VERA
VERgleichsArbeiten
in 4. Grundschulklassen

Ergebnisbericht VERA 2005:

Bremen

Prof. Dr. A. Helmke, Jun.Prof. Dr. I. Hosenfeld
J. Groß Ophoff, A. C. Halt, J. Hochweber, K. Isaac, U. Koch und F. Scherthan
Universität Koblenz - Landau, Campus Landau

Stand: 24.05.2006

Inhalt

1 Einleitung und Überblick

Am 27. und 29. September 2005 schrieben im Rahmen des Projektes VERA mehr als 300 000 Schülerinnen und Schüler der vierten Klassenstufe in sieben deutschen Bundesländern Vergleichsarbeiten in den Fächern Mathematik und Deutsch. Im Folgenden werden Hintergrund und Anlage des Projekts sowie einige empirische Ergebnisse deskriptiv dargestellt. Vertiefende Analysen müssen späteren Veröffentlichungen vorbehalten bleiben.

1.1 Ziele

Anders als beispielsweise TIMSS, PISA oder IGLU ist VERA kein bloßes „System Monitoring“ (dazu würde eine Stichprobe genügen), sondern umfasst neben der Bestandsaufnahme ausdrücklich auch die Schul- und Unterrichtsentwicklung:

So soll die aktive Einbeziehung der Lehrkräfte in den gesamten Prozess ein Anstoß für fachdidaktische Diskussion und Kooperation zwischen den Lehrkräften sein. Damit wird einem immer wieder geäußerten Ruf nach mehr schulinterner Kooperation und Teamarbeit Rechnung getragen. Die Rückmeldung des Leistungsstandes sowie von Informationen zur diagnostischen Kompetenz der Lehrkräfte und zu Fehlermustern der Schülerinnen und Schüler sollen pädagogische Impulse geben und damit die schulinterne Diskussion von (Bildungs-)Standards, der Schul- und Unterrichtsentwicklung oder der Beurteilungspraxis anregen. Darüber hinaus können die Informationen über die Fähigkeitsniveaus in den Fächern Deutsch und Mathematik als ergänzende Information zur Beratung der Eltern herangezogen werden.

Da die Aufgabenauswahl, die Auswertung und die Ergebnisrückmeldung über das Internet erfolgen, trägt das Projekt VERA zu einem Aufschwung in der effizienten Nutzung des Internet für die schulische Qualitätssicherung bei und leistet somit auch einen Beitrag zur Förderung der Medienkompetenz.

Die Vergleichsarbeiten verfolgen also mehrere Ziele:

- Unterrichtsentwicklung: Nutzung pädagogischer und fachdidaktischer Impulse
- Schulentwicklung: Intensivierung schulinterner Kooperation und Teamarbeit
- Professionalisierung der Lehrkräfte im Hinblick auf diagnostische Kompetenzen
- Standardsicherung
- Ergänzende Information zur Beratung der Eltern
- Erleichterung und Beschleunigung der Umsetzung moderner Kerncurricula, Lehr- und Rahmenpläne
- Effizienzsteigerung bei der Nutzung des Internet für die schulische Qualitätssicherung

1.2 Organisation

Durchgeführt wird das Projekt VERA von der Projektgruppe Empirische Bildungsforschung an der Universität in Landau (Leitung: Prof. Dr. Andreas Helmke und Jun.Prof. Dr. Ingmar Hosenfeld) in enger Zusammenarbeit mit den beteiligten Ministerien und Landesinstituten, Projekten und Institutionen der empirischen Bildungsforschung.

Im Herbst 2005 nahmen alle Grundschulen in den Bundesländern Berlin, Brandenburg, Bremen, Mecklenburg-Vorpommern, Nordrhein-Westfalen, Rheinland-Pfalz und Schleswig-Holstein verpflichtend an der Untersuchung teil. Darüber hinaus nahmen etliche deutsche Auslandsschulen mit vierten Klassen aus allen fünf Erdteilen das Angebot zur Teilnahme an VERA an. Abbildung 1 verdeutlicht die Vernetzung verschiedener Funktionen und Gruppen.

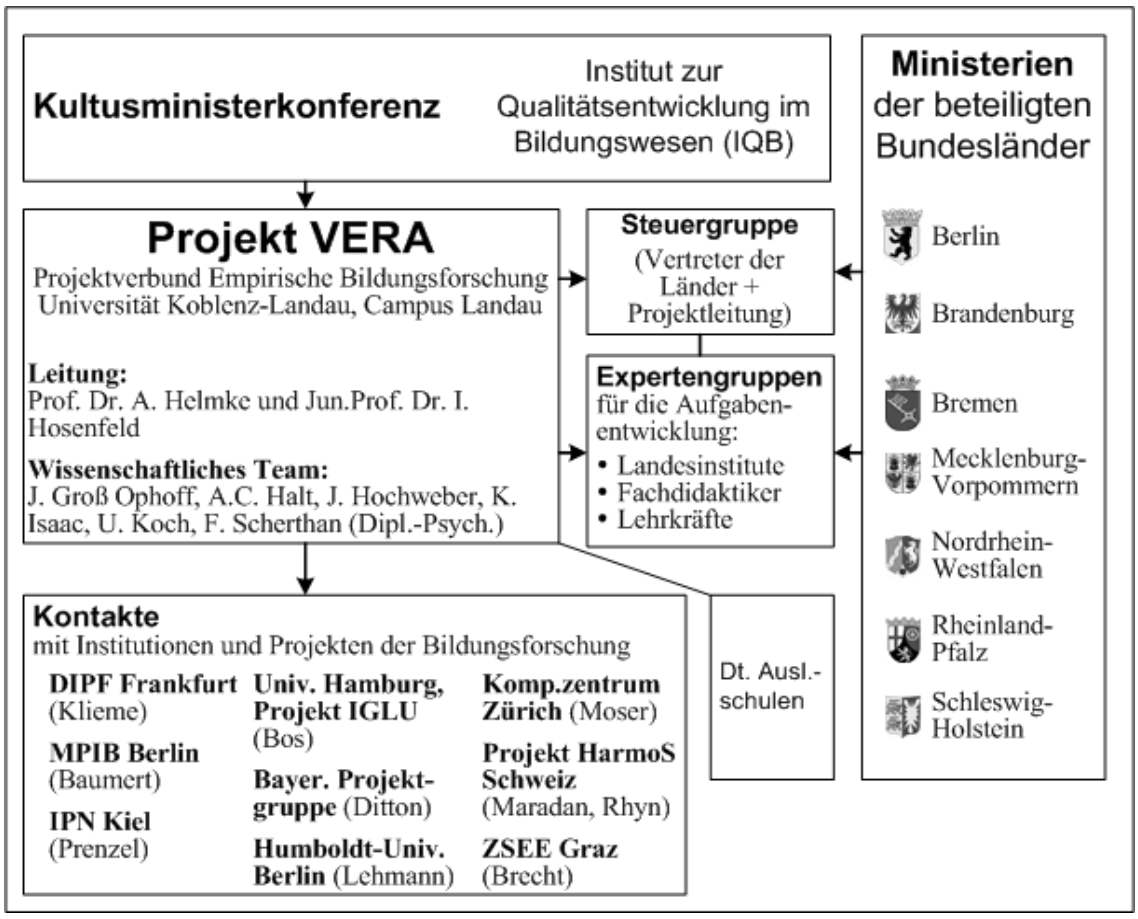

Abbildung 1: *Bildungspolitischer Kontext von VERA*

Im Vergleich zum Jahr 2004 wurden die Durchführungsmodalitäten für die Vergleichsarbeiten verändert, insbesondere um den Aufwand bei der Durchführung und Auswertung der Vergleichsarbeiten zu reduzieren. So sollen z.B. die Fächer Mathematik und Deutsch im jährlichen Wechsel versetzt in unterschiedlichem Umfang erhoben werden. Für das Jahr 2005 bedeutete dies konkret, dass das Fach Deutsch nur im Bereich „Leseverständnis“ getestet wurde, wohingegen das Fach Mathematik (Arithmetik, Geometrie, Sachrechnen) in vollem Umfang erhoben wurde. 2006 sollen im Fach Deutsch mehrere

Inhaltsbereiche, dagegen Mathematik nur anteilig geprüft werden. Darüber hinaus wurde in den Vergleichsarbeiten im Jahr 2005 auf die Auswahl von Testaufgaben sowie die aufwändige Fehleranalyse im Fach Deutsch verzichtet.

2 Aufgabenentwicklung und Definition der Fähigkeitsniveaus

Die im Rahmen von VERA verwendeten Aufgaben wurden von zwei Expertengruppen (je eine für Mathematik und Deutsch) entwickelt und im Vorfeld der Vergleichsarbeiten im Jahr 2005 umfangreich empirisch untersucht. In einem ersten Schritt wurden ausgewählte Aufgaben durch die Referenzschulen beurteilt und zwar in Bezug auf Anwendbarkeit, typische Fehler sowie schulische Akzeptanz. Basierend auf den Rückmeldungen wurden erste Aufgabenentwürfe verbessert bzw. weitere Aufgaben entwickelt. Die eigentlichen empirischen Pilotierungs- und Normierungsstudien erfolgten im Mai 2005 (ca. 10 000 Schülerinnen und Schüler) und dienten der Realisierung folgender Ziele:

- Verbesserung des entwickelten Aufgabenmaterials und der Korrekturanweisungen (v.a. Pilotierung)
- Erprobung der von Didaktikern und Lehrkräften entwickelten Aufgaben (insbesondere in Hinsicht auf die Testsituation sowie Anregung zu Schul- und Unterrichtsentwicklung)
- Bestimmung der Aufgabenschwierigkeiten und Personenparameter als Basis für die Zuordnung zu den Fähigkeitsniveaus im Rahmen der Vergleichsarbeiten 2005
- Weiterentwicklung der Fähigkeitsniveau-Definitionen, darauf aufbauend Entwicklung der didaktischen Erläuterungen der Materialien
- Untersuchung der Art und Häufigkeit von Fehlern (Falschlösungen) der Aufgaben
- Gewinnung eines Referenzrahmens für die obligatorische Aufgabenauswahl durch die Kollegien
- Erhebung und Analyse von Kontextmerkmalen, die einen „fairen Vergleich“ ermöglichen.

Dazu wurden Zufallsstichproben von Schulen aus allen Bundesländern gezogen und jeweils mit einem Drittel dritten und zwei Dritteln vierten Klassen in die Untersuchung einbezogen. Innerhalb jeder Klasse kamen sowohl in der Pilotierung als auch in der Normierung je Fach zehn verschiedene Testhefte zum Einsatz. Jedes der Hefte enthielt eine Reihe von Aufgaben, die nur in dem jeweiligen Heft präsentiert wurden und weitere Aufgaben, die in mindestens einem weiteren Heft enthalten waren (Multi-Matrix-Sampling). So konnten in beiden Fächern jeweils mehr als 200 Aufgaben untersucht werden, während jedem einzelnen Kind nur eine überschaubare Anzahl von Aufgaben vorgelegt wurde. Zur Bearbeitung der Aufgaben standen in Deutsch in der Pilotierung 90 Minuten (mit zehn Minuten Pause) und in der Normierung 60 Minuten (mit fünf Minuten Pause) zur Verfügung. Im Fach Mathematik wurden 50 Minuten Bearbeitungszeit festgesetzt. Darüber hinaus wurden wichtige Hintergrundmerkmale der Schülerinnen und Schüler (Alter, Geschlecht, zu Hause gesprochene Sprache etc.) im Rahmen eines kurzen Schülerfragebogens erhoben.

Im Folgenden wird der Prozess der Aufgabenentwicklung, getrennt für Mathematik und Deutsch, dargestellt.

2.1 Deutsch

2.1.1 Aufgabenentwicklung

Die Deutschaufgaben wurden von einer länderübergreifenden Expertengruppe entwickelt. Die Gruppe setzt sich zusammen aus erfahrenen Lehrkräften, Mitarbeitern der Landesinstitute, Herrn Prof. Dr. Albert Bremerich-Vos als fachdidaktischem Berater und Mitarbeitern des Landauer Projektteams als psychologische und psychometrische Experten. Die Entwicklung begann im Frühjahr 2003 und orientierte sich während des gesamten Entwicklungsprozesses auch an den im Oktober 2004 verabschiedeten Bildungsstandards für den Primarbereich.

Im Rahmen der Aufgabenentwicklung wurde ein besonderes Augenmerk darauf gelegt, die Entwicklung einer neuen Unterrichts- und Aufgabenkultur in den Schulen zu unterstützen. Das bedeutet konkret für die Aufgabenentwicklung, dass die generierten Items und die damit verbundenen Anforderungen nicht nur Testgütekriterien genügen müssen, sondern z.B. durch einen thematischen Rahmen und integrative Formate zu Entwicklungsprozessen anregen sollen.

Wie bereits erwähnt wurde in den Vergleichsarbeiten 2005 im Fach Deutsch nur der Bereich „Leseverständnis“ erfasst und dementsprechend in Voraberhebungen normiert. Aus dem Anspruch des integrativen Deutschunterrichts heraus wurde für VERA 2005 der thematische Rahmen „Reisen“ bestimmt. Dabei wurde in curricularer Hinsicht darauf geachtet, dass die den Testaufgaben zugrunde liegenden Lesetexte das Spektrum von fiktionalen sowie von kontinuierlichen und diskontinuierlichen Sachtexten abdeckten.

Darüber hinaus wurden alle für die Vergleichsarbeiten ausgewählten Texte und Aufgaben auf schwierigkeitsbestimmende Merkmale hin überprüft. Kriterien hierbei waren in Bezug auf den Text u.a.

- Textlänge und –komplexität
- Gestaltung (Schriftauswahl, Zeilenabstand, Gliederung, Zeilennummern)
- Unterstützende Elemente (Bilder, Grafiken)
- Syntaktische Strukturen
- Vertrautheit des Wortschatzes (Text, Aufgabe)
- Vertrautheit und Interesse am Thema.

In Bezug auf die Aufgaben wurden des weiteren die folgenden Merkmale berücksichtigt:

- Streubreite der Aufgabenschwierigkeiten und Aufgabenformate (Multiple-Choice, offene Antworten unterschiedlicher Länge, Markierungen, Zuordnungen, Sortierungen)

- Vertrautheit mit Aufgabenformat
- Anordnung und Reihenfolge der Aufgaben
- Formulierung des Aufgabentextes (Eindeutigkeit).

Zunächst wurde eine umfangreiche Sammlung von Aufgaben erzeugt, die dann in sukzessiven Auswahlrunden reduziert und optimiert wurde. Dabei spielten neben klassischen Kriterien der Testentwicklung (Objektivität, Reliabilität, Validität) auch Überlegungen zur Durchführbarkeit, der erforderlichen Bearbeitungszeit und der Auswertungsökonomie sowie zur Relevanz für Schul- und Unterrichtsentwicklung eine Rolle. Insbesondere mit Blick auf den letztgenannten Gesichtspunkt, also die Anregung zu schulischen Entwicklungsprozessen, wurden die Hintergründe zu den VERA 2005-Aufgaben sowie Möglichkeiten des Umgangs mit den zurückgemeldeten Ergebnissen ausführlich in Form von „Didaktischen Erläuterungen“ (VERA-Projektgruppe Deutsch, 2005, siehe http://www.uni-landau.de/vera/aufgaben.htm) dargestellt.

2.1.2 Skalierung

Die Verwendung unterschiedlicher, über gemeinsame Aufgaben miteinander verbundener Testhefte erfordert den Einsatz probabilistischer Testmodelle, um Aufgaben sowie Schülerinnen und Schüler über die Testhefte hinweg auf einer gemeinsamen Dimension abzubilden. Hintergrund und Ziel dieser Methode können hier nicht ausführlich dargestellt werden (s. dazu Helmke & Hosenfeld, 2004). Nach der Erfassung, Aufbereitung und Qualitätskontrolle wurde 2004 für die vier Inhaltsbereiche jeweils ein einparametrisches Modell (sog. Rasch-Modell; Rasch, 1960) angepasst. Als Ergebnis erhält man für jeden Inhaltsbereich eine Skala, auf der gleichzeitig die Aufgabenschwierigkeiten und die Personenfähigkeiten abgetragen werden können. Dabei gilt entsprechend des probabilistischen Charakters des Modells, dass sich aus der Differenz zwischen Personen(fähigkeit) und Aufgaben(schwierigkeit) angeben lässt, mit welcher Wahrscheinlichkeit eine Aufgabe gelöst wird. Hierbei wurde festgelegt (wie bei PISA 2000), dass eine 62,25-prozentige Lösungswahrscheinlichkeit resultiert, wenn Personenfähigkeit und Aufgabenschwierigkeit numerisch identisch sind, d.h. es ist gewährleistet, dass Personen entsprechende Aufgaben mit 'hinreichender' Sicherheit lösen können.

Dass Personenfähigkeiten und Aufgabenschwierigkeiten auf einer gemeinsamen Skala angeordnet werden, bietet einen großen Vorteil für die Interpretation der Ergebnisse: Die Fähigkeitsdimension lässt sich in Zonen (Fähigkeitsniveaus) einteilen, die sich anhand der ihnen zugeordneten Aufgaben kriterial beschreiben lassen. So kann verdeutlicht werden, welche Anforderungen von den zugeordneten Personen typischerweise bewältigt werden. Natürlich wurden auch die im Jahr 2005 normierten Aufgaben auf der gleichen Fähigkeitsdimension verankert, die im Rahmen der Normierungsstudien von VERA 2004 ermittelt wurde. Das wiederum ermöglicht es, die Fähigkeitsniveau-Beschreibungen fortschreitend zu präzisieren.

2.1.3 Fähigkeitsniveaus

Was bedeuten die Fähigkeitsniveaus*?

Ebenso wie in den großen internationalen Vergleichsstudien TIMSS, DESI, PISA oder PIRLS und IGLU ist es bei Lernstandserhebungen und anderen vergleichenden Leistungsmessungen im schulischen Kontext „state of the art", inhaltlich definierte Kriterien zur Beschreibung und zum Vergleich des Leistungsstandes (Kompetenzen) heranzuziehen. In der Expertise zur „Entwicklung nationaler Bildungsstandards" (Klieme et al., 2003) werden Kompetenzen als situationsangemessene bzw. domänenspezifische Problemlösefähigkeiten beschrieben.

Üblicherweise wird die Ausprägung der Kompetenz in Anlehnung an das Anforderungsniveau der damit korrespondierenden Aufgaben in sog. Kompetenzstufen (PISA, z.B. Baumert et al., 2002), Kompetenzniveaus (Lernstandserhebungen Klasse 9, NRW[1]) oder wie bei DESI und VERA in Fähigkeitsniveaus unterteilt. Das Kontinuum lässt sich von sehr gering ausgeprägten bis hin zu sehr entwickelten Fähigkeiten unterteilen. Dabei liegt die Vorstellung zugrunde, dass verschiedene Schwierigkeitsniveaus in einem hierarchischen Verhältnis zueinander stehen. Die Aussage „Schülerinnen und Schüler befinden sich in einem Inhaltsbereich X auf Niveau 2" heißt soviel wie:

- Aufgaben mittleren Anforderungsniveaus (Fähigkeitsniveau 2) werden in diesem Inhaltsbereich mit hinreichender Sicherheit (Lösungswahrscheinlichkeit von 62,25 Prozent) gelöst.
- Aufgaben des Fähigkeitsniveaus 1, also einfachere Aufgaben mit grundlegenderen Anforderungen, werden mit höherer Sicherheit gelöst.
- Anspruchsvollere Aufgaben, die sich auf dem Fähigkeitsniveau 3 befinden, werden dagegen mit geringerer Wahrscheinlichkeit gelöst.

Im Projekt VERA wurden die Fähigkeitsniveaus auf der Basis inhaltlicher und theoretischer Vorgaben in Kooperation mit Experten (Fachdidaktiker, Fachwissenschaftler, praktisch tätige Lehrkräfte und Psychologen) und ausführlicher empirischer Normierungsstudien im Rahmen der Vergleichsarbeiten 2004 entwickelt und basierend auf den Normierungsergebnissen im Jahr 2005 weiterentwickelt. Die Beschreibungen für den jeweiligen Inhaltsbereich in Deutsch und Mathematik wurden in Hinsicht auf die mit hinreichender Sicherheit zu bewältigenden Anforderungen konkretisiert.

Der Entwicklungsprozess führt also empirisch erfasste Informationen und theoretische Überlegungen zusammen:

- Auf der einen Seite benötigt man die empirische Schwierigkeitsverteilung der Aufgaben, d.h. Informationen darüber, von wie vielen Schülerinnen und Schülern

---

* Für jeden der Inhaltsbereiche der Fächer Deutsch und Mathematik wurden jeweils drei Fähigkeitsniveaus festgelegt.

[1] http://www.learn-line.nrw.de/angebote/lernstand9; Zugriff am 29. März 2005

die Aufgaben richtig gelöst wurden.

- Auf der anderen Seite sind Annahmen über Denkprozesse erforderlich, die der Aufgabenbearbeitung zugrunde liegen, sowie Informationen darüber, was Aufgaben eigentlich unterschiedlich leicht oder schwer macht („schwierigkeitsbestimmende Merkmale"). Dies erfordert fachdidaktische und psychologische Expertise: Aufgrund welcher Merkmale, welcher Anforderungen an Denk- und Gedächtnisprozesse lassen sich Aufgaben charakterisieren?

Die Entwicklung der VERA-Fähigkeitsniveaus erfolgte in technischer Hinsicht in mehreren Schritten: Zunächst wurden die Aufgaben jedes Inhaltsgebietes hinsichtlich des geschätzten Schwierigkeitsparameters aufsteigend sortiert und vorläufig zu Gruppen zusammengefasst. Dann wurden die Aufgabengruppen von Experten in Hinsicht auf Aufgabeninhalte und Anforderungsbereiche (elementare, erweiterte und fortgeschrittene Fähigkeiten) analysiert. Dabei stand die Differenzierung gegenüber dem nächst niedrigeren Niveau im Fokus der Beschreibung. Es resultierte eine erste holistische Beschreibung der typischen Anforderungen jedes Niveaus. Diese Beschreibung wurde im zweiten Schritt verfeinert, indem zusätzlich Prozessüberlegungen (z.B. notwendige Problemlöseprozeduren) und schwierigkeitsbestimmende Merkmale der Aufgaben (z.B. Eindeutigkeit der Lösung, sprachliche Komplexität, Antwortformat) einbezogen wurden. Als Abschluss der Pilotierungs- und Normierungsstudien im Jahr 2005 wurden die aus den Vergleichsarbeiten 2004 hervorgegangenen Fähigkeitsniveau-Beschreibungen noch einmal mit den neu entwickelten Aufgaben abgeglichen und ggf. präzisiert.

In diesem Zusammenhang wird deutlich, dass die Beschreibungen von Fähigkeitsniveaus einen Beitrag zur Ausbildung von bereichsspezifischen Kompetenzmodellen darstellen, jedoch zur Zeit noch nicht von einer abgeschlossenen Theoriebildung gesprochen werden kann, wie es die Verwendung des „Kompetenz"- Begriffs nahe legt. Die bei VERA erhobenen Leistungen sind als fachlich definierte Fähigkeiten beschrieben, die sich auf das Beherrschen der gesamten Breite des jeweiligen Inhaltsgebietes (inkl. der notwendigen Vorkenntnisse) beziehen, also nicht nur – wie bei Klassenarbeiten – auf die Meisterung des vor kurzem durchgenommenen Unterrichtsstoffs.

Beschreibung der Fähigkeitsniveaus im Fach Deutsch

Für das Fach Deutsch wurden bei VERA Fähigkeitsniveaus in den vier Inhaltsbereichen Leseverständnis, Schreiben, Sprachbetrachtung und Orthografie entwickelt. Das aufgefundene Fähigkeitsspektrum wurde, wie auch in Mathematik, in drei Bereiche unterteilt. Für jede Schülerin bzw. jeden Schüler wird gemäß der erfassten Testleistung jeweils ein Fähigkeitsniveau in den vier Inhaltsbereichen (Leseverständnis, Schreiben, Sprachbetrachtung und Orthografie) ermittelt. Die Zuordnung besagt, dass die für dieses Niveau formulierten Anforderungen mit hinreichender Sicherheit bewältigt werden.

Die Fähigkeitsniveaus können bereichsübergreifend wie folgt beschrieben werden:

- **Fähigkeitsniveau 1: Elementare bzw. grundlegende Fähigkeiten**
  Einfache Aufgaben mit grundlegenden Anforderungen werden hinreichend sicher gelöst.

- **Fähigkeitsniveau 2: Erweiterte Fähigkeiten**
  Aufgaben mittleren Anforderungsniveaus werden hinreichend sicher gelöst.
- **Fähigkeitsniveau 3: Fortgeschrittene Fähigkeiten**
  Es werden auch anspruchsvollere Aufgaben hinreichend sicher gelöst.

Liegen keine oder extrem unvollständige Daten vor, ist eine Zuordnung zu den beschriebenen Fähigkeitsniveaus nicht möglich. Dies wird bei VERA als „nicht auswertbare Leistung (n.a.L.)“ bezeichnet.

Da in den Vergleichsarbeiten 2005 im Fach Deutsch der Inhaltsbereich Leseverständnis fokussiert wurde, werden im Folgenden dafür die konkreten Anforderungen der Fähigkeitsniveaus dargestellt.

Leseverständnis

Fähigkeitsniveau 1: Elementare Fähigkeiten (Gewinnung von Einzelinformationen)

- Im Rahmen von Auswahlaufgaben werden einzelne Informationen im Lesetext gefunden, wenn sie in der Aufgabenstellung fast wortgleich oder in etwa bedeutungsgleich formuliert sind.
- Es kann angegeben werden, wie einzelne Stellen im Text unabhängig vom Kontext verstanden werden sollten.
- In Auswahlaufgaben können auf der Basis von elementarem Wissen einfache Schlüsse gezogen werden.
- Auf der Basis von elementarem Wissen und/oder Bildinformationen können einfache Schlüsse gezogen werden.

Fähigkeitsniveau 2: Erweiterte Fähigkeiten (Verknüpfung von mehreren Informationen)

- Über den Text verteilte Informationen können gefunden und miteinander verknüpft werden. Die zentrale Aussage eines Textes kann wiedergegeben werden.
- Im Rahmen von Auswahlaufgaben kann angegeben werden, wie einzelne Stellen im Text verstanden werden sollten – selbst dann, wenn dazu weitere Informationen im Text herangezogen werden müssen.
- Texte können ansatzweise plausibel beurteilt werden: Dabei wird vor allem auf inhaltliche Aspekte und persönliche Vorlieben Bezug genommen.
- Nahe liegende Schlüsse können unter Nutzung von einzelnen oder mehreren Informationen im Text und weit verbreitetem Wissen gezogen werden und/oder mit Textstellen belegt werden.

Fähigkeitsniveau 3: Fortgeschrittene Fähigkeiten (Komplexere Schlussfolgerungen)

- Über den Text verteilte Informationen können gefunden und miteinander verknüpft werden. Das gelingt auch dann, wenn diese Informationen in der Aufgabenstellung weder wortgleich noch sinngemäß vorkommen.

- Komplexere Schlussfolgerungen können gezogen und dargestellt werden, auch wenn ein Kurztext zu schreiben ist. Dazu wird spezielleres Sach- und manchmal auch Sprachwissen benötigt. Texte können angemessen beurteilt werden. Dabei wird vor allem sowohl auf inhaltliche Aspekte und persönliche Vorlieben als auch auf formale Aspekte des Textes Bezug genommen.

2.2 Mathematik

2.2.1 Aufgabenentwicklung

Die Mathematikaufgaben waren ursprünglich von einer Gruppe erfahrener Lehrkräfte, maßgeblich beraten durch den Fachdidaktiker Prof. Dr. Jens-Holger Lorenz, auf der Basis des Grundschul-Rahmenplans des Landes Rheinland-Pfalz entwickelt worden. Dazu wurde ein zweidimensionales Raster zur Klassifikation der Aufgaben entworfen. Die beiden Dimensionen orientierten sich zum einen an den drei klassischen mathematischen *Inhaltsbereichen* in der Grundschule und zum anderen an verschiedenen *Tätigkeitsanforderungen*.

Für die Vergleichsarbeiten im Herbst 2005 erfolgte erstmalig eine explizite Orientierung der Aufgabenkonstruktion an den im Dezember 2004 von der KMK verabschiedeten Bildungsstandards[2]. Alle Aufgaben wurden dafür entlang der fünf definierten **inhaltsbezogenen** mathematischen Kompetenzen (Zahlen und Operationen, Raum und Form, Muster und Strukturen, Größen und Messen, Daten, Häufigkeit und Wahrscheinlichkeit) und den definierten **allgemeinen** mathematischen Kompetenzen (Problemlösen, Kommunizieren, Argumentieren, Modellieren, Darstellen) entwickelt. Zusätzlich wurden alle Aufgaben wie im Jahr 2004 den „traditionellen“ Sachgebieten Arithmetik, Geometrie, Sachrechnen/Größen zugeordnet. Die genaue inhaltliche Klassifikation der Aufgaben gewährleistet eine umfassende Repräsentation der Bildungsstandards sowie aller Bereiche des Curriculums und der allgemeinen Anforderungen des Mathematikunterrichts.

Zunächst wurde eine umfangreiche Sammlung von Aufgaben (> 500) erzeugt, die dann in sukzessiven Auswahlrunden reduziert und optimiert wurde. Dabei spielten neben klassischen Kriterien der Testentwicklung (Objektivität, Reliabilität, Validität) auch Überlegungen zur Durchführbarkeit, der erforderlichen Bearbeitungszeit und der Auswertungsökonomie eine Rolle. So wurden z.B. alle Aufgaben, die besondere technische Hilfsmittel wie beispielsweise einen Zirkel erfordern, ausgeschlossen, da nicht zu gewährleisten war, dass jede Schülerin und jeder Schüler in der Vergleichsarbeit über die entsprechenden Hilfsmittel würde verfügen können.

[2] vgl. http://www.kmk.org/schul/Bildungsstandards/Grundschule_ Mathematik_BS_307KMK.pdf

2.2.2 Skalierung und Fähigkeitsniveaus des Mathematiktests

Die Skalierung in Mathematik erfolgte analog zur oben geschilderten Skalierung in Deutsch, d.h. es wurde für jeden der drei Inhaltsbereiche ein ein-parametrisches Modell (Rasch-Modell) angepasst. Wie auch in Deutsch wird von einer hierarchischen Beziehung der Fähigkeitsniveaus ausgegangen. Die Aussage „Schülerinnen und Schüler befinden sich im Bereich Arithmetik auf Fähigkeitsniveau 2“ heißt soviel wie:

- Aufgaben mittleren Anforderungsniveaus (Fähigkeitsniveau 2) werden in diesem Inhaltsbereich mit hinreichender Sicherheit (Lösungswahrscheinlichkeit von 62,25 Prozent) gelöst.
- Aufgaben des Fähigkeitsniveaus 1, also Aufgaben mit grundlegenderen Anforderungen, werden mit höherer Sicherheit gelöst.
- Anspruchsvollere Aufgaben, die sich auf dem Fähigkeitsniveau 3 befinden, werden dagegen mit geringerer Wahrscheinlichkeit gelöst.

Das Vorgehen bei der Definition der Fähigkeitsniveaus umfasste – genau wie bei Deutsch – im Wesentlichen zwei Schritte: Im ersten Schritt wurden die Aufgaben jedes Inhaltsgebietes hinsichtlich des geschätzten Schwierigkeitsparameters aufsteigend sortiert und vorläufig zu Gruppen zusammengefasst. Dann wurden die Aufgabengruppen über eine Analyse der Aufgabeninhalte und Anforderungen charakterisiert, wobei jeweils die Differenzierung gegenüber der nächst niedrigeren Stufe im Fokus der Beschreibung stand. Es resultierte eine holistische Beschreibung der typischen Anforderungen jeder Stufe.

Im Folgenden werden die Fähigkeitsniveaus getrennt für die einzelnen Bereiche Arithmetik, Geometrie und Sachrechnen/Größen dargelegt.

Arithmetik

Fähigkeitsniveau 1: Elementare Fähigkeiten (Grundlegende Kenntnisse arithmetischer Verfahren)

- Schriftliche Additionsaufgaben werden gelöst.
- Subtraktionsaufgaben ohne Übertrag werden gelöst und Ergänzungen zum nächsten Hunderter/Tausender gelingen.
- Einfache kombinatorische Aufgaben können gelöst werden (z.B. das Zusammensetzen von Zahlen aus drei Ziffern).
- Die Beziehung zwischen Zahlen und ihrer Darstellung in der Stellentafel kann hergestellt werden.
- Das Zahlbildungsprinzip wird beherrscht, innerhalb einfacher Aufgaben auch bei Zahlenfolgen.
- Die Aufgabenlösung gelingt bei sprachlich einfachen und kurzen Texten.
- In Gleichungen kann eine fehlende Rechenoperation eingesetzt werden.

Fähigkeitsniveau 2: Erweiterte Fähigkeiten (Umfassende Kenntnis der Addition und Subtraktion)

- Schriftliche Addition gelingt auch mit Überträgen in unüblichen Formaten (z.B. Lückenaufgaben).
- Schriftliche Subtraktion gelingt auch mit Überträgen oder in unüblichen Formaten (z.B. Lückenaufgaben).
- Kombinatorische Aufgaben, die zusätzlich einfache Rechnungen erfordern, können gelöst werden.
- Einsicht in das Stellenwertsystem liegt vor und der Stellenwert kann in unterschiedlichen Darstellungsformen erkannt werden.
- Das Vervollständigen von Zahlenreihen gelingt, wenn die zugrunde liegende Regel vorwiegend Strichrechnung erfordert.
- Aufgaben mit sprachlich komplexeren und längeren Texten werden gemeistert.
- Arithmetische Aufgaben mit ausgeprägten sprachlichen oder bildlichen Anteilen können gelöst werden.

Fähigkeitsniveau 3: Fortgeschrittene Fähigkeiten (Flexible Beherrschung der Grundrechenarten)

- Zahlen und Operationen können flexibel kombiniert werden. Hierbei werden mathematische Kenntnisse (z.B. Rechengesetze, Teilbarkeitsregeln) korrekt angewendet.
- Schriftliche Subtraktion gelingt auch mit Überträgen in unüblichen Formaten (z.B. Lückenaufgaben).
- Eine kombinatorische Problemstellung kann vollständig modelliert werden.
- Einsicht in das Stellenwertsystem liegt vor und Veränderungen können, z.B. durch geeignete arithmetische Operationen, vorgenommen werden.
- Das Erkennen und Benennen einer Zahlenfolge gelingt.
- Überschlagsrechnungen und Rundungen zum Tausender können vorgenommen werden.
- Mehrschrittige Rechnungen werden unter Berücksichtigung der Regel „Punktrechnung vor Strichrechnung“ bewältigt.
- Das Finden, Erklären und Korrigieren von Fehlern in schriftlichen Additionen oder Subtraktionen gelingt.
- Zahlen können durch geeignete Operationen zu einer Zielzahl kombiniert werden.

Geometrie

Fähigkeitsniveau 1: Elementare Fähigkeiten (Kenntnisse grundlegender geometrischer Formen und Abbildungen)

- Aufgaben zu Umfang, Flächeninhalt und Volumen, die sich durch einfache Operationen wie z.B. Abzählen lösen lassen, werden bewältigt.
- Ebene Figuren werden in Körpern wieder erkannt.
- Geometrische Körper werden in Alltagsgegenständen wieder erkannt.
- Ebene geometrische Formen können durch Verschiebung/Drehung zu anderen Formen kombiniert werden.
- Einfache geometrische Muster werden erkannt und können fortgeführt werden.
- Die Aufgabenlösung gelingt bei sprachlich einfachen und kurzen Texten.
- Das Ergänzen zu achsensymmetrischen Figuren gelingt.

Fähigkeitsniveau 2: Erweiterte Fähigkeiten (Erkennen und Zuordnen von ebenen und räumlichen Figuren)

- Flächeninhalte können verglichen werden, wenn verschiedene Antwortalternativen vorgegeben sind.
- Aufgaben zu Körpernetzen können gelöst werden.
- Geometrische Begriffe sind bekannt und können angewendet werden.
- Rechtecke können nach vorgegebenen Maßen konstruiert werden.
- Komplexere geometrische Muster können fortgesetzt werden.
- „Fehler“ in Mustern werden erkannt.
- Strecken können in ihrer Ausdehnung mit einer vorgegebenen Maßeinheit erfasst werden, auch wenn dabei halbe Maßeinheiten berücksichtigt werden müssen.
- Baupläne und Würfelbauten können einander zugeordnet werden.
- Über eine Raumvorstellung kann ein Perspektivenwechsel vollzogen werden, wenn die jeweiligen Ansichten keine Überdeckungen der einzelnen Körper aufweisen.
- Das Navigieren in Feldern bei vorgegebenen Koordinaten gelingt.

Fähigkeitsniveau 3: Fortgeschrittene Fähigkeiten (Konstruktionen in der Ebene und im Raum; Erkennen von Beziehungen zwischen geometrischen Begriffen)

- Flächeninhalte können in offenen Aufgabenstellungen bestimmt werden.
- Würfelnetze, Muster und räumliche Darstellungen einfacher Körper können vervollständigt werden.
- Körpereigenschaften können sachgerecht bezeichnet werden.
- Das Zerlegen einer Fläche in vorgegebene Figuren gelingt.

- Figuren in rechtwinkligen Rastern können in nicht-rechtwinklige Raster abgebildet (verzerrt) werden.
- Auch sprachlich komplexere, mehrschrittige Aufgaben werden gelöst.
- Es werden alle Spiegelachsen in ebenen Figuren erkannt.
- Das Erstellen von Bauplänen gelingt.
- Über eine Raumvorstellung kann ein Perspektivenwechsel vollzogen werden, auch wenn die jeweiligen Ansichten Überdeckungen der einzelnen Körper aufweisen.
- Punkte im Koordinatensystem werden gefunden.
- Die Relationen „senkrecht zueinander“ und „waagerecht zueinander“ werden beherrscht.
- Einfache Sachrechenaufgaben mit geometrischen Inhalten können gelöst werden.
- Seltenere Körper (z.B. Pyramide, Kegel) sind bekannt.

Sachrechnen/Größen

Fähigkeitsniveau 1: Elementare Fähigkeiten (Grundlegende Fähigkeiten im Umgang mit Größen, Darstellungen und Wahrscheinlichkeit)

- Vertraute Maßeinheiten können bei einfachen Aufgaben geordnet, umgerechnet und verglichen und es kann mit ihnen gerechnet werden (Längen-, Zeit-, Gewichts- und Geldeinheiten).
- Die Anwendung einschrittiger Operationen in authentischen Aufgaben gelingt. Die Anwendung von Addition (auch wiederholte Additionen) und Subtraktion in authentischen Aufgaben gelingt bei Aufgaben mit Auswahl aus vorgegebenen Lösungen.
- Daten können aus übersichtlich gestalteten Tabellen und Darstellungen entnommen und ggf. Folgerechnungen (einschrittige Addition/Subtraktion) vorgenommen werden.
- Wahrscheinlichkeitsaussagen zu den Begriffen sicher / unmöglich / möglich, aber nicht sicher werden korrekt kategorisiert.
- Einfache Größenvorstellungen sind vorhanden.
- Offensichtlich unlösbare Aufgaben werden erkannt.

Fähigkeitsniveau 2: Erweiterte Fähigkeiten (Entwickelte Fähigkeiten im Umgang mit Größen, Darstellungen und Wahrscheinlichkeit)

- Im Umgang mit vertrauten Maßeinheiten (Längen, Zeit-, Gewichts- und Geldeinheiten) können Aufgaben bis in den Tausender-Zahlenraum gelöst werden.
- Lösungen von authentischen Aufgaben, die Umrechnungen von Maßeinheiten erfordern, gelingen.
- Die Zuordnung arithmetischer Operationen/Relationen zu Sachsituationen gelingt.

- Daten können aus komplexeren Tabellen und Darstellungen entnommen und ggf. Folgerechnungen (Addition/Subtraktion) korrekt durchgeführt werden.
- Kenntnisse zur Wahrscheinlichkeit sind vorhanden und können in alltagsbezogenen Sachsituationen angewendet werden.
- Der Umgang mit elementaren Brüchen gelingt.
- Rundungen und Schätzungen gelingen bei Aufgaben mit vorgegebenen Lösungen.
- Verknüpfungen von Operationen werden bewältigt.
- Aufgaben mit mehreren zu verarbeitenden Größen werden gemeistert.

Fähigkeitsniveau 3: Fortgeschrittene Fähigkeiten (Eigenständige Problemlösungen)

- Funktionale Beziehungen zwischen Maßen können eigenständig hergestellt und verglichen werden.
- Aufgaben, die mehrere Teilschritte umfassen, werden beherrscht.
- Sprachlich formulierte Relationen können in arithmetische Terme übersetzt werden.
- Bei Aufgaben ohne vorgegebene Fragestellung kann eigenständig eine Aufgabe formuliert und bearbeitet werden.
- Unlösbare Aufgaben, die eine mentale Vorstellung des geschilderten Szenarios erfordern, werden erkannt.
- Die mathematische Modellierung problemhaltiger Sachsituationen gelingt.

Um den Zusammenhang zwischen den Aufgaben und der Zuordnung zu den Fähigkeitsniveaus insbesondere für die Lehrkräfte zu verdeutlichen und um Anreize zur Unterrichtsgestaltung zu geben, wurden sowohl für Mathematik als auch für Deutsch „Didaktische Erläuterungen“ (vgl. http://www.uni-landau.de/vera/aufgaben2005.htm) im Internet zur Verfügung gestellt.

3 Vergleichsarbeiten im September 2005

3.1 Ablauf

Von Ende August bis Anfang September 2005 fanden in den sieben beteiligten Ländern Erprobungsphasen mit Fokus auf der Internetnutzung statt. Vom 12. September bis zum 16. September 2005 konnten die Schulen in Berlin, Brandenburg, Bremen, Mecklenburg-Vorpommern, Nordrhein-Westfalen und Rheinland-Pfalz internetbasiert und menügesteuert einen Teil der Mathematik-Aufgaben auswählen (zehn Aufgaben, die Hälfte der insgesamt erfassten Aufgaben). Vom 12. September bis zum 27. September hatten alle Lehrkräfte außer in Nordrhein-Westfalen außerdem die Möglichkeit, die Lösungshäufigkeiten für die gewählten Aufgaben einzuschätzen, um freiwillig die eigene Diagnosegenauigkeit überprüfen zu können. In Nordrhein-Westfalen und Bremen wurden die Deutsch-Testhefte zentral gedruckt und kurz vor den Vergleichsarbeiten an die Schulen ausgeliefert. In allen anderen Ländern sowie im Fach

Mathematik konnten die Testhefte (Mathematik: Zentral- und Wahlaufgaben, Deutsch: nur Zentralaufgaben) von den Schulen vier Tage vor der Durchführung der Vergleichsarbeit in Mathematik heruntergeladen werden. Die Schulen in Schleswig-Holstein erhielten nur zentral vorgegebene Testaufgaben.

Zeitgleich am 27. September 2005 wurde in allen öffentlichen und in den meisten privaten Grundschulen der beteiligten Bundesländer die Vergleichsarbeit im Fach Mathematik geschrieben, am 29. September in Deutsch. Die Bearbeitungszeit betrug in beiden Fächern 50 Minuten. Nach der Durchführung der Vergleichsarbeiten konnten die Lehrkräfte neben diversen Informationsmaterialien (z.B. didaktische Erläuterungen) die Korrekturanweisungen für die zentral vorgegebenen und die von ihnen ausgewählten Mathematik-Aufgaben vom Landauer Universitätsserver herunterladen. Nach der Korrektur durch die Lehrkräfte konnten diese die Ergebnisse offline in ein Excel-Sheet oder online direkt im geschützten Internet-Bereich von VERA auf dem Universitätsserver eingeben. Die offline eingegebenen Werte wurden dann zum Server hochgeladen und konnten auf den VERA-Internetseiten eingesehen werden. Zwei Wochen nach der vollständigen Dateneingabe konnten die Lehrkräfte die ersten Ergebnisse ihrer Schülerinnen und Schüler abrufen. Für den Abschluss der Dateneingabe hatten die Bremer Schulen bis zum 04. November Zeit.

3.2 Elemente der Ergebnisrückmeldungen

Die Ergebnisse wurden den Schulen in zwei Wellen über das Internet zurückgemeldet, ab dem 19. Dezember 2005 standen die Zusatzinformationen zur Verfügung. Die Lehrerinnen und Lehrer sollten anschließend die Informationen in geeigneter Weise an Eltern, Schulleitung oder Schulaufsicht weiterleiten und erläutern.

Zeitnah zur vollständigen Dateneingabe konnten die Lehrkräfte die **Basisinformationen** (erste Rückmeldewelle) einsehen:

- Fähigkeitsniveaus auf Ebene der einzelnen Schülerinnen und Schüler als Tabellenübersicht, auf Klassen- und auf Schulebene als Verteilungen (gestaffelte Balkendiagramme)

Nur im Fach Mathematik:

- Analysen zur Fehlerhäufigkeit: Die in der jeweiligen Klasse aufgetretenen Fehler wurden den Fehlerhäufigkeiten aus der Normierungsstichprobe gegenübergestellt, um so Informationen über Fehlerschwerpunkte und damit Ansatzpunkte für differenzierte Interventionsmaßnahmen in der Klasse zu gewinnen.
- Analysen zur diagnostischen Kompetenz: Die Einschätzungen der Lösungshäufigkeiten für die Wahlaufgaben wurden der tatsächlichen Lösungshäufigkeit gegenüber gestellt.

In der zweiten Rückmeldewelle wurden den Lehrkräften ab dem 19. Dezember 2005 weitere **Zusatzinformationen** zur Verfügung gestellt, und zwar:

- Fähigkeitsniveaus im Vergleich zum eigenen Bundesland: Die Gegenüberstellung

der Verteilung der Fähigkeitsniveaus in der eigenen Klasse und Schule wurde kontrastiert mit der Darstellung der Ergebnisse des jeweiligen Bundeslandes.

- Fähigkeitsniveaus im „fairen Vergleich“: Zusätzlich erhielten die Lehrkräfte Ergebnisse zum Vergleich der eigenen Klasse mit einer länderunspezifischen kontextähnlichen Vergleichsgruppe.
- Lösungshäufigkeiten in den Zentralaufgaben bei VERA 2005: In dieser Darstellung wurden die richtigen Lösungen der jeweiligen Klasse im Vergleich zur Gesamtpopulation verortet.
- Analysen zur diagnostischen Kompetenz: Aus der Rangordnung der geschätzten und tatsächlichen Lösungshäufigkeiten in den Wahlaufgaben wurde ein Genauigkeitsindex für die eigene Einschätzung (Korrelation) abgeleitet. Dieser Wert wurde in einer Häufigkeitsverteilung der Population von VERA 2005 gegenüber gestellt.

Als Neuerung wurden den Schulen ab Mitte Januar zwei Gesamtdokumente mit allen Ergebnissen aus VERA 2005 zur Verfügung gestellt (Teil 1: Ergebnisse zu den Fähigkeitsniveaus; Teil 2: Aufgabenbezogene Ergebnisse). Diese können von den Schulen als schulinterne Dokumentation der Ergebnisse in den Vergleichsarbeiten 2005 oder als Basis für Berichte an z.B. die Schulaufsicht genutzt werden.

3.3 Ergebnisse zur Durchführung der Vergleichsarbeiten

Erstmalig wurden in Landau nach den Vergleichsarbeiten 2005 gezielt Informationen zur Nutzung des Internets sowie zum Supportaufkommen während der Durchführung der Vergleichsarbeiten ausgewertet. Außerdem wurde eine stichprobenartige Überprüfung der Ergebnis-Eingaben der Lehrkräfte vorgenommen, unter anderem mit dem Ziel, Erkenntnisse über die Verständlichkeit der Korrekturanweisungen zu erlangen.

In den folgenden Unterkapiteln soll etwas näher auf diese Ergebnisse eingegangen werden.

3.3.1 Auswertung in den Schulen

Um eine Einschätzung für die Genauigkeit der Ergebnis-Eingabe durch die Lehrkräfte sowie damit verbunden für die Verständlichkeit und Eindeutigkeit der in VERA verwendeten Korrekturanweisungen zu bekommen, wurde eine zufällige Stichprobe[3] der Schulen gezogen, die ihr Test-Material zur Verfügung stellen sollten. Insgesamt wurden 50 Schulen in den sieben teilnehmenden Ländern von uns angeschrieben und darum gebeten, Kopien der Testhefte (Mathematik und Deutsch) von jeweils zwei per

[3] Bei dieser Stichprobe handelt es sich um eine der Ländergröße proportionalen Zufallsstichprobe der Gesamtheit aller an VERA teilnehmenden Klassen. Repräsentativität für einzelne Länder kann daher nicht beansprucht werden.

Zufall bestimmten Schülerinnen bzw. Schülern einzusenden. Dieser Aufforderung kamen 49 Schulen nach.

Diese wurden in Landau erneut korrigiert und mit den entsprechenden Eingaben durch die Lehrkräfte abgeglichen.

Deutsch

Die von uns gefundenen Abweichungsarten sind in Tabelle 1 schematisch dargestellt.

Tabelle 1: *Abweichungsarten*

		Korrekte Lösung lt. Korrekturanweisung		
		richtig	falsch	nicht bearbeitet
Lehrerurteil	richtig		1. und 2.	kommt nicht vor
	falsch	4.		kommt nicht vor
	nicht bearbeitet	kommt nicht vor	3.	

Bei allen Schulen lagen Abweichungen (Korrekturmängel) verschiedener Arten vor, die wie folgt unterschieden wurden:

Tabelle 2: *Prozentuale Verteilung aller Abweichungen in Deutsch*

	Abweichungsarten	N	in Prozent der vorliegenden Aufgabenkorrekturen
1.	falsch gelöst und als „richtig“ bewertet	68	3,65
2.	großzügige Bewertung	36	1,93
3.	falsch gelöst und als „n.b.“ bewertet	16	0,86
4.	richtig gelöst und als „falsch“ bewertet	10	0,54
Anzahl der Beanstandungen		130	6,98
Anzahl vorliegende Aufgabenkorrekturen		1862	100

Falsch gelöste Aufgaben, die als „richtig“ bewertet wurden:

Diese Art der Abweichung beschreibt Fälle, in denen als „falsch“ zu bewertende Bearbeitungen als „richtig“ bewertet wurden. Sie machen mit 68 Stück oder 3,65 Prozent den Löwenanteil aller Korrekturmängel aus. Diese Art der Abweichung kam bei 37 Schulen vor. Es kamen jedoch auch einzelne Schulen (N = 5) vor, bei denen zusätzlich ebenfalls Abweichungen der Art vorkamen, dass richtig gelöste Aufgaben als „falsch“ bewertet wurden. Diese Mischfälle könnte man als unaufmerksame Korrekturen auffassen. 19 Schulen zeigen neben der hier dargestellten Abweichung auch großzügige Bewertungen.

Großzügige Bewertung:

Hier handelt es sich um Grenzfälle, in denen die Bearbeitung streng genommen als „falsch“ hätte eingestuft werden müssen und in denen die Lehrkräfte die Korrekturan-

weisungen etwas liberaler ausgelegt haben. Diese Abweichungen können im Wohlwollen der Lehrkräfte oder in bestimmten Formulierungen in den Korrekturanweisungen (Interpretationsspielraum, der insbesondere für offene Deutsch-Aufgaben typisch ist) begründet sein. Diese Abweichungen entsprechen 1,93 Prozent (36 Fälle) und kamen in 29 Schulen vor.

Falsch gelöste Aufgaben, die als „nicht bearbeitet (n. b.)" gewertet wurden:

In diesem Fall wurden Aufgaben, die aus unserer Sicht eindeutig als „falsch" einzustufen waren, von den Lehrkräften als „nicht bearbeitet" gewertet. Diese Abweichungen entsprechen 0,86 Prozent bzw. 16 Fällen und kamen in zehn Schulen vor.

Richtig gelöste Aufgaben, die als „falsch" bewertet wurden:

Diese Art der Abweichung benachteiligt letztlich den Schüler, da hier eine richtige Antwort von der Lehrkraft als „falsch" eingestuft wurde. Diese Form der Fehlzuordnung kam nur in 0,54 Prozent der Fälle vor (neun Schulen). In acht Schulen trat dieser Fehler nur einmal auf. Nur in einem Fall wurde dreimal eine richtige Antwort irrtümlicherweise als „falsch" eingestuft.

Fazit:

In Bezug auf einzelne Schulen und Bundesländer sind keine auffälligen Fehlermuster zu berichten. Es scheint keine systematische Tendenz zur „besseren Darstellung" der eigenen Schüler zu geben, selbst wenn es nachweislich einen geringen Prozentsatz an Aufgabenkorrekturen gibt, den man als absichtliche Verfälschung interpretieren könnte. Viele der Beanstandungshäufigkeiten gehen vermutlich auf Konzentrationsprobleme oder nicht auf als Täuschung zu interpretierendes Wohlwollen seitens der Lehrkräfte zurück. Darüber hinaus ist insbesondere der Bereich Deutsch mit der Tatsache konfrontiert, dass die Korrekturanweisungen bestimmte Interpretationsspielräume zulassen müssen (z. B. bei offenen Aufgaben), was sich in den am zweithäufigsten vertretenen Beanstandungen, nämlich den „großzügigen Bewertungen" ausdrückt. Dass knapp 60 Prozent aller 130 Beanstandungen auf nur drei der 19 Teilaufgaben entfallen, kann als deutliches Indiz für diese These aufgefasst werden.

Mathematik

Für 43 der 49 Schulen wurden Abweichungen zwischen der Landauer Korrektur und der Ergebnisübermittlung durch die Lehrkraft gefunden. Unterscheiden lassen sich insgesamt drei verschiedene Abweichungsarten: 1. falsch gelöste Aufgaben wurden nicht dem richtigen Fehlertypen zugeordnet, 2. falsch gelöste Aufgaben wurden als „richtig" bewertet, 3. richtig gelöste Aufgaben wurden als „falsch" bewertet. Die folgende Tabelle zeigt die prozentuale Verteilung aller Abweichungen (N = 183) auf die drei Abweichungstypen:

Tabelle 3: *Prozentuale Verteilung aller Abweichungen in Mathematik*

Abweichungsart	Anzahl Beanstandungen	Prozent Beanstandungen	in Prozent der vorliegenden Aufagbenkorrekturen
falsch gelöst, falsche Fehlertypangabe (1)	110	60,1	2,7
falsch gelöst, als „richtig“ bewertet (2)	43	23,5	1,1
richtig gelöst, als „falsch“ bewertet (3)	30	16,4	0,7
Anzahl der Beanstandungen	183		4,6
Anzahl vorliegender Aufgabenkorrekturen	4018[4]		

Am häufigsten wurde der Fehlertyp der Falschlösungen nicht entsprechend der Vorgaben in den Korrekturanweisungen bestimmt. Insgesamt kam es bei dieser Art der Abweichung zu 110 falschen Fehlertypzuordnungen, die sich auf 39 der 43 überprüften Schulen verteilten. In 43 Fällen, verteilt auf 26 Schulen, wurden Falschlösungen als „richtig“ eingestuft, in 30 Fällen, verteilt auf 19 Schulen, die richtige Aufgabenlösung als „falsch“ gewertet. Mögliche Ursachen hierfür könnten in einer Fehlinterpretation der Korrekturanweisungen bzw. in empfundenen Widersprüchen in der Art der vorgegebenen Aufgabenkorrektur liegen.

Tabelle 4 gibt einen Überblick über die Häufigkeiten der Angabe falscher Aufgabenlösungen als „richtig“ bzw. umgekehrt.

Tabelle 4*: Auftretenshäufigkeit der Abweichungsarten (2) und (3)*

Abweichungsart	Anzahl
alleiniges Auftreten von Abweichungstyp (2)	17
alleiniges Auftreten von Abweichungstyp (3)	9
gleichzeitiges Auftreten von Abweichungstyp (2) und (3)	19

Das gehäufte gleichzeitige Auftreten von Abweichungstyp (2) und (3) in vielen der Schulen legt nahe, dass die Fehlkategorisierung von falsch gelösten Aufgaben als „richtig“ bzw. richtig gelöster Aufgaben als „falsch“ nicht Konsequenz absichtlicher Verfälschung ist, sondern aus unterschiedlich strengen Bewertungsmaßstäben und möglicherweise auftretenden Konzentrationsproblemen bei der Dateneingabe resultiert. Die meisten Fehlzuordnungen (auch bei der Falscheinordnung von Fehlertypen) konzentrieren sich auf wenige, in der Auswertung aufwändige Aufgaben mit evtl. verbesserungswürdigen Korrekturanweisungen (z.B. FV093: Zuordnung von Alltagsgegenständen zu geometrischen Formen; FV103: Zerlegen einer Figur in Teilfiguren).

[4] Die Anzahl der Teilaufgaben variiert durch die Wahlmöglichkeiten von Schule zu Schule. Der Zentralteil enthält 23 Teilaufgaben, der Wahlteil umfasst zwischen 14 und 29 Teilaufgaben, der häufigste Wert ist 22. Die hier angenommenen 18 Teilaufgaben für den Wahlteil stellen also eine sehr konservative Schätzung dar.

Keine der Schulen und keines der Bundesländer fällt durch eine gehäufte Anzahl von Fehlkategorisierungen auf.

Insgesamt machen die Abweichungsarten, die zu veränderten Leistungsschätzungen der Schülerinnen und Schüler führen könnten (Abweichungsarten (2) und (3)), den kleineren Teil der Beanstandungen aus und treten, setzt man sie ins Verhältnis zu der Gesamtzahl der durch die Lehrkräfte insgesamt im Rahmen der Vergleichsarbeiten für eine Klasse korrigierten (Teil-)Aufgaben, in deutlich unter zwei Prozent der Fälle auf. Daraus lässt sich folgern, dass die Materialien zur Korrektur der VERA-Mathematikaufgaben weitestgehend eindeutig und verständlich sind und von den Lehrkräften gewissenhaft angewendet werden.

3.3.2 Ergebnisse zur Nutzung des Internet

Grundlegend für die Durchführung von flächendeckenden und bundesländerübergreifenden Vergleichsstudien ist die Nutzung des Internet: So wählen die an VERA teilnehmenden Lehrkräfte über das Internet einen Teil der Testaufgaben aus, laden das dynamisch generierte Aufgabenheft herunter und geben im Anschluss die Resultate der Schülerinnen und Schüler online ein. Die aufbereiteten Ergebnisse und Informationsmaterialien werden ebenfalls über das Internet distribuiert. Ergänzend nehmen die Lehrkräfte an einer Online-Befragung zum sozialen Hintergrund ihrer Klassen ein und haben die Möglichkeit, in einer freiwilligen Befragung Lob und Kritik am Projekt zu äußern.

Mit der Nutzung des Internet als Instrument für Schulleistungsstudien wurde durch das Projekt VERA Neuland betreten: Obwohl inzwischen so gut wie alle Schulen über einen Zugang zum Internet verfügen, bedeutet die Nutzung des Internet für einige Schulen eine Umstellung. Sie soll jedoch zu einem Aufschwung in der effizienten Nutzung des Internet für die schulische Qualitätssicherung führen und damit auch einen Beitrag zur Förderung der Medienkompetenz leisten.

Soft- und Hardware-Voraussetzungen der VERA-Nutzer

Sowohl an die Hardware als auch an die Software gibt es grundlegende Anforderungen, damit der VERA-Prozess von den Schulen vollständig durchlaufen werden kann. Der Begriff Hardware wird dabei als Oberbegriff für die maschinentechnische Ausrüstung eines Computersystems verwendet. Von besonderer Bedeutung ist in diesem Zusammenhang die Bildschirmauflösung der für die Durchführung und Auswertung der Vergleichsarbeiten verwendeten Computer.

Entsprechende Informationen werden mit Hilfe von JavaScript auf der Startseite des geschützten Bereichs erfasst. Die in Tabelle 5 dargestellten Häufigkeiten beziehen sich demnach nur auf den Aufruf der Startseite von August 2005 bis Februar 2006. Alles in allem haben wir 53 unterschiedliche Bildschirmauflösungen registriert. Da die VERA-Seiten für 800 x 600 Pixel optimiert wurden, sind bei niedrigeren Auflösungen Teile der Seite wie z.B. Navigationselemente weder sicht- noch bedienbar. Entsprechend proble-

matische Bildschirmauflösungen wurden insgesamt 215-mal festgestellt, also in ca. 0,1 Prozent der Fälle.

***Tabelle 5**: registrierte Bildschirmauflösungen*

		Häufigkeiten (Seiten-aufrufe)	Prozent
Bildschirmauflösung	1024 x 768	136 778	67,4
	800 x 600	35 899	17,7
	1280 x 1024	18 418	9,1
	1152 x 864	5 032	2,5
	Gesamt	196 127	96,7

Gewissermaßen als Komplement zur Hardware stellt die Software die nichtphysischen Funktionsbestandteile eines Computers dar (Computerprogramme sowie die zur Verwendung mit Computerprogrammen bestimmten Daten). Dabei wird zwischen Systemsoftware und Anwendungssoftware unterschieden: Während Systemsoftware für das ordentliche Funktionieren des Computers erforderlich ist (v.a. Betriebssysteme), unterstützt Anwendungssoftware den Benutzer bei der Ausführung seiner Aufgaben (z.B. Internet-Browser).

Auch diese Informationen wurden von uns geloggt und zwar in Bezug auf den Aufruf der Seiten im geschützten Bereich von VERA. In Bezug auf die verwendeten Betriebssysteme konnten wir 18 unterschiedliche Programme identifizieren. Zu den ältesten Betriebssystemen gehören Windows 3.11 und OS/2. Da für diese Versionen keine aktuellen Browser verfügbar sind, kann das Webangebot nur sehr eingeschränkt benutzt werden. Diese Versionen wurden von uns jedoch nur 247-mal registriert.

***Tabelle 6:** Betriebssysteme*

		Häufigkeiten (Seitenaufrufe)	Prozent
Betriebssysteme	Windows XP	9 711 158	51,0
	Windows 2000	4 179 498	21,9
	Windows 98	3 435 934	18,0
	Unix und Derivate	188 415	1,0
	MacIntosh	172 767	0,9
	Gesamt	17 687 772	92,8

In Bezug auf die Browser-Software haben wir 21 Programme unterschiedlicher Version registriert. Die häufigsten Browser können der Tabelle 7 entnommen werden. Für eine reibungslose internetgestützte Durchführung der Vergleichsarbeiten sind auch hier bestimmte Mindestanforderungen zu erfüllen. Ältere Browser wie z.B. Netscape 4.75 unterstützen aktuelle Webstandards nicht ausreichend und können demnach nur eine eingeschränkte Funktionalität des VERA-Webangebots gewährleisten. Entsprechende problematische Versionen lassen sich bei etwa 10,5 Prozent der Seitenaufrufe nachwei-

sen. Mit anderen Worten bedeutet dies, dass über 89 Prozent der Seitenaufrufe mit einem ausreichend aktuellen Browser erfolgt sind.

Tabelle 7: *Browser*

		Häufigkeiten (Seitenaufrufe)	Prozent
Browser	Internet Explorer	12 470 571	85,2
	davon veraltete Versionen (3.0 bis 5.23)	1 509 956	10,4
	Gecko: Netscape, Mozilla, Firefox	1 670 741	11,4
	davon veraltete Versionen (Netscape 3.01 bis 4.79)	10 176	0,1
	Sonstige	498 560	3,4

Aufrufe der Seiten im geschützten Bereich

Die Betrachtung der Nutzungsstatistiken unter zeitlicher Perspektive (in Abhängigkeit von Tageszeit und Monat) ist insbesondere aussagekräftig für den Unterstützungsbedarf im Rahmen der Durchführung der Vergleichsarbeiten. Die Abbildung 2 zeigt deutlich, dass praktisch den gesamten Tag und auch während der Nacht Aktivitäten im geschützten Bereich zu verzeichnen waren. Die Nutzung steigt gegen sieben Uhr morgens auf etwa fünf Prozent an und nimmt bis 13 Uhr zu (11,7 Prozent). Durch die Besetzung der Support-Hotline sowie des E-Mail-Supports von sieben Uhr bis 15 Uhr konnten etwa 68 Prozent der Aktivitäten bei Bedarf begleitet werden. Möglicherweise wäre es zu überlegen, den Support auf Beginn acht Uhr zu verlegen und bis 18 Uhr zu verlängern: So hätten etwa 87 Prozent der Aktivitäten abgedeckt werden können.

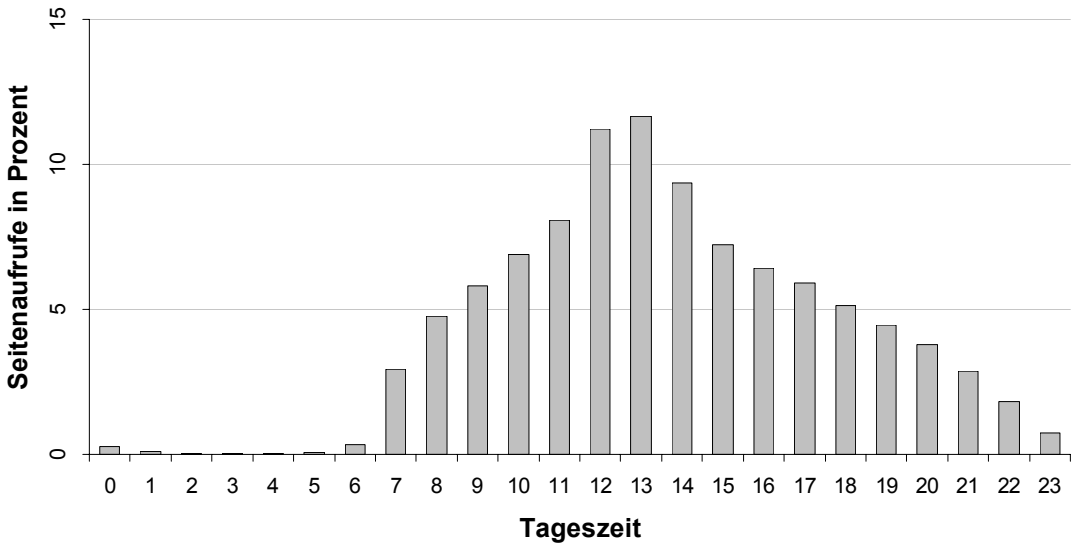

Abbildung 2: *Häufigkeiten der Seitenaufrufe in Abhängigkeit von der Tageszeit*

Ebenso aussagekräftig wie der Blick auf die Nutzung in Abhängigkeit von der Tageszeit sind die Nutzungsstatistiken im Verlauf der VERA-Durchführung. In Tabelle 8 zeigt

sich deutlich, dass sich die Aktivitäten im geschützten Bereich v.a. auf die Monate September, Oktober und November konzentrieren. Während dieser Zeit war der Telefon- und E-Mail-Support durch die Universität Landau voll besetzt und wurde erst im Lauf des Novembers allmählich reduziert.

Angesichts der Tatsache, dass 99 Prozent der teilnehmenden Schulen den VERA-Prozess vollständig durchlaufen haben, ist es über den zeitlichen Verlauf hinaus interessant, welche Seiten besonders häufig aufgerufen wurden. Denn für das Erreichen der dargestellten Ziele von VERA (vgl. 1.1, S. 4) sollen sich die Lehrkräfte und Schulen intensiv mit den Ergebnissen und zur Verfügung gestellten Informationsmaterialien auseinandersetzen. Im Folgenden werden daher für spezifische Seiten Nutzungsstatistiken berichtet.

Tabelle 8: *Nutzungsstatistiken im VERA-Zyklus*

		Aufgaben	Häufigkeiten (Seitenaufrufe)	Prozent
Monat	August	• Anmeldung im geschützten Bereich	211 120	3,2
		• Probelauf in den Schulen		
	September	• Eingabe der Schul- und Schülerdaten	1 778 866	27,0
		• Aufgabenauswahl Mathematik		
		• Download der Testhefte und Korrekturanweisungen		
		• Schreiben der Vergleichsarbeiten		
	Oktober	• schulinterne Auswertung	1 871 574	28,4
		• Dateneingabe der Ergebnisse durch die Lehrkräfte		
		• Klassen- und Schul-Ergebnisse		
	November	• Befragung der Zentralstichprobe und aller Bremer Schulen zum Kontext der Schule	1 541 646	23,4
	Dezember	• länderbezogene Ergebnisse und "fairer Vergleich"	627 952	9,5
		• Rückmeldung der Ergebnisse an die Eltern		
	Januar	• innerschulische Auseinandersetzung mit den Ergebnissen	432 205	6,5
	Februar		136 576	2,1

Die folgende Tabelle 9 gibt den prozentualen Anteil der an VERA teilnehmenden Schulen wieder, die bestimmte Seiten im geschützten Bereich aufgerufen haben. Dazu wurden die Webseiten differenziert nach Inhaltsbereich (Allgemein, Mathematik, Deutsch), Rückmeldungszeitpunkt (Basis- vs. Zusatzinformationen) sowie nach spezifischem Inhalt. In den beiden letzten Tabellenspalten sind die landesspezifischen Nutzungsstatistiken den länderübergreifenden Prozentwerten gegenüber gestellt. Bei letzteren Werten ist zu berücksichtigen, dass es sich um einen gewichteten Wert handelt, d.h. dass die

einzelnen Länderwerte gleich gewichtet eingehen. So wird vermieden, dass Bundesländer mit größeren Schülerzahlen (wie z.B. Nordrhein-Westfalen) die länderübergreifenden Vergleichswerte dominieren. Gleichzeitig sind aufgefundene Unterschiede in den einzelnen Ländern eindeutiger zu interpretieren.

Zunächst einmal fällt auf, dass die Zahl der Schulen, die sich Ergebnisse ansehen, im zeitlichen Verlauf abnimmt. So werden die Basisinformationen von rund 80 bis 90 Prozent der Schulen aufgerufen – die Zusatzinformationen jedoch deutlich seltener. Dabei ist wenig überraschend, dass die Seite mit den allgemeinen Materialien am häufigsten und mit einem Wert von rund 95 Prozent durch nahezu alle Schulen aufgerufen wurde. Auf dieser Seite finden sich u.a. Informationsmaterialien, die für den Ablauf der Vergleichsarbeiten grundlegend sind (Handreichung zur Durchführung der Vergleichsarbeiten, Erläuterungen zur Dateneingabe) sowie die Druckversion der Ergebnisse, Elternbrief und Elternrückmeldeformular.

Tabelle 9: *Nutzungsstatistiken für den geschützten Bereich,* *Angaben in Prozent*

			Bremen	**Gesamt**
			N = 106*	**N = 6624****
Allgemein		Informationsmaterialien	93,4	94,9
Mathematik	Basis	Fähigkeitsniveaus Klasse/Schule	93,4	91,8
		Fehleranalyse	88,7	83,8
		Diagnosegenauigkeit	86,0	81,5
	Zusatz	Fähigkeitsniveaus im Landesvergleich	71,7	66,0
		Fairer Vergleich	62,3	46,1
		Lösungshäufigkeiten	53,8	55,5
		Diagnosegenauigkeit im Landesvergleich	42,1	35,1
		Informationsmaterialien	82,1	77,2
Deutsch	Basis	Fähigkeitsniveaus Klasse/Schule	90,6	89,8
	Zusatz	Fähigkeitsniveaus im Landesvergleich	69,8	67,7
		Fairer Vergleich	51,9	36,8
		Lösungshäufigkeiten	34,0	36,9
		Informationsmaterialien	69,8	61,9

* Schulen in Bremen, die Eingaben für die Diagnosegenauigkeit gemacht haben: $N_{Diag.} = 57$

** Schulen gesamt, die Eingaben für die Diagnosegenauigkeit gemacht haben: $N_{Diag.} = 1531$

Da es sich bei den Bremer Schulen um eine kleine Stichprobe von 106 Schulen handelt, sollten alle Unterschiede in den Nutzungsstatistiken mit Vorsicht interpretiert werden. Es ist aber auf jeden Fall ein deutlicher Trend dahin gehend zu verzeichnen, dass die Bremer Schulen vergleichsweise aktiver als die Gesamtstichprobe im VERA-Web gewesen sind. Vor allem den Ergebnissen zum „fairen Vergleich“ (Mathematik: 62,3 Prozent; Deutsch: 51,9 Prozent) sowie den Deutsch-Informationsmaterialien (69,8 Prozent) wurde verstärkt Aufmerksamkeit gewidmet.

Alles in allem können die Häufigkeiten für die Seitenaufrufe durchaus als erfreuliches Ergebnis gewertet werden. Denn neben der hohen Aktivität der Bremer Schulen ist außerdem zu verzeichnen, dass insbesondere die klassenspezifische Fähigkeitsniveauverteilung und die fachspezifischen Informationsmaterialien, die wir als grundlegend für Schul- und Unterrichtsentwicklungsprozesse erachten, in Bremen mit am häufigsten aufgerufen wurden.

3.3.3 Ergebnisse zum Supportaufkommen

Zur Verfolgung der einzelnen Supportanfragen wurde im Jahr 2005 ein Ticketsystem eingesetzt. Die Wahl fiel hierbei auf OTRS (Open source Ticket Request System), das unter der GNU General Public License, GPL, vertrieben wird. Dieses System ist frei verfügbar und bietet durch das Webinterface die Möglichkeit, auch die bei den Länder-supports eingehenden Anfragen zentral zu verfolgen. Über entsprechende Konfigurationen konnte das System auf den hier vorliegenden Anwendungsfall angepasst werden.

Damit konnten sowohl E-Mail- als auch Telefonanfragen gemeinsam verwaltet werden. Die Abarbeitung der einzelnen Anfragen konnte durch die Zuordnung zu verschiedenen Warteschlangen prioritäts- und inhaltsbezogen gesteuert werden. Tabelle 10 gibt einerseits einen Überblick über das gesamte Supportaufkommen und andererseits über das aus Ihrem Bundesland.

Ein Ticket symbolisiert hierbei eine inhaltlich abgeschlossene Anfrage. Bei der Beantwortung und endgültigen Bearbeitung eines solchen Tickets kommt es teilweise zu mehreren Kontakten mit einer Schule. Eine Telefonanfrage entspricht dabei normalerweise einem Kontakt, Mailanfragen zwei Kontakten (Anfrage und Antwort). Vom Beginn des Probelaufs bis zum 15. März 2006 gingen insgesamt 3824 Anfragen (Tickets) ein. Die 55 Anfragen aus Bremen kommen von 32 verschiedenen Schulen. Der VERA-Support stand somit mit 32,7 Prozent der Bremer Schulen in Verbindung. Pro Anfrage ergaben sich dabei im Durchschnitt 2 Kontakte. 71,4 Prozent dieser Kontakte erfolgten per Mail, entsprechend 28,6 Prozent per Telefon. In Bremen ist der E-Mail-Anteil damit höher als im Durchschnitt über die sieben Bundesländer (59,8 Prozent). Im Durchschnitt dauerte es 12,5 Stunden, bis eine Anfrage mit dem Status „erfolgreich beantwortet“ geschlossen werden konnte. Da in diese Berechnung auch die Wartezeiten von Tickets eingehen, die außerhalb der Supportzeiten eintrafen (z.B. abends oder am Wochenende), zeigt dies die schnelle Reaktion auf Anfragen aus den Schulen.

Bei der inhaltlichen Zuordnung ergibt sich folgendes Bild: 30 Prozent der Anfragen bezogen sich auf technische Fragestellungen, vier Prozent beinhalteten positives oder negatives Feedback und 66 Prozent waren allgemeiner Natur (Fragen zum Prozedere, zu den Aufgaben usw.).

Tabelle 10: *Informationen über das Supportaufkommen*

	Gesamt	Bremen
Anzahl der Tickets	3824	55
Anzahl der Schulen mit Ticket %	35,4	32,7
Anzahl Kontakte	6631	112
Kontakte pro Ticket	1,7	2,0
Mailkontakte %	59,8	71,4
Telefonkontakte %	40,2	28,6

3.4 Gesamt-Ergebnisse für die Fähigkeitsniveaus

3.4.1 Gesamt-Verteilung

Die zusammenfassende Darstellung der teilnehmenden Bundesländer ermöglicht einen allgemeinen Blick auf relative Stärken und Schwächen. Zu berücksichtigen ist dabei, dass die Fähigkeitsniveaus abhängig von den verschiedenen Inhaltsbereichen beschrieben sind. Für den Vergleich zwischen den Fächern wird daher auf die Verteilung der Fähigkeitsniveaus innerhalb der einzelnen Inhaltsbereiche zurückgegriffen.

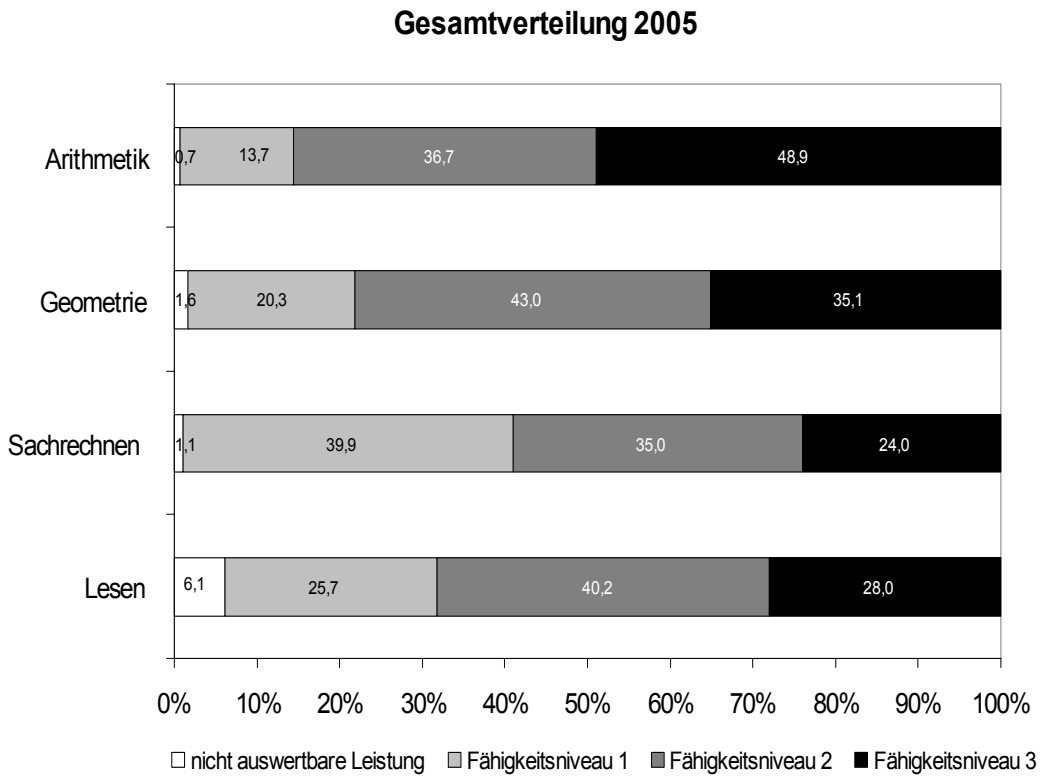

Abbildung 3: *Gesamtverteilung der Fähigkeitsniveaus über alle Länder hinweg;* *Angaben in Prozent*

Für das Fach *Mathematik* wird deutlich, dass in den Inhaltsbereichen Arithmetik und Geometrie der Großteil der Schülerinnen und Schüler erweiterte bis fortgeschrittene Fähigkeiten aufweist, also das Fähigkeitsniveau 2 und 3 erreicht (für Arithmetik deutlich mehr als 80 Prozent, für Geometrie knapp unter 80 Prozent). Der Anteil an Kindern auf dem dritten Fähigkeitsniveau ist für Arithmetik am höchsten (48,9 Prozent). Demgegenüber fällt auf, dass im Sachrechnen verhältnismäßig wenige Schülerinnen und Schüler dem dritten Fähigkeitsniveau (24 Prozent) und knapp 40 Prozent der Schülerinnen und Schüler einem Niveau, das höchstens das Beherrschen elementarer Aufgaben umfasst, zugeordnet werden können. Für das Fach *Deutsch* erreicht ein großer Anteil der Schülerinnen und Schüler im Inhaltsbereich Leseverständnis höchstens ein Fähigkeitsniveau, das elementaren Fähigkeiten entspricht (31,8 Prozent).

Aus der Verteilung der Ergebnisse kann zusammenfassend geschlussfolgert werden, dass vor allem beim Leseverständnis und beim Sachrechnen ein Förderbedarf besteht.

3.4.2 Zusammenhänge zwischen den Inhaltsbereichen

Die folgende Tabelle 11 gibt den Zusammenhang (Interkorrelationen) zwischen den Leistungen in den verschiedenen Inhaltsbereichen wieder. Darin entsprechen die dunkelgrauen Felder den Interkorrelationen innerhalb des Faches Mathematik und die hellgrauen Felder den Zusammenhängen zwischen den Inhaltsbereichen für Mathematik und dem Leseverständnis.

Die Korrelationen liegen in einem mittleren Wertebereich von r = ,51 (Leseverständnis/Geometrie) bis r = ,59 (Arithmetik/Sachrechnen) – dabei finden sich höhere Korrelationen innerhalb der Fachs Mathematik Die Korrelationen zwischen dem Leseverständnis und den drei Mathematik-Inhaltsbereichen könnten als Hinweis für die fächerübergreifende Bedeutung des Leseverständnisses interpretiert werden.

Tabelle 11: *Interkorrelationen zwischen den Leistungen in den Inhaltsbereichen (r)*2005*

	Arithmetik	**Geometrie**	**Sachrechnen**
Lesen	0,53	0,51	0,52
Arithmetik		0,55	0,59
Geometrie			0,53

* minimale Stichprobengröße: N = 280224 (Kinder)

Alles in allem sprechen die mittleren, jedoch nicht perfekten Zusammenhänge zwischen den vier Inhaltsbereichen dafür, nicht von einer Gesamtfähigkeit (z.B. schulische Fähigkeit im Lesen und Mathematik) zu sprechen, sondern vier getrennte Fähigkeitsdimensionen anzunehmen.

4 Landesspezifische Ergebnisse

4.1 Fähigkeitsniveaus

Im Folgenden werden die Fähigkeitsniveaus unter verschiedenen Gesichtspunkten diskutiert: Zunächst wird allgemein die Verteilung der Schülerinnen und Schüler auf den einzelnen Niveaus dargestellt (vgl. 4.1.1, S. 31) sowie Veränderungstrends für die Vergleichsarbeiten in den Jahren 2004 und 2005 diskutiert (vgl. 4.1.2, S. 32). Im Anschluss werden Unterschiede zwischen (bzw. innerhalb) den untersuchten Klassen/Schulen (siehe 4.1.3, S. 33), Geschlechtsunterschiede (siehe 4.1.4, S. 35), Unterschiede zwischen Schülern mit Deutsch als dominanter vs. nicht dominanter Sprache (vgl. 4.1.5, S.36), sowie landesspezifische Vergleiche (z.B. 4.1.6, S.38) genauer beleuchtet. Alle Angaben zu den Individualmerkmalen der Schülerinnen und Schüler stammen von den Lehrkräften und wurden vor der Durchführung der Vergleichsarbeiten online im „geschützten Bereich“ erfasst (siehe 3.1).

4.1.1 Verteilung der Fähigkeitsniveaus in den Ländern

Für das Fach *Mathematik* wird deutlich, dass in den Inhaltsbereichen Arithmetik und Geometrie der Großteil der Schülerinnen und Schüler erweiterte bis fortgeschrittene Fähigkeiten aufweisen, also das Niveau zwei und drei erreichen (für Arithmetik 84,9 Prozent, für Geometrie 78,0 Prozent). In Arithmetik erreicht über die Hälfte der Kinder das Fähigkeitsniveau drei (50,8 Prozent). Demgegenüber fällt auf, dass im Sachrechnen am wenigsten Schülerinnen und Schüler dem dritten Fähigkeitsniveau und über 40 Prozent einem Niveau, das höchstens das Beherrschen elementarer Aufgaben umfasst, zugeordnet werden.

Im Inhaltsbereich *Leseverständnis* findet sich ein besonders großer Anteil an Schülerinnen und Schülern, die höchstens das Fähigkeitsniveau 1 erreichen (32,8 Prozent, davon 6,4 Prozent nicht auswertbare Leistung). Der hohe Anteil an Kindern mit nicht auswertbarer Leistung steht beim Leseverständnis vermutlich im Zusammenhang mit den Kontextmerkmalen (siehe 4.2, S. 39).

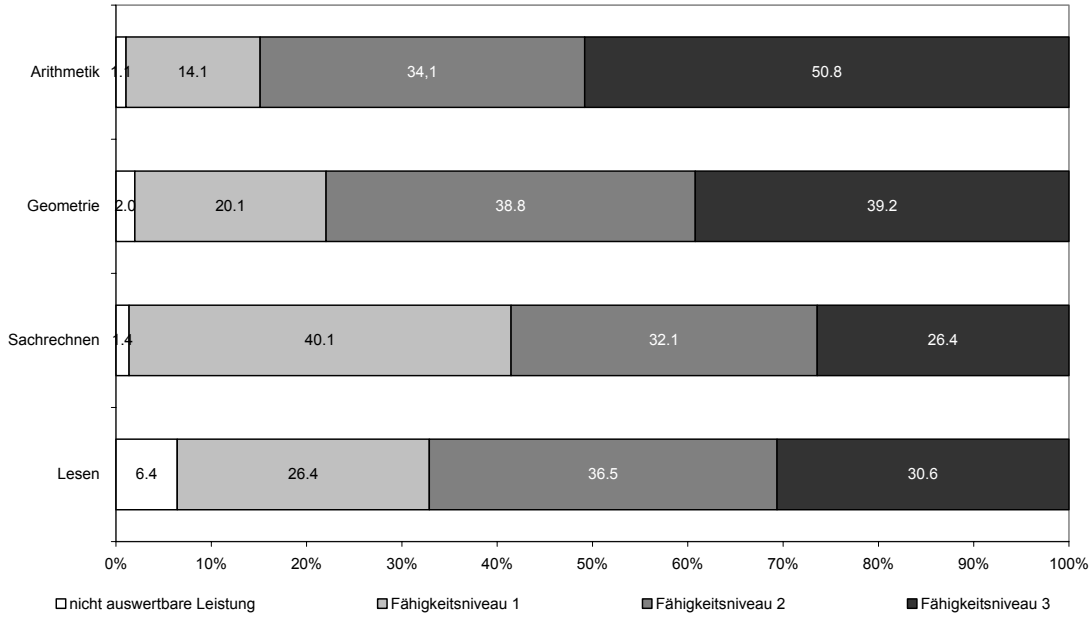

Abbildung 4: *Gesamtverteilung der Fähigkeitsniveaus (2005);* ***Angaben in Prozent***

4.1.2 Veränderungstrends 2004–2005

Die Verteilung der Schülerinnen und Schüler auf die einzelnen Fähigkeitsniveaubereiche ist in 2005 verglichen mit der Verteilung in 2004 (Abbildung 5) im Bereichen Mathematik in etwa gleich geblieben. Im Bereich Leseverständnis sind die Leistungen der Schülerinnen und Schüler gestiegen. So verringerte sich die Anzahl der Kinder mit nicht auswertbaren Leistungen und mit Fähigkeitsniveau 1, die Anzahl der Kinder mit den Fähigkeitsniveaus 2 und 3 ist dagegen angewachsen. Es ergeben sich somit beim Leseverständnis signifikante Unterschiede der Fähigkeitsniveauverteilungen zwischen den Jahren.

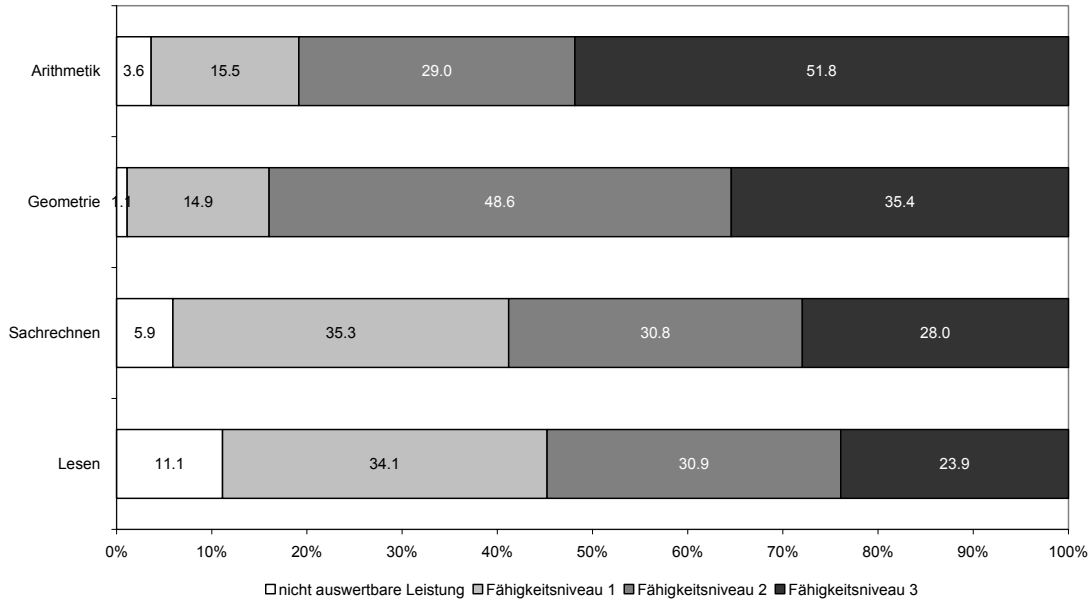

Abbildung 5: *Gesamtverteilung der Fähigkeitsniveaus (2004);* *Angaben in Prozent*

4.1.3 Unterschiede innerhalb und zwischen den Klassen

Leistungsunterschiede zwischen Schülern gehen nicht ausschließlich auf Merkmale zurück, welche mit dem Schüler als einzelner Person verknüpft sind (z.B. Geschlecht oder Erstsprache), sondern sind zu einem mehr oder minder großen Anteil auf die Zugehörigkeit zu einer bestimmten Schulklasse und einer bestimmten Schule zurückzuführen. Wissenschaftlich gesprochen, lässt sich die Leistungsvarianz also in Anteile der Individual-, der Klassen- und der Schulebene zerlegen. Ein immer wiederkehrender empirischer Befund aus Schulleistungsstudien ist, dass schulische Leistungsunterschiede zu einem überwiegenden Anteil auf interindividuelle Differenzen zurückgeführt werden können (vgl. Helmke & Weinert, 1997). Dies stellt eine teilweise, aber keineswegs weitgehende Relativierung der Bedeutung von Schule dar, da Unterricht zwar einen begrenzten, durch die Zusammenfassung in Klassen- und Schulverbänden jedoch weit streuenden Effekt hat: Von ungünstigen Unterrichts- und Kontextbedingungen ist jeweils nicht einer, sondern sind eine Reihe von Schülern betroffen.

Zerlegt man die (interindividuelle) Leistungsvarianz der Schüler, welche an den Vergleichsarbeiten teilgenommen haben, so resultieren die in Tabelle 12 dargestellten Prozentanteile.

Wie erwartet ist die Leistung in den Vergleichsarbeiten vorrangig mit Merkmalen der Individualebene verknüpft. Die Varianzanteile liegen zwischen 54,9 Prozent (Geometrie) und 71,0 Prozent (Sachrechnen). Ebenfalls erhebliche Anteile gehen auf schulische Merkmale zurück, sie bewegen sich zwischen 20,5 Prozent (Sachrechnen) und 33,4 Prozent (Geometrie). Den vergleichsweise geringsten Beitrag leistet die Zugehörigkeit zu einer Schulklasse (hinter der Unterschiede des Unterrichts und des Klassenkontextes stehen) mit 8,5 Prozent (Sachrechnen) bis 11,7 Prozent (Geometrie).

Tabelle 12: *Zerlegung in die Varianz auf den drei Ebenen Schule, Klasse und Individuum*

Bereich	Schulebene	Klassenebene	Individualebene
Arithmetik	25,0	9,7	65,3
Geometrie	33,4	11,7	54,9
Sachrechnen	20,5	8,5	71,0
Lesen	26,8	9,4	63,8
N min	90	251	4973

Nimmt man die Effekte der Schul- und Klassenebene zusammen, resultieren durchaus erhebliche Einflüsse schulischer Qualitätsmerkmale. Zudem können Effekte der Individualebene durch Merkmale der Klassen- bzw. Schulebene moderiert werden, der Einfluss des sozioökonomischen Hintergrunds kann beispielsweise von Schule zu Schule und von Klasse zu Klasse variieren. Andererseits sind Unterschiede zwischen Schulen und Klassen nicht etwa unabhängig von individuellen Faktoren, sondern sind zum Teil auf den Einfluss aggregierter Individualvariablen (z.B. den mittleren sozialen Status der Schülerschaft) zurückzuführen. Zusammenfassend darf der hohe Varianzanteil auf Individualebene nicht dazu verleiten, Unterricht und Schule für nebensächlich oder gar unbedeutend zu halten. Zum einen beeinflussen schulische Lernumgebungen nicht nur den einzelnen Schüler, sondern jeweils gesamte Schul- und Klassenverbände. Damit sind auch kleine Effekte bedeutsam, da sie immer eine größere Anzahl an Schülern betreffen. Zum anderen sollte die Wirkung von Schule nicht ausschließlich mit Blick auf Leistungsunterschiede beurteilt werden: Ohne Unterricht in Schulen erscheint der Aufbau persönlich und gesellschaftlich unentbehrlichen Wissens und vielfältiger kognitiver Fertigkeiten nahezu unmöglich (vgl. Helmke, Hosenfeld & Schrader, S. 420f.).

Der verhältnismäßig umfangreiche Varianzanteil der Schulebene, verglichen mit der Klasse, ist auf den ersten Blick überraschend und widerspricht den Ergebnissen anderer Studien, z.B. MARKUS (Hosenfeld, Helmke, Ridder & Schrader, 2001). Er erklärt sich einerseits aus der Möglichkeit, einen Teil der zu bearbeitenden Aufgaben selbst auszuwählen. Die gemeinsame Auswahl der Aufgaben begünstigt infolge der notwendigen Abstimmung im Kollegium eine Leistungshomogenisierung *innerhalb* der Schulen, während Unterschiede *zwischen* den Schulen durch unterschiedliche Vorgehensweisen bei der Auswahl akzentuiert werden können.

Andererseits reflektiert der hohe Varianzanteil auf Schulebene auch Unterschiede in den Kontextbedingungen (insbes. Einzugsgebiet, soziotopisches Profil) der Schulen. So finden sich *innerhalb* der Schülerschaft einer Schule oft keine allzu ausgeprägten Differenzen bezüglich Sozialschicht, Erwerbstätigkeit der Eltern usw., während diese *zwischen* den Schulen als Folge unterschiedlicher Einzugsgebiete erheblich variieren können.

Auf Klassen- und Schulebene fällt der relativ große (und auf Individualebene entsprechend kleine) Varianzanteil im Bereich Geometrie ins Auge. Dieser Effekt der Klassen- bzw. Schulzugehörigkeit könnte die curriculare „Stiefkindrolle“ der Geometrie widerspiegeln (vgl. Blum et al., 2004, S. 66): Da dieser Bereich in den Lehrplänen in der

Vergangenheit traditionell eine eher untergeordnete Rolle gespielt hat (im Gegensatz insbesondere zur Arithmetik), oblagen Entscheidungen zu Umfang und Art der Behandlung dieses Stoffgebiets verstärkt den Lehrkräften selbst. Diese relativ großen Handlungsspielräume könnten sich im Sinne einer Verstärkung von Leistungsunterschieden im Bereich Geometrie auswirken. Dass beträchtliche Unterschiede im Geometrieleistungsniveau gerade auch auf *Schul*ebene zu finden sind, könnte ein Indikator dafür sein, dass Entscheidungen zu Umsetzung und Ausgestaltung der Lehrpläne nicht nur Sache der einzelnen Lehrkraft sind, sondern auch im Rahmen der Organisationseinheit Schule getroffen werden.

4.1.4 Leistungen von Mädchen und Jungen

Das Geschlecht von Schülerinnen und Schülern ist ein weiterer schulleistungsrelevanter Bedingungsfaktor, welcher sich auf gut gesicherte Erkenntnisse über Unterschiede im kognitiven Bereich bezieht. So wäre etwa die Leistungsüberlegenheit von Jungen im räumlichen Denken und die von Mädchen im sprachlichen Bereich zu nennen.

Im Fach Mathematik wurden in der Regel etwas bessere Leistungen der Jungen nachgewiesen, während Mädchen in einschlägigen Studien im Leseverständnis bessere Werte aufweisen (vgl. Zimmer, Burba & Rost, 2004; Hosenfeld, Helmke, Ridder & Schrader, 2002). Obwohl diese Unterschiede in der Regel stabil sind, können sie dessen ungeachtet als marginal eingestuft werden.

In Tabelle 13 sind die Geschlechterunterschiede in den jeweiligen Inhaltsbereichen dargestellt. Als Maß für die Bedeutsamkeit eines Unterschieds gilt die Effektstärke d*, bei der die Unterschiede zwischen den Gruppen auf die Streuung der Testwerte standardisiert werden. Ein positiver d-Wert bedeutet eine Überlegenheit der Mädchen, ein negativer d-Wert umgekehrt eine Überlegenheit der Jungen.

Bei Betrachtung der geringen Effektstärken wird ersichtlich, dass die Leistungsunterschiede zwischen den bei VERA teilnehmenden Schülerinnen und Schülern erwartungsgemäß eher gering sind.

In Mathematik zeigt sich ein bedeutsamer Vorsprung (d = -0,30) der Jungen nur im Bereich Sachrechnen - dort unterscheiden sich Mädchen und Jungen im höchsten Fähigkeitsniveau um einen Anteil von 10,1 Prozentpunkten. In Arithmetik und im Lesever-

* Als Faustregel gelten in der experimentellen Forschung Werte für d um 0,2 als kleine, um 0,5 als mittlere und um 0,8 als große Effektstärken. Im Kontext nicht-experimenteller pädagogisch-psychologischer Forschung sind auch kleinere Effekte beachtenswert und interpretationswürdig (vgl. Ditton, 1990). Da allerdings die jeweilige Forschungslage zu berücksichtigen ist, dürfen die angegebenen Werte nicht dogmatisch als absolute Grundlage der Bewertung aufgefasst werden. Effektstärkemaße werden unter anderem deshalb verwendet, weil Aussagen über die Signifikanz eines Effekts u.a. von der Stichprobengröße abhängen (bei großen Stichproben werden schon sehr kleine Effekte statistisch signifikant). Die Effektstärke ist dagegen weitgehend unabhängig von der Stichprobengröße.

ständnis zeigen sich immerhin noch geringe Unterschiede in den Leistungsverteilungen der Jungen und Mädchen (d = -0,14 bzw. 0,14). Während in Arithmetik wiederum die Jungen etwas besser abschneiden, zeigt sich im Leseverständnis ein Vorsprung der Mädchen.

Für die Geometrieleistung hingegen spielt das Geschlecht kaum noch eine Rolle.

Tabelle 13: *Verteilung der Fähigkeitsniveaus Mathematik, getrennt nach Geschlecht; Angaben in Prozent*

		n.a.L.*	FN1	FN2	FN3	N (Kinder)	d**
Arithmetik	Mädchen	1,1	15,2	34,9	48,8	2435	-0,14
	Jungen	1,0	12,2	32,4	54,4	2565	
Geometrie	Mädchen	1,9	20,9	39,1	38,2	2435	-0,07
	Jungen	2,0	18,4	38,0	41,7	2565	
Sachrechnen	Mädchen	1,9	45,1	31,3	21,7	2435	-0,30
	Jungen	0,8	33,6	33,8	31,8	2565	
Lesen	Mädchen	5,5	24,7	36,3	33,6	2438	0,14
	Jungen	7,3	28,2	36,7	27,8	2536	

* nicht auswertbare Leistung
** Maß für die Effektstärke

4.1.5 Migrationshintergrund

Die Sprachbeherrschung hängt vermutlich stärker mit der vorherrschenden Familiensprache zusammen als mit dem Geburtsort des jeweiligen „nicht-deutschen" Elternteils. Anhand der Unterscheidung in „Deutsch dominant" vs. „Deutsch nicht-dominant" (N = 3986 bzw. 988) wird bei VERA der Sprachherkunft Rechnung getragen. Dabei entspricht „Deutsch nicht-dominant" zweisprachigen Schülerinnen und Schülern, bei denen - unabhängig von Nationalität und Geburtsort – Deutsch nicht die hauptsächlich gesprochene Sprache ist (vgl. Helmke & Reich, 2001). Mit dieser Unterscheidung soll dem Sachverhalt Rechnung getragen werden, dass ein Teil der Schülerschaft zwar in Deutschland geboren ist, aber nicht in erster Linie Deutsch spricht bzw. nicht in Deutschland geboren ist, jedoch hauptsächlich Deutsch spricht.

In Abbildung 6 und Abbildung 7 sind die prozentualen Schülerleistungen jeweils nach Deutsch als dominante und nicht-dominante Sprache dargestellt. Tabelle 14 zeigt die Effektstärken der Leistungsunterschiede.

Es zeigen sich bedeutsame Unterschiede zwischen Kindern mit Deutsch als dominanter und Kindern mit Deutsch als nicht-dominanter Sprache sowohl in Mathematik (d = 0,50 bis 0,62) als auch in Deutsch (d = 0,75). Während Bremer Schülerinnen und Schüler mit Deutsch als nicht-dominante Sprache in Arithmetik und Geometrie verhältnismäßig gut abschneiden, erreichen im Leseverständnis nur wenige Kinder das Fähigkeitsniveau 3

(10,5 Prozent). Im Leseverständnis und Sachrechnen fällt vor allem der hohe Anteil von Schülerinnen und Schülern auf, die ein Niveau erreichen, das höchstens das Beherrschen elementarer Aufgaben umfasst (42,8 Prozent in Leseverständnis und 60,4 Prozent in Sachrechnen).

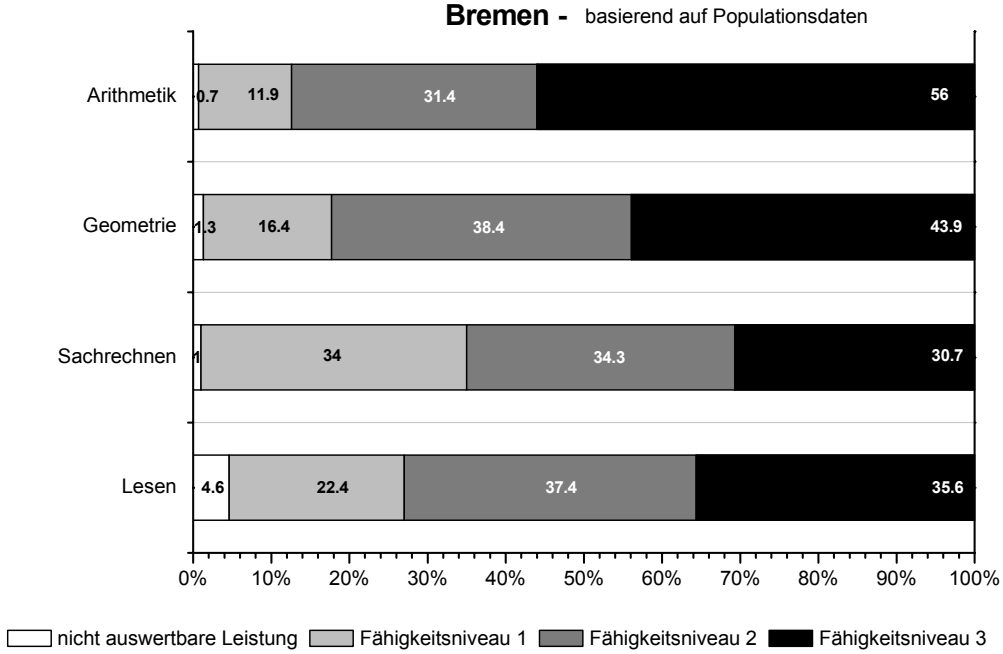

Abbildung 6: *Gesamtverteilung der Fähigkeitsniveaus für Deutsch als dominante Sprache; Angaben in Prozent*

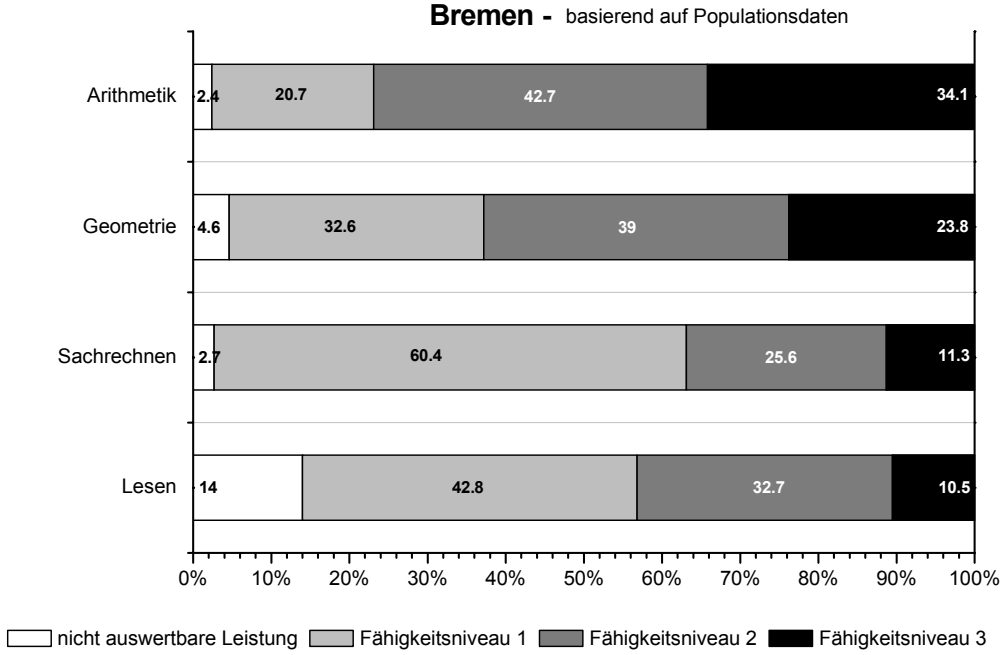

Abbildung 7: *Gesamtverteilung der Fähigkeitsniveaus für Deutsch als nicht-dominante Sprache; Angaben in Prozent*

Insbesondere im Leseverständnis kann infolge des hohen Anteils von Kindern mit nicht auswertbarer Leistung (14 Prozent) von einem substanziellen Unterschied zu Kindern, deren dominante Sprache Deutsch ist, gesprochen werden.

Tabelle 14: *Effektstärken der Leistungsunterschiede von Deutsch dominante vs. nicht-dominante Sprache*

	Arithmetik	Geometrie	Sachrechnen	Lesen	N min (Kinder)
Effektstärke (d)	0,50	0,52	0,62	0,74	3986

Die Ergebnisse bestätigen die Vermutung, dass Merkmale der Sprachherkunft für Schülerinnen und Schüler bereits in der Klassenstufe vier mit erheblichen Leistungsunterschieden gekoppelt sind (vgl. Schwippert, Bos & Lankes, 2003).

In Tabelle 15 ist die geschlechtsspezifische Verteilung von Kindern mit Deutsch als nicht-dominanter Sprache dargestellt. Es zeigen sich in gleicher Weise die bereits in Kapitel 4.1.4 aufgeführten geringen Leistungsunterschiede zwischen Mädchen und Jungen. Insbesondere in Sachrechnen fällt die etwas größere Geschlechterdifferenz (d = -0,38) auf, während im Lesen der Vorsprung der Mädchen geringfügig kleiner ist als in der Gesamtbetrachtung (d = 0,12).

Tabelle 15: *Verteilung der Fähigkeitsniveaus für Deutsch als nicht-dominante Sprache, getrennt nach Geschlecht;* *Angaben in Prozent*

		n.a.L.*	FN1	FN2	FN3	N (Kinder)	d**
Arithmetik	Mädchen	2.8	23.4	41.6	32.2	457	-0.17
	Jungen	2.1	18.3	43.7	35.9	524	
Geometrie	Mädchen	4.2	35.0	39.8	21.0	457	-0.05
	Jungen	5.0	30.5	38.4	26.1	524	
Sachrechnen	Mädchen	3.5	67.6	21.0	7.9	457	-0.38
	Jungen	1.9	54.2	29.6	14.3	524	
Lesen	Mädchen	12.9	40.3	35.1	11.8	459	0.12
	Jungen	14.9	45.0	30.6	9.5	529	

* nicht auswertbare Leistung

** Maß für die Effektstärke

4.1.6 Bremen vs. Bremerhaven

Im Folgenden sollen die Ergebnisse der Kinder aus Bremen mit denen aus Bremerhaven verglichen werden. Die Zuordnung zu Bremen bzw. Bremerhaven erfolgte an Hand der Schulnummer.

Tabelle 16 zeigt, dass die Schülerinnen und Schüler in Bremen ein etwas höheres Fähigkeitsniveau erreichen als die in Bremerhaven. Besonders deutlich wird dies z.B. in

Sachrechnen. Hier erreichen 28,7 Prozent der Schülerinnen und Schüler in Bremen das höchste Fähigkeitsniveau, in Bremerhaven sind dies nur 17,2 Prozent. Dagegen gibt es in Bremerhaven mehr Kinder, die nur das Fähigkeitsniveau 1 erreichen. Dies führt zu einer Effektstärke von d = 0,65 (positive d-Werte stehen für günstigere Werte der Schülerschaft in Bremen). Hier kann man von einem großen Effekt ausgehen. In den Inhaltsbereichen Arithmetik und Geometrie des Bereiches Mathematik ergibt sich das gleiche Bild: Die Schülerinnen und Schüler in Bremen erreichen ein höheres Fähigkeitsniveau.

Tabelle 16: *Verteilung der Fähigkeitsniveaus Mathematik und Deutsch, Bremen vs. Bremerhaven; Angaben in Prozent*

		n.a.L.*	FN 1	FN 2	FN 3	N (Klassen)	d**
Arithmetik	Bremen	1,1	12,7	33,4	52,8	206	
	Bremerhaven	1,3	20,0	35,3	43,4	46	0,48
Geometrie	Bremen	2,0	18,7	37,7	41,6	206	
	Bremerhaven	1,9	24,0	44,2	29,9	46	0,38
Sachrechnen	Bremen	1,1	37,4	32,8	28,7	206	
	Bremerhaven	2,1	50,7	30,1	17,2	46	0,65
Lesen	Bremen	6,2	25,5	36,7	31,6	206	
	Bremerhaven	8,8	33,4	36,4	21,4	46	0,45

* nicht auswertbare Leistung
** Maß für die Effektstärke

Ein vergleichbares Muster findet sich im Bereich Leseverständnis. Auch hier kann eine bessere Leistung in Bremen festgestellt werden.

Im Ganzen kann damit ein höheres Fähigkeitsniveau in Bremen festgestellt werden. Dies kann sich eventuell durch die Bevölkerungszusammensetzung ergeben. In Bremerhaven ist der Anteil an arbeitslosen Eltern und an Eltern, die Sozialleistungen erhalten, größer als in Bremen (siehe 4.2.1, S. 40). Diese Faktoren werden bei der Bestimmung des „fairen Vergleichs“ berücksichtigt. Die Auswirkungen der entsprechenden Faktoren werden im Kapitel 4.2, S. 39 näher beleuchtet.

4.2 „Fairer Vergleich“

Die großen Surveys der letzten Jahre, insbesondere PISA und IGLU/PIRLS, haben gezeigt, dass Merkmale des sozialen, ökonomischen und kulturellen Kapitals von Familien einen überwiegenden Einfluss auf die Leistungsfähigkeit der Kinder ausüben. Auf der Ebene von Klassen und Schulen entspricht dies einer wichtigen Rolle des Schuleinzugsgebietes und der Klassenzusammensetzung. Von Schulen „im sozialen Brennpunkt“ spricht man – obwohl es keine verbindliche Definitionen gibt – gemeinhin dann, wenn verschiedene unterrichts- und lernerschwerende Faktoren in konzentrierter Form auftreten, etwa bei Schulen, deren Klientel durch stark überdurchschnittliche pro-

zentuale Anteile mit Migrationshintergrund, geringer Bildungsnähe, soziale Unterschicht, Arbeitslosigkeit und Erhalt von Sozialhilfe gekennzeichnet ist.

Anders als Lernstandserhebungen und Forschungsprojekte vom Typ IGLU, MARKUS oder PISA wurden an dieser Stelle der VERA- Erhebung mit einem Lehrerfragebogen Angaben zur Klassenzusammensetzung und zum Einzugsgebiet der Schule in erster Linie zu dem Zweck erfasst, um für den „fairen Vergleich" eine fundierte Datenbasis zu erzeugen. Alle Daten beruhen demnach auf Lehrerangaben und entsprechen nicht notwendigerweise den amtlichen Schulstatistiken. Aus diesem Grund sind die folgenden Ergebnisse keinesfalls als systematische Analyse kontextueller Bedingungen schulischer Leistungen zu verstehen.

4.2.1 Beschreibung ausgewählter Kontextmerkmale

Es werden zunächst die auf Klassenebene (Ausnahme: Sprachdominanz) erfragten Daten zum sozioökonomischen und Sprachhintergrund berichtet (Tabelle 17). Anschließend folgt mit Tabelle 18 eine Darstellung der Zusammenhänge zwischen Kontextmerkmalen und Schülerleistungen. In Tabelle 19 sind die Schülervariablen dargestellt.

In Bremen wurde eine Aufteilung in Klassen aus Bremen und Bremerhaven vorgenommen. In Bremen beträgt der Anteil von Schülerinnen und Schülern, für die, unabhängig von Nationalität und Geburtsort, Deutsch nicht die dominante Sprache ist, 19 Prozent, in Bremerhaven 25,1 Prozent. Dabei ergibt sich in Gesamt-Bremen eine Korrelation mit der Schülerleistung im Leseverständnis von r = -,36.

Tabelle 17: *Kontextvariablen (Vergleich Bremen – Gesamtstichprobe)*; durchschnittliche Klassenanteile in Prozent*

	Deutsch nicht- dominante Sprache[1,a)]	Schule liegt in sozialem Brennpunkt[2)]	Familie gehört zur Grundschicht[2,b)]	Familie von Arbeitslosigkeit betroffen[2)]	Familie bezieht Sozialhilfe[2)]
Bremen	19,0	52,6	38,7	22,2	23,0
Bremerhaven	25,1	64,1	44,4	30,8	29,5
alle Länder	10,6	32,2	27,0	15,9	14,6

* Angaben beruhen auf Lehrerangaben und entsprechen nicht notwendigerweise den amtlichen Schulstatistiken

1) erfragt auf Individualebene in der Population

2) erfragt auf Klassenebene in der Zentralstichprobe, in Bremen in allen Klassen

a) Wortlaut: „Kinder, für die, unabhängig von Nationalität und Geburtsort, Deutsch die nicht-dominante Sprache darstellt."

b) Wortlaut: „sog. Unterschicht oder Grundschicht: un- und angelernte Arbeiter, Landarbeiter, alle un- und angelernten Berufe aus dem manuellen Bereich sowie Dienstleistungstätigkeiten mit weitgehend manuellem Charakter und geringem Anforderungsniveau."

Über 50 Prozent aller beteiligten Klassen in Bremen und knapp über 64 Prozent in Bremerhaven liegen im so genannten „sozialen Brennpunkt". Den Lehrerangaben zufolge sind in Bremerhaven etwa 30 Prozent der Familien ihrer Schülerinnen und Schüler von Arbeitslosigkeit betroffen oder beziehen Sozialhilfe.

Tabelle 18: *Zusammenhänge zwischen Kontextmerkmalen und Schülerleistungen (Schulklassenebene)*

	Arithmetik	Geometrie	Sachrechnen	Lesen	min. N
Prozentsatz Grundschicht	-0,27	-0,29	-0,34	-0,31	227
Prozentsatz Erhalt von Sozialhilfe	-0,32	-0,24	-0,38	-0,39	227
Prozentsatz Arbeitslosigkeit in der Familie	-0,34	-0,24	-0,41	-0,36	229
Prozentsatz deutsch nichtdominant	-0,26	-0,25	-0,38	-0,36	252

Bei den Zusammenhängen zwischen Kontextmerkmalen und Schülerleistungen wurden insgesamt niedrige bis mittlere Korrelationen gefunden (zwischen r = -,24 und r = -,41), wobei die höchsten im Sachrechnen und Leseverständnis auftreten.

Der Anteil der Klassenwiederholer liegt in Bremen laut Lehrerangaben bei 6,1 Prozent, in Bremerhaven bei 7,6 Prozent. Erwähnenswert ist die hohe Abwesenheitsquote am Tag der Durchführung in Bremerhaven (in Deutsch und Mathematik 7,5 Prozent), die vor allem in Mathematik um 6,5 Prozent höher liegt als in Bremen.

Tabelle 19: *Schülervariablen (Vergleich Bremen – Bremerhaven – Gesamtstichprobe)*; durchschnittliche Klassenanteile in Prozent*

	ungenügende Sprachbeherrschung[1,a)]	nicht anwesend beim Deutschtest[1)]	nicht anwesend beim Mathetest[1)]	Klassenwiederholer[1)]
Bremen	2,0	5,5	1,0	6,1
Bremerhaven	3,8	7,5	7,5	7,6
alle Länder	0,9	6,4	5,9	2,4

* Angaben beruhen auf Lehrerangaben und entsprechen nicht notwendigerweise den amtlichen Schulstatistiken

1) erfragt auf Individualebene in der Population

a) Wortlaut: „Kinder, die vor weniger als 12 Monaten nach Deutschland eingewandert sind und die deutsche Sprache noch nicht hinreichend beherrschen."

4.2.2 Bildung der Kontextgruppen

Um den beteiligten Lehrkräften einen „fairen" Vergleich anbieten zu können, indem Kontextunterschiede zwischen Klassen berücksichtigt werden, wurden drei landesspezifische Gruppen gebildet und mit Blick auf die Kontextmerkmale und die Leistungen beschrieben. Jede Lehrkraft hatte die Möglichkeit, auf der Grundlage einer Einschätzung der Zusammensetzung ihrer Klasse eine Zuordnung zu einer der drei definierten Kontextgruppen und somit einen „fairen" Vergleichsmaßstab für die in der Klasse erzielten Leistungen zu erhalten. Die Bildung dieser Kontextgruppen erfolgte mit Hilfe eines regressionsanalytischen Mehrebenen-Ansatzes. Dieses Vorgehen stellt im Vergleich zum 2004 gewählten Ansatz eine Veränderung in drei Details dar: Berück-

sichtigung der hierarchischen Struktur der Daten, länderspezifische Analysen und Differenzierung in drei Kontextgruppen (je Land), die so gewonnen Kontextgruppen lassen sich daher nicht auf die Daten des Jahres 2004 beziehen.

Konkret wurde zunächst ein Gesamtleistungswert ermittelt, in den die beiden Fächer Mathematik und Deutsch jeweils zu 50 Prozent eingingen, d. h. zunächst wurden die drei mathematischen Inhaltsbereiche gemittelt und dieser Wert wurde dann gleichgewichtig mit dem Leseverständnis (als einzigem verfügbaren Indikator im Fach Deutsch) zu einem Gesamtwert verrechnet. Im Anschluss daran wurden per Mehrebenenanalyse diejenigen Merkmale ermittelt, die mit diesem Gesamtleistungsindex im Zusammenhang stehen. In diese Analyse aufgenommen wurden auf Schülerebene die Merkmale Geschlecht, Deutsch als nicht-dominante Sprache, Wiederholung der vierten Klasse, Teilleistungsstörung Mathematik, Teilleistungsstörung Deutsch und kombinierte Teilleistungsstörungen in Mathematik und Deutsch. Auf der Ebene der Klassen wurden zwei verschiedene Blöcke von Informationen genutzt. Zum einen sind dies die auf Klassenebene aggregierten Schülermerkmale: die Geschlechterverteilung (Jungenanteil), der Anteil von Klassenwiederholern, der Anteil von Kindern mit Teilleistungsstörungen (getrennt für Mathematik und Deutsch), der Anteil von Kindern mit Deutsch als nicht-dominanter Sprache, der Anteil von Kindern mit sonderpädagogischem Förderbedarf[6] und der Anteil von Kindern mit ungenügender Sprachbeherrschung[7]. Zum anderen sind dies die von den Lehrkräften im Lehrerfragebogen angegebenen Informationen zur Charakterisierung der Klasse: der Anteil der Grundschichtfamilien, der Anteil von Familien betroffen von Arbeitslosigkeit, der Anteil von Familien mit Bezug von Sozialleistungen.

In einem ersten Modell ohne Prädiktoren wurde die Verteilung der Varianz auf die beiden untersuchten Ebenen (Individuen; Klassen) ermittelt. Gut 63 Prozent der Leistungsvarianz liegt innerhalb der Klassen, etwa 37 Prozent sind mit der Klassenzugehörigkeit assoziiert. In der Folge wurde eine Serie von Analysen durchgeführt, um über schrittweise Erweiterungen des Modells die signifikanten Prädiktoren zu ermitteln.

Für Bremen erwiesen sich folgende Variablen als signifikante Prädiktoren: Deutsch als nicht dominante Sprache, Teilleistungsstörung Mathematik, Teilleistungsstörung Deutsch, kombinierte Teilleistungsstörung Deutsch *und* Mathematik, Klassenwiederholer; auf Ebene der Klasse: Anteil der Grundschichtfamilien, Schule im sozialen Brennpunkt und der *Anteil* von Kindern mit Teilleistungsstörungen in Deutsch. Kinder mit Teilleistungsstörungen in Mathematik erzielen niedrigere Leistungen als Kinder ohne Teilleistungsstörungen. Gleiches gilt auch für Kinder mit Teilleistungsstörungen

[6] Kinder mit sonderpädagogischem Förderbedarf nahmen nach Ermessen der unterrichtenden Lehrkraft an den Vergleichsarbeiten teil. Wenn entsprechende Kinder Aufgaben der Vergleichsarbeiten bearbeiteten, wurden auch Rückmeldungen für diese Kinder erstellt, ihre Leistungen gingen jedoch keinesfalls in die Leistungswerte der Klasse ein.

[7] Unter die Definition „ungenügende Sprachbeherrschung“ fallen bei VERA nur solche Kinder, die weniger als sechs Monate in Deutschland zur Schule gehen und daher die deutsche Sprache noch nicht ausreichend beherrschen. Für die Rückmeldung und Wertung der Leistungen dieser Kinder gilt Analoges wie bei den Kindern mit sonderpädagogischem Förderbedarf.

in Deutsch. Kinder mit Teilleistungsstörungen in Mathematik *und* Deutsch sind ebenfalls leistungsschwächer als Kinder ohne Teilleistungsstörungen, jedoch nicht so leistungsschwach wie es sich aus der reinen Addition der Leistungsrückstände der beiden Formen von Teilleistungsstörungen ergeben würde. Ein ähnlicher Effekt ergibt sich auf der Ebene der Klasse. Der Einfluss des Klassenanteils mit Teilleistungsstörungen in Deutsch ist als Korrekturfaktor anzusehen: Klassen mit hohen Anteilen an Kindern mit Teilleistungsstörungen im Fach Deutsch schneiden *besser* ab als es auf Grund der reinen addierten Leistungsrückstände der einzelnen Schüler mit Teilleistungsstörung zu erwarten wäre.

Insgesamt können mit diesem Modell gut 14 Prozent der Leistungsvarianz auf Individualebene und gut 22 Prozent der Variation zwischen Klassen erklärt werden. Erwartungsgemäß fällt der Anteil der erklärten Varianz auf Ebene der Schülerinnen und Schüler gering aus. Der erklärte Varianzanteil zwischen Klassen liegt auf dem Niveau wie in fast allen anderen Bundesländern und kann angesichts der verfügbaren Prädiktoren als zufrieden stellend gelten.

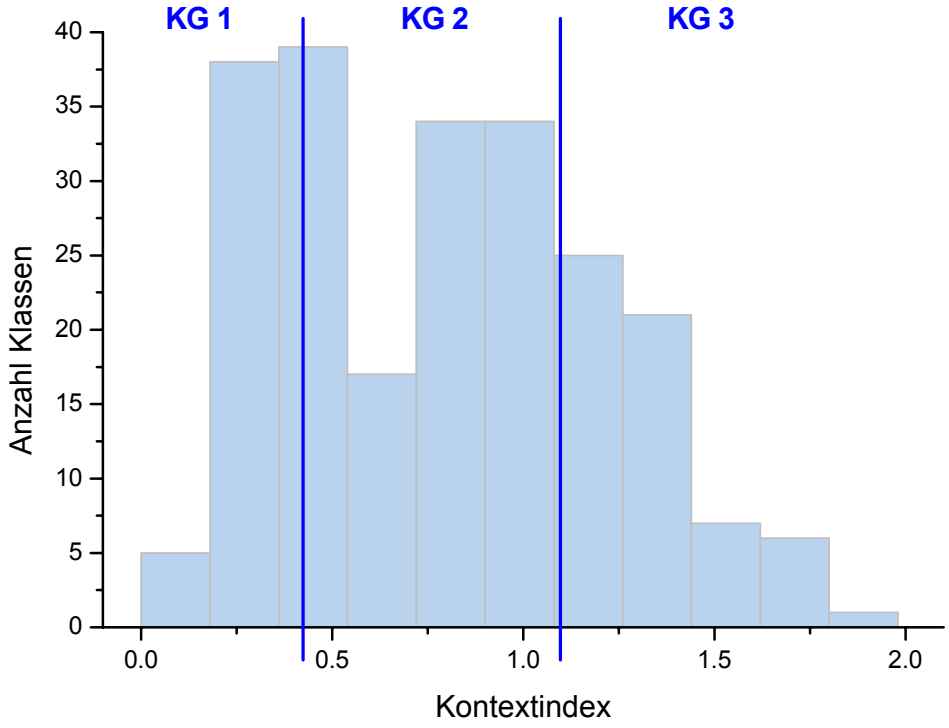

Abbildung 8: *Verteilung der Kontextwerte inkl. der Grenzen zwischen den drei Kontextgruppen (KG 1 = günstige bis KG 3 = ungünstige Kontextbedingungen)*

Im nächsten Schritt wurden diese Informationen gemäß der ermittelten Regressionsgewichte zu einem Kontextwert kombiniert. Abbildung 8 zeigt die Verteilung dieser Werte der Bremer Klassen, für die vollständige Daten vorliegen und die daher berücksichtigt werden konnten (N = 216). Niedrige Werte stehen dabei für günstige, hohe Werte für ungünstige Kontextbedingungen.

Tabelle 20: *Korrelation zwischen der mittleren Klassenleistung und dem Kontextwert in Bremen**

Mathematik	**r**	**Deutsch**	**r**
Arithmetik	-0,36	Leseverständnis	-0,44
Geometrie	-0,34		
Sachrechnen/Größen	-0,49		

* Stichprobengröße: N = 216 (Klassen mit vollständigen Daten)

Die Korrelationen dieses Kontextwertes mit den Leistungen in den verschiedenen Inhaltsbereichen sind in Tabelle 20 dargestellt. Erwartungsgemäß zeigen sich geringere Zusammenhänge zu den Leistungen im Bereich Arithmetik und Geometrie und engere Zusammenhänge mit den stärker sprachgebundenen Bereichen des Leseverständnisses und des Sachrechnens. Insgesamt fallen die Zusammenhänge zufrieden stellend aus.

Anschließend wurden die Klassen in drei Kontextgruppen aufgeteilt, wobei die „günstigste“ Kontextgruppe 1 und die „ungünstigste“ Kontextgruppe 3 jeweils etwa 25 Prozent umfassen, während die mittlere Gruppe ca. 50 Prozent der Klassen umfasst. Durch die kompensatorische Art der Verknüpfung der Einzelinformationen können unterschiedliche Kontextkonfigurationen zur Zuweisung in die gleiche Kontextgruppe führen. Tabelle 21 gibt die Durchschnittswerte zentraler Kontextmerkmale der drei Gruppen wieder.

Tabelle 21: *Beschreibung der drei Kontextgruppen anhand durchschnittlicher Merkmalsausprägungen; Angaben in Prozent*

	Kontextgruppe		
	1	**2**	**3**
Anteil der Kinder mit Deutsch als nicht dominanter Sprache	~3	~7	~46
Anteil der Klassen im sozialen Brennpunkt	~0	~60	~100
Anteil der Kinder aus Familien der Grundschicht	~10	~41	~73

4.2.3 Verteilung der Fähigkeitsniveaus nach Kontextgruppen

Diese so gebildeten Kontextgruppen unterscheiden sich nicht nur mit Blick auf Merkmale der Klassenzusammensetzung, sondern auch hinsichtlich der erzielten Leistungen. In Tabelle 22 sind die Verteilungen auf die Fähigkeitsniveaus in Mathematik und Deutsch nach den drei Kontextgruppen differenziert dargestellt.

Tabelle 22: *Verteilung der Fähigkeitsniveaus, aufgeschlüsselt nach Kontextgruppen; Angaben in Prozent*

	Kontextgruppe	n.a.L.*	FN 1	FN 2	FN 3	N (Klassen)
Arithmetik	1	0,2	8,8	27,7	63,2	54
	2	0,6	13,2	34,5	51,7	108
	3	2,0	19,7	37,9	40,4	54
Geometrie	1	0,8	13,0	34,2	52,0	54
	2	1,7	20,1	39,2	39,0	108
	3	3,2	24,8	42,7	29,3	54
Sachrechnen	1	0,9	26,3	33,5	39,2	54
	2	0,8	37,1	35,6	26,5	108
	3	1,9	55,3	26,9	15,8	54
Lesen	1	2,4	16,9	36,8	43,9	54
	2	5,4	26,9	38,6	29,1	108
	3	11,1	34,8	34,0	20,1	54

* nicht auswertbare Leistung

Es wird deutlich, dass sich die Leistungen in den drei Kontextgruppen erwartungskonform unterscheiden. Der Anteil von Kindern mit fortgeschrittenen Fähigkeiten steigt mit zunehmender Günstigkeit des Kontextes stetig an, während der Anteil von Kindern mit nicht auswertbaren Leistungen nahezu stetig abnimmt. Angesichts der gefundenen Korrelationen zwischen dem gebildeten Kontextindex und den Schülerleistungen überrascht dies nicht.

4.2.4 Vergleich tatsächliche vs. erwartete Leistung

Die folgende Gegenüberstellung der erwarteten Leistung und der tatsächlichen Leistung soll verdeutlichen, in welchem Ausmaß die gemittelte Leistung der Klassen über bzw. unter dem Erwartungswert liegt. Zu diesem Zweck wurden die Klassen, für die entsprechende Daten vorlagen (N = 227), nach dem Kontextindex sortiert, so dass die Klassen mit dem ungünstigsten Kontext ganz links und die Klassen mit dem günstigsten Kontext ganz rechts stehen. Entsprechend steigt der erwartete, gemittelte Leistungswert der Klassen von links nach rechts an. In der Abbildung dargestellt sind die Differenzen zwischen erwartetem und tatsächlichem Leistungswert. Jede Klasse wird durch einen senkrechten Balken dargestellt, je länger der Balken, desto größer die Differenz. Die Werte sind hierbei anhand der tatsächlichen Leistung aller 252 Bremer Klassen z-standardisiert. Liegt die tatsächliche Leistung über der erwarteten, so ist der Balken grün dargestellt und weist vom Erwartungswert aus nach oben, liegt die gezeigte Leistung unter dem Erwartungswert, so ist der Balken rot und geht nach unten. Die dicken blauen Linien kennzeichnen die Grenzen zwischen den drei Kontextgruppen (KG 1, KG 2, KG 3).

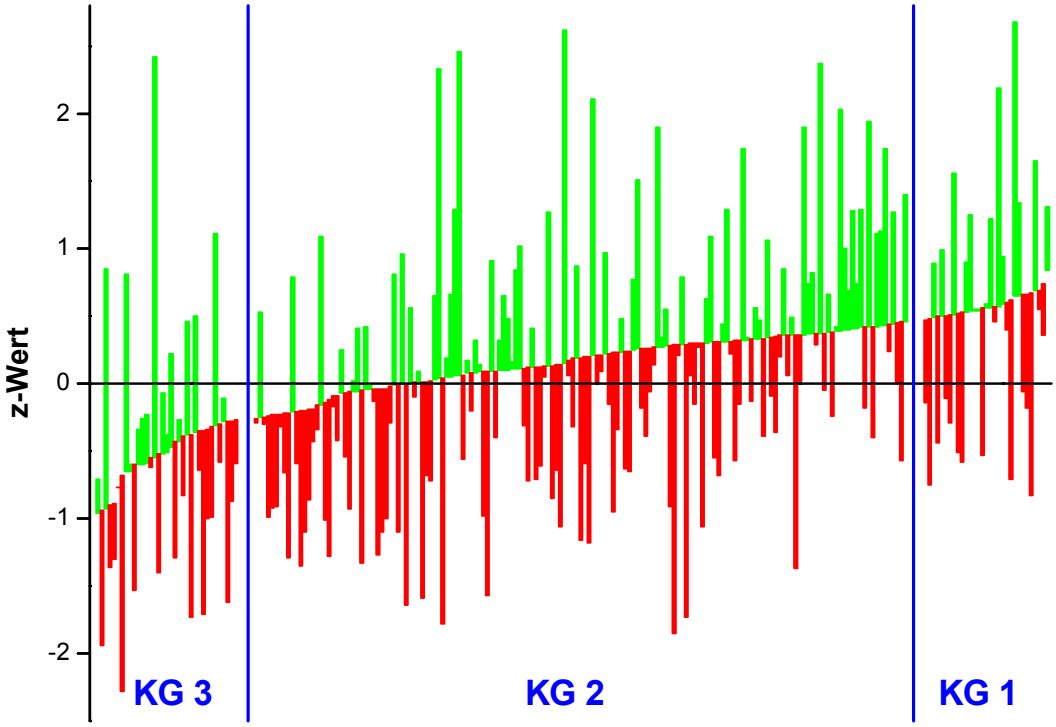

Abbildung 9: *Differenzen zwischen erwarteter und tatsächlicher über beide Fächer gemittelter Leistung (N = 227 Klassen mit vollständigen Kontextdaten); KG = Kontextgruppe*

Auf den ersten Blick wird deutlich, dass es in allen Kontextbereichen deutliche Abweichungen vom Erwartungswert sowohl nach oben als auch nach unten gibt. Dabei liegen diese Abweichungen für etwa die Hälfte der Klassen in einer Zone von -0,67 bis +0,49 Standardabweichungen. Die schwächsten zehn Prozent unterschreiten ihren Erwartungswert jeweils um mindestens 1,1 Standardabweichungen, die leistungsstärksten zehn Prozent erzielen gemittelte Leistungen, die jeweils mindestens 1,0 Standardabweichungen über ihrem Erwartungswert liegen. Das Ausmaß dieser Abweichungen ist in Bremen vergleichbar mit den anderen an VERA teilnehmenden Ländern. Über die Ursachen dieser Abweichungen kann anhand der vorliegenden Daten keine belastbare Aussage getroffen werden. Die bereits berücksichtigten Kontextmerkmale scheiden jedoch als mögliche Ursache der Unterschiede aus.

4.3 Diagnosegenauigkeit

Im Schatten der Frage nach den Fähigkeitsniveaus und dem Leistungsstand der Viertklässler, aber aus pädagogischen Gründen gleichwohl bedeutsam, steht die Frage nach der Diagnosegenauigkeit von Lehrkräften – ein wichtiger Aspekt der diagnostischen Kompetenz. Wenn man die bei VERA vorgenommene Einschätzung der Aufgabenschwierigkeiten bei der Aufgabenwahl (in Mathematik) mit den tatsächlichen Aufgabenschwierigkeiten miteinander in Beziehung setzt, dann ergeben sich zwei interessante Kennwerte der Diagnosegenauigkeit: (a) das Ausmaß der *Unter- oder Überschätzung* der Aufgabenschwierigkeit, d.h. des Leistungsniveaus der Schulklasse und (b) die *Kor-*

relation zwischen beiden Rangreihen, also die Ähnlichkeit der Rangordnung geschätzter vs. realer Aufgabenlösungen.

Man muss diese beiden Aspekte der Diagnosegenauigkeit unbedingt unterscheiden. Eine Lehrkraft kann z.B. alle Aufgaben konstant etwas über- oder unterschätzen, aber gleichwohl kann die Rangordnung der von ihr geschätzten Aufgabenschwierigkeit identisch mit der Rangordnung der gelösten Aufgaben sein - oder umgekehrt. Eine hohe Korrelation der geschätzten mit der empirischen Aufgabenrangreihe sagt etwas über die *fachdidaktische* Kompetenz der Lehrkraft aus, ist am ehesten Ausdruck einer zutreffenden Orientiertheit über Schwierigkeitsunterschiede zwischen Aufgaben. Eine geringe Abweichung der durchschnittlichen geschätzten Aufgabenschwierigkeit von der empirischen Schwierigkeit sagt dagegen eher etwas über die *pädagogisch-psychologische* Diagnostik aus, d.h. wie gut die Lehrkraft über das durchschnittliche Leistungsniveau der Klasse im Bilde ist. Wir sprechen deshalb im Folgenden vereinfacht von der fachdidaktischen und der pädagogischen Komponente der Diagnosegenauigkeit. Die beiden Kennwerte sind nicht nur konzeptuell, sondern auch statistisch fast vollkommen unabhängig voneinander: Die Korrelation beträgt über alle Länder hinweg r = ,08.

Die Frage der Diagnosegenauigkeit zu vertiefen, würde allerdings den Rahmen dieses Berichtes sprengen. Hierzu liegen nicht nur Handreichungen für die an VERA beteiligten Lehrkräfte, sondern auch Publikationen der VERA-Autoren vor[8].

a) Pädagogische Komponente: Unter- vs. Überschätzungstendenz

Wir berichten in der folgenden Abbildung das Ausmaß, in dem die Lehrkräfte die Leistungen ihrer Schülerinnen und Schüler unter- oder überschätzen. Hierzu liegen verwertbare Angaben von 122 Lehrkräften vor.

[8] Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2003). Diagnosekompetenz in Ausbildung und Beruf entwickeln. Karlsruher Pädagogische Beiträge (55), 15-34.

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold & C. Griese (Hrsg.), Schulleitung und Schulentwicklung (S. 119-144). Hohengehren: Schneider-Verlag.

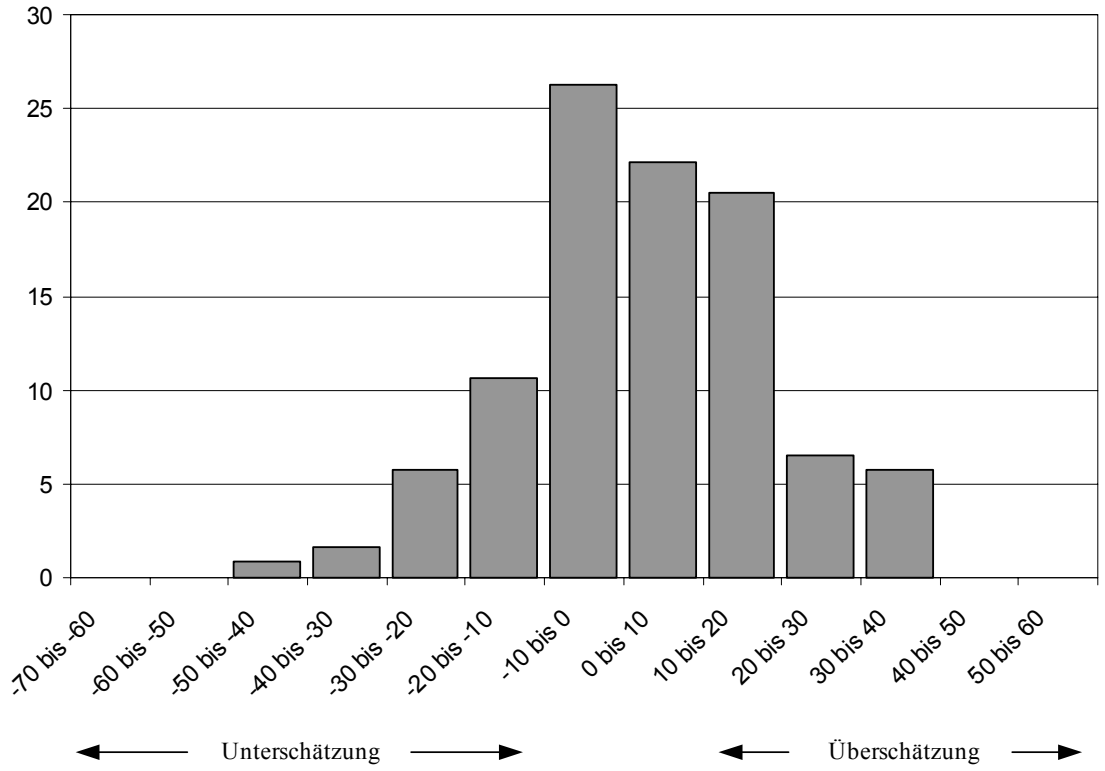

Abbildung 10: *Prozentuale Unter- vs. Überschätzung des mathematischen Leistungsniveaus der Klasse (pädagogisch-psychologischer Aspekt der Diagnosegenauigkeit) durch die Lehrkräfte; Angaben in Prozent*

Die Abbildung zeigt, dass die Lehrkräfte die Leistungen ihrer Klassen etwas häufiger *über*schätzen als *unter*schätzen. Im Durchschnitt wird die tatsächliche Schülerleistung um 2,8 Prozent überschätzt.

Legt man das Ergebnis (die durchschnittliche Einschätzung) unter die Lupe und fragt, in welchem der drei Teilgebiete der Mathematik die geringsten und die größten Überschätzungen stattfinden, dann zeigt sich folgendes: Die Schülerleistungen im Bereich Sachrechnen/Größen werden um 4,2 Prozent *unter*schätzt, die in Arithmetik um 4,6 Prozent und die im Bereich der Geometrie um 9,7 Prozent *über*schätzt. Die Lösungshäufigkeiten der Sachrechen-Aufgaben wurden damit 2005 realistischer eingeschätzt als jene der Arithmetik- und Geometrie-Aufgaben – eine Verkehrung gegenüber den Ergebnissen von 2004. Möglicherweise ist dieser Befund als Wirkung der Berichtlegung von 2004 zu interpretieren, in der auf die deutlichen Fehleinschätzungen im Bereich Sachrechnen aufmerksam gemacht wurde. An Plausibilität gewinnt dieser Schluss, wenn man die ausgeprägte Freiwilligkeit der Teilnahme in 2005 hinzunimmt – es dürfte sich bei den teilnehmenden Lehrkräften vor allem um die besonders interessierten und gut informierten gehandelt haben.

b) Fachdidaktische Komponente: Vergleich der Aufgaben-Rangordnungen

Im Land Bremen liegen verwertbare Angaben von 122 Lehrkräften zur Diagnosegenauigkeit vor. Die durchschnittliche Korrelation liegt bei r = ,37. Wenn man das Auflösungsniveau erhöht, zeigt sich die folgende Verteilung.

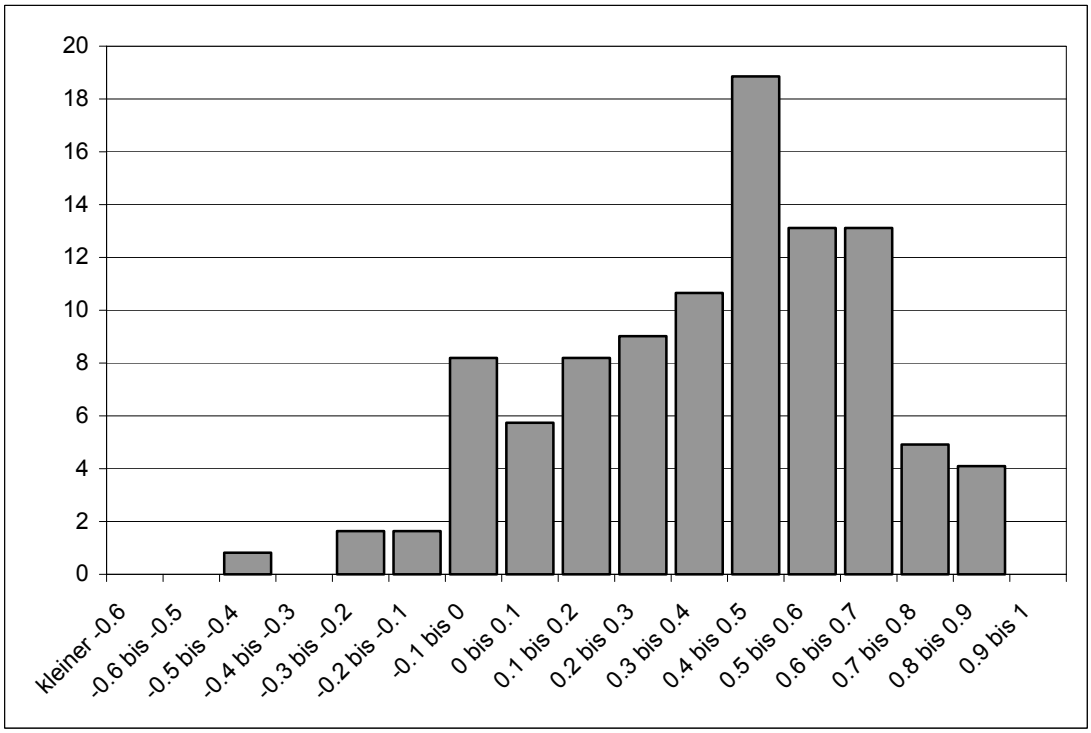

Abbildung 11: *Verteilung der Korrelationskoeffizienten: Ähnlichkeit geschätzter und empirischer Aufgabenschwierigkeiten (fachdidaktischer Aspekt der Diagnosegenauigkeit);* *Angaben in Prozent*

Es zeigt sich ein sehr großes Spektrum, von Korrelationen um Null herum (dies ergibt sich, wenn man nach Zufall antwortet) und sogar einigen wenigen negativen Korrelationen, bis hin zu wenigen Klassen, deren Lehrkräfte eine sehr gute Diagnosegenauigkeit aufweisen (Korrelationen zwischen r = ,80 und ,90). Korrelationen in Höhe von r = ,50 und höher würden wir als akzeptabel, Korrelationen in Höhe von r = ,70 und höher als gut bezeichnen, d.h. ca. neun Prozent der Lehrkräfte verfügen über eine gute oder sehr gute Diagnosegenauigkeit, ca. 26 Prozent verfügen über eine akzeptable Diagnosegenauigkeit, aber ca. 65 Prozent der Lehrkräfte liegen mit ihrer Diagnosegenauigkeit zum Teil deutlich unterhalb akzeptabler Werte.

Die 2005 gegebene Möglichkeit, auf Ebene der einzelnen Teilaufgaben Einschätzungen für die voraussichtliche Häufigkeit der richtigen Lösungen durch Schüler vorzunehmen, reduziert die Gefahr deutlicher Fehleinschätzungen, da auf der Ebene von Teilaufgaben ein feiner abgestimmtes Urteil möglich ist. Dies könnte als Ursache dafür angesehen werden, dass der Grad der Verschätzung (nach oben und nach unten) in diesem Jahr insgesamt geringer ausfällt. Andererseits wird es durch die größere Anzahl an (Teil-)Aufgaben schwieriger, die Häufigkeiten der richtigen Lösung in eine Rangreihe zu bringen. Dies dürfte zumindest zum Teil erklären, warum der Zusammenhang zwi-

schen geschätzter und tatsächlicher Lösungshäufigkeit in diesem Jahr geringer ausfällt als 2004.

4.4 Lehrerfragebogen

Der Lehrerfragebogen sollte in Bremen für 252 Klassen in 91 Schulen bearbeitet werden. An die Vollständigkeit des Rücklaufs des Lehrerfragebogens können unterschiedlich strenge Maßstäbe angelegt werden. An dieser Stelle sollen drei Kriterien berichtet werden:

- *Absolute Vollständigkeit*: Alle im Lehrerfragebogen erfragten numerischen Angaben wurden von der Lehrkraft gemacht. Lediglich die offenen Antwortfelder gehen nicht in die Wertung ein.
- *Relative Vollständigkeit*: Fehlen höchstens zwei der numerischen Angaben, so liegen weitgehend vollständige Angaben vor, denn es werden hier beispielsweise Fälle mit aufgenommen, deren fehlende Werte aus versehentlichem Auslassen o.Ä. resultieren.
- *Vollständigkeit zentraler Angaben*: Die wichtigsten Angaben für die Kontextuierung einer Klasse sind solche zum
  - Anteil an Familien mit Grundschichtzugehörigkeit
  - Anteil an sozialhilfeempfangenden Familien
  - Anteil an Familien mit arbeitslosen Eltern(teilen)
  - Sozialen Brennpunkt.

  In diesem Sinne besteht Vollständigkeit, wenn Angaben zu diesen vier Kriterien gemacht wurden.

Für diese Vollständigkeitsmaße resultieren die in Tabelle 23 dargestellten Werte.

Tabelle 23: *Vollständigkeit des Rücklaufs des Lehrerfragebogens; Angaben in Prozent*

	absolut vollständig	relativ vollständig	4 zentrale Angaben
komplett	79,4	84,5	85,7
unvollständig	20,6	15,5	14,3

Im Folgenden sind die Fähigkeitsniveauverteilungen in Bezug auf ausgewählte Variablen aus dem Lehrerfragebogen dargestellt. Zum Einen werden bestimmte Lehrermerkmale (Kontinuität des Unterrichts, Unterrichtserfahrung, grundständiges Fach) diskutiert, die einen Einfluss auf die Schülerleistungen haben können. Zum Anderen soll genauer beleuchtet werden, inwiefern VERA 2005 zumindest mittelfristig zu einer erhöhten Kooperation zwischen den beteiligten Lehrkräften geführt hat.

4.4.1 Kontinuität des Unterrichts in Mathematik und Deutsch und Fähigkeitsniveaus

Der Vergleich zwischen Klassen aus Schulen, in welchen *kein regulärer* Lehrerwechsel stattfindet, mit Klassen aus Schulen *mit regulärem* Lehrerwechsel nach der zweiten oder dritten Klasse, kann angesichts der teilweise geringen Gruppengrößen (N < 10) nicht zuverlässig interpretiert werden (s. Tabelle 24).

Tabelle 24: *Lehrerwechsel während der ersten vier Grundschuljahre in den Fächern Mathematik und Deutsch; Angaben in Prozent*

Bereich		n.a.L.*	FN 1	FN 2	FN 3	N (Klassen)
Arithmetik	keine Regel	0,7	14,4	35,0	49,9	49
	kein Lehrerwechsel	1,1	13,3	33,8	51,9	156
	Lehrerwechsel nach der 2. Klasse	0,0	16,8	28,9	54,3	9
	Lehrerwechsel nach der 3. Klasse	0,0	20,8	34,4	44,7	5
	eine andere Regel	0,5	22,0	35,3	42,2	8
Geometrie	keine Regel	2,3	22,7	41,4	33,6	49
	kein Lehrerwechsel	1,7	18,2	38,4	41,7	156
	Lehrerwechsel nach der 2. Klasse	0,6	21,9	41,2	36,3	9
	Lehrerwechsel nach der 3. Klasse	0,0	7,3	55,3	37,3	5
	eine andere Regel	4,3	28,3	38,7	28,8	8
Sachrechnen	keine Regel	0,7	40,4	30,0	28,9	49
	kein Lehrerwechsel	1,3	39,3	32,9	26,5	156
	Lehrerwechsel nach der 2. Klasse	0,0	35,9	30,9	33,2	9
	Lehrerwechsel nach der 3. Klasse	0,0	43,1	42,6	14,3	5
	eine andere Regel	2,8	42,2	36,5	18,5	8
Lesen	keine Regel	6,3	26,2	35,5	32,0	49
	kein Lehrerwechsel	6,6	26,9	36,6	29,8	156
	Lehrerwechsel nach der 2. Klasse	1,9	17,0	42,5	38,5	9
	Lehrerwechsel nach der 3. Klasse	2,6	27,1	46,5	23,8	5
	eine andere Regel	9,5	36,4	36,7	17,3	8

* nicht auswertbare Leistung

Betrachtet man die Frage nach der Zahl der durch die einzelnen Lehrkräfte unterrichteten Halbjahre (Tabelle 25), so können keine systematischen Zusammenhänge gefunden werden. Ebenfalls gibt es wieder teilweise geringe Gruppengrößen, weshalb auf eine weiterreichende Interpretation verzichtet werden sollte.

Tabelle 25: *Durch die Lehrkraft unterrichtete Halbjahre in den Fächern Mathematik und Deutsch,* *Angaben in Prozent*

Bereich		n.a.L.*	FN 1	FN 2	FN 3	N (Klassen)
Arithmetik	Beginn 1. Klasse	1	13,5	33,7	51,7	142
	Mitte 1. Klasse	0,0	8,1	45,5	46,3	4
	Beginn 2. Klasse	2,6	13,9	29,4	54,1	8
	Mitte 2. Klasse	0,0	6,8	34,8	58,3	2
	Beginn 3. Klasse	0,6	15,9	30,3	53,2	36
	Mitte 3. Klasse	0,4	14, 8	39,4	45,4	10
	Beginn 4. Klasse	2	18,2	35,7	44,1	23
Geometrie	Beginn 1. Klasse	1,7	18,7	38,5	41,1	142
	Mitte 1. Klasse	0,0	20,6	42,1	37,3	4
	Beginn 2. Klasse	2,5	14,9	42,9	39,8	8
	Mitte 2. Klasse	0,0	10,2	65, 2	24, 7	2
	Beginn 3. Klasse	1,7	20,2	39,1	39	36
	Mitte 3. Klasse	1,7	13,8	46,3	38,2	10
	Beginn 4. Klasse	4,4	30,8	38,2	26,6	23
Sachrechnen	Beginn 1. Klasse	1,2	39,6	31,7	27,5	142
	Mitte 1. Klasse	0,0	33,1	33,4	33,4	4
	Beginn 2. Klasse	0,0	37,2	38,9	24	8
	Mitte 2. Klasse	0,0	28,2	46,1	25,8	2
	Beginn 3. Klasse	1	38,9	32,3	27,6	36
	Mitte 3. Klasse	0,9	35,4	37,6	26,2	10
	Beginn 4. Klasse	2,5	43,8	32,7	21	23
Lesen	Beginn 1. Klasse	6,1	26,8	37,1	30,0	161
	Mitte 1. Klasse	9,0	40,9	30,2	19,9	3
	Beginn 2. Klasse	4,1	16,6	30,2	49,2	9
	Mitte 2. Klasse	12,2	39,8	27,5	20,4	6
	Beginn 3. Klasse	6,4	22,0	38,3	33,3	22
	Mitte 3. Klasse	5,0	28,9	37,9	28,1	9
	Beginn 4. Klasse	9,1	29,9	40,9	20,1	16

* nicht auswertbare Leistung

4.4.2 Unterrichtserfahrung, grundständige Ausbildung und Fähigkeitsniveaus

Aufgrund der verschiedenen in den Tabellen berücksichtigten Merkmalsebenen (z.B. Fach, Inhaltsbereich und entsprechende Kategorien) und der vielfältigen damit zusammenhängenden Determinanten (z.B. bestimmte Kontextmerkmale, wie Grundschichtzugehörigkeit) ist eine Interpretation komplex. Da sich darüber hinaus ähnlich wie bei MARKUS (Helmke, Hosenfeld & Schrader, 2002) keine systematischen, insbesondere keine linearen (vom Typ „je...desto“) Unterschiede in Abhängigkeit von der Unterrichtserfahrung (Tabelle 26) sowie der grundständigen Ausbildung (Tabelle 27) ergeben, werden die entsprechenden Fähigkeitsniveauverteilungen der Übersicht halber im Folgenden wiedergegeben, ohne sie im Detail zu diskutieren.

Tabelle 26: *Durch die Lehrkraft unterrichtete Jahre in den Fächern Mathematik und Deutsch, Angaben in Prozent*

Bereich		n.a.L.*	FN 1	FN 2	FN 3	N (Klassen)
Arithmetik	bis 2 Jahre	0,4	15,3	33,5	50,9	25
	2 bis 5 Jahre	1,4	13,4	33,0	52,2	32
	6 bis 10 Jahre	0,8	10,2	33,3	55,8	32
	11 bis 15 Jahre	1,2	12,7	28,6	57,5	30
	16 bis 20 Jahre	0,3	15,2	28,2	56,3	17
	21 bis 25 Jahre	0,0	18,1	42,5	39,3	11
	26 bis 30 Jahre	0,9	11,0	32,9	55,1	22
	31 bis 35 Jahre	1,8	15,0	34,4	48,8	34
	36 bis 40 Jahre	1,8	22,3	36,9	38,9	19
	mehr als 40 Jahre	0,0	12,0	56,0	32,0	1
Geometrie	bis 2 Jahre	2,1	18,3	44,7	35,0	25
	2 bis 5 Jahre	1,4	17,6	42,5	38,5	32
	6 bis 10 Jahre	2,3	16,8	35,8	45,2	32
	11 bis 15 Jahre	1,8	13,8	38,9	45,5	30
	16 bis 20 Jahre	0,5	17,4	32,3	49,7	17
	21 bis 25 Jahre	1,7	29,0	36,9	32,4	11
	26 bis 30 Jahre	1,3	13,2	41,0	44,6	22
	31 bis 35 Jahre	2,7	24,1	37,0	36,1	34
	36 bis 40 Jahre	3,2	35,6	41,8	19,4	19
	mehr als 40 Jahre	0,0	0,0	52,0	48,0	1

Bereich		n.a.L.*	FN 1	FN 2	FN 3	N (Klassen)
Sachrechnen	bis 2 Jahre	0,7	39,4	34,5	25,3	25
	2 bis 5 Jahre	1,2	42,0	32,7	24,1	32
	6 bis 10 Jahre	0,5	39,5	33,2	26,8	32
	11 bis 15 Jahre	0,8	33,3	33,7	32,2	30
	16 bis 20 Jahre	0,9	34,0	34,6	30,5	17
	21 bis 25 Jahre	1,2	43,0	37,9	17,9	11
	26 bis 30 Jahre	1,2	34,4	30,9	33,4	22
	31 bis 35 Jahre	1,7	42,6	30,3	25,4	34
	36 bis 40 Jahre	2,8	47,3	27,3	22,6	19
	mehr als 40 Jahre	0,0	28,0	36,0	36,0	1
Lesen	bis 2 Jahre	7,4	30,7	34,1	27,8	20
	2 bis 5 Jahre	5,3	26,0	39,5	29,1	31
	6 bis 10 Jahre	6,8	25,9	33,3	33,9	25
	11 bis 15 Jahre	4,7	22,9	40,6	31,8	31
	16 bis 20 Jahre	5,5	22,9	37,1	34,5	28
	21 bis 25 Jahre	10,2	29,9	33,7	26,2	19
	26 bis 30 Jahre	3,3	27,7	37,0	32,0	25
	31 bis 35 Jahre	9,2	28,1	36,8	25,9	31
	36 bis 40 Jahre	6,5	29,2	36,8	27,5	14
	mehr als 40 Jahre	-	-	-	-	0

* nicht auswertbare Leistung

Tabelle 27: *Mathematik und Deutsch als grundständiges Fach, Angaben in Prozent*

Bereich		n.a.L.*	FN 1	FN 2	FN 3	N (Klassen)
Arithmetik	nein	1,3	15,2	33,8	49,7	112
	ja	0,8	13,3	33,0	52,9	112
Geometrie	nein	2,2	20,7	37,9	39,2	112
	ja	1,6	18,7	40,7	39,0	112
Sachrechnen	nein	1,3	40,1	32,2	26,4	112
	ja	1,1	38,7	32,7	27,6	112
Lesen	nein	6,1	27,9	36,2	29,8	77
	ja	6,6	25,9	37,2	30,2	147

* nicht auswertbare Leistung

4.4.3 Vorbereitung auf die Vergleichsarbeiten

Mit Blick auf die im Projekt VERA formulierten Ziele beziehen sich gewünschte Konsequenzen im Wesentlichen darauf, dass durch die zurückgemeldeten Ergebnisse eigene „blinde Flecke“ in Bezug auf die Klasse verdeutlicht und damit möglicherweise Maßnahmen zur Unterrichtsentwicklung angeregt werden. Dabei stellen die Vergleichsarbeiten als externe Evaluation sowohl für Lehrkräfte als auch für Schülerinnen und Schüler i.d.R. eine neuartige und aufregende Situation dar. Vor diesem Hintergrund ist es ein nachvollziehbares Bedürfnis, die eigene Klasse auf die Vergleichsarbeiten so vorzubereiten, dass das bisher Gelernte gezeigt werden kann und Ängste oder Belastungsreaktionen reduziert werden. In diesem Zusammenhang bieten sich unterschiedliche Zugangsweisen an wie z.B.

- *Vermittlung von Vertrautheit mit dem Testverfahren*: Vorbereitung auf den Ablauf, Besprechung von Aufgabenformaten und bei VERA eingesetzten Korrekturkriterien bzw. Anforderungen in den Korrekturanweisungen
- *Inhaltliche Vorbereitung*: Besprechung von VERA 2004-Aufgaben oder von typischen Inhalten der Vergleichsarbeiten
- *Vermittlung von Teststrategien*

Hier ist jedoch zu betonen, dass insbesondere mit der inhaltlichen Vorbereitung keinesfalls ein „teaching to the test“ von z.B. den herunter geladenen aktuellen Testaufgaben gemeint ist. Ein entsprechendes Vor-Üben der VERA-Aufgaben ist in keiner Hinsicht sinnvoll, insbesondere, da die zurückgemeldeten Ergebnisse unter solchen Voraussetzungen allenfalls über die Reproduktionsleistung der Schülerinnen und Schüler Auskunft geben können.

Die folgende Tabelle 28 fasst die Häufigkeiten für die unterschiedlichen Formen der Vorbereitung zusammen.

Tabelle 28: *Formen der Vorbereitung; Angaben in Prozent*

		Häufigkeiten N = 224*	**Prozent**
Vertrautheit mit dem Testformat	Ablauf	173	77,2
	Aufgabenformate	77	34,4
	Korrekturkriterien	5	2,2
Inhaltliche Vorbereitung	Aufgaben 2004	74	33
	VERA-Inhalte	49	21,9
Teststrategien	Teststrategien	120	53,6

* Mehrfachantworten waren bei dieser Frage zugelassen.

Dabei scheint in Bremen der Fokus der Vorbereitung v. a. auf der Vertrautheit mit dem Testformat gelegen zu haben. So wurde am häufigsten angegeben, dass die Schülerinnen und Schüler auf den Ablauf der Vergleichsarbeiten vorbereitet wurden (77,2 Prozent). Dieses Ergebnis korrespondiert mit den Aufrufstatistiken für die Webseite mit

den allgemeinen Materialien (z.B. Handreichung zur Durchführung der Vergleichsarbeiten, Erläuterungen zur Dateneingabe, vgl. 0, S. 19). Ungeklärt bleibt in diesem Zusammenhang, ob und aus welchem Grund in den anderen Fällen darauf verzichtet wurde. So ist es durchaus vorstellbar, dass die Lehrkräfte ihren Klassen auf diesem Wege das Entstehen von Angst ersparen wollten.

Ebenfalls häufig wurde angegeben, die Schülerinnen und Schüler in Bezug auf Teststrategien (53,6 Prozent) und Aufgabenformate bzw. VERA 2004-Aufgaben (über 30 Prozent) vorbereitet zu haben. Besonders selten berichteten die Lehrkräfte dagegen die Vorbereitung auf VERA-Korrekturkriterien: Dieses Vorgehen wurde nur in fünf Fällen berichtet. Alles in allem zeigen die Ergebnisse jedoch deutlich, dass bei den Lehrkräften das Bedürfnis besteht, ihre Schülerinnen und Schüler auf die Vergleichsarbeiten vorzubereiten.

4.4.4 Kooperation bei der Aufgabenauswahl und -auswertung

Neben Informationen zum sozialen Kontext wurde im Lehrerfragebogen auch erfragt, inwiefern die Auswahl der Wahlaufgaben in Mathematik sowie die Auswertung der Vergleichsarbeiten in Kooperation mit Kolleg/innen erfolgte. Da die Kooperation (z.B. zur Besprechung der VERA-Ergebnisse) für die Tiefe der Verarbeitung wichtig ist und gerade die Auseinandersetzung mit den Inhalten und Konzepten der Vergleichsarbeiten den Blick auf Ansatzpunkte für Unterrichtsentwicklung eröffnet, wurde eine Kooperation durch das Prozedere der Aufgabenauswahl (pro Schule nur eine Aufgabenkombination erlaubt) und der Vergleichsangebote in den Ergebnisrückmeldungen forciert. Die Antworten der Lehrkräfte zu diesen Fragen können als Hinweis zu den schulischen Bedingungen, insbesondere dem Evaluations- und Kooperationsklima interpretiert werden. Tabelle 29 belegt, dass bei der *Aufgabenauswahl* in den meisten Fällen kooperiert wurde. Demgegenüber erfolgte die *schulinterne Auswertung* seltener in Zusammenarbeit mit Kolleg/innen. Die von VERA intendierte Kooperation im Zusammenhang mit den Vergleichsarbeiten ist hier noch nicht überzeugend gelungen. Ein Grund könnte sein, dass die Lehrkräfte sich für die zeitintensivere Auswertung außerhalb des Unterrichts verabreden mussten. Dies lässt sich mit dem Arbeitsalltag von Lehrkräften u. U. schwer in Einklang bringen. Eine andere Erklärung wäre, dass immer noch sehr viel Angst davor besteht, die Klassenergebnisse und damit den eigenen Unterricht vor anderen offen zu legen. Bei der Auswertung in Deutsch ergibt sich erfreulicherweise eine Steigerung gegenüber dem Vorjahr von 19,3 Prozent auf 28,3 Prozent, in Mathematik von 22,1 auf ebenfalls 28,3 Prozent.

Tabelle 29: *Kooperation im Rahmen der Vergleichsarbeiten;* *Angaben in Prozent*

		ja	nein	N (Klassen)
Aufgabenauswahl im Team	Mathematik	93,7	6,3	252
Auswertung im Team	Deutsch	28,3	71,7	252
	Mathematik	28,3	71,7	252

5 Ausblick

Das Projekt VERA hat in vielfacher Hinsicht Neuland betreten, insbesondere mit dem auf Kommunikation und Kooperation zielenden Prinzip der schulinternen Auswahl von Aufgaben und mit der konsequenten Nutzung des Internet, die im letzten Jahr mit dem Ziel einer Verringerung der Arbeit für die beteiligten Lehrkräfte noch weiter ausgebaut und zugleich vereinfacht wurde. Die Vergleichsarbeiten der kommenden Jahren werden sich konsequent an den bundeslandübergreifenden Bildungsstandards orientieren.

An dieser Stelle möchten wir nochmals an die in Kapitel 1.1 skizzierten bildungspolitischen Ziele von Vergleichsarbeiten erinnern. Vergleichsarbeiten sind kein Selbstzweck, sondern ein Werkzeug zur Bestandsaufnahme, Sicherung und Verbesserung der Bildungsqualität. Der eigentliche Wert von VERA als einem bundeslandübergreifenden und flächendeckend angelegten Unternehmen, das im jährlichen Zyklus angelegt ist, wird sich danach bemessen, ob VERA zu einer Verbesserung insbesondere des Unterrichts (Nutzung pädagogischer und fachdidaktischer Impulse der Ergebnisrückmeldungen, der Handreichungen und kommentierten Aufgabenbeispiele), der Schulentwicklung (Anstöße zur Verbesserung der Evaluations- und Kooperationskultur) sowie der Professionalisierung der Lehrerschaft im Bereich der pädagogischen Diagnostik (Erfassung und Verbesserung von Aspekten der Diagnosekompetenz) beiträgt. Für *System Monitoring* eignen sich Studien, die – wie TIMSS, PISA, IGLU oder DESI – anders angelegt sind als VERA.

Seit der „empirischen Wende“ ist der Grundsatz der Wirkungsorientierung im Bereich von Schule und Unterricht unbestritten: Das Ausmaß der Erreichung (oder Verfehlung) bildungspolitischer Ziele muss nachgewiesen, also empirisch überprüft werden. Einen wichtigen Schritt in diese Richtung haben wir in Gestalt internetbasierter Lehrerbefragungen zur Rezeption und zum Nutzen von VERA getan. Die Ergebnisse dieser Befragungen, deren Ergebnisse nicht Teil dieses Berichtes sind, zeigen, welches Potenzial die Vergleichsarbeiten für die Reflektion und Veränderung des eigenen Unterrichts, die Professionalisierung und die Intensivierung der schulinternen Kooperation besitzen. Wenn dazu als flankierendes Maßnahmenbündel ein entfaltetes Unterstützungssystem kommt, dann wären dies günstige Bedingungen dafür, dass Evaluation nicht bei der Standortbestimmung stehen bleibt, sondern für Innovation genutzt wird. Ein zweiter Schritt in Richtung "Unterrichtsentwicklung" besteht in einer an VERA angedockten Studie des Grundschulunterrichts in den Fächern Deutsch und Mathematik, die im laufenden Schuljahr in zwei Bundesländern durchgeführt wird. Diese als Zweipunktmessung angelegte Unterrichtsstudie, deren Kern eine Videographie des Unterrichts ist, nutzt die VERA-Erhebung vom Herbst 2005 als "base line" und führt am Ende des Schuljahres eine weitere Lernstandserhebung in Mathematik und Deutsch (Leseverständnis) durch. Die Nutzung der so gewonnenen Daten über unterrichtliche Bedingungen der Kompetenzentwicklung im Verlaufe der vierten Klassenstufe ist ein weiterer Mosaikstein, um die Vergleichsarbeiten stärker für die Verbesserung des Lehrens und Lernens zu nutzen.

6 Glossar

aggregieren, Aggregation, aggregierte Ebene, aggregierte Effekte, das Aggregieren bezeichnet einen datentechnischen Vorgang, bei dem mehrere Fälle einer Gruppe zu einem neuen Fall zusammengefasst („aggregiert“) werden. Beispielsweise lassen sich in der vorliegenden Untersuchung die Daten von allen Schülerinnen und Schülern einer Klasse zu →arithmetischen Mitteln auf Klassenebene aggregieren. Neben der Aggregierung von der Individualebene (Angaben einzelner Schüler) auf die Klassenebene sind auch Aggregierungen auf die Ebene der Schule oder der Schulart denkbar.

arithmetisches Mittel, arithmetischer Mittelwert, Durchschnittswert
→Mittelwert.

DESI, Deutsch Englisch Schülerleistungen International. Projekt der Kultusministerkonferenz, in dem es - als Komplement zu →PISA - um die aktive Beherrschung der Muttersprache und des Englischen als Fremdsprache geht. Das Projekt DESI wird von einem Konsortium unter der Leitung des DIPF (Deutsches Institut für Internationale Pädagogische Forschung) Frankfurt durchgeführt.

Effektstärke, Maß für die Größe bzw. die praktische Bedeutsamkeit eines Effekts (d.h. eines Unterschieds zwischen Mittelwerten, Streuungen, Korrelationen usw.). Es gibt verschiede Effektstärkemaße: Beim Vergleich der Mittelwerte zweier Gruppen (→t-Test) kann das Maß d verwendet werden: Größe des Unterschieds zwischen beiden Gruppenmittelwerten, dividiert durch die gemittelte Streuung. Als Faustregel gelten in der experimentellen Forschung Werte für d um 0,2 als kleine, um 0,5 als mittlere und um 0,8 als große Effektstärken. Im Kontext nicht-experimenteller pädagogisch-psychologischer Forschung sind auch kleinere Effekte beachtenswert und interpretationswürdig. Da allerdings die jeweilige Forschungslage zu berücksichtigen ist, dürfen die angegebenen Werte nicht dogmatisch als absolute Grundlage der Bewertung aufgefasst werden. Effektstärkemaße werden unter anderem deshalb verwendet, weil Aussagen über die Signifikanz eines Effekts u.a. von der Stichprobengröße abhängen (bei großen Stichproben werden schon sehr kleine Effekte statistisch signifikant). Die Effektstärke ist dagegen weitgehend unabhängig von der Stichprobengröße.

erklärte Varianz, aufgeklärte Varianz, die erklärte Varianz ist derjenige prozentuale Anteil der →Varianz der Werte einer Variablen x, der aufgrund der Werte einer anderen Variable y erklärbar ist. Bei einer Korrelationsrechnung wird die erklärte Varianz durch das Quadrat des →Korrelationskoeffizienten bestimmt.

IEA, International Association for the Evaluation of Educational Achievement. Diese Organisation hat die weltweit meisten internationalen Vergleichsstudien, darunter →TIMSS, durchgeführt.

IGLU, Internationale Grundschul-Lese-Untersuchung. Deutsche Teilstudie der Studie PIRLS (Progress in International Reading Literacy Study) der →IEA, ergänzt um Mathematik und naturwissenschaftliche Teilkomponenten (IGLU-E). Die Hauptuntersuchung fand 2001 statt, die ersten Ergebnisse werden 2003 publiziert. Alle 16 Bundesländer haben sich an IGLU beteiligt, 13 an IGLU-E.

Intervallskala, intervallskalierte Variable, Skala, bei der gleich große Unterschiede zwischen den Skalenwerten gleich große Merkmalsunterschiede anzeigen (z.B. ist der Temperaturunterschied zwischen den Skalenwerten 16° und 18° genau so groß wie der zwischen 21° und 23°); bei einer Verhältnisskala (z.B. Länge, Gewicht) ist darüber hinaus der Nullpunkt eindeutig festgelegt.

Item, Bezeichnung für die Aufgaben eines Tests oder die Fragen/Feststellungen eines Fragebogens. Die Items werden häufig zu einer →Skala zusammengefasst.

Koeffizient, ein Koeffizient ist ein statistischer bzw. mathematischer Kennwert. Pearsons r ist z.B. ein →Korrelationskoeffizient, d. h. ein statistisches Zusammenhangsmaß.

Korrelation, korrelieren, Korrelationskoeffizient, eine Korrelation ist ein statistisches Maß für den Grad des linearen Zusammenhangs zwischen zwei →Variablen (Merkmalen) x und y. Für →intervallskalierte Daten ist das Korrelationsmaß der Pearsonsche Produkt-Moment-Korrelationskoeffizient r_{xy} (kurz „**Pearsons r**“ oder nur „**r**“). r_{xy} hat einen Wertebereich von –1 bis +1. Ein hohes negatives r_{xy} besagt: Je höher das eine Merkmal ausgeprägt ist, desto niedriger ist das andere Merkmal, und je niedriger das eine Merkmal, desto höher das andere Merkmal. Ein hohes positives r_{xy} besagt sinngemäß entsprechend: Je höhere Werte das eine Merkmal annimmt, desto höhere hat auch das andere (bzw. je niedriger, desto niedriger). Ein r_{xy} von Null sagt aus, dass zwischen den beiden Merkmalen kein linearer Zusammenhang besteht. r_{xy}^2 ist ein Maß für die →erklärte Varianz.

Kriterium(s-Variable)
Ein anderer Begriff für →abhängige Variable, siehe auch →Regressionsanalyse.

Lösungswahrscheinlichkeit; die Lösungswahrscheinlichkeit einer Aufgabe gibt an, wie groß die Wahrscheinlichkeit ist, dass ein Schüler bzw. eine Schülerin diese Aufgabe löst. Die Lösungswahrscheinlichkeit wird mit dem Wert p (vom englischen *probability*) angegeben und liegt zwischen 0 und 1. Eine Lösungswahrscheinlichkeit von p = 0,47 beispielsweise besagt, dass 47 Prozent der Schülerinnen und Schüler einer definierten Gruppe diese Aufgabe lösen.

M
Abkürzung für →Mittelwert (engl. Mean).

M ± SD, Wertebereich, der durch eine Streuungseinheit (→SD) oberhalb und unterhalb des Mittelwerts (M) abgedeckt wird.

MARKUS, Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext. Gegenstand ist eine Vollerhebung der Mathematikleistungen der Schüler in der 8. Jahrgangsstufe sowie zu individuellen Lernvoraussetzungen, zum persönlichen Lernhintergrund und zu Unterrichtsmerkmalen (Mathematiktests, Schüler-, Lehrer- und Schulleiterfragebogen). Durchgeführt von einer Forschergruppe der Universität Landau.

Median, der Median beschreibt den Wert einer Verteilung, ober- und unterhalb dessen 50% aller Fälle oder Werte angesiedelt sind.

Mittelwert, Kurzbezeichnung für das arithmetische Mittel. Es entspricht der Summe der Einzelwerte aller Fälle dividiert durch die Zahl der Fälle. Der Mittelwert ist ein sinnvolles Maß, wenn mindestens →intervallskalierte Daten vorliegen.

N, bezeichnet die Anzahl der untersuchten Personen.

Normalverteilung, auch Gaußsche Verteilung oder Glockenkurve genannt, ist eine symmetrische, glockenförmige Verteilung, die anhand von nur zwei Kennwerten vollständig beschrieben wird. Diese Kennwerte sind der →Mittelwert und die →Standardabweichung. Bei einer Normalverteilung entfallen 68 % aller Fälle auf das Intervall von einer Standardabweichung unterhalb bis einer Standardabweichung oberhalb des Mittelwertes.

Objektivität, Objektivität ist ein Gütekriterium für sozialwissenschaftliche Messungen. Sie besagt, dass die Ergebnisse der Messung unabhängig vom Untersucher sind.

Objektivität in Schulleistungsuntersuchungen ist gegeben, wenn für alle Schülerinnen und Schüler gleiche Aufgabenstellungen, Bearbeitungszeiten, Erläuterungen der Aufgaben, Arbeitsmaterialien u. ä. gelten und wenn Auswertung und Interpretation nach klaren Kriterien, die unabhängig von der Person des Auswerters sind, erfolgen.

PISA, Programme for International Student Assessment. Studie der OECD (1998 - 2007) zur Lesekompetenz, zur mathematisch-naturwissenschaftlichen Grundbildung und zu fächerübergreifenden Kompetenzen mit vielfältigen Indikatoren für Lernergebnisse bei 15jährigen Schülern. Federführend für den ersten Zyklus (PISA 2000) mit dem Schwerpunkt Leseverständnis: MPI für Bildungsforschung Berlin; für den zweiten Zyklus (PISA 2003) mit dem Schwerpunkt Mathematik: das IPN Kiel.

Populationsdaten, diese Daten repräsentieren die untersuchte Gesamtanzahl von Individuen (Grundgesamtheit oder auch Grundpopulation). Bei bestimmten Fragestellungen wird aus pragmatischen Erwägungen normalerweise nicht die Grundgesamtheit, sondern eine Stichprobe untersucht, die für die Grundgesamtheit repräsentativ ist.

Rasch-Modell, ein Messmodell im Rahmen der probabilistischen →Testtheorie, mit dessen Hilfe Personen unterschiedlicher Fähigkeit und Aufgaben unterschiedlicher Schwierigkeit auf einer gemeinsamen Skala bzw. Dimension abgebildet werden.

Regressionsanalyse, die (multiple) Regressionsanalyse ist ein Analyseverfahren, das den Zusammenhang zwischen einer →intervallskalierten abhängigen (zu erklärenden) Variable (dem so genannten Kriterium) und mehreren, ebenfalls intervallskalierten unabhängigen (erklärenden) Variablen (den so genannten Prädiktoren) aufdeckt. Bei der Berechnung der Regressionsgleichung werden die →Korrelationen der Prädiktoren untereinander berücksichtigt.

Reliabilität, Reliabilität ist ein Gütekriterium für sozialwissenschaftliche Messungen, das die Zuverlässigkeit einer Messung kennzeichnet. Reliabel ist ein Test oder eine Skala, wenn nur geringe Messfehler auftreten.

SD, Standard deviation: englisch für Streuung oder →Standardabweichung.

Signifikanz, signifikant, Signifikanzniveau, von einem signifikanten oder statistisch bedeutsamen Ergebnis spricht man im allgemeinen dann, wenn die Irrtumswahrscheinlichkeit sehr gering (in der Regel höchstens 5%) ist.

Skala, 1. Kurzbezeichnung für eine Einschätz- oder Beurteilungsskala (Ratingskala). So entsprechen z.B. die Antwortmöglichkeiten von 0 = „nie“ bis 4 = „sehr oft“ im Lehrerfragebogen zur Einschätzung der inneren Differenzierung einer fünfstufigen Skala. **2**. Inhaltlich zusammenpassende Einzelitems oder –fragen (→Items) können, z. B. durch Aufsummieren oder Mittelwertbildung, zu einer Skala zusammengefasst werden. Ein Beispiel ist die Skala „Innere Differenzierung“, bei der für jede Lehrkraft der Mittelwert ihrer Antworten auf 7 Fragen des Lehrerfragebogens berechnet wurde, um ein Maß für ihre Bereitschaft zu erhalten, Maßnahmen der inneren Differenzierung einzusetzen.

Standardabweichung, SD, Die Standardabweichung ist ein so genanntes Streuungsmaß, das für intervallskalierte Daten Auskunft darüber gibt, wie homogen oder heterogen eine Merkmalsverteilung ist. Je kleiner die Standardabweichung ist, desto enger gruppieren sich die Werte der einzelnen Fälle um den →Mittelwert; je größer sie ist, desto weiter streuen sie um den Mittelwert. Liegt eine →Normalverteilung vor, so lässt sich über die Verteilung folgendes sagen: Im Bereich Mittelwert ± eine Standardabweichung liegen etwa 68 Prozent der Fälle; im Bereich Mittelwert ± zwei Standardabweichungen liegen etwa 95 Prozent der Fälle.

Streuung
→Standardabweichung

t-Test, beim t-Test handelt es sich um ein statistisches Testverfahren, mit dessen Hilfe geprüft wird, ob sich die →Mittelwerte zweier Gruppen statistisch →signifikant voneinander unterscheiden. So könnte z.B. geprüft werden, ob sich die mittlere Testleistung der Mädchen statistisch signifikant von der der Jungen unterscheidet. Als Prüfgröße wird der t-Wert berechnet. Das analoge statistische Verfahren für den Vergleich der Mittelwerte mehrerer Gruppen ist die →Varianzanalyse.

t-Wert
Statistische Prüfgröße bei einem →t-Test.

Testtheorie, die der Konstruktion von psychologischen und pädagogischen Tests zugrundeliegende mathematisch-statistische Theorie. Die Testtheorie befasst sich u.a. mit der Frage, wie empirische Testwerte und die zu messenden Merkmalsausprägungen zusammenhängen. Aus den Annahmen einer Testtheorie können Gütekriterien wie →Reliabilität, →Validität und →Objektivität abgeleitet werden. Man kann z.B. mit Hilfe einer Testtheorie prüfen, ob eine →Skala statistisch akzeptiert werden kann.

TIMSS, Third International Mathematics and Science Study. Diese Studie setzt die Reihe der international vergleichenden Schulleistungsuntersuchungen fort, die seit 1959 von der →IEA durchgeführt werden. TIMSS umfasste drei Altersgruppen: Population I (Ende der Grundschule), II (Sekundarstufe I) und III (Ende der Pflichtschulzeit, Sekundarstufe III) und fokussierte auf naturwissenschaftliche und mathematische Leistungen. In Deutschland wurden nur die Populationen II und III untersucht.

Validität, Validität ist ein Gütekriterium für sozialwissenschaftliche Messungen. Validität gibt die Gültigkeit eines Messinstruments, z. B. eines Tests, an. Ein Test ist valide, wenn er tatsächlich das misst, was er zu messen beansprucht.

Varianz, die Varianz entspricht dem Quadrat der →Standardabweichung. Mathematisch ist die Varianz der Durchschnitt aus den quadrierten Abweichungen aller Einzelwerte vom Mittelwert.

Varianzanalyse, die Varianzanalyse (ANOVA, analysis of variance) ist ein Verfahren zur statistischen Überprüfung von Mittelwertsunterschieden zwischen verschiedenen Gruppen und stellt damit die Verallgemeinerung des →t-Tests auf mehr als 2 Gruppen dar. So könnte z.B. geprüft werden, ob sich die mittleren Testleistungen der Schülerinnen und Schüler aus den 4 Bildungsganggruppen Gymnasium, Realschule, Hauptschule A-Kurs und Hauptschule G-Kurs statistisch signifikant voneinander unterscheiden.

7 Literatur

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F. & Carstensen, C. H. (2004). Mathematische Kompetenz. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 47-92). Münster: Waxmann.

Deutsches PISA - Konsortium (Hrsg.). (2003). *PISA 2000 - Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland*. Opladen: Leske + Budrich.

Helmke, A. & Hosenfeld, I. (2004). Vergleichsarbeiten - Kompetenzmodelle - Standards. In M. Wosnitza, A. Frey & R. S. Jäger (Hrsg.), *Lernprozesse, Lernumgebungen und Lerndiagnostik. Wissenschaftliche Beiträge zum Lernen im 21. Jahrhundert* (S. 56-75). Landau: Verlag Empirische Pädagogik.

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2002). Unterricht, Mathematikleistung und Lernmotivation. In A. Helmke & R. S. Jäger (Hrsg.), *Die Studie MARKUS - Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext.* (S. 413-480). Landau: Verlag Empirische Pädagogik.

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2003). Diagnosekompetenz in Ausbildung und Beruf entwickeln. *Karlsruher Pädagogische Beiträge* (55), 15-34.

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold & C. Griese (Hrsg.), *Schulleitung und Schulentwicklung* (S. 119-144). Hohengehren: Schneider-Verlag.

Helmke, A. & Reich, H. H. (2001). Die Bedeutung der sprachlichen Herkunft für die Schulleistung. *Empirische Pädagogik, 15* (4), 567-600.

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (Enzyklopädie der Psychologie, Pädagogische Psychologie, Vol. 3, S. 71-176). Göttingen: Hogrefe.

Hosenfeld, I., Helmke, A., Ridder, A. & Schrader, F.-W. (2001). Eine mehrebenenanalytische Betrachtung von Schul- und Klasseneffekten. *Empirische Pädagogik, 15* (4), 513-534.

Hosenfeld, I., Helmke, A., Ridder, A. & Schrader, F.-W. (2002). Die Rolle des Kontextes. In A. Helmke & R. S. Jäger (Hrsg.), *Die Studie MARKUS - Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext.* (S. 155-256). Landau: Verlag Empirische Pädagogik.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Frankfurt a. M.: DIPF.

Peek, R. & Dobbelstein, P. (2006). Benchmarks als Input für die Schulentwicklung. In H. Kuper & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen - Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 41-58). Münster: Waxmann.

Projektgruppe VERA-Deutsch. (2005). *Didaktische Erläuterungen Lesen*. Landau.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests* (Studies in Mathematical Psychology). Copenhagen: Nielsen & Lydiche.

Schwippert, K., Bos, W. & Lankes, E. M. (2003). Heterogenität und Chancengleichheit am Ende der vierten Jahrgangsstufe im internationalen Vergleich. In W. Bos, E. M. Lankes, M. Prenzel, K. Schwippert, G. Walther & R. Valtin (Hrsg.), *Erste Ergebnisse aus IGLU* (S. 265-302). Münster: Waxmann.

Zimmer, K., Burba, D. & Rost, J. (2004). Kompetenzen von Jungen und Mädchen. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 211-223). Münster: Waxmann.