise zu grob (Beispiel: Teiltest *Hören* in Englisch) oder lassen sich nur mittelbar auf den Unterricht beziehen (Beispiel: Teil I der Deutscharbeit). Und Hinweise, die sich aus Einzelitems ergeben, verharren inhaltlich zu sehr im Detail und zudem ist ihre statistische Aussagekraft gering, da Unterschiede zwischen den Werten der eigenen Klasse und (schulartspezifischen) Bezugswerten auf Landesebene häufig nicht signifikant sein dürften.

Der für das Kapitel B des vorliegenden Berichts gewählte Ansatz versucht daher, eine mittlere Aggregationsebene zu finden, sich also nicht im Einzelnen zu verlieren, noch im Gesamten das Profil verschwinden zu lassen, wobei durch die Abkehr von der äußeren Struktur der Vergleichsarbeit hin zu den konzeptionellen Aspekten der Aufgabenentwicklung der inhaltliche Bezug zur Unterrichtspraxis gestärkt werden sollte. Dies ist nur teilweise gelungen.

²³ Vgl. Tabelle 1.1 im September-Bericht.

Allein in Mathematik liegen mit den Leitideen Einteilungsgesichtspunkte vor, die unmittelbar mit den Unterrichtsinhalten zusammenhängen, während in den Fächern Deutsch und Englisch mit den Verstehensdimensionen, Niveaustufen etc. dimensional andere Ordnungsaspekte vorliegen, die Anforderungen und Kompetenzen thematisieren. Erschwerend kommt hinzu, dass die Items nicht eindeutig nach den Einteilungskriterien gruppiert werden können, dass es also verschiedene Gründe haben kann, wenn eine Aufgabe nicht bewältigt wird.

Um die praktische Relevanz der Vergleichsarbeiten zu erhöhen, ist es erforderlich, möglichst trennscharfe Items zu finden, "analytische" Aufgaben, aus denen sich eine gewissermaßen atomare Struktur der Vergleichsarbeit ergibt. Durch eine weitgehend eindeutige Zuordnung von Aufgabe zur damit verbundenen Anforderung lässt sich dann im Idealfall unmittelbar aus dem über- oder unterdurchschnittlichen Abschneiden folgern, wo die Stärken und Schwächen der eigenen Klasse liegen.

Vergleichsarbeiten sind keine Klassenarbeiten, sie stellen einen prinzipiell anderen Typ der Leistungsfeststellung dar. Beide Typen ergänzen sich nicht nur deswegen, weil sie unterschiedliche Zeiträume abdecken. Die mehr gesamtheitlich konzipierten Aufgaben der Klassenarbeit sind sinnvoll, weil sie dem Vorgehen im Unterricht entsprechen. Sie sind aber in ihrer diagnostischen Funktion sehr beschränkt, die bei den Vergleichsarbeiten im Vordergrund steht.

Das Entwickeln derartiger zur Diagnose geeigneten Vergleichsarbeiten dürfte wesentlich durch ein zweidimensionales Schema erleichtert werden, das neben den kognitiven Anforderungen (Beispiel: Kompetenzbereiche, Niveaustufen) Inhalte, Themen berücksichtigt (Beispiel: Leitideen). Es ist vordringlich, dass für alle Fächer derartige verbindliche Schemata gefunden werden, deren einzelne Zellen gezielt durch entsprechende Aufgaben abzudecken sind.²⁴ Dabei ist durchaus denkbar, dass nicht alle Zellen in einer Vergleichsarbeit wg. des beschränkten Umfangs repräsentiert sein können, so dass von Jahr zu Jahr die Module, aus denen eine Vergleichsarbeit sich zusammensetzt, wechseln.

Die Relevanz der Vergleichsarbeiten für die Unterrichtspraxis, ihre diagnostische Funktion muss gestärkt werden. Das verlangt Anstrengungen von beiden Seiten, von den "Produzenten" und von den "Abnehmern". Die Entwicklerteams sind gefordert, bei der Entwicklung der Aufgaben dem spezifischen Charakter von Vergleichsarbeiten Rechnung zu tragen und möglichst trennscharfe, möglichst analytische Aufgaben zu finden. Die Lehrkräfte müssen in bislang wenig geübter Praxis eine diagnostisch orientierte Auswertung der Arbeit vornehmen, indem nicht allein die Gesamtergebnisse in Bezug zu den Vergleichswerten gesetzt werden, sondern auch von wohldefinierten Teilbereichen oder - falls erforderlich - von Einzelaufgaben.

²⁴ Die Items einer Zelle lassen sich dann zu einer Skala zusammenfassen, wie das hier beispielhaft für die Leitideen gezeigt wurde. Ob die Skalenbildung zulässig ist, lässt sich allerdings nur nachträglich empirisch-statistisch anhand der schulischen Rückmeldungen überprüfen oder durch eine vorgängige Pilotierung.

A N H A N G

Migrationshintergrund (Sprachkombination) und Schulart

Z1 Tabelle: Verteilung der Stichprobenteilnehmer/innen an der Vergleichsarbeit Deutsch auf die Schularten differenziert nach der Kombination aus Herkunfts- und Verkehrssprache zu Hause.²⁵

Angegeben werden absolute Häufigkeiten und die prozentualen Anteile innerhalb der Sprachengruppen (Spaltenprozente). H/V: Herkunfts-/Verkehrssprache; d: deutsch, nd: nichtdeutsch, t: türkisch, r: russisch, a: andere Sprachen.

Sprachkombination	dH - dV	ndH - dV	ndH - tV	ndH - rV	ndH - aV	gesamt
Gesamtschule	395 27%	16 31%	37 22%	19 37%	35 26%	502 27%
Hauptschule	98 7%	2 4%	27 16%	2 4%	11 8%	140 8%
Realschule	321 22%	10 20%	44 26%	4 8%	24 18%	403 22%
Gymnasium	477 33%	19 37%	28 17%	18 35%	47 35%	589 32%
Berufsschule (OBF)	168 12%	4 8%	34 20%	9 17%	19 14%	234 13%
gesamt	1 459 100%	51 100%	170 100%	52 100%	136 100%	1 868 100%

Die unterschiedlichen Verteilungen auf die Schularten lassen sich Indizien für unterschiedliche Bildungsansprüche deuten, wobei allerdings das unterschiedliche Leistungsniveau berücksichtigt werden müsste. Zu untersuchen wäre z.B., ob bei derselben Leistung in allen Gruppen dieselbe Chance besteht, auf das Gymnasium zu gehen. Aufgrund der geringen Gruppengrößen lässt sich keine aussagekräftige Analyse durchführen.

Die Werte der Tabelle Z1 haben eine hinweisende, eine hypothesenspendende Funktion. Beispiel: Die Gruppe ndH/dV besitzt im Vergleich zur Gesamtstichprobe (rechte Spalte) überdurchschnittliche Anteile bei der Gesamtschule und dem Gymnasium, was sich als überdurchschnittliche Bildungsaspiration interpretieren ließe, eine Vermutung, die auf der Grundlage einer erheblich größeren Datenbasis zu überprüfen ist.

²⁵ Diese Verteilung auf die Schularten gilt in etwa auch für die anderen Fächer, denn die Datensätze überlappen sich hinsichtlich der Schüler/innen weitestgehend; vgl. die Bemerkung im Anschluss an die Tabelle A1.

Leistung und Geschlecht

Z2-a Tabelle: Deutsch.

Punktwerte differenziert nach Schulart und Geschlecht.

Angegeben wird jeweils der Mittelwert, für das Gesamtergebnis ebenfalls der mittlere Lösungsanteil und bei der Bestehensquote stellvertretend für alle Ergebnisse die Fallzahlen.

	O/FE	O/GA	OH	OR	OG	OBF	gesamt
Teil I							
männlich	25,8	22,1	17,5	24,7	27,8	21,7	24,3
weiblich	25,6	19,1	19,6	24,9	28,6	22,0	25,2
gesamt	25,7	21,0	18,3	24,8	28,2	21,8	24,8
II/Inhalt							
männlich	5,7	3,8	2,4	5,3	6,5	3,0	4,9
weiblich	5,9	3,2	3,8	5,6	7,3	4,0	5,7
gesamt	5,8	3,6	2,9	5,4	6,9	3,5	5,3
II/Darstellung							
männlich	3,0	2,0	1,2	2,7	3,6	2,2	2,7
weiblich	3,1	1,6	1,7	2,9	3,9	2,4	3,0
gesamt	3,0	1,9	1,4	2,8	3,8	2,3	2,9
Teil II							
männlich	8,7	5,8	3,6	7,9	10,1	5,3	7,6
weiblich	8,9	4,7	5,5	8,5	11,1	6,4	8,8
gesamt	8,8	5,4	4,3	8,2	10,7	5,8	8,2
Gesamtpunktzahl							
männlich	34,5 69%	27,9 56%	21,1 42%	32,7 65%	37,9 76%	26,9 54%	31,9 64%
weiblich	34,6 69%	23,9 48%	25,1 50%	33,4 67%	39,8 80%	28,4 57%	34,0 68%
gesamt	34,5 69%	26,5 53%	22,6 45%	33,0 66%	38,9 78%	27,6 55%	32,9 66%
Bestehensquote							
männlich	95% (N=95)	67% (N=124)	31% (N=93)	88% (N=206)	100% (N=286)	65% (N=117)	81% (N=921)
weiblich	95% (N=116)	44% (N=68)	53% (N=53)	85% (N=187)	97% (N=311)	70% (N=105)	84% (N=840)
gesamt	95% (N=211)	59% (N=192)	39% (N=146)	87% (N=393)	98% (N=597)	67% (N=222)	82% (N=1761)

Z2-b Tabelle: Englisch.**Punktwerte differenziert nach Schulart und Geschlecht.**

Angegeben wird jeweils der Mittelwert, für das Gesamtergebnis ebenfalls der mittlere Lösungsanteil und bei der Bestehensquote stellvertretend für alle Ergebnisse die Fallzahlen.

	O/FE	O/GA	OH/A	OH/B,C	OR	OG	OBF	gesamt
Hören								
männlich	19,8	15,4	15,6	13,4	19,1	22,4	17,5	19,4
weiblich	19,0	15,0	15,2	16,3	18,8	22,2	17,1	19,5
gesamt	19,4	15,2	15,4	14,0	19,0	22,3	17,3	19,4
Lesen								
männlich	19,5	15,3	16,3	12,6	19,7	21,4	17,8	19,1
weiblich	18,6	14,6	16,3	14,8	19,2	21,4	16,2	19,0
gesamt	19,0	15,0	16,3	13,1	19,4	21,4	17,0	19,1
Schreiben								
männlich	16,2	10,8	13,5	7,6	15,9	19,2	13,0	15,7
weiblich	17,4	11,3	13,9	9,0	16,1	19,7	9,2	16,2
gesamt	16,8	11,0	13,6	7,9	16,0	19,4	11,0	16,0
Gesamtpunktzahl								
männlich	55,4 74%	41,4 55%	45,4 61%	33,6 45%	54,7 73%	63,0 84%	48,3 64%	54,3 72%
weiblich	54,9 73%	40,8 54%	45,4 61%	40,1 53%	54,1 72%	63,3 84%	42,5 57%	54,7 73%
gesamt	55,2 74%	41,1 55%	45,4 61%	35,0 47%	54,4 73%	63,1 84%	45,3 60%	54,5 73%
Bestehensquote								
männlich	91% (N=127)	46% (N=122)	54% (N=28)	15% (N=53)	86% (N=188)	97% (N=373)	63% (N=98)	79% (N=989)
weiblich	87% (N=123)	40% (N=112)	68% (N=22)	33% (N=15)	88% (N=184)	98% (N=384)	48% (N=101)	81% (N=941)
gesamt	89% (N=250)	43% (N=234)	60% (N=50)	19% (N=68)	87% (N=372)	97% (N=757)	55% (N=199)	80% (N=1930)

Z2-c Tabelle: Mathematik. Aufteilung der Vergleichsarbeiten Mathematik auf die drei Wahlbereiche Körperberechnung (K), Trigonometrie (T) und Sachrechnen (S) differenziert nach Schulart und Geschlecht. Angegeben sind die absoluten Häufigkeiten, in den *Summenzeilen* die Zeilenprozent, ansonsten Spaltenprozent je jeweils bezogen auf die Untergruppen *männlich*, *weiblich* und *gesamt*.

	O/FE	O/GA	OH/A	OH/B,C	OR	OG	OBF	gesamt
männlich								
K	50 48%	57 53%	21 64%	17 77%	69 33%	122 42%	18 19%	354 42%
T	50 48%	19 18%	5 15%	3 14%	137 67%	165 58%	15 16%	394 46%
S	4 4%	32 30%	7 21%	2 9%	0 0%	0 0%	61 65%	106 12%
Summe	104 12%	108 13%	33 4%	22 3%	206 24%	287 34%	94 11%	854 100%
weiblich								
K	41 43%	51 52%	5 36%	8 53%	64 35%	113 35%	34 32%	316 38%
T	49 52%	14 14%	1 7%	4 27%	121 65%	206 65%	1 1%	396 48%
S	5 5%	33 34%	8 57%	3 20%	0 0%	0 0%	70 67%	119 14%
Summe	95 11%	98 12%	14 2%	15 2%	185 22%	319 38%	105 13%	831 100%
gesamt								
K	91 46%	108 52%	26 55%	25 68%	133 34%	235 39%	52 26%	670 40%
T	99 50%	33 16%	6 13%	7 19%	258 66%	371 61%	16 8%	790 47%
S	9 5%	65 32%	15 32%	5 14%	0 0%	0 0%	131 66%	225 13%
Summe	199 12%	206 12%	47 3%	37 2%	391 23%	606 36%	199 12%	1685 100%

Die beiden Tabellen Z2-c und Z2-d wurden vor allem aus Platzgründen getrennt, bilden aber eine Einheit, auch wenn Z2-c einige Informationen enthält (keine wesentlichen Unterschiede in der Verteilung der Geschlechter auf die Schularten/Kurse), die für sich von Bedeutung sind. Tabelle Z2-d jedenfalls ist nur dann angemessen zu interpretieren, wenn die Fallzahlen berücksichtigt werden, d.h. wenn beiden Tabellen zusammen gelesen werden.

Z2-d Tabelle: Mathematik.**Punktwerte differenziert nach Schulart und Geschlecht.**

Angegeben wird jeweils der Mittelwert, für das Gesamtergebnis ebenfalls der mittlere Lösungsanteil. Wg. der teilweise sehr kleinen Fallzahlen ist Tabelle Z2-c zu beachten!

	O/FE	O/GA	OH/A	OH/B,C	OR	OG	OBF	gesamt
Teil 1 (Aufgaben 1 bis 13)								
männlich	21,0	13,8	17,6	9,7	21,5	26,6	15,2	21,0
weiblich	17,4	10,9	13,9	9,8	19,4	24,9	13,5	19,2
gesamt	19,3	12,4	16,5	9,8	20,5	25,7	14,3	20,1
Wahlbereich Körperberechnung								
männlich	5,3	3,2	3,6	0,9	7,2	8,3	2,5	5,9
weiblich	5,2	2,7	1,6	0,9	6,3	7,3	1,9	5,3
gesamt	5,3	3,0	3,2	0,9	6,8	7,8	2,1	5,6
Wahlbereich Trigonometrie								
männlich	5,9	1,0	1,0	1,0	6,8	8,0	1,9	6,6
weiblich	6,1	1,7	2,0	0,0	6,3	7,5	0,0	6,7
gesamt	6,0	1,3	1,2	0,4	6,5	7,7	1,8	6,6
Wahlbereich Sachrechnen								
männlich	7,8	2,1	1,4	1,5	---	---	3,1	2,8
weiblich	6,4	1,6	1,9	2,7	---	---	0,8	1,4
gesamt	7,9	1,9	1,7	2,2	---	---	1,9	2,1
Gesamtpunktzahl								
männlich	26,7 59%	16,3 36%	20,9 46%	11,0 24%	28,4 63%	34,7 77%	18,0 40%	26,9 60%
weiblich	23,1 51%	13,1 29%	16,3 36%	10,8 24%	25,7 57%	32,3 72%	14,7 33%	24,6 55%
gesamt	25,0 56%	14,8 33%	19,5 43%	10,9 24%	27,1 60%	33,5 74%	16,2 36%	25,7 57%
Bestehensquote								
männlich	80%	24%	44%	13%	84%	98%	36%	72%
weiblich	63%	14%	19%	6%	72%	93%	22%	64%
gesamt	72%	19%	37%	10%	78%	96%	29%	68%

Herausfinden bemerkenswerter Items am Beispiel Mathematik: Itemschwierigkeiten in Untergruppen

Nicht nur im generellen Leistungsniveau können Untergruppen (z.B.: Jungen/Mädchen, Schularten) sich unterscheiden, sondern auch im Umgang mit einzelnen Items. Im ersten Fall müssten die Lösungshäufigkeiten in der einen Gruppe sich bei allen Aufgaben in etwa um einen stets gleichbleibenden Betrag von den Lösungshäufigkeiten in der anderen Gruppe unterscheiden. Die Differenz wäre bei allen Items (in etwa) gleich. Wie stellt der zweite Fall sich dar?

Tabelle Z2-d zeigt, dass zwar die Mädchen in der Regel nicht so hohe Leistungswerte wie die Jungens erzielen. Das schließt aber nicht aus, dass es eine ganze Reihe Aufgaben geben mag, bei denen die Abstände zwischen den Geschlechtern geringer als ansonsten sind oder bei denen die Mädchen besser als die Jungens abschneiden. Es liegt dann eine Wechselwirkung zwischen Aufgabe und Geschlecht vor, d.h. offensichtlich lassen sich in solchen Fällen Leistungsunterschiede zwischen den Geschlechtern nicht einfach durch einen allgemein wirksamen Fähigkeitsunterschied erklären, sondern beruhen auf unterschiedlichen Wahrnehmungen und Herangehensweisen.

Derartige Items sind nicht nur von besonderem testtheoretischem wie testpraktischem, sondern auch von fachdidaktischem Interesse,²⁶ können sie doch Hinweise auf differenzierte Auseinandersetzungen mit Anforderungen liefern und somit auch Hinweise auf die Erfordernis differenzierter Formen der Vermittlung. Nachstehend soll am Beispiel der Mathematikaufgaben des ersten Teils der Vergleichsarbeit, also der Aufgaben 1 bis 13, gezeigt werden, wie sich solche Items grafisch bemerkbar machen.

Tabelle Z3-a listet die Lösungshäufigkeiten, die Indikatoren für die Itemschwierigkeit, getrennt nach Geschlecht auf. In der nachfolgenden Abbildung Z3-b sind diese Lösungshäufigkeiten Item für Item eingetragen. Die Lösungshäufigkeiten bilden demnach die Koordinaten der Punkte, die für die einzelnen Items stehen. (Die beiden Punkte links unten stehen für die Items 12-b und 12-c mit den Lösungshäufigkeiten $J(\text{ungen})=27\%/M(\text{ädchen})=26\%$ bzw. $J=29\%/M=26\%$.) Durch den so entstehenden Punkteschwarm wird eine möglichst gut angepasste Gerade gelegt. Bestünde nur eine konstante Differenz in den Lösungshäufigkeiten der Geschlechter müssten alle Punkte auf der Geraden liegen.

Items, die im oben skizzierten Sinne darüber hinaus in Wechselwirkung mit dem Geschlecht stehen, weichen von der Geraden ab, wobei für die Itemanalyse alleine jene interessieren, die am weitesten weg liegen. Diese müssen anhand der Tabelle Z3-a bestimmt werden. Es sind dies Item 08-c ($J=47\%$, $M=51\%$), 12-a ($J=54\%$, $M=36\%$) und 10-a ($J=67\%$, $M=71\%$). Diese Items wären nun unter fachlichen Gesichtspunkten auf ihre Anforderungen und mögliche Lösungsansätze hin zu analysieren.

²⁶ vgl. LIND & KNOCHE 2004. Das hier skizzierte Vorgehen ist stark vereinfachend.

**Z3-a Tabelle: Mathematik.
Lösungshäufigkeiten differenziert nach Geschlecht.**

ITEM	männlich	weiblich	Berlin: Insgesamt
01: Fahnenmast a: immer langsamer	80%	69%	75%
b: immer schneller	77%	63%	70%
c: gleichbleibend	91%	82%	87%
d: am schnellsten	87%	78%	83%
02: Grundstücksansicht: Perspektivwechsel	95%	93%	94%
03: Geschwindigkeit a: Fußgänger	79%	70%	75%
b: Haarwachstum	85%	78%	82%
c: Brieftaube	72%	65%	69%
04: Näherungswert	84%	84%	84%
05: Funktionsgrafen a: Parabel	76%	76%	76%
b: Gerade nach oben	66%	63%	65%
c: Gerade nach oben	74%	75%	74%
d: Gerade nach unten	65%	60%	62%
06: Computerkauf a: Endpreis 1	80%	72%	76%
b: Gespartes Geld	82%	72%	77%
07: Lineare Gleichung: Grundmenge Q	46%	41%	44%
08: Potenzgesetze a: Multiplizieren	64%	63%	64%
b: Dividieren	56%	52%	54%
c: Zahlen kürzen	47%	51%	49%
d: Variable kürzen	31%	32%	31%
09: Dreieck konstruieren: Konstruierbarkeit	83%	78%	80%
10: Kongruenzsatz a: Lösung	67%	71%	69%
b: Begründung	44%	42%	43%
11: Vereinskasse a: Anzahl der Scheine	45%	44%	43%
b: Ansatzgleichung	44%	35%	39%
c: Bestimmung von a	43%	36%	40%
d: Bestimmung von b	44%	38%	41%
e: Antwortsatz formuliert	62%	51%	57%
12: Bevölkerung a: Jahresangabe	54%	36%	45%
b: Formelumgang	27%	26%	27%
c: Berechnung von p	29%	26%	27%
d: Antwortsatz formuliert	39%	37%	38%
13: Streckenlänge a: Pythagoras angewendet	36%	28%	32%
b: Richtige Lösung	34%	27%	30%

Z3-b Abbildung: Mathematik.
Lösungshäufigkeiten differenziert nach Geschlecht.

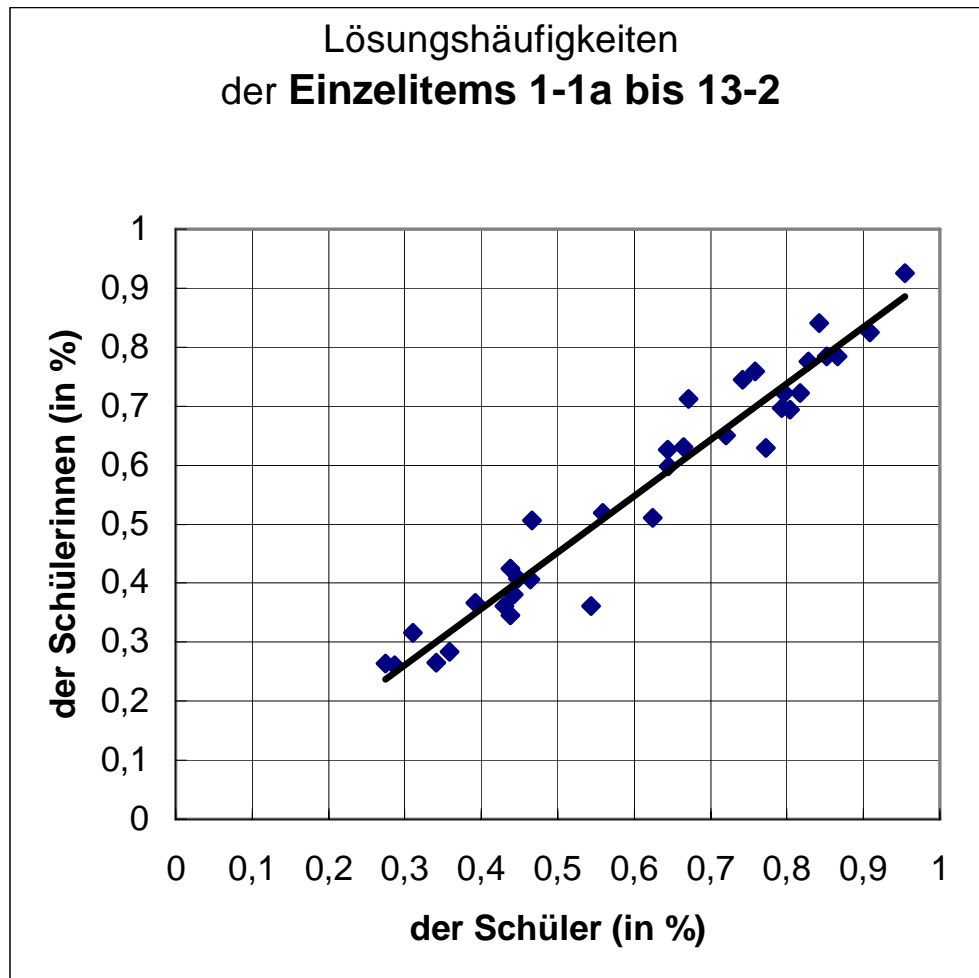


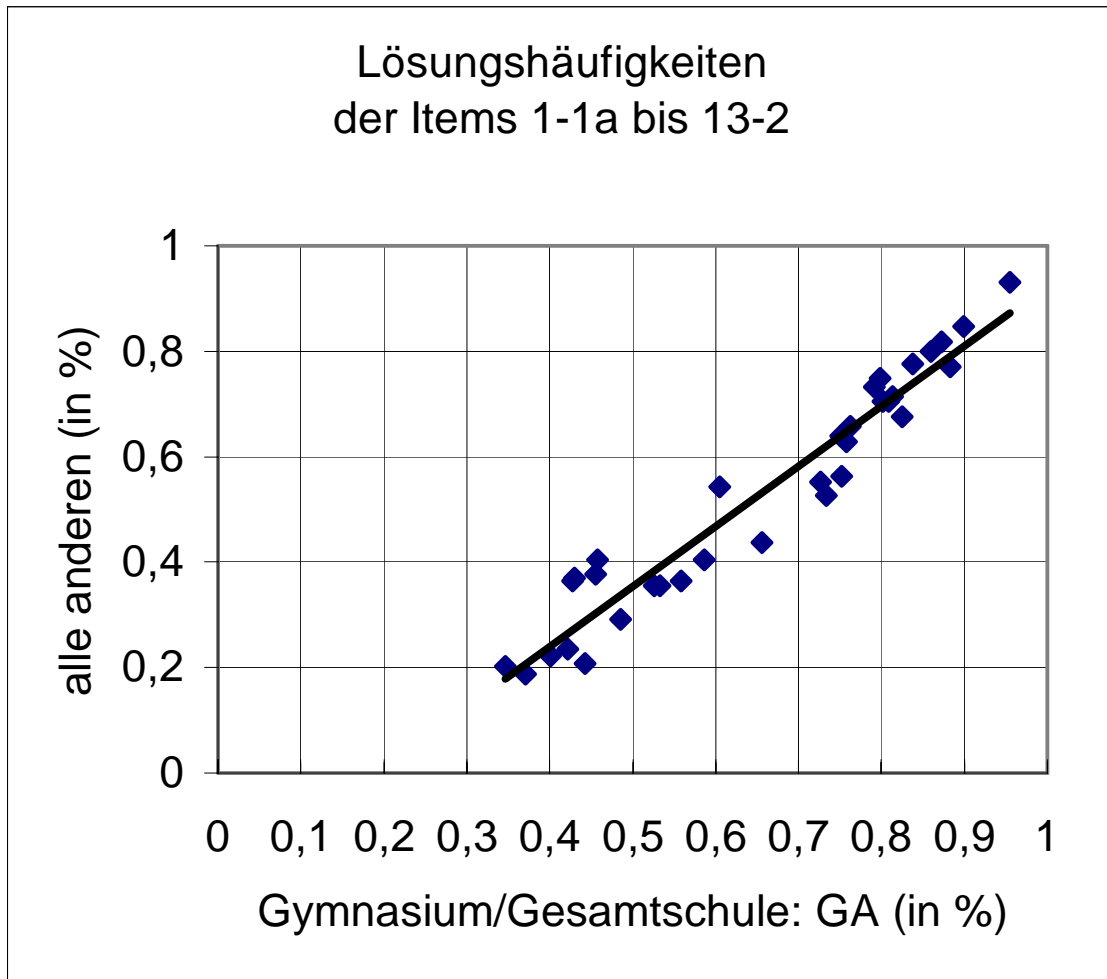
Tabelle und Abbildung Z4 setzen das Beispiel fort, indem nach Schulart differenziert wird. Dabei werden nur zwei Gruppen betrachtet, jene der Schüler/innen aus dem Gymnasium und dem Kursniveau GA aus der Gesamtschule sowie jene aus allen anderen Schularten. Das Vorgehen ist völlig analog zum eben Skizzierten.

Es gibt keine größeren Abweichungen von der Geraden. Bemerkenswert scheint die Dreiergruppe aus der Aufgabe 11 (Item 11-b: Gym=43%/sonst=36%, 11-c: 43%/37%, 11-d: 46%/38%) und Item 08-b: 66%/44%.

**Z4-a Tabelle: Mathematik.
Lösungshäufigkeiten differenziert nach Schulart (dichotomisiert).**

ITEM	OG/O: GA	alle anderen	Berlin: Insgesamt
01: Fahnenmast a: immer langsamer	81%	70%	75%
b: immer schneller	76%	66%	70%
c: gleichbleibend	90%	85%	87%
d: am schnellsten	86%	80%	83%
02: Grundstücksansicht: Perspektivwechsel	95%	93%	94%
03: Geschwindigkeit a: Fußgänger	80%	70%	75%
b: Haarwachstum	88%	77%	82%
c: Brieftaube	75%	64%	69%
04: Näherungswert	87%	82%	84%
05: Funktionsgraf a: Parabel	81%	71%	76%
b: Gerade nach oben	75%	56%	65%
c: Gerade nach unten	83%	68%	74%
d: Gerade nach unten	73%	53%	62%
06: Computerkauf a: Endpreis 1	79%	73%	76%
b: Gespartes Geld	80%	75%	77%
07: Lineare Gleichung: Grundmenge Q	53%	36%	44%
08: Potenzgesetze a: Multiplizieren	73%	55%	64%
b: Dividieren	66%	44%	54%
c: Zahlen kürzen	59%	40%	49%
d: Variable kürzen	44%	21%	31%
09: Dreieck konstruieren: Konstruierbarkeit	84%	78%	80%
10: Kongruenzsatz a: Lösung	76%	63%	69%
b: Begründung	53%	35%	43%
11: Vereinskasse a: Anzahl der Scheine	46%	40%	43%
b: Ansatzgleichung	43%	36%	39%
c: Bestimmung von a	43%	37%	40%
d: Bestimmung von b	46%	38%	41%
e: Antwortsatz formuliert	60%	54%	57%
12: Bevölkerung a: Jahresangabe	56%	36%	45%
b: Formelumgang	35%	20%	27%
c: Berechnung von p	37%	19%	27%
d: Antwortsatz formuliert	49%	29%	38%
13: Streckenlänge a: Pythagoras angewendet	42%	23%	32%
b: Richtige Lösung	40%	22%	30%

Z4-b Abbildung: Mathematik.
 Lösungshäufigkeiten differenziert nach Schulart (dichotomisiert).



Zusammenhang zwischen Zensuren und den Ergebnissen in den Vergleichsarbeiten

Zu den Daten der Rückmeldung aus den Schulen gehörten die Zensuren, die im Halbjahreszeugnis Ende Januar vergeben worden waren. Zensuren sind ebenso Quantifizierung des Leistungsniveaus wie die Gesamtpunktzahlen aus den Vergleichsarbeiten. Tabelle Z5 zeigt, wie hoch der korrelative Zusammenhang zwischen den beiden Maßen ist.

Z5 Tabelle: Zusammenhang zwischen den Zensuren aus dem Halbjahreszeugnis Januar 2004 und den Gesamtpunktzahlen aus den Vergleichsarbeiten im Frühjahr 2004.

Angegeben werden Korrelationskoeffizienten²⁷ und die zugrundeliegenden Fallzahlen.

Fach	Deutsch	Englisch	Mathematik
Gesamtschule	- .53 (N= 492)	- .48 (N= 484)	- .35 (N= 432)
Hauptschule	- .50 (N= 136)	- .57 (N= 100)	- .41 (N= 95)
Realschule	- .51 (N= 391)	- .53 (N= 339)	- .51 (N= 391)
Gymnasium	- .47 (N= 572)	- .51 (N= 754)	- .55 (N= 576)
Berufsschule (OBF)	- .45 (N= 201)	- .61 (N= 169)	- .26 (N= 176)
gesamt	- .53 (N=1792)	- .49 (N=1846)	- .39 (N=1670)

Die Korrelationskoeffizienten sind alle statistisch signifikant von Null verschieden. (Da hohe Werte bei den Zensuren niedrige Leistungen bedeuten, während für die Gesamtpunktzahlen das Gegenteil gilt, sind die Koeffizienten alle negativ.) Sie sind mittelhoch, d.h. es liegt ein substantieller Zusammenhang vor, der aber nicht sehr stark ist. Die Einschätzung des Leistungsniveaus anhand der Zensuren ist ähnlich, aber bei weitem nicht identisch mit der der Vergleichsarbeiten.

Dies ist durchaus plausibel, denn in die Zensuren fließen eine Vielzahl von Beobachtungen ein, während die Vergleichsarbeit eine Momentaufnahme ist, auch wenn sie ein breiteres

²⁷ SPEARMAN's ρ .

Korrelationskoeffizienten nehmen Werte zwischen -1 und +1 an, dürfen aber nicht als Prozentwerte missinterpretiert werden. Die Größe des Koeffizienten ist ein Gradmesser der Stärke des Zusammenhanges. Das Vorzeichen des Koeffizienten beschreibt dessen Richtung: Positiv für einen gleichsinnig proportionalen Zusammenhang (je größer das eine, desto größer das andere), negativ für einen umgekehrt proportionalen Zusammenhang.

Spektrum als eine Klassenarbeit abdeckt.²⁸ Vergleichsarbeiten haben einen anderen Zugriff auf die Kenntnisse und Fähigkeiten der Schüler/innen als Klassenarbeiten oder der Prozess der Zensurengebung; vgl. Kapitel C. Von besonderem Interesse sind daher die Unterschiede, die zwischen den achtzehn Korrelationskoeffizienten der Tabelle Z2 auftreten.

In Deutsch und in Englisch ist der Zusammenhang zwischen Zensur und Gesamtergebnis in etwa gleich stark ausgeprägt, in Mathematik deutlich schwächer; vgl. unterste Zeile. Weiter als in den anderen Fächern scheinen in Mathematik die Aufgaben der Vergleichsarbeit von jenen im Unterricht behandelten oder in Klassenarbeit vorkommenden entfernt zu sein.

Zugleich treten im Fach Mathematik die größten Unterschiede in der Höhe der Korrelationen zwischen den Schularten auf; vgl. die jeweils kleinsten und größten Werte in den drei Fächerspalten der Tabelle Z5. Die schulartspezifische Unterrichtspraxis ist anscheinend unterschiedlich weit weg von dem, was in den Aufgaben der Vergleichsarbeit zum Ausdruck kommt.

²⁸ Die mittelhohen Korrelationskoeffizienten ließen sich auch als diagnostische Schwächen der Lehrkräfte deuten. Eine derart weitreichende Interpretation überforderte jedoch das vorliegende Datenmaterial. Eine kurze Einführung in die Problematik liefert SCHRADER 2001.

Zusammenhänge zwischen den Ergebnissen in den drei Fächern

Über die Schülernummer können die drei Datensätze Deutsch, Englisch und Mathematik miteinander verknüpft werden, so dass die fachspezifischen Zensuren und Ergebnisse in den Vergleichsarbeiten miteinander verglichen werden können.

Z6 Tabelle: Zusammenhang der Leistungsmaße zwischen den Fächern.
Grundlage: Zensuren aus dem Halbjahreszeugnis Januar 2004 und den Gesamtpunktzahlen aus den Vergleichsarbeiten im Frühjahr 2004.
 Angegeben werden Korrelationskoeffizienten²⁹.

Fächerpaar	Korrelation der	
	Zensuren	Gesamtwerte
Deutsch - Englisch	.58	.66
Deutsch - Mathematik	.31	.54
Englisch- Mathematik	.30	.59

Die Korrelationen zwischen den Gesamtwerten der drei Vergleichsarbeiten liegen höher, teilweise deutlich höher, als die Korrelationen zwischen den Zensuren. Sie liegen auch höher als die fachspezifischen Korrelationen zwischen den Zensuren und den Gesamtpunktzahlen; vgl. Tabelle Z6.

Erwartungsgemäß sind die Korrelationen der Zensuren zwischen den beiden sprachlichen Fächern erheblich höher als mit der Zensur in Mathematik. Dies deckt sich mit der weit verbreiteten Annahme, dass sprachliche Fähigkeiten zu einem großen Teil unabhängig von mathematischen vorhanden sein können. Desto bemerkenswerter die Korrelationen der Ergebnisse aus den Vergleichsarbeiten, die nicht nur durchweg höher sind als jene der Zensuren, sondern alle drei in etwa derselben Größenordnung. Offensichtlich enthalten die Aufgaben aus allen drei Vergleichsarbeiten Anforderungen, die weniger fachabhängig sind als jene, die sich in den Bewertungen der Lehrkräfte als Zensuren niederschlagen.

²⁹ SPEARMAN's ρ .

Alle Koeffizienten in der Tabelle Z6 sind statistisch signifikant, also überzufällig von Null verschieden.

Quellenangaben:

LIND, Detlef & KNOCHE, Norbert: Testtheoretische Modelle und Verfahren bei PISA-2000-Mathematik.

In NEUBRAND 2004, 51-69

NEUBRAND, Michael (Hg.): mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland. Vertiefende Analyse im Rahmen von PISA 2000.

Wiesbaden: Verlag für Sozialwissenschaften 2004.

ROST, D.H. (Hg.): Handwörterbuch Pädagogische Psychologie.

Weinheim: Psychologie Verlags Union (Beltz) 2001²

SCHRADER, Friedrich-Wilhelm: Diagnostische Kompetenz von Eltern und Lehrern.

In ROST 2001, 91-96

Abkürzungen:

O: Gesamtschule

OH: Hauptschule

OR: Realschule

OG: Gymnasium

OBF: Berufsfachschule

MW: Mittelwert

Nachstehend wurden die Ergebnisblätter aus dem Frühjahr 2004 entsprechend den Einteilungen aus dem Kapitel B abgewandelt, ein Vorschlag, wie die hier für die Werte der Stichprobe vorgenommenen Auswertungen für die eigene Klasse nachvollzogen werden können.

Das Fach Englisch wird hierbei nicht berücksichtigt, da zum einen die Vorabzuordnung der einzelnen Aufgaben zu den Niveaustufen sich teilweise als empirisch nicht haltbar herausstellte, vgl. Abschnitt B2, zum anderen weil mit den einzelnen Aufgabenteilen - pro Fähigkeitsbereich Hören, Lesen, Schreiben drei Aufgaben - bereits eine modulare Einteilung vorliegt, deren Vergleichswerte im ersten Bericht samt seiner Ergänzung ausgewiesen wurden und die so fein untergliedert ist, dass nach eigenen Bedürfnissen Zusammenstellungen erfolgen können.

VA 2004/Deutsch: Daten nach Verstehensdimensionen. Klasse: (S-1, S-2 etc.: Schreibaufgabe, Kriterium 1, 2 etc.)

	Verstehensdimension 2													Σ-2	gesamt
	4	6	7	9	10	13	14	15	16	S-1	S-6	S-8	S-9		
	(≤2)	(≤2)	(≤2)	(≤2)	(≤2)	(≤1)	(≤2)	(≤4)	(≤2)	(≤2)	(≤1)	(≤1)	(≤1)		
01															
02															
03															
04															
05															
06															
07															
08															
09															
10															
11															
12															
13															
14															
15															
16															
17															
18															
19															
20															
21															
22															
23															
24															
25															
26															
27															
28															
29															
30															



