

Norbert Knoche, Detlef Lind,  
Werner Blum, Elmar Cohors-Fresenborg, Lothar Flade, Wolfgang Löding, Gerd Möller,  
Michael Neubrand, Alexander Wynands  
(Deutsche PISA-Expertengruppe Mathematik, PISA-2000)

## Die PISA-2000-Studie, einige Ergebnisse und Analysen

**Zusammenfassung:** Die Pisa-2000-Studie setzt sich in der Komponente „Mathematik“ aus zwei Tests zusammen, dem „Internationalen“ Test und dem „Nationalen Ergänzungstest“.

Der folgende Beitrag stellt die Konzeptionen beider Tests und Analysen der Ergebnisse vor. Dabei wird in die Betrachtungen auch eine Darstellung der messtheoretischen Verfahren, die in die Konzeptionen der Tests wie in die Analysen eingehen, so weit aufgenommen, dass der Leser die vorgestellten Analysen mit Blick auf beide Komponenten – die Konzeption und das Analyseverfahren – selbst nachvollziehen kann.

Bei den Analysen konzentriert sich die Arbeit auf die Themenkomplexe: „Mathematik, Naturwissenschaften und Lesen im Vergleich“ – „Schwierigkeitsgenerierende Faktoren“ – „Mathematik und Geschlecht“.

**Abstract:** The "Mathematics" component of the Pisa-2000-study is based on two tests, the "International Test" and the "National Supplementary Test". The following paper explains the concepts of both tests and the analysis procedures of the results. It also contains information about the theoretical measurement procedures that are the basis for the concepts and analyses.

This allows the reader to retrace the analyses with respect to both the concepts and the analysis procedures. The analyses concentrate on the topics "Mathematics, Science and Reading in comparison", "Difficulty generating factors", and "Mathematics and Gender".

### Einleitung

Die PISA-Studie erscheint auf den ersten Blick so wie die TIMS-Studien als Leistungsstudie, wenn man nur auf die Tests zur Messung von Fähigkeiten abhebt.

Schon die Analysen der TIMS-Studie zeigten aber, dass es möglich ist, aus den erhobenen „Leistungsdaten“ Rückschlüsse auf einzelne die Leistung bedingende Faktoren zu ziehen und damit *fachdidaktische* Fragestellungen zu untersuchen, das um so mehr, wenn man in die Betrachtungen Ergebnisse aus Begleituntersuchungen einbezieht.

Das gleiche Ziel – Leistungsvergleiche im internationalen und nationalen Bereich und das Erkennen von Kompetenzmerkmalen der untersuchten Schülerpopulation – verfolgt auch die PISA-Studie. In der folgenden Darstellung konzentrieren wir uns primär auf den Leistungstest Mathematik. Andere Tests, wie die Tests zur Messung der Lesekompetenz, zur Messung der Naturwissenschaftlichen Grundbildung und die begleitenden Untersuchungen werden nur dort angesprochen, wo sie der Interpretation der Ergebnisse der mathematischen Leistungsstudie dienen (an weiteren Informationen interessierte Leser seien auf Baumert, J. et al. (2001a) und OECD (Hrsg.) (2001) verwiesen).

Im Unterschied zur zweiten TIMS-Studie (im Folgenden stets kurz TIMSS II genannt), in der die Konzeption des Tests unter weitgehender Beachtung von *Curriculumvalidität* der Aufgaben und der Berücksichtigung von Stoffgebieten erfolgte, verfolgt die internationale PISA-Studie in allen drei Domänen *Textverständnis*, *Mathematik* und *Naturwissenschaften* das Ziel, Leistung an einem normativ festgelegten Begriff von „Literacy“ zu messen, im Mathematik-Test also an „Mathematical Literacy“.

Erweitert und vertieft wird dieser Test durch einen „Nationalen Ergänzungstest“. Dieser Test ist in seiner Konzeption, Leistung an „Literacy“ zu messen, mit dem internationalen Test vergleichbar. Er erweitert aber den internationalen Literacy-Begriff auf den Begriff der „Mathematischen Grundbildung“. Zudem berücksichtigt er in seiner Aufgabenstruktur auch curriculare Aspekte des Unterrichts und orientiert Aufgaben und Aufgabensequenzen auch an fachdidaktischen Fragestellungen.

Eine Beurteilung der Analysen der Tests ist nur möglich, wenn man neben den konzeptionellen Vorstellungen und den zu untersuchenden Fragestellungen bei der Entwicklung der Tests mit den Modellvorstellungen vertraut ist, die messtheoretisch die geplanten Analysen der erhobenen Daten im Blick haben. Verständnisschwierigkeiten bei Diskussionen von „Ergebnissen“ beruhen oft auf fehlender Gesamtschau beider Komponenten.

Zentral bei psychometrischen Untersuchungen ist die Unterscheidung zwischen *manifesten* Variablen, die das beobachtbare Verhalten der Probanden beschreiben (z. B. Testscores) und *latenten* Variablen, die Persönlichkeitsparameter (z. B. Kompetenzen) beschreiben und auf die über die beobachtete Testleistung „zurückgeschlossen“ wird. Das ist nur mit Modellannahmen über die Verbindung solcher Variablen möglich.

In den folgenden Kapiteln werden beide Komponenten der Tests so dargestellt, dass der Leser die vorgestellten Analysen in ihrer Aussagekraft auch mit Blick auf die Methodik der Auswertung nachvollziehen kann. Wir verwenden dazu nur freigegebene Aufgaben. Das Copyright der Aufgaben des internationalen Tests liegt bei *PISA-2000 - International, OECD Paris*, das der nationalen Aufgaben bei *PISA-2000 – Deutschland, MPI-Bildungsforschung Berlin*.

## **I Zur Konzeption der PISA-Studie**

Die folgenden Ausführungen stellen die Grundprinzipien der Konzeption des internationalen Tests und des nationalen Ergänzungstests dar, soweit sie für die folgenden Betrachtungen benötigt werden. Umfassende Darstellungen findet der Leser in Klieme et al. (2001), Neubrand et al. (2001) und Neubrand (2001).

### **I.1 Zum Internationalen Mathematiktest**

Wie bereits in der Einleitung erwähnt, verfolgt die internationale PISA-Studie in allen drei Domänen Textverständnis, Mathematik und Naturwissenschaften das Ziel, Leistung an „Literacy“ zu messen. Mit Blick auf die Mathematik wird *Literacy* als die Fähigkeit gesehen, mathematisches Wissen „funktional“, d.h. ideenreich, mit Einsicht und flexibel bei der Bearbeitung kontextbezogener Aufgaben einsetzen und die Rolle von Mathematik in unserer Welt beurteilen zu können (vgl. OECD (1999)). Dabei wird *Literacy*

durch Fähigkeiten („Competencies“) charakterisiert, von denen *Modellieren* im Sinne des Übersetzens zwischen Realität und Mathematik (vgl. Blum (1996)) die wichtigste ist.

Was bedeutet das für die Konzeption der Aufgaben und für ihre Anforderungsmerkmale, d.h. die für ihre Lösung erwarteten Fähigkeiten?

Das Konzept des internationalen Tests ist geprägt von einer integrativen Sicht vom Bearbeiten mathematischer Aufgaben. Das bedeutet, dass man im internationalen Test sehr wenige Aufgaben zur Untersuchung einzelner Fertigkeiten oder Fähigkeiten findet, wie sie in der in Abschnitt I.2 skizzierten Konzeption des nationalen Ergänzungstests bewusst verlangt werden.

Die Gesamtstruktur des internationalen PISA-Tests wird durch sogenannte *Kompetenzklassen* beschrieben, denen die Aufgaben zugeordnet werden können. Diese insgesamt drei Klassen repräsentieren unterschiedliche kognitive Anforderungen, in die der Prozess des mathematischen Arbeitens zerlegt werden kann. Eine solche Kompetenzklasse enthält gemäß der angesprochenen integrativen Sicht vom Bearbeiten mathematischer Aufgaben dann immer ein ganzes Bündel einzelner mathematischer Kenntnisse, Vorstellungen, Fertigkeiten und Fähigkeiten, die für alle Bereiche und Ebenen der mathematischen Bildung relevant sind.

- Klasse 1** beinhaltet Aufgaben, zu deren Lösung Kenntnisse von Fakten und einfachen Berechnungen benötigt werden (kurz: „Reproduction“).
- Klasse 2** umfasst Aufgaben, zu deren Lösung auch Querverbindungen zwischen unterschiedlichen mathematischen Inhalten oder zwischen Mathematik und Realität herzustellen sind (kurz: „Connection“).
- Klasse 3** umfasst Aufgaben, deren Lösung einsichtsvolles mathematisches Denken und strukturelles Verallgemeinern erfordert (kurz: „Reflection“).

Die Kompetenzklassen charakterisieren zunächst die gestellten Aufgaben. Sie stellen a priori keine Hierarchie mit Blick auf die Schwierigkeit einer Aufgabe dar, sondern strukturieren das Spektrum mathematischer Leistungsanforderungen insgesamt.

Im internationalen Test sind die Aufgaben nicht den üblichen Stoffgebieten (Algebra, Geometrie, u.s.w.) sondern übergeordneten Leitideen zugeordnet, von denen in PISA-2000 bei der Aufgabenkonstruktion nur „Raum und Form“ und „Veränderung und Wachstum“ berücksichtigt werden konnten.

### **I.2 Zum nationalen Ergänzungstest**

In der Umsetzung des Konstrukts „Mathematical Literacy“ konzentriert sich die internationale Studie weitgehend auf Aufgaben, die in einen *außermathematischen* Kontext eingebunden sind. Der Diskussion des Begriffsinhalts von mathematischer Grundbildung in Deutschland entsprechend (vgl. Klieme et al. (2001), Neubrand et al. (2001), Winter (1995)), nach der auch mathematische Begriffsvorstellungen oder das Erschließen von Zusammenhängen innerhalb der Mathematik Komponenten mathematischer Grundbildung sind, bezieht das Rahmenkonzept des nationalen Ergänzungstests auch

mathematisches Arbeiten gezielt mit ein, das in *innermathematische* Kontexte eingebunden ist.

Der nationale Test enthält außerdem mehrere Aufgaben, die Faktenwissen und/oder Kenntnisse der Probanden über mathematische Verfahren ohne Einbindung der Aufgabe in einen Kontext abfragen und daher einer Kompetenzklasse „Technische Fertigkeiten“ zugeordnet sind, die im Zuge einer Ausdifferenzierung neu gebildet wurde. Im internationalen PISA-Konzept werden solche Fertigkeiten und Fähigkeiten als notwendige Voraussetzungen mathematischer Grundbildung verstanden. Aber eben weil sie das sind, sind Kenntnisse über den Leistungsstand der Probanden in diesen Fertigkeiten mit Blick auf die Beurteilung von Unausgewogenheiten im Leistungsprofil der Probanden ebenfalls wichtig. Defizite im „funktionalen Einsatz von Mathematik“ lassen sich nur dann als solche ausmachen, wenn die notwendigen Voraussetzungen für den funktionalen Einsatz der Mathematik, also mathematische Kenntnisse und Fertigkeiten selbst gesichert sind.

Der nationale Ergänzungstest bemüht sich schließlich auch, die gängigen mathematischen Stoffgebiete ausgewogen in Aufgaben zu repräsentieren, um bei der empirischen Analyse die Leistungen differenziert auch nach Stoffgebieten betrachten zu können.

Ebenso wie beim internationalen Test lassen sich die Aufgaben des nationalen Tests in *Kompetenzklassen* (jetzt jedoch 5 Klassen) einordnen<sup>1</sup>:

- Klasse 1A** Zu dieser Klasse gehören Aufgaben, die nur *technische Fertigkeiten* und/oder den Abruf von *Faktenwissen* erfordern.
- Klasse 1B** Zu dieser Klasse gehören Aufgaben, die eine *einschrittige Modellierung* erfordern, was oft direkt auf einen Algorithmus führt.
- Klasse 2A** Zur Lösung ist überwiegend ein einziger, *begrifflicher* Schritt erforderlich.
- Klasse 2B** Die Struktur der *Modellierung* ist *mehrschrittig* in dem Sinn, dass bei der Lösung Wissen aus mehreren mathematischen Zusammenhängen einzusetzen ist oder mehrfach gleichartige Schritte vorzunehmen sind.
- Klasse 3** Diese Klasse umfasst Aufgaben, deren Lösung einsichtsvolles mathematisches Denken, Begründen und/oder strukturelles Verallgemeinern erfordert.

Mit Blick auf die mathematischen Tätigkeiten bei der Lösung der Aufgaben hat sich die nationale PISA-Expertengruppe entschlossen, in diesem *Rahmenkonzept* die konkrete Bearbeitung und Lösung einer Aufgabe, abgesehen von den Aufgaben des Typs „Technische Fertigkeiten“ (1A) als einen Prozess der Erstellung, Verarbeitung und Interpretation eines mathematischen Modells zu sehen. Dabei wird der Begriff des *Modellierungsprozesses* weiter gefasst als üblich. Er ist nicht auf Aufgaben beschränkt, die in

---

<sup>1</sup> Eine detaillierte Beschreibung dieser Klassen findet der Leser in Neubrand et al. (2001).

einen außermathematischen Kontext eingebunden sind. Dort wird der Begriff des Modellierungsprozesses bekanntlich mit den Teilprozessen *Mathematisieren der Problemsituation – Arbeiten im mathematischen Modell - Interpretieren der Ergebnisse - Validierung* identifiziert (vgl. Blum (1996), Klieme et al. (2001), Schupp (1988)).

Von einem kognitionstheoretischen Standpunkt aus (vgl. Cohors-Fresenborg (1996), Cohors-Fresenborg & Sjuts (2001) und J. Neubrand (2002)) kann man die Strukturierung einer innermathematischen Problemsituation, die Erstellung des „Ansatzes“ aus einem Sach- oder Strukturzusammenhang als äquivalent zur „Mathematisierung“ einer außermathematisch gegebenen Problemsituation sehen. Bei einer Beschreibung des Lösungsprozesses einer im *innermathematischen* Kontext gegebenen Aufgabe erhalten dann die oben genannten Teilprozesse eine andere Bedeutung, z. B. entspricht der *Validierung* eine *Rückschau*, ob der verwendete *Ansatz* „elegant“ und effektiv“ war.

Die angesprochene Erweiterung des Modellierungsbegriffs ist keineswegs unumstritten, da der Begriff „Modellieren“ in der mathematikdidaktischen Diskussion der letzten 20 Jahre überwiegend im Sinne des Übersetzens zwischen *Realsituationen* und Mathematik verstanden wurde (vgl. Blum (1996) und Kaiser (1995)). Mögliche Positionen dazu sollen jedoch in dieser Arbeit nicht diskutiert werden.

Zu Analyse Zwecken sind alle Aufgaben, *nationale* wie *internationale*, sowohl nach den nationalen als auch den internationalen Kompetenzklassen eingeordnet worden.

### **I.3 Erweiterung der Aufgabenkategorienraster.**

Bisher haben wir das internationale und das nationale Rahmenkonzept dargestellt und über Kompetenzklassen charakterisiert. Die Testaufgaben lassen sich je nach didaktischer Fragestellung auch nach anderen Aspekten, z. B. Anforderungsprofilen, kategorisieren<sup>2</sup>, also unter anderem nach

- der Anzahl der bei der Bearbeitung angesprochenen *Wissensbereiche* und / oder der Art und dem Umfang des erforderlichen *stofflichen Wissens*<sup>3</sup>,
- den *mathematischen Tätigkeiten* bei der Bearbeitung der Aufgabe,
- dem *Kontext* der Aufgabenstellung,
- den geforderten *kognitiven Fähigkeiten*<sup>4</sup>,
- Art und Umfang der zum Bearbeiten erforderlichen „*Grundvorstellungen*“<sup>5</sup>.

Solche detaillierten Zuordnungen dienen dem Zweck, fachspezifische und fachdidaktisch interessante Analysen der Aufgaben durchführen zu können und z. B. *schwierigkeitsgenerierende Faktoren*<sup>6</sup> zu finden. Innerhalb der deutschen PISA-Expertengruppe

---

<sup>2</sup> Diese Gesichtspunkte gingen zum Teil auch schon in die Definition der nationalen Kompetenzklassen ein.

<sup>3</sup> vgl. hierzu J. Neubrand (2002)

<sup>4</sup> vgl. hierzu Cohors-Fresenborg & Sjuts (2001)

<sup>5</sup> vgl. hierzu vom Hofe (1995)

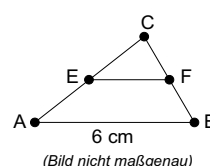
<sup>6</sup> zu diesem Begriff vgl. Klieme (1989, 2000)

wurde dazu ein Kategorienraster entwickelt, das z. B. in dem geplanten ausführlichen thematischen Bericht im kommenden Jahr zur Anwendung kommen wird.

Eine a priori - Kategorisierung von Aufgaben hat natürlich auch immer eine subjektive Komponente. Man sollte sich das generell bei Analysen von Aufgaben unter Kategorisierungen vergegenwärtigen. Betrachten wir dazu folgende Aufgabe aus dem nationalen Test:

### DREIECK

Die Seite  $\overline{AB}$  des Dreiecks  $ABC$  ist 6 cm lang. Es werden die Mittelpunkte  $E$  und  $F$  der Seiten  $\overline{AC}$  und  $\overline{BC}$  eingezeichnet. Wie lang ist  $\overline{EF}$ ?



Denkt man nur an Probanden, die sich an den „Satz von den Mittellinien“ des Dreiecks oder Spezialfälle des Strahlensatzes erinnern, so würde man die Aufgabe der Klasse 1A oder 2A zuordnen, je nachdem, wie sehr man den Schritt „bilde die Hälfte“ als begrifflich ansieht. Wer sich hier nur Probanden vorstellt, die Verhältnismäßigkeiten aufstellen und lösen, würde die Aufgabe der Klasse 1B zuordnen.

Um solche Probleme weitgehend zu vermeiden, erfolgte die nationale Einteilung der Aufgaben bei PISA durch ein *Rating*, bei dem sich die *Rater* Aufgabenlöser mit „Expertenwissen“ vorzustellen hatten (vgl. dazu auch J. Neubrand (2002)).

Sieht man die mit der Bearbeitung einer im Kontext gegebenen Aufgabe verbundene Tätigkeit als einen Modellierungsprozess, so kann man zwischen zwei Kategorien mathematischen Arbeitens unterscheiden, „rechnerischer Modellierung“ (Kompetenzklassen 1B und 2B) und „begrifflicher Modellierung“ (Kompetenzklassen 2A und 3) je nachdem, ob bei der Lösung der Aufgabe prozedurales Wissen oder konzeptuelles Wissen dominiert. Als dritte Kategorie „technische Aufgaben“ bleiben dann nur noch die Aufgaben der Kompetenzklasse 1A, d.h. die Aufgaben, bei deren Bearbeitung überwiegend technische Fertigkeiten oder Faktenwissen benötigt werden. Damit können drei Typen mathematischen Arbeitens unterschieden werden.

Diese Kategorisierung wird in Klieme et al. (2001) gewählt und wir werden sie auch in unsere Betrachtungen aufnehmen.

## II Methodische Aspekte der Testkonzeption

### II.1 Modellbetrachtungen und Skalen

Wie bereits in der Studie TIMSS II erfolgten die Testkonstruktion und Interpretation der Testergebnisse auf der Grundlage *logistischer* Testmodelle. Die dabei verwendete

Software CONQUEST (vgl. Wu (1998)) kann die Aufgabenparameter und Probandenparameter auch dann schätzen, wenn ein großer Aufgabenvorrat so auf mehrere Testhefte verteilt wird, dass ein Testheft immer nur einen Teil der Aufgaben enthält, es aber in je zwei Testheften immer gemeinsame Aufgaben gibt. Solche verbindenden Aufgaben nennt man *Ankeraufgaben*.

Für das Verständnis der momentan veröffentlichten Ergebnisse reicht es, wenn man das sogenannte zweikategorielle *Raschmodell* und seine Ableger kennt. Wir stellen diese Modelle in knapper Form vor und verweisen bezüglich ihrer Rechtfertigungsproblematik auf Knoche & Lind (2000), bezüglich der Modelldefinition und Schätzproblematik auf Rasch (1960), Lord & Novick (1968), Andersen (1991), Fischer & Molenaar (1995), Rost (1996) und Wu (1998).

**Zweikategorielles Raschmodell:** Bei der Modellierung der Testsituation in einer Probandenpopulation  $P$  wird unterstellt, dass die Bearbeitung einer Testaufgabe  $i$  durch einen Probanden  $k$  aus  $P$  als Zufallsexperiment angesehen werden kann, das die möglichen Ergebnisse 1 (akzeptable Bearbeitung) und 0 (nicht akzeptable Bearbeitung) besitzt. Die zugehörige Zufallsgröße mit den möglichen Werten 0 und 1 sei mit  $X_i^{(k)}$  bezeichnet.

Besteht  $P$  aus  $N$  Probanden und enthält der vorgelegte Test  $A$  insgesamt  $n$  Aufgaben, so wird als Modellannahme *postuliert*:

- (1) Jeder Aufgabe  $i \in A$  lässt sich ein Aufgabenparameter  $\delta_i \in \mathbb{R}$  (auch *Schwierigkeitsparameter* oder kurz *Aufgabenindex* genannt) zuordnen und jedem Probanden  $k \in P$  lässt sich ein Personenparameter  $\theta_k \in \mathbb{R}$  (auch *Fähigkeitsparameter* genannt) zuordnen, so dass für alle  $i \in A$  und alle  $k \in P$  gilt:

$$P(X_i^{(k)} = 1) = \frac{1}{1 + \exp(\delta_i - \theta_k)} .$$

( *logistische Aufgabencharakteristik* )

- (2) Für jeden Probanden  $k \in P$  sind die Zufallsgrößen  $X_1^{(k)}, \dots, X_n^{(k)}$  unabhängig.  
( *lokale stochastische Unabhängigkeit* )
- (3) Für jede Aufgabe  $i \in A$  sind die Zufallsgrößen  $X_i^{(1)}, \dots, X_i^{(N)}$  unabhängig. Dies gilt auch für die Antwortvektoren  $\vec{X}^{(1)}, \dots, \vec{X}^{(N)}$  aller Probanden.  
( *globale stochastische Unabhängigkeit* )

Wenn diese Modellvorstellung auf die Bearbeitung eines Tests  $A$  durch eine Probandenpopulation  $P$  zutrifft, so kann man davon sprechen, dass dieser Test ein *formal eindimensionales* Probandenmerkmal definiert. Dies bedeutet *nicht*, dass alle Aufgaben die gleichen Bearbeitungstechniken verlangen.

Da  $\theta_k$  ( $k=1, \dots, N$ ) *latente* Personenvariablen sind, kann man ihre Werte nur schätzen und benötigt dazu beobachtbare Indikatoren in Form von Testrohwerten und Lösungsquoten von Testaufgaben.

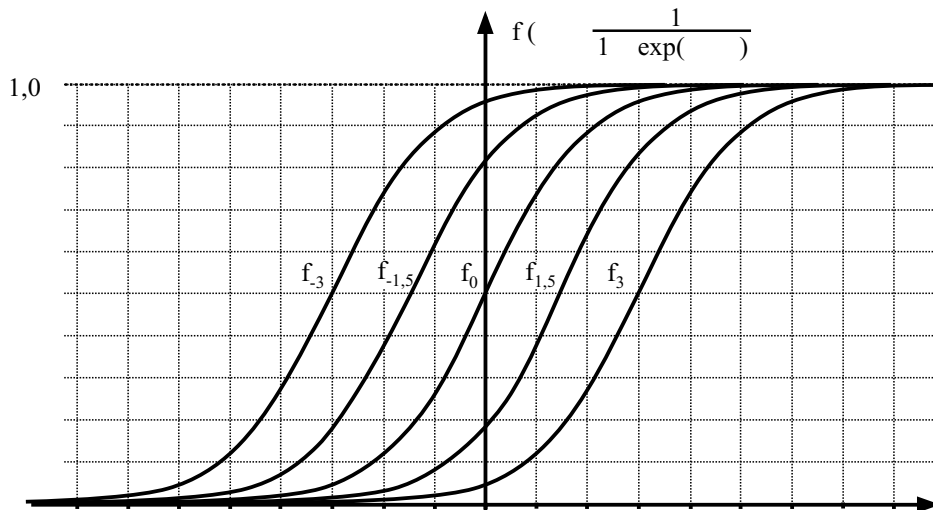
Man nennt  $P(X_i^{(k)} = 1)$  die *Lösungswahrscheinlichkeit* des Probanden  $k$  für die Aufgabe  $i$ . Offensichtlich hängt diese Wahrscheinlichkeit unter der genannten Modellvorstellung nur von der Differenz  $\delta_i - \theta_k$  ab. Daher ändert sich nichts an den Lösungswahrscheinlichkeiten, wenn man sowohl zu den Schwierigkeitsparametern als auch den Fähigkeitsparametern eine gemeinsame Konstante addiert. Bei der Schätzung der Modellparameter muss daher entweder einer der Parameter oder eine Parametersumme fixiert werden.

Vertritt man Probanden innerhalb des Modells durch ihre Fähigkeitsparameter und Aufgaben durch ihre Schwierigkeitsparameter, so nennt man die Funktion  $f_\delta: \mathbb{R} \rightarrow \mathbb{R}$  mit der Vorschrift

$$(II.1.1) \quad f_\delta(\theta) := \frac{1}{1 + \exp(\delta - \theta)} \quad (\text{für alle } \theta \in \mathbb{R})$$

die *Aufgabencharakteristik* einer Aufgabe mit Schwierigkeitsparameter  $\delta$ .

Bei der Modellprüfung setzt man *Itemfits* ein. Dies sind Aufgabenkennwerte, die auf dem Vergleich von theoretisch erwarteten mit tatsächlich beobachteten Lösungsquoten von Probanden in unterschiedlichen Leistungsgruppen beruhen.



**Abbildung 1:** Fünf Aufgabencharakteristiken eines Raschmodells



Wie Abb. 1 verdeutlicht, ist die Lösungswahrscheinlichkeit im Falle  $\theta = \delta$  gleich 0,5. An sich ist diese Modellvorstellung für *multiple choice Aufgaben* (im Folgenden stets kurz MC-Aufgaben genannt) unangemessen, wenn Probanden bei fehlendem Wissen Alternativen rein zufällig ankreuzen. In diesem Fall geht die Lösungswahrscheinlichkeit einer solchen Aufgabe nicht gegen Null, sondern liegt bei  $m$  Antwortalternativen für leistungsschwache Probanden in der Nähe von  $\frac{1}{m}$ . Wird darauf geachtet, dass die *Distraktoren* bekannten Fehlertypen entsprechen oder wenigstens für Unkundige attraktiver als die richtige Antwort wirken, so lässt sich der Rateeffekt soweit abschwächen, dass die Schätzung der Schwierigkeitsparameter kaum noch verzerrt wird. *Sinnhafte* Distraktoren haben zudem noch den Vorteil, diagnostische Aussagen zu ermöglichen. Zur Verdeutlichung sollen zwei nationale technische Aufgaben dienen:

#### MULTIPLIKATION

<p>Multipliziere aus und kreuze die richtige Antwort an: <math>(2x - 3y)^2 =</math></p> <p> <input type="checkbox"/> <math>4x^2 - 9y^2</math>      <input type="checkbox"/> <math>4x^2 + 6xy + 9y^2</math>      <input type="checkbox"/> <math>4x^2 - 6xy + 9y^2</math>  <input type="checkbox"/> <math>4x^2 - 12xy + 9y^2</math>      <input type="checkbox"/> <math>4x^2 - 12xy - 9y^2</math> </p>
--

Hier sind vier der üblichen Fehlermuster unter den Distraktoren vertreten. Die Lösungshäufigkeit lag an deutschen Gymnasien immerhin bei 66 %, über alle Schulformen dagegen nur bei 35 %. Dass die Aufgabe nicht zum Raten animierte, zeigen die bevorzugte Fehllösung „ $4x^2 - 9y^2$ “ und der sehr gute *Itemfit*.

#### RECHNUNG

<p>Berechne und kreuze die richtige Lösung an!</p> <p><math>4 + 3 \cdot (2 + 1) =</math>      <input type="checkbox"/> 11      <input type="checkbox"/> 13      <input type="checkbox"/> 14      <input type="checkbox"/> 15      <input type="checkbox"/> 21</p>
---

Hier gibt es für jede der Fehlantworten 11, 14, 15 und 21 eine Möglichkeit, sie durch falsches Umgehen mit der Bklammerung zu erhalten.

Mit einer Lösungsquote von rund 61 % war die Aufgabe „Rechnung“ so leicht, dass sie wohl ebenfalls nur sehr wenige Probanden zum Raten animierte. Dies lässt sich auch daraus schließen, dass ihr *Itemfit* noch akzeptabel war.

Bei der Modellparameterschätzung mit CONQUEST wird die theoretische Wahrscheinlichkeit  $L$  der beobachteten Datenmatrix unter der Annahme maximiert, dass die Personenparameter approximativ nach einem vorgegebenen Verteilungstyp (zum Beispiel einer Normalverteilung mit  $\mu_0 = 0$  und einer zu schätzenden Standardabweichung  $\sigma_0$ ) verteilt sind. Man nennt  $L$  die *Likelihood* der Datenmatrix. Dabei definiert jedes Testheft

als Teilttest seine eigenen Schätzgleichungen und es wird als Nebenbedingung bei der Schätzung verlangt, dass der Schätzwert  $\delta_a$  für den Schwierigkeitsparameter einer *Ankeraufgabe* in allen Teilttests der gleiche ist, in denen  $a$  vorkommt.

Da auf der Probandenseite unabhängig von der Populationsgröße nur wenige Verteilungsparameter zu schätzen sind, sind die Schätzungen der Schwierigkeitsparameter konsistent.

Werden für Detailanalysen auch Fähigkeitsparameter von Einzelpersonen gebraucht, so kann für einen Probanden mit dem Antwortvektor  $(x_1, \dots, x_n) \in \{0;1\}^{\times n}$  zu  $n$  Aufgaben mit den vorab mit CONQUEST geschätzten Schwierigkeitsparametern  $\delta_1, \dots, \delta_n$  der Fähigkeitsparameter  $\theta$  nach der Maximum-Likelihood-Methode durch Lösen folgender Gleichung geschätzt werden:

$$(II.1.2) \quad \sum_{i=1}^n \frac{1}{1 + \exp(\delta_i - \theta)} = \sum_{i=1}^n x_i.$$

Dieses Verfahren (kurz *ML-Schätzung* genannt) hat den Nachteil, dass Probanden mit 0 oder  $n$  richtig bearbeiteten Aufgaben kein Schätzwert für  $\theta$  zugewiesen werden kann. So sind nur  $n-1$  interpretierbare Werte für  $\theta$  möglich. Außerdem muss der Schätzfehler als relativ groß angesehen werden, da die bei der Schätzung verwendete Aufgabenzahl  $n$  jeweils nur die des bearbeiteten Testhefts ist. Für die PISA-Studie wurde daher wie schon in den TIMS-Studien das Verfahren der *Plausible Values* verwendet. Wir werden es inhaltlich im Anschluss an die Testmodelle beschreiben.

**Mehrdimensionales Raschmodell mit Aufgabengruppen** (vgl. Wu (1998)): Zerlegt man einen Test in Teilttests  $A_1, \dots, A_d$  und vermutet, dass zwar jeder dieser Tests für sich raschmodellierbar ist, dafür jedoch jedem Probanden spezifische Fähigkeitsparameter  $\theta_1, \dots, \theta_d$  zugewiesen werden müssen, so kann man eine  $d$ -fache Raschmodellierung vornehmen und regressionsanalytisch die Zusammenhänge zwischen den Fähigkeitsparametern untersuchen. Die Software CONQUEST bietet auch diese Option und schätzt sogar „fehlerkorrigierte“ Korrelationen zwischen den latenten Variablen  $\theta_1$  bis  $\theta_d$ .

Ein auf diese Weise angepasstes  $d$ -dimensionales Modell kann hinsichtlich der Anpassungsgüte mit dem eindimensionalen Modell verglichen werden, bei dem alle Aufgaben zusammengefasst wurden. In diesem Fall ist bekanntlich das Zweifache der Differenz der natürlichen Logarithmen der Datenlikelihood  $L_1$  des eindimensionalen und der Datenlikelihood  $L_d$  des  $d$ -dimensionalen Modells näherungsweise  $\chi^2$ -verteilt. Als Anzahl der Freiheitsgrade ist dabei der Unterschied in der Anzahl zu schätzender Parameter zu wählen.

**Weitere Modelle:** Hinsichtlich Verallgemeinerungen des Raschmodells auf mehr als zwei Antwortkategorien und der Einbeziehung von Begleitvariablen in die Testmodel-

lierung müssen wir auf spätere Veröffentlichungen verweisen. Es soll jetzt nur soviel gesagt werden, dass man einerseits dem Fähigkeitsparameter auf der Grundlage von Begleittests und erhobenen Probandenmerkmalen (zum Beispiel dem Geschlecht) Strukturen aufprägen kann und dann ein allgemeineres Verteilungsmodell für die Population erhält, andererseits jedoch auch die Schwierigkeitsparameter der Aufgaben von solchen Merkmalen abhängig machen kann.

**Plausible Values:** Hinter diesem Begriff verbirgt sich ein Verfahren, das mit folgendem Prinzip arbeitet: Wurden die Aufgabenparameter und die Verteilungsparameter der Population geschätzt, so kann man zu jedem Antwortvektor  $\bar{x}$  eines Probanden  $k$  die *bedingte Wahrscheinlichkeitsverteilung*  $\Psi(\bar{x})$  des zu schätzenden Fähigkeitsparameters  $\theta_k$  bestimmen<sup>7</sup>. Mit Hilfe geeignet transformierter Zufallszahlen kann man einen oder mehrere „zufällige“ Werte  $\theta_k^{(1)}, \dots, \theta_k^{(m)}$  für den Probanden  $k$  generieren, von denen jeder nach  $\Psi(\bar{x})$  verteilt ist. Wenn zusätzliche Begleitmerkmale in die Konstruktion der jeweiligen bedingten Verteilung eingehen, ist die Varianz dieser Verteilung deutlich kleiner als die Varianz des ML-Schätzers (vgl. Mislevy et al. (1992a,1992b)).

Im Rahmen von Regressionsanalysen bietet die Verwendung von *Plausible Values* den Vorteil, dass man sich gegen die Gefahr der *Artefaktbildung* absichern kann, indem man beispielsweise derartige Analysen mit dem ersten Plausible Value durchführt und mit den anderen Plausible Values wiederholt.

Will man nur mit einem einzigen Schätzwert für den Fähigkeitsparameter jedes Probanden arbeiten, so kann man den *Erwartungswert*  $\theta^{\text{cap}}$  von  $\Psi(\bar{x})$  als Schätzwert für  $\theta$  nehmen, nimmt dabei jedoch in Kauf, dass  $\theta^{\text{cap}}$  keine erwartungstreue Schätzung der Populationsvarianz von  $\theta$  liefert. Die Abkürzung „cap“ steht für *expected ability parameter*.

**Der OECD-PISA-Index:** In Baumert et al. (2001b) heißt es, dass „die Skala der mathematischen Grundbildung international so normiert wurde, dass über alle OECD-Staaten hinweg der Mittelwert von  $\theta$  gleich 500 und die Standardabweichung von  $\theta$  gleich 100 ist.“

Bei solchen Transformationen geht es vor allem um die bessere Lesbarkeit von tabellarischen Aufstellungen. Es lässt sich nämlich erreichen, dass die transformierten Schätzwerte der Fähigkeitsparameter alle positiv sind und ohne nennenswerte Vergrößerung des Schätzfehlers ganzzahlig gerundet werden können.

---

<sup>7</sup> Bei der Anwendung eines Populationsmodells liegt nicht nur der Antwortvektor zugrunde. In die bedingte Verteilung gehen auch die Messwerte der Begleitvariablen ein.

Wenn  $\bar{\theta}$  der Mittelwert und  $\sigma$  die Standardabweichung der geschätzten  $\theta$ -Werte in der OECD-Gesamtpopulation sind, so verwendet man für diese Normierung die lineare Transformation  $\tau : \theta \mapsto \theta_{\text{OECD}}$  mit der Vorschrift

$$(II.1.2) \quad \tau(\theta) := 500 + 100 \frac{\theta - \bar{\theta}}{\sigma}.$$

Umgekehrt lässt sich mit der Vorschrift

$$(II.1.3) \quad \tau^{-1}(\theta_{\text{OECD}}) := \bar{\theta} + \sigma \frac{\theta_{\text{OECD}} - 500}{100}$$

aus dem OECD-Fähigkeitswert eines Probanden der ursprüngliche Schätzwert  $\theta$  auf der ursprünglichen Modellskala (auch *logit*-Skala genannt) berechnen.

Auch die Schwierigkeitsparameter wurden mit der Vorschrift (II.1.2) umgerechnet. Danach wurde jedoch im Zuge einer Kompetenzstufenbildung (Näheres dazu in Abschnitt II.2) der Schwierigkeitsparameter  $\delta_{\text{OECD}}$  einer Aufgabe auf der OECD-Skala so undefiniert, dass ein Proband mit  $\theta_{\text{OECD}} = \delta_{\text{OECD}}$  für diese Aufgabe die Lösungswahrscheinlichkeit 0,62 besitzt.

Damit ergibt sich aus der Standardabweichung  $\sigma$  und dem Mittelwert der Fähigkeits-schätzwerte  $\bar{\theta}$  in der OECD-Population aus (II.1.1) durch Anwenden von (II.1.2) auf beide Modellparameter folgende Gleichung einer Rasch-Charakteristik, in der zwecks einfacherer Notation der *PISA-Aufgaben-Index* mit  $\delta$  und der auf der OECD-Skala gemessene Fähigkeitsparameters mit  $\theta$  bezeichnet sind<sup>8</sup>:

$$(II.1.4) \quad f_{\delta}(\theta) = \frac{1}{1 + \frac{0,38}{0,62} \exp\left(\frac{\delta - \theta}{76,27}\right)}$$

Als Beispiel für die Verwendung dieser Beziehung soll folgende Aufgabe dienen:

GLASFABRIK, Version 2

Eine Glasfabrik stellt Flaschen her. 2 % der Flaschen sind fehlerhaft; dies sind 160 Flaschen. Wie viele Flaschen wurden insgesamt hergestellt?

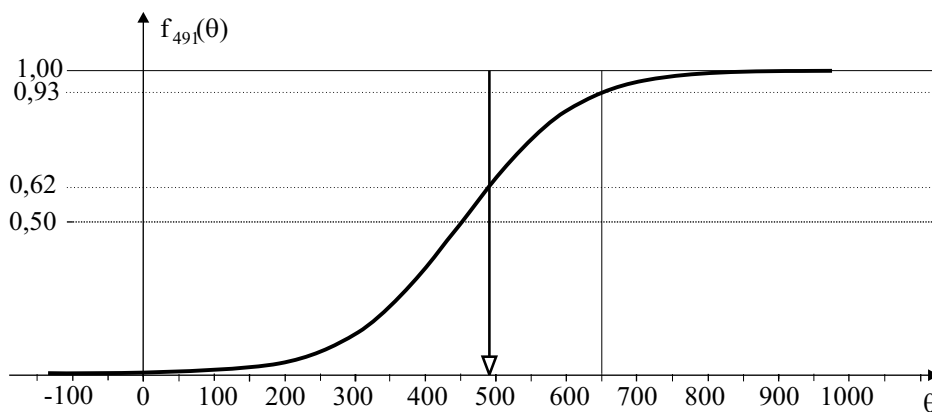
320 Flaschen     3200 Flaschen     800 Flaschen     8000 Flaschen

12 500 Flaschen

Diese Aufgabe hat den PISA-Index 491, sie wurde von etwa 65 % der Probanden gelöst.

---

<sup>8</sup> Diese Bezeichnungen sollen ab jetzt durchgängig verwendet werden, wenn nichts anderes gesagt wird.



**Abbildung 2:** Aufgabencharakteristik der Aufgabe „Glasfabrik, Version 2“.

Wie aus der Gleichung  $f_{491}(\theta) = \frac{1}{1 + \frac{0,38}{0,62} \exp(\frac{491-\theta}{76,27})}$  ablesbar ist, hat ein Proband mit

einem Fähigkeitswert von  $\theta = 491$  bei dieser Aufgabe eine Lösungswahrscheinlichkeit von 0,62. Ein Proband mit einem  $\theta$ -Wert von 650 hat dagegen bereits eine Lösungswahrscheinlichkeit von etwa 0,93.

## II.2 Die Bildung von Kompetenzstufen

Da es bei großen Untersuchungen nur wenig Sinn macht, viele Einzelwerte zu diskutieren, bietet sich eine Unterteilung sowohl des Aufgabenschwierigkeitsbereichs als auch des Fähigkeitswertebereichs in *Intervalle* an. Bei PISA-2000 einigte man sich dazu auf die Bildung von 5 sogenannten *Kompetenzstufen*<sup>9</sup>, die durch Intervalle für den Aufgabenparameter  $\delta$  vertreten werden sollten. Die Stufen wurden simultan für Probanden und Aufgaben durch die Forderung festgelegt, dass Probanden am unteren Ende einer Stufe die leichtesten Aufgaben der Stufe und Probanden am oberen Ende einer Stufe die schwersten Aufgaben der Stufe mit der Wahrscheinlichkeit 0,62 lösen können (vgl. Artelt et al. (2001b), S. 95 und Baumert et al. (2001b), S. 52). Das gewählte Verfahren lässt sich wie folgt charakterisieren:

<sup>9</sup> Eigentlich wäre die Bezeichnung *Leistungsstufe* angebracht, da es zunächst einmal nur um Intervalle für den Schwierigkeitsparameter geht. Wir werden im Folgenden trotzdem den Begriff *Kompetenzstufe* beibehalten, weil er dem Sprachgebrauch entspricht (Verfahren dieser Art werden auch unter dem englischen Begriff *Proficiency Scaling* diskutiert).

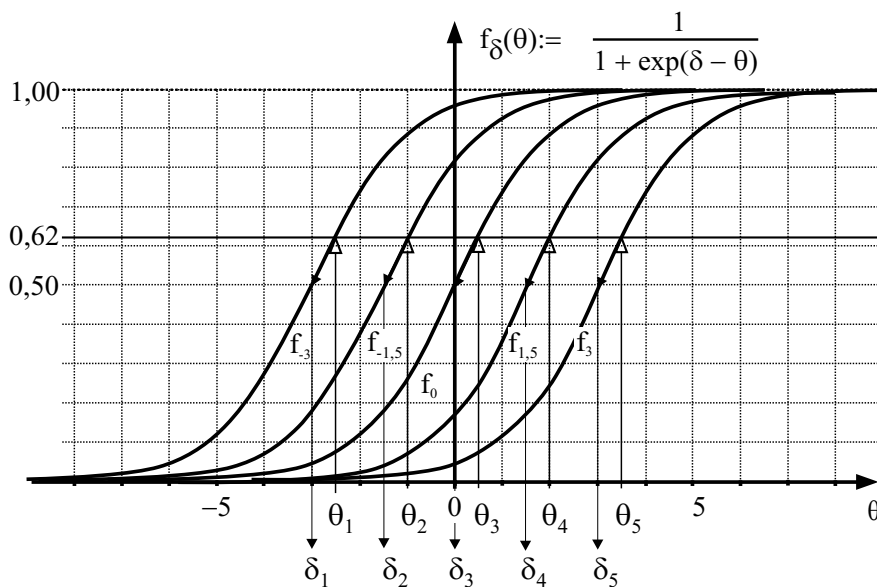
Der Begriff *Kompetenzstufe* darf nicht mit dem *Kompetenzklassen*begriff verwechselt werden!

**Kompetenzstufendefinition, Variante 1:** Man unterteilt einen für wesentlich gehaltenen Abschnitt des beobachteten Leistungsintervalls  $[\theta_{\min}; \theta_{\max}]$  in  $n$  Teilintervalle, von denen die ersten  $I_1, I_2, \dots, I_{n-1}$  gleich lang sind und  $I_n$  ein nach oben offenes Randintervall ist (bei PISA-2000 wurde  $n$  gleich 5 gewählt):

$$I_1 := [\theta_1; \theta_2[ , I_2 := [\theta_2; \theta_3[ , I_3 := [\theta_3; \theta_4[ , I_4 := [\theta_4; \theta_5[ , \dots , I_n := [\theta_n; \infty[ .$$

Vor  $I_1$  hat man sich ein ebenfalls offenes Randintervall  $I_0 := ] - \infty; \theta_1[$  vorzustellen, in dem „völlig inakzeptable“ Fähigkeitswerte liegen.

Nach Festlegung einer *Mindestlösungschance*  $p_{\min}$ , die das akzeptable Beherrschen einer Aufgabe kennzeichnen soll, stellt man sich zu jeder Grenze  $\theta_j$  ( $j=1, \dots, n$ ) eine fiktive Testaufgabe  $A_j$  vor, bei der die Lösungswahrscheinlichkeit eines Probanden mit dem Leistungswert  $\theta_j$  gleich  $p_{\min}$  ist. Daraus ergibt sich jeweils ein Schwierigkeitsparameter  $\delta_j$  für diese Aufgabe. Die Schwellen  $\delta_1, \delta_2, \delta_3, \delta_4, \dots, \delta_n$  teilen die *Schwierigkeitskala* in Intervalle, *Kompetenzstufen* genannt, (vgl. Abb. 3) und man kann diesen Stufen tatsächliche Testaufgaben zuordnen.



**Abbildung 3:** Festlegung von fünf Kompetenzstufengrenzen durch Unterteilung der Fähigkeitskala.

Die Abbildung verdeutlicht das Verfahren am Beispiel PISA-2000. Dabei werden Modellparameter auf der ursprünglichen Skala des Raschmodells angegeben.

Bei der Wahl von  $p_{\min}$  orientiert man sich daran, dass es im Test Aufgaben zu jeder der Stufen geben muss. Da der internationale Test nur 31 Mathematikaufgaben enthielt, musste für eine befriedigende Klassenbesetzung  $p_{\min}$  mit 0,62 etwas kleiner als in der Studie TIMSS II gewählt werden (dort war  $p_{\min}$  gleich 0,65).

Wird der *Schwierigkeitsparameter* einer Aufgabe gleich dem Fähigkeitsparameter eines Probanden gesetzt, der diese Aufgabe mit der Wahrscheinlichkeit  $p_{\min}$  löst, so stimmen die Intervallgrenzen auf der *Fähigkeitsskala* und der *Schwierigkeitsskala* dem Wert nach überein.

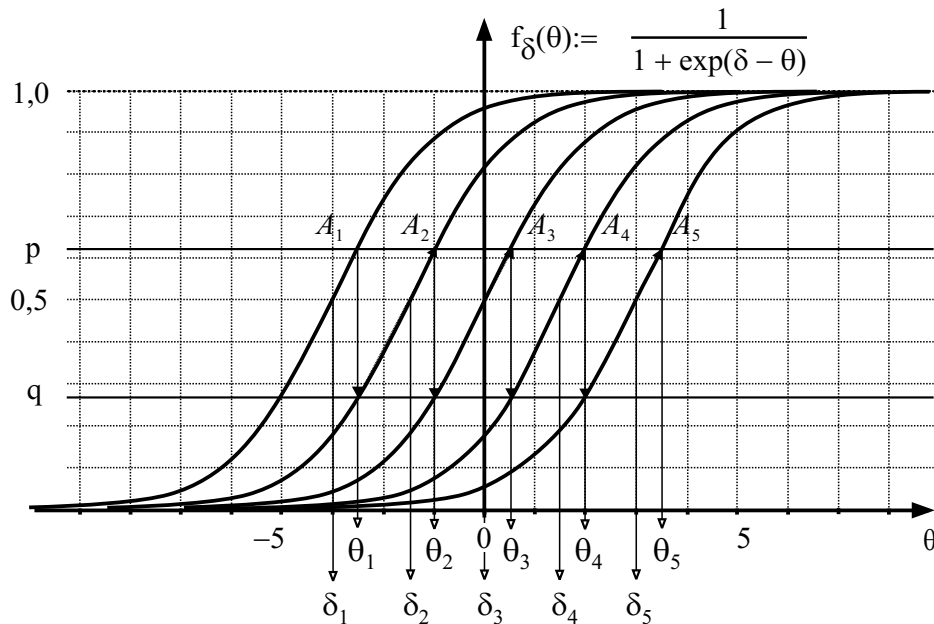
In Anlehnung an Beaton und Allen (1992) hätte bei mehr Testaufgaben auch folgendes Verfahren gewählt werden können:

**Kompetenzstufendefinition, Variante 2:** Man legt einen Wert für den Schwierigkeitsparameter  $\delta_1$  einer fiktiven Testaufgabe  $A_1$  fest, welche die unterste Kompetenzstufe (I) vertreten soll. Nach Festlegung einer (Mindest-)Lösungswahrscheinlichkeit  $p_{\min}$ , die das akzeptable Beherrschen von Aufgaben kennzeichnen soll, werden die nächsten fiktiven Testaufgaben folgendermaßen bestimmt:

1. Es wird ein fiktiver „Modellproband“ betrachtet, dessen Fähigkeitsparameter  $\theta_1$  gerade so groß ist, dass er  $A_1$  mit der Wahrscheinlichkeit  $p_{\min}$  löst.
2. Die fiktive Testaufgabe  $A_2$  für die Kompetenzstufe (II) wird so definiert, dass es der Modellproband mit Leistungswert  $\theta_1$  nur mit einer vorgegebenen geringen Wahrscheinlichkeit  $q < p_{\min}$  löst. Daraus ergibt sich der Schwierigkeitsparameter  $\delta_2$  für die Aufgabe  $A_2$ .
3. Die Schritte 1 und 2 werden sinngemäß mit fixierten Werten  $p_{\min}$  und  $q$  so lange wiederholt, bis man die gewünschte Zahl  $n$  von Stufen definiert hat.

Auch dieses Verfahren soll an einer Abbildung für fünf Kompetenzstufen verdeutlicht werden. Wie Abb. 4 zeigt, ergeben sich auch hier äquidistante Schwellenwerte  $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5$  mit zugeordneten Leistungsgrenzen  $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5$ .

Im Gegensatz zur ersten Variante ist jetzt jedoch die Breite der *Fähigkeitsintervalle* nicht vorgegeben, da sie sich aus  $p_{\min}$  und  $q$  über die „Konstruktion“ der *Markieraufgaben*  $A_1, A_2, A_3, A_4$  und  $A_5$  ergibt.



**Abbildung 4:** Festlegung von fünf Kompetenzstufengrenzen durch rekursive Bestimmung von Markieraufgaben.

### III Inhaltliche Beschreibung der Tests

In Abschnitt I wurden die grundlegenden Konzepte der beiden Tests kurz umrissen. Im vorangegangenen Abschnitt II wurden dann modelltheoretische Betrachtungen durchgeführt, die unter Berücksichtigung der Konzepte der Tests ebenfalls für die Auswahl der Aufgaben, jetzt mit Blick auf die technische, methodische Analyse der Daten von Bedeutung waren. Im Folgenden wollen wir Beispielaufgaben aus den beiden Tests vorstellen, auf die wir dann zurückgreifen können. Dabei soll auch deutlich werden, wie sich die Aufgaben in die Rahmenkonzeptionen der beiden Tests einordnen.

#### III.1 Der internationale Test

In Klieme et al. (2001) wurde die in Abbildung 5 gezeigte Aufgabe zur inhaltlichen Beschreibung des internationalen Pisa-Konzepts ausgewählt und kommentiert.<sup>10</sup>

<sup>10</sup> Bisher sind erst wenige Aufgaben zu Demonstrationszwecken freigegeben, da die meisten Aufgaben bei PISA-2003 nochmals verwendet werden sollen. Der Leser sollte also nicht verwundert darüber sein, dass er in Berichten zu PISA nahezu immer die gleichen Beispielaufgaben sieht.



Wie die meisten der internationalen PISA-Aufgaben beginnt die Aufgabe mit einer allgemeinen Beschreibung einer Problemsituation. Daran schließen sich Fragen an, die immer weiter in einen Prozess des mathematischen Modellierens hineinführen.

Die Aufgabe gehört nach dem Kategorisierungsschema des internationalen Tests zur Leitidee „Veränderung und Wachstum“, die Teilaufgaben „Äpfel 1“ und „Äpfel 3“ wurden der internationalen Kompetenzklasse<sup>11</sup> 2 und die Teilaufgabe „Äpfel 2“ der Kompetenzklasse 1 zugeordnet.

Unter dem Blickwinkel des nationalen Ergänzungstests wurden die Teilaufgaben „Äpfel 1“ und „Äpfel 3“ der Kompetenzklasse 2A, d.h. der Kategorie „begriffliches Modellieren“, die Teilaufgabe „Äpfel 2“ der Klasse 1B, d.h. der Kategorie „Rechnerisches Modellieren“ zugeordnet.

### ÄPFEL:

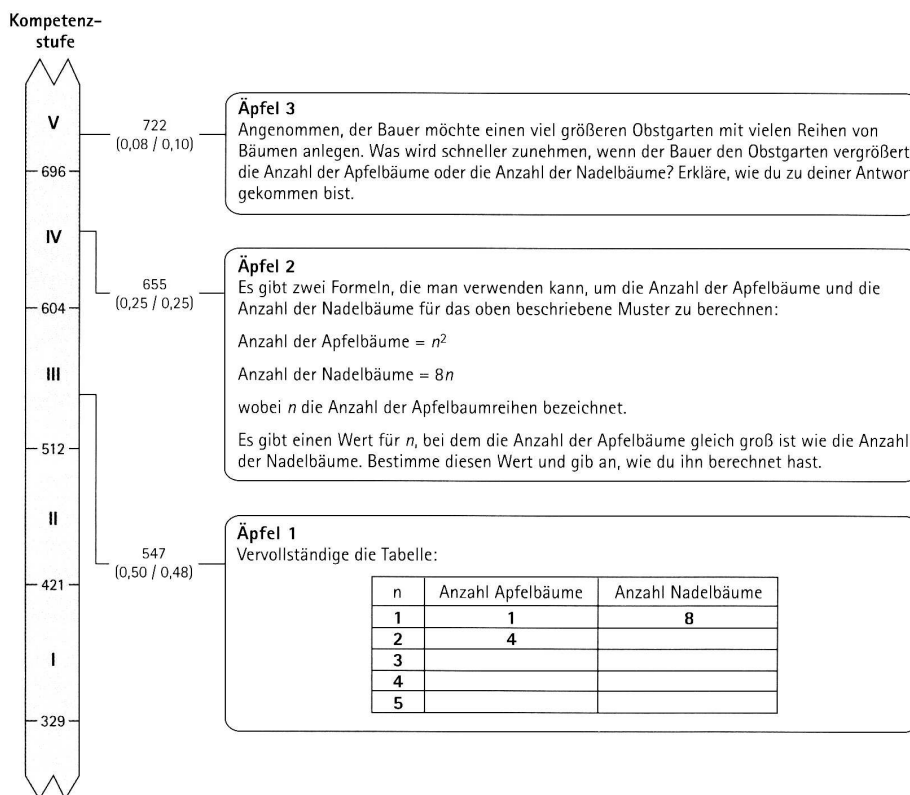
Ein Bauer pflanzt Apfelbäume an, die er in einem quadratischen Muster anordnet. Um diese Bäume vor dem Wind zu schützen, pflanzt er Nadelbäume um den Obstgarten herum.

Im folgenden Diagramm siehst du das Muster, nach dem Apfelbäume und Nadelbäume für eine beliebige Anzahl ( $n$ ) von Apfelbaumreihen gepflanzt werden:

n = 1	n = 2	n = 3	n = 4
x x x	x x x x x	x x x x x x x	x x x x x x x x x
x • x	x • • x	x • • • x	x • • • • x
x x x	x • x	x • • x	x • • • • x
	x • • x	x • • • x	x • • • • x
	x x x x x	x • • • x	x • • • • x
		x x x x x x x	x • • • • x
			x • • • • x
			x x x x x x x x x

x = Nadelbaum  
• = Apfelbaum

<sup>11</sup> Dieser Begriff wurde in Abschnitt I.1 als Teil eines Kategorisierungsschemas beschrieben und darf nicht mit dem empirisch festgelegten Kompetenzstufenbegriff verwechselt werden!



**Abbildung 5:** Der Aufgabenblock ÄPFEL aus dem internationalen PISA-Test. An den Verbindungslinien werden jeweils der PISA-Index und in Klammern die Lösungsquoten OECD/national angegeben.  
 (nach Abb. 3.2 aus Klieme et al. (2001), S. 148)

Ein zweites Beispiel ist die Aufgabe „Bauernhöfe“ (Abb. 6). Diese Aufgabe gehört zur Leitidee „Raum und Form“, die Teilaufgabe „Bauernhöfe 1“ gehört zur internationalen Kompetenzklasse 1, die Teilaufgabe „Bauernhöfe 2“ zur Kompetenzklasse 2.

Unter dem Blickwinkel des nationalen Ergänzungstests wurden beide Teilaufgaben der Klasse 1B, d.h. der Kategorie „Rechnerisches Modellieren“ zugeordnet.

Kompetenzstufe

V

696

IV

604

III

512

II


421

I

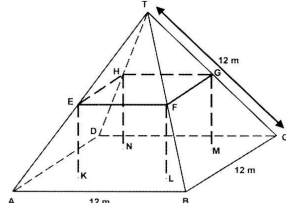
329

**Bauernhöfe**

Hier siehst du ein Foto eines Bauernhauses mit pyramidenförmigem Dach.



Nachfolgend siehst du eine Skizze mit den entsprechenden Maßen, die eine Schülerin vom **Dach** des Bauernhauses gezeichnet hat.



Der Dachboden, in der Skizze ABCD, ist ein Quadrat. Die Balken, die das Dach stützen, sind die Kanten des Quaders (rechtwinkliges Prisma) EFGHLMN. E ist die Mitte von  $\overline{AT}$ , F ist die Mitte von  $\overline{BT}$ , G ist die Mitte von  $\overline{CT}$  und H ist die Mitte von  $\overline{DT}$ . Jede Kante der Pyramide in der Skizze misst 12 m.

492  
(0,61 / 0,51)

524  
(0,55 / 0,41)

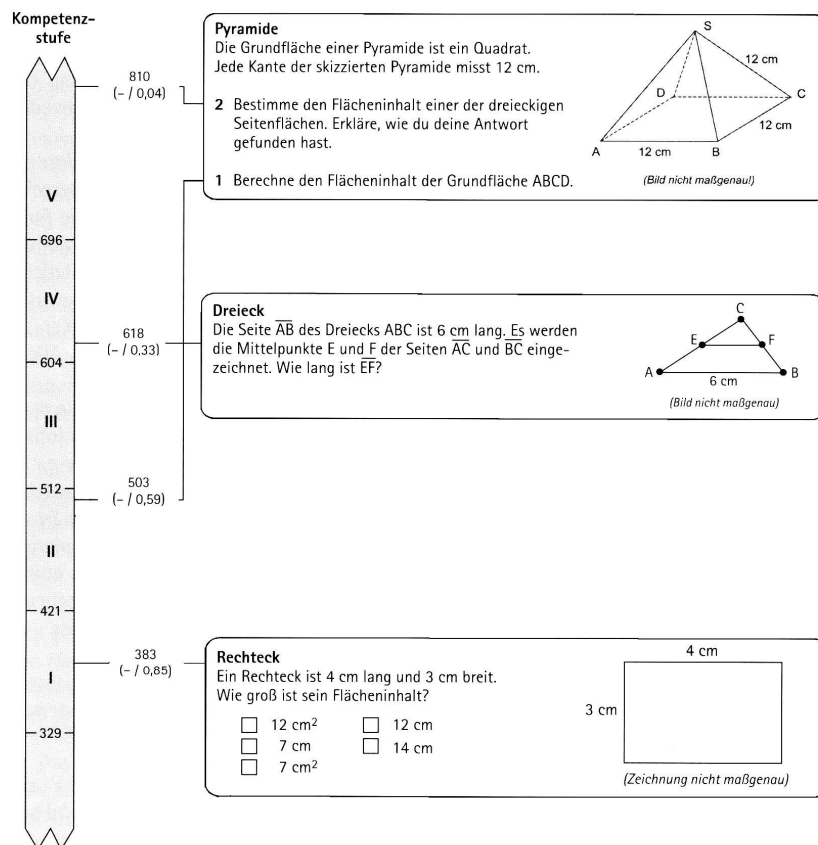
2 Berechne die Länge von  $\overline{EF}$ , einer der waagerechten Kanten des Quaders.  
Die Länge von  $\overline{EF}$  = \_\_\_\_\_ m

1 Berechne den Flächeninhalt des Dachbodens ABCD.  
Der Flächeninhalt des Dachbodens ABCD = \_\_\_\_\_ m<sup>2</sup>

**Abbildung 6:** Der Aufgabenblock BAUERNHÖFE aus dem internationalen PISA-Test. (Auszug aus Abb. 3.3 in Klieme et al. (2001), S. 152)

### III.2 Der nationale Ergänzungstest

In I.2 wurde das Konzept skizziert, unter dem der nationale Ergänzungstest das Konzept des internationalen Tests ergänzt. Einige nationale Aufgaben wurden bereits in I.3 und II.1 vorgestellt. Weitere Aufgaben zeigt die folgende Abbildung, in der die Aufgabe „Dreieck“ noch einmal zwecks Mitteilung ihres Schwierigkeitsparameters aufgenommen wurde:



**Abbildung 7:** Geometrische Aufgaben aus dem nationalen PISA-Test.  
(Auszug aus Abb. 3.3 in Klieme et al. (2001), S. 152)

Die Aufgabe „Pyramide 1“ stellt eine Wiederholung der internationalen Aufgabe Bauernhöfe 1 ohne Realkontext dar.

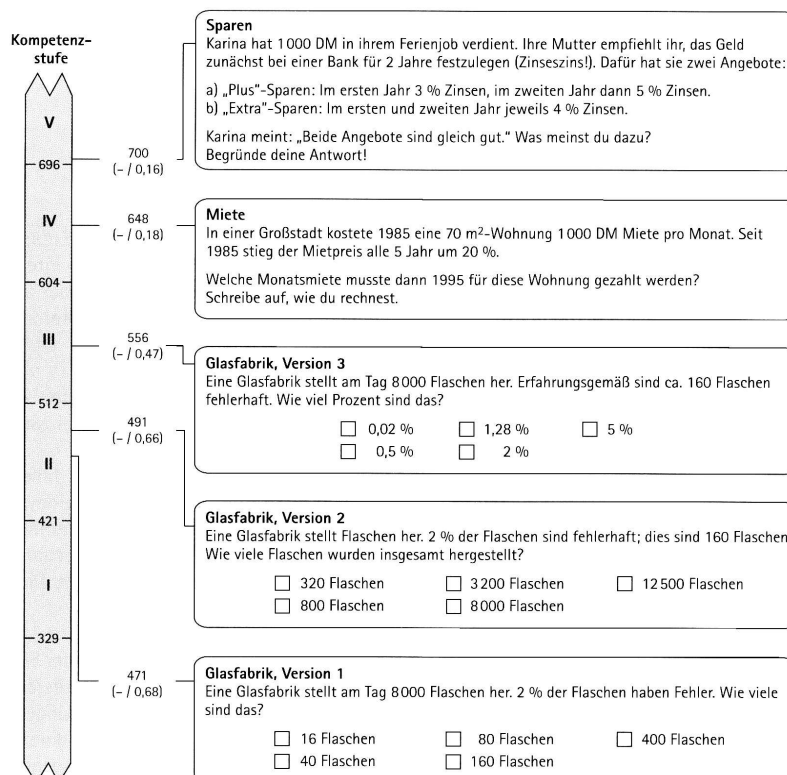
Die nationale Aufgabe „Dreieck“ ist in ihrer Struktur mit „Bauernhöfe 2“ vergleichbar, auch wenn sie nicht im strengen Sinn eine Parallelversion mit innermathematischem Kontext ist. Bei der Aufgabe „Bauernhöfe 2“ liegt ein gleichseitiges Dreieck vor, bei der Aufgabe „Dreieck“ ist das nicht der Fall.

Bei der Aufgabe „Pyramide 2“ aus dem nationalen Test handelt es sich um eine anspruchsvolle Aufgabe, die neben geometrischem Wissen auf relativ hohem Niveau auch eine Strukturierung des Lösungsprozesses erfordert.

Die Aufgaben „Rechteck“ und „Dreieck“ wurden der nationalen Kompetenzklasse 1A („Technische Aufgaben“) zugewiesen, die Teilaufgabe „Pyramide 1“ der Klasse 1B

und die Teilaufgabe „Pyramide 2“ der Klasse 2B, d.h. beide der Kategorie „Rechnerisches Modellieren“ zugeordnet.

Ein zweites Beispiel stellt die folgende Aufgabensequenz dar, in der die bereits vorgestellte Aufgabe „Glasfabrik, Version 2“ noch einmal aufgeführt ist.



**Abbildung 8:** Aufgaben zur Prozentrechnung aus dem nationalen PISA-Test. (Auszug aus Abb. 3.4 in Klieme et al. (2001), S. 154)

Die Aufgabensequenz ist dem Themenbereich „Proportionalität“ entnommen. Das Ziel dieser Aufgabenfolge ist, das Lösungsverhalten der Probanden bei einer Stufung zunehmend komplexerer Sachsituationen aus ein und demselben Stoffbereich zu verfolgen. Alle Aufgaben gehören der Klasse „Rechnerisches Modellieren“ an.

Den Abschluss sollen einige Beispiele aus der Klasse „Technische Fertigkeiten“, also der nationalen Aufgabenklasse 1A, bilden:

Wie bereits die schon angesprochenen Aufgaben „Rechteck“ (PISA-Index 383), „Rechnung“ (PISA-Index 504) und „Multiplikation“ (PISA-Index 612) zeigen, streuten die Aufgabenschwierigkeiten innerhalb dieser Aufgabenklasse über einen weiten Bereich. Dies hatte die nationale Expertengruppe bei der Testkonstruktion auch so beabsichtigt. Dass sogar in allen Kompetenzstufen Aufgaben aus der Klasse 1A vertreten sind, zeigen die beiden folgenden Aufgaben, die den Kompetenzstufen III bzw. V zuzuordnen sind:

FUNKTION, Teilaufgabe (3) (PISA-Index 583)

Die Funktion mit der Gleichung  $y = 2x - 1$  soll untersucht werden.

(3) Berechne für  $x = 100$  den  $y$ -Wert.

QUADRATISCHE GLEICHUNG (PISA-Index 797)

Löse die Gleichung  $4x + 4 = 3x^2$ .

### III.3 Vergleich der Teststrukturen

Die Verteilung der Aufgaben nach ihrer Schwierigkeit zeigt, dass es bei den internationalen Aufgaben keine Aufgaben gibt, die den deutschen Probanden *extrem* schwer fallen:

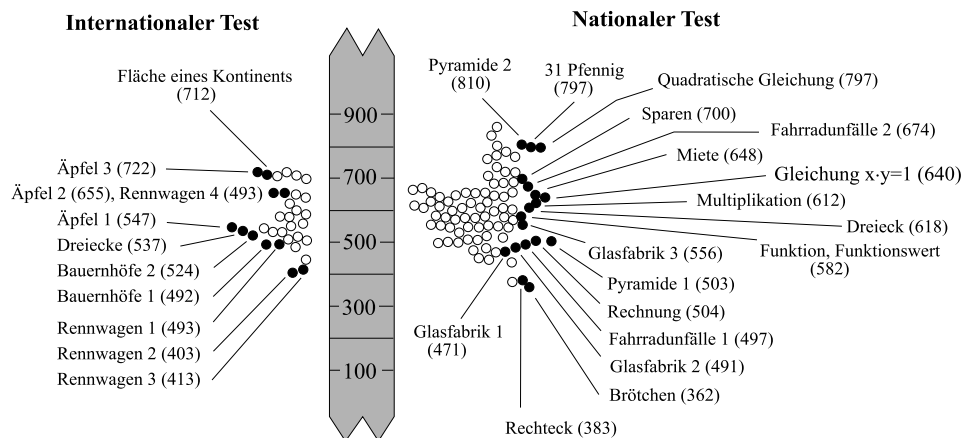
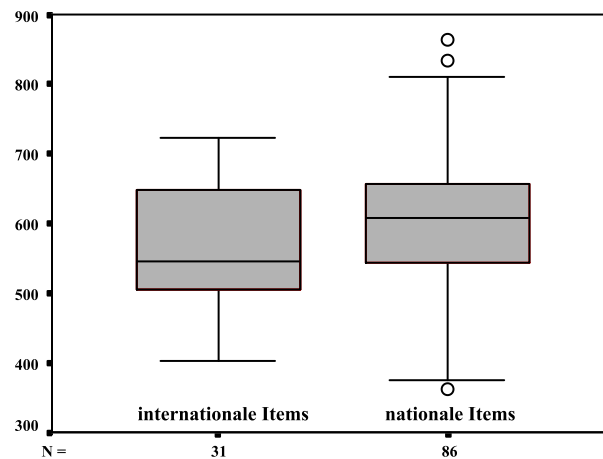


Abbildung 9: Verteilung der Aufgaben nach ihrem PISA-Index .

Dass der nationale Test eine etwas breitere Streuung der Aufgabenschwierigkeiten (im Sinne der Schwierigkeitsspannweite) besitzt, geht auch aus den beiden folgenden Boxplots hervor.



**Abbildung 10:** Verteilung der Schwierigkeitsparameter

Die Lage der Medianwerte macht deutlich, dass im internationalen Test die Verteilung der Aufgabenschwierigkeitsparameter unsymmetrischer als im nationalen Zusatztest ist. Zusätzlich dokumentiert Abb. 10, dass der internationale Test für die deutsche Schülerpopulation etwas *leichter* (!) als der nationale Test war. Für die Kombination beider Tests lässt sich nach Abb. 9 und Abb. 10 sagen, dass die Testaufgaben hinsichtlich ihrer Schwierigkeitsverteilung gut geeignet sind, um innerhalb der deutschen Gesamtpopulation zwischen Mathematikleistungen zu differenzieren.

Im rechten Boxplot werden drei Aufgaben als Ausreißer genannt. Am unteren Ende der Schwierigkeitsskala handelt es sich dabei um die leichteste Aufgabe des gesamten Tests:

BRÖTCHEN (PISA-Index 362)

7 Brötchen kosten 3,15 DM. Was kosten 11 Brötchen?

5,05 DM   
  4,95 DM   
  4,85 DM   
  4,75 DM   
  4,65 DM

Die Verschiedenartigkeit beider Verteilungen legt sofort die Frage nahe, ob denn die beiden Teiltests verschiedene „Arten der Mathematikleistung“ im Sinne von Dimensio-

nen messen. Diese Frage kann im Hinblick auf ein *allgemeines Mathematikleistungs-konstrukt* verneint werden, da sich die mit dem Gesamttest erhobenen Daten gut durch eines der in Abschnitt II.1 beschriebenen formal eindimensionalen Testmodelle erklären lassen (vgl. Klieme et al. (2001), S. 156ff).

#### IV Befunde und Analysen

In II.2 wurde das Verfahren zur Entwicklung der Fähigkeitsskala aus den „Schwierigkeitsparametern“ der Aufgaben erläutert und es wurde deutlich, wie diese Skala in *Kompetenzstufen* unterteilt worden ist. Ein wichtiger Schritt bei der Interpretation dieser Stufen ist ihre inhaltliche Spezifikation. Das geschieht über eine Analyse der Anforderungsmerkmale der Aufgaben, deren „Schwierigkeitsstufen“ den betreffenden Stufen der Fähigkeitsskala entsprechen.

Den Begriff Kompetenzstufe darf man nicht a priori mit kognitionstheoretischen Betrachtungen zum individuellen Lösungsverhalten verbinden. Aufgaben können nämlich aus unterschiedlichen Gründen „schwer“ sein, wie die Analysen in Neubrand et al. (2002) zeigten. Dies wird auch die Analyse ausgewählter Items in Abschnitt IV.2. zeigen.

Daher scheint es angemessen und inhaltlich konsistenter, eine Aufschlüsselung der Inhalte der einzelnen Kompetenzstufen je nach dem angesprochenen Typ mathematischen Arbeitens - *technische Aufgaben* oder *rechnerische* bzw. *begriffliche Modellierungsaufgaben* (vgl. I.3) - vorzunehmen. Eine Beschreibung der Kompetenzstufen über alle Aufgaben gleich welchen Typs hinweg, wie noch approximativ in Klieme et al. (2001) geschehen, verwischt die unterschiedlichen Merkmale, die die Schwierigkeit von Aufgaben beeinflussen. In knapper Darstellung lassen sich die einzelnen Stufen demnach wie folgt beschreiben, wobei die von ACER angegebene Stufengliederung (vgl. II.2, Variante 1) zu Grunde gelegt wird. Die zur jeweiligen Kompetenzstufe zu zählenden Aufgaben des PISA-Tests erfordern im wesentlichen die in den Zellen angegebenen Kompetenzen<sup>12</sup>.

	<i>technische Items</i>	<i>rechnerische Modellierungsaufgaben</i>	<i>begriffliche Modellierungsaufgaben</i>
<b>Stufe I</b> [329;421]	In den Aufgaben wird Wissen über einfache Fakten abgefragt, die i.a. bereits in der Grundschule zur Verfügung stehen.	Standardmathematisierungen sind auf einem Niveau vorzunehmen, das meist bereits in der Grundschule erreicht wird.	(nur 1 Aufgabe: einfache räumliche Vorstellung)

<sup>12</sup> Eine Beschreibung der Stufen über Anforderungsmerkmale von Aufgaben dieser Form bedeutet, dass die Mehrzahl der Aufgaben, die zur Beschreibung einer Stufe dienen, einen Schwierigkeitsindex haben, dem ein Wert des Fähigkeitsparameters auf der betreffenden Stufe entspricht.



<b>Stufe II</b> [422;511]	Zu den Anforderungen in Stufe I kommt Wissen über einfache Konventionen, ebenfalls auf niedrigem curricularem Niveau hinzu.	Es sind Standardmathematisierungen vorzunehmen, die bereits auf elementares Wissen aus der Sekundarstufe I zurückgreifen können.	In den Aufgaben sind kleinere nichtrechnerische Schritte durchzuführen, wobei im Text der Aufgabe selbst bereits Hinweise, meist visueller Art, gegeben sind; auch einfache Aufgaben zur räumlichen Vorstellung treten hier auf.
<b>Stufe III</b> [512;603]	Umgehen mit elementaren mathematischen Begriffen und Verfahren der Sekundarstufe I wird verlangt, etwa Basiswissen über Funktionsvorschriften (z.B. Zahlenwerte einsetzen können).	Auf der Basis des Stoffes der Sekundarstufe I werden Standardmodellierungen vorgenommen; insbesondere ist auf dieser Stufe das Umgehen mit linearen Modellen (z.B. Dreisatz, Grundaufgaben der Prozentrechnung) erforderlich.	Die Aufgaben verlangen es, bereits vorgegebene Zusammenhänge zu verstehen und begonnene Strukturen fortzusetzen. Oft wird dies auch noch durch informelle, meist visuelle Unterstützung aus der Situation der Aufgabe heraus erleichtert.
<b>Stufe IV</b> [604;695]	Mathematische Grundtechniken, wie etwa Termumformungen, sind auszuführen.	Noch auf der Basis der Standardkenntnisse aus der Sekundarstufe I sind nun umfangreichere Modellierungen vorzunehmen, die auf <i>mehrschrittige</i> , hier meist <i>repetitive</i> Verarbeitungen <sup>13</sup> führen.	Es sind mathematische Zusammenhänge und Begriffe zu beurteilen und Verknüpfungen vorzunehmen, die für die Sekundarstufe I typisch sind. Insbesondere sind funktionale Zusammenhänge zu erkennen und einzuschätzen, z.B. lineares gegen quadratisches Wachstum abzugrenzen.
<b>Stufe V</b> [696; ...]	Höhere algebraische Techniken, etwa die Lösung quadratischer Gleichungen, sind zu beherrschen.	Auf anspruchsvollem curricularem Niveau sind komplexe Modellierungen, auch solche, die <i>integrative</i> Verarbeitung <sup>14</sup> erfordern, durchzuführen.	Probleme sind selbständig zu strukturieren, eine allgemeine Strategie ist zu entwerfen und durchzuhalten, Lösungen sind argumentativ zu begründen oder es sind Beweise zu führen.

Die Trennung der Kompetenzstufenbeschreibung nach den Typen mathematischen Arbeitens macht auch einsichtig, dass kognitiv so unterschiedliche Aufgaben wie „Quadratische Gleichung“ (siehe III.2) und „31 Pfennig“ (siehe IV.2) gleichermaßen der Stufe V zugerechnet werden müssen. Je nach Typ mathematischen Arbeitens können z. B. der

<sup>13</sup> vgl. Neubrand et al. (2001)

<sup>14</sup> vgl. ebenda

curriculare Ort der Aufgabe oder der Anspruch an den Modellierungsprozess oder verschiedene Aspekte der kognitiven Komplexität (z.B. multiple Lösbarkeit oder die Notwendigkeit, *formale Werkzeuge*<sup>15</sup> zur Präzisierung einzusetzen) die ausschlaggebenden Merkmale für die Schwierigkeit der Aufgabe sein. Dies wurde in Neubrand et al. (2002) auf der Grundlage ausgewählter zentraler Aufgabenmerkmale empirisch untermauert.

In IV.2 wird der Frage der Erfassung des Spektrums mathematischer Fähigkeiten weiter nachgegangen, indem das Verhalten der Schülerinnen und Schüler auf den drei Typen mathematischen Arbeitens untersucht wird.

Bei den nun folgenden Analysen beschränken wir uns auf die deutsche Stichprobe der 15-Jährigen des internationalen Vergleichs (N=5073). Dem Bericht Klieme et al. (2001), S. 169 ist das folgende Globalbild zu entnehmen:

Kompetenzstufe	< I	I	II	III	IV	V
Anteil der 15-Jährigen	7 %	17 %	32 %	31 %	12 %	1 %

**Tabelle 1:** Prozentuale Verteilung von Schülerinnen und Schülern auf die Kompetenzstufen (Werte auf volle Prozent gerundet).

Es ist nicht zufriedenstellend, dass der Anteil der Population unterhalb der Kompetenzstufe I bei immerhin 7 % liegt und größer als der Anteil auf Kompetenzstufe V ist.

Informativ ist eine Aufschlüsselung dieser Verteilung nach Bildungsgängen, nach der man von einem „Grundbildungsgefälle“ sprechen könnte:

Kompetenzstufe	Hauptschule	Integrierte Gesamtschule	Realschule	Gymnasium
V	0,0	0,6	0,5	4,2
IV	0,4	4,1	6,5	31,9
III	6,5	24,2	36,1	48,0
II	37,1	40,7	42,4	14,8
I	38,6	24,6	12,7	1,1
< I	17,4	6,2	2,0	0,0

**Tabelle 2:** Mathematische Kompetenzen nach Bildungsgängen (nach Klieme et al. (2001), S. 181) (prozentuale Anteile an der jeweiligen Teilpopulation)

In Klieme et al. (2001) wird die Stufe III als „Standardstufe“ gekennzeichnet, d.h. als die Leistungsstufe, ab welcher man von einem *ausreichenden* Niveau an mathematischer Grundbildung sprechen sollte. Begründen lässt sich dies einmal über einen Ver-

<sup>15</sup> Zum Begriff formaler Werkzeuge vgl. Cohors-Fresenborg & Sjuts (2001).

gleich der in den Lehrplänen unserer Bildungsgänge genannten Lernziele und dem Anforderungsprofil der Aufgaben, die der Stufe III zugeordnet sind.

Eine zweite Begründung lieferte eine Expertenbefragung in den Bundesländern. Jedes Bundesland hatte pro Schulform einen Experten benannt, der sowohl mit dem Lehrplan als auch mit der Unterrichtspraxis und typischen Anforderungen der jeweiligen Schulform besonders vertraut sein sollte. Sie beurteilten auf vierstufigen Skalen

- a) inwieweit die Schüler mit dem in der Aufgabe angesprochenen Stoff vertraut sein sollten,
- b) inwieweit die Schüler mit der Aufgabenstellung (Verwendung von Formulierungen, Symbolen, der Kontextualisierung) vertraut sein sollten,
- c) welche Bedeutung die mit dieser Aufgabe geprüfte Fähigkeit oder Fertigkeit für den jeweiligen Abschluss des betreffenden Bildungsgangs hat.

Dabei stellte sich heraus, dass sich erst auf Stufe III mehr als 50 % der von den Experten als abschlussrelevant eingestuften Aufgaben lösen lassen.

Normativ - jedoch inhaltlich begründet - die Stufe III als Standardniveau zu definieren, mag als „zu anspruchsvoll“ gelten, weil nur 44 % der deutschen Schülerinnen und Schüler diese Grenze erreichen oder überschreiten. Andererseits gibt es Länder, auch solche, in denen die Mathematikdidaktik kaum andere Wege als bei uns geht, die wesentlich größere Anteile der Schülerschaft auf bzw. über die Stufe III bringen, etwa Österreich zu etwa 52 % und die Schweiz zu etwa 58 % (vgl. die detaillierte Übersicht in Klieme et al. (2001). Diese Vergleiche zeigen Strukturen in den Leistungsverteilungen auf, die möglicherweise auf unterschiedliche didaktische Schwerpunktsetzungen im Unterricht verweisen und zum didaktischen Handeln herausfordern.

Leistungen im Fach Mathematik werden nicht nur durch fachimmanente Faktoren beeinflusst, sondern auch durch fachübergreifende Kompetenzen und Faktoren wie z. B. Einstellungen zum und Interesse am Fach. Außerdem spielen u.a. das *mathematische Selbstkonzept* (d.h. die Selbsteinschätzung mathematischer Kompetenz), motivationale und emotionale Aspekte, Wertschätzung von (schulischer) Bildung, soziokulturelles Umfeld und familiäre Bedingungen eine Rolle. Eine Frage, die auch immer wieder gestellt wird, ist die nach dem Geschlecht als Faktor. Wir wollen dem Einfluss dieser Faktoren nachgehen und beginnen mit einem Vergleich der drei PISA-Domänen *Mathematik*, *Naturwissenschaften* und *Lesen*.

#### **IV.1 Mathematik, Naturwissenschaften und Lesen im Zusammenhang**

Ein wichtiger Befund der PISA-2000-Studie sind die hohen Korrelationen zwischen den Leistungen in den Bereichen *Mathematik*, *Lesen* und *Naturwissenschaften* (vgl. Prenzel et al. (2001), S. 222). Da in Prenzel et al. (2001) nicht der *Mathematikgesamttest* herangezogen wird, haben wir die entsprechenden Korrelationen neu berechnet und geben sie in der folgenden Tabelle an<sup>16</sup>.

---

<sup>16</sup> Es soll angemerkt werden, dass die Korrelationen in dieser und den folgenden Tabellen jeweils mit dem ersten Plausible Value berechnet wurden. (vgl. auch II.1 „Plausible Values“). Die Tests

Dabei bezeichnen wir die betrachteten Tests mit

$M$  (internationaler und nationaler Test zur Messung der „mathematischen Grundbildung“ als Gesamtest),

$NW_i$  (internationaler Test zur Messung der „naturwissenschaftlichen Grundbildung“),

$NW_n$  (nationaler Test zur Messung der „naturwissenschaftlichen Grundbildung“),

$L$  (internationaler Test zur Messung der „Lesekompetenz“).

	$\rho(M,L)$	$\rho(M,NW_i)$	$\rho(M,NW_n)$	$\rho(NW_i,L)$	$\rho(NW_n,L)$
Korrelationskoeffizient	0,83	0,82	0,87	0,86	0,83

**Tabelle 3:** Korrelationen zwischen Fähigkeiten in Mathematik, Naturwissenschaften und Lesen

Die Höhe dieser Werte erweckt den Eindruck, dass die drei Bereiche eng miteinander verwoben sind. Erwartungsgemäß korrelieren die Leistungen in den naturwissenschaftlichen Tests hoch mit dem Mathematiktest. Auch die hohe Korrelation von 0,86 zwischen dem Lesetest und dem internationalen naturwissenschaftlichen Test ist „plausibel“. Einerseits setzt der internationale naturwissenschaftliche Test in einem hohen Maße Lesekompetenz voraus, da alle Aufgaben durch einen längeren Text eingeleitet werden, dessen Verständnis eine wichtige Voraussetzung für das Lösen der Aufgaben darstellt, und die Lösung der Aufgaben dann zum Teil auch noch in Textform gegeben werden muss (vgl. Prenzel et al. (2001)). Auf der anderen Seite stützt sich der Test zur Messung der Lesekompetenz in besonderem Maße auf naturwissenschaftliche Texte.

Betrachtet man den nationalen Ergänzungstest  $NW_n$ , so bleibt auch die Korrelation zwischen diesem Test und dem Lesetest auf vergleichbarem Niveau, obwohl in Prenzel et al. (2001) darauf hingewiesen wird, dass der Textaufwand im nationalen Test gegenüber dem internationalen reduziert wurde.

Das gleiche Bild zeigt ein Vergleich des Mathematiktests mit dem Test auf Lesekompetenz. Die in einem Kontext – ob inner- oder außermathematisch – formulierten Aufgaben setzen natürlich Lesekompetenz im Sinn der Fähigkeit einer aktiven Auseinandersetzung mit dem Text, der Einordnung des Gelesenen in das Wissen des Lesenden (vgl. Artelt et al. (2001b)) voraus. Aber die Texte der Aufgaben haben nur einen geringen Umfang, die Aufgaben zur Klasse „Technische Fertigkeiten“ insbesondere sind weitgehend textfrei formuliert. Dies zeigt, dass die Korrelation zwischen Lesekompetenz und Mathematikleistung wohl auch Ausdruck einer bereichsübergreifenden Kompetenz ist.

Zu einer vergleichbaren Aussage kommt auch der Bericht über die Ergebnisse zu dem Naturwissenschaftlichen Test in PISA (vgl. Prenzel et al. (2001), S. 221).

---

zu Naturwissenschaften wurden wegen der in Prenzel et al. (2001) dokumentierten Unterschiede getrennt belassen. Beim Lesetest gibt es bisher für den nationalen Test keinen Plausible Value für den Gesamtest. Daher wird in der Tabelle nur der internationale Test verwendet.

Es fällt auf, dass die Korrelationen zwischen dem Mathematiktest und dem Lesetest, dem  $NW_n$ -Test und Lesetest sowie zwischen dem Mathematiktest und dem  $NW_i$ -Test etwa gleich groß sind. Auch dieses Ergebnis legt die Vermutung nahe, dass die Korrelationen durch fachübergreifende Faktoren beeinflusst sind.

Einer dieser Faktoren dürfte der Bildungsgang sein, da die Testkorrelationen sinken, wenn man sie innerhalb der einzelnen Bildungsgänge berechnet:

Korr.koeff.	$\rho(M,L)$	$\rho(M,NW_i)$	$\rho(M,NW_n)$	$\rho(NW_i,L)$	$\rho(NW_n,L)$
HS	0,69	0,61	0,75	0,75	0,74
IGS	0,74	0,73	0,83	0,78	0,77
RS	0,67	0,69	0,79	0,75	0,69
Gymnasium	0,60	0,63	0,77	0,70	0,63

**Tabelle 4:** Korrelationen in den Bildungsgängen

Dass an den Korrelationen zwischen jeweils zwei der drei Domänen stets die dritte Domäne beteiligt ist, erkennt man an den *Partialkorrelationen* der drei Domänen. Das sind die Korrelationen je zweier Domänen, die ohne den „linearen“ Einfluss der dritten Domäne vorhanden sind<sup>17</sup>

In der folgenden Tabelle beschränken wir uns auf den internationalen Naturwissenschaftstest und bezeichnen partielle Korrelationskoeffizienten wie im folgendem Beispiel des partiellen Korrelationskoeffizienten zwischen M und L nach Auspartialisierung von  $NW_i$  in der Form  $\rho(M,L|NW_i)$ . Es ergeben sich bei Verwendung der auch in Tabelle 3 verwendeten Plausible Values folgende Werte in der Gesamtpopulation:

	$\rho(M,L NW_i)$	$\rho(M,NW_i L)$	$\rho(NW_i,L M)$
Part. Korrelation	0,42	0,38	0,57

**Tabelle 5:** Partielle Korrelationen.

Die Werte haben sich gegenüber Tabelle 3 deutlich verkleinert.

Berechnet man die partiellen Korrelationskoeffizienten auch wieder in den Bildungsgängen (Tabelle 6), so ergeben sich zwischen Tabelle 6 und Tabelle 4 tendenziell vergleichbare Verhältnisse, wie zwischen den Tabellen 5 und 3:

<sup>17</sup> Zur Definition partieller Korrelationskoeffizienten:

Seien  $X, Y, Z, V$  Zufallsvariable und  $r_{XZV} = a_1 + a_2 z + a_3 v$  und  $r_{YZV} = b_1 + b_2 z + b_3 v$  die Gleichungen der Regressionsgeraden von  $X$  bzw.  $Y$  auf  $Z$  und  $V$ . Dann misst  $X - r_{XZV}$  für ein Wertepaar  $(x_i, z_i, v_i)$  die Differenz zwischen  $x_i$  und dem „linearen Anteil“ des über  $z_i, v_i$  geschätzten  $x$ -Wertes. Analoges gilt für  $Y - r_{YZV}$ . Der „partielle Korrelationskoeffizient zwischen  $X$  und  $Y$  nach Auspartialisierung von  $Z$  und  $V$ “ ist dann definiert als  $\rho(X, Y|Z, V) := \rho(X - r_{XZV}, Y - r_{YZV})$ .

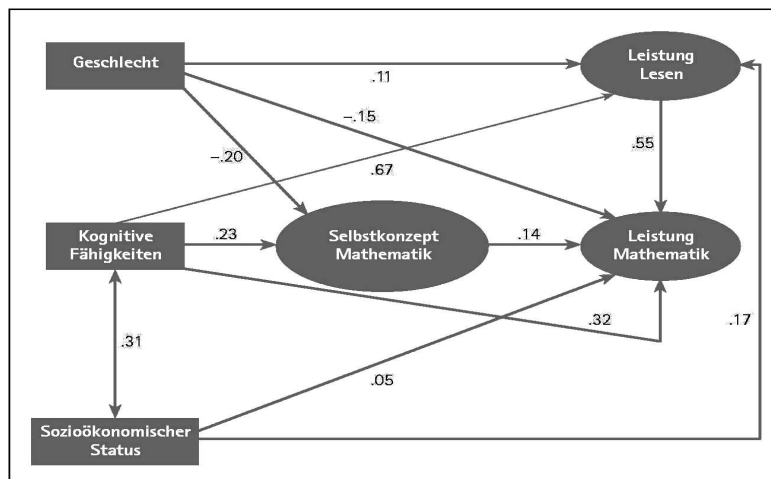
Wird die Regression der Variablen  $X$  und  $Y$  auf nur eine Variable  $Z$  betrachtet, so ist der Koeffizient  $\rho(X, Y|Z)$  analog definiert.

Part. Korrelation	$\rho(M,L   NW_i)$	$\rho(M,NW_i   L)$	$\rho(NW_i, L   M)$
HS	0,37	0,23	0,60
IGS	0,40	0,37	0,52
RS	0,33	0,38	0,53
Gymnasium	0,30	0,36	0,51

**Tabelle 6:** Partielle Korrelationen nach Schulform.

Einige auffallende Details<sup>18</sup> zeigen, dass sich zukünftige Detailanalysen in den Bildungsgängen lohnen dürften .

In Klieme et al. (2001) findet man ein Pfaddiagramm, das den Einfluss fachübergreifender Faktoren aufnimmt. Dieses Pfaddiagramm ist die grafische Darstellung eines „Linearen Strukturgleichungssystems“ (LISREL-System):



**Abbildung 11:** Pfadmodell zur Erklärung der Mathematikleistung (vgl. Klieme et al. (2001), S. 184).

Mit den Variablen  $X_1$ : *Geschlecht*,  $X_2$ : *Kognitive Fähigkeiten*,  $X_3$ : *Sozioökonomischer Status*,  $Y_1$ : *Selbstkonzept Mathematik*,  $Y_2$ : *Leseleistung*,  $Y_3$ : *Leistung Mathematik* zur

<sup>18</sup> Man vergleiche z. B. die partiellen Korrelationskoeffizienten für die Hauptschulpopulation mit den entsprechenden Werten in der Gesamtpopulation.

Kennzeichnung der im Modell auftretenden Faktoren hat dieses Gleichungssystem die Form

$$\begin{aligned} Y_1 &= c_{11} X_1 + c_{12} X_2 + c_{13} X_3 + F_1 \\ Y_2 &= c_{21} X_1 + c_{22} X_2 + c_{23} X_3 + F_2 \\ Y_3 &= a_{31} Y_1 + a_{32} Y_2 + c_{31} X_1 + c_{32} X_2 + c_{33} X_3 + F_3. \end{aligned}$$

Unter der Voraussetzung  $\text{cov}(F_i, F_j) = 0$  und  $\text{cov}(F_i, X_j) = 0$ , also der üblichen Voraussetzung einer Regressionsanalyse, liefert der Datensatz die in Abbildung 11 angegebenen Regressionskoeffizienten, wobei nur die Werte eingetragen sind, die signifikant von Null verschieden sind.

Dieses Modell beschreibt den Einfluss nicht im Fach liegender Faktoren auf die Mathematikleistung, wobei das Modell, die Auswahl der Faktoren und ihr Zusammenspiel über die Regressionsgleichungen *a priori* vorgegeben wurde. Der Datensatz liefert dann „optimale“ Schätzwerte für die Regressionskoeffizienten im Sinn der kleinsten-Quadrat-Schätzung oder Maximum-Likelihood-Schätzung. Interpretationen dieses Diagramms, d.h. die Interpretation der Regressionskoeffizienten, sind nur im Rahmen des vorgegebenen Modells zu sehen, das zum Beispiel den Faktor NW-Test nicht berücksichtigt.

Das Diagramm zeigt, dass im Rahmen dieses Modells die kognitiven Fähigkeiten neben der Leseleistung einen bedeutsamen Einfluss, zum Teil vermittelt über die Leseleistung, auf die Mathematikleistung haben.

Berechnet man zur Kontrolle die Korrelation zwischen den beiden Faktoren „Mathematikleistung“ und „Kognitive Fähigkeiten“, im Folgenden abgekürzt mit KF, über Plausible Values des in PISA verwendeten Begleittests zur Untersuchung kognitiver Fähigkeiten der Probanden, so erhält man mit  $\rho(M, KF) = 0,78$  einen Wert, der nur wenig niedriger ist als die Korrelation zwischen der Mathematikleistung und dem Lesetest.

Die partielle Korrelation zwischen der Mathematikleistung und der Variablen „Kognitive Fähigkeiten“ bei Auspartialisierung von Lesekompetenz und Naturwissenschaftlichem Test beträgt  $\rho(M, KF | L, NW_i) = 0,37$ . Auch dieser Wert ist vergleichbar mit den partiellen Korrelationen zwischen den Tests Mathematik und Naturwissenschaften sowie Mathematik und Lesekompetenz bei Auspartialisierung jeweils eines der beiden verbleibenden Tests (vgl. Tabelle 3). Wir erweitern die Tabelle 5, indem wir auch noch die Variable KF auspartialisieren:

	$\rho(M, L   NW_i, KF)$	$\rho(M, NW_i   L, KF)$	$\rho(NW_i, L   M, KF)$
Part. Korrelation	0,28	0,31	0,53

**Tabelle 7:** Partielle Korrelationen mit KF als zusätzlicher Kontrollvariable.

Vergleicht man die Entwicklungen in den Tabellen 3 bis 7, so wird zunehmend erkennbar, inwieweit die Korrelationen zwischen je zwei Testleistungen durch Hintergrundvariablen beeinflusst sind. Insbesondere scheint die Korrelation zwischen den Variablen M und KF substanzieller zu sein als die zwischen L und KF bzw. NW<sub>i</sub> und KF.

In Artelt et al. (2001a) wird gezeigt, dass das Interesse an Mathematik in Deutschland in allen Bildungsgängen gering ist. Betrachtet man die Populationen in den einzelnen Bildungsgängen, so „bekunden“ allerdings Jungen durchweg ein größeres Interesse an Mathematik als Mädchen. Insgesamt zeigt sich aber kein bedeutender Einfluss des Faktors „Interesse an Mathematik“ auf die Mathematikleistung<sup>19</sup>. Dieses Ergebnis unterscheidet sich von den Ergebnissen im Lesetest, wo sich ein deutlicher Zusammenhang zwischen dem Faktor *Interesse* und dem *Kompetenzerwerb* zeigt. Die Begründung für dieses Ergebnis in Artelt et al. (2001a), S. 284, ist nachvollziehbar: Auch außerhalb des Unterrichts gibt es vielfältige Gelegenheiten zum Lesen, wodurch natürlich Kompetenz erworben wird. Vergleichbare Möglichkeiten findet man in der Mathematik kaum.

Im Pfadmodell wird das „mathematische Selbstkonzept“ als ein die Leistung beeinflussender Faktor deutlich. Das belegen auch die Analysen in Artelt et al. (2001a). Danach liegen deutsche Schülerinnen und Schüler mit einem sehr positiven Selbstkonzept - das sind Probanden, welche die oberen 25 % der Verteilung des Merkmals auf der Mess-Skala bilden – auf der Mathematikfähigkeitsskala bis zu 50 Punkte höher als die Schülerinnen und Schüler, die zum untersten Quartil der Verteilung gehören, also ein schwaches bis negatives Selbstkonzept haben (zu weiteren Einzelheiten vgl. Artelt et al. (2001a)).

Blickt man auf die Bildungsgänge, so zeigen sich wieder keine nennenswerten Unterschiede zwischen den Populationen der einzelnen Bildungsgänge. Die Selbsteinschätzung erfolgt offenbar mit Blick auf die Bezugsgruppe der Mitschüler (vgl. Köller (2001)).

Betrachtet man die Populationen innerhalb der Bildungsgänge, so haben die Jungen jeweils ein deutlich höheres mathematisches Selbstkonzept als Mädchen. Das gilt dann wegen des Fehlens von Unterschieden zwischen den Populationen der Bildungsgänge auch bei Betrachtung der Gesamtpopulation und erklärt das negative Vorzeichen des Pfadkoeffizienten von „Geschlecht“ zu „Selbstkonzept“ in dem Pfaddiagramm.

Auf den Einfluss des Geschlechts als beeinflussenden Faktor der Mathematikleistung gehen wir im Abschnitt IV.3 ein.

## IV.2 Zur Frage schwierigkeitsgenerierender Faktoren

In III hatten wir Anforderungsprofile angesprochen, die unter didaktischem Aspekt der Analyse von schwierigkeitsgenerierenden Faktoren dienen können. Im folgenden wollen wir einigen Fragen nachgehen (weitere Analysen werden an anderer Stelle folgen).

Schwierigkeitsanalysen im Sinne eines Aufspürens von Schwierigkeit generierenden Faktoren sind diffizil. Informationen über Lösungswahrscheinlichkeiten einzelner Aufgaben liefern erste Indikatoren, wenn man Aufgabengruppen betrachtet, in denen bei inhaltlich aufeinander bezogenen Aufgaben unterschiedliches Lösungsverhalten sich aus einer Stufung der Anforderungsmerkmale ergibt. Ein Beispiel liefert die Aufgabengruppe in Abb. 8.

---

<sup>19</sup> In den in Prenzel et al. (2001) vorgetragenen Analysen wird die Mathematikleistung im internationalen PISA-Test verwendet.



Notwendig sind daneben auch Informationen über Lösungsprozesse, Lösungsansätze und auch Fehlschlüsse im Verlauf von Lösungsprozessen. Es ist grundsätzlich nicht unmöglich, solche Informationen teilweise auch aus den Testergebnissen selbst zu gewinnen, z. B. aus den Distraktoren bei MC-Aufgaben oder über geeignete Kodierungen von „offenen“ Aufgaben, bei denen die Rechnung dargestellt bzw. die Antwort begründet werden musste. Die folgenden Beispiele mögen das verdeutlichen.

GLASFABRIK, Version 1 (MC-Aufgabe)

<p>Eine Glasfabrik stellt am Tag 8000 Flaschen her. 2 % der Flaschen haben Fehler. Wie viele sind das?</p> <p><input type="checkbox"/> 16 Flaschen   <input type="checkbox"/> 40 Flaschen   <input type="checkbox"/> 80 Flaschen   <input type="checkbox"/> 160 Flaschen   <input type="checkbox"/> 400 Flaschen</p>
--

Der Leser wird unschwer erkennen, wie aus dem Ankreuzen falscher Distraktoren auf das dahinter stehende Vorgehen des Probanden bei der Bearbeitung der Aufgabe zurück geschlossen werden kann (vgl. auch die Beispiele in Abschnitt II.1).

Bei MC-Aufgaben ist bei Ankreuzen der richtigen Alternative natürlich der Lösungsweg, den der Proband beschritten hat, nicht erkennbar. Das ist aber bei der Suche nach Schwierigkeiten generierenden Faktoren auch nicht das Thema.

SPAREN (offene Aufgabe und zugehörige Kodierungsanweisungen)

<p>Karina hat 1.000 DM in ihrem Ferienjob verdient. Ihre Mutter empfiehlt ihr, das Geld zunächst bei einer Bank für 2 Jahre festzulegen (Zinseszins!). Dafür hat sie zwei Angebote:</p> <p>a) "Plus-Sparen": Im ersten Jahr 3 % Zinsen, im zweiten Jahr 5 % Zinsen.</p> <p>b) "Extra-Sparen": Im ersten und zweiten Jahr jeweils 4 % Zinsen.</p> <p>Karina meint: "Beide Angebote sind gleich gut." Was meinst du dazu? Begründe deine Antwort!</p>	
Code	Antwort
01	"Extra-Sparen ist besser" ohne oder mit falscher Begründung (ein Beispiel für eine falsche Begründung wäre etwa: in beiden Fällen sind es insgesamt 8 %, bei "Plus-Sparen" aber zuerst 3 %, das kann durch 5 % im 2. Jahr nicht mehr aufgeholt werden).

02	<p>Falsche Antwort:</p> <ul style="list-style-type: none"> <li>• "Plus-Sparen" ist besser mit oder ohne Begründung</li> <li>• "Karina hat Recht" / "beide Angebote sind gleich gut" ohne Begründung.</li> </ul>
11	<p>Richtige Antwort: Angebot 2 ("Extra-Sparen") ist besser. Es wird der Kontostand konkret schrittweise ausgerechnet:</p> <p>bei "Plus-Sparen": 3 % von 1000DM ... 5 % von 1030 DM, also 1081,50 DM</p> <p>bei "Extra-Sparen": 4 % ... , also 1081,60 DM</p>
12	<p>Richtige Antwort: Angebot 2 ("Extra-Sparen") ist besser. Es wird mit Zinsoperatoren gerechnet, aber nach wie vor auf das konkrete Kapital:</p> <ul style="list-style-type: none"> <li>• bei "Plus-Sparen": <math>1000 \text{ DM} * 1,03 * 1,05 = 1081,50 \text{ DM}</math></li> <li>• bei "Extra-Sparen": <math>1000 \text{ DM} * 1,04 * 1,04 = 1081,60 \text{ DM}</math></li> </ul>
13	<p>Richtige Antwort: Angebot 2 ("Extra-Sparen") ist besser. Es wird nur mit den Zinsoperatoren, ohne Bezug zum konkreten Fall gerechnet.</p> <ul style="list-style-type: none"> <li>• <math>1,03 * 1,05 = 1,0815 / 1,04 = 1,0816</math></li> </ul>
14	<p>Richtige Überlegung und Berechnung, jedoch als Antwort: "gleich gut", mit der Begründung, dass 10 Pfennig keinen Unterschied machen (das wäre auch eine sehr vernünftige Antwort!).</p>

Ein Ergebnis der Studie TIMSS II war, dass deutsche Schülerinnen und Schüler bei technischen Aufgaben relativ gut abschneiden, aber Schwächen bei der Modellierung anspruchsvoller Kontexte zeigen (vgl. z. B. Hrsg. Blum & Neubrand (1998)). Folgerungen aus den Ergebnissen waren Forderungen nach einem Unterricht

- der in größerem Maße Kreativität und inhaltlich nicht standardisiertes Argumentieren fördert und der weniger auf das Abarbeiten von Routinen und Kalkül zielt,
- der sich um ein „verständnisvolles Lernen“, um ein Verstehen von Konzepten und strukturellen Zusammenhängen bemüht, um eine Vernetzung von Stoffen und Begriffen und ihre Entwicklung auch in außermathematischen Kontexten.

Vergleicht man die Analysen von PISA mit denen von TIMSS II, so werden die Ergebnisse aus TIMSS II durch PISA weitgehend bestätigt.<sup>20</sup> Man betrachte z. B. noch einmal die Aufgabensequenz in Abb. 8:

„Glasfabrik, Version 1“ (Lösungshäufigkeit in Deutschland 69 %) fragt nach dem Prozentwert, also Standardwissen der Sekundarstufe I.

„Miete“ (Lösungshäufigkeit in Deutschland 20 %) fordert das gleiche Faktenwissen, nur muss es jetzt mehrfach angewendet werden.

Die Lösung der Aufgabe „Sparen“ (Lösungshäufigkeit in Deutschland 19 %) beruht auf dem gleichen Faktenwissen wie die vorausgegangenen Aufgaben, der Berechnung von Prozentwerten, fordert nun aber komplexere Überlegungen. Zur Lösung ist die Verknüpfung mehrerer Berechnungen notwendig<sup>21</sup> und zusätzlich ist ein Vergleich der beiden Resultate, also ein eigenständiges Argumentieren notwendig.

Ein weiteres Beispiel liefert die folgende Aufgabe:

FAHRRADUNFÄLLE (PISA-Index von (1): 497, PISA-Index von (2): 674)

Eine Zeitung meldet:

**70 % aller mit dem Fahrrad verunglückten Kinder sind Jungen. Jungen auf dem Rad sind also stärker gefährdet als Mädchen.**

Die Zeitungsmeldung beruht auf folgender Tabelle, in der die 10 000 Schülerinnen und Schüler einer Region, die mit dem Fahrrad zur Schule fahren, nach Geschlecht und Unfallbeteiligung aufgeführt sind.

	Verunglückt	Nicht verunglückt	Insgesamt Jungen/Mädchen
Jungen	70	8 400	8 470
Mädchen	30	1 500	1 530
Kinder insgesamt	100	9 900	10 000

Beurteile die Zeitungsmeldung mittels der Tabelle:

<sup>20</sup> Die „Kalkülorientierung“ des Unterrichts ist allerdings nebenbei bemerkt kein charakteristisch deutsches Phänomen. Im internationalen PISA-Rahmenkonzept (vgl. OECD (1999)) liest man „However, school mathematics is often offered to students as a strictly compartmentalized science, and overemphasizes computation and formulae“.

<sup>21</sup> Eine ähnliche Aufgabe wird in Cohors-Fresenborg & Sjuts (2001) unter dem Aspekt *kognitive Komplexität* analysiert.

- (1) Die Zeitungsmeldung, dass 70 % aller mit dem Fahrrad verunglückten Kinder Jungen sind, ist
- richtig,       falsch,       nicht anhand der Tabelle zu beantworten.
- (2) Begründe: Die Zeitungsmeldung, dass Jungen stärker gefährdet als Mädchen sind, ist
- richtig, weil ...                       falsch, weil ...

Teil (1) dieser Aufgabe gehörte zur nationalen Kompetenzklasse 2A und wurde von rund 64 % der Probanden gelöst. Teil (2) gehörte zur Klasse 2B und wurde nur noch von 21 % der Schüler gelöst. Die Aufgabe ist gut konstruiert, da sie in Teil (1) nur die Kenntnis der Prozentbegriffsdefinition und in Teil (2) argumentatives Umgehen mit Anteilen prüfen sollte. Teil (1) hatte einen schlechten Itemfit, da hier die richtige Lösung für Unkundige zu attraktiv war. Der Itemfit von Teil (2) war jedoch sehr gut. Es gab daher keinen Grund, diese Aufgabe aus dem Test zu nehmen.

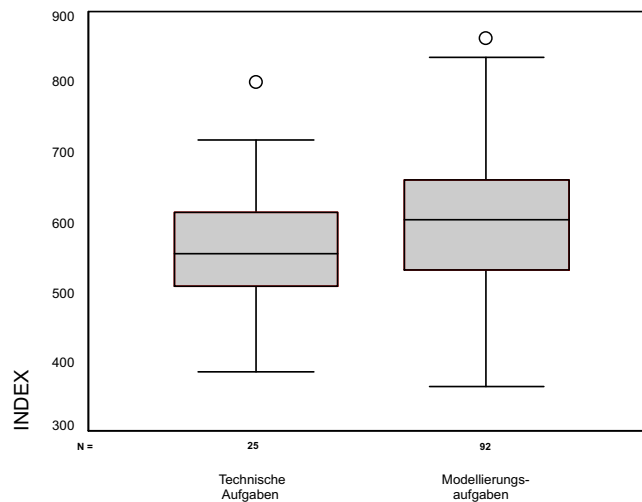
Ein besonders aufschlussreiches Beispiel ist die Aufgabe „31 Pfennig“, die der höchsten Kompetenzstufe angehört. Diese Aufgabe ist allein durch Strukturierung des Problems zu lösen. Dabei geht es um die Entwicklung einer Strategie, während der rechnerische Anspruch offensichtlich sehr gering ist. Da das Denken bei der Bearbeitung genau organisiert und überwacht werden muss, sind auch in hohem Masse *metakognitive Kompetenzen* (vgl. dazu Cohors-Fresenborg & Sjuts (2001)) erforderlich.

#### 31 PFENNIG (PISA-Index 797)

Wie kannst du einen Geldbetrag von genau 31 Pfennig hinlegen, wenn du nur  
10-Pfennig-, 5-Pfennig- und 2-Pfennig-Münzen  
zur Verfügung hast? Gib **alle** Möglichkeiten an.

18 % der deutschen Probanden finden vier oder fünf Möglichkeiten. Nur 3 % finden alle sechs Möglichkeiten.

Betrachtet man die Verteilung der Schwierigkeitsparameter der Aufgaben im Gesamttest, so erkennt man an den beiden folgenden Abbildungen, dass die Aufgaben der Klasse „Technische Aufgaben“ (1A) nicht nur im Durchschnitt leichter scheinen als die „Modellierungsaufgaben“ (2A, 1B, 2B und 3), sondern auch eine kleinere Schwierigkeitsstreuung (im Sinne der Schwierigkeitsspannweite) aufweisen:



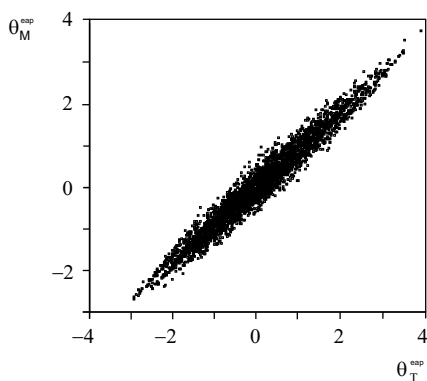
**Abbildung 12:** Vergleich technischer Aufgaben mit den übrigen Aufgaben .

Das darf jedoch nicht zu dem Fehlschluss verleiten, dass diese beiden Aufgabengruppen verschiedene Dimensionen im Rahmen der Rasch-Modellierung definieren.

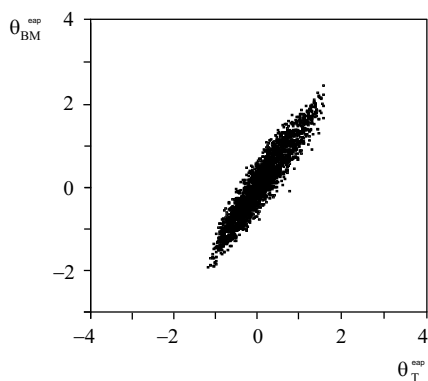
Ein versuchsweise zweidimensional gerechnetes Modell (1. Dimension: „Technische Aufgaben“, 2. Dimension „Modellierungsaufgaben“) ohne Einbeziehung von Begleitvariablen liefert fast die gleichen Aufgabenschwierigkeitsschätzungen wie das eindimensionale Modell und zeigt eine sehr hohe Korrelation zwischen dem Lösungsverhalten der Probanden auf der Teilgruppe „Technische Aufgaben“ und ihrem Verhalten auf der Teilgruppe „Modellierungsaufgaben“ (vgl. Abb. 13). Die Korrelation zwischen dem eap-Schätzer  $\theta_T^{\text{eap}}$  auf der ersten und dem eap-Schätzer  $\theta_M^{\text{eap}}$  auf der zweiten Dimension beträgt  $\rho = 0,98$ .

Zwar fordern die mit dem Begriff „Modellierungsaufgaben“ gekennzeichneten Aufgaben gegenüber den Aufgaben, zu deren Lösung nur technische Fertigkeiten oder Faktenwissen erforderlich ist, zusätzliche Kompetenzen, was sie etwas (de facto minimal) schwieriger macht. Aber wer bei diesen Aufgaben durchschnittlich schwache Leistungen zeigt, zeigt auch Schwächen im Kalkül und umgekehrt.

Dieses Bild ändert sich, wenn man nur die Teilgruppen „Technische Aufgaben“ (1A) und „Begriffliche Modellierungsaufgaben“ (Klassen 2A und 3) in ein zweidimensionales Modell aufnimmt und den so definierten Teilstest betrachtet. Da die Entwicklung der Fähigkeit zum begrifflichen Arbeiten auch an den Erwerb von Wissen über Techniken gebunden ist, sind der eap-Schätzer  $\theta_T^{\text{eap}}$  auf der ersten und der eap-Schätzer  $\theta_{BM}^{\text{eap}}$  auf der zweiten Dimension mit  $\rho = 0,95$  hochkorreliert. Die Streuungen der Leistungsschätzwerte auf den beiden Dimensionen sind jedoch nun deutlich verschieden, wie Abbildung 14 zeigt:



**Abbildung 13:** Die eap-Schätzungen der Fähigkeiten im Gesamttest bei einer zweidimensionalen Modellierung (*technische Fertigkeiten* gegen übrige Aufgaben).



**Abbildung 14:** Die eap-Schätzungen der Fähigkeiten bei einem Teilttest, in dem nur *technische Fertigkeiten* und *begriffliches Arbeiten* geprüft werden.

Dieser Befund deutet darauf hin, dass die technischen Aufgaben und die begrifflichen Aufgaben auch unterschiedliche Fähigkeiten ansprechen.

Ein solcher Unterschied wird auch in der Studie Neubrand et al. (2002) sichtbar, in der mit Hilfe von Regressionsanalysen dem Einfluss einzelner Faktoren auf den Schwierigkeitsindex einer Aufgabe als abhängige Variable nachgegangen wird. Betrachtet werden dabei unter anderem die Faktoren

- „Komplexität der Modellierung“, mit den internationalen Kompetenzklassen K1, K2 und K3 als Kategorien<sup>22</sup>
- „Curriculare Wissensstufe“ mit den Kategorien: *Grundkenntnisse; Einfaches Wissen der Sek.I; Anspruchsvolles Wissen der Sek.I.*

Ein Ergebnis dieser Studie ist, dass der Schwierigkeitsgrad einer Aufgabe der Kategorie „Technische Aufgaben“ im wesentlichen durch den Faktor „Curriculare Wissensstufe“ bestimmt wird. Das war zu erwarten. Aber auch innerhalb der Kategorie „Rechnerische Modellierungsaufgaben“ dominiert dieser Faktor, wobei jetzt zusätzlich der Einfluss des Faktors „Komplexität der Modellierung“ und der Umfang der Bearbeitung der Aufgabe eine Rolle spielt. In der Kategorie „Begriffliche Modellierungsaufgaben“ verliert der Faktor „Wissensstufe“ seinen Einfluss. Dafür gewinnt der Faktor „Kontext“ (kategorisiert durch: *außermathematisch – innermathematisch - ohne Kontext*) an Bedeutung.

Betrachtet man in Tabelle 2 die Verteilung der Probanden auf die Kompetenzstufen nach Schulformen, so wird deutlich, dass man der Frage nach der Bedeutung einzelner Anforderungsmerkmale als Schwierigkeit generierender Faktor wohl auch in den ein-

<sup>22</sup> d.h. die Aufgaben wurden nach den Kompetenzklassen, denen sie zugeordnet waren, kodiert

zelen Bildungsgängen nachgehen muss. Dies kann hier nicht geleistet werden und muss späteren eigenständigen Detailanalysen vorbehalten bleiben.

### IV.3 Mathematik und Geschlecht

Unterschiede zwischen Jungen und Mädchen in den verbalen und mathematischen Fähigkeiten werden in der Regel als gesichert angesehen. Mädchen übertreffen die Jungen in Lesekompetenz. In Mathematik und den Naturwissenschaften, abgesehen vielleicht von Biologie, ist es anders herum. TIMSS II lieferte dann aber das Ergebnis, dass sich in Deutschland für das Fach Mathematik Leistungsunterschiede zwischen Jungen und Mädchen nicht mehr nachweisen lassen, wenn man die Gesamtpopulation betrachtet<sup>23</sup>.

Damit hätte man zufrieden sein können. Betrachtet man aber die einzelnen Bildungsgänge, so sieht das Ergebnis anders aus. Die folgende Tabelle gibt die Differenzen der Mittelwerte der erreichten Punktzahlen in den Bildungsgängen in TIMSS II an. Aufgelistet sind die Leistungsvorsprünge der Jungen.<sup>24</sup>

	Hauptschule	Realschule	Integrierte GS	Gymnasium
Leistungsvorsprung Jungen	25	18	12	10

**Tabelle 8:** Leistungsunterschiede zwischen Jungen und Mädchen nach Schulform (Differenz der Mittelw., nach Abb. D22 in Baumert et al. (1997), S. 156).

Auf der Ebene der Schulformen findet man also konsistent Leistungsunterschiede zwischen Jungen und Mädchen, die in die erwartete Richtung zeigen.

Dass sich dieser Effekt in der Gesamtpopulation nicht bemerkbar macht, ist ein Beispiel für das als *Simpsonsches Paradoxon* in der Wahrscheinlichkeitstheorie und Statistik bekannte Phänomen. Wie die folgende Tabelle zeigt, waren in TIMSS II die Schülerinnen in den „leistungsstärkeren“ Bildungsgängen stärker vertreten als Schüler, während die Situation in den „leistungsschwächeren“ Schulformen umgekehrt war.

Geschlecht	Hauptschule	Realschule	Integrierte GS	Gymnasium
Mädchen	21	29	9	41
Jungen	33	27	10	30

**Tabelle 9:** Schüler der 8. Jahrgangsstufe nach Schulform und Geschlecht in Prozent der Schüler eines Geschlechts (nach Tab. D4 in Baumert et al. (1997), S. 153).

<sup>23</sup> vgl. dazu Hosenfeld et al. (1999)

<sup>24</sup> **Zur Erinnerung:** Wie in PISA war in TIMSS die Leistungsskala auf einen internationalen Mittelwert von 500 und eine Standardabweichung von 100 normiert. Auf dieser Skala betrug der Mittelwert der deutschen Teilnehmer für die Mathematikleistung 509 bei einem Standardfehler von 4,5. Die Mittelwerte für die verschiedenen Schulformen waren: Hauptschule 446; Realschule 504; Gymnasium 573.

Wie sieht das bei PISA aus? Im Gegensatz zu den Ergebnissen von TIMSS II zeigen sich nun wieder Unterschiede zugunsten der Jungen, auch in der Gesamtpopulation. In Stanat & Kunter (2001) findet man die Daten für die Ergebnisse des internationalen Tests, angegeben auf der nationalen Skala, die auf den Mittelwert 100 und die Standardabweichung 30 normiert ist.

Im Globalvergleich, d.h. bei Betrachtung der Gesamtpopulation, wird für den internationalen Test auf dieser Skala ein Punktvorsprung von 4 Punkten zugunsten der Jungen angegeben.

Bei Betrachtung der einzelnen Bildungsgänge ergeben sich wie bei TIMSS II höhere Differenzen:

	Hauptschule	Realschule	Int. GS	Gymnasium
Leistungsvorsprung Jungen	10 (33)	8 (27)	10 (33)	9 (30)

**Tabelle 10:** Leistungsunterschiede zwischen Jungen und Mädchen im internationalen Test (Differenz der Mittelwerte auf der nationalen Skala nach Abb. 5.3 in Stanat & Kunter (2001), S. 259; Differenzen auf der internationalen Skala in Klammern)

Der Grund dafür, dass bei PISA der Leistungsunterschied zwischen den Geschlechtern in der Gesamtpopulation relativ gering ist, liegt wie bei TIMSS II an der Verteilung der Geschlechter in den Schulformen:

Geschlecht	HS	RS	Int. GS	Gymnasium	Sonstige
Mädchen	19	27	9	33	12
Jungen	20	25	9	24	22

**Tabelle 11:** 15-jährige Schüler nach Schulform und Geschlecht in Prozent der Schüler eines Geschlechts

Um den Einfluss des nationalen Tests zu prüfen, haben wir für den nationalen Test den mittleren Leistungsunterschied auf der *nationalen Skala* bestimmt. Es ergab sich eine Differenz der Mittelwerte der erreichten Leistungswerte von 5 zugunsten der Jungen.

Anschließend wurde die Differenz der Mittelwerte der erreichten Leistungswerte in den einzelnen Bildungsgängen bestimmt:

	Hauptschule	Realschule	Integrierte GS	Gymnasium
Leistungsvorsprung Jungen	9	9	6	10

**Tabelle 12:** Mittlere Leistungsunterschiede von Jungen und Mädchen nach Schulform im nationalen Test (nationale Skala).



Die Werte in Tabelle 12 haben sich gegenüber den Werten in Tabelle 10 verändert, ein weiterer Hinweis darauf, dass es sinnvoll ist, die Daten der Studie PISA-2000 auch getrennt nach Bildungsgängen zu analysieren.

Um mittlere Leistungsdifferenzen beurteilen zu können, muss man sie in Relation zur Leistungsbandbreite sehen, die wir hier nur in Form von Leistungsmittelwerten in den Schulformen dokumentieren:

	Hauptschule	Realschule	Int. GS	Gymnasium
Mittl. Leistung Jungen	78	103	88	133
Mittl. Leistung Mädchen	69	94	82	123

**Tabelle 13:** Mittlere Leistungswerte von Jungen und Mädchen im nationalen Test nach Bildungsgängen (nationale Skala).

Die voraufgegangenen Tabellen geben nur einen ersten Überblick. Es drängen sich sofort Fragen nach spezifischen Stärken und Schwächen von Jungen und Mädchen auf. Lassen sich z. B. Aufgabentypen, Kategorien, Stoffgebiete identifizieren, in denen sich Positiva oder Defizite der einen oder anderen Gruppe zeigen?

Mit den vorgestellten Analysen ist die Frage nach den Ursachen für die Ergebnisse noch nicht beantwortet. In die Beantwortung dieser Frage gehen wesentlich Hintergrundvariablen ein, angefangen vom schon erwähnten Interesse über das Selbstkonzept bis hin zu soziokulturellen Daten der Familie, deren Einstellung zu Mathematik sicher einen bedeutenden Einfluss auf die gerade genannten Variablen und damit auch auf Faktoren wie Motivation, Leistungsbereitschaft, u.s.w. hat. Die hier geforderte Ursachenforschung kann im Rahmen dieser Arbeit nicht geleistet werden. Sie kann nur in einer eigenständigen Detailanalyse erfolgen, in die wesentlich die begleitenden Fragebögen der Leistungsstudie eingehen müssen.

Das Gleiche gilt für die Untersuchung fachspezifischer Faktoren, z. B. der Frage ob sich Jungen und Mädchen unterschiedlich in ihrem Leistungsverhalten in Abhängigkeit vom Stoffgebiet verhalten oder bezüglich anderer oben genannter Anforderungsmerkmale der Aufgaben.

## Schlussbemerkung

Wir haben in den beiden ersten Abschnitten versucht, in knapper Form die Intentionen und Strukturen des internationalen Tests und des nationalen Ergänzungstests darzustellen, sowie das Untersuchungsdesign, das der Auswertung der Tests zugrunde liegt. Auf der Grundlage dieser Darstellung haben wir dann Analysen vorgetragen, wobei es uns wesentlich darauf ankam, dass der Leser die Argumentationen mit den bereitgestellten Informationen nachvollziehen kann. Diese Analysen dienten zwei Zielen:

Zunächst sollten einige Ergebnisse der nationalen und internationalen Studie PISA-2000 gegeben werden. Darüber hinaus sollte diskutiert werden, mit welchen Methoden Informationen aus den Daten von PISA-2000 gewonnen werden können.

Die vorgestellten Analysen behandeln die gestellten Fragen sicher nicht abschließend. Ganze Bereiche, wie z. B. der motivationale Bereich, haben hier keine Aufnahme gefunden, wengleich zu erwarten steht, dass hier Faktoren zu finden sind, die das Leistungsverhalten beeinflussen. Diesbezügliche Detailuntersuchungen konnten im Rahmen dieser Arbeit nicht zur Sprache kommen und müssen Folgearbeiten vorbehalten bleiben.

Wir möchten abschließend Johanna Neubrand und Rudolf vom Hofe aus der Experten-  
gruppe PISA-2003, Jürgen Baumert und dem JMD-Gutachter Rolf Biehler für wertvolle  
Hinweise und Anregungen danken.

Unser Dank gilt auch Oliver Lüdtke vom Max Planck Institut für Bildungsforschung in  
Berlin für die Einführung in die Schätzsoftware Conquest.

## Literatur

- Andersen, E. B. (1991): *The Statistical Analysis of Categorical Data*. Springer, Berlin Heidelberg New York Tokyo 1991.
- Artelt, C., Demmrich, A., Baumert, J. (2001a): Selbstreguliertes Lernen. In Baumert, J. et al. (2001a), 271-298
- Artelt, C., Stanat, P., Schneider, W., Schiefele, U. (2001b): Lesekompetenz: Testkonzeption und Ergebnisse. In Baumert, J. et al. (2001a), 69-137
- Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., Weiß, M. (Deutsches PISA-Konsortium) (2001a): *PISA 2000 - Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Leske + Budrich, Opladen 2001.
- Baumert, J., Lehmann, R. H., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller, O. & Neubrand, J. (1997): *TIMSS – Mathematisch - naturwissenschaftlicher – Unterricht im Vergleich*. Leske + Budrich, Opladen 1997.
- Baumert, J., Stanat, P., Demmrich, A. (2001b): PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In Baumert, J. et al. (2001a), 15-68
- Beaton, A.E., Allen, N.L. (1992): Interpreting scales through scale anchoring. *Journal of Educational Statistics*, 17(2), 191-204.
- Blum, W. (1996): Anwendungsbezüge im Mathematikunterricht - Trends und Perspektiven. In: *Beiträge zum 7. Internationalen Symposium zur Didaktik der Mathematik in Klagenfurt, Sept. 1994*. Hölder-Pichler-Tempsky. Wien 1996, 15-38
- Blum, W. & Neubrand, M. (Hrsg.) (1998): *TIMSS und der Mathematikunterricht: Informationen, Analysen, Konsequenzen*. Schroedel, Hannover 1998.
- Cohors-Fresenborg, E. (1996): Mathematik als Werkzeug zur Wissensrepräsentation. In: *Beiträge zum 7. Internationalen Symposium zur Didaktik der Mathematik in Klagenfurt, Sept. 1994*. Hölder-Pichler-Tempsky. Wien 1996, 85-90

- Cohors-Fresenborg, E., Sjuts, J. (2001): Die Berücksichtigung von kognitiver und metakognitiver Dimension bei zu erbringenden und zu beurteilenden Leistungen im Mathematikunterricht. In: Solzbacher C. & Freitag, C. (Hrsg.): *Anpassen, Verändern, Abschaffen? Schulische Leistungsbewertung in der Diskussion*. Klinkhardt, Bad Heilbrunn 2001, 147-162
- Fischer, G. H. & Molenaar, I. W. (Hrsg.) (1995): *Rasch Models. Foundation, recent developments, and applications*. Springer, New York 1995.
- vom Hofe, R. (1995): *Grundvorstellungen mathematischer Inhalte*. Spektrum, Heidelberg 1995.
- Hosenfeld, I., Köller, O. & Baumert, J. (1999): Why sex differences in mathematics achievement disappear in German secondary schools: A reanalysis of then German TIMSS-data. *Studies in Educational Evaluation*, 25, 143-162
- Kaiser, G. (1995): Realitätsbezüge im Mathematikunterricht – Ein Überblick über die aktuelle und historische Diskussion. In: *Materialien für einen realitätsbezogenen Mathematikunterricht, Bd. 2* (Hrsg. Graumann et al.). Franzbecker Hildesheim 1995, 66-84
- Klieme, E. (1989): *Mathematisches Problemlösen als Testleistung*. Lang, Frankfurt 1989.
- Klieme, E. (2000): Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte. In: Baumert, J., Bos, W., Lehmann, R. H (Hrsg.): *TIMSS/III - Dritte Internationale Mathematik- und Naturwissenschaftsstudie, Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn, Bd. 2*, 57-128. Leske + Budrich, Opladen 2000.
- Klieme, E., Neubrand, M., Lüdtke, O. (2001): Mathematische Grundbildung: Testkonzeptionen und Ergebnisse. In Baumert, J. et al. (2001a), 141-190.
- Knoche, N., Lind, D. (2000): Eine Analyse der Aussagen und Interpretationen von TIMSS unter Betonung methodologischer Aspekte. *Journal für Mathematikdidaktik*, 21 (1), 3-27.
- Köller, O. (2001): *Leistungsgruppierungen, soziale Vergleiche und selbstbezogene Fähigkeitskognitionen in der Schule*. Habilitationsschrift, Universität Potsdam.
- Lord, F. M. & Novick, M. R. (1968): *Statistical theories of mental test scores*. Addison-Wesley, Reading MA 1968.
- Mislevy, R.J., Beaton, A.E., Kaplan, B., Sheehan, K.M. (1992a): Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133-161.
- Mislevy, R.J., Johnson, E.G., Musaki, E. (1992b): Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2), 131-154.
- Neubrand, J. (2002): *Eine Klassifikation mathematischer Aufgaben zur Analyse von Unterrichtssituationen: Selbsttätiges Arbeiten in Schülerarbeitsphasen in den Stunden der TIMS-Videostudie*. Franzbecker, Hildesheim 2002.
- Neubrand, M. (2001): PISA – „Mathematische Grundbildung“ / „mathematical literacy“ als Kern einer internationalen und nationalen Leistungsstudie. In: Kaiser, G., Knoche, N., Lind, D., Zillmer, W.: *Leistungsvergleiche im Mathematikunterricht – ein Überblick über aktuelle nationale Studien*. Franzbecker, Hildesheim 2001.

- Neubrand, M., Biehler, R., Blum, W., Cohors-Fresenborg, E., Flade, L., Knoche, N., Lind, D., Löding, W. Möller, G., Wynands, A. (2001): Grundlagen der Ergänzung des internationalen PISA-Mathematiktests in der deutschen Zusatzerhebung. *Zentralblatt für Didaktik der Mathematik*, Vol.33 (2), 45-49.
- Neubrand, M., Klieme, E., Lüdtke, O., Neubrand, J., (2002): Kompetenzstufen und Schwierigkeitsmodelle für den Pisa-Test zur mathematischen Grundbildung. *Unterrichtswissenschaft*, 30 (1), 100-119 .
- OECD (Ed.) (1999): *Measuring student knowledge and skills. A new framework for assessment*. Paris: OECD Publication Service. [Deutsches PISA-Konsortium (Hrsg.) (2000): *Schülerleistungen im internationalen Vergleich: Eine neue Rahmenkonzeption für die Erfassung von Wissen und Fähigkeiten*. Berlin: MPI für Bildungsforschung. ]
- OECD (Hrsg.) (2001): *Lernen für das Leben: Erste Ergebnisse der internationalen Schulleistungsstudie PISA-2000*. Paris: OECD Publication Service.
- Prenzel, M., Rost, M., Senkbeil, M. Häußler, P. & Klopp, A. (2001) : Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse. In Baumert, J. et al. (2001a), 191-248.
- Rasch, G. (1960): *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press, Chicago 1980 (Nachdruck der Veröffentlichung aus dem Jahr 1960).
- Rost, J. (1996): *Lehrbuch Testtheorie, Testkonstruktion*. Huber, Bern 1996.
- Schupp, H. (1988): Anwendungsorientierter Mathematikunterricht in der Sekundarstufe I zwischen Tradition und neuen Impulsen. In: *Der Mathematikunterricht*, 34(6), 5-16
- Stanat, P., Kunter, M. (2001): Geschlechterunterschiede in Basiskompetenzen. In Baumert et al. (2001) Baumert, J. et al. (2001a), 249-269
- Winter, H. (1995): Mathematikunterricht und Allgemeinbildung. In: *Mitteilungen der Gesellschaft für Didaktik der Mathematik*, Nr. 61, 37-46.
- Wu, M.L. (1998): *ACER ConQuest: generalised item response modelling software manual*. 1998, ACER, Melbourne.

Für die PISA-2000 Expertengruppe:

Norbert Knoche, Universität Essen, Fachbereich Mathematik und Informatik,  
Didaktik der Mathematik, Universitätsstraße 3, 45141 Essen  
norbert.knoche@uni-essen.de

Detlef Lind, Bergische Universität Wuppertal, Fachbereich Mathematik, Didaktik der  
Mathematik, Gaußstraße 20, 42097 Wuppertal  
lind@uni-wuppertal.de