

Manfred Prenzel
Jürgen Baumert (Hrsg.)

Vertiefende Analysen zu PISA 2006

Zeitschrift für
Erziehungswissenschaft.

Sonderheft 10 | 2008

Zeitschrift für
ZfE Erziehungs-
wissenschaft

SONDERHEFT 10 | 2008



VS VERLAG FÜR SOZIALWISSENSCHAFTEN

Manfred Prenzel · Jürgen Baumert (Hrsg.)

Vertiefende Analysen zu PISA 2006

Zeitschrift für Erziehungswissenschaft
Sonderheft 10 | 2008

Manfred Prenzel
Jürgen Baumert (Hrsg.)

Vertiefende Analysen zu PISA 2006

Zeitschrift für
Erziehungswissenschaft

Sonderheft 10 | 2008



VS VERLAG FÜR SOZIALWISSENSCHAFTEN

Zeitschrift für Erziehungswissenschaft

herausgegeben von:

Jürgen Baumert (Schriftleitung), Hans-Peter Blossfeld, Yvonne Ehrenspeck, Ingrid Gogolin (Schriftleitung), Stephanie Hellekamps, Frieda Heyting (1998–2003), Heinz-Hermann Krüger (Schriftleitung), Dieter Lenzen (Schriftleitung, Geschäftsführung), Meinert A. Meyer, Manfred Prenzel, Thomas Rauschenbach, Hans-Günther Roßbach, Uwe Sander, Annette Scheunflug, Christoph Wulf

Herausgeber des Sonderheftes Vertiefende Analysen zu PISA 2006:

Manfred Prenzel und Jürgen Baumert

Redaktion:

Friedrich Rost (und Rezensionen), Thorsten Junge

Anschrift der Redaktion:

Zeitschrift für Erziehungswissenschaft
c/o Freie Universität Berlin, Arbeitsbereich Philosophie der Erziehung,
Arnimallee 10, D-14195 Berlin
Tel.: (+49)030 838-55888; Fax: -55889
E-Mail: zfe@zedat.fu-berlin.de URL: <http://zfe-online.de> Volltexte: <http://zfe-digital.de>

Beirat:

Neville Alexander (Kapstadt), Jean-Marie Barbier (Paris), Jacky Beillerot † (Paris), Wilfried Bos (Dortmund), Elliot W. Eisner (Stanford/USA), Frieda Heyting (Amsterdam), Axel Honneth (Frankfurt a. M.), Marianne Horstkemper (Potsdam), Ludwig Huber (Bielefeld), Yasuo Imai (Tokyo), Jochen Kade (Frankfurt a. M.), Anastassios Kodakos (Rhodos), Gunther Kress (London), Sverker Lindblad (Uppsala), Christian Lüders (München), Niklas Luhmann † (Bielefeld), Joan-Carles Mèlich (Barcelona), Hans Merckens (Berlin), Klaus Mollenhauer † (Göttingen), Christiane Schiersmann (Heidelberg), Wolfgang Seitter (Marburg), Rudolf Tippelt (München), Gisela Trommsdorff (Konstanz), Philip Wexler (Jerusalem), John White (London), Christopher Winch (Northampton)

VS Verlag für Sozialwissenschaften | GWV Fachverlage GmbH | Abraham-Lincoln-Str. 46 | 65189 Wiesbaden

Geschäftsführer: Dr. Ralf Birkelbach (Vors.)
Albrecht F. Schirmacher

Gesamtleitung Anzeigen: Thomas Werner
Gesamtleitung Produktion: Ingo Eichel
Gesamtleitung Vertrieb: Gabriel Göttlinger

Leserservice: Martin Gneupel, Telefon (0611) 7878-151, Telefax (0611) 7878-423

E-Mail: Martin.Gneupel@gwv-fachverlage.de

Abonnentenbetreuung: Ursula Müller, Telefon (05241) 801965, Telefax (05241) 809620

E-Mail: Ursula.Mueller@gwv-fachverlage.de

Marketing: Ronald Schmidt-Serrière M.A., Telefon (0611) 7878-280, Telefax (0611) 7878-440

E-Mail: Ronald.Schmidt-Serriere@vs-verlag.de

Anzeigenleitung: Yvonne Guderjahn, Telefon (0611) 7878-369, Telefax (0611) 7878-430

E-Mail: Yvonne.Guderjahn@gwv-fachverlage.de

Anzeigendisposition: Monika Dannenberger, Telefon (0611) 7878-148, Telefax (0611) 7878-443

E-Mail: Monika.Dannenberger@gwv-fachverlage.de

Produktion/Layout: Frieder Kumm, Telefon (0611) 7878-175, Telefax (0611) 7878-468

E-Mail: Frieder.Kumm@gwv-fachverlage.de

Bezugsmöglichkeiten 2008: Jährlich erscheinen 4 Hefte. Jahresabonnement / privat (print+online) € 89,-; Jahresabonnement / privat (nur online) € 64,-; Jahresabonnement / Bibliotheken/Institutionen € 139,-; Jahresabonnement / Studenten/Emeritus (print+online) – bei Vorlage einer Studienbescheinigung € 35,-. Alle Print-Preise zuzüglich Versandkosten. Alle Preise und Versandkosten unterliegen der Preisbindung. Die Bezugspreise beinhalten die gültige Mehrwertsteuer. Kündigungen des Abonnements müssen spätestens 6 Wochen vor Ablauf des Bezugszeitraumes schriftlich mit Nennung der Kundennummer erfolgen.

Jährlich können Sonderhefte (Beihefte) erscheinen, die nach Umfang berechnet und den Abonnenten des laufenden Jahrgangs mit einem Nachlass von 25% des jeweiligen Ladenpreises geliefert werden. Bei Nichtgefallen können die Sonderhefte innerhalb einer Frist von drei Wochen zurückgegeben werden.

Zuschriften, die den Vertrieb oder Anzeigen betreffen, bitte nur an den Verlag.

© VS Verlag für Sozialwissenschaften | GWV Fachverlage GmbH, Wiesbaden 2009

VS Verlag für Sozialwissenschaften ist Teil der Fachverlagsgruppe Springer Science+Business Media.

Alle Rechte vorbehalten. Kein Teil dieser Zeitschrift darf ohne schriftliche Genehmigung des Verlages vervielfältigt oder verbreitet werden. Unter dieses Verbot fällt insbesondere die gewerbliche Vervielfältigung per Kopie, die Aufnahme in elektronische Datenbanken und die Vervielfältigung auf CD-ROM und allen anderen elektronischen Datenträgern.

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier.

Printed in the Netherlands

ISBN 978-3-531-15929-4

10. Jahrgang, Sonderheft 10/2008

Inhaltsverzeichnis

Manfred Prenzel und Jürgen Baumert	Editorial: Vertiefende Analysen zu PISA 2006	7
Claus H. Carstensen, Manfred Prenzel und Jürgen Baumert	Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt?	11
Cordula Artelt, Barbara Drechsel, Wilfried Bos und Tobias C. Stubbe	Lesekompetenz in PISA und PIRLS/IGLU – ein Vergleich	35
Katrin Schöps, Martin Senkbeil und Kerstin Schütte	Umweltbezogene Einstellungen von Jugendlichen in Deutschland – Ergebnisse aus PISA 2006	53
Päivi Taskinen, Regine Asseburg und Oliver Walter	Wer möchte später einen naturwissenschaftsbezogenen oder technischen Beruf ergreifen? Berufserwartungen und Schülermerkmale in PISA 2006	79
Martin Senkbeil und Jörg Wittwer	Antezedenzen und Konsequenzen informellen Lernens am Beispiel der Mediennutzung von Jugendlichen	107
Timo Ehmke	Welche Bedeutung haben lernförderliche Prozesse und naturwissenschaftsbezogene Einstellungen im Elternhaus für die Erklärung sozialer Disparitäten in der naturwissen- schaftlichen Kompetenz?	129
Oliver Walter	Herkunftsassoziierte Disparitäten im Lesen, der Mathematik und den Naturwissenschaften: ein Vergleich zwischen PISA 2000, PISA 2003 und PISA 2006	149

Oliver Walter	Ethno-linguale Kompositionseffekte in neunten Klassen: Befunde aus der Klassenstichprobe von PISA 2006	169
Oliver Walter und Päivi Taskinen	Naturwissenschaftsbezogene Motivationen und Kompetenzen von Schülerinnen und Schülern mit Migrationshintergrund in Deutschland: der Einfluss der Generation, der Herkunft und des Elternhauses	185
Jörg Wittwer	What Influences the Agreement Among Student Ratings of Science Instruction?	205
Eckhard Klieme und Brigitte Steinert	Schulentwicklung im Längsschnitt. Ein Forschungsprogramm und erste explorative Analysen	221

ZfE Beiheft „Vertiefende Analysen zu PISA 2006“

MANFRED PRENZEL und JÜRGEN BAUMERT

Editorial:

Vertiefende Analysen zu PISA 2006

Die Erhebungswellen von PISA landeten mit einiger Verzögerung an, nämlich dann, wenn die ersten Berichte der Untersuchungen veröffentlicht werden. So wurden im Dezember 2007 die Ergebnisse der Erhebungen im Sommer 2006 publiziert, die im Rahmen des „Programme for International Student Assessment“ (PISA) in 57 Staaten durchgeführt worden waren.

Alle drei Jahre werden in der von der OECD (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung) koordinierten internationalen Vergleichsstudie Kompetenzen fünfzehnjähriger Schülerinnen und Schüler in den Bereichen Lesen, Mathematik und Naturwissenschaften getestet sowie weitere Schul- und Schülermerkmale mit Fragebögen erfasst. Die OECD verspricht den teilnehmenden Staaten empirisch gestütztes Wissen, das ihnen bei der Steuerung ihrer Bildungssysteme helfen soll. Die Grundanlage der Studie mit umfangreichen repräsentativen Stichproben und ausgezeichneten Testverfahren macht die Untersuchung aber auch aus einer wissenschaftlichen Sicht attraktiv. Für die Forschung interessant wird PISA insbesondere dann, wenn Stichproben ergänzt und zusätzliche Erhebungen durchgeführt werden können, um gezielt weiterführende wissenschaftliche Fragen zu beantworten. Entsprechende zusätzliche Erhebungen und Auswertungen helfen außerdem, die Befunde der internationalen Studie abzusichern und gewissenhaft zu interpretieren. Auch konfrontiert ein Erhebungsprogramm wie PISA immer wieder mit einigen grundlegenden, zum Teil methodischen Fragen, deren Beantwortung nicht nur für die Studie selbst, sondern für die Forschung generell bedeutsam wird. In diesem Sinn haben die in Deutschland für die Durchführung und Auswertung der Studie verantwortlichen Gruppen von Wissenschaftlerinnen und Wissenschaftlern („PISA-Konsortien“) bei den bisherigen drei PISA-Runden konsequent die Perspektive der empirischen Bildungsforschung vertreten (BAUMERT u.a. 2001; PRENZEL u.a. 2004). Die zusätzlichen Erhebungen und Analysen der deutschen PISA-Konsortien ergänzten damit die internationale Berichterstattung der OECD. Gelegentlich mussten jedoch auch auf der Basis vertiefter Analysen manche Aussagen und Interpretationen zurechtgerückt werden.

Die PISA-Erhebungswellen konfrontieren die Öffentlichkeit aber auch mit einer zunehmenden Flut an Informationen. Der erste Bericht der OECD umfasste bei PISA 2006 zwei Bände mit jeweils mehr als 400 Seiten (OECD 2007). Dieser „Initial report“ erfüllt in bestimmter Hinsicht auch die Aufgabe einer Dokumentation von Ergebnissen, nämlich Mittelwerten, Streuungsmaßen oder Kennzahlen für (meist bivariate) Zusammenhänge zwischen Merkmalen. Internationale thematische Berichte, die nachfolgen, werden stärker inhaltlich und theoretisch orientiert sein. Der Bericht zu PISA 2006, der vom PISA-Konsortium in Deutschland (PRENZEL u.a. 2007) zeitgleich mit dem OECD-Report veröffentlicht wurde, berücksichtigte demgegenüber stärker Fragen, die aus der nationalen Perspektive relevant erschienen und bemühte sich von Anfang an darum, diese in einem wissenschaftlichen und theoretisch begründeten Kontext zu beantworten.

Die internationale Datenbasis für die statistischen Analysen und für die nachfolgende Berichterstellung erhalten die nationalen Projektleiter erst vier bis fünf Monate vor dem offiziellen Veröffentlichungstermin. Die Analysen für den ersten Bericht müssen sich deshalb auf Aspekte konzentrieren, die für die Interpretation der internationalen Vergleichsergebnisse vordringlich zu klären sind. Viele Fragen, die aus einer Bildungsforschungsperspektive relevant erscheinen, können zunächst noch nicht aufgegriffen werden. Antworten auf diese Fragen sind aber auch eher für die wissenschaftliche Community interessant und kaum für Schlagzeilen geeignet. Deshalb liegt es auf der Hand, Ergebnisse vertiefender wissenschaftlicher Analysen zu einem späteren Zeitpunkt in einschlägigen wissenschaftlichen Publikationsorganen zu veröffentlichen. In diesem Beiheft stellen wir somit eine Reihe von aktuellen Befunden aus PISA 2006 vor, einer Erhebungsrunde, in der die Naturwissenschaften als Hauptdomäne erstmals den größten Anteil der Testzeit erhielten. Aus diesem Grund stehen in mehreren Beiträgen die Naturwissenschaften im Zentrum. Aber auch die anderen Domänen finden eine angemessene Berücksichtigung.

Der Beitrag von CLAUS CARSTENSEN, MANFRED PRENZEL und JÜRGEN BAUMERT thematisiert einige grundlegende methodische Fragen zur Analyse von Trends über die dreijährlichen Erhebungsrunden. Trendanalysen betreffen die Frage, ob und inwieweit sich die Testleistungen in einem Staat beziehungsweise in allen OECD-Staaten über die Erhebungsrunden verbessert oder verschlechtert haben. Der Beitrag erläutert und diskutiert Voraussetzungen und Möglichkeiten von Trendanalysen in einer internationalen Vergleichsstudie wie PISA.

Über PISA hinaus weist der Artikel von CORDULA ARTELT, BARBARA DREHSEL, WILFRIED BOS und TOBIAS C. STUBBE. Sie vergleichen die Befunde zur Lesekompetenz in PISA und in PIRLS/IGLU – einer Studie, die Schülerinnen und Schüler am Ende der Grundschule untersuchte. In ihrem Beitrag arbeiten sie Besonderheiten der beiden Erhebungen heraus, die einen systematischen Vergleich einschränken und die bei der gemeinsamen Interpretation von Befunden aus beiden Studien berücksichtigt werden sollten.

Mit dem Schwerpunkt Naturwissenschaften bei PISA 2006 bot es sich an, umweltbezogene Einstellungen von Jugendlichen genauer zu untersuchen. KATRIN SCHÖPS, MARTIN SENKBEIL und KERSTIN SCHÜTTE berichten über Umweltwissen, Wahrnehmung von Umweltproblemen sowie Einstellungen zu Umweltmaßnahmen der Schülerinnen und Schüler in Deutschland und beziehen diese Einstellungen, einem theoretischen Modell folgend, auf weitere Merkmale der Personen und ihrer Umgebungen.

Die Schwerpunktsetzung Naturwissenschaften wird ebenfalls im Beitrag von PÄIVI TASKINEN, REGINE ASSEBURG und OLIVER WALTER genutzt. Auf der Grundlage differenzierter Auswertungen der Berufsvorstellungen von Jugendlichen untersuchen sie, welche Schülermerkmale eine Vorhersage von Präferenzen für einen naturwissenschaftsbezogenen oder technischen Beruf erkennen lassen. Diese Frage gewinnt vor dem Hintergrund aktueller Debatten über die Sicherung des Nachwuchses für die naturwissenschaftlich-technischen Berufsfelder besondere Bedeutung.

Einflüsse informeller Lernprozesse auf die Kompetenzentwicklung untersuchen MARTIN SENKBEIL und JÖRG WITTEWITZ anhand der Daten aus PISA 2006. Sie betrachten hier insbesondere die freizeitbezogenen Medienaktivitäten der Jugendlichen unter dem Aspekt von informellem Lernen und berücksichtigen Einflüsse der sozialen Herkunft auf die Mediennutzung.

Während soziale Herkunft meist unter dem Aspekt vertikaler Ungleichheit betrachtet wird, erweitert TIMO EHMKE diese Sicht, indem er nach horizontalen Disparitäten fragt. In seinem Beitrag analysiert er die Bedeutung lernförderlicher und naturwissenschaftsbezogener Einstellungen und Prozesse im Elternhaus für die Erklärung sozial bedingter Unterschiede in der naturwissenschaftlichen Kompetenz.

Mehrere Beiträge im Sonderheft befassen sich mit dem Herkunftsmerkmal Migrationshintergrund und Zusammenhängen mit der Kompetenzentwicklung. OLIVER WALTER greift auf die Daten früherer PISA-Runden zurück, um Zusammenhänge zwischen Migrationsstatus und eth-

nischer Herkunft von PISA 2000 bis PISA 2006 für die deutsche Stichprobe zu beschreiben. Seine Analysen zeigen, dass sich die beobachteten Disparitäten zu einem beträchtlichen Teil auf die soziale Herkunft und den Sprachgebrauch in der Familie zurückführen lassen.

In einem weiteren Beitrag nutzt OLIVER WALTER Daten aus einer Stichprobenerweiterung in Deutschland, bei der in PISA 2006 komplette Klassen der neunten Jahrgangsstufe einbezogen wurden. Mit dieser Klassenstichprobe untersucht er Einflüsse der Zusammensetzung der Schulklasse (in Hinsicht auf den Anteil von Schülerinnen und Schülern mit Migrationshintergrund) auf die drei bei PISA untersuchten Kompetenzbereiche.

Zusammen mit PÄIVI TASKINEN analysiert OLIVER WALTER dann naturwissenschaftsbezogene Kompetenzen und Motivationen von Schülerinnen und Schülern mit Migrationshintergrund. Hintergrund der Analysen sind Annahmen über eher optimistische bzw. pessimistische Einstellungen von Zuwanderern in Deutschland. Die Auswertungen berücksichtigen Einflüsse der Generation, der Herkunft und des Elternhauses.

Die erweiterte Stichprobe kompletter Klassen steht ebenfalls im Zentrum der Analysen von JÖRG WITTE. Er geht der methodisch bedeutsamen Frage nach, inwieweit Einschätzungen des naturwissenschaftlichen Unterrichts innerhalb einer Klasse soweit geteilt werden, dass eine Aggregation der Individualwerte zu einem aussagekräftigen Klassenwert gerechtfertigt erscheint. Seine Analysen liefern Evidenz für diese Annahme, zeigen aber auch, dass das Ausmaß der Übereinstimmung zwischen den Schülereinschätzungen von mehreren Faktoren abhängt.

Der Beitrag von ECKHARD KLIEME und BRIGITTE STEINERT schließlich weist auf eine Besonderheit der PISA-Erhebungen in Deutschland hin: Aufgrund der Stichprobenerweiterungen für Ländervergleiche gibt es eine ganze Reihe von Schulen, die mehrmals bei den PISA-Erhebungsrunden getestet wurden. Die Besonderheit bietet die Chance, Veränderungen der Schulen im Verlauf der Zeit abzubilden und diese auch im Zusammenhang mit möglichen Veränderungen der Kompetenzwerte zu betrachten. Anhand von ersten explorativen Auswertungen weisen die Autoren auf innovative Möglichkeiten hin, theoriegeleitet die Schulentwicklung über längere Zeiträume in diesem besonderen Sample zu untersuchen.

Literatur

- BAUMERT u.a. = BAUMERT, J./KLIEME, E./NEUBRAND, M./PRENZEL, M./SCHIEFELE, U./SCHNEIDER, W./STANAT, P./TILLMANN, J./WEISS, M. (Hrsg.), (2001): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. – Opladen: Leske & Budrich.
- OECD (2007): PISA 2006. Schulleistungen im internationalen Vergleich. Naturwissenschaftliche Kompetenzen für die Welt von morgen. Paris – OECD.
- PRENZEL u.a. 2007 = PRENZEL, M./ARTELT, C./BAUMERT, J./BLUM, W./HAMMANN, M./KLIEME, E./PEKRUN, R. (Hrsg.), (2007): PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie. – Münster.
- PRENZEL u.a. 2004 = PRENZEL, M./BAUMERT, J./BLUM, W./LEHMANN, R./LEUTNER, D./NEUBRAND, M./PEKRUN, R./ROLFF, H.-G./ROST, J./SCHIEFELE, U. (Hrsg.), (2004): PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. – Münster.

Claus H. Carstensen¹, Manfred Prenzel¹ und Jürgen Baumert²

Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt?

Zusammenfassung:

Die Veröffentlichungen von Ergebnissen der letzten PISA-Erhebung geben Anlass, Möglichkeiten und Grenzen von Trendanalysen in internationalen Vergleichsstudien mit komplexen Testdesigns zu untersuchen. Der Beitrag schließt an Trendanalysen an, die in Deutschland bei PISA 2006 durchgeführt wurden. In einem ersten Schritt werden Voraussetzungen für Trendanalysen bei Studien wie PISA benannt, um dann anhand der Daten aus den drei Erhebungen in Deutschland zu prüfen, ob sie erfüllt sind. Im Zentrum des Beitrags steht die Diskussion eines Modells für Trendanalysen innerhalb eines Staates, das auch dann angewendet werden kann, wenn bestimmte Voraussetzungen für Trendanalysen mit den internationalen Kompetenzindikatoren nicht gegeben sind. Die Ergebnisse einer Anwendung dieses Modells auf die deutschen Daten für alle drei Kompetenzbereiche (Lesen, Mathematik, Naturwissenschaften) werden vorgestellt und interpretiert. Dabei werden die Einschränkungen bezüglich einer international vergleichenden Interpretation hervorgehoben. Im Ausblick werden weiterführende Möglichkeiten diskutiert, in einer Studie wie PISA mit denselben Kompetenzindikatoren sowohl querschnittliche Vergleiche als auch tragfähige Trendanalysen durchführen zu können.

Abstract:

The publication of the results of the last PISA study prompts an examination of the possibilities and limitations of trend analyses of international comparative studies with complex test designs. This paper follows up on trend analyses which were carried out in Germany for PISA 2006. In a first step, the requirements for trend analyses of studies such as PISA will be stated in order to then test whether they are fulfilled by using the data from the three PISA cycles in Germany. The discussion of a model for trend analyses within a country which can even be used if certain requirements for trend analyses using the international competency indicators are not fulfilled is at the core of this paper. The results of an application of this model to the German data for all three domains (reading, mathematics, science) will be presented and interpreted. The limitations with regard to an interpretation at the level of an international comparison will thereby be highlighted. With regard to future prospects, further possibilities of using the same competency indicators to carry out both cross-sectional comparisons and also sustainable trend analyses in a study such as PISA will be discussed.

1 Internationale Vergleiche und die Analyse von Trends

Das „Programme for International Student Assessment“ (PISA) erhebt im Abstand von drei Jahren grundlegende Kompetenzen fünfzehnjähriger Schülerinnen und Schüler in den Bereichen Lesen, Mathematik und Naturwissenschaften. Die von der OECD koordinierte Studie untersucht, inwieweit es den Bildungssystemen der teilnehmenden Staaten gelingt, Jugendliche auf Herausfor-

¹ Leibniz-Institut für die Pädagogik der Naturwissenschaften an der Universität Kiel

² Max-Planck-Institut für Bildungsforschung, Berlin

derungen der Wissensgesellschaft vorzubereiten (OECD 1999, 2006). Wichtige Bezugspunkte für die Analysen liefert der *internationale Vergleich*. Die Kompetenzen der getesteten Jugendlichen werden an der Gruppe der OECD-Staaten normiert. Anhand der Berichtsskalen (mit einem Mittelwert von 500 und einer Standardabweichung von 100 Punkten) können die in den Staaten erreichten Leistungen verglichen und auf Kontextindikatoren (z.B. soziale Herkunft) bezogen werden. Die bei PISA verwendete Skala gibt den teilnehmenden Staaten Auskunft über ihre Position im internationalen Vergleich und über ihre relativen Schwächen und Stärken.

Aber PISA verspricht noch mehr Information, denn in allen bisherigen PISA-Runden wurden die drei Kompetenzen auf der Grundlage vergleichbarer Testkonzeptionen erhoben. Zwar wurde jeweils eine Domäne als Schwerpunktgebiet ausführlicher untersucht, doch stand auch für die anderen Gebiete ausreichend Testzeit zur Verfügung, um Vergleiche über die Erhebungsrunden zu ermöglichen (OECD 2001). Um die Vergleichbarkeit zu gewährleisten, wurde in allen Erhebungsrunden eine Menge gleicher (geheim gehaltener) Aufgaben eingesetzt. Die Untersuchung war und ist prinzipiell so angelegt, dass *Kompetenzveränderungen über die Zeit* (Trends) analysiert werden können. Diese Trendinformationen berichten, inwieweit die fünfzehnjährigen Jugendlichen in aufeinander folgenden Erhebungsrunden ein höheres, gleiches oder niedrigeres Kompetenzniveau erreicht haben. Die Leistungen der einzelnen Staaten zu den Erhebungszeitpunkten werden also an einem inhaltlichen Maßstab beurteilt.

Diese besondere Anlage des Erhebungsdesigns wurde bereits bei den Auswertungen zu PISA 2003 ausgeschöpft, als erstmals Trendanalysen möglich waren. Der entsprechende Bericht der OECD (2004) präsentiert Trends für alle drei Domänen: in den Bereichen Lesen und Naturwissenschaften für die Gesamtskalen, in der Mathematik für zwei Subskalen. Im Bericht zu PISA 2006 stellt die OECD (2007a) wiederum Trends dar, jedoch unterscheiden sich die Auswertungsverfahren zwischen den Domänen und zum Teil auch von denen in früheren Berichten erheblich. Nur im Bereich Lesen wird der Trend von PISA 2000 zu PISA 2003 mit den neuen Zahlen von PISA 2006 fortgeschrieben. Für die mathematische Kompetenz wird ein Trend berichtet, der nun mit der Erhebung im Jahr 2003 (als Mathematik Schwerpunktgebiet war) beginnt. Für die Naturwissenschaften wird argumentiert, dass ein Trend erst ab der Erhebung im Jahr 2006 (mit den Naturwissenschaften als Schwerpunkt) analysiert werden kann (OECD 2007a, S. 26). Dessen ungeachtet findet sich im Anhang des Berichtsbandes (OECD 2007a, S. 370) eine Tabelle, die zwei unterschiedliche Trends für die Naturwissenschaften zwischen PISA 2003 und PISA 2006 darstellt. Die Kennwerte dieser Tabelle werden jedoch nicht erläutert oder kommentiert.

Wie unübersichtlich die Lage ist, lässt sich an einigen Beispielen darstellen: Legt man zum Beispiel die Ergebnisberichte für die *Naturwissenschaften* aus den internationalen Berichten der OECD (2001, 2004, 2007a) nebeneinander, dann erhält Deutschland auf der PISA-Skala im Jahr 2000 487 Punkte, im Jahr 2003 502 Punkte und im Jahr 2006 516 Punkte. Folgt man den Angaben der OECD im zweiten Berichtsband (OECD 2004), dann beziehen sich die Kennwerte für PISA 2000 und PISA 2003 auf die gleiche Skala. Der Abstand von 15 Punkten, der für Deutschland zwischen 2000 und 2003 errechnet wird, kann also auch als substantieller Kompetenzzuwachs verstanden werden. Für die bei PISA 2006 gemessenen 516 Punkte ist es laut OECD jedoch nicht möglich, sie auf die früheren Kennwerte zu beziehen, um einen Trend abzulesen. Sie können zunächst nur als Ergebnis des Vergleichs zwischen den Staaten zu diesem Messzeitpunkt verstanden werden. Zu diesem Erhebungszeitpunkt im Jahr 2006 schneidet Deutschland – was die Position im OECD-Vergleich anbetrifft – freilich gegenüber den früheren Erhebungen sehr gut ab, nämlich überdurchschnittlich (für PISA 2003 waren durchschnittliche, für PISA 2000 unterdurchschnittliche Werte berichtet worden). Ähnlich kompliziert ist die Lage in der *Mathematik* für die Erhebungen bei PISA 2000 und PISA 2003. Zwar werden für beide Runden Gesamtleistungen in der Mathematik berichtet (nämlich 490 Punkte bei PISA 2000 und 503 Punkte bei PISA 2003), doch sind diese wiederum nur jeweils für sich innerhalb einer Studie im Sinne eines Staa-

tenvergleichs zu interpretieren. Wenn man etwas über Trends erfahren will, muss man die Ergebnisse auf zwei Teilskalen betrachten.

Diese Unübersichtlichkeit der Befundlage ist keineswegs beabsichtigt. Sie lässt sich vielmehr darauf zurückführen, dass PISA mit einem methodisch komplexen und ambitionierten Erhebungs- und Auswertungsansatz an mehreren Stellen Neuland betreten hat und manchmal mit unvorhergesehenen Problemen konfrontiert wurde. Da PISA als Indikatorenprogramm für längere Zeiträume geplant ist, gilt es diese Probleme zu meistern. Damit stellt sich die Frage, ob es Möglichkeiten und Wege gibt, die Probleme von Trendanalysen bei internationalen Vergleichsstudien besser zu beherrschen, so dass auch im Fall von Komplikationen die vorhandene Information soweit wie möglich genutzt wird, um Aussagen über substantielle Veränderungen treffen zu können.

Die Wissenschaftlerinnen und Wissenschaftler, die in Deutschland (parallel zur OECD) an eigenen Auswertungen arbeiteten (PRENZEL u. a. 2007a), hatten bei PISA 2006 einen Ausweg gesehen, für die Bereiche Naturwissenschaften und Mathematik doch Trendanalysen auf nationaler Ebene durchzuführen. In der Beschreibung dieses Zugangs (CARSTENSEN u. a. 2007; PRENZEL u. a. 2007b) war unterstrichen worden, dass dieser Zugang nur landesspezifisch möglich ist. Aus diesem Grund war für den Bericht über Zuwächse nicht die übliche PISA-Metrik verwendet worden. Die Veränderungen wurden vielmehr in Effektstärken ausgedrückt.

Der vorliegende Beitrag schließt an diese Analysen an. Am Beispiel der in Deutschland erhobenen Daten wird eine Methode zur Analyse von Trends erläutert, mit deren Hilfe Kompetenzveränderungen in einem Staat über alle Erhebungen berichtet werden können, ohne gleichzeitig den Anspruch zu erheben, die Vergleichbarkeit zwischen den Staaten aufrecht zu erhalten. Mit diesem Beitrag wollen wir nicht nur das in Deutschland bei PISA 2006 gewählte Vorgehen darlegen, sondern uns genereller mit den Möglichkeiten von Trendanalysen über die PISA-Erhebungen hinweg befassen. Dazu sollen die Fragerichtungen und Voraussetzungen eines Staatenvergleichs zu einem Erhebungszeitpunkt einerseits und einer Analyse von Veränderungen über Erhebungen andererseits gegenübergestellt werden. Auf dieser Grundlage kann dann geklärt werden, ob sich anhand der in Deutschland mit PISA erhobenen Daten belastbare Trendindikatoren für die drei Domänen Lesen, Mathematik und Naturwissenschaften über alle bisherigen Erhebungsrunden konstruieren lassen.

2 Skalierung und Trendanalyse

Aus den in PISA erhobenen Daten werden Indikatoren für Kompetenzen konstruiert, die zwischen den teilnehmenden Staaten verglichen werden können. Die Konstruktion dieser Indikatoren ist an eine Reihe von Voraussetzungen geknüpft, die im Folgenden benannt werden, um dann zu klären, welche Voraussetzungen für Vergleiche *zwischen* Staaten zu einem Erhebungszeitpunkt und für Trendanalysen *innerhalb* eines Staates gelten.

Die Konstruktion valider Kompetenzindikatoren im Rahmen internationaler Vergleichsstudien beruht auf einigen grundlegenden Annahmen. Die zu untersuchenden Kompetenzen müssen in allen Ländern in vergleichbarer Weise erfassbar sein. Deshalb werden theoretische Rahmenkonzeptionen für die Testentwicklung von internationalen Expertengruppen erarbeitet (OECD 1999, 2003, 2006). Sie strukturieren international vergleichbare Kompetenzbereiche und Felder für die Aufgabenentwicklung. Mit den Aufgaben (und mit dem Testdesign) müssen in allen teilnehmenden Staaten in gleicher Weise die intendierten Kompetenzen erfasst werden können. Die Testaufgaben müssen also in ihrer Gesamtheit in allen Staaten dasselbe Konstrukt messen, das heißt zum Beispiel, keine Aufgabe darf in einzelnen Staaten aufgrund kultureller Unterschiede unterschiedliche Bedeutung erhalten. Andernfalls wäre nicht entscheidbar, ob ein beobachteter Unter-

schied auf das durchschnittliche Können der Schülerinnen und Schüler oder auf kulturelle Traditionen zurückzuführen ist.

Um die Kompetenzverteilungen in den Staaten auf der Grundlage einer ausreichend großen Zahl von Aufgaben schätzen zu können, wird eine Anzahl von Testheften eingesetzt, welche jeweils unterschiedliche Teilmengen der Testaufgaben enthalten. Der Testaufbau bei PISA folgt einem Multi-Matrix-Design, bei dem jede Schülerin und jeder Schüler jeweils nur einen Teil der gesamten Aufgabenmenge (bei PISA 2006 31%) bearbeitet. Da die Testhefte aus unterschiedlichen (systematisch rotierten) Aufgaben zusammengestellt werden, müssen die Schwierigkeiten der einzelnen Aufgaben bei der Konstruktion der Kompetenzindikatoren berücksichtigt werden. Dazu wird in PISA das so genannte Rasch-Modell (RASCH 1960), eines der grundlegenden Modelle der Item Response-Theorie, als Messmodell verwendet. Sind die mit dem Rasch-Modell verbundenen Annahmen für den Test in der untersuchten Population in einer erhobenen Domäne erfüllt, so kann von einer homogenen, d. h. eindimensionalen Kompetenzskala ausgegangen werden und es lassen sich Messwerte auf dieser Skala errechnen. Die Messwerte sind unter diesen Voraussetzungen zwischen allen Staaten und Subgruppen innerhalb der Staaten vergleichbar.

Zu den Annahmen des Rasch-Modells zählen die Itemhomogenität, die Personenhomogenität und die lokale stochastische Unabhängigkeit (RASCH 1960; ROST 2004). Mit den ersten beiden Annahmen werden Voraussetzungen der Eindimensionalität eines Testes bezüglich der Zielpopulation beschrieben. Für alle Schülerinnen und Schüler muss dasselbe Messmodell gelten, d. h. die Schwierigkeiten der Testaufgaben müssen für alle Schülerinnen und Schüler gleich sein. Umgekehrt muss jede einzelne Aufgabe die gleichen Messeigenschaften aufweisen (die gemessene Kompetenz für alle Personen ist unabhängig von der Auswahl der Aufgaben). Das heißt, bei jeder beliebigen Auswahl von Aufgaben aus dem Test darf sich die Reihenfolge der Personen hinsichtlich ihrer Kompetenz, zum Beispiel gemessen durch den Anteil von richtigen Lösungen, nicht unterscheiden. Aus diesen Annahmen lässt sich die so genannte Suffizienz der Rohwerte ableiten: Wenn die geforderte Eindimensionalität der Testaufgaben vorliegt, lassen sich die Fähigkeitsunterschiede bereits an den Rohwerten, d. h. an der Zahl der richtig gelösten Aufgaben, ablesen.

Weiter wird mit dem Rasch-Modell die so genannte lokale stochastische Unabhängigkeit angenommen: Es wird vorausgesetzt, dass die Schwierigkeit einer Aufgabe nicht von anderen Aufgaben im Test abhängt. So darf zum Beispiel die Lösung einer Aufgabe nicht von der Lösung der vorangehenden Aufgabe abhängen. Bei der Konstruktion der 13 Testhefte in PISA 2003 und PISA 2006 wurden die einzelnen Testhefte jeweils mit vier aus insgesamt dreizehn Aufgabenblöcken gebildet, wobei jeder Aufgabenblock einmal an jeder der vier möglichen Positionen im Testheft angeordnet war. Mit einer solchen Zuordnung ändert sich der Kontext eines Aufgabenblocks durch den jeweils vorangehenden Aufgabenblock. Potentielle Auswirkungen dieses Kontexteffekts auf das Testergebnis sind mit dem gewählten Testdesign nicht kontrollierbar und es muss angenommen werden, dass diese zu vernachlässigen sind. Mögliche Verzerrungen, die sich mit dem Testheftdesign ergeben könnten, werden durch eine zufällige und gleichverteilte Zuordnung von Schülerinnen und Schülern zu den Testheften in jedem teilnehmenden Staat ausgeschlossen. Damit ist die Vergleichbarkeit zwischen den Staaten *innerhalb einer PISA-Erhebung* gegeben.

Will man die Kompetenzergebnisse auch für Trendanalysen heranziehen, muss nachgewiesen werden, dass alle Voraussetzungen für die Anwendung des Rasch-Modells gleichzeitig *über alle Erhebungszyklen* hinweg gegeben sind. Nur dann sind die Messwerte zwischen den Erhebungen vergleichbar. Da mit den Erhebungsrunden die Schwerpunkte wechseln, werden Kompetenzen in den einzelnen Domänen einmal mit einem umfangreichen Test von 84 oder mehr Aufgaben erfasst und in anderen Zyklen lediglich mit etwa 30 Aufgaben. Da nicht alle Aufgaben in den folgenden Erhebungen wieder verwendet werden, ist die Zahl der gemeinsamen Aufgaben zwischen zwei Erhebungen meistens nicht größer als 25. Für Vergleiche über die Erhebungen hinweg ist deshalb sicherzustellen, dass die unterschiedlichen Testformen, gleichgültig ob Lang-

oder Kurzform, dasselbe Konstrukt erfassen. Dieser Gesichtspunkt wird bei der Auswahl der Aufgaben für die folgenden Studien durch die Expertengruppen berücksichtigt. Außerdem lässt sich der Zusammenhang zwischen den kürzeren und längeren Testformen empirisch bestimmen. Schließlich muss gewährleistet sein, dass die Aufgaben und Aufgabenblöcke in den unterschiedlichen Erhebungen unter denselben Bedingungen vorgegeben werden. Ideal wäre es daher, dieselben Testhefte zu verwenden. Da sich jedoch mit den Erhebungen und ihren Schwerpunkten die Zahlen von Aufgaben je Domäne ändern, muss bei der Konstruktion neuer Testhefte sichergestellt werden, dass die Aufgaben unter möglichst vergleichbaren Bedingungen eingesetzt werden. Da die Position der Aufgaben innerhalb der Aufgabenblöcke im Testdesign von PISA nicht variiert wird, ist eine empirische Kontrolle möglicher Auswirkungen von unterschiedlichen Positionen der Aufgaben in den Testheften nicht möglich. Insofern müssen wir annehmen, dass mit dem Testdesign sichergestellt wurde, dass mögliche Positioneffekte von Aufgaben, die zur Schätzung von Trends verwendet werden, über die Erhebungen zu vernachlässigen sind. Auch die bereits erwähnten Kontexteffekte werden mit dem PISA-Testdesign nicht variiert und sind somit nicht empirisch kontrollierbar. Hier muss ebenfalls angenommen werden, dass diese Effekte zwischen den Erhebungen zwar geringfügig zur Fehlervarianz beitragen, aber keine Verzerrung der Ergebnisse zugunsten einer Erhebung mit sich bringen.

Sind die genannten Voraussetzungen nicht erfüllt, können unterschiedliche Kompetenzergebnisse in den Erhebungsrunden nicht allein auf Veränderungen der Kompetenzen der Jugendlichen in den Ländern zurückgeführt werden. Vielmehr können auch andere Faktoren, zum Beispiel eine vermehrte oder verringerte Beschäftigung mit den Themenbereichen bestimmter Aufgaben in einzelnen Staaten zu bestimmten Erhebungszeitpunkten, Unterschiede in den verwendeten Testformen oder Positionseffekte der Aufgaben zu Unterschieden in den Kompetenzergebnissen beitragen.

Faktisch mussten in allen Testdomänen bei Übergängen von einer Haupt- zu einer Nebenkomponekte oder von einer Neben- zu einer Hauptkomponente aus inhaltlichen Gründen Aufgabenpositionen in erheblichem Maße verändert werden. Aufgaben wurden zum Teil zu neuen Clustern zusammengestellt, und Cluster erhielten unterschiedliche Positionen im Testablauf. Das OECD-Dokument EDU/PISA/GB(2007)42 (OECD, 2007b) gibt einen Überblick über diese Veränderungen. Welche Auswirkungen die Testmodifikationen auf Trendschätzungen in den einzelnen Domänen haben, lässt sich nicht exakt angeben. Mit nennenswerten Trendverschätzungen ist zu rechnen, wenn Positions- und Kontextänderungen systematisch und über alle Länder hinweg die Schwierigkeiten der gemeinsamen Aufgaben zwischen den Erhebungen beeinflussen und – technisch gesprochen ein „item drift“ auftritt, der nicht wirklich korrigierbar ist. An diesem Punkt werden die Grenzen des internationalen Testdesigns sichtbar.

Die hier benannten Annahmen sind komplex und in ihrer Gesamtheit sehr restriktiv. Allerdings kann eine Konstruktion von Kompetenzindikatoren vertretbar sein, auch wenn nicht alle Annahmen vollständig erfüllt sind. Dann muss freilich bei der Konstruktion der Indikatoren und bei der Interpretation der Ergebnisse berücksichtigt werden, inwieweit sich Einschränkungen auf die intendierten Vergleiche auswirken können. Je nach Art der beabsichtigten Vergleiche können sich unterschiedliche Einschränkungen ergeben. Wie durch die Berichterstattung zu PISA 2006 deutlich geworden ist, sind Trendindikatoren nicht so leicht zu konstruieren wie Kompetenzindikatoren für querschnittliche Vergleiche zwischen Staaten zu einem Erhebungszeitpunkt.

Die Voraussetzungen für Vergleiche innerhalb der PISA-Studien sind normalerweise gegeben. Durch die Rahmenkonzeptionen werden international vergleichbare Kompetenzen beschrieben, die durch die von Expertengruppen entwickelten und sorgfältig überprüften Aufgaben erfasst werden können (ADAMS/ WU 2002; OECD 2005, in Vorbereitung). Für die Skalierung werden die für die OECD-Staaten gültigen Aufgabenschwierigkeitsparameter in allen Staaten verwendet. Damit wird das erfasste Konstrukt in jeder Domäne in jedem Staat identisch definiert. Bei der