Christoph Kulgemeyer und Horst Schecker

PISA 2000 bis 2006 – Ein Vergleich anhand eines Strukturmodells für naturwissenschaftliche Aufgaben

From PISA 2000 to PISA 2006 – A Model-Based Comparison of Scientific Literacy Items

Zusammenfassung
Anhand eines Strukturmodells zur Beschreibung naturwissenschaftlicher Aufgaben wird ein Vergleich der veröffentlichten PISA-Units bisheriger Testdurchläufe vorgenommen. Dabei wird untersucht, welche Veränderungen zwischen den Testdurchläufen 2000 und 2003 einerseits sowie 2006 andererseits vorgenommen wurden. Dies wurde möglich, da zu PISA 2006 weit mehr naturwissenschaftliche Aufgaben veröffentlicht sind als zu den Vorgängerstudien. Für die Durchführung des Vergleichs wurden bestehende Kriterien zur Beschreibung von Aufgaben genutzt und ergänzend Erkenntnisse kognitionspsychologischer Textverständlichkeitsforschung sowie Untersuchungen zur Verbesserung der Aufgabenkultur und der Modellierung naturwissenschaftlicher Kompetenz einbezogen. Die Ergebnisse führen zu Empfehlungen, wie Aufgaben gestaltet werden müssen, die den Anspruch haben, PISA-ähnlich zu sein.
Schlüsselworte: PISA-Aufgaben, Aufgabenparameter, Textverständlichkeit, Kompetenzmodell

Abstract
This paper describes how published PISA-units changed their characteristics during the three test periods. We differentiate between PISA 2000 and 2003 on the one hand and PISA 2006 on the other. The units are analysed following a structural model of item-characteristics. The study became possible because of the large number of published scientific-literacy-units in connection with PISA 2006. Existing criteria for the description of science items were revised by integrating aspects of text comprehension and approaches for modelling students' competence in science. The results are used to formulate guidelines for the construction of PISA-similar items.
Keywords: PISA-units, item characteristics, text comprehension, competence model

1 Einleitung

Seit im Jahr 2000 die erste PISA-Studie durchgeführt wurde und der anschließende Schock über das mäßige Abschneiden deutscher Schüler einen allgemeinen Reflexionsprozess ausgelöst hat, sind auch die Aufgaben der Tests im Fokus der Öffentlichkeit. PISA-Aufgaben lösen zu können, wird in der Öffentlichkeit als Bildungsziel wahrgenommen – obwohl die so bezeichneten Aufgaben mit den wirklichen PISA-Aufgaben oft wenig gemein haben. Auch in der fachlichen Diskussion um PISA sind die Testaufgaben des Öfteren herber Kritik ausgesetzt. Diese Kritik entzündet sich jedoch zumeist an Details (Schmidt, 2004), an der Praxis des Testens an sich (Meyerhöfer, 2005, 15ff.) oder an testtheoretischen Einzelheiten (Rindermann, 2006). Eine systematische Auseinandersetzung mit PISA-Aufgaben hat jedoch bisher kaum stattgefunden. In der Naturwissenschaftsdidaktik ist der Grund dafür wohl, dass aus den Durchgängen 2000 und 2003 zusammen nur sehr wenige Naturwissenschaftsaufgaben veröffentlicht wurden. Mit dem Durchgang 2006, seinem Fokus auf naturwissenschaftlicher Kompetenz und der entsprechend hohen Zahl auch veröffentlichter Aufgaben bestehen für eine Analyse nunmehr deutlich bessere Voraussetzungen, auch wenn die Fallzahlen nach wie vor gering sind. In dieser Arbeit werden Charakteristika veröffent-

lichter PISA-Aufgaben der bislang stattgefundenen Testdurchläufe 2000, 2003 und 2006 analysiert. Untersuchungsgegenstand sind dabei die naturwissenschaftlichen Aufgaben. Der Fokus der Arbeit liegt dabei auf der Beschreibung von Veränderungen, die zwischen den Aufgaben verschiedener Testdurchläufe messbar sind, nicht darin, PISA-Aufgaben mit anderen Aufgabentypen zu vergleichen. Als Voraussetzung dafür müssen geeignete Kategorien für die Aufgabenbeschreibung gefunden werden. Auf Basis der gefundenen Aufgabenmerkmale sollen schließlich Empfehlungen für die Konstruktion PISA-ähnlicher Aufgaben gegeben werden.

2 Forschungsstand

Die formale Behandlung von Aufgaben kann in der Physikdidaktik unter zwei Gesichtspunkten gegliedert gesehen werden:

1. Ansätze zur *Beschreibung* von Aufgaben
2. Ansätze zur *Beurteilung* von Aufgaben

Der erste Ansatz hat den Anspruch, möglichst viele Charakteristika einer Aufgabe zu erfassen und darzustellen. Das Resultat ist ein Datenblatt der Aufgabe – und somit eine Art Aufgaben-Steckbrief (Fischer & Draxler, 2002). Der zweite Ansatz versucht, Aussagen über den didaktischen Wert von Aufgaben zu treffen. Dies geschieht zum Beispiel mithilfe eines aus empirischen Erkenntnissen gewonnenen Kriterienkatalogs oder durch Expertisen und bezieht auch Rahmenbedingungen, also beispielsweise die unterrichtliche Vernetzung, mit ein. Überlegungen dieser Art stehen im Zusammenhang mit globalen Aussagen über idealtypische Aufgaben und ihre Verwendung. Untersuchungen zur so genannten „neuen Aufgabenkultur“ lassen sich hier einordnen (z.B. Häußler & Lind, 1998).

Zur systematische Erfassung von Aufgabenmerkmalen können beide Ansätze verbunden werden: Kriterien, die anerkanntermaßen gute Aufgaben ausmachen, werden in einem Kategoriensystem festgehalten und anhand dieses Systems untersucht, zu welchem Grade die eingestuften Aufgaben die Kategorien erfüllen.

Das Modell von Fischer und Draxler (2002) kann in einigen Punkten diesen Anspruch erfüllen – es ist von der Voraussetzungen her besonders systematisch und umfangreich. Darüber hinaus ist es verbreitet und gut dokumentiert. In dieser Arbeit muss es jedoch verändert und ergänzt werden, um den Ansprüchen der Beschreibung von PISA-Aufgaben zu genügen. Dies geschieht hier mithilfe einiger Kriterien aus Arbeiten zur neuen Aufgabenkultur, Ergebnissen der Textverständlichkeitsforschung und unter Einbezug des zu den neuen Bildungsstandards anschlussfähigen Bremen-Oldenburger Kompetenzmodells (im Detail dargestellt in Kapitel 3.1).

2.1 Das Modell von Fischer und Draxler

Das Modell von Fischer und Draxler (2002) ist als ein Instrument zur kriterienorientierten Einschätzung bestehender Aufgaben gedacht. Seine Bedeutung sehen die Autoren in der systematischen Auflistung einer Vielzahl unterschiedlicher Ansprüche, die an Aufgaben in Lehr- und Prüfungssituationen gestellt werden. Das Modell ist sehr umfangreich, detailreich und gut dokumentiert – es erfüllt daher die Voraussetzungen, eine Vielzahl unterschiedlicher Test- und Lernaufgaben adäquat zu beschreiben.

Fischer und Draxler bedienen sich für ihr Modell im wesentlichen zweier Quellen, einerseits den im Rahmen des BLK-Programms zur Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts ermittelten Ansätzen zum Umgang mit Aufgaben (Häußler & Lind, 1998) und andererseits Untersuchungen im Rahmen von TIMSS/III (Klieme, 2000). Daraus folgern sie sechs zur Beschreibung von Aufgaben zentrale Kategorien, die im Folgenden vorgestellt werden.

Inhaltliche und curriculare Einordnung:

Diese Kategorie soll die Zuordnung von Aufgaben zu den physikalischen Teilgebieten wie Mechanik, Elektrodynamik, etc. beinhalten (inhaltliche Einordnung). Außerdem soll eine Einordnung gemäß der in den jeweiligen Curricula vorgeschriebenen Sachthe-

men erfolgen (curriculare Einordnung). Dazu gehört auch eine Bewertung der interesseweckenden alltagsweltlichen Aspekte.

Lösungswege:
Es wird in dieser Kategorie unterschieden zwischen „experimentellen Lösungen", „halbquantitativen Lösungen", die durch die Interpretation von Graphen oder Wertetabellen geschehen, „rechnerischen Lösungen", die unter Zuhilfenahme einer Formel gegebene Daten behandelt und „theoretischen Lösungen", die die Verwendung physikalischer Konzepte voraussetzen. Wenn mehrere Lösungswege möglich sind, erfolgt eine multiple Einschätzung.

Antwortformat, Offenheit und Experimentierverhalten:
Fischer und Draxler unterscheiden zwischen drei möglichen Antwortformaten: „Multiple Choice-Aufgaben", „Kurzantwort-Aufgabe", die die eigenständige Formulierung eines Satzes, einer Zahl oder einer kurzen Rechnung fordern und „Aufgaben mit erweitertem Antwortformat", die ausführliche Rechnungen, Beweise oder sogar Aufsätze benötigen. Des Weiteren soll eine Einschätzung der Eindeutigkeit des vorgegebenen Lösungsweges vorgenommen werden. Es werden drei Stufen unterschieden, die mit abnehmender Offenheit die Aufgabe beschreiben – von Stufe 1, bei der mehrere Lösungswege möglich sind und die Aufgabe keinen impliziert, bis Stufe 3 bei der ein eindeutiger Lösungsweg skizziert wird. Im Falle der Stufen 2 oder 3 wird zusätzlich zwischen den Intensitäten der Vorgabe fein differenziert. Für experimentelle Aufgaben wird ein darauf angepasstes Kriteriensystem verwendet.

Kompetenzstufen:
Die Kompetenzstufenzuweisung zu einer Aufgabe soll bei Fischer und Draxler zunächst ein Maß für die Schwierigkeit dieser Aufgabe sein. In der neueren fachdidaktischen Forschung wird dies als ein explizites (Teil-)Ziel zu entwickelnder Kompetenzmodelle gesehen (Klieme et al., 2003), das aktuell verstärkt erforscht wird (z.B. Einhaus, 2007). Klar ist jedoch, dass der bei Fischer und Draxler skizzierte Weg mithilfe eines eindimensionalen Kompetenzmodells einer empirischen Überprüfung nicht standhalten kann und weitergehende empirische Untersuchungen nötig sind – die Frage der Stufung von Kompetenz muss aktuell als noch nicht gelöst angesehen werden. Die Kompetenzstufenzuweisung soll außerdem eine genaue Angabe der in die Aufgabe einzubringenden oder an ihr zu erlernenden Kompetenzen ermöglichen. Fischer und Draxler unterscheiden sechs Kompetenzstufen, die ein hierarchisches System bilden und sich an post hoc Untersuchungen zu TIMSS orientieren (Klieme, 2000). Die Stufen erstrecken sich von „Anwenden naturwissenschaftlichen Alltagswissens" (Stufe 1) bis „Überwinden von Fehlvorstellungen" (Stufe 6).

Anforderungsmerkmale:
Die Anforderungsmerkmale beschreiben detailliert die zur Lösung der Aufgabe notwendigen Fähigkeiten. Dazu wird ein Katalog von Merkmalen der Aufgaben entworfen, der sich an Klieme (2000, 72ff.), anlehnt. Es wird vorgeschlagen, jedem der 16 Merkmale einen Wert auf einer Skala von 0 (= nicht von Bedeutung) bis 2 (= ohne dieses Merkmal nicht zu lösen) zuzuordnen (Fischer & Draxler, 2002, 310). Ausgewählte Merkmale lauten:

1. Überwindung von Fehlvorstellungen
2. Kenntnis älterer Unterrichtsinhalte
3. Fähigkeiten des Problemlösens

Fischer und Draxler weisen darauf hin, dass die Zuordnung zu den Anforderungsmerkmalen nicht unabhängig von den Lösungswegen erfolgen könne. Aus diesem Grunde sei es vonnöten – genau wie bei der Zuweisung der Kompetenzstufen – jeden Lösungsweg einzeln einzuschätzen.

Unterrichtsphasen:
Aufgaben können in unterschiedlichen Unterrichtsphasen eingesetzt werden und haben demzufolge unterschiedliche Intentionen. Fischer und Draxler unterscheiden drei Möglichkeiten:

1. Erarbeitungsphase
2. Übungsphase
3. Leistungsmessungsphase

Inzwischen haben Fischer und Draxler (2006) das Modell modifiziert und vor allem die Kategorie „Lesekompetenz“ neu eingeführt. Für unsere Untersuchung ist diese Erweiterung nicht relevant, da Kriterien der Textverständlichkeitsforschung bei uns einen eigenen Analyseschwerpunkt bilden.

2.2 Zur Textverständlichkeitsforschung

Aus der kognitiven Psychologie sind einige Ansätze zur Beschreibung des Verstehens von geschriebenen Texten bekannt. Gut evaluiert sind die Theorien von Kintsch und van Dijk, die ein zyklisches Modell des Textverstehens entwarfen (Kintsch & van Dijk, 1978) und das „Hamburger Verständlichkeitskonzept“ von Langer, Schulz von Thun und Tausch, die das Textverständnis aufgrund von Textmerkmalen untersuchten (Langer, Schulz von Thun & Tausch, 1974).
Kintsch und van Dijk (1978) entwickelten ein Modell des Textverstehens, das sowohl das Verstehen als auch das Erinnern von geschriebenem Text beschreibt. Im Gegensatz zum rein deskriptiven „Hamburger Verständlichkeitskonzept“ beruhen ihre Ergebnisse auf einer breit gefächerten theoretischen Basis. Danach wird Text verstanden, indem aktuell aufgenommene Propositionen des Textes mit früheren Propositionen verknüpft werden. Dabei ist eine Proposition „die kleinste Wissenseinheit, die eine selbstständige [...] Aussage bilden kann“ (Anderson, 1996, 141). Am einfachsten zu verstehen ist ein Text, wenn die Verknüpfung der Propositionen ohne Überbrückungsschluss (Inferenz) – Folgerungen aus bereits bekanntem Wissensinhalt – möglich ist. Der Leser versucht stets, Propositionen sinnvoll aneinander zu knüpfen und zu aktiv im Gedächtnis befindlichen Propositionsstrukturen hinzuzufügen, wobei die Anzahl der aktiven Propositionen limitiert ist. Der Schlüssel zu optimalem Textverständnis ist demnach also ein Text, bei dem sich die gelieferten Informationen möglichst lückenlos aufeinander beziehen, bzw. der Überbrückungsschlüsse direkt aus vorhandenem Wissen zulässt.
Im Gegensatz dazu formuliert das „Hamburger Verständlichkeitskonzept“ lediglich einen Katalog von Textmerkmalen, die in empirischen Untersuchungen zu einem besseren Textverständnis führten (Wellenreuther, 2005, 184). Der Leser wird dabei nicht berücksichtigt; der Text wird – im Gegensatz zum Modell von Kintsch und van Dijk – als eigenständig angenähert. Das Modell ist auch deshalb auch häufig kritisiert worden (z.B. Groeben, 1982; Hochhaus, 2004), seine Stärken liegen darin, dass es durch seine einfachen Annahmen sehr pragmatisch ist (Groeben, 1982).
Empirische Untersuchungen zeigen anhand der Bearbeitung von Lehrbuchtexten, dass Optimierungen im Sinne beider Theorien zu einem besseren Textverständnis führen und dass dieser Effekt – in gewissen Grenzen – umso stärker ausgeprägt ist, je geringer die Vorkenntnis des Lesers ist (Wellenreuther, 2005, 212).
Der Schluss, dass auch Testergebnisse verfälscht werden, wenn sie auf ein gesteigertes Textverständnis angewiesen sind und ein umfangreiches Textmaterial mitliefern, ist nahe liegend. Je inkonsistenter der Text geschrieben wurde, desto mehr ist die eigentlich getestete Kompetenz das Leseverständnis: „Ein konsistenter Vorspann erleichtert die Bildung einer Textbasis und damit das Beantworten von Testfragen, die von einer guten Textbasis abhängen.“ (Wellenreuther, 2005, 196) Die Qualität der Texte war tatsächlich auch bei PISA einer der Kritikpunkte, der immer wieder betont wurde (z.B. Schmidt, 2004).
Aus den Experimenten von Britton und Gülgöz (1991) auf der Basis der Theorie von Kintsch und van Dijk sowie Schulz von Thun, Göbel und Tausch (Wellenreuther, 2005, 186) – Vertretern des Hamburger Verständlichkeitskonzepts – lassen sich einige einfache Regeln zur Optimierung von Lehrbuchtexten ableiten:

Gliederungs-Ordnung:
möglichst hohe optische Übersichtlichkeit und geordnete, vollständige, Textinhalte; dazu sinnhafte Absätze.

Kürze/ Prägnanz:
Balance zwischen sprachlicher Ausschmückung und Lehrziel (nicht zu kurz, nicht zu viele Redundanzen, nicht zu hypotaktisch bzw. zu parataktisch).

Zusätzliche Stimulanz:
Integration anregender Textgestaltungselemente wie wörtlicher Rede und Bildern, aber auch lebensnaher Beispiele.

Kohärenz:
Kohärente Satzfolge durch Einbau von Verbindungselementen (Partikeln, Konjunktionen, substantivische Anknüpfung an den vorherigen Satz (Rekurrenz)). Innerhalb eines Satzes soll zunächst das Bekannte (Thema), danach das Neue (Rhema) genannt werden (Thema-Rhema-Gliederung). Textlinguistisch wird die Thema-Rhema-Gliederung des Öfteren zur so genannten Textkohäsion gerechnet, die von der Textkohärenz abgegrenzt ist. Diese Unterscheidung ist jedoch umstritten und nicht immer eindeutig, sodass in dieser Arbeit ein allgemeiner und umfassender Begriff der Textkohärenz verwendet wird (Glück, 2000, 352).

Zwar sind einige der Punkte – wie die Abfolge von Thema und Rhema – im Deutschen bereits in der Mehrzahl der Fälle natürlich, dennoch ist eine strukturierte und bewusste Befolgung anzuraten – insbesondere weil den schwächeren Schülerinnen und Schülern hier ein Vorteil zugute kommt und dem „Matthäus-Effekt" entgegengewirkt werden kann.

Beispielanalyse
„Ich gehe morgen in die Stadt. Auf dem Weg zur Stadt liegt ein Schnellrestaurant."

Gliederungs-Ordnung:
Hier nicht aussagekräftig, dieses Kriterium gewinnt Bedeutung bei längeren Texten.
Kürze/ Prägnanz:
Die Folge der beiden Sätze ist parataktisch. Dies ist für die Verständlichkeit nicht optimal, kommt aber erst bei mehreren Sätzen zum Tragen.
Zusätzliche Stimulanz:
Durch wörtliche Rede indirekt vorhanden, da der Text aber komplett aus wörtlicher Rede besteht, ist dieses Kriterium hier nicht aussagekräftig.
Kohärenz:
Der Text ist besonders wegen der Rekurrenz des Substantivs „Stadt" sehr kohärent. Dies ist der Verständlichkeit förderlich. Auch die Thema-Rhema-Gliederung wird beachtet, da im zweiten Satz „Stadt" als bekannter Anknüpfungspunkt vor „Schnellrestaurant" genannt wird.
Fazit:
Die Sätze sind sehr kohärent, die anderen Verständlichkeitskriterien kommen erst bei längeren Texten wirklich zum Tragen. Die Sätze sind demnach nach den Textverständlichkeitskriterien als leicht verständlich einzustufen.

Auch in der Fachdidaktik der Physik wird das Potenzial zur Kenntnis genommen und die Auswirkung der Textgestaltung auf den Wissenserwerb untersucht. Vorläufige Erkenntnisse lassen – ganz im Einklang mit Kintsch und van Dijk – die Textkohärenz als besonders bedeutend für das Verstehen physikalischer Texte erscheinen (Rabe & Mikelskis, 2004, 297). Starauschek (2006) verwendet den Begriff der Textkohäsion für textuelle Verständlichkeitskriterien, die die mentale Repräsentation des Lesers vernachlässigen. Dies ist als weitgehend gleichwertig zu dem hier verwendeten Kohärenzbegriff zu verstehen. Er zeigt, dass dieses Kriterium als bedeutend für das Verständnis von Schulbuchtexten der Physik anzusehen ist, bzw. dahingehend optimierte Texte von Schülern als verständlich eingeschätzt werden. Es bildet sich jedoch keine eindeutige Präferenz für diese optimierten Texte aus.

2.3 Das Bremen-Oldenburger Kompetenzmodell (BOlKo)

2.3.1 Konzeption des Modells

An den Universitäten Bremen und Oldenburg wird an einem Modell zur Beschreibung der Struktur naturwissenschaftlicher Kompetenz

gearbeitet (Schecker & Parchmann, 2006). Anders als beim rein normativen Modell der nationalen Bildungsstandards werden empirische Befunde über Schülerkompetenzen herangezogen, um das Modell in Richtung eines deskriptiven Modells weiterzuentwickeln. Eine umfangreiche empirische Studie ist dazu gerade abgeschlossen worden (Einhaus, 2007).

Das Modell umfasst fünf Dimensionen der Kompetenzmodellierung: „Inhaltsbereiche /Basiskonzepte", „Prozess/Handlung", „Kontext", „Ausprägung" und „kognitive Anforderungen". Die Dimension „Prozess/Handlung" entspricht dabei im Wesentlichen der Dimension „Kompetenzbereiche" der Bildungsstandards. Ihre Komponenten heißen „Fachwissen nutzen", „Erkenntnisse gewinnen", „Kommunizieren" und „Bewerten".

In der Dimension „Ausprägung" wurde den Anforderungsbereichen der Bildungsstandards eine weitere Komponente hinzugefügt, sodass die vier Stufen „lebensweltlich", „nominell/ reproduktiv", „aktiv anwenden" und „konzeptuell vertieft" resultieren. Mit der Dimension „Kontext" wird unter anderem Untersuchungen zum Conceptual Change (z.B. Caravita & Hallden, 1994) Rechnung getragen. Daher stammt die Erkenntnis, dass der Kontext Schülererklärungen beeinflusst; darüber hinaus konnten Interessenstudien zeigen, dass durch den Kontext auch die affektive Komponenten von Kompetenz verändert wird (z.B. Hoffmann, Häußler & Lehrke, 1998). Im Modell wird eine Unterteilung in innerunterrichtliche Kontexte, persönlich relevante Kontexte und professionelle Anforderungssituationen vorgeschlagen.

In Anlehnung an die Ausdifferenzierung des Facettendesigns von PISA (Rost et al., 2005, 197-199) wird die Dimension „kognitive Anforderungen" hinzugefügt. Von den von Rost et al. (2005) genannten sieben Kompetenzen werden jedoch nur die vier Kategorien „konvergentes Denken", „divergentes Denken", „Umgang mit mentalen Modellen" und „Umgang mit Zahlen" berücksichtigt, da die Kategorien „Bewerten", „Sachverhalte verbalisieren" und „Umgang mit Graphen" bereits kongruent mit einigen Komponenten der anderen Dimensionen des Bremen-Oldenburger Kompetenzmodells sind.

Für empirische Untersuchungen müssen diese Dimensionen reduziert werden, da nicht alle Zellen der fünfdimensionalen Matrix mit Items ausreichender Anzahl bestückt werden können. So kann beispielsweise, ein Inhaltsbereich konstant gehalten und Kontexte sowie kognitive Anforderungen als Co-Variaten erfasst werden (Theyßen et al., 2006, 2). Hierbei müssten nur noch die Dimensionen „Handlung/Prozess" und „Ausprägungen" variiert werden. Aus der fünfdimensionalen Matrix wird so für die Einzelstudie eine zweidimensionale, die der der Bildungsstandards ähnlich ist und ebenso zur Charakterisierung von Aufgaben genutzt werden kann.

Die Charakterisierung von Aufgaben durch Einstufung in das Kompetenzmodell ist in der Praxis jedoch recht aufwändig. Für eine reliable Einstufung sollten Items durch mehrere Experten den Zellen der Matrix zugeordnet werden. Das erweist sich als schwierig: „Untersuchungen zur Einschätzung von Items des TIMSS III-Tests zur voruniversitäten Physik [...] haben gezeigt, dass die Übereinstimmung verschiedener Rater bei der Einschätzung von Aufgaben anhand von Merkmalen selbst dann gering ist, wenn die Merkmale kleinschrittig aufgeschlüsselt werden." (Theyßen et al., 2006, 3)

Aus diesem Grunde stellen Theyßen et al. (2006) ein Verfahren vor, das auf Basis eines indikatorbasierenden Einordnungsschemas für das Bremen-Oldenburger Kompetenzmodell in bisherigen Erprobungen zu guten Ergebnissen geführt hat führt (Theyßen et al., 2006, 131: η-Koeffizienten zwischen 0,49 und 1,00, ab η =0,4 wird bei diesem Koeffizienten von Übereinstimmung gesprochen).

Die direkte Einordnung wird dabei durch ein sukzessives Vorgehen ersetzt. Dabei wird eine Aufgabenantwort zunächst anhand von „Prozess-Indikatoren" grundlegend einem Prozess, z.B. „Fachwissen nutzen" oder „Bewerten", zugeordnet. Die feinere Aufschlüs-

selung nach den Ausprägungen innerhalb eines Prozesses und damit die Zuordnung zu einer Zelle der Kompetenzmatrix (oder auch mehreren Matrixzellen) erfolgt anhand weiterer Indikatoren. Für die 16 Zellen (4 Prozesse, 4 Ausprägungen) liegen über 40 verschiedene Zellindikatoren vor, deren Formulierung aus der Analyse der Bildungsstandards, der Einheitlichen Prüfungsanforderungen für die Abiturprüfung Physik und zahlreicher Musteraufgaben resultiert. Ein großer Vorteil des BOlKo liegt also in seiner feinen Operationalisierung sowie seinem erprobten und validierten System zur Aufgabeneinstufung.

2.4 Aufgabenkontexte

Ein zentraler Aspekt der „neuen Aufgabenkultur" ist die Einbettung von Aufgaben in Kontexte (z.B. Leisen, 2005, 307). „Bei den meisten Aufgaben lässt sich eine Tiefenstruktur von einer Oberflächenstruktur unterscheiden. Die Tiefenstruktur bezieht sich auf das zugrunde liegende Prinzip, durch dessen sinngemäße Anwendung eine Lösung herbeigeführt werden kann. Die Oberflächenstruktur umfasst die konkreten in der Aufgabe beschriebenen Objekte" (Häußler & Lind, 1998, 21).

Der Energieerhaltungssatz kann beispielsweise im Zusammenhang mit einem Fadenpendel oder einem die schiefe Ebene herunter rollenden Fass behandelt werden. Diese beiden Aufgaben hätten dieselbe Tiefenstruktur – den Energieerhaltungssatz – aber unterschiedliche Oberflächenstrukturen. Wenn von „Kontexten" die Rede ist, so meint dies hier die Oberflächenstruktur. Es hat sich gezeigt, dass insbesondere Mädchen von Kontexten profitieren, die ihre Interessengebiete berühren (Labudde, 1999, 6) – ohne dass Jungen benachteiligt werden.

Nach Häußler und Lind (1998) ergeben sich folgende interessestiftende Kontexte:

- Kontexte, die sich auf alltägliche Erfahrungen oder die Umwelt beziehen. Dies ist jedoch für Mädchen nur dann förderlich, wenn sie bereits Erfahrungen mit diesen Sachverhalten haben – technische Bezüge sind dazu meist kontraproduktiv.
- Kontexte, die emotional positiv gefärbt sind (z. B. Phänomene, die zum Staunen anregen, Naturphänomene).
- Kontexte, die die gesellschaftliche Bedeutung von Naturwissenschaft in den Vordergrund stellen.
- Kontexte, die den menschlichen Körper behandeln (z. B. medizinische Anwendungen oder die Funktion der Sinnesorgane).
- Kontexte, die einen Anwendungsbezug aufzeigen. Der Sinn einer Anwendung muss dabei erkennbar sein.

3 Kategorien zur Beschreibung von PISA-Aufgaben und methodisches Vorgehen

3.1 Verwendete Kategorien

Aus den beschriebenen theoretischen Grundlagen können nunmehr Kategorien gewonnen werden, die eine sinnvolle Charakterisierung von PISA-Aufgaben sowie deren Vergleich unter einander ermöglichen. Dabei bildet das Modell von Fischer und Draxler (2002) wegen seines Detailreichtums und seines hohen Grades an Bekanntheit die Grundlage. Es kann aus diesem Modell jedoch auf alle Kategorien verzichtet werden, die direkt auf Unterricht und praktischen Einsatz Bezug nehmen, da sie für die Beschreibung einer autarken Aufgabe irrelevant sind. In der Folge werden die verwendeten Kategorien mit ihren jeweiligen Kriterien vorgestellt.

A) Aufgabenkultur

Der Ursprung dieser Kategorie ist das Kriterium „Interesse" aus der Kategorie „inhaltliche und curriculare Einordnung" des Modells von Fischer und Draxler (2002). Dort wird der Aufgabenkontext als bedeutsam für die Entwicklung von Motivation der Schülerinnen und Schüler eingestuft und dessen Alltagsnähe untersucht. In dem hier verwendeten Modell wird in stärkerem Maße Bezug auf die Motivations- und Interessenforschung genommen. Die alleinige Beschränkung des Kontextes auf eine möglichst hohe Alltagsnähe ist nicht ausreichend, da unter anderem Aspekte der unterschiedlich gelagerten

Interessen von Mädchen und Jungen nur unzureichend Berücksichtigung finden.
Das Interesse wird in die Kategorie „Aufgabenkultur“ integriert und stark ausdifferenziert. Im Einklang mit den beschriebenen Erkenntnissen über interessefördernde Kontexte resultiert daraus das Kriterium *Bezug*. Die Unterpunkte sind wie in Kap. 2.4 aufgeführt „Alltag“, „Natur“, „Mensch“, „Gesellschaft“ und „Anwendung“. Es werden also besonders gendersensible Aufgabenbezüge eingeschätzt. Dies erscheint bei einem Vergleich von PISA-Aufgaben als bedeutsam, weil dem Kontext eine hohe Bedeutung für die Gestaltung von Aufgaben eingeräumt wird (Häußler & Lind, 1998). Ebenso erscheint dies als bei PISA-Aufgaben beobachtbar, da bei diesen die Skizzierung des Kontextes oftmals umfangreich ist.

B) Textbarriere

Für Aufgaben, die sehr textlastig sind, ist es lohnenswert, zur Beschreibung textbezogene Kriterien hinzuzuziehen. Textgestaltung wird bei Fischer und Draxler (2002) höchstens peripher mit dem Anforderungsmerkmal „Textverständnis“ berücksichtigt. Dies ist jedoch in den Bereich der Lesekompetenz zu zählen und trifft keinerlei Aussagen über die Qualität des Textes, sondern darüber, ob die Aufgabenlösung die Informationsentnahme aus einer textuellen Quelle benötigt oder nicht. Aussagekräftiger ist es, Qualitätskriterien aus den kognitionspsychologischen Ansätzen zur Verarbeitung geschriebenen Textes einzubeziehen. Eine Bewertung der Qualität des Aufgabentextes gelingt nur auf diese Weise.
Für in großem Umfang textbezogene Aufgaben könnten – wie bei Fischer und Draxler (2006) – ebenfalls die Stufungen der Lesekompetenzen, wie sie bei PISA ermittelt wurden, mit berücksichtigt werden. Dies ist sinnvoll, wenn eine Aufgabe ihre zentrale Schwierigkeit daraus zieht, dass sie Textverständnis benötigt. Bei naturwissenschaftlichen Aufgaben soll dies jedoch in der Regel nicht der Fall sein. Auch bei PISA wird getrennt zwischen der Erhebung von Textverständnis und der Erhebung naturwissenschaftlicher Kompetenz, denn im Idealfall soll keine Aufgabe mehrere Kompetenzen auf einmal messen. Ansonsten ist es schwierig, festzustellen, welcher Kompetenz Erfolg oder Misserfolg zuzuschreiben ist, was die Validität des Testes in Frage stellt. Wenn bei der Einstufung in ein Kategoriensystem zur Beurteilung von Aufgaben also die Textverständlichkeit als Kriterium verwendet wird und nicht die Lesekompetenz, so kommt dies einer Umorientierung gleich: Nicht die Leser haben maximal (lese-)kompetent zu sein, sondern die Aufgaben maximal verständlich. Gerade bei PISA-Aufgaben erscheint die Untersuchung der Aufgabentexte wegen des oft umfangreichen Kontextmaterials bedeutsam.

C) Kompetenzzuordnung

Das Bremen-Oldenburger Kompetenzmodell ist von entscheidender Bedeutung für das Beschreibungsmodell von Aufgaben, besonders zur Ermittlung der in die Aufgaben einzubringenden Kompetenzen.
Zur Erfassung dieser Kompetenzen müssen die Aufgaben schließlich in ein entsprechendes Modell eingestuft werden. Das von Fischer und Draxler (2002) dazu verwendete eindimensionale Kompetenzmodell nach TIMSS hat sich z.B. bei PISA 2003 als nicht tragfähig erwiesen (Rost, 2004, 665). Verwendet wird daher in dieser Arbeit das BOlKo mit seinen Dimensionen „Prozess“ und „Ausprägung“ (Abb. 1). Theyßen et.al. (2006) haben gezeigt, dass Testaufgaben verlässlich in das Modell eingestuft werden können.

D) Aufgabenformat

Die Kategorie „Aufgabenformat“ ist aus der Kategorie „Antwortformat, Offenheit und Experimentierverhalten“ des Modells von Fischer und Draxler (2002) entstanden. Es erscheint an dieser Stelle sinnvoll, zunächst eine dichotome Unterscheidung von geschlossenen und offenen Aufgabenfor-

		Prozess			
		Fachwissen nutzen	Erkenntnisse gewinnen	Kommunizieren	Bewerten
Ausprägung	lebensweltlich				
	nominell / reproduktiv				
	aktiv anwenden				
	konzeptuell vertieft				

Abb. 1: Prozess-Ausprägung-Matrix des Bremen-Oldenburger Kompetenzmodells

maten zu treffen. Dies ist ertragreich, weil geschlossene Aufgabenformate generell andere kognitive Fähigkeiten erfordern als offene (Kircher et al., 2001, 306). Auch bei PISA ist diese Unterscheidung vorgenommen worden. Die Anschlussfähigkeit an PISA wird durch diese generelle Trennung also erleichtert. Bei geschlossenen Aufgaben wird hier zwischen Multiple Choice-Aufgaben, bei denen nur eine Antwortmöglichkeit ausgewählt werden muss, und Multiple Select-Aufgaben, bei denen mehrere Behauptungen aus einer gegebenen Anzahl von Behauptungen richtig sein können, unterschieden. Bei den offenen Aufgaben wird zwischen Kurzsatz- und Lang- bzw. Aufsatz-Aufgaben differenziert.

Bei der Einschätzung der PISA-Aufgaben wurden die Kriterien „Offenheit“ und „Experimentierverhalten“ aus Fischer und Draxler (2002) nicht berücksichtigt. In einem Vergleich der Testdurchläufe wären sie nicht von Wert, da PISA-Testaufgaben möglichst nur einen Lösungsweg aufweisen sollten und bei PISA nicht mit Experimenten gearbeitet wurde.

E) Inhaltsrepräsentation

Das Bremen-Oldenburger Kompetenzmodell beschreibt im Prozess „Kommunikation“ – anders als die Bildungsstandards – nur die aktive Form der Kommunikation. Das Erschließen von Informationen hingegen wird im Modell als Zusatzkodierung erhoben. Bei Aufgaben, die viele Sachinformationen zur Aufgabenlösung in sich tragen und in verschiedener Form kodiert haben, ist Informationserschließung jedoch von großer Bedeutung. Aus diesem Grunde wird in das Strukturmodell zur Beschreibung von Aufgaben die Kategorie „Inhaltsrepräsentation“ eingeführt. Zugrunde liegt der Kategorie die weit gefasste Definition von Texten, wie sie bei PISA verwendet wird. Hier wird zwischen kontinuierlichen und diskontinuierlichen Texten differenziert:

„Kontinuierliche Texte bestehen normalerweise aus Sätzen, die in Absätzen organisiert sind. […] Nicht-kontinuierliche Texte liegen häufig im Matrixformat vor und beruhen auf Kombinationen von Listen.“ (PISA-Konsortium Deutschland, 2000, 29).

Innerhalb dieser Kriterien wird eine Unterscheidung zwischen „fachlichen“ und „alltäglichen“ Texten vorgenommen. Ein Beispiel für einen alltäglichen, kontinuierlichen Text wäre ein Zeitungsartikel, eines für einen alltäglichen, diskontinuierlichen Text die Bundesligatabelle. Für einen fachlichen, diskontinuierlichen Text kann ein Energieflussdiagramm als Beispiel genannt werden und für einen fachlichen, kontinuierlichen Text ein Lehrbuchtext aus einem Physikbuch.

Des Weiteren kann angegeben werden, ob die Information zur Lösung der Aufgaben in den jeweiligen Texten bereits vorhanden ist („ablesen“) oder ob zusätzliche Information hinzugefügt werden muss („ergänzen“). Dies ist eine wichtige Unterscheidung, da viele (Test-)Aufgaben darauf abzielen, bereits erworbenes Wissen zu aktivieren, hier

wird also nur ein Teil der Information in der Inhaltsrepräsentation bereits geliefert. Es findet somit eine Vernetzung von Aufgabentext und Lerninhalten statt. Andere Aufgaben hingegen liefern alle Informationen mit sich; diese müssen nur geschickt gefunden oder zusammengesetzt werden – dies gilt für PISA-Aufgaben sogar als besonderes Merkmal (Petri & Einhaus, 2006). Der Unterschied ist also in etwa so wie zwischen einem Kreuzworträtsel, das zusätzliche Information aus der Allgemeinbildung verlangt, und einem Sudoku, bei dem bereits aus der Anfangsposition die Lösung determiniert ist und die vorhandene Information nur geschickt entschlüsselt werden muss. Eine Einstufung in diese Kategorie soll jedoch nur erfolgen, wenn tatsächlich wesentliche Informationen zur Aufgabenlösung aus einem Text entnommen werden müssen. Die bloße Formulierung der Aufgabenstellung reicht hierfür nicht aus.

3.2 Methodik

Zur Beschreibung der Aufgaben müssen Einstufungen in die in Abschnitt 3.1 beschriebenen Kategorien und Kriterien vorgenommen werden. Dies ist bei den unterschiedlich festgelegten Kriterien jedoch nicht einheitlich möglich. Eine Differenzierung kann vorgenommen werden, indem Kriterien, die einander ausschließen *(absolute Kriterien)* von solchen unterschieden werden, die graduell gestuft sind *(gestufte Kriterien)*. Bei ersteren trifft immer eines der Kriterien einer Kategorie zu. Hier können Häufigkeiten ausgezählt werden, um eine Aussage über die Gesamtheit der Aufgaben zu machen.

Die gestuften Kriterien hingegen bieten die Möglichkeit, zumindest als heuristisches Mittel der Illustration von Unterschieden Durchschnittswerte oder Mediane zu errechnen. Für die Gesamtheit der Aufgaben ergibt sich dadurch der Grad, inwieweit ein Kriterium zutrifft. Die gewählten Skalen gehen dabei stets von 0 bis 1. In den Abbildungen 2 und 3 sind die verwendeten absoluten bzw. abgestuften Kriterien aus dem Abschnitt 3.1 dargestellt worden.

Kategorie	Kriterium	Kodierung
Bezug (A)	Anwendung	0;1
	Gesellschaft	0;1
	Mensch	0;1
	Natur	0;1
	Alltag	0;1
Textbarriere (B)	Kohärenz	0; 0,5; 1
	zus. Stimulanz	0; 0,5; 1
	Satzbau	0; 0,5; 1
	Gliederung	0; 0,5; 1

Abb. 2: Gestufte Kriterien, nach denen die Aufgaben beschrieben wurden. Hier ist eine Durchschnittsbildung möglich, dabei entspricht in der Kategorie Bezug 0 „nicht vorhanden" und 1 „vorhanden", während in der Kategorie *Textverständlichkeit* von 0 bis 1 je nach Grad des Übereinstimmens mit theoretischen Kriterien gestuft wird. In der Spalte „Kategorie" wird auf die Unterteilungen des Abschnitts 3.1 verwiesen.

Kategorie	absolute Kriterien
Kompetenzzuordnung(C)	• Zellen der Matrix (Abbildung 1)
Aufgabenformat (D)	• Multiple Choice • Multiple Select • Kurzsatz • Lang-/Aufsatz
Inhaltsrepräsentation - kontinuierlich (E)	• fachlich - ergänzen • alltäglich - ergänzen • fachlich - ablesen • alltäglich - ablesen
Inhaltsrepräsentation - diskontinuierlich (E)	• fachlich - ergänzen • alltäglich - ergänzen • fachlich - ablesen • alltäglich - ablesen

Abb. 3: Absolute Kriterien, nach denen die Aufgaben beschrieben werden. Die Kriterien einer Zelle schließen einander aus, sodass hier Häufigkeiten ausgezählt werden können. In der Spalte „Kategorie" wird auf die Unterteilungen des Abschnitts 3.1 verwiesen.

Die Einstufung in die Prozess-Ausprägung-Matrix des Bremen-Oldenburger Kompetenzmodells geschieht mithilfe des vorgestellten indikatorbasierenden Einordnungsschemas durch die Autoren. An dieser Stelle sind geringe Fehlerquellen zu erwarten, da das Einstufungsschema bei mehreren Ratern zu einer sehr großen Übereinstimmung führt (Theyßen et al., 2006, 131: η-Koeffizienten zwischen 0,49 und 1,00, ab η =0,4 wird bei diesem Koeffizienten von Übereinstimmung gesprochen). Hier wurde also ein bereits validiertes Verfahren verwendet.

Auch die Einstufungen in den absoluten Kategorien, die lediglich formale Kriterien aufweisen (Rahmenbedingungen – Intention, Rahmenbedingungen – Einbindung, Aufgabenkultur – Kooperation, Aufgabenformat) ist durch die Autoren vorgenommen worden. Die beiden absoluten Kriterien der Inhaltsrepräsentation sowie die gestuften Kriterien wurden von den Autoren anhand eines Systems von Indikatoren eingeschätzt, das möglichst einfache Entscheidungen ermöglicht (Kulgemeyer, 2007, IX-XII), indem die Kriterien auf formale, ablesbare Indikatoren reduziert werden. Ein Beispiel dafür ist hier aus der Kategorie Textbarriere angeführt (Abb. 4).

Bei der Inhaltsrepräsentation erfolgt eine Einstufung nur bei Aufgaben, die Informationen entweder aus dem Text heraus mit zu memorierenden Inhalten verknüpfen („ergänzen“) oder eine Verknüpfung von im Text gegebenen Informationen („ablesen“) verlangen. Zu letzterer Kategorie zählen auch reine Textverständnisaufgaben. Dazu ist noch eine Unterteilung zwischen fachlichen und alltäglichen Texten sowie diskontinuierlichen bzw. kontinuierlichen Texten getroffen worden, für die obiges zutrifft. Anzumerken ist hier, dass auch eine doppelte Einstufung getroffen werden konnte, wenn sowohl aus diskontinuierlichen als auch aus kontinuierlichen Texten Informationen entnommen werden. Wenn die Aufgabe keine wesentliche Information benötigt, die aus dem Text entnommen werden muss, wurde gar keine Einstufung vorgenommen. Selbstverständlich benötigt jede Art von Aufgabe, die in einem Papier-und-Bleistift-Test auftritt, irgendeine Form von Inhaltsrepräsentation in textueller Form.

Ähnlich wie bei der Einstufung im Bremen-Oldenburger Kompetenzmodell wurde jedoch der Fokus der Aufgabe als Anhaltspunkt genommen. Wenn die Aufgabe also ihren Fokus nicht darauf legt, dass textuelle Information zur Lösung benötigt wird, kann keine Einstufung erfolgen. Dies führt dazu, dass die Summe der Anteile nicht zwingend 100 % sein muss.

Zur Sicherung der Objektivität des Verfahrens wurden Quereinstufungen von mehreren Ratern vorgenommen. Dazu wurde etwa ein fünftel (N=15) der gerateten Aufgaben von drei Experten in jeweils ein Kriterium, das eine dichotome Kodierung erfordert und eines, das eine dreistufige Kodierung verlangt, eingestuft. Zu diesem Zwecke wurden das Kriterium „Mensch“

Kriteritum	Unterpunkt 1	Unterpunkt 2	Unterpunkt 3
Gliederung	Gut (1) Mehrere Absätze (Optik), Absätze thematisch getrennt	Mittel (0,5) Mind. 1 Absatz thematisch zentriert, ges. mind. 2 Absätze	Schlecht (0) Keine Absätze (Optik), Absätze nicht thematisch fest

Abb. 4: Indikatoren für das Kriterium „Gliederung“ der Kategorie Textbarriere.

aus der Kategorie „Aufgabenkultur – Bezug“ sowie das Kriterium „Gliederung“ aus der Kategorie „Textbarriere“ ausgewählt. Der Grund für die Auswahl gerader dieser Kriterien zur Quereinstufung ist ihr besonderer Charakter. Sie repräsentieren jeweils einen der beiden Typen von Kriterien, die in dieser Studie verwendet werden: „Aufgabenkultur-Bezug“ ist absolut und dichotom, „Gliederung“ gestuft. Zudem stammen sie aus den beiden Kategorien der Studie, deren Einstufung hoch inferent ist und in denen die Auswahl noch nicht durch ein validiertes Verfahren gestützt wird. Bei beiden Kriterien wurden gute bzw. sehr gute Übereinstimmungen erreicht („Mensch“ – Fleiss' Kappa (κ)=1,00; „Gliederung“ – κ =0,78). Als Maß für die Interraterreliabilität wurde hier Fleiss' Verallgemeinerung von Cohens Kappa auf mehrere Rater verwendet (Fleiss, 1971). Die Experten waren alle in Textarbeit erfahrene Didaktiker bzw. Lehrer. Das verwendete Ratingverfahren kann nach diesen Ergebnissen also als ausreichend objektiv angesehen werden.

Zum Vergleich werden einerseits die Aufgaben aus PISA 2006 (Cresswell & Vaysettes, 2006) und andererseits die aus PISA 2000 und 2003 (PISA-Konsortium Deutschland, 2000 und PISA-Konsortium Deutschland, 2003) gruppiert. Dies ist sinnvoll, da bei PISA 2006 der Fokus auf Scientific Literacy lag. Es ist zu prüfen, ob damit in diesem Bereich Veränderungen vorgenommen wurden, die sich auf die Form der Aufgaben auswirken. Zum Vergleich wurden sowohl die Verteilungen im Einzelnen dargestellt und interpretiert als auch Chi-Quadrat-Tests durchgeführt.

4 Ergebnisse

4.1 Vergleich der Aufgaben verschiedener PISA-Durchgänge

Wir werden im Folgenden zeigen, dass sich beim Vergleich der Testdurchläufe 2000 und 2003 einerseits sowie 2006 andererseits einige interessante Veränderungen zeigen lassen. Dabei werden wir oft die Mittelwerte als heuristisches Mittel des Vergleichs nutzen, jedoch in der Folge auch die Ergebnisse eines Chi-Quadrat-Tests zur Feststellung der Signifikanz von Veränderungen nennen. Bei der Analyse können wir nur auf veröffentlichte Aufgaben zurückgreifen, sodass unsere Stichprobe limitiert ist (2000/2003: N=16, 2006: N=56). Dies ist im Vergleich zu den verwendeten Aufgaben jedoch ein nicht unmaßgebliche Anzahl (z.B. 2000: N=46 (Prenzel, Rost, Senkbeil, Häußler & Klopp, 2001, 210); 2006: N=108 (PISA-Konsortium, 2007, 336)). Zwar sind nicht alle veröffentlichten Aufgaben auch verwendet worden, nach Selbstauskunft des PISA-Konsortiums seien sie dennoch repräsentativ für die Gesamtheit der Aufgaben – alle Ergebnisse sind also unter diesem Vorbehalt zu verstehen.

Bei der Analyse des Aspektes „Bezug“ sind erste Differenzen zu konstatieren (Kap. 3.1, Kategorie A, Abb. 5 u. 6). Gleich geblieben ist zwar die Häufigkeitsrangfolge der Bezüge („Natur“, vor „Alltag“ und „Mensch“), allerdings ist die Gewichtung verändert. 2000/2003 ist ein Naturbezug mit einem Anteil von 0,75 der bei weitem dominierende. Dieser Wert ist 2006 auf 0,63 zurückgegangen, gleichzeitig sind die Bedeutungen von „Alltag“ (0,50) und „Mensch“ (0,57) als Bezüge stark gestiegen.

2000/2003 sind beide mit jeweils 0,25 noch von weit geringerer Bedeutung. Der Bezug „Gesellschaft“ spielt 2006 fast keine Rolle mehr (2000/2003: 0,19). Es kann also konstatiert werden, dass sich die Prioritäten bei den verschiedenen Kontexten verändert haben und mehr in Richtung einer Einbindung in die „Natur“ gehen. Die Veränderungen in den Kriterien „Gesellschaft“ und Mensch“ sind nach Chi-Quadrat-Test

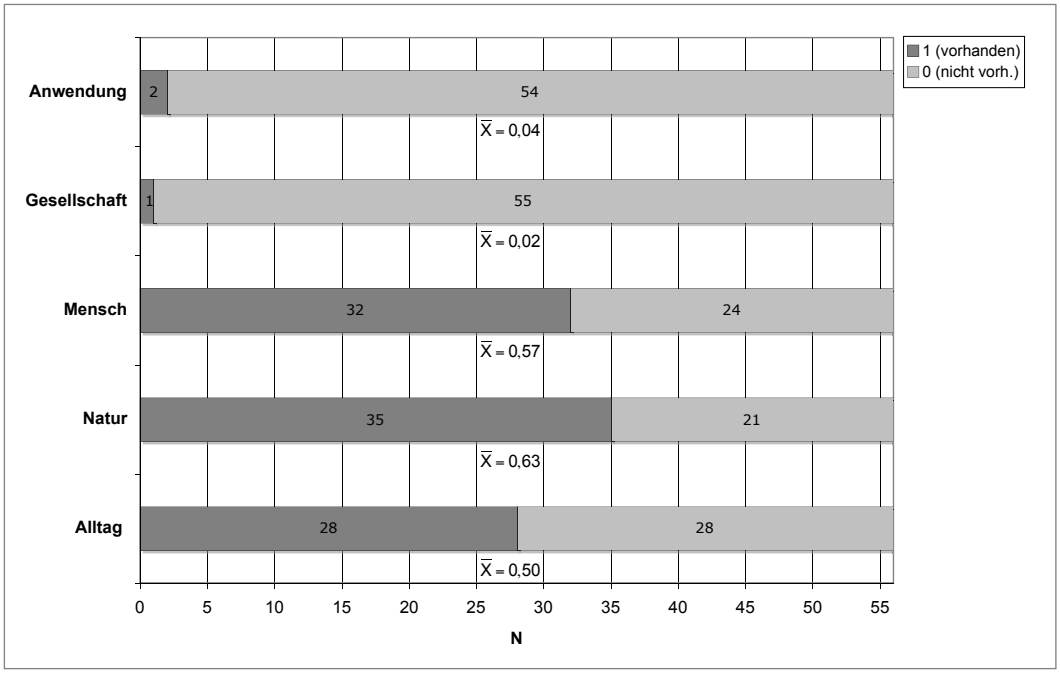

Abb. 5: Verteilung der Kontextbezüge der Aufgaben aus PISA 2006 (Cresswell & Vaysettes, 2006). Unter den Balken sind jeweils die Durchschnittswerte der Kategorien als heuristisches Mittel des Vergleichs angegeben.

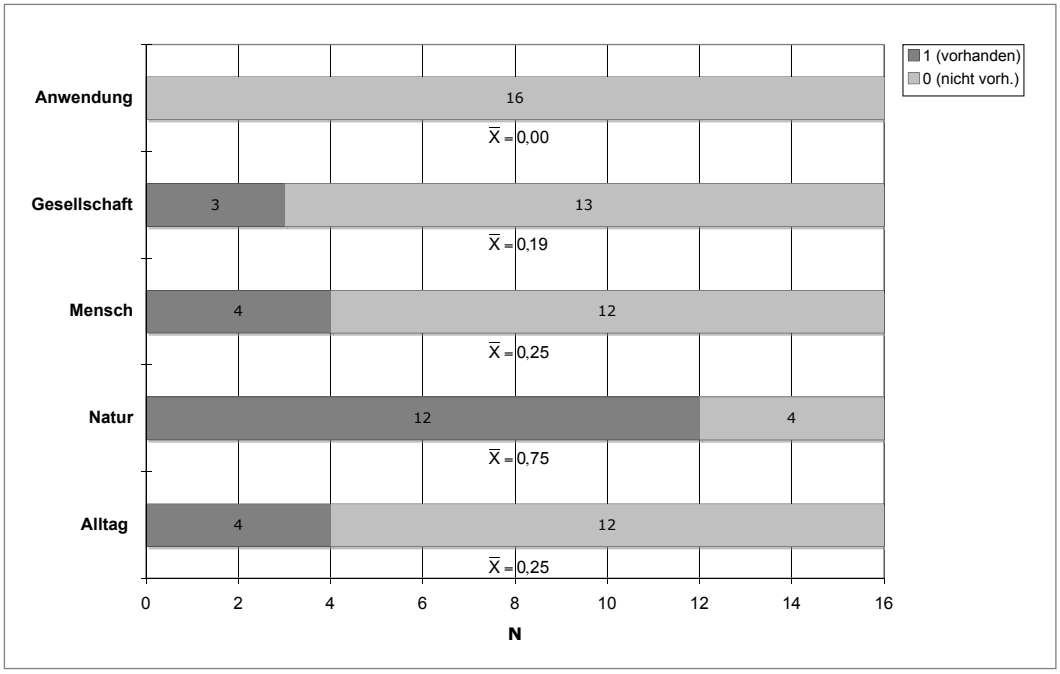

Abb. 6: Verteilung der Kontextbezüge der Aufgaben aus PISA 2000 und PISA 2003 (PISA-Konsortium Deutschland, 2000 und PISA-Konsortium Deutschland, 2003). Unter den Balken sind jeweils die Durchschnittswerte der Kategorien als heuristisches Mittel des Vergleichs angegeben.

signifikant (Abb. 16). Es folgt auch, dass die PISA-Aufgaben 2006 in ihrer kontextuellen Einbindung breiter gestreut sind als die von 2000 bzw. 2003, denn der einseitige Bezug auf die Natur ist durch drei vorrangig verwendete Bezüge ersetzt worden.

Noch ertragreicher ist der Vergleich der „Textbarriere“ (Kap. 3.1, Kategorie B, Abb. 7 u. 8), die in den Aufgaben der verschiedenen Jahrgänge erstellt wurde. Die Aufgaben erreichen 2006 hier fast durchweg bessere Werte, d. h. eine bessere Verständlichkeit.

Eine Ausnahme stellt nur die Einbindung zusätzlicher Stimulanz dar. Hier wurde 2000/2003 mit 0,69 ein höherer Wert erreicht (2006: 0,61). Innerhalb der textuellen Formulierung waren die Aufgaben 2006 jedoch offensichtlich optimiert. Sowohl die „Gliederung“ als auch der „Satzbau“ und die „Kohärenz“ sind im Durchschnitt deutlich besser, d.h. verständlicher, als 2000/2003 – die Veränderungen sind nach Chi-Quadrat-Test durchweg signifikant (Abb. 16). Es ist also möglich, dass das Modell der Textverarbeitung von Kintsch und van Dijk bei der Konzeption der Aufgaben eine bedeutende Rolle spielte oder die Aufgaben nachträglich dahingehend optimiert wurden – auch wenn Textverständlichkeit nicht explizit als Optimierungskriterium genannt wird (Prenzel, Carstensen, Frey, Drechsel & Rönnebeck, 2007); gerade die hohe Kohärenz eines Textes ist aber ein zentraler Gedanke dieses Modells als Voraussetzung für verständliche Texte.

Auf jeden Fall ist es von theoretischer Warte aus gesehen 2006 so, dass die Probanden die textuelle Hürde leichter überwinden können und somit nicht bereits beim Textverständnis scheitern. Es kann vermutet werden, dass der Naturwissenschaftsteil von PISA 2006 geringer durch Leseverständnis konfundiert ist als 2000 bzw. 2003.

Passend zu der These einer Verminderung der textuellen Barriere ist auch der Vergleich der durchschnittlichen Anzahl an Worten, die

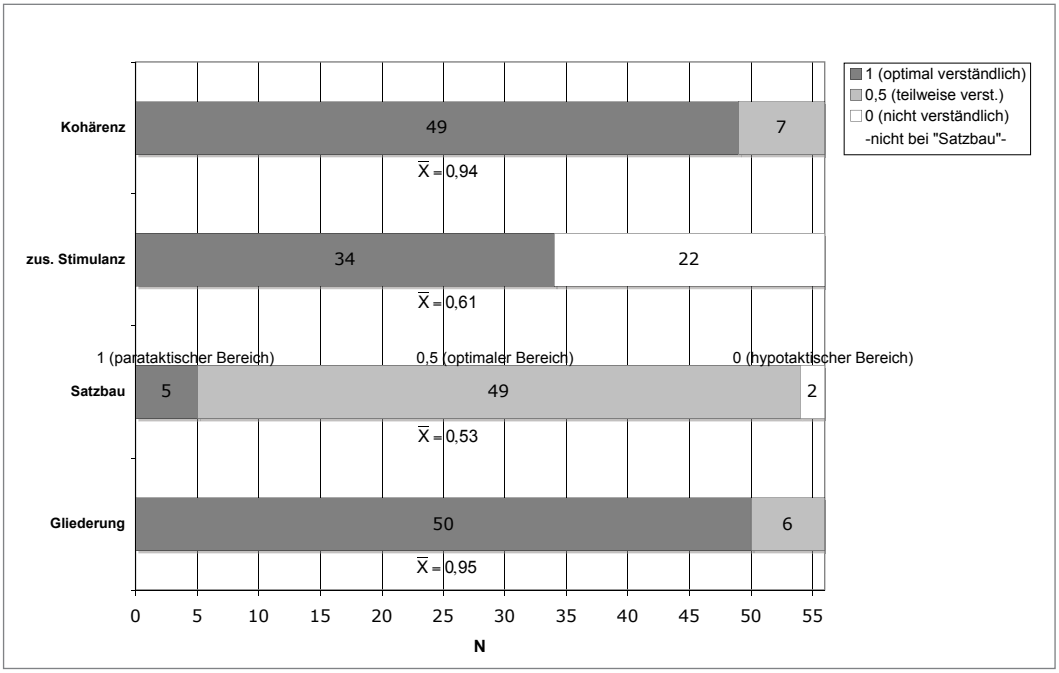

Abb. 7: Verteilung der Textverständlichkeitskriterien der Aufgaben aus PISA 2006 (Cresswell & Vaysettes, 2006). Unter den Balken sind jeweils die Durchschnittswerte der Kategorien als heuristisches Mittel des Vergleichs angegeben.

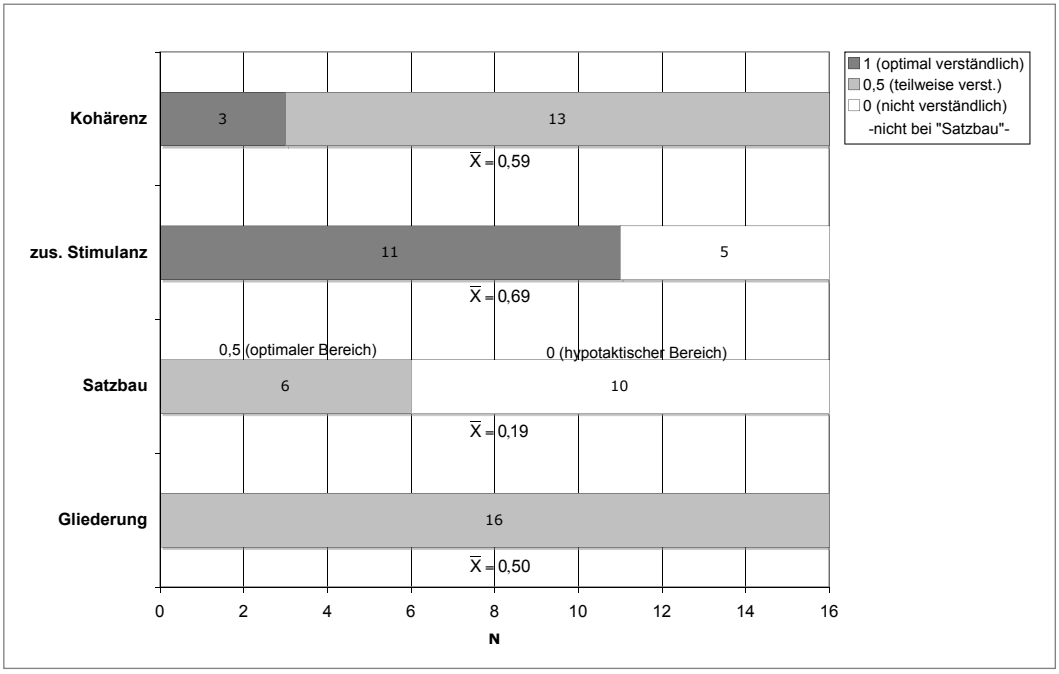

Abb. 8: Verteilung der Textverständlichkeitskriterien aus PISA 2000 und PISA 2003 (PISA-Konsortium Deutschland, 2000 und PISA-Konsortium Deutschland, 2003). Unter den Balken sind jeweils die Durchschnittswerte der Kategorien als heuristisches Mittel des Vergleichs angegeben.

Durchlauf	Mittelw. Worte / Kontext	
2000	268	
2003	157	
2000 + 2003	212	≙ 100 %
2006	96	≙ 45 %

Abb. 9: Vergleich des Wortumfangs der Kontextbeschreibungen

pro kontextuelle Darstellung verwendet werden (Abb. 9).

2006 wurden bei den veröffentlichten Aufgaben im Durchschnitt 55 % weniger Worte benutzt, der Umfang des Kontextes also etwa halbiert. Nimmt man hinzu, dass auch weniger zusätzliches Stimulanzmaterial Teil des Tests ist, dann lässt sich auf eine insgesamt geringere Bedeutung des Kontextmaterials schließen.

Darüber hinaus wird auch weniger Raum benötigt, um den Kontext zu umreißen. Folgerungen hieraus lassen sich schwer ziehen, zumal aus vorangegangenen Testdurchläufen geschlossen wurde, dass die Länge des Kontextes die Aufgabenschwierigkeit nicht nennenswert erhöht (Prenzel et al., 2002, 132) – vermutlich waren testpragmatische Gründe dafür verantwortlich.

Auch der Vergleich im Bremen-Oldenburger Kompetenzmodell (Kap. 3.1, Kategorie C) zeigt leichte Akzentverschiebungen bei den Aufgaben (Abb. 10+11). Von 2000/2003

auf 2006 hat sich die Bedeutung des Prozesses „Fachwissen nutzen“ erhöht. Der Anteil der Aufgaben mit Hauptschwierigkeit in diesem Bereich steigt von 44 % auf 61 %. Die Summe der Anteile der einzelnen Kompetenzbereiche muss hier nicht 100 % betragen, da einige Aufgaben reine Textverständnisaufgaben waren und somit nach Interpretation des Bremen-Oldenburg Kompetenzmodells keine naturwissenschaftliche Kompetenz erforderten. PISA 2006 ist also leicht stärker auf die Nutzung von Fachwissen konzentriert und in seiner Bandbreite der getesteten Kompetenzen nicht so vielseitig wie seine Vorgänger, die Unterschiede sind jedoch nicht signifikant (Abb. 16). Die veröffentlichten PISA-Aufgaben sind allgemein stärker als angenommen auf Fachwissen fokussiert.

Mit Einschränkungen könnte an dieser Stelle behauptet werden, dass PISA 2006 in der Tendenz ein wenig „klassischer“ in seinen Aufgaben geworden ist. Dies zeigt auch die Betrachtung der verwendeten Aufgabenformate (Abb. 12 u. 13).

2006 wurden bei den veröffentlichten Aufgaben mehr geschlossene Formate verwendet (71 %) als 2000/2003 (50 %). Vor allem der Anteil der Langsatz- oder Aufsatzaufgaben sinkt von 19 % auf 10 %.

Die verwendeten Aufgabenformate wurden tendenziell also einseitiger. Es kann festgehalten werden, dass PISA 2006 einen höheren Akzent auf leicht zu korrigierende Aufgabenformate legt, möglicherweise aus Gründen der dadurch steigenden Auswerteobjektivität. Auch hier sind die Unterschiede nach Chi-Quadrat-Test jedoch als nicht signifikant zu beurteilen.

Im Vergleich der „Inhaltsrepräsentation“ von Aufgaben aus PISA 2006 und PISA 2000/2003 zeigen sich ebenfalls interessante Veränderungen (Abb. 14 u. 15). Bei den ersten beiden PISA-Durchläufen waren mehr Aufgaben den beiden Stufen von „ablesen“ zuzuordnen (N=6) als denen von „ergänzen“ (N=5). Dies ist bei PISA 2006 in der Tendenz anders, jedoch nicht signifikant. Bemerkenswert ist, dass Prenzel et al. (2002) bereits festgestellt haben, dass für PISA 2000 die nationalen Ergänzungsaufgaben einen anderen Charakter haben als die internationalen Haupttestaufgaben – sie benötigen nämlich weniger textuelle Information (Prenzel et al., 2002, 128). PISA 2006 tendiert hier also in eine Richtung, die die nationalen Ergänzungsaufgaben Deutschlands gewiesen haben. Außerdem schätzen Prenzel et al. (2002) den Anteil der Aufgaben mit lösungsrelevanter Information aus dem Text mit ca. 65 % in einer ähnlichen Größenordnung der Ergebnisse hier ein („ergänzen“ + „ablesen“ = 69 %) (Prenzel et al., 2002, 128).

Zum Vergleich der Verteilungen in den Kategorien bei PISA 2000 und 2003 einerseits sowie PISA 2006 andererseits wurden Chi-Quadrat-Tests durchgeführt, um die Unabhängigkeit der Verteilungen zu überprüfen. Dabei ergeben sich die Daten aus Abb. 16. Es zeigt sich, dass mindestens signifikante Veränderungen in den Kategorien „Bezug“ (Kriterien „Gesellschaft“ und „Mensch“) und „Textbarriere“ (Kriterien „Kohärenz“, „Satzbau“ und „Gliederung“) gefunden werden können. Der Chi-Quadrat-Test unterstützt somit die anschauliche Interpretation über die Mittelwerte.

Wegen der geringen verfügbaren Itemzahlen mussten bei diesen Tests die BOlKo-Kategorien „Fachwissen nutzen“ und „Erkenntnisse gewinnen“ sowie „Kommunizieren“ und „Bewerten“ zusammengefasst werden – dies ist jedoch auch inhaltlich sinnvoll und möglich.

		Prozess			
		Fachwissen nutzen	Erkenntnisse gewinnen	Kommuni-zieren	Bewerten
Ausprägung	lebensweltlich	8 (14 %)	-	-	-
	nominell / reproduktiv	21 (38 %)	1 (2 %)	1 (2 %)	8 (14 %)
	aktiv anwenden	3 (5 %)	11 (20 %)	-	-
	konzeptuell vertieft	2 (4 %)	-	-	1 (2%)
	Summe:	34 (61 %)	12 (21 %)	1 (2 %)	9 (16 %)

Abb.10: Einordnung der Aufgaben aus PISA 2006 in die Prozess-Ausprägung-Matrix des BOlKo.

		Prozess			
		Fachwissen nutzen	Erkenntnisse gewinnen	Kommuni-zieren	Bewerten
Ausprägung	lebensweltlich	1 (6 %)	-	-	-
	nominell / reproduktiv	5 (31 %)	-	2 (13 %)	2 (13 %)
	aktiv anwenden	1 (6 %)	3 (19 %)	-	-
	konzeptuell vertieft	-	-	-	1 (6 %)
	Summe:	7 (44 %)	3 (19 %)	2 (13 %)	3 (19 %)

Abb. 11: Einordnung der Aufgaben aus PISA 2000/2003 in die Prozess-Ausprägung-Matrix des BOlKo (eine Aufgabe konnte nicht eingestuft werden).

Aufgabentyp	Anzahl	Anteil
Multiple Choice	23	41%
Multiple Select	17	29%
Kurzsatz	11	20%
Lang-/Aufsatz	5	10%
∑ geschl.:	**40**	**71%**
∑ offen:	**16**	**29%**

Aufgabentyp	Anzahl	Anteil
Multiple Choice	6	38%
Multiple Select	2	13%
Kurzsatz	5	31%
Lang-/Aufsatz	3	19%
∑ geschl.:	**8**	**50%**
∑ offen:	**8**	**50%**

Abb. 12 (links): Unterscheidung der Aufgaben aus PISA 2006 nach ihrem Format.
Abb. 13 (rechts): Unterscheidung der Aufgaben aus PISA 2000/2003 nach ihrem Format.

	Fachlich / ergänzen	**Alltäglich / ergänzen**	**Fachlich / ablesen**	**Alltäglich / ablesen**
kontinuierlich	1 (2 %)	18 (31 %)	-	6 (11 %)
diskontinuierlich	1 (2 %)	7 (13 %)	8 (14 %)	3 (5 %)
	$\sum = 27\ (48\ \%)$		$\sum = 17\ (30\ \%)$	

Abb. 14: Unterscheidung der Aufgaben aus PISA 2006 nach den Bereichen der Inhaltsrepräsentation.

	Fachlich / ergänzen	Alltäglich / ergänzen	Fachlich / ablesen	Alltäglich / ablesen
kontinuierlich	-	4 (25 %)	-	4 (25 %)
diskontinuierlich	1 (6 %)	-	2 (13 %)	-
	$\sum = 5$ (31 %)		$\sum = 6$ (38 %)	

Abb. 15: Unterscheidung der Aufgaben aus PISA 2000/2003 nach den Bereichen der Inhaltsrepräsentation.

Kategorie	Kriterium	df	Wert
Bezug	Anwendung	1	0,588
	Gesellschaft	1	6,826**
	Mensch	1	5,143*
	Natur	1	0,858
	Alltag	1	3,150
Textbarriere	Kohärenz	1	29,319**
	zus. Stimulanz	1	0,343
	Satzbau	1	31,144**
	Gliederung	1	46,753**
Kompetenzzuordnung	BOlKo - Prozess	1	0,804
Aufgabenformat	offen/ geschlossen	1	2,571
Inhaltsrepräsentation	Information ablesen/ ergänzen	1	0,915

Abb. 16: Ergebnisse des Chi-Quadrat-Tests nach Pearson zum Vergleich der Verteilungen der einzelnen Kategorien bei PISA 2000/2003 einerseits und PISA 2006 andererseits (df: Anzahl der Freiheitsgrade). Schwellenwerte für die Signifikanz: signifikant (*) p<0,05, hoch signifikant (**) p<0,01, jeweils bezogen auf die Wahrscheinlichkeit der Identität der Verteilungen.

5 Zusammenfassung

Betont werden muss vorab, dass die hier vorgelegten Analysen auf Basis der veröffentlichten PISA-Aufgaben erfolgt sind. Es muss daher unterstellt werden, dass die zugänglichen Units von ihrer Anlage her den in der Durchführung verwendeten Aufgaben entsprechen. Alle Ergebnisse sind unter dieser Annahme zu betrachten.

Über die Testläufe PISA 2000 bis 2006 identisch geblieben ist die grobe Konzeption der Aufgaben als Units mit einer organisierten Serie von Items und unterstützendem Kontextmaterial. In der Gestaltung der Aufgaben zeigen sich jedoch zum Teil interessante Veränderungen.

Ein zentraler Unterschied zwischen den PISA Durchläufen 2000/2003 und 2006 betrifft die Gestaltung des Kontextmaterials. Die Kontextbeschreibungen wurden weit weniger umfangreich (55 % weniger Worte, d.h. der Umfang wurde etwa halbiert) und beinhalteten weniger Stimulationsmaterial. Dafür wurden textuelle Kohärenz, Gliederung und Satzbau nahezu optimiert, was dazu führt, dass die (kontinuierlichen) Texte 2006 signifikant (nach Chi-Quadrat-Test) verständlicher sind als 2000/2003. Die Gestaltung lehnt sich der Anschauung nach stark an das Textverständnismodell von Kintsch und van Dijk an. Gerade die starke Verbesserung der Textkohärenz – als zentrale Folgerung dieses

Modells – weist in diese Richtung. Daraus folgt, dass der Naturwissenschaftsteil von PISA 2006 geringer durch Leseverständnis konfundiert sein könnte als seine Vorgänger 2000/ 2003. Damit dringen tendenziell mehr Probanden zu den naturwissenschaftlichen Fragestellungen vor und scheitern nicht bereits an der textuellen Hürde. Die Aufgaben sollten im Vergleich mit den Aufgaben aus den Durchgängen 2000 und 2003 also stärker die naturwissenschaftliche Kompetenz messen als das Textverständnis. Die Reduzierung in Umfang und Verständnisanforderung führt insgesamt dazu, dass die Bedeutung des Kontextmaterials zurückgeht.

Die veröffentlichten Aufgaben aus PISA 2006 sind vornehmlich in Bezüge zur Natur eingebunden. Ebenfalls häufig aufzufinden sind Bezüge zum Alltag und zum menschlichen Körper. Dies ist 2000/2003 nicht grundsätzlich anders – hier ist die Dominanz des Naturbezugs zwar größer, der Unterschied ist jedoch nicht signifikant. Nach Chi-Quadrat-Test signifikante Veränderungen lassen sich in der Kategorie „Bezüge“ jedoch im Anteil der Bezugskriterien „Mensch“ und „Gesellschaft“ finden.

Nach den Kriterien des Bremen-Oldenburger Kompetenzmodells zeigt sich, dass die Aufgaben aus PISA 2006 den Akzent ebenso wie die Aufgaben aus PISA 2000/2003 auf den Prozess „Fachwissen nutzen“ legen. Der Anteil dieser Aufgaben stieg von 44 % (2000/2003) auf 61 % (2006). Im Vergleich sank 2006 besonders der Anteil der Aufgaben, die dem Prozess „Kommunizieren“ zugeordnet werden können. Der Unterschied in den durch die Aufgaben angesprochenen Prozesse im Bremen-Oldenburger Kompetenzmodell ist allerdings nicht signifikant.

Die Analyse der Form der Inhaltsrepräsentation zeigt nur eine leichte Verschiebung von 2000/2003 zu 2006. 2000/2003 waren von den Aufgaben, die Informationen des Textes zur Lösung heranzogen, tendenziell die meisten so gestaltet, dass sie alle Informationen zur Lösung bereits mit dem Text mitlieferten. Bei den Aufgaben aus PISA 2006 ist dies geringfügig anders: Die meisten Aufgaben benötigen hier eine Verknüpfung von memorierten und textuell gelieferten Informationen, um zur Lösung zu gelangen. Vergleicht man die Veränderungen zwischen PISA 2000/2003 und PISA 2006 in diesem Kriterium mittels eines Chi-Quadrat-Tests, so ist die Veränderung jedoch nicht signifikant. Es zeigt sich, dass sowohl die Aufgaben aus PISA 2006 als auch die aus PISA 2000/2003 nicht alle Informationen zur Lösung mitliefern – ein oft für ein Merkmal von PISA-Aufgaben gehaltenes Kriterium (Petri & Einhaus, 2006) ist nicht haltbar.

Bei den Aufgaben aus PISA 2006 stieg der Anteil an geschlossenen Aufgabenformaten leicht an. Bei den Aufgaben aus den Durchläufen 2000/2003 sind 50 % geschlossenen Formats, während es bei den veröffentlichten Aufgaben aus dem Durchlauf 2006 72 % sind. Diese Veränderungen sind jedoch nicht signifikant.

Nimmt man alle diese Ergebnisse zusammen, so lässt sich festhalten, dass die Naturwissenschaftsaufgaben bei PISA von 2000/2003 bis 2006 Veränderungen unterworfen haben. Die Units und Items von PISA 2006 wurden in ihrer kontextuellen Formulierung verständlicher und eindeutiger angelegt. Darüber hinaus wurden die kontextuellen Bezüge „Gesellschaft“ und „Mensch“ häufiger verwendet. Die Veränderungen in der Kategorie „Textbarriere“ (Kriterien „Kohärenz“, „Satzbau“ und „Gliederung“) sowie in den Kriterien „Gesellschaft“ und „Mensch“ der Kategorie „Bezug“ sind statistisch signifikant und könnten somit dazu beitragen, dass deutsche Schülerinnen und Schüler bei PISA 2006 besser abschneiden als in den Jahren 2000 und 2003. Durch die Veränderungen in der Gestaltung von Umfang und Bezug des Kontextmaterials wird den Aufgaben auf jeden Fall ein anderer Charakter verliehen.

Zum Abschluss sollen die Eigenschaften der Aufgaben aus PISA 2006 zusammenfassend dargestellt werden, auch um die Konstruktion von Aufgaben zu ermöglichen, die deren Charakteristika nachbilden („PISA-ähnliche Aufgaben“, Abb. 17). Hier sind fünf Eigenschaftsfelder aufgeführt, die auf Basis der

Aufgabenanalysen erstellt wurden: Allgemeines, Kontextgestaltung, Kontextbezug, Itemgestaltung und Kompetenzzuordnung. Innerhalb der Felder werden durch die farbig abgehobenen Felder Aufzählungspunkte repräsentiert, die wichtige Eigenschaften genauer fassen. Sofern die Aufzählungspunkte in einem bestimmten Verhältnis zueinander stehen, sind sie in ein Diagramm einsortiert. Zur Gestaltung PISA-ähnlicher Aufgaben sollte darauf geachtet werden, dass die Verhältnisse für eine große Anzahl von Aufgaben zutreffen, bei der Einzelaufgabe jedoch als Richtwert gesehen werden.

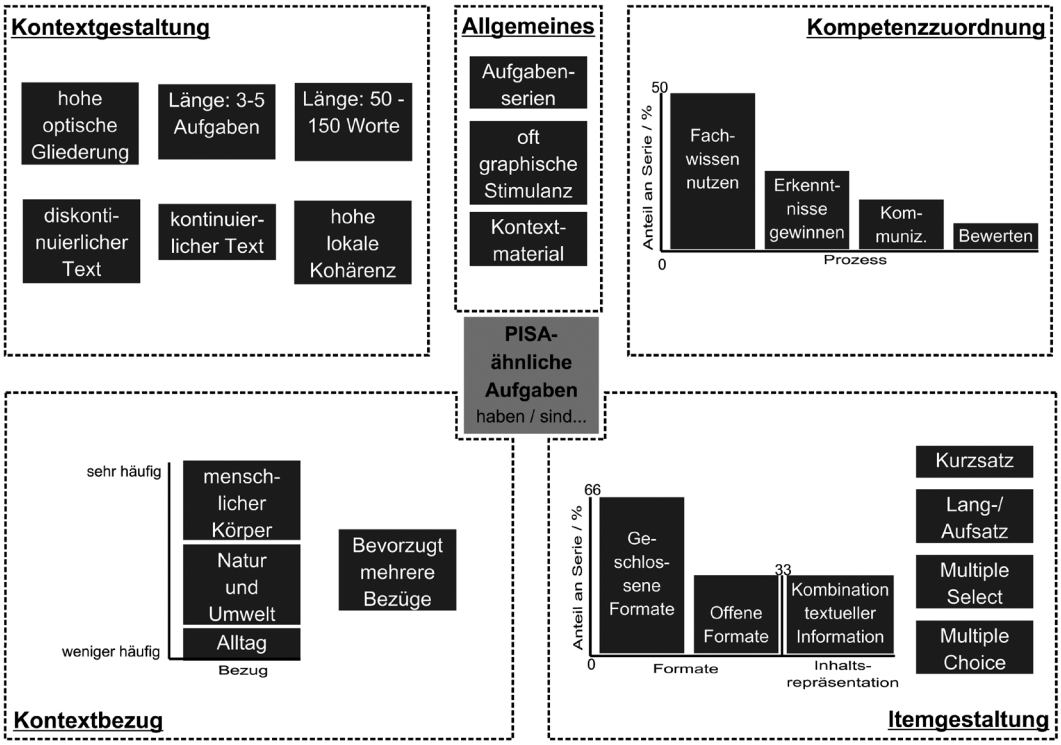

Abb. 17: Fünf Eigenschaftsfelder zur Zusammenfassung der Charakteristika von PISA 2006-Aufgaben und als Empfehlung zur Gestaltung PISA-ähnlicher Aufgaben.

6 Literatur

Anderson, J. (1996). *Kognitive Psychologie*. Heidelberg: Spektrum.

Britton, B. & Gülgöz, S. (1991). Using Kintsch's Computational Model to Improve Instructional Text: Effects of Repairing Inference Calls on Recall and Cognitive Structures. *Journal of Educational Psychology, 83*, 329-345.

Caravita, S. & Hallden, O. (1994). Re-framing the problem of conceptual change. *Learning and Instruction, 4*(1), 89-111.

Cresswell, J. & Vaysettes, S. (2006). *Assessing Scientific, Reading and Mathematical Literacy. A Framework for PISA 2006*. Paris : OECD.

Duit, R. (2006). Initiativen zur Verbesserung des Physikunterrichts in Deutschland. *Physik und Didaktik in Schule und Hochschule, 2*, 83–96.

Einhaus, E., Kulgemeyer, C., Marks R. & Petri, J. (2005). *Wege zu einer neuen Aufgabenkultur. Beispiele aus dem Bereich der naturwissenschaftlichen Grundbildung- Band 2*. Bremen: Der Senator für Bildung und Wissenschaft.

Einhaus, E. (2007). *Schülerkompetenzen im Bereich Wärmelehre*. Berlin: Logos.

Fischer, H. & Draxler, D. (2002). Konstruktion und Bewertung von Physikaufgaben. In E. Kircher & W. Schneider (Eds.), *Physikdidaktik in der Praxis*, (pp. 300 – 322). Berlin: Springer.

Fischer, H. & Draxler, D. (2006). Konstruktion und Bewertung von Physikaufgaben. In E. Kircher, R. Girwidz & P. Häußler (Eds.), *Physikdidaktik. Theorie und Praxis*, (pp. 639-655). Berlin: Springer.

Fleiss, J. (1971). Measuring Nominal Scale Agreement among many Raters. *Psychological Bulletin, 76,* 378-382.

Glück, H. (2000). *Metzler Lexikon Sprache*. Stuttgart: Metzler.

Groeben, N. (1982). *Leserpsychologie: Textverständnis – Textverständlichkeit*. Münster: Aschendorff.

Häußler, P. & Lind, G. (1998). *BLK-Programmförderung „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts". Erläuterungen zu Modul 1 mit Beispielen für den Physikunterricht. Weiterentwicklung der Aufgabenkultur im mathematisch-naturwissenschaftlichen Unterricht*. Forschungsbericht, IPN Kiel.

Häußler, P. & Lind, G. (2000). „Aufgabenkultur" – Was ist das? *Praxis der Naturwissenschaften – Physik, 49*(4), 2-10.

Hochhaus, S. (2004). *Der verständliche Text. Perspektiven auf die Textoptimierung*. Schriftliche Hausarbeit für die Magisterprüfung. Ruhr-Universität Bochum.

Hoffmann, L., Häußler, P., & Lehrke, M. (1998). *Die IPN-Interessenstudie Physik*. Kiel: IPN.

Kintsch, W. & van Dijk, T. A. (1978). Towards a Model of Text Comprehension and Production. *Psychological Review, 85*, 363-394.

Kircher, E., Girwidz, R, & ; Häußler, P. (2001). *Physikdidaktik. Eine Einführung*. Berlin: Springer.

Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: theoretische Grundlagen, Kompetenzstufen und Unterrichtschwerpunkte. In J. Baumert, W. Bos & R. Lehmann (Eds.), *TIMSS/III Dritte Internationale Mathematik und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn- Band 2* (pp. 57-128). Opladen: Leske und Budrich.

Klieme, E et al. (2003). *Zur Entwicklung nationaler Bildungsstandards*. Bonn: BMBF.

Kulgemeyer, C. (2007). *Weiterentwicklung und weitere Entwicklung von PISA-ähnlichen Aufgaben*. Hausarbeit zum Ersten Staatsexamen, Universität Bremen.

Labudde, P. (1999): Mädchen und Jungen auf dem Weg zur Physik. Reflexive Koedukation im Physikunterricht. *Unterricht Physik 49*(10), 4-10.

Langer, I., Schulz von Thun, F., & Tausch, R. (1974). *Verständlichkeit in Schule, Verwaltung und Politik*. München: Ernst Reinhard.

Leisen, J. (2005). Zur Arbeit mit Bildungsstandards. Lernaufgaben als Einstieg und Schlüssel. *Der mathematische und naturwissenschaftliche Unterricht, 58*, 306-308.

Meyerhöfer, W. (2005). *Tests im Test: Das Beispiel PISA*. Opladen : Budrich.

Petri, J. & Einhaus, E. (2006). Aufgabenbeispiele zur naturwissenschaftlichen Grundbildung gemäß der PISA-Konzeption. *Der mathematische und naturwissenschaftliche Unterricht, 59*, 300.

PISA-Konsortium (2007): PISA 2006. Volume2: Data. http://www.oecd.org/data-oecd/30/17/39703267.pdf (Abgerufen: 17.12.2007)

PISA-Konsortium Deutschland (2000). *PISA 2000. Beispielaufgaben aus dem Naturwissenschaftstest*. OECD-Druck.

PISA-Konsortium Deutschland (2003). *PISA 2003. Beispielaufgaben aus dem Naturwissenschaftstest*. OECD-Druck.

Prenzel, M.,Carstensen, C., Frey, A., Drechsel, B., Rönnebeck, S. (2007). PISA 2006 – Eine Einführung in die Studie. In PISA-Konsortium Deutschland (Eds.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (pp. 31-59). Münster: Waxmann.

Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2000). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft. Zeitschrift für Lernforschung, 30*, 120-133.

Prenzel, M., Rost, J., Senkbeil, M. Häußler, P. & Klopp, A. (2001). Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse. In PISA-Konsortium Deutschland (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 192-248). Oplade: Leske und Budrich.

Rost, J. (2004). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik, 50*(5), 662-678.

Rost, J., Walter, O., Carstensen, C., Senkbeil, M. & Prenzel, M. (2005). Der nationale Naturwissenschaftstest PISA 2003. *Der mathematische und naturwissenschaftliche Unterricht, 58*, 196-204.

Rabe, T. & Mikelskis, H. (2004). Selbsterklärung und Textkohärenz beim Wissenserwerb zur Physik mit Multimedia. In A. Pitton (Ed.*), Relevanz fachdidaktischer Forschungsergebnisse für die Lehrerbildung. Jahrestagung der GDCP in Heidelberg* (pp. 396-398). Münster: Lit.

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten oder allgemeine Intelligenz? *Psychologische Rundschau, 57*(2), 69-86.

Schecker, H. Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenzen. In: *Zeitschrift für Didaktik der Naturwissenschaften, 12*, 45-66.

Schmidt, H. (2004). Analyse der veröffentlichten Chemie-Aufgaben von PISA. *Der mathematische und naturwissenschaftliche Unterricht, 57*, 180-183.

Starauschek, E. (2006). Der Einfluss von Textkohäsion und gegenständlichen externen piktoralen Repräsentationen auf die Verständlichkeit von Texten zum Physiklernen. *Zeitschrift für Didaktik der Naturwissenschaften, 12*, 128-157.

Theyßen, H., Schmidt, M., Einhaus, E. & Schecker, H. (2006). Ein indikatorenbasiertes Verfahren zur Einstufung von Testaufgaben in ein Kompetenzmodell. *Physik und Didaktik in Schule und Hochschule*, 2(5), 123-134.

Wellenreuther, M. (2005). *Lehren und Lernen – aber wie? Empirisch-experimentelle Forschung zum Lehren und Lernen im Unterricht.* Hohengehren : Schneider.

Kontakt
Christoph Kulgemeyer
Universität Bremen, Institut für Didaktik der Naturwissenschaften, Abt. Physikdidaktik,
Fachbereich 1 Physik/Elektrotechnik
Postfach 330440
D-28334 Bremen
kulgemeyer@physik.uni-bremen.de

Autoreninformation
Christoph Kulgemeyer ist wissenschaftlicher Mitarbeiter und Doktorand in der Arbeitsgruppe von Horst Schecker. Er hat in Bremen Physik und Germanistik für das Lehramt an Gymnasien studiert.
Prof. Dr. Horst Schecker arbeitet zurzeit an der Modellierung physikalischer Kompetenzstrukturen bei Schülern sowie daran orientierten Lernaufgaben und Leistungstests. Ein weiterer aktueller Arbeitsbereich ist die Reform der Physiklehrerausbildung. Langfristige Forschungsgebiete sind Schülervorstellungen und Lernprozesse im Physikunterricht sowie Studien zur Lernwirksamkeit virtueller Lehr- und Lernmedien. Horst Schecker ist Sprecher des Vorstands der Gesellschaft für Didaktik der Chemie und Physik (GDCP).