



**Ergebnisbericht
VERA 2006/2007:**

Berlin

Andreas Helmke, Ingmar Hosenfeld

Jana Groß Ophoff, Ann Christin Halt, Florian Henk, Kevin Isaac, Ursula Koch,
Frank Scherthan, Franziska Thonke und Sonja Wagner

Universität Koblenz - Landau, Campus Landau

Stand: 31.03.2007

Inhalt

1. Berliner Ergebnisse

1.1	Fähigkeitsniveaus	3
	<i>1.1.1. Verteilung der Fähigkeitsniveaus in den Ländern</i>	<i>3</i>
	<i>1.1.2. Unterschiede innerhalb und zwischen den Klassen</i>	<i>4</i>
	<i>1.1.3. Leistungen von Mädchen und Jungen</i>	<i>6</i>
	<i>1.1.4. Migrationshintergrund</i>	<i>7</i>
1.2	"Fairer Vergleich"	9
1.3	Diagnosegenauigkeit	10
1.4	Lehrerfragebogen.....	14
	<i>1.4.1. Grundlegende Merkmale der Lehrkräfte</i>	<i>15</i>
	<i>1.4.2. Inhaltsbereiche im Unterricht</i>	<i>16</i>
	<i>1.4.3. Akzeptanz der Vergleichsarbeiten</i>	<i>18</i>
	<i>1.4.4. Vorbereitung auf die Vergleichsarbeiten</i>	<i>19</i>
	<i>1.4.5. Auswertung der Vergleichsarbeiten</i>	<i>22</i>
2	Literatur	23
3	Glossar	24

1 Berliner Ergebnisse

Der Bericht 2006 führt die Veröffentlichung der Senatsverwaltung für Bildung, Wissenschaft und Forschung zu den Ergebnissen der Vergleichsarbeiten am Anfang der Jahrgangsstufe 4 im Schuljahr 2005/2006 fort. Detaillierte Ausführungen zu den Zielen und zur Organisation von VERA, der Aufgabenentwicklung und Definition der Fähigkeitsniveaus sind dort zu finden.

1.1 Fähigkeitsniveaus

Im Folgenden werden die Fähigkeitsniveaus unter verschiedenen Gesichtspunkten für Berlin im Schuljahr 2006/2007 diskutiert: Zunächst wird allgemein die Verteilung der Schüler auf den einzelnen Niveaus dargestellt (vgl. 1.1.1). Im Anschluss werden Unterschiede zwischen (bzw. innerhalb) den untersuchten Klassen/Schulen (siehe 1.1.2), Geschlechtsunterschiede (siehe 1.1.3) sowie Unterschiede zwischen Schülerinnen und Schülern mit Deutsch als dominanter vs. nicht dominanter Sprache (vgl. 1.1.4) 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.

1.1.1. Verteilung der Fähigkeitsniveaus in den Ländern

Im Fach Mathematik wird deutlich, dass in den Standardbereichen „Zahlen und Operationen“ und „Größen und Messen“ etwa zwei Drittel der Schülerinnen und Schüler erweiterte bis fortgeschrittene Fähigkeiten aufweisen (für „Zahlen und Operationen“ 69,2% und für „Größen und Messen“ 62,6%). Während sich für „Zahlen und Operationen“ in etwa eine Gleichverteilung auf die Fähigkeitsniveaus eins und drei zeigt (28,9% bzw. 25,0%), besetzen in „Größen und Messen“ mehr Kinder ein Niveau, das höchstens das Beherrschen elementarer Aufgaben umfasst.

Im Fach Deutsch zeigt sich, dass im Bereich „Sprache und Sprachgebrauch untersuchen“ 70 Prozent der Schülerinnen und Schüler erweiterte bis fortgeschrittene Fähigkeiten erreichen, während dieser Anteil im Bereich „Lesen – mit Texten und Medien umgehen“ deutlich darunter liegt (50,6%). In beiden Inhaltsbereichen findet sich darüber hinaus, dass für einen großen Anteil der Kinder die Leistungen als „nicht auswertbar“ eingestuft wurden (Sprache und Sprachgebrauch untersuchen: 7,0%; Leseverständnis: 8,2%).

Im Vergleich mit der länderübergreifenden Fähigkeitsniveauverteilung fällt auf, dass in Berlin in allen erfassten Bereichen weniger Kinder dem dritten Fähigkeitsniveau und mehr Kinder dem Fähigkeitsniveau eins bzw. nicht auswertbaren Leistungen zugeordnet wurden: In den Bereichen „Lesen“ und „Größen und Messen“ scheint angesichts der häufigen Besetzung des unteren Fähigkeitsniveaus ein besonders großer Förderbedarf zu bestehen.

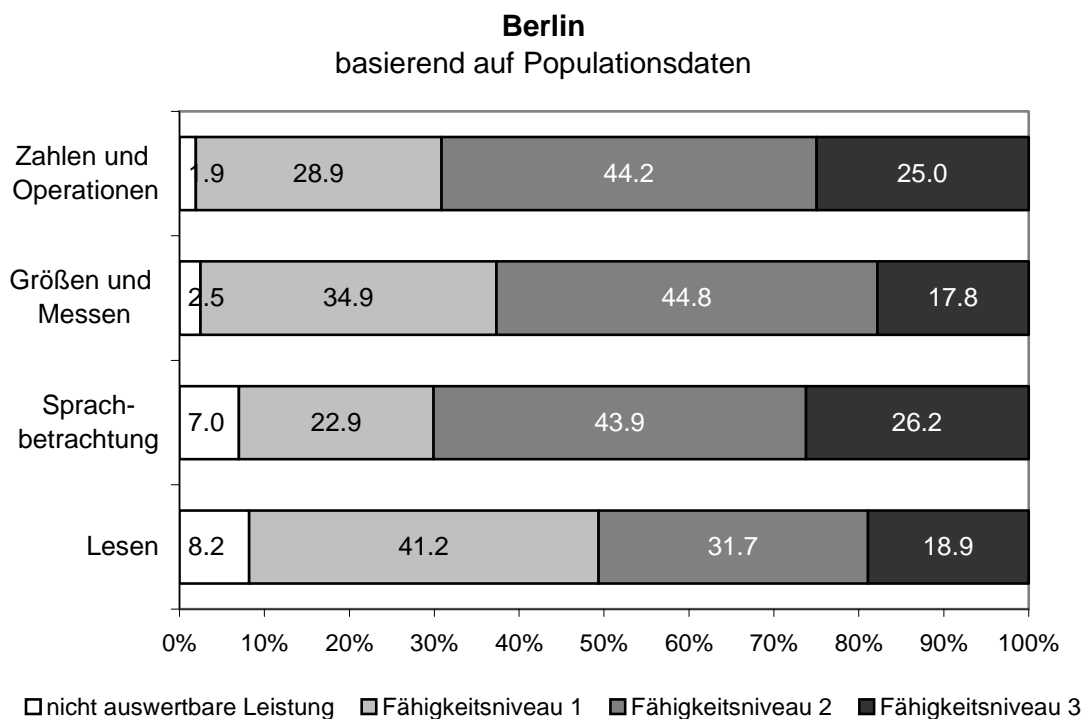


Abbildung 1: Gesamtverteilung der Fähigkeitsniveaus (2006); Angaben in Prozent

1.1.2. 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 (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 1 dargestellten Prozentanteile.

Table 1: Zerlegung in die Varianz auf den drei Ebenen Schule, Klasse und Individuum;
Angaben in Prozent

Bereich	Schulebene	Klassenebene	Individualebene
Zahlen und Operationen	18.9	8.0	73.1
Größen und Messen	20.8	7.7	71.5
Sprache und Sprachgebrauch untersuchen	21.4	8.8	69.8
Lesen	23.6	9.8	66.5
N min	366	1012	22310

Wie erwartet ist die Leistung in den Vergleichsarbeiten vorrangig mit Merkmalen der Individualebene verknüpft. Die Varianzanteile liegen zwischen 66,5 Prozent („Lesen“) und 73,1 Prozent („Zahlen und Operationen“). Ebenfalls erhebliche Anteile gehen auf schulische Merkmale zurück, sie bewegen sich zwischen 18,9 Prozent („Zahlen und Operationen“) und 23,6 Prozent („Lesen“). Den vergleichsweise geringsten Beitrag leistet die Zugehörigkeit zu einer Schulklasse (hinter der Unterschiede des Unterrichts und des Klassenkontextes stehen) mit 7,7 Prozent („Größen und Messen“) bis 9,8 Prozent („Lesen“).

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. Demgegenüber sind Unterschiede zwischen Schulen und Klassen nicht etwa unabhängig von individuellen Faktoren, sondern 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 (Helmke, Hosenfeld & Schrader, 2002, 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), bestätigt aber die Ergebnisse vom letzten Jahr. Er reflektiert möglicherweise 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 *zwi-*

schen den Schulen als Folge unterschiedlicher Einzugsgebiete erheblich variieren können.

1.1.3. 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 (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 2 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.

Tabelle 2: Verteilung der Fähigkeitsniveaus, getrennt nach Geschlecht; Angaben in Prozent

		n.a.L.*	FN1	FN2	FN3	N (Kinder)	d^{**}
Zahlen und Operationen	Mädchen	2,5	32,5	44,4	20,6	11171	-0,25
	Jungen	1,4	25,3	44,0	29,3	11219	
Größen und Messen	Mädchen	2,9	42,0	43,3	11,8	11171	-0,40
	Jungen	2,0	27,8	46,4	23,8	11219	

* 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.

		n.a.L.*	FN1	FN2	FN3	N (Kinder)	d**
Sprache und Sprachgebrauch untersuchen	Mädchen	6,8	21,8	44,3	27,1	11150	0,05
	Jungen	7,2	24,1	43,4	25,3	11161	
Lesen	Mädchen	7,7	40,8	32,4	19,1	11153	0,04
	Jungen	8,7	41,4	31,1	18,8	11162	

* nicht auswertbare Leistung

** Maß für die Effektstärke

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

Hier schneiden die Mädchen nur in vergleichsweise geringem Maße besser ab als die Jungen ($d = 0,05$ für „Sprache und Sprachgebrauch“ untersuchen bzw. $d = 0,04$ für „Lesen“). Im Fach Mathematik hingegen zeigt sich mit $d = -0,40$ noch der bedeutsamste Unterschied mit einem Vorsprung der Jungen im Bereich „Größen und Messen“. Dort unterscheiden sich Mädchen und Jungen im höchsten Fähigkeitsniveau um einen Anteil von 12,0 Prozentpunkten. In „Zahlen und Operationen“ unterscheiden sich die Leistungen der Jungen und Mädchen ebenfalls, aber in etwas geringerem Maße ($d = -0,25$).

1.1.4. 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“ ($N_{\text{minimal}} = 17195$ bzw. 5120) 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 (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 2 und Abbildung 3 sind die prozentualen Schülerleistungen jeweils nach Deutsch als dominante und nicht-dominante Sprache dargestellt. Tabelle 3 zeigt die Effektstärken der Leistungsunterschiede.

Berlin - Deutsch als dominante Sprache
basierend auf Populationsdaten

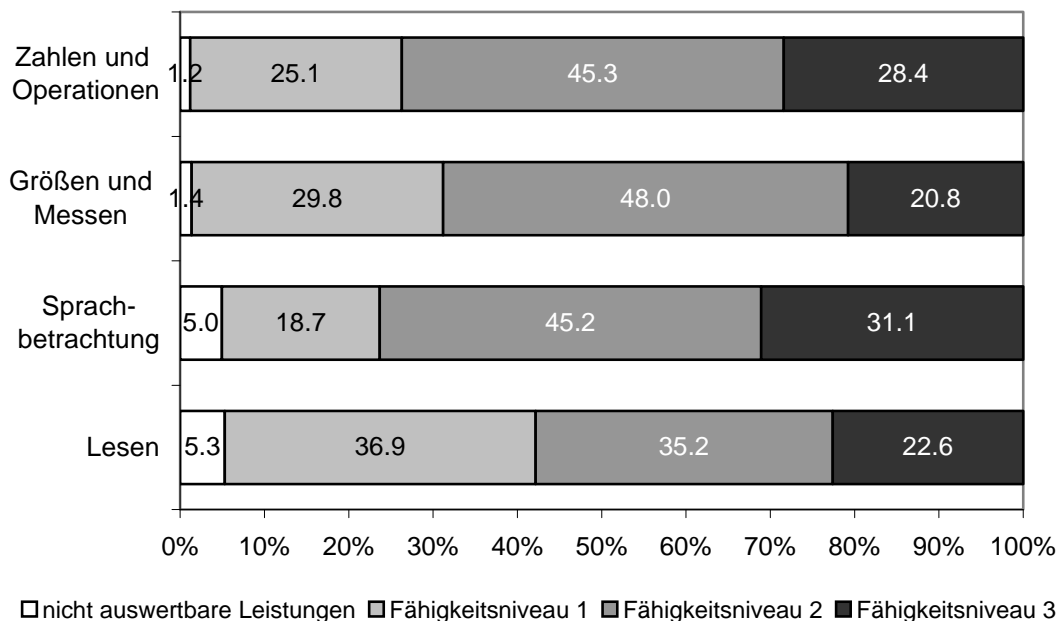


Abbildung 2: Gesamtverteilung der Fähigkeitsniveaus für Deutsch als dominante Sprache; Angaben in Prozent

Berlin - Deutsch als nicht-dominante Sprache
basierend auf Populationsdaten

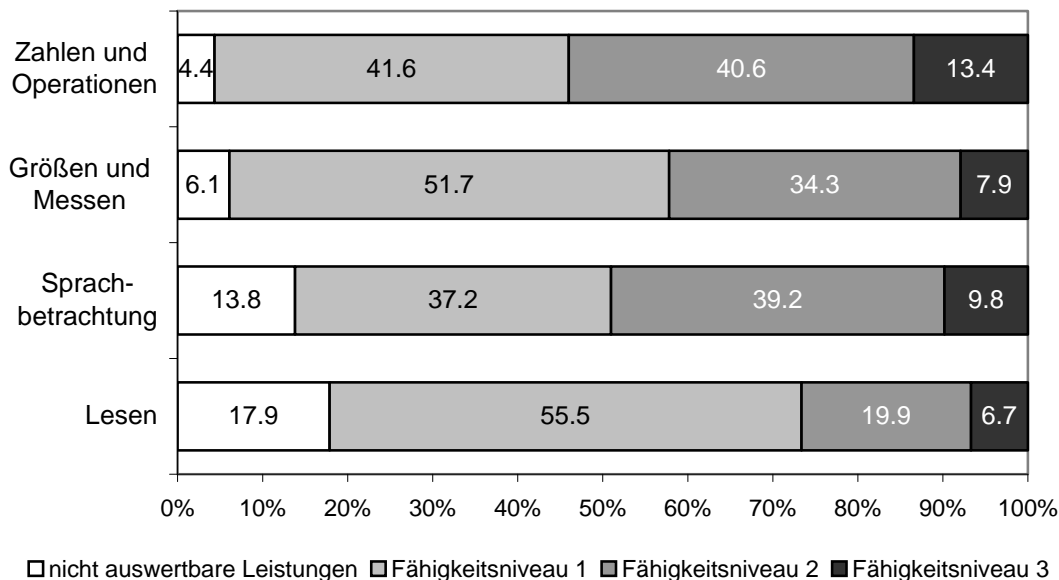


Abbildung 3: Gesamtverteilung der Fähigkeitsniveaus für Deutsch als nicht-dominante Sprache; Angaben in Prozent

Es zeigen sich bedeutsame Unterschiede zwischen Kindern mit Deutsch als dominanter und Kindern mit Deutsch als nicht-dominanter Sprache sowohl in Mathematik ($d = 0,51$ und $0,65$) als auch in Deutsch ($d = 0,49$ und $0,77$). So erreichen in drei der vier Bereichen von den Schülerinnen und Schülern mit Deutsch als nicht-dominanter Sprache nur unter zehn Prozent das höchste Fähigkeitsniveau, im Bereich „Lesen“ sind es 6,7 Prozent. Dabei fällt vor allem der hohe Anteil von Kindern auf, die in „Größen und Messen“ und „Lesen“ ein Niveau erreichen, das höchstens das Beherrschen elementarer Aufgaben umfasst (51,7 Prozent in „Größen und Messen“ und 55,5 Prozent im Lesen). Insbesondere beim Lesen kann infolge des hohen Anteils von Kindern mit nicht auswertbarer Leistung (knapp 18 Prozent) von einem substantiellen Unterschied zu Kindern, deren dominante Sprache Deutsch ist, gesprochen werden.

Die Resultate bestätigen die Vermutung, dass Merkmale der Sprachherkunft für Schülerinnen und Schüler bereits in der Klassenstufe 4 mit erheblichen Leistungsunterschieden gekoppelt sind (Schwippert, Bos & Lankes, 2003).

Table 3: Effektstärken der Leistungsunterschiede von Deutsch dominante vs. nicht-dominante Sprache

	Zahlen und Operationen	Größen und Messen	Sprache und Sprachgebrauch untersuchen	Lesen	N min (Kinder)
Effektstärke (d)	0,51	0,65	0,49	0,77	22311

1.2 „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 erheblichen Einfluss auf die Leistungsfähigkeit der Kinder ausüben. Auf der Ebene von Klassen und Schulen entspricht dies der wichtigen Rolle des Schuleinzugsgebietes und der Klassenzusammensetzung. Von Schulen „im sozialen Brennpunkt“ spricht man – obwohl es keine verbindlichen Definitionen gibt – gemeinhin dann, wenn verschiedene unterrichts- und lernerschwerende Faktoren in konzentrierter Form auftreten, etwa bei Schulen, deren Klientel durch stark überdurchschnittliche prozentuale 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 erfasst, um für den „fairen Vergleich“ eine fundierte Datenbasis zu erzeugen.

Aufgrund der niedrigen Erfüllungsquote bei den Angaben zum Kontext konnte der faire Vergleich dieses Jahr nicht angeboten werden.

1.3 Diagnosegenauigkeit

Im Rahmen der Vergleichsarbeiten lassen sich nicht nur Aussagen über den Leistungsstand und die erreichten Fähigkeitsniveaus der beteiligten Schülerinnen und Schüler machen, sondern auch über die Diagnosegenauigkeit der teilnehmenden Lehrkräfte. Diese stellt einen Teilaspekt diagnostischer Kompetenz dar, der vor allem für die Binnendifferenzierung im Unterricht und für die Auswahl geeigneten Lernmaterials eine besondere Bedeutung hat. Erforderlich hierfür sind u. a. Wissen über Personen und Wissen über Aufgabenmerkmale (diagnostisches und fachdidaktisches Wissen).

Bei VERA werden die Lehrkräfte daher gebeten, Aussagen darüber zu machen, wie viele ihrer Schüler die einzelnen Testaufgaben lösen werden. Diese von den Lehrkräften vorgenommene Einschätzung der Schwierigkeiten der Testaufgaben wird dann mit den tatsächlichen (empirisch) ermittelten Aufgabenschwierigkeiten in Beziehung gesetzt. Zurückgemeldet werden zwei verschiedene Kennwerte:

- (a) das Ausmaß der *Unter- oder Überschätzung* der Aufgabenschwierigkeit, d.h. Aussagen über das Leistungsniveau der Schulklasse (diagnostisches Wissen)
- (b) die *Korrelation* zwischen den beiden Rangreihen der Aufgabenschwierigkeiten, also die Ähnlichkeit der Rangordnung geschätzter vs. tatsächlicher (empirischer) Aufgabenlösungen (fachdidaktisches Wissen).

Diese beiden Aspekte der Diagnosegenauigkeit sollten unbedingt unterschieden werden. Eine Lehrkraft kann z.B. alle Aufgaben konstant etwas über- oder unterschätzen, die Rangordnung der von ihr geschätzten Aufgabenschwierigkeit aber identisch mit der tatsächlichen Rangordnung der gelösten Aufgabe sein.

Eine geringe Abweichung der durchschnittlichen geschätzten Aufgabenschwierigkeit von der empirischen Schwierigkeit sagt etwas über die *pädagogisch-psychologische* Diagnosefähigkeit aus, d.h. wie genau die Lehrkraft das durchschnittliche Leistungsniveau der Klasse einschätzen kann. Eine hohe Korrelation der geschätzten mit der empirischen Aufgabenrangreihe sagt dagegen eher etwas über die *fachdidaktische* Kompetenz der Lehrkraft aus und ist am ehesten Ausdruck einer zutreffenden Orientiertheit über Schwierigkeitsunterschiede zwischen Aufgaben. Wir sprechen deshalb im Folgenden vereinfacht von der pädagogischen und der fachdidaktischen Komponente der Diagnosegenauigkeit. Die beiden Kennwerte sind nicht nur konzeptuell, sondern auch statistisch fast vollkommen unabhängig voneinander (Korrelation über alle Länder: $r = ,03$ in Mathematik, $r = -,04$ in Deutsch).

Zur weiteren Auseinandersetzung mit Fragen der Diagnosegenauigkeit liegen nicht nur Handreichungen für die an VERA beteiligten Lehrkräfte, sondern auch diverse Publikationen der VERA-Mitarbeiter vor (z.B. Schrader, 2006; Schrader, Helmke, Hosenfeld, Halt & Hochweber, 2006; Helmke, Hosenfeld & Schrader, 2003, 2004).

a) Pädagogische Komponente: Unter- vs. Überschätzungstendenz

Wir berichten in der folgenden Abbildung für das Fach Mathematik das Ausmaß, in dem die Lehrkräfte die Leistungen ihrer Schülerinnen und Schüler unter- oder überschätzen. Hierzu liegen verwertbare Angaben von 64 Lehrkräften vor.

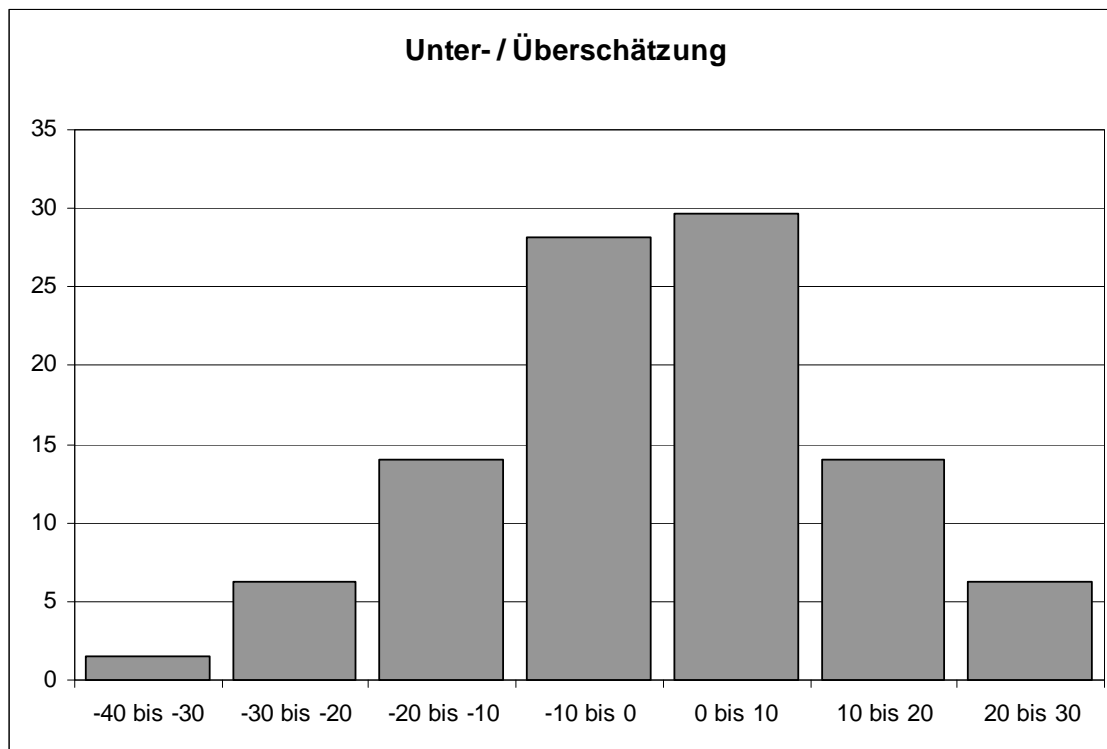


Abbildung 4: Prozentuale Unter- vs. Überschätzung des mathematischen Leistungsniveaus der Klasse (pädagogisch-psychologischer Aspekt der Diagnosegenauigkeit) durch die Lehrkräfte; Angaben in Prozent

Über alle Mathematikaufgaben hinweg, liegt die durchschnittliche Verschätzung der Lehrkräfte bei nicht einmal einem Prozent (-0,77 Prozent). Abbildung 4 zeigt außerdem, dass es in etwa gleichem Maße zu Über- und Unterschätzungen kommt.

Betrachtet man die beiden Inhaltsbereiche „Zahlen und Operationen (ZO)“ und „Größen und Messen (GM)“ getrennt voneinander, dann zeigt sich folgendes: während die Schülerleistungen im Bereich ZO im Durchschnitt recht genau eingeschätzt werden (die Abweichung beträgt 0,57 Prozent), wird sie im Bereich GM leicht *unterschätzt* (die Abweichung beträgt -1,6 Prozent).

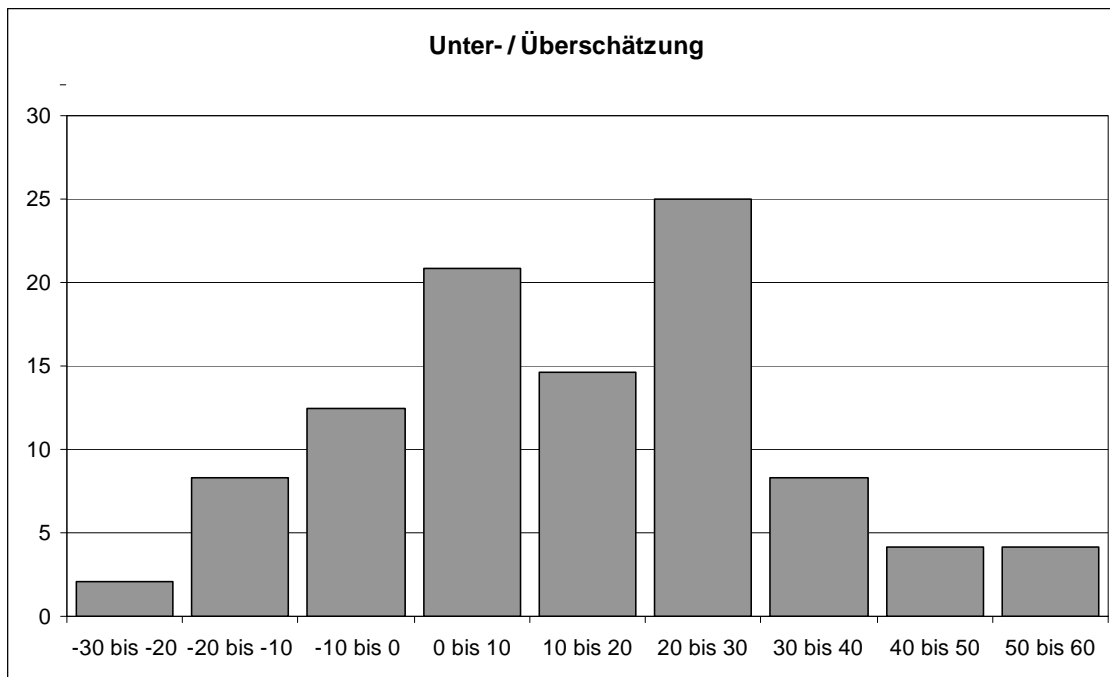


Abbildung 5: *Prozentuale Unter- vs. Überschätzung des Leistungsniveaus der Klasse (pädagogisch-psychologischer Aspekt der Diagnosegenauigkeit) im Bereich Lesen durch die Lehrkräfte; Angaben in Prozent*

Im Fach Deutsch bestand für die Lehrkräfte erstmalig die Möglichkeit eine Einschätzung einzugeben. Dies erfolgte für die Aufgaben im Inhaltsbereich Lesen. Von 48 Lehrkräfte liegen hierfür Daten vor. Die Lehrkräfte überschätzten im Durchschnitt die Lösungen um 14,5 Prozent. Die Überschätzung liegt damit über der in Mathematik. Hier zeigt sich, dass die mehrjährige Erfahrung der Lehrkräfte bei der Einschätzung im Bereich Mathematik zu einem genaueren Urteil führt.

b) Fachdidaktische Komponente: Vergleich der Aufgaben-Rangordnungen

Auch für diese Komponente liegen für das Fach Mathematik verwertbare Angaben von 64 Lehrkräften vor. Die durchschnittliche Korrelation liegt hier bei $r = ,36$.

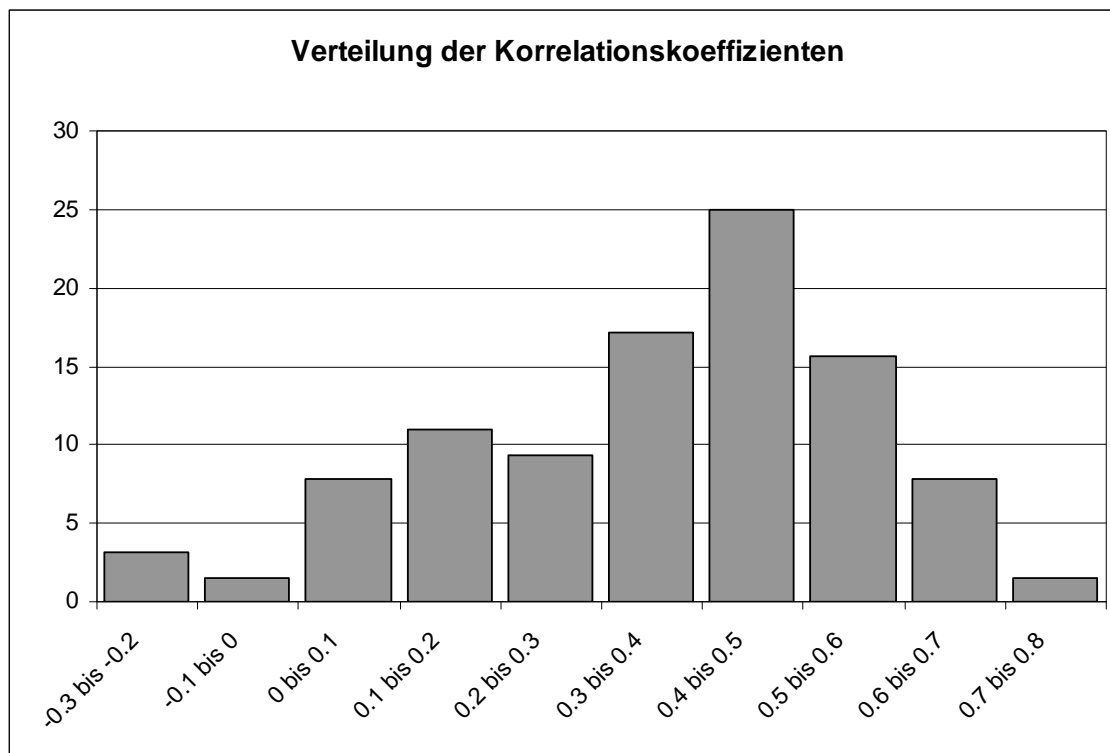


Abbildung 6: Verteilung der Korrelationskoeffizienten: Ähnlichkeit geschätzter und empirischer Aufgabenschwierigkeiten (fachdidaktischer Aspekt der Diagnosegenauigkeit) im Fach Mathematik; Angaben in Prozent

Betrachtet man die Verteilung in Abbildung 11 wird deutlich, dass für ca. 14 Prozent der Lehrkräfte Korrelationen um Null (diese kommen zustande, wenn nach Zufall geantwortet wird) oder aber sogar negative Korrelationen (diese bedeuten, dass eine der tatsächlichen Rangordnung entgegen gesetzte Rangordnung eingeschätzt wurde) festgestellt wurden. Es gibt kaum Klassen, deren Lehrkräfte eine gute oder sehr gute Diagnosegenauigkeit aufweisen (Korrelationen in Höhe von $r = ,70$ bzw. zwischen $r = ,80$ und $,90$). Über eine zufrieden stellende diagnostische Genauigkeit (Korrelationen zwischen $r = ,50$ und $r = ,60$) verfügen ca. 16 Prozent der Lehrkräfte, ca. 42 Prozent weisen positive Korrelationen auf, die noch als akzeptable Übereinstimmungen angesehen werden können (zwischen $r = ,30$ bis $r = ,50$).

Die Verteilung der Übereinstimmungen zwischen den Aufgabenrangreihen erstreckt sich von $- ,26$ bis $,74$. Diese Spannweite verdeutlicht, dass es zwischen den Lehrkräften beträchtliche individuelle Unterschiede in dieser fachdidaktischen Komponente der Diagnosegenauigkeit gibt.

Für den Inhaltbereich Deutsch ergibt sich eine durchschnittliche Korrelation von $r = ,23$. Auch hier zeigt sich eine etwas schlechtere Einschätzung als in Mathematik. Die Verteilung erstreckt sich dabei von $r = -,40$ bis $,90$, die Spannweite ist damit etwas größer als in Mathematik.

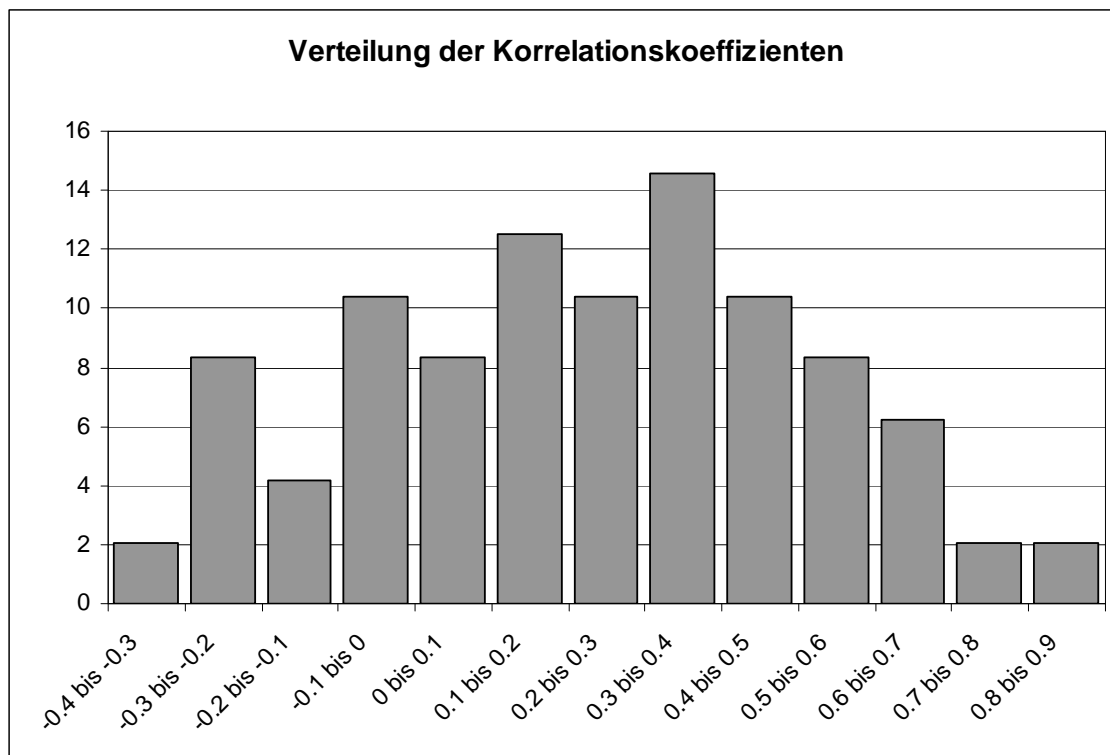


Abbildung 7: Verteilung der Korrelationskoeffizienten: Ähnlichkeit geschätzter und empirischer Aufgabenschwierigkeiten (fachdidaktischer Aspekt der Diagnosegenauigkeit) im Bereich Lesen; Angaben in Prozent

1.4 Lehrerfragebogen

Der Lehrerfragebogen konzentrierte sich in den Vergleichsarbeiten 2003, 2004 und 2005 im Wesentlichen darauf, schulleistungsrelevante Bedingungen (u.a. Standortvoraussetzungen, Klassenzusammensetzung) zu ermitteln. Zusätzlich zu den auf vier Fragen verkürzten Kontextinformationen konnten weitere Themen in den Blick genommen werden wie z.B. die Akzeptanz von Vergleichsarbeiten, die Vorbereitung, Korrektur und Auswertung der Vergleichsarbeiten, Unterrichtserfahrung und Ausbildung sowie Schwerpunktsetzung im Unterricht.

Im Folgenden werden die Ergebnisse aus dem Lehrerfragebogen für Berlin den Ergebnissen aus den weiteren Bundesländern gegenüber gestellt. Auf Grund der vielfältigen Bedingungen, die mit den hier dargestellten Variablen in Zusammenhang stehen (siehe auch Helmke, 2004, 2006), sollten die Ergebnisse mit Vorsicht interpretiert werden.

Der Lehrerfragebogen wurde in Berlin für 98,8 Prozent der Klassen aus den Zentralstichprobenschulen (N = 163) ausgefüllt, was 223 Fachlehrer/-innen aus 58 Schulen entspricht. Dabei wurde der Fragebogen von 23,3 Prozent der Lehrkräfte vollständig ausgefüllt. In unvollständig ausgefüllten Fragebögen wurden im Durchschnitt 6,9 Angaben (SD = 4,13) ausgelassen.

1.4.1. Grundlegende Merkmale der Lehrkräfte

26 Prozent der Berliner Lehrkräfte gaben in dem Lehrerfragebogen an, dass sie ihre Klasse sowohl in Mathematik als auch in Deutsch unterrichten. In den anderen VERA-Ländern unterrichten fast doppelt so viele Lehrkräfte beide Fächer (siehe Tabelle 4). Dabei haben in Berlin 52,9 Prozent derjenigen Lehrkräfte, die Deutsch unterrichten, dieses Fach auch grundständig studiert. In Mathematik ist der Anteil geringer: In Mathematik liegt der Anteil der Mathematik unterrichtenden Lehrkräfte, die dieses Fach grundständig studiert haben, bei 49,3 Prozent.

Tabelle 4: *Unterrichtetes Fach*

	Berlin		andere VERA-Länder ¹	
	Prozent	N (Lehrkräfte)	Prozent	N (Lehrkräfte)
Deutsch	36,32%	81	29,62%	279
Mathematik	37,67%	84	25,27%	238
beide Fächer	26,01%	58	45,12%	425
Gesamt	100%	223	100%	942

¹ Brandenburg, Bremen, Mecklenburg-Vorpommern, Rheinland-Pfalz (VERA 2006)

In der diesjährigen Zentralstichprobe zeigt sich (Tabelle 5), dass die Berliner Lehrkräfte seit etwas mehr Jahren tätig sind als in den anderen VERA-Ländern.

Tabelle 5: *Tätigkeit als Lehrkraft und Unterrichtserfahrung in Mathematik und Deutsch (in Jahren)*

Seit wie vielen Jahren...	Berlin		andere VERA-Länder ¹	
	M	SD	M	SD
tätig als Lehrkraft	24,2	8,90	21,8	12,00
Deutsch-Unterricht	20,2	10,92	20,0	12,14
Mathematik-Unterricht	20,6	11,17	19,8	12,01

¹ Brandenburg, Bremen, Mecklenburg-Vorpommern, Rheinland-Pfalz (VERA 2006)

In dem Histogramm für den Zeitraum der Tätigkeit als Lehrkraft (vgl. Abbildung) wird zusätzlich deutlich, dass sich die Verteilung einer Normalverteilungskurve nähert. Die meisten Lehrkräfte unterrichten demnach seit 20 bis 22 Jahren.

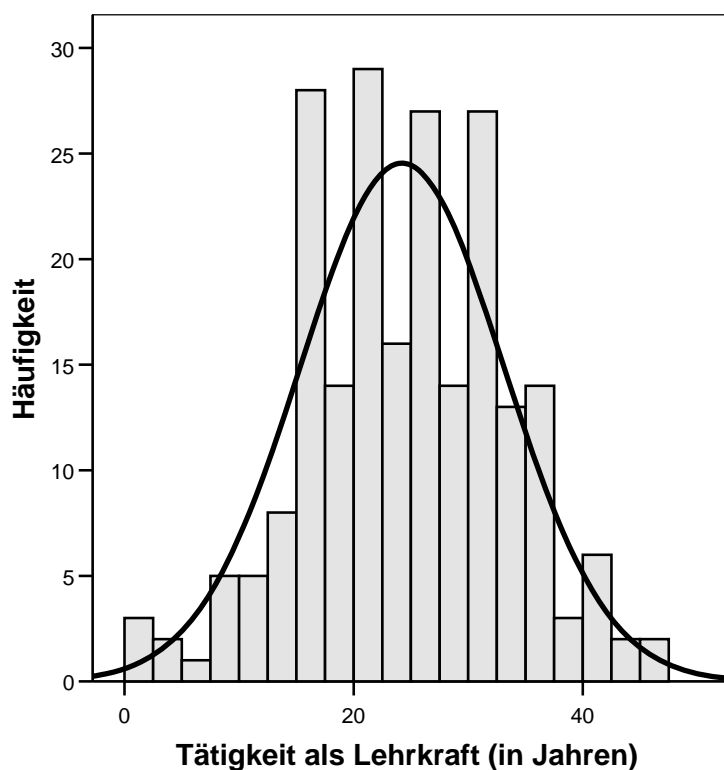


Abbildung 8: Histogramm für die Verteilung der Lehrertätigkeit in Jahren

1.4.2. Inhaltsbereiche im Unterricht

Nach Ditton (2000) sind v.a. die Unterrichtsbedingungen, also die unmittelbare Lehr- und Lernsituation, maßgeblich für den Lernerfolg. Um zu klären, ob bestimmte Unterrichtsformen oder -schwerpunkte in Zusammenhang mit den in den Vergleichsarbeiten ermittelten Leistungen stehen, wurde u.a. erfragt, in welchen Anteilen die fachspezifischen Inhaltsbereiche unterrichtet werden.

In Tabelle zeigt sich, dass im Fach Deutsch die Anteile gleichmäßig auf die unterschiedlichen Inhaltsbereiche verteilt sind und auch nur gering streuen. Etwas seltener werden in Berlin die Bereiche „Schreiben“ und „Sprache und Sprachgebrauch untersuchen“ unterrichtet. Im Vergleich zu den anderen VERA-Ländern ergeben sich darüber hinaus nur geringe Unterschiede.

Im Fach Mathematik sind die Anteile deutlich ungleichmäßiger verteilt: Während „Zahlen und Operationen“ den größten Anteil des Unterrichts ausmacht, werden die Bereiche „Muster und Strukturen“ sowie „Daten, Häufigkeit und Wahrscheinlichkeit“ in geringeren Anteilen unterrichtet.

Tabelle 6: Mittlerer Anteil der Inhaltsbereiche im Unterricht in Deutsch und Mathematik (in Prozent)

Anteile der Inhaltsbereiche (%) im Unterricht:	Berlin		andere VERA-Länder ¹	
	M	SD	M	SD
Deutsch				
Sprechen und Zuhören	21,1%	6,99	19,8%	6,69
Schreiben	19,1%	4,63	19,5%	5,11
Rechtschreibung	20,7%	5,24	21,6%	5,61
Lesen	22,4%	5,29	21,8%	5,55
Sprache untersuchen	16,7%	5,83	17,3%	4,19
Mathematik	M	SD	M	SD
Zahlen und Operationen	43,6%	14,50	44,4%	12,85
Raum und Form	16,4%	5,38	15,6%	5,34
Muster und Strukturen	11,0%	4,92	10,7%	4,59
Größen und Messen	19,1%	5,11	20,2%	5,90
Daten, Häufigkeit und Wahrscheinlichkeit	10,9%	5,48	9,5%	4,61

¹Brandenburg, Bremen, Mecklenburg-Vorpommern, Rheinland-Pfalz (VERA 2006)

Tabelle 7: Korrelation der Anteile im Unterricht in Deutsch

	Sprechen und Zuhören	Schreiben	Rechtschreibung	Lesen	Sprache untersuchen
Sprechen und Zuhören	1	,028	-,513**	-,351**	-,424**
Schreiben		1	-,209*	-,242**	-,451**
Rechtschreibung			1	-,140	,010
Lesen				1	-,160
Sprache untersuchen					1

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

* Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

In Tabelle 7 sind die Korrelationen zwischen den angegebenen Anteilen der Inhaltsbereiche im Fach Deutsch angegeben. Erwartungsgemäß überwiegen negative Zusammenhänge deutlich, d.h. durch eine umfangreichere Beschäftigung mit einem Inhaltsbereich verbleibt weniger Unterrichtszeit für die anderen Bereiche. Ausgeprägt zeigt sich dieser kompensatorische Effekt für „Lesen“.

Für Mathematik (Tabelle 8) ergibt sich ein etwas anderes Bild. Es wird deutlich, dass die Beschäftigung mit anderen Inhaltsbereichen fast ausschließlich zu Lasten der Zeit für den Bereich „Zahlen und Operationen“ geht.

Tabelle 8: Korrelation der Anteile im Unterricht in Mathematik

	Zahlen und Operationen	Raum und Form	Muster und Strukturen	Größen und Messen	Daten, Häufigkeit und Wahrscheinlichkeit
Zahlen und Operationen	1	-,533**	-,720**	-,494**	-,701**
Raum und Form		1	,293**	-,001	,116
Muster und Strukturen			1	,182*	,468**
Größen und Messen				1	,089
Daten, Häufigkeit und Wahrscheinlichkeit					1

** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

* Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

Es ergeben sich in keinem der beiden Fächer bedeutsame direkte Zusammenhänge zwischen der Leistung und den Anteilen der unterrichtlichen Behandlung der Inhaltsbereiche.

1.4.3. Akzeptanz der Vergleichsarbeiten

In Bezug auf die Akzeptanz der Vergleichsarbeiten wurden die Lehrkräfte gebeten, für die in Tabelle 9 dargestellten Aussagen (vgl. Peek & Nilshon, 2004) ihre Zustimmung auf vier Stufen („1“ stimme gar nicht zu bis „4“ stimme voll zu“) auszudrücken. Im Mittel weisen die Lehrkräfte aus Berlin einen ähnlichen Akzeptanz-Wert auf wie die Lehrerinnen und Lehrer aus den anderen VERA-Ländern ($M = 2,4^*$; $SD = ,64$).

Tabelle 9: Akzeptanz der Vergleichsarbeiten

	M	SD
Vergleichsarbeiten...		
sind für die schulische Qualitätssicherung nützlich.	2,5	0,81
sind für die Arbeit der Lehrkräfte wichtig.	2,5	0,84
Tragen dazu bei, dass man sich an den Schulen mehr bemüht.	2,0	0,82
informieren objektiv darüber, wo eine Klasse bzw. eine Schule steht.	2,4	0,84
Nützen für meine eigentliche Arbeit als Lehrer wenig.	2,6	0,91
erzeugen unnötigen Druck auf die Schulen.	2,8	0,94
Gesamt	2,4*	0,64

* In dem Mittelwert wurden die beiden letzten Items als umgepolte Werte berücksichtigt.

Auch in Bezug auf die Akzeptanz von Vergleichsarbeiten ergeben sich keine bedeutsamen direkten Zusammenhänge mit der Leistung.

1.4.4. 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. Da die Vergleichsarbeiten als externe Evaluation sowohl für Lehrkräfte als auch für Schülerinnen und Schüler vermutlich eine aufregende Situation darstellen, 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. Dazu bieten sich unterschiedliche Zugangsweisen an wie z. B.

- *Inhaltliche Vorbereitung*: Üben von VERA 2004- und 2005-Aufgaben oder Behandlung von VERA-Inhalten im Unterricht
- *Vertrautmachen mit dem Testverfahren*: Vorbereitung auf den Ablauf, Besprechung von Aufgabenformaten und bei VERA eingesetzten Korrekturkriterien bzw. Anforderungen in den Korrekturanweisungen
- *Vermittlung von Teststrategien*: u.a. Tipps zu typischen Aufgabenformaten oder zur Bearbeitungsreihenfolge

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 solches 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.

Um die eingesetzten Vorbereitungsformen zu erfassen, wurde in dem Zentralstichprobenfragebogen eine entsprechende Frage aufgenommen. Tabelle 10 fasst die Häufigkeiten für die unterschiedlichen Formen der Vorbereitung zusammen.

Tabelle 10: Formen der Vorbereitung

Vorbereitung	Berlin		andere VERA-Länder ¹	
	Prozent	N	Prozent	N
gar nicht	26,5%	58	22,0%	202
Inhaltlich	16,0%	35	12,6%	116
Testverfahren	7,3%	16	8,9%	82
Teststrategien	10,5%	23	12,1%	111
inhaltlich + Testverfahren	5,5%	12	5,4%	50
inhaltlich + Teststrategien	9,6%	21	10,7%	98

Vorbereitung	Berlin		andere VERA-Länder¹	
	Prozent	N	Prozent	N
Testverfahren + Teststrategien	5,5%	12	12,8%	118
alle 3 Formen	19,2%	42	15,5%	142
Gesamt	100%	219	100%	919

¹Brandenburg, Bremen, Mecklenburg-Vorpommern, Rheinland-Pfalz (VERA 2006)

Demnach haben 73,5 Prozent der Lehrkräfte ihre Klassen auf die Vergleichsarbeiten vorbereitet. Dem Vergleich zu den anderen Ländern zufolge scheinen die berlinerischen Lehrkräfte den Schwerpunkt insbesondere auf die inhaltliche Vorbereitung bzw. auf die Kombination der drei Formen gelegt zu haben. Vergleichsweise selten wurden die Schülerinnen und Schüler zugleich mit dem Testverfahren vertraut gemacht und in Testbearbeitungsstrategien unterrichtet.

In den folgenden Abbildungen werden die Fähigkeitsniveaueverteilungen von Schülerinnen und Schülern, die auf die Vergleichsarbeit vorbereitet wurden, denen gegenüber gestellt, die laut Angaben der Lehrkräfte nicht vorbereitet wurden.

Vorbereitung in Berlin (Deutsch)

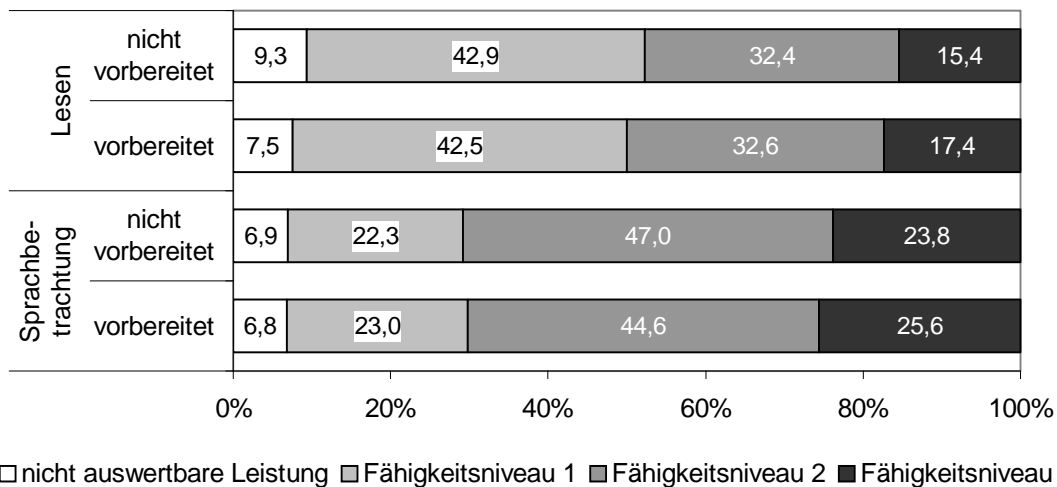


Abbildung 9: Gesamtverteilung der Fähigkeitsniveaus im Fach Deutsch in Abhängigkeit von der Vorbereitung; Angaben in Prozent

Vorbereitung in Berlin (Mathematik)

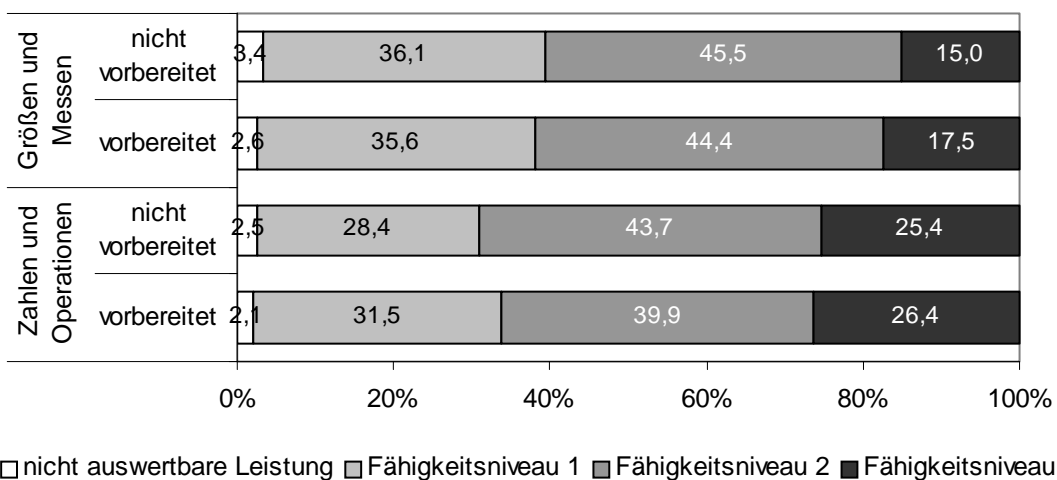


Abbildung10: Gesamtverteilung der Fähigkeitsniveaus im Fach Mathematik in Abhängigkeit von der Vorbereitung; Angaben in Prozent

Berücksichtigt man zusätzlich die unterschiedlichen Vorbereitungsformen bzw. Kombinationen daraus, ergeben sich nur vereinzelt und unsystematisch signifikante Effekte auf die Leistung in den unterschiedlichen Inhaltsbereichen. Daher wird der Übersichtlichkeit halber auf eine ausführliche Darstellung der entsprechenden Ergebnisse an dieser Stelle verzichtet.

1.4.5. Auswertung der Vergleichsarbeiten

Da eine Kooperation bei der Durchführung und Auswertung der Vergleichsarbeiten die Auseinandersetzung mit den Ergebnismeldungen positiv beeinflusst (Koch, Groß Ophoff & Helmke, 2006) und darüber zu Unterrichtsentwicklung angeregt werden kann, wurde im Lehrerfragebogen auch erfragt, ob die Auswertung der Vergleichsarbeiten in Kooperation mit Kolleg/innen erfolgte. Zusätzlich wurde erfasst, wie viel Zeit für die Auswertung der Bearbeitungen benötigt wurde.

Tabelle 11 belegt, dass eine schulinterne Auswertung im Team, wie sie z.B. in der Durchführungshandreichung angeregt wird, von etwa 30 Prozent der Lehrkräfte durchgeführt wurde. Auffällig ist hierbei, dass in Berlin vergleichsweise weniger Lehrkräfte im Rahmen der Auswertung beider Fächer kooperiert haben..

Für die Auswertung der Vergleichsarbeiten im Fach Deutsch wurde insgesamt etwas mehr Zeit benötigt. Diesbezüglich ergeben sich nur geringe Unterschiede zwischen Berlin und den anderen VERA-Ländern.

Tabelle 61: Kooperation im Rahmen der Vergleichsarbeiten; Angaben in Prozent

	Berlin		andere VERA-Länder ¹	
Auswertung im Team				
• Deutsch	30,2%	n = 42	41,3%	n = 290
• Mathematik	25,4%	n = 36	37,8%	n = 249
Zeit für Korrektur				
• Deutsch	6,4h	SD = 5,02	6,9 h	SD = 5,77
• Mathematik	4,9 h	SD = 4,56	5,6 h	SD = 5,40

¹Brandenburg, Bremen, Mecklenburg-Vorpommern, Rheinland-Pfalz (VERA 2006)

2 Literatur

- Baumert, J., Artelt, C., Carstensen, C. H., Sibberns, H. & Stanat, P. (2002). Untersuchungsgegenstand, Fragestellungen und technische Grundlagen der Studie. In Deutsches PISA - Konsortium (Hrsg.), *PISA 2000 - Die Länder der Bundesrepublik Deutschland im Vergleich* (S. 11-38). Opladen: Leske + Budrich.
- Helmke, A. (2004). Von der Evaluation zur Innovation: Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. *Seminar*, 2, 90-112.
- Helmke, A. (2006). Was wissen wir über guten Unterricht? *Pädagogik (Große Serie 2006: Forschung-Schule-Unterricht. Befunde und Konsequenzen)*, 2, 42-45.
- 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.), *Das Projekt Markus: Mathematikgesamterhebung 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).
- 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), S. 567-600.
- Helmke, A. & Weinert, F. E. (1997). Unterrichtsqualität und Leistungsentwicklung. Ergebnisse aus dem SCHOLASTIK-Projekt. In F. E. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S. 241-251). Weinheim: Psychologie Verlags Union.
- Hosenfeld, I., Helmke, A., Ridder, A. & Schrader, F.-W. (2001). Eine mehrbenenanalytische 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.
- Isaac, K., Eichler, W., Hosenfeld, I. & Groß Ophoff, J. (2006, September). *Sprache und Sprachgebrauch untersuchen als Gegenstand von Vergleichsarbeiten in der Grundschule*. Beitrag präsentiert bei 68. Tagung der Arbeitsgruppe der Empirischen Bildungsforschung (AEPF) Sektion Empirische Bildungsforschung (DGfE). Ludwig-Maximilians-Universität München.
- 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.
- Koch, U., Groß Ophoff, J. & Helmke, A. (2006, September). *Bedingungen der Rezeption von Ergebnisrückmeldungen - am Beispiel der Evaluation von VERA 2005*. Beitrag präsentiert bei 68. Tagung der Arbeitsgruppe der Empirischen Bildungsforschung (AEPF) Sektion Empirische Bildungsforschung (DGfE), Ludwigs-Maximilian-Universität München.
- Peek, R. & Nilshon, I. (2004). *Schulrückmeldungen von Schulleistungsstudien am Beispiel des QuaSUM-Projektes. Zwei Untersuchungen zur Wirksamkeit*. Potsdam: Land Brandenburg, Ministerium für Bildung, Jugend und Sport.
- Projektgruppe VERA-Deutsch. (2006). *Didaktische Erläuterungen "Sprache und Sprachgebrauch untersuchen" und "Leseverständnis" (VERA 2006)*, [PDF].

- Universität Koblenz-Landau (Campus Landau). Information: www.uni-landau.de/vera/ (Material) [01.02.2007].
- Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests* (Studies in Mathematical Psychology). Copenhagen: Nielsen & Lydiche.
- Schrader, F.-W. (2006). *Diagnostische Kompetenz von Eltern und Lehrern*. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (2. Aufl., S. 68-71). Weinheim: Psychologie Verlags Union. → neuere Auflage, wenn Buch gekommen ist!
- Schrader, F.-W., Helmke, A., Hosenfeld, I., Halt, A. C. & Hochweber, J. (2006). Komponenten der Diagnosegenauigkeit von Lehrkräften: Ergebnisse aus Vergleichsarbeiten in der Grundschule. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF* (S. 265-278). Münster: Waxmann.
- 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 Bildungsstandard der Jugendlichen in Deutschland. Ergebnisse des zweiten internationalen Vergleichs* (S. 211-222). Münster: Waxmann Verlag.

3 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 Aggregation von der Individualebene (Angaben einzelner Schüler) auf die Klassenebene sind auch Aggregationen 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 verschiedene 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 \rightarrow 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 \rightarrow Korrelationskoeffizienten bestimmt.

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

IGLU, Internationale Grundschul-Lese-Untersuchung. Deutsche Teilstudie der Studie PIRLS (Progress in International Reading Literacy Study) der \rightarrow 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 \rightarrow Skala zusammengefasst.

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

Korrelation, korrelieren, Korrelationskoeffizient, eine Korrelation ist ein statistisches Maß für den Grad des linearen Zusammenhangs zwischen zwei \rightarrow Variablen (Merkmalen) x und y . Für \rightarrow 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 \rightarrow erklärte Varianz.

Kriterium(s-Variable)

Ein anderer Begriff für \rightarrow abhängige Variable, siehe auch \rightarrow 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 \rightarrow Mittelwert (engl. Mean).

$M \pm SD$, Wertebereich, der durch eine Streuungseinheit ($\rightarrow SD$) 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 \rightarrow 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 \rightarrow Mittelwert und die \rightarrow 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 \pm eine Standardabweichung liegen etwa 68 Prozent der Fälle; im Bereich Mittelwert \pm 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.