PISA 2003

Der Bildungsstand der Jugendlichen in Deutschland –
Ergebnisse des zweiten internationalen Vergleichs

Manfred Prenzel, Jürgen Baumert, Werner Blum,
Rainer Lehmann, Detlev Leutner, Michael Neubrand,
Reinhard Pekrun, Hans-Günter Rolff, Jürgen Rost
und Ulrich Schiefele (Hrsg.)

PISA-Konsortium Deutschland

PISA 2003

Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs

Waxmann
Münster/New York
München/Berlin

Bibliografische Informationen Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

ISBN 3-8309-1455-5

Postfach 8603, D-48046 Münster
Waxmann Publishing Co.
P.O. Box 1318, New York, NY 10028, USA

www.waxmann.com
E-Mail: info@waxmann.com

Buchumschlag: Christian Averbeck, Münster
Satz: Stoddart Satz- und Layoutservice, Münster
Druck: Runge GmbH, Cloppenburg
Gedruckt auf alterungsbeständigem Papier DIN 6738

Inhalt

Vorwort der Präsidentin der Kultusministerkonferenz

Wie ist die Qualität des deutschen Bildungswesens einzuschätzen? Sind die Schülerinnen und Schüler gut vorbereitet auf die Herausforderungen der Zukunft? Diese Fragen hat die Kultusministerkonferenz mit ihrem Konstanzer Beschluss vom Oktober 1997 aktiv aufgegriffen. Sie wollte die Leistungen des deutschen Schulsystems auf den Prüfstand stellen und anhand empirischer Ergebnisse und Analysen die Qualitätsentwicklung im Bildungswesen zielgerichtet voranbringen.

Die Länder in der Bundesrepublik Deutschland beteiligen sich an den drei Erhebungsrunden des OECD-Programms PISA, die in den Jahren 2000, 2003 und 2006 stattfinden. PISA stellt Daten zu Leistungsergebnissen und ihren Bedingungen im internationalen Vergleich zur Verfügung, die für Entscheidungen zur Verbesserung der Bildungssysteme herangezogen werden können. PISA erfasst die Lesekompetenz, die mathematische und naturwissenschaftliche Grundbildung und bezieht auch fächerübergreifende Kompetenzen von Fünfzehnjährigen ein. Die Kultusministerkonferenz führt zudem Leistungsuntersuchungen in der neunten Klasse in den Fächern Deutsch und Englisch (DESI) durch und beteiligt sich an der internationalen Grundschulstudie PIRLS/IGLU.

Die bisher vorliegenden Ergebnisse internationaler Schulleistungsuntersuchungen haben deutlich gemacht, vor welchen zentralen Herausforderungen bildungspolitisches Handeln in den Ländern der Bundesrepublik Deutschland steht. Auf der Grundlage der Ergebnisse von PISA 2000 hat die Kultusministerkonferenz im Dezember 2001 sieben vorrangige Handlungsfelder festgelegt:

- Verbesserung der Sprachkompetenz bereits im vorschulischen Bereich,
- bessere Verzahnung von vorschulischem Bereich und Grundschule mit dem Ziel einer frühzeitigen Einschulung,
- Verbesserung der Grundschulbildung und durchgängige Verbesserung der Lesekompetenz und des grundlegenden Verständnisses mathematischer und naturwissenschaftlicher Zusammenhänge,
- wirksame Förderung bildungsbenachteiligter Kinder, insbesondere auch der Kinder und Jugendlichen mit Migrationshintergrund,
- konsequente Weiterentwicklung und Sicherung der Qualität von Unterricht und Schule auf der Grundlage von verbindlichen Standards sowie eine ergebnisorientierte Evaluation,
- Verbesserung der Professionalität der Lehrtätigkeit, insbesondere im Hinblick auf diagnostische und methodische Kompetenz als Bestandteil systematischer Schulentwicklung und

- Weiterentwicklung von schulischen und außerschulischen Ganztagsangeboten mit dem Ziel erweiterter Bildungs- und Fördermöglichkeiten, insbesondere für Schülerinnen und Schüler mit Bildungsdefiziten oder besonderen Begabungen.

Die Länder haben vielfältige Aktivitäten in diesen Handlungsfeldern ergriffen, die auch im ersten Bildungsbericht der Kultusministerkonferenz dargestellt sind. Von zentraler Bedeutung sind dabei die gemeinsamen Bildungsstandards und die Vereinbarungen zu deren regelmäßiger Überprüfung.

Die Kultusministerkonferenz hat zwischenzeitlich das Institut zur Qualitätsentwicklung im Bildungswesen als An-Institut an der Humboldt-Universität zu Berlin gegründet, um die Normierung und Überprüfung der Standards länderübergreifend zu gewährleisten. Sie hat gemeinsam mit dem Bund die Bildungsberichterstattung etabliert, die – aus einer wissenschaftlich unabhängigen Perspektive – die Kontext-, Prozess- und Wirkungsfaktoren von Bildungssystemen regelmäßig in den Blick nehmen wird.

Im Zentrum der zweiten PISA-Erhebung im Jahr 2003 stand die Mathematik. Die nun vorliegenden Ergebnisse sind auch geeignet, die nach TIMSS seit dem Jahr 1997 eingeleiteten Maßnahmen zur Verbesserung der mathematisch-naturwissenschaftlichen Bildung daraufhin zu überprüfen, ob erste Wirkungen erkennbar sind, wo wir unsere Anstrengungen verstärken müssen und inwieweit neue Gewichtungen vorgenommen werden müssen. Die Ergebnisse sind auch für die Standardentwicklung und -überprüfung von Bedeutung. Die Befunde werden außerdem in den kommenden Jahren eine der Grundlagen für die Bildungsberichterstattung über den schulischen Bereich sein.

Bereits PISA 2000 hat bestätigt, dass die gesellschaftliche Wertschätzung des Lernens und die gemeinsame Verantwortung für Bildung zentral für den Erfolg von Bildungssystemen sind. Alle an Bildung Beteiligten müssen intensiv an dem – nach den Erfahrungen erfolgreicher Länder – langen Prozess der Qualitätsentwicklung mitarbeiten und sich gemeinsam sowohl für die gegenwärtigen Bildungsleistungen als auch für künftige Verbesserungen einsetzen. Dabei ist noch stärker darauf zu achten, dass eine wirksamere individuelle Förderung erfolgt und die Bildungschancen aller gewahrt werden.

Bonn, im Dezember 2004

Doris Ahnen
Präsidentin der Ständigen Konferenz der Kultusminister der Länder

1 PISA 2003 – eine Einführung

Manfred Prenzel, Barbara Drechsel,
Claus H. Carstensen und Gesa Ramm

Der vorliegende Band berichtet über die internationalen Ergebnisse der zweiten PISA-Erhebung. Er beantwortet die Frage, wie Deutschland in PISA 2003 abgeschnitten hat. Dieser nationale Bericht unterscheidet sich von den zeitgleich erscheinenden internationalen Berichten der *Organisation für wirtschaftliche Zusammenarbeit und Entwicklung* (OECD, 2004a; OECD, 2004b) durch eine andere Perspektive: Vom Bericht des PISA-Konsortiums Deutschland erwartet man, dass er die internationalen Befunde aufgreift und diese aus einer nationalen Sicht bespricht. Selbstverständlich muss der nationale Bericht die Ergebnisse für Deutschland ausführlicher und differenzierter behandeln als der internationale Report. Außerdem nutzen wir die Gelegenheit, erste Befunde aus Zusatzerhebungen in Deutschland vorzustellen, mit denen die Ergebnisse des internationalen Vergleichs ergänzt und besser eingeordnet werden können. Weitere nationale Berichte folgen.

Das Einleitungskapitel beschreibt die Ziele und den Untersuchungsansatz des *Programme for International Student Assessment* (PISA) der OECD. Es informiert über zusätzliche Erhebungen, die in Deutschland durchgeführt wurden. Das Kapitel führt aber auch in Begriffe, Methoden und Darstellungsformen ein, die in den folgenden Kapiteln häufig verwendet werden.[1]

1.1 Die Erhebung

Das *Programme for International Student Assessment* untersucht, wie gut die jungen Menschen in den teilnehmenden Staaten auf Herausforderungen der Wissensgesellschaft vorbereitet sind. Zielgruppe des Programms sind die fünfzehnjährigen Jugendlichen, die sich in zahlreichen Staaten dem Ende der Pflichtschulzeit nähern. PISA konzentriert die Erhebungen auf zentrale und grundlegende Kompetenzen, die für die individuellen Lern- und Lebenschancen ebenso bedeutsam sind wie für die gesellschaftliche, politi-

1 Alle Abkürzungen, die in diesem Buch verwendet werden, sind in einem Abkürzungsverzeichnis im Anhang erklärt.

sche und wirtschaftliche Weiterentwicklung. Mit dieser Ausrichtung interessiert es besonders, ob die Jugendlichen ihre Kompetenzen flexibel und situationsgerecht bei der Lösung von Aufgaben nutzen können, die vielfältige alltags-, ausbildungs- wie berufsbezogene Anforderungen repräsentieren. Die Kompetenzen verwendet PISA als Kriterium, um die Leistungsfähigkeit von Bildungssystemen zu vergleichen und zu beurteilen.

PISA ist der zentrale Teil eines umfassenden Indikatorensystems der OECD. Es dient dazu, die Mitgliedstaaten über Stärken und Schwächen ihrer Bildungssysteme zu informieren. PISA gestattet es, Bildungsergebnisse nach internationalen Maßstäben zu beurteilen. Die Studie liefert Bezugspunkte, an denen Möglichkeiten für eine Weiterentwicklung der Bildungssysteme geklärt werden können. In diesem Sinne dient PISA einem *Benchmarking* im Bildungsbereich. Da die Erhebungen regelmäßig in einem Abstand von drei Jahren durchgeführt werden, informieren sie auf längere Sicht darüber, inwieweit ergriffene Maßnahmen die angestrebten Wirkungen erreichen konnten. Unter diesem Aspekt kann PISA als Verfahren zur Dauerbeobachtung von Bildungssystemen im Sinne eines *Bildungsmonitoring* beitragen. PISA stellt den Regierungen als Auftraggebern der Studie empirisch fundierte Erkenntnisse in Aussicht, die ihnen helfen sollen, ihre Bildungssysteme auf der Basis umfassender und zuverlässiger Daten zu steuern.

Befunde über die Stärken und Schwächen des jeweiligen Bildungssystems im internationalen Vergleich sind jedoch nicht nur für Bildungsadministration und Bildungspolitik bedeutsam. Auch für die Öffentlichkeit, die Eltern und vor allem die Personen, die sich professionell mit Schule und Bildungsfragen befassen, ist es wichtig zu erfahren, inwieweit junge Menschen im Verlauf der Schulzeit auf zukünftige Anforderungen vorbereitet werden (vgl. Abbildung 1.1).

Das Anliegen von PISA richtet sich deshalb *nicht* darauf, Staaten in der Art eines olympischen Leistungswettbewerbs zu vergleichen und mit Rangplätzen auszuzeichnen. Die bei PISA erfassten Kompetenzen beschreiben ja keine Endzustände eines Trainings für einen Leistungsvergleich, sondern vielmehr entscheidende Voraussetzungen für weiterführende Lernprozesse mit herausragender Bedeutung für das Individuum und die Gesellschaft. Mit Blick auf die langfristigen Konsequenzen gilt es deshalb, Erkenntnisse über Probleme, notwendige und mögliche Verbesserungen im Bildungsbereich der Teilnehmerstaaten zu erhalten. Deshalb gewinnen bei PISA alle Staaten, unabhängig davon, wie sie im internationalen Vergleich abschneiden, durch ihre Teilnahme nützliches Wissen. Der internationale Vergleich erweist sich dabei als sehr hilfreiches Verfahren, um relative Stärken und Schwächen zu erkennen. In diesem Band greifen wir auf zusätzliche Bezugsmaßstäbe zurück, um Bildungsergebnisse zu beurteilen. So werden die Testergebnisse auch an inhaltlichen Gütekriterien der Kompetenzentwicklung eingeordnet. Eine weitere Beurteilungsperspektive ergibt sich zum Beispiel aus einem Vergleich der Ergebnisse, die bei den Erhebungsrunden in den Jahren 2000 und 2003 erzielt wurden.

Abbildung 1.1: PISA 2003 im Überblick

Inhaltsbereiche

- PISA untersucht die Kompetenzen von fünfzehnjährigen Schülerinnen und Schülern in den Bereichen Mathematik (Schwerpunktgebiet 2003), Lesen und Naturwissenschaften. Die Erhebungen beruhen auf einem Testansatz, der in allen Inhaltsbereichen zwischen Konzepten, Prozessen und Situationen beziehungsweise Kontexten unterscheidet. Die Testkonzeption ist an einer Vorstellung von lebenslangem Lernen orientiert und betont das Verstehen und die flexible, situationsgerechte Anwendung des Wissens.
- Neben bereichsspezifischen werden bereichsübergreifende Kompetenzen untersucht. Im Zentrum dieser so genannten Cross-Curricular-Competencies steht 2003 das Problemlösen. Neben den Tests zur Problemlösekompetenz ergänzen Erhebungen zu Lernstrategien, Lernmotivation und zur Vertrautheit mit Informationstechnologien den fächerübergreifenden Untersuchungsbereich.

Erhebungsverfahren

- Die Erhebung erfolgte an einem Testtag an der Schule, in Gruppen und unter Aufsicht.
- Die Tests bestehen aus Mehrfachauswahlfragen (Multiple-Choice). Sie werden mit Fragen kombiniert, die von den Schülerinnen und Schülern mit eigenen Worten oder Darstellungen beantwortet werden müssen (offene Fragen). Die Einzelfragen (Items) sind thematisch zu Aufgaben gruppiert, die sich auf eine durch einen kurzen Text beschriebene Situation beziehen.
- Die reine Testzeit für jeden Schüler beziehungsweise jede Schülerin betrug zwei Stunden.
- Die Schülerinnen und Schüler erhielten Testhefte mit unterschiedlichen Aufgabenpaketen. Auf diese Weise kann insgesamt Itemmaterial für mehr als sechseinhalb Stunden Testzeit eingesetzt werden. Der größte Teil dieser Testzeit (3,5 Stunden) entfiel 2003 auf das Schwerpunktgebiet Mathematik.
- Die Schülerinnen und Schüler bearbeiteten weiterhin einen Fragebogen (30 Minuten), der sich auf ihre Herkunft und Umgebung, ihre Lerngewohnheiten und Motivation bezieht. Optional ist ein weiterer Fragebogen zur Vertrautheit mit Computern und zu Vorstellungen über die eigene Bildungskarriere (15 Minuten).
- Die Schulleitungen wurden gebeten, einen Fragebogen zu Merkmalen ihrer Schule (z.B. Ressourcen, Qualifikation der Lehrkräfte, Schulklima) auszufüllen.

Stichprobe

- International wurden in den 41 an PISA teilnehmenden Staaten (30 OECD-Staaten und 11 Partnerländer) ca. 250 000 Schülerinnen und Schüler getestet.
- Die teilnehmenden Schulen und die teilnehmenden Schülerinnen und Schüler wurden mit einem Zufallsverfahren gezogen.

Ergebnisse

- Ein Profil von wichtigen Kompetenzen, über die Fünfzehnjährige verfügen
- Informationen über Zusammenhänge zwischen Kompetenzen und Merkmalen der sozialen Herkunft
- Erste Einschätzungen der Veränderungen in den Kompetenzen zwischen den Erhebungen in den Jahren 2000 und 2003

Optionen für Ergänzungen und Erweiterungen

- Auch 2003 konnten internationale Optionen (z.B. ergänzende Fragebögen) gewählt werden.
- Die internationale Erhebung konnte durch nationale Komponenten (erweiterte Stichproben, zusätzliche Testtage und Erhebungsverfahren) ergänzt werden.

Ausblick

- Der Schwerpunkt der nächsten Erhebung im Jahr 2006 wird auf dem Bereich Naturwissenschaften liegen. Im Jahr 2009 soll dann wieder die Lesekompetenz im Zentrum stehen.
- In zukünftigen Erhebungen sollen zumindest zum Teil computerbasierte Erhebungen stattfinden, um das Fähigkeitsspektrum besser abbilden zu können.

Teilnehmende Staaten (Staaten, die an PISA 2000 und 2003 teilnahmen, sind kursiv)
OECD: Australien, *Belgien*, *Dänemark*, *Deutschland*, *Finnland*, *Frankreich*, *Griechenland*, *Island*, *Irland*, *Italien*, *Japan*, *Kanada*, *Korea*, *Luxemburg*, *Mexiko*, *Niederlande*, *Neuseeland*, *Norwegen*, *Österreich*, *Polen*, *Portugal*, *Schweden*, *Schweiz*, Slowakische Republik, *Spanien*, *Tschechische Republik*, Türkei, *Ungarn*, *Vereinigtes Königreich*, *Vereinigte Staaten*
Partnerstaaten: Brasilien, Hongkong-China, Indonesien, *Lettland*, *Liechtenstein*, Macao-China, *Russische Förderation*, Serbien und Montenegro, Thailand, Tunesien, Uruguay

1.2 Der theoretische Rahmen: Kompetenzbereiche und Hintergrundmerkmale

Die theoretischen Grundlagen für PISA 2003 wurden von Expertengruppen für die verschiedenen Erhebungsbereiche in einer Rahmenkonzeption dargelegt (OECD, 2003). Diese Konzeption begründet und beschreibt die Anforderungen an die Konstruktion der Testverfahren und Fragebögen. Gegenüber der Rahmenkonzeption der ersten Erhebungswelle (Baumert et al., 2001; OECD, 1999) sind einige Differenzierungen und Ergänzungen zu verzeichnen, die insbesondere der veränderten Schwerpunktsetzung (Mathematik sowie Problemlösen) Rechnung tragen. Die Erhebungskonzeption ist jedoch in der Struktur gleich geblieben, so dass die Ergebnisse aus den verschiedenen Testwellen verglichen werden können.

Eine Vorstellung von Grundbildung: Als Bezugspunkt für die Testentwicklung dient bei PISA eine Vorstellung von Grundbildung, die im Englischen als *Literacy* bezeichnet wird. Dieser Begriff wird seit geraumer Zeit in der internationalen fachdidaktischen Diskussion verwendet, um Ansprüche an eine *Grundbildung für alle* zu charakterisieren (z.B. American Association for the Advancement of Science (AAAS), 1993; Kirsch, 1995; National Council of Teachers of Mathematics (NCTM), 2000): Literacy im (engeren) Sinne einer Lesekompetenz befähigt, an einer Kultur teilzuhaben, deren Wissen in Texten vorliegt. Das Beispiel Lesekompetenz zeigt die Tragweite einer „kulturellen Teilhabe", die umfassend die persönlichen Handlungsmöglichkeiten im Alltag, Beruf und gesellschaftlichen Leben betrifft. Auf ähnliche Weise können grundlegende mathematische und naturwissenschaftliche Kompetenzen im (übertragenen) Sinne einer „Literacy" bestimmt und beschrieben werden. Sie sind erforderlich, um sich an der aktuellen Kultur beteiligen zu können, die heute ebenfalls stark durch Mathematik, Naturwissenschaften und Technik geprägt ist. Die OECD spricht damit den Kompetenzbereichen Mathematik, Lesen und Naturwissenschaften eine Schlüsselstellung für die gesellschaftliche Teilhabe und Weiterentwicklung zu.

Die PISA zugrunde liegende Bildungskonzeption orientiert sich damit zunächst an der *Funktion* von Kompetenzen im kulturellen und gesellschaftlichen Zusammenhang. So wird die Ausarbeitung der Rahmenkonzeption bestimmt durch die Frage, welches Konzeptwissen („Wissen, dass") und welches Prozesswissen („Wissen, wie") für die kulturelle Teilhabe bedeutsam sind und in welchen Situationen entsprechendes Wissen genutzt werden sollte. Ein weiterer Gesichtspunkt kommt hinzu: Um einzugrenzen, welche Kompetenzen für Fünfzehnjährige in Hinblick auf zukünftige Herausforderungen wichtig werden, rückt PISA die *Voraussetzungen für ein weiterführendes Lernen* beziehungsweise für eine weiterführende Auseinandersetzung mit diesen Kulturbereichen in den Blickpunkt. Hier gilt es zu klären, welches begriffliche Verständnis und welches Prozesswissen gute Chancen bieten, neue Informationen zu verstehen, einzuordnen und damit weiter zu lernen. PISA bemüht sich besonders darum, das für die Kompetenzbereiche grundlegende und anschlussfähige Wissen zu identifizieren und in Aufgaben umzusetzen.

Diese Gesichtspunkte begrenzen die Rahmenkonzeption und eröffnen zugleich einen pragmatischen Weg, theoretisch anspruchsvolle Bildungsergebnisse empirisch zu erfassen und zwischen den teilnehmenden Staaten zu vergleichen. In Anbetracht einer großen internationalen curricularen Vielfalt verzichtet PISA auf eine enge Orientierung an den Lehrplänen. Die Studie konzentriert sich vielmehr auf eine Untersuchung von Kompetenzen in Schlüsselbereichen, die als *notwendige Voraussetzungen für weiterführende Lern- und Bildungsprozesse* angesehen werden.

PISA hat in das Spektrum der zu untersuchenden Kompetenzen von Anfang an *bereichsübergreifende Kompetenzen* mit einbezogen. Die Erhebungsrunde 2003 lenkt hier die Aufmerksamkeit auf das Lösen von Problemen, die typisch für viele Situationen im alltäglichen Leben sind. Sie können nicht einem Schulfach oder einer Disziplin zugeordnet werden und verlangen eine gründliche Analyse der Anforderungen und intelligente Verknüpfung von Wissensbeständen. Als Aspekte fächerübergreifender Kompetenzen werden außerdem Strategien des Lernens und der motivationalen Steuerung betrachtet. Auch Einstellungen zum Lernen und zu bestimmten Gegenstands- beziehungsweise Kulturbereichen werden bei PISA als wichtige Ziele schulischer Bildungsprozesse betrachtet. Der Testansatz von PISA beschränkt sich damit keineswegs auf drei Kernbereiche, sondern erfasst ein relativ breit gefasstes Spektrum von Kompetenzen.

Die Rahmenkonzeption für PISA (OECD, 2003), die diese Ansprüche für die verschiedenen Bereiche umzusetzen versucht, beschreibt damit einen Testansatz, der wichtige Aspekte von *Grundbildung* (im Sinne etwa der deutschsprachigen Diskussion) aufgreift (vgl. Tenorth, 2004). Grundbildungskonzeptionen zeichnen sich dadurch aus, dass sie Ansprüche formulieren, die möglichst von allen Schülerinnen und Schülern erreicht werden sollten. Auch diesem Anspruch wird die internationale Vorstellung von Literacy gerecht. Freilich muss die Testkonzeption auch in der Lage sein, unterschiedliche Niveaus in der Kompetenzentwicklung abzubilden, die von fünfzehnjährigen Schülerinnen und Schülern erreicht werden können. Damit können auch Aussagen über Gruppen getroffen werden, die sich durch besonders niedrige oder hohe Kompetenz auszeichnen. Da sich die Vorstellungen von Literacy und Grundbildung in vielfacher Hinsicht überschneiden, werden wir in diesem Band die Begriffe synonym verwenden.

Kompetenz als Potential: PISA vergleicht das Grundbildungsniveau, das in den verschiedenen nationalen Bildungssystemen über den Verlauf der Schulzeit erreicht wird. Damit das angeeignete Wissen beim Weiterlernen oder unter Anforderungen außerhalb der Schule genutzt werden kann, muss es in unterschiedlichen Situationen aktiviert und auf die jeweiligen Anforderungen bezogen werden. Deshalb unterscheidet sich PISA von herkömmlichen Verfahren, Schul- und Buchwissen abzufragen. Die Jugendlichen werden vielmehr mit vielfältigen, realitätsnahen Aufgaben konfrontiert, die eine flexible, auf die besondere Situation angepasste Anwendung ihres Wissens verlangen. Die Tests zielen darauf ab, die Möglichkeiten der Schülerinnen und Schüler auszuloten, in einem bestimmten Gebiet Anforderungen zu bewältigen, Probleme zu lösen und weiter zu lernen. Letztlich soll also das entsprechende *Potential* der Schülerinnen und Schüler erfasst werden. Dieses Potential wollen wir im vorliegenden Band als *Kompetenz* be-

zeichnen, auch wenn dieser Begriff in der wissenschaftlichen Literatur nicht immer einheitlich gebraucht wird (Klieme et al., 2003; Weinert, 1999). Aus den Antworten auf die Testfragen (aus der Testleistung oder „Performanz") zu einem bestimmten Teilgebiet wird also die mathematische oder naturwissenschaftliche Kompetenz beziehungsweise die Lese- oder Problemlösekompetenz erschlossen. Der Begriff „Kompetenz" wird dem Anliegen von PISA, die Vorbereitung auf Anforderungen zu erfassen, sehr viel besser gerecht als die Bezeichnung „Schulleistung". Freilich waren den Expertengruppen die üblichen curricularen Schwerpunktsetzungen durchaus bewusst, als sie Testanforderungen und Kompetenz im Rahmen einer Grundbildungskonzeption spezifizierten, die das anschlussfähige Wissen und die kulturelle Teilhabe betont.

Struktur der Kompetenzerhebungen: Die bei PISA 2000 verwendete Grundstruktur bestimmt auch den Aufbau der Testkonzeptionen der aktuellen Erhebung. Der Testansatz für Problemlösen musste 2003 neu erarbeitet, der für den Bereich Mathematik als Schwerpunktgebiet ausgebaut werden. Im Bereich Naturwissenschaften sind einige kleinere Änderungen zu verzeichnen. Die Testkonzeptionen unterscheiden jeweils drei Aspekte oder Dimensionen:

- die *Inhalte* oder die Konzepte („Wissen, dass"), über die Schülerinnen und Schüler verfügen sollten;
- die *Prozesse* und Prozeduren („Wissen, wie"), die von den Schülerinnen und Schülern verstanden und beherrscht werden sollten;
- die *Situationsklassen* und Kontexte, in denen die Schülerinnen und Schüler ihr Wissen anwenden können sollten.

Überblick über die Kompetenzbereiche: Die Abbildung 1.2 greift diese Unterscheidung auf, um die wichtigsten Merkmale für die in PISA 2003 untersuchten Kompetenzbereiche zu skizzieren.

Die in Abbildung 1.2 vorgestellten Definitionen lassen die Betonung von Kompetenz im Sinne eines Fähigkeitspotentials ebenso erkennen wie die Orientierung auf die kulturelle Teilhabe im Sinne von Literacy. Die Testkonzeption für das Schwerpunktgebiet *Mathematik* orientiert sich unter anderem am Ansatz einer „realistischen Mathematik" (Freudenthal, 1977) und unterstreicht den funktionalen Gebrauch von Mathematik als Werkzeug, um mathematische Probleme in unterschiedlichen Zusammenhängen zu erkennen, zu formulieren und zu lösen. Im Bereich *Naturwissenschaften* wird unter anderem auf Vorstellungen von „Scientific Literacy" (z.B. Bybee, 1997) zurückgegriffen, die das begriffliche Verständnis, das naturwissenschaftliche Herangehen und die Berücksichtigung von Evidenz bei der Betrachtung von Phänomenen oder Problemen hervorheben. Die Testkonzeption zum Bereich *Lesen* stellt das Ermitteln von Informationen, das Interpretieren und Reflektieren von Texten in den Vordergrund (vgl. Kirsch, 1995).

Die in Abbildung 1.2 enthaltenen Angaben zu den Inhalts- und Prozessaspekten vermitteln eine erste Vorstellung davon, was PISA in den verschiedenen Bereichen als bedeutsam für Grundbildung erachtet. Die einzelnen Testkonzeptionen werden in den entsprechenden Kapiteln dieses Bandes ausführlicher erläutert.

Abbildung 1.2: Überblick über die Testkonzeptionen (vgl. OECD, 2003)

	Mathematik	**Naturwissenschaften**	**Lesen**
Definition	Die Fähigkeit einer Person, die Rolle zu erkennen und zu verstehen, die Mathematik in der Welt spielt, fundierte mathematische Urteile abzugeben und sich auf eine Weise mit der Mathematik zu befassen, die den Anforderungen des gegenwärtigen und künftigen Lebens als konstruktivem, engagiertem und reflektierendem Bürger entspricht.	Die Fähigkeit, naturwissenschaftliches Wissen anzuwenden, naturwissenschaftliche Fragen zu erkennen und aus Belegen Schlussfolgerungen zu ziehen, um Entscheidungen zu verstehen und zu treffen, welche die natürliche Welt und die durch menschliches Handeln an ihr vorgenommenen Veränderungen betreffen.	Die Fähigkeit geschriebene Texte zu verstehen, zu nutzen und über sie zu reflektieren, um eigene Ziele zu erreichen, das eigene Wissen und Potential weiterzuentwickeln und am gesellschaftlichen Leben teilzunehmen.
„Inhalte“	Übergreifende Ideen: – Quantität – Raum und Form – Veränderung und Beziehungen – Unsicherheit	Bereiche naturwissenschaftlichen Wissens und naturwissenschaftlicher Konzepte wie zum Beispiel – Kraft und Bewegung – Artenvielfalt – physiologische Veränderungen	Textarten: – kontinuierliche Texte, z.B. Erzählungen, Beschreibungen, Argumentationen – nichtkontinuierliche Texte, z.B. Diagramme, Formulare und Listen
„Prozesse“	Die Kompetenzcluster definieren mathematische Fertigkeiten, unterteilt in drei Niveaubereiche: – Reproduktion (Ausführen einfacher Standardtätigkeiten, die direkt aus der Situation entnommen werden können) – Verbindungen (überschaubare Tätigkeiten, welche bereits mehrere Schritte oder die Verknüpfung mehrerer Aufgabenelemente erfordern) – Reflexion (komplexe Tätigkeiten, Verallgemeinerungen oder Reflexionen gefordert)	Die Fähigkeit, naturwissenschaftliches Wissen und Verständnis zu nutzen, naturwissenschaftliche Befunde zu erheben, zu interpretieren und nach ihnen zu handeln. Sie umfasst: – das Beschreiben, Erklären und Vorhersagen naturwissenschaftlicher Phänomene – das Verstehen wissenschaftlicher Forschung – das Interpretieren wissenschaftlicher Befunde und Schlussfolgerungen	Leseaufgaben: – Informationen heraussuchen – eine Interpretation entwickeln und/oder – über Inhalt und Form eines Textes reflektieren PISA untersucht eher das „Lesen, um zu lernen“ als das „Lesen lernen“ selbst. Daher werden die grundlegendsten Lesefertigkeiten nicht getestet.
„Situationen“	Das Rahmenkonzept unterscheidet folgende Situationen: – persönliche – ausbildungs- und berufsbezogene – gesellschaftsbezogene – wissenschaftliche	Naturwissenschaftliche Anwendungen in den Bereichen – Leben und Gesundheit – Erde und Umwelt – Naturwissenschaft in Technologien	Der Text dient – privaten Zwecken (z.B. Brief) – öffentlichen Zwecken (z.B. offizielles Dokument) – der beruflichen Qualifikation (z.B. Berichte)

Die im Feld der *fächerübergreifenden Kompetenzen* angesiedelte Erhebung zum *Problemlösen* zielt auf die Fähigkeit, realistische Problemstellungen zu bearbeiten und zu lösen, in denen ein Lösungsweg nicht unmittelbar erkennbar ist und Wissen aus mehreren Domänen genutzt werden muss. Die Problemtypen „Entscheidungen treffen", „Systeme analysieren" und „Fehler entdecken" messen dem analytischen Problemlösen besondere Bedeutung bei (OECD, 2003; OECD, 2004b).

Dem Bereich der fächerübergreifenden Kompetenzen sind weitere Erhebungen zum *Selbstregulierten Lernen*, zur *Lernmotivation* und zu *Einstellungen zum Lernen*, zur *Vertrautheit mit Computern und Informationstechnologien* sowie zur bisherigen und antizipierten weiteren *Bildungskarriere* zugeordnet. Für diese Erhebungen werden nicht – wie bei den anderen Kompetenzen – Testverfahren verwendet, sondern Fragebögen und Einschätzskalen. Diese Erhebungen liefern gleichwohl Informationen über Einstellungen, Überzeugungen und Fähigkeiten von Jugendlichen, die als Bildungsergebnisse gegen Ende der Pflichtschulzeit verstanden werden können.

Hintergrundmerkmale: Die Erhebungen bei PISA beschränken sich nicht nur auf Bildungsergebnisse. Unter dem Begriff „Hintergrundmerkmale" verbergen sich Merkmale, die mit den Bildungsergebnissen der Schüler und Schülerinnen assoziiert sind und deren Lern- und Lebensumgebungen zugeordnet werden können. Es werden Lern- und Entwicklungsbedingungen erhoben, die vor allem das Elternhaus, die Schule und den Unterricht charakterisieren. Mit der Erhebung von Hintergrundmerkmalen auf diesen unterschiedlichen Ebenen kann PISA nicht nur aufschlussreiche Informationen über Bedingungen des Aufwachsens und diesbezügliche Unterschiede liefern. Die Erhebungen geben auch die Möglichkeit, die unter verschiedenen (und aus einer theoretischen Sicht unterschiedlich unterstützenden) Bedingungen entwickelte Kompetenz von Schülerinnen und Schülern zu vergleichen, auf der nationalen wie internationalen Ebene. Damit können auch Aussagen darüber getroffen werden, inwieweit bestimmte Lebensbedingungen (z.B. Merkmale der Herkunft) in den einzelnen Ländern systematisch mit Kompetenzunterschieden – also unterschiedlichen Chancen auf eine erfolgreiche Kompetenzentwicklung – verknüpft sind. Da der Erhebungsschwerpunkt im Jahr 2003 auf Mathematik liegt, werden verstärkt Hintergrundmerkmale berücksichtigt, die für die Entwicklung der mathematischen Kompetenz als bedeutsam erscheinen.

Für die Erfassung dieser Merkmale nutzt PISA zwei *Informationsquellen*: Die Schülerinnen und Schüler selbst und die Schulleitungen. Im *Schülerfragebogen* geben die Jugendlichen Auskunft über Merkmale ihres Elternhauses (ihrer sozialen Herkunft), über Einstellungen, Aktivitäten und über die Wahrnehmung ihres (vor allem Mathematik-)Unterrichts und ihrer Schule (z.B. Schulklima). Der an die Schulleitung gerichtete *Schulfragebogen* erhebt unter anderem die Größe und Ausstattung der Schule, Ressourcen, thematisiert aber auch das Schulmanagement, Kooperationen und Verfahren der Qualitätssicherung und fragt nach dem Schulklima und besonderen Unterstützungsangeboten.

1.3 Nationale Ergänzungen und Erweiterungen

Das *Design für die internationale Erhebung* sieht einen Testtag vor, an dem die fünfzehnjährigen Schülerinnen und Schüler in Gruppen an ihrer Schule unter Aufsicht getestet werden. Die reine Testzeit beträgt 120 Minuten, hinzu kommen 30 Minuten für das Ausfüllen des Schülerfragebogens. Weitere 15 Minuten werden für zusätzliche Fragebögen (Vertrautheit mit Computern, Vorstellungen über die Bildungskarriere) benötigt, die als internationale Option angeboten waren.

Das PISA-Konsortium Deutschland hat in Abstimmung mit der Auftraggeberin (der KMK) die verfügbaren *internationalen Optionen* gewählt und umgesetzt. Das Konsortium hat aber auch mit tatkräftiger Unterstützung der Auftraggeberin Möglichkeiten genutzt, die internationale PISA-Erhebung zu ergänzen und zu erweitern. Die Ergänzungen und Erweiterungen wurden so vorgenommen, dass an den deutschen Schulen die Testung an einem ersten Testtag exakt den internationalen Vorgaben entsprach. Zusätzliche Erhebungen wurden an einem nachfolgenden Testtag durchgeführt. Die Erweiterungen in Deutschland betrafen

- den Umfang und die Zusammensetzung der Stichprobe,
- die Erhebungsinstrumente und die befragten Gruppen (zusätzliche Befragungen von Lehrkräften und Eltern),
- die Erhebungszeitpunkte (Messwiederholung in einer Teilstichprobe).

Mit diesen Erweiterungen wurden mehrere Zielstellungen verfolgt. Durch zusätzliche Erhebungen und eine Erweiterung des Kreises der Befragten sollten ergänzende Informationen gewonnen werden, unter anderem, um die international gewonnenen Daten abzusichern und zu validieren. Erweiterungen der Stichprobe dienten dazu, die internationale Studie um einen Vergleich zwischen den deutschen Ländern zu bereichern. Andere Erweiterungen von Stichprobe und Design waren wissenschaftlich motiviert, um Bedingungsanalysen durchführen zu können, mit denen Erklärungsmodelle geprüft werden können. Insgesamt sollten also theoretisch wie praktisch bedeutsame Erkenntnisse gewonnen werden, die weit über die vorwiegend deskriptiven PISA-Befunde hinausreichen.

Erweiterung der Stichprobe: Um Vergleiche zwischen den deutschen Ländern durchführen zu können, die dem internationalen Vergleich entsprechen, musste der Stichprobenumfang in Deutschland um ein Vielfaches vergrößert werden. Dabei wurden außerdem zusätzliche Schülerinnen und Schüler mit Migrationshintergrund in die Stichprobe mit einbezogen, um zuverlässige Aussagen über Teilgruppen (unterschiedlicher Herkunft bzw. unterschiedlicher Generationen) treffen zu können. In einer Teilstichprobe von Schulen (den Schulen, die für den internationalen Vergleich ausgewählt waren) wurde die international verbindliche Stichprobe der Fünfzehnjährigen so ergänzt, dass wir dort zusätzlich zwei komplette Klassen (Jahrgangsstufe 9) untersuchen konnten. An diesen Schulen wurde auch eine Zufallsauswahl von Lehrkräften gezogen (vorzugsweise Lehrkräfte, die das Fach Mathematik unterrichten).

Zusätzliche Erhebungsinstrumente: An einer Teilstichprobe der Schulen (internationale Schulstichprobe) hatten die Schülerinnen und Schüler an einem zweiten Testtag nationale Tests für die Bereiche Mathematik, Naturwissenschaften, Lesen und Vertrautheit mit dem Computer zu bearbeiten. Sie bearbeiteten außerdem einen nationalen Schülerfragebogen mit Fragen zu mathematikbezogenen Einstellungen, Emotionen, Motivation wie zu Freizeitaktivitäten und zur Wahrnehmung von Elternhaus und Mathematikunterricht. Diese Tests und Fragebögen waren vom PISA-Konsortium Deutschland eigens für die Erhebung entwickelt worden. An diesen Schulen erhielten auch die Eltern der teilnehmenden Schülerinnen und Schüler einen Fragebogen, der insbesondere Angaben zu Ressourcen, zur Wahrnehmung von Unterricht wie Schule und zu mathematikbezogenen Unterstützungen erbat. Weiterhin wurde die Stichprobe der Lehrkräfte an diesen Schulen zu ihrer Wahrnehmung von Schule und Schulleitung, von Belastungen sowie zur Beteiligung an Schulentwicklung und an professionellen Kooperationen befragt. Alle Schulen, die in Deutschland an PISA teilnahmen, erhielten einen zusätzlichen Fragebogen für die Schulleitungen, der auf differenzierte Auskünfte über die Schule und Ansätze zur Schul- und Qualitätsentwicklung abzielte. Für eine kleine Teilstichprobe von Schulen wurde außerdem ein computergestützter Test entwickelt, der mit komplexen (mathematikbezogenen) Problemstellungen konfrontierte und insbesondere dazu diente, die Strategien zur Lösung des Problems zu rekonstruieren.

Zusätzliche Testzeitpunkte: Die Teilstichprobe der Schulen, die für den internationalen Vergleich gezogen wurde, durfte 2003 nicht nur an einem zweiten Testtag zusätzliche Tests bearbeiten. Diese Stichprobe, die pro Schule jeweils auch zwei komplette Klassen (neunter Jahrgang) umfasste, wurde ein Jahr später (also auf der zehnten Klassenstufe) noch einmal mit dem Schwerpunkt Mathematik getestet. Bei diesem zweiten Messzeitpunkt wurden zusätzliche Aufgaben einbezogen, die dem Niveau der zehnten Klassenstufe entsprachen. Diese Erhebung gibt Auskunft über die Kompetenzentwicklung bei den einzelnen Schülerinnen und Schülern beziehungsweise den jeweiligen Klassen im Verlauf eines Schuljahres.

Insgesamt kann man die Studie mit den Erweiterungen, die in Deutschland wahrgenommen wurden, als *drei Teilstudien* begreifen:

Die *erste* Teilstudie setzt den internationalen Vergleich nach den Standardregeln um und wird *PISA-I* („PISA-International“) bezeichnet.

Die *zweite* Teilstudie erweitert den internationalen Vergleich um den Vergleich zwischen den Ländern der Bundesrepublik Deutschland. Hier wird die Stichprobe drastisch erweitert, ergänzt durch ein Oversampling nach Migrationshintergrund. Die Schülerinnen und Schüler (Fünfzehnjährige) bearbeiten nur die internationalen Tests (ein Testtag). Diese Teilstudie wird als *PISA-E* (Erweiterungsstudie) etikettiert.

Die *dritte* Teilstudie konzentriert sich auf die kompletten neunten Klassen (aus den Schulen der internationalen Stichprobe). Die Schülerinnen und Schüler bearbeiten an einem zweiten Testtag zusätzliche Tests, es werden die Eltern und Lehrkräfte befragt. Diese Studie liefert somit systematische Informationen über Einflüsse auf den Ebenen Elternhaus, Klasse (auch Klassenlehrkraft Mathematik) und Schule (Kollegium, Schul-

leitung). Dieselben Schülerinnen und Schüler (bzw. Klassen) wurden 2004 noch einmal getestet, um ihre Kompetenzzuwächse im Verlauf eines Schuljahres messen zu können. In dieser Studie – *PISA-I-Plus* – kann geprüft werden, welche Effekte Einflussgrößen auf den Ebenen Individuum, Elternhaus, Klasse, Lehrkraft und Schule auf die Kompetenzentwicklung im Fach Mathematik haben.

Diese drei Teilstudien liefern äußerst umfangreiche Daten, die schrittweise geprüft und ausgewertet werden müssen. Deshalb werden wir über diese nationalen Teilstudien getrennt und zu unterschiedlichen Zeitpunkten berichten. Der vorliegende Band enthält im Wesentlichen die Ergebnisse des internationalen Vergleichs (PISA-I), angereichert durch Befunde, die zur Stabilisierung, Validierung und Differenzierung der internationalen Ergebnisse beitragen. Die Ergebnisse des Ländervergleichs (PISA-E) mit ausführlichen Analysen der Zusammenhänge zum Migrationshintergrund erscheinen als zweiter nationaler Bericht im Herbst 2005. Im Frühjahr 2006 werden schließlich die Hauptergebnisse aus der Messwiederholungsstudie mit kompletten Klassen präsentiert. Diese Studie PISA-I-Plus gestattet es, Annahmen über Bedingungsfaktoren auf unterschiedlichen Ebenen und ihre Effekte auf die Kompetenzentwicklung zu prüfen. Analysen, die Wirkungen kausal relevanter Faktoren klären, sind Teil zukünftiger Arbeiten und werden dem dritten Berichtsband vorbehalten bleiben.

An dieser Stelle soll auch darauf aufmerksam gemacht werden, dass aus dem Konsortium und Mitarbeiterkreis heraus weitere Forschungsprojekte begonnen wurden, die eng mit PISA-Fragestellungen verkoppelt sind. Diese Untersuchungen, die meist im DFG-Schwerpunktprogramm „Bildungsqualität von Schule“ gefördert werden, zielen auf Wissen, das zur Erklärung der PISA-Befunde beitragen kann (Ehmke, 2003; Krauss et al., 2004, in Druck; Lipowsky, Rakoczy, Klieme, Pauli & Reusser, 2004, in Druck; Pekrun et al., 2004, in Druck; Seidel & Prenzel, 2004, in Druck).

1.4 Anlage der Untersuchung

Internationale Vergleichsstudien, die über die Qualität von Bildungssystemen Auskunft geben sollen, müssen anspruchsvollen methodischen Standards genügen. Die folgenden Abschnitte stellen in knapper Weise dar, wie die Untersuchung angelegt, durchgeführt und ausgewertet wurde. Dabei soll die Fachterminologie auf das Notwendigste beschränkt werden. Über methodische und technische Details informiert unter anderem der technische Bericht (Adams, in Vorbereitung).

1.4.1 Untersuchungspopulation und Ziehung der Stichprobe

Die Zielpopulation von PISA wird durch das Lebensalter definiert. Auf diese Weise werden Jugendliche verglichen, die sich – je nach Schulsystem – auf unterschiedlichen Klassenstufen befinden können. Die Frage, welche Kompetenz Jugendliche bis zu einem

vergleichbaren Alter entwickeln, führt den Umgang mit Lebenszeit als Kriterium für den internationalen Vergleich ein. In Anbetracht sehr unterschiedlicher schulorganisatorischer Regelungen in den Teilnehmerstaaten wird mit der Altersstichprobe eine pragmatische und aussagekräftige Vergleichsmöglichkeit geschaffen.

Der Altersbereich für die *Zielpopulation der Fünfzehnjährigen* ist genau definiert (zwischen 15 Jahren 3 Monaten und 16 Jahren 2 Monaten). Die Grundgesamtheit der Untersuchung wird durch ein zweites Kriterium bestimmt: Der oder die Fünfzehnjährige muss sich noch *im Schulsystem* befinden (zumindest in einem Teilzeitschulverhältnis). In fast allen Staaten befinden sich nahezu alle Fünfzehnjährigen noch in einer Form der schulischen Ausbildung. In Deutschland sind dies 96,6 Prozent (vgl. Tabelle 1.1). Es gibt allerdings einige OECD-Staaten, bei denen die Anteile der Fünfzehnjährigen, die sich in keiner schulischen Ausbildung befinden, fünf Prozent und mehr betragen. Bei PISA 2003 sind dies die OECD-Länder Australien (6,5 Prozent), Mexiko (41,9 Prozent), Österreich (5,7 Prozent), Portugal (5,8 Prozent), Spanien (7,9 Prozent) und die Türkei (46,2 Prozent). Da anzunehmen ist, dass eher leistungsschwächere Schülerinnen und Schüler mit fünfzehn Jahren das Schulsystem verlassen haben, kann in einigen Staaten tendenziell mit einer gewissen Überschätzung der tatsächlichen Kompetenz gerechnet werden.

Ausschöpfung und Ausschlussgründe: Die Zielpopulation von Fünfzehnjährigen in schulischer Ausbildung sollte von den Staaten möglichst vollständig ausgeschöpft werden. PISA lässt nur wenige, genau definierte Ausschlussgründe zu: Die Schülerinnen und Schüler sind aus körperlichen, emotionalen oder geistigen Gründen nicht in der Lage, selbständig den Test zu bearbeiten, oder die Testsprache ist nicht ihre Muttersprache und sie wurden bisher weniger als ein Jahr in dieser Testsprache unterrichtet. Entsprechend können Schulen, die ausschließlich entsprechend definierte Schülergruppen unterrichten, ausgeschlossen werden. Ausschlüsse unter diesen Kriterien dürfen auf der Ebene der Schulen wie auf der Ebene der Schülerinnen und Schüler innerhalb der Schule jeweils maximal 2,5 Prozent der Zielpopulation erreichen. Unter Anwendung dieser Kriterien wurden in Deutschland nur die Schulen für geistig, körperlich und mehrfach Behinderte und Kranke ausgeschlossen. Insgesamt wurde in Deutschland eine *Ausschöpfung der Zielpopulation* von 96,2 Prozent erreicht (vgl. Tabelle 1.1).

Ziehung der Stichprobe: PISA schreibt ein festes Verfahren für die Ziehung der Stichprobe vor. Beim Stichprobenplan für Deutschland mussten einige Besonderheiten berücksichtigt werden (Details vgl. Kapitel 12.1.1), die mit der Gliederung des deutschen Schulsystems nach Schulformen zu tun haben, die aber auch durch die nationalen Erweiterungen (Ländervergleiche) bedingt waren. In Deutschland wurde eine mehrfach stratifizierte Wahrscheinlichkeitsstichprobe von Schulen gezogen, in denen dann eine zufällig ausgewählte Anzahl von Fünfzehnjährigen getestet wurde.

Für den *internationalen Vergleich* wurden in der ersten Stufe nach einem festgelegten Ziehungsplan 220 Schulen ausgewählt. Auf der zweiten Stufe wurden innerhalb dieser Schulen Zufallsstichproben von Fünfzehnjährigen gezogen. Die Stichprobengröße umfasste an den beruflichen Schulen wie Förder-/Sonderschulen alle Fünfzehnjährigen; in

Tabelle 1.1: Zielpopulation und Ausschöpfungsgrad 2003 in Deutschland

		Ausschlüsse auf Schulebene		Ausschlüsse in den Schulen		Erreichte Population	
Population der 15-Jährigen[a]	15-Jährige in Schulausbildung[b]	absolut[c]	in Prozent der Zielpopulation	absolut	in Prozent der Zielpopulation	absolut	in Prozent der Zielpopulation
951 800	919 017	5 600	0.61	11 533	1.29	884 358	96.2

PISA-Zielpopulation

[a] Quelle: Statistisches Bundesamt
[b] Quelle: Diese Angabe stellt eine Schätzung aus der Liste aller Schulen (Sampling frame) dar, die zur Stichprobenziehung erstellt wurde.
[c] Quelle: Schätzung aus dem Sampling frame

allen anderen Schulen wurden jeweils 25 Fünfzehnjährige in die Stichproben einbezogen. Für die *nationalen Ergänzungsstudien* wurden an den Schulen der internationalen Stichprobe zusätzlich Schülerinnen und Schüler auf der neunten Klassenstufe gezogen, um zwei komplette neunte Klassen testen zu können. Für die Ländervergleiche wurden eine sehr viel größere Schulstichprobe einbezogen sowie nach einem festgelegten Verfahren zusätzlich Fünfzehnjährige mit Migrationshintergrund gezogen (Kriterien: nicht in Deutschland geboren, Eltern nicht in Deutschland geboren, Umgangssprache im Elternhaus ist nicht Deutsch).

In Deutschland beteiligten sich alle Schulen, die für die internationale Vergleichsstichprobe gezogen worden waren[2]; vier Schulen sind nach Zusammenstellung der Schulliste geschlossen worden, so dass 216 Schulen teilnahmen. Damit beträgt die *Ausschöpfungsquote der Schulstichprobe* 100 Prozent. Als *Ausschöpfungsquote auf der Schülerebene* wurde eine Beteiligung von 92 Prozent der Fünfzehnjährigen erreicht. Eine Aufstellung der Beteiligungsquoten nach Schulform sowie Jahrgangsstufe der getesteten Fünfzehnjährigen wird in der Tabelle 1.2 präsentiert. Hier kann angemerkt werden, dass in Deutschland 2003 die Anstrengungen kräftig verstärkt wurden (z.B. durch mehrtägige Testleiterbesuche und Testgelegenheiten), in den beruflichen Schulen höhere Beteiligungsquoten zu erzielen.

Die in Deutschland getestete Stichprobe erfüllt damit alle internationalen Anforderungen. Diese Qualitätsanforderungen wurden ebenfalls von fast allen anderen Staaten erreicht. Für die Mittelwertsvergleiche nicht berücksichtigt werden konnte das Vereinigte Königreich, das mit einer Ausschöpfung von 63 Prozent auf Schulebene (nach Replacement 77 Prozent) und 78 Prozent auf Schülerebene die Anforderungen für eine Vergleichbarkeit nicht erfüllte.

Das PISA-Konsortium Deutschland hat 2003 eine Kontrollmöglichkeit genutzt, um eine mögliche Verzerrung der Ergebnisse durch eine selektive Schülerbeteiligung auszuschließen. Auch bei Beteiligungsquoten in der Größenordnung von 80 Prozent ist nicht

2 Zwei Schulen verweigerten die Teilnahme. Für sie wurden nach dem internationalen Reglement Ersatzschulen gezogen, die teilnahmen.

Tabelle 1.2: Untersuchungsbeteiligung der Fünfzehnjährigen nach Schulform und Jahrgangsstufe[3]

	Haupt-schule	Schule mit mehreren Bildungs-gängen	Real-schule	Inte-grierte Gesamt-schule	Gym-nasium	Berufs-schule	Förder-/ Sonder-schule	Summen
keine Zuordnung						127		127
7	33	14	10	3	2		11	73
8	233	72	189	55	69		36	654
9	514	329	691	263	859		54	2710
10 und 11	60	111	281	125	512		7	1096
getestet	840	526	1171	446	1442	127	108	4660
Ausschöpfung in %	85	93	92	91	95	64	72	90

auszuschließen, dass in der Restgruppe der ausgewählten, aber nicht am Test teilnehmenden Schülerinnen und Schüler eher Leistungsschwächere überrepräsentiert sind. Deshalb wurden die Testkoordinatorinnen und -koordinatoren an den Schulen gebeten, für alle Schülerinnen und Schüler der Stichprobe die Schulnoten in den Fächern Mathematik, Deutsch, Biologie, Chemie und Physik in eine Liste einzutragen. Anhand dieser Listen können die Durchschnittsnoten zwischen den ausgewählten Schülerinnen und Schülern, die tatsächlich am Test teilnahmen, und jenen, die nicht teilnahmen, verglichen werden. Die Ergebnisse dieses Vergleichs sind in Tabelle 1.3 dargestellt.

Tabelle 1.3: Schulnoten (Mittelwerte und Streuungen) der für den Test ausgewählten und der am Test teilnehmenden Fünfzehnjährigen, nach Schulform

	In die Stichprobe aufgenommene Fünfzehnjährige		Teilnehmende Fünfzehnjährige	
	MW	SD	MW	SD
Mathematik				
Hauptschule	3.44	1.04	3.37	1.01
Integrierte Gesamtschule	3.37	1.07	3.30	1.05
Realschule	3.30	1.01	3.27	1.00
Gymnasium	3.10	1.03	3.10	1.03
Alle Schulformen	3.27	1.04	3.23	1.02
Deutsch				
Hauptschule	3.48	0.87	3.42	0.84
Integrierte Gesamtschule	3.25	0.91	3.21	0.88
Realschule	3.24	0.81	3.24	0.81
Gymnasium	2.92	0.83	2.91	0.83
Alle Schulformen	3.17	0.87	3.14	0.86

3 Die Tabelle berichtet die ungewichteten realisierten Stichproben. Verzerrungen der realisierten Stichprobe aufgrund des Stichprobenplans und der Beteiligungsraten werden durch Gewichtungen ausgeglichen (vgl. Kapitel 12.1.3).

Wie der Tabelle 1.3 entnommen werden kann, unterscheiden sich die Durchschnittsnoten der Schülergruppen, die am Test teilnahmen beziehungsweise nicht teilnahmen, nur geringfügig. Es kann deshalb die Vermutung zurückgewiesen werden, dass die Ergebnisse in Deutschland durch eine selektive Beteiligung leistungsstarker Schülerinnen und Schüler verzerrt sein könnten (Kienzl, in Vorbereitung).

1.4.2 Test- und Fragebogenentwicklung sowie Testdesign

Wie bereits angesprochen, erfolgte die Entwicklung von Testverfahren und Fragebögen auf der Grundlage von Rahmenkonzeptionen, die von international besetzten Expertengruppen erarbeitet worden waren. Diese Expertengruppen wirkten beratend bei der Testentwicklung mit und überprüften die Materialien sorgfältig unter inhaltlichen und methodischen Gesichtspunkten.

Das Aufgabenmaterial für die Kompetenztests wurde an Instituten des Internationalen Konsortiums (vornehmlich ACER, CITO, ETS) entwickelt. In die *Testentwicklung* aufgenommen wurden zahlreiche *Aufgabenvorschläge* aus den Teilnehmerstaaten, die Nationale Projektmanager eingereicht hatten. Die Testaufgaben wurden in einem mehrstufigen Verfahren überprüft und gegebenenfalls weiterentwickelt. Nach ersten Erprobungen der Aufgaben an Schülergruppen erfolgte eine *Aufgabenbeurteilung* durch die jeweilige Expertengruppe und durch alle Nationalen Projektmanager. Kriterien der Beurteilung waren unter anderem fachliche Richtigkeit, Bezüge zum Curriculum, eventuelle kulturelle oder geschlechterbezogene Benachteiligung, Interessantheit und Schwierigkeit der Aufgabe. Aufgaben, die diese Beurteilung erfolgreich passierten, wurden in einem *Feldtest* überprüft, der 2002 in allen Staaten mit einer hinreichend großen Stichprobe durchgeführt wurde. Aufgrund der Itemanalysen der Daten aus dem Feldtest und einer neuerlichen Expertenbeurteilung wurden die Aufgaben für den Haupttest 2003 zusammengestellt.

Die Tests und die Fragebögen wurden den Nationalen Projektmanagern in einer englischen und französischen Version vorgelegt. Diese Versionen mussten von zwei unabhängigen Übersetzungsteams in die Testsprache übertragen und von einem dritten Übersetzer in eine Schlussversion gebracht werden. Die Erstellung der deutschsprachigen Vorlage wurde in einer Zusammenarbeit der Projektleitungen aus Luxemburg, Österreich, der Schweiz und Deutschland vorgenommen. Die Versionen für die deutschsprachigen Schülerinnen und Schüler unterschieden sich allenfalls in einigen landesspezifischen Formulierungen. Die Tests wurden in Deutschland abschließend von Experten geprüft.[4] Die von den Nationalen Projektmanagern fertig gestellten Test- und Fragebogenvorlagen wurden zuletzt von Übersetzungsexperten des internationalen Konsortiums kontrolliert und abgenommen.

4 Wir bedanken uns bei Jochen Borchert, Gerd Boysen, Marcus Hammann, Reinders Duit und Peter Nentwig für ihre Unterstützung.

Das PISA-Konsortium Deutschland hat auch 2003 zusätzlich die Bezüge der Mathematik- und Naturwissenschaftsaufgaben zu deutschen Lehrplänen prüfen lassen. Die *curriculare Validität* wurde für die internationalen und für die zusätzlichen nationalen Tests von Expertengruppen aus den Mathematik- und Naturwissenschaftsdidaktiken geprüft. Die Ergebnisse dieser Prüfung werden in den entsprechenden Kapiteln berichtet.

Die Zusammenstellung der Aufgaben in den Testheften und die Gestaltung der Testtage weisen bei PISA einige Besonderheiten auf. PISA soll auf ökonomische Weise in mehreren umfassenden Inhaltsbereichen zuverlässige Aussagen über die Kompetenzen treffen, die in den Bildungssystemen der Länder erreicht werden. Um die Testzeit und Testbelastung klein zu halten, dabei aber breite Kompetenzbereiche untersuchen zu können, wird bei PISA das *Testdesign des Multi-Matrix-Sampling* gewählt. Dabei werden nach bestimmten Prinzipien unterschiedliche Testhefte mit systematisch variierten Aufgabenblöcken erstellt und zufällig den Schülerinnen und Schülern zugewiesen. Bei PISA 2003 wurden insgesamt 13 solcher Testheftversionen eingesetzt. Das Multi-Matrix-Sampling gestattet es, Aussagen über die Kompetenz zu treffen, die in hinreichend großen Gruppen (bis zur Population in einem Land) erreicht wird. Durch dieses Verfahren gewinnt man eine Datenmatrix, die mit speziellen Auswertungsverfahren (der probabilistischen, so genannten Item-Response-Testtheorie) analysiert werden kann.

Abbildung 1.3: Das Design des internationalen Tests (1. Testtag)

Testheft	1	2	3	4	5	6	7	8	9	10	11	12	13
30 Min.	M1	M2	M3	M4	M5	M6	M7	N1	N2	L1	L2	P1	P2
30 Min.	M2	M3	M4	M5	M6	M7	N1	N2	L1	L2	P1	P2	M1
30 Min.	M4	M5	M6	M7	N1	N2	L1	L2	P1	P2	M1	M2	M3
30 Min.	L1	L2	P1	P2	M1	M2	M3	M4	M5	M6	M7	N1	N2
47 Min	FB	FB	FB	FB	FB	FB	FB	FB	FB	FB	FB	FB	FB

M1– M7 bezeichnet Aufgabenblöcke in Mathematik, N1 und N2 in Naturwissenschaften, L1 und L2 in Lesen, P1 und P2 in Problemlösen und FB bezeichnet den Schülerfragebogen

Der Abbildung 1.3 kann entnommen werden, wie die *Aufgabenblöcke* mit jeweils dreißig Minuten Bearbeitungszeit auf die Testhefte verteilt wurden und somit der internationale Test am ersten Testtag zusammengestellt war. Es wurden sieben Blöcke mit Mathematikaufgaben (insgesamt 210 Minuten Testzeit) und jeweils zwei Blöcke mit Aufgaben zum Lesen, Problemlösen und zu den Naturwissenschaften vorgegeben. Am Schluss des Tests hatten die Schülerinnen und Schüler die Fragebögen auszufüllen. Die Schülerinnen und Schüler der erweiterten Stichproben in Deutschland, die einen zweiten Testtag mit nationalen Tests und Fragebögen zu bearbeiten hatten, erhielten ebenfalls dreizehn Testhefte, die wiederum rotierte Aufgabenblöcke enthielten und mit den gleichen Verfahren konzipiert und ausgewertet wurden.

In den internationalen wie den nationalen Tests für die Kompetenzbereiche wurden unterschiedliche *Itemformate* verwendet. In allen Bereichen wurde etwas mehr als die Hälfte der Testfragen in einem Format gestellt, das bei der Auswertung unmittelbar die richtige Antwort erkennen ließ. Dabei kamen unterschiedliche Auswahlformate (z.B. einfache und komplexe Multiple-Choice-Items) zur Anwendung. Die andere (knappe) Hälfte der Testfragen verlangte von den Schülerinnen und Schülern, eine eigene Antwort zu formulieren, die aus einem einzigen Wort (Kurzantwortfragen) oder mehren Sätzen oder Zeichnungen (offene Fragen) bestehen konnte. Für diese Fragen lag jeweils ein sehr differenziertes Auswertungsschema vor, mit dem die Lösungen klassifiziert und bewertet werden mussten. Die Antworten auf diese Fragen wurden von mehreren (bis zu vier) unabhängigen Beurteilern bewertet, die ein spezielles Kodiertraining erhalten hatten. Um zu überprüfen, ob die Auswertungen in allen Ländern nach gleichen Maßstäben und zuverlässig erfolgten, wurde eine Auswahl von Testheften und Auswertungen aus jedem Land international überprüft. Der Vergleich zeigte, dass bei den Testauswertungen über alle Länder hinweg eine sehr gute Konsistenz erzielt wurde. Die Beurteiler stimmten im Durchschnitt aller Teilnehmerländer je nach Domäne in 90 bis 95 Prozent der Fälle überein, in Deutschland waren es 92,9 Prozent.

1.4.3 Durchführung der Erhebung in Deutschland

Die Erhebung in Deutschland wurde in enger Zusammenarbeit zwischen dem nationalen Konsortium (der Koordinierungsstelle am IPN als Nationalem Projektmanager), dem mit der Durchführung der Erhebungen beauftragten Data Processing Center der IEA (DPC) und den für die Schulen zuständigen Administrationen der Länder vorbereitet und durchgeführt. Ebenfalls einbezogen wurden die Datenschutzbeauftragten der Länder. Sie erhielten alle Informationen über die Stichprobenziehung, die Anschreiben an die Schulen, Eltern wie Schülerinnen und Schüler, über die einzusetzenden Fragebögen und die Verfahren zur Sicherung der Anonymität, so dass in allen Ländern die geltenden Richtlinien für die Datensicherheit eingehalten werden konnten. Die Administrationen wurden über die gezogenen Schulen informiert.

Je nach Gesetzen und Richtlinien der Bundesländer war (zum Testzeitpunkt) die Teilnahme an den Tests für alle Schülerinnen und Schüler verpflichtend oder freiwillig und von der Zustimmung der Eltern abhängig. Die Beantwortung der Fragebögen, für die auch die Zustimmung der Eltern eingeholt werden musste, war in 15 Ländern freiwillig. In einem Bundesland (Brandenburg) war die Teilnahme an Test und Fragebogen für alle Schülerinnen und Schüler verpflichtend. In anderen Ländern (Baden-Württemberg, Bayern, Niedersachsen, Nordrhein-Westfalen, Saarland, Schleswig-Holstein), in denen die Testteilnahme für die Schülerinnen und Schüler freiwillig war, mussten die Einverständniserklärungen der Eltern spätestens am Testtag vorliegen. Die unterschiedlichen Regelungen einer verbindlichen versus freiwilligen Teilnahme können natürlich die Ausschöpfung der Stichprobe in den einzelnen Ländern beeinflussen. Der

oben berichtete Vergleich der Schulnoten zwischen der Gruppe der gezogenen und der teilnehmenden Schülerinnen und Schüler liefert keine Hinweise auf systematische Verzerrung bei der internationalen Stichprobe (220 Schulen). Allerdings werden wir diese Regelungen insbesondere beim Vergleich der Ergebnisse der Länder in Deutschland im nächsten Bericht des PISA-Konsortiums Deutschland berücksichtigen. In Anschreiben an die Schulen wurde auf die Bedeutung der Studie hingewiesen. Die Schulen wurden gebeten, an der Erhebung teilzunehmen und bei den gezogenen Schülerinnen und Schülern und deren Eltern für eine Beteiligung zu werben.

Von den Schulen wurden Koordinatorinnen und Koordinatoren benannt, die bei der Zusammenstellung der Schülerlisten für die Stichprobenziehung mitwirkten und zusätzlich eine Reihe von organisatorischen Aufgaben (z.B. Bereitstellung von Räumen) übernahmen. Das Verfahren war so organisiert, dass die Schülerinnen und Schüler zwar anhand von Namenslisten mit einer Codenummer für den Test gezogen wurden. Die Zuordnung zwischen Namen und Codenummer verblieb jedoch an der Schule, so dass in den Testheften und Fragebögen keine Namensangaben mehr enthalten waren. Dieses Kodifizierungsverfahren gestattete es den Koordinatorinnen und Koordinatoren an den Schulen auch, den Schülerlisten bei der zweiten Testung (am zweiten Testtag bzw. bei der Messwiederholung) ein Testheft mit dem gleichen Code zuzuordnen.

Die Erhebungen an den Schulen der internationalen Stichprobe (PISA-I-Schulen) wurden von schulexternen Testleitern durchgeführt, die vom DPC ausgewählt und für die Aufgabe ausgebildet worden waren. Den Testleitungen wurden die für die jeweilige Schule kodierten Testhefte zugestellt. Für die Erhebungen an den zusätzlich für den Ländervergleich gezogenen Schulen (PISA-E-Schulen) rekrutierten die Länder Personen für die Testleitung (z.B. Schulpsychologen), die ebenfalls vom DPC auf ihre Aufgaben im Detail vorbereitet und geschult wurden. Diese Testleiterinnen und Testleiter hatten jeweils eine größere Anzahl von Schulen zu testen. Für die Vorbereitung und Durchführung der Tests lag ein detailliertes Manual vor, das genau zu befolgen war. Zur Gewährleistung der Aufsichtspflicht ordneten die Schulen jeweils eine Lehrkraft ab, die neben der Testleitung im Testraum anwesend war. Lehrkräfte erhielten jedoch keine Möglichkeit, die Testhefte zu studieren. In allen Stufen des Verfahrens wurde streng auf die Geheimhaltung des Testmaterials geachtet. Alle Personen, die Gelegenheit hatten, Einblick in die Testhefte zu nehmen (z.B. Testleiterinnen und -leiter) hatten Vertraulichkeitserklärungen zu unterzeichnen.

Um sicherzustellen, dass die Testung an allen Schulen nach den verbindlichen Regeln erfolgte, entschied sich das Konsortium, zusätzliche Qualitätskontrollen in Auftrag zu geben. So erhielten per Zufall ausgewählte Schulen der internationalen und der erweiterten Stichprobe am Testtag unangekündigt Besuch von erfahrenen Kolleginnen und Kollegen, die wiederum nach einem festgelegten Verfahren die Erhebung vor Ort kontrollierten. Aus der internationalen Stichprobe wurden 15 Schulen (von einer internationalen Gruppe), aus der Erweiterungsstichprobe 100 Schulen überprüft. Im internationalen Testbericht wurden keine nennenswerten Beanstandungen genannt. Auch der Bericht über das Qualitätsmonitoring bei den Schulen der Erweiterungsstichprobe

gelangt insgesamt zur Aussage, dass die Testung in allen entscheidenden Punkten dem vorgesehenen Reglement entsprach (vgl. Köller, 2003). In einzelnen Fällen wurden kleinere Abweichungen vom Erhebungsskript festgestellt.

1.4.4 Auswertung und Skalierung

Die Auswertung der Fragebögen und Tests vollzog sich über eine Reihe von Schritten. So mussten die Testhefte eingelesen und die „offenen" Fragen kodiert werden, um die insgesamt sehr umfangreiche Datenbasis (ca. 535 Variablen im internationalen Teil und 1365 Variablen in der nationalen Ergänzung) zu erhalten. Diese Arbeiten erfolgten in den einzelnen Staaten. Die Daten wurden dann an das internationale Konsortium beziehungsweise an dessen federführendes Institut, das Australian Council for Educational Research (ACER), zur zentralen Auswertung weitergegeben. In Deutschland wurde zeitgleich an der Auswertung dieser nationalen Daten und der zusätzlich erhobenen Daten des zweiten Testtags und der Klassenerhebungen gearbeitet.

Als entscheidenden Schritt bei einem internationalen Leistungsvergleich kann man die *Skalierung der Tests* bezeichnen. Ziel der Skalierung ist es, die Antworten der Schülerinnen und Schüler auf die Testfragen so zu analysieren, dass ihre Kompetenz verglichen und – auf einer Skala – abgestuft werden kann. Will man die Kompetenz von Personen auf einer Skala einordnen, dann muss diese Skala eindimensional sein (eine einheitliche Kompetenz abbilden). Die Skala sollte ebenfalls möglichst messgenau sein; das bedeutet, die Items sollten die Kompetenz zuverlässig erfassen. Die Skalierung kann auch dazu dienen, die Schwierigkeit der Testaufgaben zu bestimmen.

Die bei PISA verwendeten Verfahren werden in Kapitel 12 mit einigen technischen Details skizziert. International wie national wurden bei der Auswertung der Testdaten Analyseverfahren verwendet, die auf Modellen der Item-Response-Theorie beruhen (vgl. Adams & Wu, 2002; Rost, 2004). Auswertungen mit diesem Ansatz liefern Informationen über die Dimensionalität eines Tests und seine Messgenauigkeit. Anhand der Itemkennwerte können die gemessenen Kompetenzen der Schülerinnen und Schüler unter inhaltlichen Bezugskriterien interpretiert und, theoretisch begründet, so genannten Kompetenzstufen zugeordnet werden. Dieser Ansatz gestattet es aber auch, die Verteilungen von Kompetenzen zu betrachten und die Leistungen von Schülergruppen (bis hin zu nationalen Stichproben) zu vergleichen. Ein weiterer Vorteil dieses Auswertungsverfahrens besteht darin, dass man bei einem Testdesign mit rotierten Aufgabenblöcken die Kompetenzen der Schülerinnen und Schüler auf einer gemeinsamen Skala vergleichen kann. Allerdings macht das Testdesign mit mehren Testheften und unterschiedlichen Aufgabenblöcken weitere statistische Berechnungen (z.B. von „Plausible Values", da immer nur Teile des Tests beantwortet wurden) erforderlich, um die Mittelwerte und Streuungen für die Populationen korrekt zu schätzen (vgl. Adams & Wu, 2002; Mislevy, Beaton, Kaplan & Sheehan, 1992).

Nach der internationalen Skalierung der Tests, die von ACER vorgenommen wurde, erhielten die teilnehmenden Staaten zunächst ihre skalierten nationalen Datensätze zurück, später den gesamten internationalen Datensatz. Die internationale Skalierung wird in Deutschland ebenfalls als Bezugspunkt für die Auswertungen der zusätzlichen Stichproben genutzt. Angemerkt sei, dass sich die internationale Skalierung und Normierung der Tests nur auf die OECD-Staaten bezieht. Die so genannten Partnerländer (außerhalb der OECD), die an PISA teilnehmen, wurden dieser OECD-Normierung zugeordnet. Die skalierten Datensätze können wiederum mit unterschiedlichen statistischen Verfahren weiter analysiert werden (allerdings unter Berücksichtigung einiger Besonderheiten und Einschränkungen, etwa aufgrund der Imputation von „Plausible Values").

Etwas einfacher stellt sich die Auswertung der *Fragebogendaten* dar. Allerdings waren auch hier zunächst einige Kodierungsschritte bei „offenen" Fragen (z.B. zur Klassifikation der Berufsangaben) erforderlich. Eine ganze Reihe der mit Fragebögen erhobenen Daten bezieht sich auf Merkmale, die mit einer Angabe beschrieben werden können (z.B. Alter oder Geschlecht). Die Fragebögen zielen aber auch auf Merkmale („Konstrukte"), die erst über mehrere Fragen (z.B. auch über Einschätzfragen, sog. Ratings) zuverlässig eingeschätzt werden können. Typische Beispiele für solche Konstrukte sind zum Beispiel das Mathematikinteresse oder die Einschätzung des Schulklimas. Die verschiedenen Fragen (bzw. Items) zu diesem Konstrukt müssen ebenfalls skaliert werden. Auch hier informieren Analysen der Dimensionalität oder Zuverlässigkeit über die Qualität und die Eignung der Skala für weitere Auswertungen. Das PISA-Konsortium Deutschland hat die Konstrukte der nationalen und internationalen Fragebögen auch auf der Basis der Item-Response-Theorie analysiert.

Soweit in diesem Band über Befunde berichtet wird, die ebenfalls im internationalen Report vorgestellt werden, orientieren wir uns an diesen Auswertungs- und Darstellungsverfahren.

1.4.5 Berichterstattung und Darstellung

Der vorliegende erste deutsche Bericht über PISA 2003 konzentriert sich auf eine Darstellung deskriptiver Befunde. In erster Linie beschreiben wir bedeutsame *Bildungsergebnisse*, die in Deutschland erreicht wurden.

Die Bildungsergebnisse kann man aus unterschiedlichen *Perspektiven* betrachten und analysieren. Eine erste Perspektive bietet der internationale Vergleich. Wir wollen die Befunde zweitens unter inhaltlichen Kriterien beleuchten, die aus Ansprüchen an eine zeitgemäße und anschlussfähige Grundbildung im Sinne von Literacy resultieren. Unter einer dritten Perspektive liegt es nahe, die Ergebnisse zu vergleichen, die bei PISA 2000 und 2003 erzielt wurden.

PISA gibt auch die Möglichkeit, über Merkmale von schulischen und außerschulischen Lernumgebungen zu berichten, die aus einer theoretischen Sicht als Ressourcen oder Voraussetzungen für die Entwicklung von Kompetenz gelten können. Auch hier kön-

nen die Befunde für sich stehend unter Qualitätskriterien diskutiert werden. Allerdings liegt hier die Versuchung nahe, aus dem Vergleich von Bildungsergebnissen und Merkmalen von Lernumgebungen *Erklärungen* (Aussagen über Ursachen) für feststellbare Unterschiede abzuleiten. Statistische Auswertungen, die korrelative Zusammenhänge zwischen Umgebungsmerkmalen und Kompetenzen beschreiben, werden gerne als Bedingungsanalysen verstanden, obwohl sie tatsächlich keine Aussagen über Bedingungen zulassen (allenfalls Vorhersagen von Kompetenzunterschieden auf der Basis von Merkmalsunterschieden, die gelegentlich als Varianzerklärung bezeichnet werden). So anregend solche Assoziationen und Schlussfolgerungen sein mögen, sie sind nicht mehr als Spekulationen oder Vermutungen, die einer kritischen empirischen Überprüfung bedürfen. Der internationale PISA-Datensatz lässt eine angemessene Prüfung von Hypothesen über kausal relevante Bedingungen jedoch nicht zu. Unter diesem Vorbehalt sollten auch alle Aussagen des Berichts der OECD (2004a, 2004b) gelesen werden. Das PISA-Konsortium Deutschland wird in diesem Band keine Aussagen über mögliche Ursachen (kausal relevante Bedingungen) der Bildungsergebnisse treffen. Wir verweisen jedoch auf nachfolgende Analysen mit einem erweiterten nationalen Design, das durch zwei Erhebungszeitpunkte und systematische Erhebungen auf mehreren Ebenen empirisch tragfähige Bedingungsanalysen und weiterführende Erkenntnisse zulässt. Über diese Ergebnisse berichten wir im dritten Band, der im Frühjahr 2006 erscheinen wird.

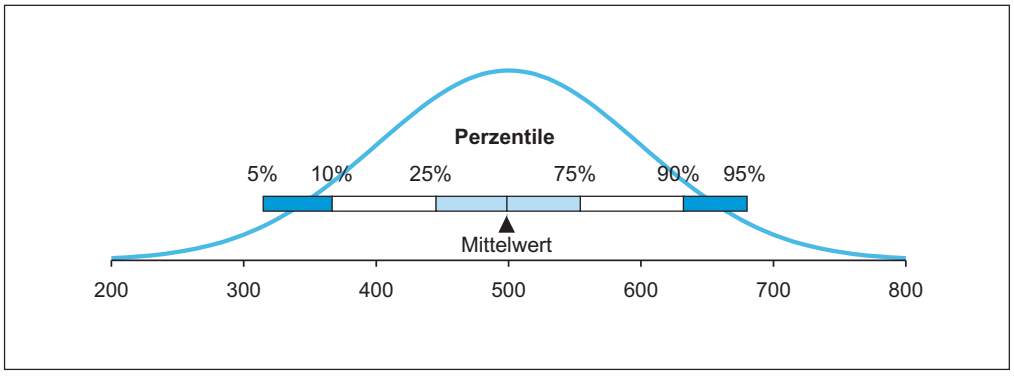
In den folgenden Abschnitten möchten wir erläutern, wie wir in diesem Band die Befunde unter den drei genannten Perspektiven darstellen.

In der PISA-Rezeption wurde der *internationale Vergleich* häufig nur als Rangordnung gelesen. Tatsächlich wurden die Rangordnungen oft unzulässig interpretiert und viele andere Informationen aus dem Vergleich übersehen. Bezugspunkt für die internationalen Vergleiche sind die international skalierten Tests. So wie es mehrere Möglichkeiten gibt, die Temperatur zu skalieren, können auch *Kompetenzskalen* unterschiedliche Zahlenwerte und Abstände zugeordnet werden. Die internationalen Kompetenzskalen sind so definiert, dass ihr *Mittelwert bei 500* liegt. Dieser Wert entspricht also dem internationalen Durchschnitt. Die internationale Skala ist weiterhin dadurch bestimmt, dass die *Standardabweichung* bei einem Wert von 100 liegt. Die Standardabweichung ist eine statistische Maßzahl, die die durchschnittliche Streuung der Werte einer Verteilung um den Mittelwert beschreibt. Bei PISA liegen im Bereich zwischen 400 und 600 (also Mittelwert ± eine Standardabweichung) ca. zwei Drittel (genau: 68,2 Prozent) der international getesteten Schülerinnen und Schüler. Bei einem Abstand von zwei Standardabweichungen (also 700 bzw. 300 Punkten) nähert man sich schon den Enden der Verteilung. Dann schneiden nur mehr jeweils 2,3 Prozent der Schülerinnen und Schüler besser (als 700 Punkte) beziehungsweise schlechter (als 300 Punkte) ab. Anhand der Maßzahlen „Mittelwert" (MW) und „Standardabweichung" (SD) können auch die Ergebnisse der einzelnen Staaten international verglichen werden. Erzielt ein Land einen Mittelwert von 600, dann liegt es über dem internationalen Durchschnitt, und zwar erheblich, denn wir wissen, dass dieser Wert international nur mehr von 15,9 Prozent der getesteten Schülerinnen und Schüler übertroffen wird. Aber auch der Kenn-

wert für die Standardabweichung in einem bestimmten Land ist aussagekräftig. Liegt die Standardabweichung über 100, dann streuen dort die Leistungen stärker als im internationalen Durchschnitt. Bei einer Standardabweichung unter 100 wären die Leistungen dementsprechend relativ homogen. Um die in Deutschland entwickelten und nur hier eingesetzten Tests von den international skalierten Tests abzuheben, haben wir uns entschieden, für die entsprechenden nationalen Skalen einen Mittelwert von 50 und eine Standardabweichung von 10 festzusetzen.

Die Tabellen für den internationalen Vergleich präsentieren zusätzlich Angaben über so genannte *Perzentile* (genauer für bestimmte Perzentilwerte, nämlich 5, 25, 75, 95). Diese Maßzahlen informieren darüber, wie hoch der Kennwert für die Kompetenz an bestimmten Abschnitten der Verteilung ist. Bei einem Perzentil von 95 wird also der Punktwert berichtet, ab dem die besten 5 Prozent einer Verteilung beginnen. Die Perzentilwerte 75 und 25 betreffen das obere oder untere Viertel der Verteilung. Die Perzentilwerte informieren somit vor allem darüber, wie stark die Leistungsspitze (die oberen 5 bzw. 25 Prozent) und wie schwach die untersten Gruppen (5 bzw. 25 Prozent) sind. Diese Angaben stellen wir graphisch auch in so genannten *Perzentilbändern* dar, um die Kompetenzverteilungen zu veranschaulichen. In Abbildung 1.4 ist ein Perzentilband dargestellt, zur Veranschaulichung ist ebenfalls eine Normalverteilung eingezeichnet. Mit der Darstellung von Perzentilbändern wird keine Normalteilung der Kennwerte vorausgesetzt.

Abbildung 1.4: Ein Beispiel für ein Perzentilband und Normalverteilung



Die Tabellen zum internationalen Vergleich (sowie zahlreiche andere) enthalten außerdem Angaben über den so genannten *Standardfehler* (S.E.) der Schätzung des Populationskennwertes. Mit Hilfe von Zufallsstichproben wird ja versucht, Aussagen über Merkmalsverteilungen in einer Population (Grundgesamtheit) zu treffen. Allerdings sind diese Schätzungen auf der Basis von Stichproben immer fehlerbehaftet. Die Größe des Fehlers wiederum lässt sich anhand der (gemessenen) Streuung in der Stichprobe im Verhältnis zur Stichprobengröße schätzen (die Wurzel aus diesem Wert ist der Standard-

fehler). Mit Hilfe des Standardfehlers kann man mit einer bestimmten *Irrtumswahrscheinlichkeit* (z.B. 5 Prozent) abschätzen, in welchem Wertebereich der „wahre" Wert der Population liegt (nämlich in einem Intervall von ± zwei Standardfehlern).[5] Die angegebenen Standardfehler kann man also auch nutzen, um anhand der geschätzten Populationswerte zu prüfen, ob sich die Mittelwerte von zwei Gruppen/Staaten überzufällig (mit einer Irrtumswahrscheinlichkeit von 5 Prozent) unterscheiden.

Wenn die bei PISA teilnehmenden Staaten nun anhand ihrer Mittelwerte in einer Tabelle angeordnet werden, erzeugt man eine *Rangfolge*. Die Ausführungen zum Standardfehler für die Populationsschätzung zeigen freilich, dass die Unterschiede in den Stichprobenmittelwerten nicht immer substantielle Unterschiede (zwischen den Populationen) abbilden. Deshalb müssen geeignete statistische Verfahren zum *Mittelwertsvergleich* angewendet werden. In den internationalen Vergleichstabellen werden entsprechend *drei Blöcke* gebildet (OECD-Durchschnitt, oberhalb und unterhalb des OECD-Durchschnittes). Innerhalb dieser Blöcke sind Unterschiede zwischen den Staaten statistisch nicht mehr zuverlässig abzusichern. Folgt man dieser Betrachtung, dann verbietet es sich, die Tabellenplätze durchzunummerieren. Im Anhang befinden sich Tabellen, die im Detail über signifikante Unterschiede bei Vergleichen einzelner Staaten berichten.

Wenden wir bei PISA die gebräuchlichen statistischen Verfahren zur Prüfung von Mittelwertsunterschieden an, müssen wir aufgrund der großen Stichprobenumfänge damit rechnen, dass sich auch kleinere Mittelwertsunterschiede als *statistisch signifikant* erweisen können (auch bei Irrtumswahrscheinlichkeiten unter 1 Prozent). Allerdings kann es sein, dass diese kleinen Mittelwertsunterschiede aus einer praktischen Perspektive kaum mehr bedeutsam sind, etwa wenn sich die Verteilungen der beiden Vergleichsgruppen für ein Merkmal fast vollständig überlappen. Aus diesem Grund werden wir bei entsprechenden Vergleichen (und signifikanten Unterschieden) auch das Maß der *Effektstärke* nutzen, um das Ausmaß des Unterschiedes statistisch darzustellen. Die Effektstärke (die Differenz der Mittelwertsunterschiede, geteilt durch die Standardabweichung) gibt an, wie stark sich Verteilungen überlappen. Das in diesem Band verwendete Effektstärkemaß (d) wird häufig so interpretiert, dass Werte in der Größenordnung von $d = 0.2$ als „kleine" Effekte, $d = 0.5$ als „mittlere" und $d = 0.8$ als „große" Effekte bezeichnet werden.

Bei der Beurteilung von Kompetenzunterschieden zwischen Gruppen bei großen Schulleistungsvergleichen hat es sich als praktisch erwiesen, diese in ungefähre Entwicklungszeiten (Schuljahre) zu übersetzen. Als Anhaltspunkt dienen hier die Informationen über durchschnittliche Kompetenzunterschiede zwischen Schülerinnen und Schülern verschiedener Klassenstufen. Im Sinne einer groben Faustregel können Unterschiede in der Größenordnung von etwa 40 Punkten in einen *Abstand von einem Schuljahr* umgerechnet werden.

---

5 In Tabellen und Abbildungen mit allen OECD-Staaten wird ein für multiple Vergleiche korrigiertes Intervall von ± 3,2 Standardfehlern verwendet.

Mit dem Kriterium einer Vorbereitung auf die Teilhabe an der Wissensgesellschaft und mit entsprechenden Testkonzeptionen für die Kompetenzbereiche beschränkt sich PISA nicht auf Mittelwertsvergleiche. Die Staaten sollen auch informiert werden, über welche *inhaltlichen Kompetenzen* ihre Schülerinnen und Schüler verfügen und wie sich solche Kompetenzen auf Schülergruppen innerhalb des Landes verteilen. Zu diesem Zweck werden bei PISA für die jeweiligen Schwerpunktgebiete Tests und *Subskalen für Teilkompetenzen* entwickelt. Die Information aus den Gesamt- und Subskalen kann im Rahmen der bei PISA durchgeführten Skalierung weiter aufgeschlüsselt werden. Die Skalierung beschreibt die Schwierigkeit der Testitems. Aus der theoretischen Sicht der Testkonzeption können wiederum mögliche Abstufungen der Kompetenz inhaltlich beschrieben und auf die vorfindbaren Items mit ihren Schwierigkeitskennwerten und Anforderungen bezogen werden. Auf diese Weise können unterschiedliche Niveaus einer Kompetenz bestimmt und anhand von Aufgabenanforderungen inhaltlich charakterisiert werden (vgl. Baumert et al., 2001; Klieme, Baumert, Köller & Bos, 2000). Formal sind die *Kompetenzstufen* so definiert, dass Schülerinnen und Schüler auf dieser Stufe zugeordnete Schwellenitems mit einer bestimmten Wahrscheinlichkeit (62 Prozent) lösen. Aufgaben, die höheren Kompetenzstufen entsprechen, werden mit einer sehr viel geringeren Wahrscheinlichkeit gelöst. Die inhaltliche Interpretation der Tests mit Hilfe von Kompetenzstufen wurde bisher bei PISA (Baumert et al., 2001; OECD, 2001; Turner, 2002) dazu genutzt, Gruppen (Anteile) von Schülerinnen und Schülern zu identifizieren, die aufgrund ihrer Kompetenz sehr schlechte Chancen für ein nachfolgendes Lernen innerhalb und außerhalb der Schule haben. Für Schülerinnen und Schüler, die auf oder unterhalb der ersten Kompetenzstufe anzusiedeln sind, ist die Prognose für die weitere Bildungskarriere (auch bezogen auf eine berufliche Ausbildung) ungünstig. Umgekehrt kann im Spitzenbereich (der obersten Kompetenzstufe) gefragt werden, wie groß die Anteile von Schülerinnen und Schülern sind, die über exzellente Kompetenz in einem Gebiet verfügen. In den entsprechenden Kapiteln werden wir deshalb über die entsprechenden Anteile von Schülerinnen und Schülern berichten.

An dieser Stelle weisen wir noch auf eine Auswertungs- und Darstellungsform hin, die im internationalen Bericht der OECD wie auch in diesem Band dann oft verwendet wird, wenn *außerhalb der Kompetenzbereiche* über skalierte Merkmale (Aussagen) von Schülerinnen und Schülern, Lehrkräften oder Schulleitungen berichtet wird. Skalen zu solchen Konstrukten (z.B. Motivation, Schulklima) werden häufig so transformiert, dass sie einen Mittelwert von 0 und eine Standardabweichung von 1 aufweisen. Entsprechend transformierte Skalen stellen anschaulich die Unterschiede zwischen Teilgruppen dar, zum Beispiel zwischen Staaten, aber auch innerhalb der Staaten zwischen beispielsweise Jungen und Mädchen oder Schulformen.

Der *Vergleich von Ergebnissen über Erhebungszeitpunkte*, der über die PISA-Zyklen möglich wird, dürfte schließlich die auf längere Sicht interessanteste Perspektive eröffnen. Über mehrere PISA-Erhebungen können die Entwicklungen von Bildungssystemen mit definierten Kriterien und geeigneten Erhebungsverfahren empirisch beschrieben werden. Die Staaten erhalten dabei auch Rückmeldung darüber, inwieweit Maßnahmen –

die möglicherweise als Reaktion auf frühere PISA-Befunde eingeleitet wurden – gegriffen haben und die beabsichtigten Wirkungen erzielen. Dabei muss in Rechnung gestellt werden, dass Eingriffe und Veränderungen in Bildungssystemen sich oft erst mit erheblicher Verzögerung in Bildungsergebnisse niederschlagen: Die Bildungsergebnisse, die PISA bei Jugendlichen misst, repräsentieren eine fünfzehnjährige Lerngeschichte. Für Maßnahmen, die unmittelbar nach einem PISA-Bericht ergriffen werden, bleibt ein Wirkungszeitraum von maximal 18 Monaten bis zur nächsten Erhebungsrunde. Dieser sehr knappe Zeitraum wird auch nur dann ausgeschöpft, wenn die Maßnahmen tatsächlich unmittelbar das Lernen der Schülerinnen und Schüler beeinflussen. Insofern dürften sich Effekte von zwischenzeitlich ergriffenen Maßnahmen erst längerfristig nachweisen lassen. Diese Bedingungen sollte man sich vor Augen führen, wenn man Erkenntnisse aus dem nun erstmals möglichen Vergleich zweier PISA-Erhebungen erwartet. Es gibt weitere methodische Einschränkungen für diesen Vergleich, die wir im folgenden Abschnitt behandeln.

1.5 Von PISA 2000 nach PISA 2003: Belastbare Aussagen über Veränderungen

Bei PISA 2000 und PISA 2003 wurden in den Staaten unterschiedliche Populationen (Kohorten) mit vergleichbaren Erhebungsverfahren getestet. Zu beiden Zeitpunkten werden die Populationen durch Stichproben repräsentiert. Diese Tatsache müssen wir ebenso berücksichtigen wie die Zuverlässigkeit von Erhebungsverfahren, wenn wir die Ergebnisse aus beiden PISA-Runden nebeneinander stellen. Letztlich gilt es, die substantiellen Unterschiede zwischen den Erhebungen – unter Kontrolle von eventuellen Stichprobenfehlern und Messungenauigkeiten – herauszuarbeiten. Gerade wenn man PISA auch als Feedback über erreichte Veränderungen nutzen möchte, müssen wir die gemessenen Ergebnisse kritisch und mit größter methodischer Sorgfalt prüfen. Es ist nicht auszuschließen, dass Veränderungen in Kennwerten, die auf den ersten Blick als bemerkenswert erscheinen, sich bei einer kritischen Betrachtung als unbedeutsam erweisen. Wenn in diesem Band Ergebnisse aus den beiden PISA-Erhebungen gegenübergestellt werden, dann sind mehrere Aspekte zu berücksichtigen beziehungsweise zu beachten.

Da PISA bisher nur einen Vergleich über zwei Messzeitpunkte gestattet, fehlt die empirische Grundlage, um von einer *Entwicklung* oder von einem *Trend* zu sprechen. Beide Bezeichnungen unterstellen eine substantielle Veränderung und suggerieren eine Progression. In Übereinstimmung mit dem internationalen Bericht plädieren wir dafür, bei der Begriffsverwendung Vorsicht walten zu lassen. Über einen Trend könnte gesprochen werden, wenn Veränderungen über drei Messzeitpunkte in die gleiche Richtung weisen. In diesem Band begnügen wir uns damit, von *Unterschieden* zwischen den Erhebungszeitpunkten beziehungsweise von *Veränderungen* zu sprechen, wenn die Unterschiede als substantiell erscheinen.

Um einzuschätzen, ob Veränderungen in Kennwerten substantielle Änderungen in den Bildungsergebnissen anzeigen, müssen wir uns mit Besonderheiten der Stichproben, der Testdurchführung und -beteiligung sowie der Zuverlässigkeit der Erhebungsverfahren befassen.

Unterschiede in den typischen Kennwerten (MW, SD) über zwei Erhebungszeitpunkte können auf Unterschiede in der *Ausschöpfung der Population und der Stichprobe* zurückzuführen sein. Veränderungen der Anteile der Zielpopulation (Fünfzehnjährige in schulischer Ausbildung) an der Population (Fünfzehnjährige) oder die Ausschöpfung der Stichprobe (z.B. Testung von Jugendlichen in beruflichen Teilzeitschulen) können sich deutlich in den Ergebnissen niederschlagen. Ebenso könnten Unterschiede auftreten, wenn die Testbeteiligung zu einem Zeitpunkt systematisch verzerrt wäre. Für Deutschland haben wir ausgeschlossen, dass 2003 systematisch leistungsschwächere Schülerinnen und Schüler von einer Testteilnahme absahen (s.o.). Um zu prüfen, ob die Unterschiede zwischen den Stichproben der Jahre 2000 und 2003 auch Unterschiede zwischen den Populationen repräsentieren, müssen wiederum (s.o.) Standardschätzfehler berücksichtigt beziehungsweise geeignete Signifikanzprüfungen durchgeführt werden.

Erhebliche Auswirkungen auf die Testergebnisse kann auch die Art und Weise der *Testadministration* haben. So bereiteten unter anderem Testheft- und Positionseffekte eine Reihe von Problemen, die Kompetenzkennwerte zu den zwei Zeitpunkten zu vergleichen. Das Beispiel Luxemburg zeigt, dass Unterschiede in der Testadministration zu den zwei Erhebungszeitpunkten (Zuordnung der Testhefte nach Sprachgruppen) mit deutlichen Unterschieden in den Ergebnissen verbunden sein können. So ist zu vermuten, dass in diesem Staat die Kompetenz durch die Testergebnisse 2000 unterschätzt war. Aus diesem Grund wird für Luxemburg kein Vergleich zwischen PISA 2000 und PISA 2003 berichtet. Entsprechend ist sorgfältig zu prüfen, ob die Testdurchführung oder auch Qualität der Übersetzungen zu beiden Erhebungszeitpunkten in jeder Hinsicht identisch war.

Auch wenn es einige Evidenz dafür gibt, dass Schulleistungsvergleiche und PISA-Befunde robust gegenüber Unterschieden in der *Testmotivation* sind (Baumert & Demmrich, 2001; O'Neill, Sugrue & Baker, 1996), empfiehlt es sich dennoch zu kontrollieren, ob die Testmotivation bei den Erhebungsrunden gleich ausgeprägt war. Die Testhefte enthielten eine Skala, an der die Schülerinnen und Schüler ihre Anstrengung bei der Testbearbeitung einschätzten. Da speziell in Deutschland aufgrund der öffentlichen Aufmerksamkeit für PISA Auswirkungen auf die Testmotivation nicht auszuschließen waren, haben wir die entsprechenden Kennwerte für die beiden PISA-Erhebungen verglichen.

Wie die Tabelle 1.4 zeigt, berichten die Schülerinnen und Schüler 2003 eine höhere Testmotivation. Dieser Unterschied stellt sich ausgeprägt (und signifikant) für die Schulformen Gymnasium und Realschule dar, etwas geringer in der Hauptschule. Für die Schülerinnen und Schüler an Integrierten Gesamtschulen, die bereits in PISA 2000 eine relativ hohe Testmotivation berichteten, ist in PISA 2003 keine Veränderung festzustellen. Ob diese Unterschiede in der Testmotivation tatsächlich für eine Leistungssteigerung verantwortlich gemacht werden können, steht dahin. Die Korrelationen zwischen Test-

Tabelle 1.4: Testmotivation für PISA 2000 und PISA 2003, Einschätzskalen am Ende der Testhefte

Schulform	PISA 2000		PISA 2003		Unterschied in Effektstärke d
	MW	SD	MW	SD	
Hauptschule	7.0	3.4	7.2	2.1	**0.06**
Integrierte Gesamtschule	7.5	3.7	7.4	2.0	0.00
Realschule	7.2	2.0	7.6	1.8	**0.15**
Gymnasium	7.2	2.0	7.4	1.9	**0.11**
Gesamt	7.2	2.5	7.5	1.9	**0.09**

Signifikante Unterschiede sind in **Fettschrift** gekennzeichnet

motivation und den Kompetenzen sind sehr gering (je nach Kompetenz und Schulform zwischen $r = 0.00$ und $r = 0.08$) und lassen keinen bedeutsamen Zusammenhang zwischen diesen Merkmalen erkennen. Die Befunde weisen darauf hin, dass die Schülerinnen und Schüler in Deutschland 2003 den PISA-Test etwas ernster nahmen als 2000. Wenn die Testmotivation tatsächlich die Testergebnisse beeinflusst, dann müsste der Effekt bei allen Kompetenzbereichen in einem vergleichbaren Ausmaß festzustellen sein. Wir haben in den folgenden Kapiteln die Möglichkeit, dies zu prüfen.

Ebenfalls im Zusammenhang mit der PISA-Debatte in Deutschland stellt sich die Frage, inwieweit die Schülerinnen und Schüler sich durch gezieltes Üben speziell auf die PISA-Tests vorbereiteten beziehungsweise vorbereitet wurden. Über den Sinn eines kurzfristigen testbezogenen Trainierens von freigegebenen PISA-Aufgaben (und Aufgaben in diesem Stil) kann gestritten werden: Wenn ein solches Training wirksam wäre, würde die tatsächliche Kompetenz überschätzt und ein unrealistisches Bild der Leistungsfähigkeit vermittelt. Die in Deutschland am Test teilnehmenden Schülerinnen und Schüler wurden deshalb gefragt, inwieweit sie gezielt für den Test trainierten. Insgesamt berichten 26 Prozent der Schülerinnen und Schüler, dass sie für den Mathematiktest trainierten. Über die Schulformen unterscheiden sich diese Angaben; in den Hauptschulen geben 17 Prozent der Jugendlichen an, für den Mathematiktest geübt zu haben, in den Integrierten Gesamtschulen sind es 28 Prozent, in den Realschulen 25 Prozent und in den Gymnasien 35 Prozent. Innerhalb der Schulformen unterscheiden sich die Leistungen nicht zwischen Schülerinnen und Schülern, die geübt haben oder nicht geübt haben. Zwölf Prozent der Schülerinnen und Schüler gaben an, für den Lesetest trainiert zu haben, diese Schülerinnen und Schüler unterscheiden sich im Mittel der Lesekompetenz nicht von denen, die berichten, nicht geübt zu haben.

Die Mittelwerte der in PISA 2000 erhobenen Kompetenz-Skalen wurden für den OECD-Durchschnitt auf einen Mittelwert von 500 und eine Standardabweichung von 100 festgelegt. Um die Ergebnisse von PISA 2003 auf die vorherige Studie beziehen zu können, muss eine gemeinsame Verankerung der Kompetenzdaten beider Studien gewählt werden. Für die *Lesekompetenz* und die *Naturwissenschaften* wurde als Bezugs-

punkt die Verankerung auf der PISA-Skala 2000 beibehalten; daher muss der OECD-Durchschnitt für diese beiden Skalen in PISA 2003 nicht 500 Punkte betragen.

Für *Mathematik* wurde PISA 2003 als Referenzpunkt gewählt, da hier ein vervollständigter Mathematiktest eingesetzt wurde. Die Daten aus PISA 2000 wurden auf diesen neuen Bezugspunkt umgerechnet und die Ergebnisse aus PISA 2000 und PISA 2003 müssen anhand der neu normierten Kompetenzwerte aufeinander bezogen werden. Die *Problemlösekompetenz* wurde in PISA 2003 erstmals erfasst und entsprechend für PISA 2003 auf einen Mittelwert von 500 und eine Standardabweichung von 100 Punkten verankert.

1.6 PISA – ein kooperatives Unternehmen

Eine aufwändige Studie wie PISA erfordert die Zusammenarbeit zahlreicher Organisationen, Institute und Personen auf nationaler und internationaler Ebene. Die politische Steuerung liegt in den Händen eines PISA Governing Board (PGB), in dem alle Teilnehmerländer vertreten sind (Vertreter des Bundes und der Länder sind Elfriede Ohrnberger, Bayerisches Staatsministerium für Unterricht und Kultus; Botho Priebe, Institut für schulische Fortbildung und schulpsychologische Beratung des Landes Rheinland-Pfalz; Hans Konrad Koch, Bundesministerium für Bildung und Forschung). Die internationale Koordination wird vom Sekretariat der OECD in Paris übernommen (verantwortlich: Andreas Schleicher). Unter der Federführung von ACER und unter der Leitung des Projektdirektors Raymond Adams wurde ein internationales Konsortium mit der wissenschaftlichen Koordination der Studien beauftragt. In diesem Konsortium arbeiten das Australian Council for Educational Research (ACER) zusammen mit dem Netherlands National Institute for Educational Measurement (CITO), dem National Institute for Educational Research, Japan (NIER), dem Educational Testing Service, Vereinigte Staaten (ETS) und WESTAT Inc., Vereinigte Staaten.

Im Auftrag des internationalen Konsortiums wurden internationale Expertengruppen eingerichtet für die Bereiche Lesekompetenz, mathematische Kompetenz, naturwissenschaftliche Kompetenz, Problemlösekompetenz, die Fragebögen und für technische Fragen (OECD, 2003).

In allen PISA-Teilnehmerländern sind Nationale Projektmanager für die ordnungsgemäße Vorbereitung und Durchführung der Studie verantwortlich. In Deutschland wurde der Auftrag für die Durchführung der Studie (nach Ausschreibung) durch die Ständige Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (KMK) an ein nationales Konsortium vergeben. Das Konsortium stimmt die Arbeit mit der Amtschefskommission „Qualitätssicherung in Schulen“ ab (Vorsitz: Ministerialdirektor Josef Erhard, Bayern, und Staatssekretär Elmar Schulz-Vanheyden, Nordrhein-Westfalen). Die Amtschefskommission bildet zusammen mit wissenschaftlichen Vertreterinnen und Vertretern (Rainer Bromme, Münster; Helmut Fend, Zürich; Kurt Heller,

München; Andreas Helmke, Landau; Klaus Klemm, Essen; Friederike Klippel, München; Kristina Reiss, Augsburg; Kaspar Spinner, Augsburg; Elke Sumfleth, Essen) einen Beirat für Vergleichsstudien in Deutschland.

Die Federführung des PISA-Konsortiums Deutschland liegt am Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN) in Kiel (http://pisa.ipn.uni-kiel.de). Dem Konsortium gehören folgende Wissenschaftler an:

- Manfred Prenzel, Kiel (Sprecher)
- Jürgen Baumert, Berlin
- Werner Blum, Kassel
- Rainer Lehmann, Berlin
- Detlev Leutner, Essen
- Michael Neubrand, Oldenburg
- Reinhard Pekrun, München
- Hans-Günter Rolff, Dortmund
- Jürgen Rost, Kiel
- Ulrich Schiefele, Bielefeld

Für die Projektkoordination für PISA 2003 am IPN sind Barbara Drechsel und Gesa Ramm verantwortlich. Der PISA-Arbeitsgruppe am IPN gehören Désirée Burba, Claus H. Carstensen, Barbara Drechsel, Timo Ehmke, Heike Heidemeier, Fanny Hohensee, Audrey McDonald-Blum, Manfred Prenzel, Gesa Ramm, Silke Rönnebeck, Jürgen Rost, Martin Senkbeil, Beate von der Heydt, Oliver Walter und Karin Zimmer sowie zahlreiche studentische Hilfskräfte an. Die Skalierung der nationalen Ergänzungstests und die Datenaufbereitung am IPN leitet Claus H. Carstensen.

Die Organisation der Datenerhebung sowie der Datenverarbeitung unternimmt das IEA-DPC in Hamburg. Zuständig waren Susan Böhmer, Regina Borchert, Falk Brese, Jens Gomolka, Steffen Knoll (Projektkoordination), Cornelia Kutter, Heiko Sibberns (Leitung) und Anja Waschk.

Die Arbeit des PISA-Konsortiums Deutschland wird durch Expertengruppen tatkräftig unterstützt. Die nationalen Expertengruppen setzen sich wie folgt zusammen.

Mathematik

Werner Blum (Sprecher), Kassel
Michael Neubrand (Sprecher), Oldenburg
Regina Bruder, Darmstadt
Elmar Cohors-Fresenborg, Osnabrück
Lothar Flade, Magdeburg
Rudolf vom Hofe, Regensburg
Alexander Jordan, Kassel
Norbert Knoche, Essen
Detlef Lind, Wuppertal
Wolfgang Löding, Hamburg
Gerd Möller, Düsseldorf
Johanna Neubrand, Vechta
Alexander Wynands, Bonn
Frauke Ulfig, Oldenburg

Leseverständnis
Ulrich Schiefele (Sprecher), Bielefeld
Cordula Artelt, Berlin
Jens Möller, Kiel
Wolfgang Schneider, Würzburg
Wolfgang Schnotz, Landau
Petra Stanat, Berlin
Lilian Streblow, Bielefeld

Naturwissenschaften
Jürgen Rost (Sprecher), Kiel
Horst Bayrhuber, Kiel
Wolfgang Bünder, Kiel
Claus H. Carstensen, Kiel
Reinders Duit, Kiel
Manfred Euler, Kiel
Hans E. Fischer, Essen
Alfred Flint, Rostock
Peter Häußler, Kiel
Marcus Hammann, Kiel
Rainer Klee, Gießen
Michael Komorek, Kiel
Armin Lude, Kassel
Jürgen Mayer, Gießen
Peter Nentwig, Kiel
Sabine Nick, Kiel
Helmut Prechtl, Kiel
Manfred Prenzel, Kiel
Martin Senkbeil, Kiel
Oliver Walter, Kiel

Problemlösen
Detlev Leutner (Sprecher), Essen
Cordula Artelt, Berlin
Joachim Funke, Heidelberg
Eckhard Klieme, Frankfurt a.M.
Stephan Kröner, Erfurt
Petra Stanat, Berlin
Joachim Wirth, Essen

Elternhaus/Schülervoraussetzungen
Reinhard Pekrun (Sprecher), München
Thomas Götz, München
Bettina Hannover, Berlin
Sabine Walper, München
Elke Wild, Bielefeld
Anne Zirngibl, München

Schulkontext
Rainer Lehmann (Sprecher), Berlin
Hans-Günter Rolff (Sprecher), Dortmund
Martin Bonsen, Dortmund
Astrid Neumann, Berlin
Hermann Pfeiffer, Dortmund
Brigitte Steinert, Frankfurt a.M.

Bedingungsbereich Mathematik
Jürgen Baumert (Sprecher), Berlin
Werner Blum, Kassel
Stefan Krauss, Berlin
Mareike Kunter, Berlin
Michael Neubrand, Oldenburg

Dass die PISA-Erhebung in Deutschland 2003 erfolgreich durchgeführt wurde, verdanken wir dem Engagement der Koordinatorinnen und Koordinatoren in den Ländern und an den einzelnen Schulen. Dieses Engagement drückt die Bereitschaft aus, sich Vergleichen zu stellen und aus Vergleichen zu lernen. Das PISA-Konsortium Deutschland möchte an dieser Stelle ebenfalls allen Schülerinnen und Schülern, Eltern, Lehrkräften und Schulleitungen für ihre Mitarbeit bei PISA 2003 herzlich danken.

1.7 Überblick über den Berichtsband „PISA 2003 – Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs"

Der vorliegende Bericht stellt erste Ergebnisse von PISA 2003 im internationalen Vergleich aus deutscher Perspektive dar. Dazu nehmen wir Bezug auf die beiden Berichte zum internationalen Vergleich, welche zeitgleich von der OECD vorgelegt werden und ergänzen diese Darstellung mit Ergebnissen zu Fragen aus deutscher Sicht. Dabei wird auch auf Ergebnisse aus in Deutschland zusätzlich erhobenen Tests und Fragebögen zurückgegriffen.

Dieser Band beginnt mit der Darstellung der untersuchten Kompetenzbereiche Mathematik (Kapitel 2), Lesen (Kapitel 3), Naturwissenschaften (Kapitel 4) und Problemlösen (Kapitel 5). Den Kapiteln über Kompetenzen folgen Berichte über die Vertrautheit mit neuen Medien (Kapitel 6), über Schülermerkmale (Kapitel 7) und über Geschlechterdifferenzen (Kapitel 8). All diese Kapitel berichten in erster Linie über Bildungsergebnisse. Die anschließenden Kapitel befassen sich dann mit Merkmalen der sozialen Herkunft (Kapitel 9) und den Lernumgebungen, Unterricht und Schule. Kapitel 10 betrachtet Unterricht und Schule aus unterschiedlichen Perspektiven, nämlich einmal aus einer institutionellen Sicht und dann aus der Wahrnehmung von Schulleitungen, Lehrkräften, Schülerinnen und Schülern. Das Kapitel 11 schließlich widmet sich zusammenfassend dem Vergleich der Ergebnisse aus den Erhebungsrunden 2000 und 2003. Für Leserinnen und Leser, die an methodischen Fragen interessiert sind, gibt das Kapitel 12 einige grundlegende Informationen.

Literatur

Adams, R. (Ed.). (in Vorb.). *PISA 2003 technical report*. Paris: OECD.

Adams, R. & Wu, M. (Eds.) (2002). *PISA 2000 technical report*. Paris: OECD.

American Association for the Advancement of Science (AAAS) (Hrsg.). (1993). *Benchmarks for science literacy. Project 2061*. New York: Oxford University Press.

Baumert, J. & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441-462.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J. & Weiß, M. (Hrsg.). (2001). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Bybee, R. W. (1997). Towards an understanding of scientific literacy. In W. Gräber & C. Bolte (Hrsg.), *Scientific literacy. An international symposium* (S. 37-68). Kiel: Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN).

Ehmke, T. (2003). *Mathematical Literacy bei Erwachsenen: Eine Studie an Eltern von PISA-Schülerinnen und -Schülern. DFG-Antrag auf Sachbeihilfe*. Kiel: Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN).

Freudenthal, H. (Hrsg.). (1977). *Mathematik als pädagogische Aufgabe*. Stuttgart: Klett.

Kienzl, A. (in Vorb.). *Überprüfung der Schulnotenverteilungen im Rahmen von PISA 2003 (Diplomarbeit)*. Universität Kiel: Psychologisches Insitut.

Kirsch, I. (1995). Literacy performance on three scales: Definitions and results. In Literacy, economy & society (Ed.), *Results of the first international adult literacy survey* (S. 27-53). Paris, Ottawa: OECD, Statistics Canada.

Klieme, E., Baumert, J., Köller, O. & Bos, W. (2000). Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 1 – Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit* (S. 85-133). Opladen: Leske + Budrich.

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

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M. & Jordan, A. (2004, in Druck). COACTIV: Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz. In J. Doll & M. Prenzel (Hrsg.), *Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung*. Münster: Waxmann.

Köller, O. (2003). Qualitätsmonitoring PISA-E 2003. Unveröffentlichtes Manuskript, Erlangen.

Lipowsky, F., Rakoczy, K., Klieme, E., Pauli, C. & Reusser, K. (2004, in Druck). Hausaufgabenpraxis im Mathematikunterricht – Ein Thema für die Unterrichtsqualitätsforschung? In J. Doll & M. Prenzel (Hrsg.), *Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung*. Münster: Waxmann.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of items. *Journal of Educational Measurement, 29*(2), 133-164.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for mathematics*. Reston,VA/USA: NCTM.

OECD (1999). *Measuring student knowledge and skills. A new framework for assessment*. Paris: OECD.

OECD (2001). *Knowledge and skills for life. First results from PISA 2000*. Paris: OECD.

OECD (2003). *The PISA 2003 assessment framework – Mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.

OECD (2004a). *Learning for tomorrow's world. First results from PISA 2003.* Paris: OECD.

OECD (2004b). *Problem solving for tomorrow's world – First measures of cross-curricular skills from PISA 2003*. Paris: OECD.

O'Neill, H. F., Jr., Sugrue, B. & Baker, E. L. (1996). Effects of motivational interventions on the national assessment of educational progress mathematics performance. *Educational Assessment, 4*(3), 135-157.

Pekrun, R., Götz, T., Vom Hofe, R., Blum, W., Jullien, S., Zirngibl, A., Kleine, M., Wartha, S. & Jordan, A. (2004, in Druck). Emotionen und Leistung im Fach Mathematik: Ziele und erste Befunde aus dem „Projekt zur Analyse der Leistungsentwicklung in Mathematik" (PALMA). In J. Doll & M. Prenzel (Hrsg.), *Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung*. Münster: Waxmann.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Hans Huber.

Seidel, T. & Prenzel, M. (2004, in Druck). Muster unterrichtlicher Aktivitäten im Physikunterricht. In J. Doll & M. Prenzel (Hrsg.), *Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung*. Münster: Waxmann.

Tenorth, H.-E. (2004). Bildungsstandards und Kerncurriculum – Systematischer Kontext, bildungstheoretische Probleme. *Zeitschrift für Pädagogik, 50*(5), 650-661.

Turner, R. (2002). Proficiency scales construction. In R. Adams & M. Wu (Eds.), *PISA 2000: Technical Report.* (S. 195-216). Paris: OECD.

Weinert, F. E. (1999). *Concepts of competence (Contribution within the OECD project. Definition and selection of competencies: Theoretical and conceptual foundations)*. Neuchâtel: DeSeCo.