Leistungen von Jungen und Mädchen bei PISA 2003 – bedeutsame Unterschiede?

Dissertation

zur Erlangung des Doktorgrades der Philosophischen Fakultät der Christian-Albrechts-Universität zu Kiel

vorgelegt von Désirée Burba
Kiel
(2006)

Erstgutachter: Prof. Dr. Jürgen Rost
Zweitgutachter: Prof. Dr. Thomas Bliesener
Tag der mündlichen Prüfung: 12. 07. 2006
Durch den zweiten Prodekan Prof. Dr. Norbert Nübler
zum Druck genehmigt am: 12. 07. 2006

Die Menschen werden wahre Gleichheit mit offenen Armen empfangen.

JC Denton, Deus Ex – Invisible Wars

Zusammenfassung

Geschlechterunterschiede stellen ein zentral und vielfältig diskutiertes Thema in der Psychologie dar. Speziell in der pädagogischen Psychologie ergeben sich wichtige Implikationen für Schule und Familie. Im Rahmen der large scale assessment - Studien liefert PISA Kernaussagen über die Leistung von 15-jährigen.

Die vorliegende Arbeit nutzt die Daten aus der internationalen Schülerstichprobe, um die Geschlechterunterschiede bei PISA 2003 genauer zu analysieren. Zunächst wurde geprüft, ob die Interessen und Fähigkeiten von deutschen Mädchen und Jungen bei PISA 2003 in stereotyper Weise variieren. Mittels einer geschlechtsspezifischen Skalierung mit dem Rasch-Modell wurden die Itemschwierigkeiten von Mädchen und Jungen für jedes Item in Mathematik, Lesen, Naturwissenschaften und Problemlösen verglichen. Daran wurde eine inhaltliche Analyse derjenigen Items angeschlossen, die einen wesentlichen Geschlechterunterschied aufwiesen, wobei 45 von 360 Items geringere Schwierigkeitsparameter für ein Geschlecht zeigten.

Mit Blick auf Inhalte und Anforderungen bestätigen sich für eine Vielzahl von Items stereotype Befunde. Die Ergebnisse wurden in sechs weiteren ausgewählten PISA-Teilnehmerstaaten (Österreich, Finnland, Island, Niederlande, USA, Japan) untersucht und mit den deutschen Befunden verglichen. Die Anzahl der in mindestens vier der sieben Staaten auffälligen Items reduzierte sich auf insgesamt neun, dabei zeigte sich eines über alle sieben Länder konsistent von Jungen leichter zu lösen. Unterschiede konnten über geschlechtstypische Anforderungen und Inhalte erklärt werden.

Um die Rolle der Geschlechtsunterschiede in den kognitiven Kompetenzen im nationalen Naturwissenschaftstest zu klären, wurden diese jeweils geschlechtsspezifisch zu einer Skala zusammengefasst. Das Ziel bestand darin, jedem Schüler auf einer Skala aus einer Differenz zwischen „männlichen“ und „weiblichen“ Teilkompetenzvorsprüngen einen kontinuierlichen Kennwert zuzuweisen, der unabhängig vom biologischen Geschlecht indiziert, welche Fähigkeiten überwiegen. Diese Skala konnte jedoch andere Zusammenhänge bei PISA 2003 nicht hinreichend erklären und stellte sich daher als nicht valide heraus.

Eine Analyse mit dem Mixed Rasch-Modell konnte die praktische Relevanz der kognitiven Teilkompetenzen für Geschlechterunterschiede nicht bestätigen. Die latente Klasseneinteilung identifizierte also weder ein latentes Geschlecht, noch entsprechen die Klassen einem verbal-bewertenden oder räumlich-grafisch-abstrakten Profil. Zur Lösung einer Aufgabe spielt das Geschlecht demnach nur eine untergeordnete Rolle. Inhaltlich ließen sich die identifizierten

Zusammenfassung

Klassen quantitativ auf ihr Leistungsniveau und qualitativ auf Unterschiede in der Bewältigung von praxisnahen, alltagsbezogenen Problemstellungen zurückführen. Während leistungsstarke Schüler bei Aufgaben mit starkem Praxisbezug in ihren Leistungen absinken, profitieren eher leistungsschwache Schüler gerade von solchen Aufgaben.
Für PISA ergibt sich die Implikation, relevante Leistungsklassen auf Basis der Antwortmuster zu bilden, um so relevante Heterogenitäten aufzudecken und genauer analysieren zu können.

Vorwort

In der Zukunft meiner Träume gibt es keine Männer und Frauen, sondern nur Menschen. In der Zukunft meiner Träume werden Arbeiten wie diese nicht mehr gebraucht.
(Désirée Burba)

Bis dahin werden viele Jahre vergehen, in denen Untersuchungen wie diese den Unterschied zwischen den Geschlechtern zum Thema ihrer Untersuchung machen, um eine optimale Förderung der Geschlechter zu erreichen, Unterschiede zu nivellieren und eines Tages Arbeiten wie diese undenkbar zu machen.
Eine solche Zukunft erscheint nur schwer vorstellbar und niemand kann entscheiden, welche Zukunft eine gute ist. Ich bitte Sie, sich einmal vorzustellen, dass die zumeist offensichtliche Teilung zwischen Mann und Frau nicht länger in unseren Köpfen existiert und unsere Handlungen nicht mehr beeinflusst. Wie werden wir uns kleiden? Wie werden wir einen Partner finden? Auf was werden wir bei einem Gegenüber **als allererstes** achten?
Philosophen diskutieren bereits den Zerfall von Bindungen durch die voranschreitende Auflösung der Geschlechterdichotomie. Treffen Rousseaus Postulate zu, in denen nur die Übernahme der typischen Geschlechterrollen das Vorankommen unserer Kultur sichert?
Die Welt ohne die Wahrnehmung des biologischen Geschlechts und ohne die übliche Dichotomisierung wird eine andere sein. Im meinen Augen eine bessere, aber vielleicht auch eine, die es niemals geben kann.
Bei meinen Studien und Diskussionen zum Thema Geschlecht hatte ich die Gelegenheit, viele interdisziplinäre Gespräche zu führen. Erstaunlicherweise ertappte ich mich immer wieder dabei, dass ich unter Naturwissenschaftlern die Bedeutung der psychosozialen Faktoren betonte, unter Sozialwissenschaftlern aber oftmals an die biologischen Erklärungsansätze erinnerte. Dies spiegelt sicher wieder, dass in meinen Augen der psychobiosoziale Ansatz die Geschlechterunterschiede am plausibelsten erklären kann. Meiner Meinung nach fügen sich erst unter dieser Perspektive alle Teile des geschlechtlichen Puzzles zusammen und es würde dem Menschen als sozialem *und* biologischem Wesen nicht gerecht, eine dieser beiden Qualitäten auszublenden.
Die Geschlechtsvariable wird oft als irrelevant betrachtet, ggf. gilt es sogar als altmodisch, überholt oder nicht mehr zeitgemäß, sich überhaupt noch mit dem Thema zu beschäftigen. Von einigen scheint die Forschung dazu als redundant gesehen zu werden. Manchmal erscheint es auf den ersten Blick, als sei schon alles zu dem Thema gesagt. Oder das Thema gilt als zu heißes Eisen, das man lieber nicht anfassen sollte. Zu weitreichend scheinen

Implikationen, wenn man Dinge feststellt, die lieber nicht herausgefunden werden sollten. Ich bin nicht davor zurückgeschreckt, schlafende Hunde zu wecken. Allerdings merkte ich schnell, dass Wahrnehmung und Weltanschauung nicht unbeeinflusst bleiben, je länger man sich mit dem Thema Geschlecht auseinandersetzt.

Schnell stellt sich die Frage, was man selbst erreichen möchte, welche Zukunft man sich für die Geschlechterthematik wünscht und welche überhaupt realisierbar wäre. Aber genau das macht für mich die Geschlechterforschung so wichtig und interessant. Das ständige Wechselspiel aus biologischer Implikation und sozialer Konstruktion ermöglicht einen Aktionismus für eine vielleicht gerechtere Welt. Denn eine rein deskriptive Feststellung von Befunden lässt die Menschheit auf der Stelle treten. Nur das Ergreifen geeigneter Maßnahmen wird Veränderungen bringen, dazu muss jedoch die Forschung die Basis für Entscheidungen liefern. Allzu leicht werden Änderungsmaßnahmen ergriffen, ohne den Nachweis erbracht zu haben, dass sich dadurch wirklich etwas ändert. Bringen z. B. sprachliche Änderungen wirklich etwas im Sinne einer Geschlechternivellierung?

Da der Beweis dafür in meinen Augen noch aussteht, verwende ich in der vorliegenden Publikation stets die männliche Form, spreche also in dieser Arbeit vorwiegend von Schülern und meine damit dann Mädchen und Jungen. Es hätte ebenso gut die weibliche Form verwendet werden können.

Für die großartige Unterstützung danke ich meinem Doktorvater Prof. Dr. Jürgen Rost, der mich herzlich und kompetent betreute, immer ein offenes Wort, einen guten Rat oder eine Idee für mich hatte und ohne den ich diese Arbeit sicher nicht hätte vorlegen können. Weiterhin möchte ich mich bei zahlreichen Doktoranden, HiWis und vielen anderen Mitarbeitern des Leibniz-Instituts für die Pädagogik der Naturwissenschaften (IPN) in Kiel für die technische Unterstützung, den fachlichen und persönlichen Gedankenaustausch bedanken, der mich in verschiedenster Weise weitergebracht hat.

Für die inhaltliche Anregung und zahlreichen Diskussionen auch über den rein fachlichen Horizont hinaus danke ich der gender studies group des interdisziplinären Promotionsstudienganges an der Christian-Albrechts-Universität (www.uni-kiel.de/gender/default.htm) und allen, die dabei mitgewirkt haben.

Für das sorgfältige Korrekturlesen, die lustigen Kommentare und tollen Smileys danke ich Ute Schröder, für die methodischen Gespräche den Doktoranden des IPN, für die persönliche Unterstützung Gesa Ramm sowie allen HiWis des PISA-Projekts. Saskia Freiberger hat mich zuverlässig und engagiert bei der Literaturbeschaffung unterstützt.

Zu guter Letzt möchte ich meinen Dank an meinen Mann Andreas Ammann richten, der mit mir lange Diskussionen führte, der mir wie immer eine moralische und motivierende Stütze war und somit auch zur Entstehung dieser Veröffentlichung beigetragen hat. Wir beide haben es bisher geschafft, eine von festgelegten Geschlechterrollen weitgehend freie Partnerschaft und Ehe zu führen und setzen alles daran, dass dies auch weiterhin möglich ist.

Zu Entstehung der vorliegenden Arbeit haben indirekt und direkt beigetragen:

Geniale Sandwiches bei Subway in den Mittagspausen, die Geduld meiner Familie, Freunde und Bekannten, wenn ich sie aus Zeitgründen vernachlässigt habe, vier kuschelige Meerschweinchen, die mir immer wieder eine angenehme Pause beschert haben, ruhige kinderlose freundliche Nachbarn, meine Mutter, die so freundlich war, mir ihr Auto dauerhaft zu leihen, Kirstin Lobemeier, die mir Tabletten aus den USA mitbrachte, so dass ich die Promotionsphase auch kulinarisch überleben konnte, Frau Marienfeld, die mich Gelassenheit lehrte, Nescafé Gold für einen wachen Start in den Morgen und konzentrierte Nächte, ein meist nicht ganz so zuverlässiger Computer, die Liebe meines Ehemannes.

1 Die Psychologie des Geschlechts – Perspektiven aus der Forschung

In Abbildung 1 ist die Gliederung des nachfolgenden Kapitels skizziert. Zunächst wird der Frage nachgegangen, was Geschlecht überhaupt bedeutet (Abschnitt 1.1). Anschließend werden die Geschlechterunterschiede in verschiedenen Teilgebieten der psychologischen Forschung vorgestellt (Abschnitt 1.2). Im Abschluss erfolgt eine genauere Betrachtung der Geschlechterforschung speziell in large scale assessment – Studien (Abschnitt 1.3).

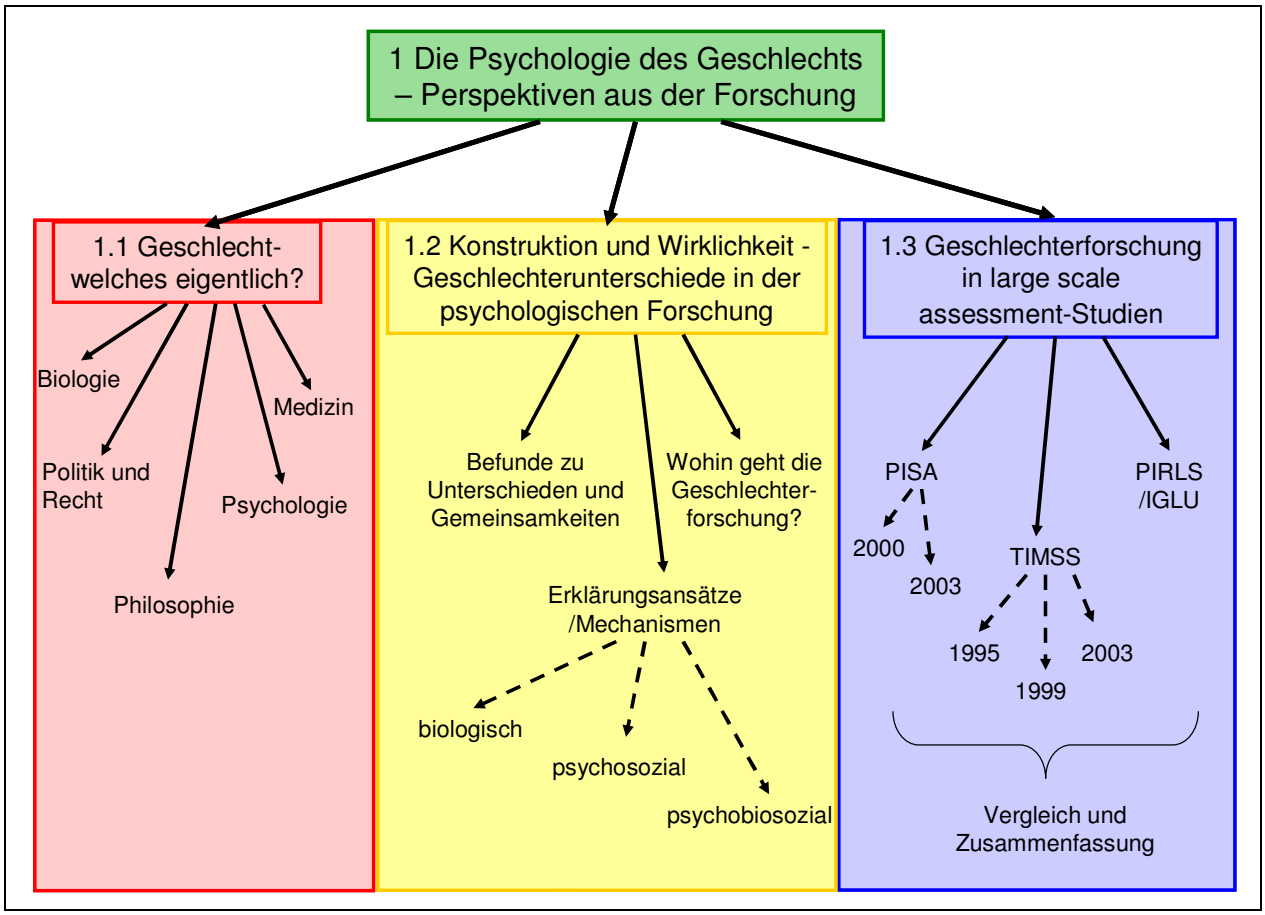

Abbildung 1 Visualisierung des Aufbaus von Abschnitt 1.

1.1 Geschlecht - Welches eigentlich?

Der Begriff „Geschlecht“ löst eine Vielzahl an Assoziationen aus. Der Leser sei an dieser Stelle gebeten, vor dem Weiterlesen einmal kurz inne zu halten und sich zu überlegen, was genau er (oder sie) mit diesem Wort spontan verknüpft, welche Erwartungen an den folgenden Text bestehen.

Etymologisch gesehen kommt das Wort „Geschlecht“ vom althochdeutschen „gishlanti“, was so viel bedeutet wie „in die selbe Richtung schlagen“.

Es soll in der vorliegenden Publikation weder um das musikalische Tongeschlecht, noch um sprachwissenschaftliche Erörterungen zum Genus gehen, sondern Unterschiede zwischen Frauen und Männern oder Mädchen und Jungen thematisiert werden. Die Kategorie Geschlecht ist jedem vertraut und bekannt, trotzdem oder gerade deshalb liefert jede wissenschaftliche Disziplin eine eigene Definition.

Das physiologisch-anatomische bzw. biologische Geschlecht ist angeboren und damit zunächst ein (nahezu) unveränderbares Merkmal eines Menschen, dennoch stellt es neben der biologischen Kategorie gleichzeitig eine bedeutende soziale Kategorie dar, die durch verschiedene Mechanismen stark variierbar und beeinflussbar ist.

Aber worüber ist das Geschlecht überhaupt definiert?

Wozu brauchen wir Geschlechter?

Gibt es mehr als zwei Geschlechter?

Wie groß sind Unterschiede zwischen Geschlechtern wirklich?

Wie weit soll die Gleichbehandlung der Geschlechter gehen?

Wie sieht eine Zukunft ohne gender *aus? Ist sie erstrebenswert und sinnvoll oder überhaupt erreichbar?*

Verpasst ein Kind, das ohne konventionelle Geschlechterrollen aufwächst, sich mit einem Geschlecht zu identifizieren? Wird es Probleme bei der Partnerwahl und Partnerfindung haben?

Aus mancher Perspektive erscheinen solche Fragen absurd, redundant, merkwürdig. Aber bei genauerer Analyse erfordern Geschlechterfragen kritische Betrachtungsweisen.

Im Folgenden sollen zunächst kurz die verschiedenen fachspezifischen Perspektiven vorgestellt werden, um zu demonstrieren, dass Geschlecht immer auch ein Konstrukt darstellt (vgl. zum Aufbau Abbildung 1). Selbstverständlich kann es bei der Darstellung der Fächer außerhalb der Psychologie nur um einen groben Überblick gehen, bei dem stark vereinfacht wird.

1.1.1 Philosophisches zur Geschlechterteilung

Schlägt man im historischen Wörterbuch der Philosophie unter dem Stichwort „männlich/weiblich“ nach (Heinz, Kranz, & Kuster, 2004), so zeigt sich eine seit der Antike Jahrhunderte, ja sogar Jahrtausende währende Debatte um die Bedeutung des Geschlechts und der Geschlechterrollen auf.

Während in der Antike die Geschlechter als Resultat der Teilung eines vollkommenen Einen gesehen werden, gibt es bereits hier Ansätze zu einer Vereinigung von männlich und weiblich in einem androgynen Geschlecht.

Dennoch sind schon früh geschlechtsspezifische Rollenbilder formuliert und in einem von Polaritäten geprägten Denken für eine lange Zeit übernommen worden. Auch christlich-theologische Ansätze ändern nichts daran, sondern bestätigen übliche Zweiteilungen mit zugehörigen Wertungen und Rollen, begründet in der biblischen Schöpfungsgeschichte.

Beginnend im 16. und 17. Jahrhundert werden die tradierten Geschlechterbilder zunehmend in Frage gestellt, zunächst mit dem Ziel, Frauen Zugang zu Bildung zu verschaffen.

Aber erst in der Neuzeit, im 18. Jahrhundert, lassen sich Ansätze zu einem echten ideologischen Geschlechterdiskurs finden. So forderte Th. Hobbes nach dem cartesianischen Gleichheitsprinzip der Irrelevanz eines Geschlechterunterschiedes gleiche Erziehung und Bildung für Männer und Frauen. Demnach ist er gleichsam als der geistige Vater des heute herrschenden Strebens nach geschlechterfairer und politisch korrekter Vermittlung von Bildung zu sehen.

Jean-Jacques Rousseau (1712-1778, französisch-schweizerischer Philosoph und Schriftsteller) hingegen argumentierte so eindrucksvoll dagegen, dass auf der Basis seiner Ausführungen Frauen von Bürgerrechten ausgeschlossen blieben. Er lieferte mit seinen Werken Legitimationen und Vorstellungen, die sich z. T. bis heute erhalten haben. Laut seiner Argumente seien Frauen schwach, zart, gefühls- und nicht vernunftgesteuert und natürlicherweise nicht zur Erschaffung geistiger Werke bestimmt, während Männer von ihm als stark und rational charakterisiert werden. Immanuel Kant (1724-1804, deutscher Philosoph) bekräftigte Rousseaus Ausführungen und sah die Aufgabe der Frau in der Erhaltung von Gattung und Kultur.

Durch Wilhelm von Humboldt (1767-1835) wird das Weibliche in romantischem Denken idealisiert. Nach ihm kommt die Frau dem Ideal, das Weibliche und Männliche zu vereinen (was ihm zufolge prinzipiell in jedem Menschen gleichzeitig angelegt ist) aber deutlich näher als der Mann. Die Geschlechterdebatte innerhalb der Philosophie konnte sich selbst im

ausgehenden 19. Jahrhundert nicht von traditionellen Rollenbildern lösen und Frauen blieb eine Gleichberechtigung selbst in Wertvorstellungen weiterhin verwehrt.
Ein echter Wandel geschlechtlicher Gleichberechtigungskonzepte setzt erst mit dem 20. Jahrhundert ein. Es wird mehr und mehr erkannt, dass Geschlecht nicht nur in der Physiologie, sondern auch in der Sozialisation fußt, also nicht nur als vererbt, sondern auch als erworben zu betrachten werden muss.
Innerhalb der Strömung des so genannten Geschlechtskonstruktivismus unterscheidet Judith Butler zwischen sex und gender. Ersteres ist biologisch-anatomisch, letzteres sozial begründet. Diese begriffliche Einteilung beherrscht die aktuelle Geschlechterforschung in beinahe allen wissenschaftlichen Disziplinen.

1.1.2 Die biologische Natur des Geschlechts

In der Biologie steht nicht nur der Mensch im Zentrum des Interesses (Ausnahmen bilden hier sicherlich Anthropologie/Humanbiologie). Der Vorteil bei einer solchen Sichtweise ist die möglicherweise weniger eingeschränkte Beurteilung von Verhaltensweisen über Artgrenzen hinaus.
Schlägt man in einer Gesamtdarstellung des Faches Biologie unter dem Stichwort *Geschlecht* (bzw. *sex*, weil englischsprachig) nach (z. B. Johnson & Raven, 1996), so finden sich verschiedene Ausführungen zur Geschlechtsdifferenzierung, Sexualität, Geschlechtsdetermination etc.
Bei den Abhandlungen über die verschiedenen Geschlechtsaspekte scheint ganz allgemein von dem chromosomalen Geschlecht ausgegangen zu werden. Größere Diskussionen zum Thema Geschlecht beschäftigen sich beispielsweise mit den Vorteilen geschlechtlicher Reproduktion im Gegensatz zu asexueller Fortpflanzung.
Der geschlechtlichen Vermehrung wird ein enormer evolutionärer Vorteil zugesprochen, weil dies einen variableren Genpool sicherstellt. Dabei existieren zumeist Unterschiede im Erscheinungsbild der Geschlechter, was auch als Geschlechtsdimorphismus bezeichnet wird. Je nach Art des Lebewesens sind die Unterschiede im Phänotyp verschieden stark ausgeprägt. In der Anthropologie wird sich das übrigens zunutze gemacht: Nach Tausenden von Jahren können männliche Skelette anhand diverser Merkmale wie eckigem Kinn, steilem Kinnwinkel oder schmalem Becken etc. von weiblichen unterschieden werden – ein Umstand, der auch in der Rechtsmedizin ganz aktuell Anwendung findet, um etwa Leichen zu identifizieren (Herrmann, Grupe, Hummel, Piepenbrink, & Schutkowski, 1990).

Die Anzahl der Geschlechter bei Organismen variiert. So gibt es Organismen, die nur über eine Geschlechtsausprägung verfügen – oder keine, was eine Frage des Standpunktes darstellt. Es gibt Lebewesen mit zwei Geschlechtsausprägungen zur gleichen Zeit wie z.B. Gastropoden (Schnecken) bis hin zu solchen mit mehr als zwei Geschlechtern.
Der Fokus des Geschlechts in der Biologie liegt also auf der Fortpflanzung und sieht den Menschen nur als einen speziellen Vertreter vieler verwandter Organismen. Dabei fällt der Mensch unter die Ordnung der Primaten und in die Familie der Menschenaffen. In Affenverbänden existieren übrigens auch verschiedene Formen der Geschlechterteilung: Weibliche Hauptgruppen, Haremsverbände oder gemischte Organisationsformen (Voland, 1993).
Wie im späteren noch deutlich werden wird, eröffnet gerade die biologische Sichtweise auf die Erklärung von Geschlechterunterschieden in der Psychologie eine übergeordnete Perspektive, die immer wieder deutlich macht, dass der Mensch einen Teil der Natur in sich trägt und nur über begrenzte Wahlmöglichkeiten verfügt. Ein Mann kann sich einer Geschlechtsumwandlung unterziehen und äußerlich und rechtlich zur Frau werden – aber das Gebären von Kindern wird ihm dennoch biologisch nicht möglich sein.

1.1.3 Die medizinisch definierten Geschlechter

Auch wenn nach Meinung einiger Sozialwissenschaftler das Geschlecht lediglich Konstruktion ist, existieren multiple physiologische Unterschiede. Medizinisch gibt es eine scharf umgrenzte Definition dessen, was Geschlecht bedeutet (Medizinisches Wörterbuch, Pschyrembel, 1994). Unterschieden wird zwischen chromosomalem (genetischem), gonadalem (Gonaden = Keimdrüsen, Ovarien oder Testes, primäre Geschlechtsmerkmale), genitalem (die sekundären Geschlechtsmerkmale betreffend, Vagina oder Penis) und psychosozialem Geschlecht (unterteilbar in psychisches und soziales Geschlecht). Durch spezifische physiologische (angeborene) Aberrationen ist manchmal nicht sicher gestellt, dass die ersten drei wirklich identisch sind.

Während über psychische Unterschiede des Geschlechts diskutiert wird, sind andere, physiologische Unterschiede selbstverständlich: Nur Frauen können Kinder empfangen, austragen, gebären, und stillen, nur Männer Kinder zeugen. Physiologischen Unterschieden bei der Reproduktion sind Ungleichheiten im Verhalten zugeordnet (Stillverhalten, Reaktion auf das Kindchenschema etc.). Niemand bezweifelt z. B., dass Stillen oder Brutpflege hormonell beeinflusst sind. Der Übergang zu anderen Verhaltensweisen ist hingegen fließend.

Es fällt schwer zu bestimmen, wo die physiologische Steuerung aufhört und der soziale Einfluss anfängt. Wie weiter unten gezeigt werden kann, geht die Frage aber weit darüber hinaus, da sich Psychologie und Physiologie gegenseitig beeinflussen.

1.1.4 Gleichheit vor dem Gesetz? Geschlecht in Politik und Recht

Eine Gesellschaft ist maßgeblich von ihren Regeln bestimmt.
Es stellt sich die Frage, welche Rolle dem Geschlecht in Politik und Recht zukommt. Ein berechtigtes Anliegen einer geschlechterfairen Gesellschaft besteht darin, gleiche Rechte für Männer und Frauen gesetzlich zu verankern. Diesem wurde in Deutschland im Jahr 1949 nachgekommen. Mit der „Gleichheit vor dem Gesetz" befasst sich Artikel 3 des Grundgesetzes für die Bundesrepublik Deutschland. Dort heißt es:

Artikel 3 [Gleichheit vor dem Gesetz]

(1) Alle Menschen sind vor dem Gesetz gleich.
(2) Männer und Frauen sind gleichberechtigt. Der Staat fördert die tatsächliche Durchsetzung der Gleichberechtigung von Frauen und Männern und wirkt auf die Beseitigung bestehender Nachteile hin.
(3) Niemand darf wegen seines Geschlechtes (…) benachteiligt oder bevorzugt werden.

Hiermit ist die Grundlage für eine Gleichberechtigung von Männern und Frauen vor dem Gesetz gegeben – zumindest aus juristischer Sicht.
Aber in Artikel 6, der sich mit Ehe, Familie und Kindern beschäftigt und Artikel 12a findet man folgende Aussagen:

Artikel 6 [Ehe - Familie - Kinder]

(4) Jede *Mutter* hat Anspruch auf den Schutz und die Fürsorge der Gemeinschaft.

Weiterhin heißt es:

Artikel 12a [Militärische und zivile Dienstpflichten]

(1) *Männer* können vom vollendeten achtzehnten Lebensjahr an zum Dienst in den Streitkräften, im Bundesgrenzschutz oder in einem Zivilschutzverband verpflichtet werden.

Es wird zunächst festgelegt, dass Männer und Frauen vor dem Gesetz gleichberechtigt sind, um dies dann in späteren Artikeln zu relativieren: Offenbar gilt in den Artikeln 6 Absatz 4 und 12a Absatz 1 keine Gleichberechtigung von Mann und Frau, in beiden Fällen zu Ungunsten der Männer. Bei Artikel 6 Abs. 4 sind Väter dem Wortlaut nach ausgenommen, während in Artikel 12a Abs.1 nur Männer zu Wehr- bzw. Zivildienst verpflichtet werden. Ein Mann wird demnach wegen seines Geschlechts benachteiligt, wenn er zu Wehr- bzw. Zivildienst eingezogen wird. Gleiches Recht geht also nicht mit gleichen Pflichten einher.
In den Artikeln des Grundgesetzes ist stets das biologische Geschlecht gemeint. Mit dem kulturell geprägten Geschlecht befasst sich die Politik im Rahmen eines Programms des so genannten gender mainstreaming (www.gender-mainstreaming.net/). Mit gender ist hierbei die sozial erlernte Geschlechterrolle gemeint. Ziel des Programms ist, neben der theoretischen, rechtlichen Gleichstellung eine tatsächliche Gleichstellung mittels geeigneter Maßnahmen anzustreben.
In dieser Arbeit wird die Meinung vertreten, dass eine rechtliche Gleichstellung eine notwendige, aber keine hinreichende Bedingung darstellen kann. So lange Interessen und Selbstkonzepte bereits im frühen Lebensalter divergieren, kann die rechtliche Eröffnung von Möglichkeiten beispielsweise bei der Besetzung von Ausbildungsplätzen nur marginal zu einer wirklichen Veränderung beitragen: Eine Geschlechtsrollentypisierung findet häufig unbemerkt statt (wie vor allem in Abschnitt 1.2.2.2 deutlich werden wird).

1.1.5 Das psychologische Geschlecht

Die verschiedenen Teilgebiete der Psychologie befassen sich mit verschiedenen Aspekten von Geschlechterunterschieden und betonen für gewöhnlich auch unterschiedliche Perspektiven. In jedem Fall stehen zwar Verhalten und Leistungen im Vordergrund, Unterschiede werden jedoch oftmals sehr verschieden interpretiert und erklärt. Die Biopsychologie schildert andere Sichtweisen als z. B. die pädagogische Psychologie: Während bei Pinel (2001) hormonelle Erklärungsmuster im Vordergrund stehen, berichten Gage & Berliner (1996) über die Geschlechterunterschiede im Rahmen von Schule und Familie.
In der Disziplin der Psychologie existiert aufgrund dieser verschiedenen Ansätze eine Diskussion um Anlage versus Umwelt oder auch Biologie versus Soziologie (nature versus nurture, z.B. Pinker, 2004).
Jede Publikation unterliegt selbstverständlich einer speziellen Sichtweise. In der vorliegenden Arbeit soll versucht werden, einer weitgehend offenen Perspektive gerecht zu werden, indem

verschiedene Ansätze vorgestellt werden. In den folgenden Kapiteln wird deshalb ausführlich auf verschieden psychologische Perspektiven des Geschlechts eingegangen.

1.2 Konstruktion und Wirklichkeit - Geschlechterunterschiede in der psychologischen Forschung

Im Folgenden werden sowohl Befunde zu Geschlechterunterschieden vorgestellt als auch Erklärungsansätze erläutert (vgl. zum Aufbau des nachfolgenden Abschnitts Abbildung 1). Spezielle Befunde aus large scale assessment - Studien sind unter Punkt 1.3 ausgeführt.

1.2.1 Befunde zu kleinen Unterschieden unter großen Gemeinsamkeiten

Berichtet man über Geschlechterunterschiede, so fällt auf, dass die Effekte oft eher gering sind und nur in Ausnahmefällen größere Ausmaße erreichen. Diskutiert man über Diskrepanzen, geraten diese derart ins Zentrum der Aufmerksamkeit, dass ihr tatsächliches Ausmaß leicht überschätzt wird. Dabei sind die Leistungsdivergenzen innerhalb eines Geschlechts größer als zwischen den Geschlechtern. Mit anderen Worten: Die psychologischen Ähnlichkeiten zwischen Männern und Frauen sind größer als die psychologischen Unterschiede. Dieser Sachverhalt ist in Abbildung 2 grafisch veranschaulicht.

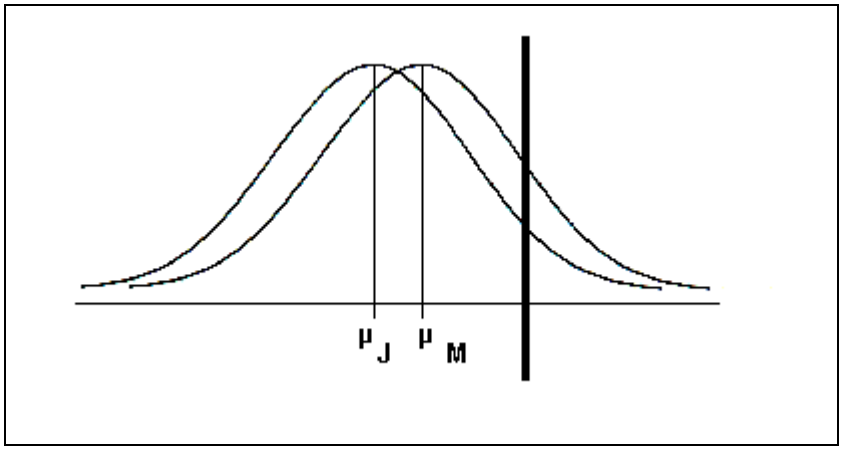

Abbildung 2 Schematische Darstellung kognitiver Geschlechterunterschiede. Obwohl sich die Populationsmittelwerte von Jungen (μ_J) und Mädchen (μ_M) unterscheiden, überlappen sich die Verteilungen stark. Der Unterschied innerhalb eines Geschlechts ist größer als der Unterschied zwischen den Geschlechtern. Betrachtet man nur den oberen Leistungsbereich (in der Abbildung rechts vom fett gedruckten Balken), so ist ein Geschlecht anteilsmäßig stärker vertreten. Je nachdem wie stark sich die Verteilungen überlappen, treten Unterschiede in den Anteilen und den Mittelwerten deutlicher hervor.

Oft sind die Unterschiede nicht einmal von praktischer Relevanz. Bei den folgenden Darstellungen ist also zu stets zu betrachten, wie hoch die Effektstärken ausfallen, zu einer inhaltlichen Einordnung wird die Einteilung nach Cohen (1988) verwendet, vgl. Tabelle 1. Um Unterschiede in ihrer praktischen Bedeutsamkeit einordnen zu können, normiert die Effektstärke d die Unterschiede zwischen den experimentellen Gruppen auf die Streuung der Testwerte.

Tabelle 1 Einordnung der Effektstärke nach ihrer praktischen Bedeutsamkeit, Einteilung nach (Cohen, 1988).

d = 0.8	**➔**	**großer Effekt**
d = 0.5	**➔**	**mittlerer Effekt**
d = 0.2	**➔**	**kleiner Effekt**

Gage & Berliner (1996) weisen auf einen wichtiges Phänomen hin: Ergebnisse von Studien, in denen Geschlechtsunterschiede gefunden wurden, werden häufiger publiziert bzw. eher zur Publikation angenommen als solche, in denen keine Geschlechterunterschiede nachgewiesen wurden. Dies entspricht möglicherweise der generellen Tendenz, signifikante Ergebnisse als besser zu bewerten und häufiger zu publizieren als nicht signifikante (Phänomen publication bias). So wird eine Überbewertung der Geschlechterunterschiede systematisch produziert. Es rücken nur die betreffenden Unterschiede in das Zentrum der Aufmerksamkeit. Gemeinsamkeiten bzw. das Fehlen von Unterschieden sind ggf. häufiger nachgewiesen, werden aber vernachlässigt und unterschätzt.
Empirisch gefundene Unterschiede in kognitiven Leistungen sind auch keineswegs widerspruchsfrei.

Im Folgenden werden (auch kontroverse) Befunde zu kognitiven Geschlechterunterschieden vorgestellt, die für die vorliegende Arbeit von zentraler Bedeutung sind. Dazu zählen die Kontroversen um Unterschiede in der allgemeinen Intelligenz, sowie die vielfach diskutierten Divergenzen in den verbalen, visuell-räumlichen und mathematischen Leistungen. Als Grundlagen kognitiver Unterschiede werden überblicksartig sensorische und motorische Fähigkeitsunterschiede und Unterschiede in Gedächtnisleistungen genannt.
Sofern nicht anders angegeben, beziehen sich die Ausführungen der folgenden Abschnitte auf Halpern (2000).

Sensorische und motorische Fähigkeitsunterschiede

Eine Fülle von Studien weist darauf hin, dass geschlechtsspezifische Diskrepanzen bereits im Bereich sensorischer und motorischer Fertig- und Fähigkeiten existieren (Baker, 1987; Morrell, Gordon-Salant, Pearson, Brant, & Fozart, 1996; Nicholson & Kimura, 1996; Rebok, 1987), die ihrerseits wesentliche Voraussetzungen für kognitive Leistungen darstellen. So kann bessere Wahrnehmung oder eine niedrigere Reizschwelle diverse Leistungen auf einem frühen Niveau modifizieren.

Dass viele physiologische Unterschiede schon im Kindesalter nachweisbar sind, heißt allerdings nicht, dass sie im weiteren Leben nicht noch beeinflussbar wären.

Zu geschlechtervariablen Wahrnehmungsleistungen zählen: Hören, Geruch und Geschmack, Berührungsschwellen, Sehschärfe, Zeitwahrnehmung. Während Frauen besser in feinmotorischen Tätigkeiten abschneiden (möglicherweise aber nur durch ihre durchschnittlich geringere Fingerbreite bedingt), zeigen sich Männer besser darin, bewegliche Ziele zu treffen. Diese Befunde können sowohl evolutionär als auch umweltbedingt durch häufigere Praxis in Ballspielen etc. vermittelt worden sein, dazu an späterer Stelle mehr.

Unterschiede in der allgemeinen Intelligenz

Schon seit geraumer Zeit ist (auch in der breiten Öffentlichkeit) von Interesse, ob sich Männer und Frauen in ihrer Intelligenz unterscheiden. Obwohl ein quantitativer Vergleich der Intelligenz z. B. ethnischer Gruppen undenkbar und politisch inkorrekt wäre, wird eine solche Fragestellung die Geschlechter betreffend selten als problematisch empfunden.

Wie bei allen Teilgruppen, deren Intelligenz man vergleicht, stellen sich auch beim Vergleich der Geschlechter einige Probleme ein. Verschiedene Altersstufen sowie Kultureinflüsse, Sozialschicht und viele weitere Faktoren (soziale Ursachefaktoren zur Erklärung von Geschlechterdiskrepanzen unter Punkt 1.2.2.2) verändern Leistungen in Tests, mit denen Intelligenz gemessen wird, so dass sich eine Vermessung möglicher Intelligenzunterschiede nicht nur als langwierig, sondern auch als störanfällig erweist. Selbstverständlich beeinflusst auch die Messmethode an sich und das jeweilige Intelligenzkonstrukt die Resultate.

Weiterhin hat es sich als weniger sinnvoll herausgestellt, Unterschiede an der globalen Intelligenz festzumachen, denn hier sind mehrheitlich keine bedeutsamen Differenzen auszumachen (Halpern, 2000). Wesentlich fruchtbarer ist die Feststellung gewisser Teilfähigkeiten, die nach Geschlecht divergieren. Unterschiede treten im jeweiligen Gebrauch bestimmter Fähigkeiten zutage. Es ist also eher von Interesse, in welchem Zusammenhang Unterschiede in kognitiven Leistungen zwischen Mädchen und Jungen zu beobachten sind.

Gedächtnisleistungen

Frauen schneiden bei zahlreichen Gedächtnisleistungen besser ab als Männer. Dazu zählen das visuelle und Kurzzeit-Gedächtnis, episodische Gedächtnisleistungen sowie Erinnerungen an räumliche Lagebestimmungen. Frauen berichten durchschnittlich über frühere Erinnerungen in der Kindheit. Sie schneiden besser in Aufgaben mit Namensassoziationen (auch in Verbindung mit Gesichtern) und Wortlisten ab, wobei dieser Effekt auf den verbalen Kompetenzvorsprüngen beruhen kann.

Verbale Geschlechtsdifferenzen

Ebenso vielfältig wie das Konzept der Gedächtnisleistungen ist auch das Feld der verbalen Leistungen, bei denen unterschiedlichste Gehirnregionen beteiligt sind. Die Ergebnisse sind allerdings kontrovers: Gage & Berliner (1996) kommen zu dem Schluss, dass keine bedeutenden Unterschiede zwischen Mädchen und Jungen in verbalen Leistungen vorhanden sind und verzichten daher auf eine Darstellung dazu. Allerdings ist in vielen Untersuchungen eine Überlegenheit weiblicher Probanden bei verbalen Anforderungen nachweisbar (Johnson, 1996; Lehmann, 1994; Maccoby & Jacklin, 1974; Richter, 1996), einige Befunde werden nachfolgend vorgestellt.

In einer Meta-Analyse berichten Hyde & Linn (1988) eine Effektstärke von 0,11 zu Gunsten der Frauen, global über alle einbezogenen Studien gerechnet. Demnach ist der Vorsprung der Frauen in verbalen Fähigkeiten nur von geringer praktischer Relevanz (vgl. Tabelle 1). Die Effektstärken schwanken je nach Fähigkeit, Test, Alter der Probanden und Setting.

Geschlechterunterschiede scheinen schon bei Kindern im Alter von einem Jahr nachweisbar zu sein, also sehr früh in der Entwicklung. Bis zum Alter von 6 Jahren schätzen Hyde & Linn (1988) die Effektstärke auf etwa 0,13.

Studien, die das Alter von sechs bis 25 Jahren berücksichtigen, zeigen in der Meta-Analyse durchschnittlich keinen Effekt. Erst ab dem Alter von über 25 Jahren steigt der Effekt wieder auf etwa 0,2 an. Ob der generelle Kompetenzvorsprung in verbalen Fähigkeiten bis ins Alter erhalten bleibt, ist nach Meinz & Salthouse (1998) unklar.

Unterscheidet man die Ergebnisse noch nach Fähigkeit, ergibt sich ein etwas anderes Bild:

In Tests zum Wortschatz schneiden Jungen zwischen 6 und 10 Jahren besser ab ($d = -0{,}26$), während sich der Befund im Alter von 19 bis 25 wieder umkehrt ($d = 0{,}23$) und Frauen deutlich im Vorteil sind. Von der größten Effektstärke wird bei der Lesenkompetenz berichtet ($d = 0{,}31$).

Betrachtet man ergänzend zu dieser Meta-Analyse Störungen verbaler Leistungen, also den eher niedrigen Kompetenzbereich, so finden sich anteilsmäßig mehr stotternde oder lese-/rechtschreibschwache Jungen als Mädchen (z.B. Vandenberg, 1987). Nach Schlaganfällen oder operativen Eingriffen ins Gehirn haben Männer mehr Sprachausfälle und erlangen ihre sprachlichen Fähigkeiten später wieder als Frauen (Kolb & Wishaw, 1993).
Verbale Kompetenzvorteile von Frauen bzw. Mädchen sind auch in neueren Studien aus dem Bereich der Bildung mehrfach nachgewiesen (Bos et al., 2003b; Stanat & Kunter, 2001; Zimmer, Burba, & Rost, 2004; vgl. dazu Abschnitt 1.3).

Bei (Halpern, 2000) wird gleichzeitig zu diesen Befunden darauf hingewiesen, dass - trotz des relativen Vorsprungs der Frauen - angesehene Berufe, in denen verbale Fähigkeiten wichtig sind, eine lange Zeit Männerdomänen waren oder noch immer sind. Dazu zählen z. B. Rechtsanwalt, Politiker, Journalist. Auf diesen Aspekt wird unter 1.2.2.2 noch ausführlich eingegangen.

Unterschiede in visuell-räumlichen Leistungen

Die visuell-räumlichen Fähigkeiten werden – wie die verbale Komponente – als unabhängiger Faktor der Intelligenz gesehen. Der visuell-räumlichen Fähigkeit liegen verschiedene Konzepte zugrunde, die Ergebnisse zu Geschlechterunterschieden variieren je nach Definition, Test und Altersgruppe (Maier, 1999). Ein Beispiel stellt die mentale Rotation eines Objekts dar.
Dass Männer in verschiedenen Tests zu visuell-räumlichen Fähigkeiten im Durchschnitt besser als Frauen abschneiden, ist oft nachgewiesen worden (Loring-Meier & Halpern, 1999; Maier, 1999; Nyborg, 1983; Richardson, 1991). Ein Vorteil männlicher Individuen wurde bei einigen visuell-räumlichen Aspekten auch im Tierversuch nachgewiesen, was biologische Ursachen vermuten lässt.
Der Leistungsvorsprung der Männer wird in verschiedenen Tests deutlich (z.B. Labyrinth-Tests, Rotation zwei- oder dreidimensionaler Objekte; Halpern, 2000; Kolb & Wishaw, 1993).
Männliche Probanden nutzen ihre besseren räumlichen Fähigkeiten offenbar auch vermehrt mittels visueller Techniken zur Lösung eines Problems. Frauen erzielen bei gleichen Problemstellungen mit den von ihnen bevorzugten verbalen Lösungsstrategien unter Umständen schlechtere Ergebnisse.

Die Effektstärke der Geschlechterunterschiede in visuell-räumlichen Leistungen wird insgesamt auf etwa 0,37 geschätzt, bei speziellen Tests deutlich höher, am höchsten mit 0,9 bei Aufgaben zur mentalen Rotation, was bereits recht große Effekte indiziert.
Entwicklungspsychologisch betrachtet treten Diskrepanzen zwischen den Geschlechtern offenbar bereits im Kleinkindalter auf, im Gegensatz zu den verbalen beginnen die visuell-räumlichen Fähigkeiten allgemein mit dem Alter stärker nachzulassen. Wie sich das zunehmende Alter auf die Geschlechterunterschiede in diesem Bereich auswirkt, ist noch nicht vollständig geklärt.
Wie auch schon bei den verbalen Fähigkeiten angedeutet, sind die im Labor getroffenen Aussagen allerdings nicht einfach auf das praktische Leben zu übertragen.

Divergenzen in den mathematischen Fähigkeiten

in vielen Studien zeigen sich für Mathematik keine Diskrepanzen zwischen den Geschlechtern. So sind nach einer Meta-Analyse von Hyde, Fennema, & Lamon (1990), in die 100 Studien einbezogen wurden, global betrachtet keine Geschlechterdifferenzen bezüglich mathematischer Leistungen festzustellen.
Unterschiede in mathematischen Fähigkeiten werden erst im Zusammenhang mit den verwendeten Tests und dem Alter der Probanden deutlich.
Differenziert nach verschiedenen Teilgebieten und Altergruppen zeichnet sich folgendes Bild: Frauen zeigen bei Beweisführungen oder beim Lösen mathematischer Gleichungen bessere Leistungen als gleichaltrige Männer. Letztere hingegen schneiden besser ab, wenn es um Geometrie, mathematische Größen, Wahrscheinlichkeit und Statistik geht (Stones, Beckmann, & Stephens, 1982). Auffällig ist hierbei, dass die Gebiete, in denen Frauen höhere Leistungen erzielen, mittels verbaler Strategien bearbeitet werden können, während in den männerspezifischen Teilaspekten eine visuell-räumliche Informationsverarbeitung von Vorteil ist. Dieser Befund passt zu den oben getroffenen Aussagen. Ergänzende, aber z. T. auch widersprüchliche Befunde finden sich bei der Darstellung der Ergebnisse aus large scale assessment - Studien in Abschnitt 1.3. Dort gibt es auch einige Hinweise darauf, dass die Kluft zwischen den Geschlechtern zunimmt, wenn man Personen des oberen Leistungsspektrums vergleicht.
Kongruent dazu treten bei hochbegabten Kindern Geschlechterunterschiede in der mathematischen Kompetenz schon im Kindergartenalter auf (Robinson, Abbott, Berninger, & Busse, 1996), sind ansonsten häufig erst verstärkt bei pubertierenden Jungen und Mädchen zu beobachten.

Nach Altersstufen betrachtet, zeigt sich erst ab der High School eine Effektstärke von d = 0,29, die im College auf d = 0,41 anwächst und sich im Erwachsenenalter auf d = 0,59 einpendelt, stets zu Gunsten der Männer. Mit zunehmendem Alter nehmen die Diskrepanzen also zu. Allerdings basiert dieser Befund auf Querschnitts- und nicht auf Längsschnittsdaten. Es zeigte sich zudem, dass in älteren Studien die Diskrepanzen höher waren als in solchen jüngeren Datums (Hyde et al., 1990).

Fennema (1996) fasst zusammen, dass sich die Geschlechterunterschiede in Mathematik mit der Zeit verringert haben und in Abhängigkeit von sozioökonomischem Status, ethnischer Zugehörigkeit sowie Schule und Lehrer variieren.

Spencer, Steele, & Quinn (1999) berichten davon, dass gefundene Leistungsschwankungen zwischen Männern und Frauen in ihren Untersuchungen von der Erwartungshaltung der Probanden abhingen. Wurden den College-Studenten vor einem Mathematik-Test explizit mitgeteilt, dass keine Geschlechterunterschiede erwartet werden, so schnitten Männer und Frauen gleich ab, während ohne diese Instruktionen deutliche Leistungsvorsprünge der Männer beobachtbar waren. Die Überzeugung, dass keine Geschlechterunterschiede in der Leistung auftreten werden, hat die Probanden offenbar so in ihrem Verhalten beeinflusst, dass die Leistungen zwischen Frauen und Männern tatsächlich nicht divergierten.

Auch die geschlechtsspezifischen Befunde in Mathematik können also als Folge diverser Ursachen interpretiert werden.

Zusammenfassung kognitiver Geschlechterunterschiede

Will man obige Befunde zusammenfassen, so wird deutlich, dass eine einheitliche Aussage nur schwer möglich ist. Der Leistungsvorsprung von Männern ist am deutlichsten in den visuell-räumlichen Fähigkeiten zu beobachten, während die verbalen Leistungen bei Frauen etwas höher liegen. Die mathematischen Geschlechterunterschiede schwanken je nach Teilgebiet, global betrachtet können hier nur geringe Unterschiede verzeichnet werden. Insgesamt scheinen die Geschlechterunterschiede mit zunehmendem Alter und zunehmender Leistung anzuwachsen, allerdings kann dies auch als ein Effekt von Querschnittsuntersuchungen interpretiert werden.

Wichtig zu beachten ist hierbei, dass keine Unterschiede in der allgemeinen Intelligenz angenommen werden können.

Die meisten Studien zu Geschlechterunterschieden in kognitiven Leistungen sind nur auf ein Land bzw. Kulturkreis bezogen, nur in wenigen Studien werden mehrere Länder parallel untersucht und verglichen. Die so genannten large scale assessment - Studien sind hingegen

gerade so angelegt, dass internationale Vergleiche ermöglicht werden. Sie bieten eine gute Möglichkeit, Geschlechterunterschiede über verschiedene Kulturen hinweg zu betrachten. Die Vorteile, aber auch die Grenzen solcher Studien finden sich in Abschnitt 1.3.
Obwohl die Unterschiede also als eher gering zu beurteilen sind, bestimmen sie dennoch Urteilen und Handeln. Zugrunde liegende Mechanismen werden in den nachfolgenden Abschnitten genauer beleuchtet.

1.2.2 Multiple Erklärungsansätze zu Geschlechterunterschieden

Für die beobachtbaren Unterschiede zwischen Männern und Frauen gibt es unterschiedliche Erklärungsansätze.
Die Erklärungsansätze lassen sich in die Anlage-Umwelt-Debatte einreihen. So gibt es biologische wie psychosoziale Ansätze, die jeweils recht plausibel verschiedene Effekte erklären. Dazu lassen sich jeweils hypothesenkonforme und widersprüchliche Ergebnisse anführen. Aber erst der psychobiosoziale Ansatz als eine Integration beider Theorien liefert zu den uneinheitlichen Befunden ein umfassendes und zufrieden stellendes Bild, wie unter 1.2.2.3 gezeigt. Im Folgenden soll jede Forschungsrichtung kurz vorgestellt werden.
(Zur Orientierung sei nochmals auf Abbildung 1 verwiesen, welche folgende Abschnitte als Gliederung wiedergibt.)

1.2.2.1 Geschlecht als biologisches Programm?

Biologische Ansätze bieten eine spezielle Perspektive der Erklärung von Geschlechterunterschieden. Aber gerade biologische Modelle werden häufig missverstanden. So kommt es vor, dass besonders unter Geisteswissenschaftlern eine biologische Sichtweise zur Erklärung psychischer Vorgänge als Diskriminierung von Frauen oder auch als naive Sichtweise der Dinge beurteilt wird.
Um ein vollständiges Bild über das Verhalten zu gewinnen, wird aber die Einbeziehung biologische Erklärungsmechanismen benötigt. Eine Ablehnung der Biologie würde heißen, zentrale wissenschaftliche Erkenntnisse außer Acht zu lassen.

Biologische Ansätze zur Erklärung von psychologischen Geschlechterunterschieden ziehen genetische Grundlagen, hormonelle Steuerungen, Gehirnorganisation und einen

Selektionsvorteil im Rahmen der Evolution heran (vgl. Abbildung 1). Der menschliche Körper unterliegt verschiedenen biologischen Mechanismen, genau dieses biologische Programm beeinflusst menschliches (geschlechtsspezifisches) Verhalten. Der Mensch hat sich über Millionen von Jahren unter einem selektiven Druck zu dem Lebewesen entwickelt, das jetzt in der Psychologie beforscht wird.
Der Mensch gehört zu der Gruppe der Säugetiere und hat viele genetische wie verhaltenstheoretische Ähnlichkeiten mit anderen Säugetieren. Dennoch werden kognitive Geschlechterunterschiede ungern mithilfe der Biologie erklärt. Biologische Gründe *erscheinen* weniger veränderbar als sozial bedingte Geschlechterstereotype.
Es gibt zahlreiche Hinweise für eine biologische Steuerung der psychologischen Geschlechterunterschiede, die nachfolgend erläutert werden sollen.

Die biologischen Erklärungsansätze sind normalerweise konfundiert: Die chromosomale/genetische Ausstattung bestimmt die Hormonproduktion, beide Faktoren wirken auf die Neuroanatomie, das Gehirn des Menschen und dessen Entwicklung. Als zentrale Grundlage für diese Mechanismen wird die Evolution des Menschen gesehen.
Alle vier Ansätze (Genetik, Hormonhaushalt, Neuroanatmomie, Evolution) sollen im Folgenden erläutert werden. Da in diesem Rahmen nur ein Überblick gegeben wird, beziehen sich folgende Ausführungen – sofern nicht anders angegeben – auf Halpern (2000).

Machen die Gene den kleinen Unterschied? Die Rolle der Erbanlagen

Unsere genetische Ausstattung entscheidet (prinzipiell), ob wir als weibliche oder männliche Individuen geboren werden. Bereits im frühen Embryonalstadium beginnt eine Differenzierung nach Geschlechtern. Die Gonosomen (Geschlechtschromosomen) liegen als XX oder XY vor. Ist ein Y vorhanden, so eröffnet in der 7. Woche nach der Konzeption eine Kaskade von biochemischen Ereignissen die Entwicklung von Hoden. Fehlt ein Y, entwickelt sich der Embryo zu einem weiblichen Organismus.
Studien über genetische Einflüsse auf kognitive Geschlechtsunterschiede fußen auf der Untersuchung genetischer Aberrationen, Zwillingsstudien und der Häufigkeitsverteilung bestimmter Merkmale in großen Populationen, um ein Vererbungsmuster solcher Unterschiede zu finden. Soll ein rein genetischer Einfluss verantwortlich sein, so muss bezogen auf die jeweiligen kognitiven Kompetenzdifferenzen ein bestimmter Erbgang

vorhanden sein. Hierbei kommen nicht nur die Gonosomen infrage, sondern auch die Autosomen (also der Chromosomensatz ohne die Geschlechtschromosomen X und Y).

Eine populäre und leicht nachvollziehbare Theorie sieht nach der rezessiv-dominanten Vererbungslehre von Mendel visuell-räumliche Fähigkeiten rezessiv auf dem X-Chromosom lokalisiert. Hierbei wird angenommen, dass die rezessive Version des Gens, das für räumliche Fähigkeiten kodiert, mit gleicher Wahrscheinlichkeit wie die dominante Version auftritt.

Daraus resultiert folgende Vorhersage: Liegt die Fähigkeit rezessiv auf dem X-Chromosom, sollte sich diese bei Männern vermehrt ausprägen können.

Da Frauen über zwei X-Chromosomen verfügen, ergeben sich vier verschiedene mögliche Genotypen. Nur wenn beide X-Chromosomen die rezessiven Genversionen enthalten, die für gute räumliche Fähigkeiten kodieren, wird sich die Fähigkeit auch durchsetzen. In allen anderen Fällen wird aufgrund der Rezessivität die räumliche Fähigkeit eher weniger gut ausgeprägt, selbst wenn sie genetisch vorhanden ist. Demnach wären bei etwa einem Viertel der Frauen, aber bei etwa der Hälfte der Männer (da diese nur ein X-Chromosom besitzen) die räumlichen Fähigkeiten gut ausgeprägt. Weiterhin müssten Söhne in ihren räumlichen Fähigkeiten phänotypisch häufiger den Müttern entsprechen, Mädchen dagegen den Vätern.

Gittler & Vitouch (1994) konnten den Zusammenhang nicht bestätigen.

Um zu demonstrieren, wie Gene und Verhalten in Form einer Rückkoppelung interagieren, nennt Halpern (2000) als Beispiel ein Kind, das sich die Umwelt nach genetischer Ausstattung selektiert: So könnte es aufgrund seiner genetischen Ausstattung für verbale Tätigkeiten besonders geeignet sein und dadurch verbale Anregung präferieren, also eher ein Buch lesen als andere Beschäftigungsmöglichkeiten wählen. Eine „genetische Umwelt“ könnte ebenso dazu führen, dass in Zwillingsstudien solche Zwillinge ähnlichere Ergebnisse liefern, die getrennt aufgewachsen sind. Sie selektieren ihre Umwelt entsprechend ihrer Genetik, während zusammen aufwachsende Zwillinge eher ihre Unterschiedlichkeit betonen und verschiedene Tätigkeiten wählen. Die Effekte sind also auf eine Interaktion der Gene mit der Umwelt zurückzuführen. Einen relativ neuen Forschungszweig stellt die Verhaltensgenetik dar (z.B. Plomin, DeFries, McClearn, & McGuffin, 2001). In interdisziplinärer Arbeit werden Methoden und Ergebnisse der Genetik auf die Erforschung von Verhalten angewendet.

Insgesamt liefert noch keine Theorie eine überzeugende rein genetische Erklärung der kognitiven Geschlechterunterschiede. Das Feld der Genetik entwickelt sich aber derzeit so rasant, so dass neue Ergebnisse sicher nicht lange auf sich warten lassen. Seit 2001 liegt das menschliche Genom im Übrigen vollständig sequenziert vor, ist also „entschlüsselt“.

Möglicherweise eröffnen sich dadurch auch neue Erkenntnisse zur Rolle der Gene bei geschlechtsspezifischen Fähigkeiten.

Definieren biochemische Botenstoffe Mann und Frau? Hormone und Geschlecht

Täglich wirken in unserem Körper zahlreiche hormonelle Einflüsse, angefangen von der Regulation des Tag-Nacht-Rhythmus bis hin zu einem kurzen Adrenalinschub, wenn wir bemerken, dass der Chef uns über die Schulter schaut.
Hormone sind biochemische Botenstoffe, die im Blut zirkulieren. Dadurch können sie also jedes Organ erreichen, sind aber nicht so schnell wie über Nerven vermittelte Signale. Sie besitzen aber bereits in geringer Konzentration starke Wirkungen, indem sie an spezielle Rezeptoren der Zielorgane binden.
Im Körper von Männern und Frauen zirkulieren identische Hormone, geschlechtsspezifische Unterschiede sind vor allem im Konzentrationsverhältnis der Hormone zu finden: Im Blutkreislauf von Männern zirkuliert eine 17fach höhere Konzentration der so genannten Androgene, der „männlichen“ Hormone wie dem Hauptvertreter Testosteron. Je nach Zyklusabschnitt ist bei Frauen 50mal mehr Östrogen als bei Männern nachweisbar, allerdings fluktuieren bei Frauen die Hormonkonzentrationen mit den Zyklusphasen sehr stark. Auch bei Männern zeigten sich leichte tägliche oder zumindest saisonale Hormonschwankungen. Generell schwanken Hormone auch mit dem Alter bzw. mit verschiedenen Entwicklungsphasen.
Es wird also im Folgenden von weiblichen und männlichen Hormonen gesprochen, auch wenn es streng genommen keine geschlechtsspezifischen Hormone gibt. Gemeint sind aber die Hormone, die jeweils bei Frauen (v.a. Östrogene und Progesteron) oder Männern (vorwiegend Testosteron) in höherer Konzentration als beim jeweils anderen Geschlecht vorkommen.
Hormonschwankungen können große Effekte nach sich ziehen, weiterhin können Hormone in einander überführt werden, so kann Progesteron in Androgene umgewandelt werden. Umgekehrt wird Testosteron im Körper in Östradiol (ebenfalls weibliches Hormon) überführt. Der Nachweis eines moderierenden Einflusses von Hormonen auf kognitive Geschlechterunterschiede gestaltet sich als komplexe Herausforderung.
Der Einfluss von Sexualhormonen wird besonders bei der Betrachtung vulnerabler Phasen in der Entwicklung deutlich, wo Hormone eine große Rolle spielen, dies wird im Anschluss erläutert.

Bei der folgenden Darstellung hormoneller Einflüsse auf Geschlechterunterschiede geht es nicht darum, festzustellen, dass Frauen weniger leisten oder weniger intelligent sind. Stattdessen ist eine deskriptive Feststellung möglicher Unterschiede beabsichtigt. Eine Variation der kognitiven Leistungen von Frauen in Abhängigkeit vom Zyklus kommt also keiner Diskriminierung oder Abwertung von Frauen gleich.

Geschwinds Theorie pränataler Hormoneffekte steigt genau an diesem Punkt ein. Ihm zufolge verlangsamt Testosteron das Zellwachstum des Gehirns, dabei ist die linke Hirnhälfte in der Embryonalentwicklung stärker betroffen als die rechte Hemisphäre. Im Mutterleib wird ein männlicher Fötus bereits stärkeren Konzentrationen von Testosteron ausgesetzt, da dieser nicht nur von der Mutter und den eigenen Nebennieren mit dem Androgen versorgt wird, sondern die Hoden sehr früh mit der Hormonproduktion beginnen. Durch eine Verlangsamung des Wachstums der linken Hemisphäre dominiert die rechte Hemisphäre in der Entwicklung. So erfolgt eine stärkere Lateralisierung der Gehirne der männlichen Föten, während weibliche Föten eine symmetrischere Hirnentwicklung aufweisen. Nimmt man Händigkeit als Indikator der Hirnhälftendominanz und Auswirkung von Lateralisierung, so müsste es gemäß der Theorie Geschwinds unter Männern mehr Linkshänder geben, dies ist auch tatsächlich der Fall. Unter Mädchen, die intrauterin mehr Testosteron ausgesetzt waren, sind ebenfalls mehr Linkshänder zu finden.
Nach Nyborg kommt es aber eher auf einen optimalen Level an bestimmten Hormonen an. Demzufolge ist ein mittlerer Level an Östradiol günstig bzw. optimal für räumliche Leistungen und es sollten feminine Männer und maskuline Frauen entsprechend besser in räumlichen Leistungen abschneiden. Auch für diese Theorie gibt es einige stützende Befunde, für die sich aber auch alternative Erklärungen als möglich erweisen.

Bei Abwesenheit eins Y-Chromosoms werden in der Embryonalentwicklung keine Hoden ausgebildet. Dementsprechend fehlt eine höhere Konzentration männlicher Hormone, das bewirkt die Entwicklung zu einem Mädchen. Bereits an diesem Punkt können diverse Störungen auftreten: Auch ohne ein Y-Chromosom kann eine hohe Testosteronkonzentration auftreten. Umgekehrt kann auch bei Vorliegen eines Y-Chromosoms die Testosteron-Konzentration zu niedrig sein. Die Ausbildung des Geschlechts ist also bedingt und beeinflusst durch verschiedene Faktoren.
Die Sexuelle Differenzierung des Gehirns beginnt möglicherweise schon an dieser Stelle, starke Hormonausschüttungen sind noch 6-12 Monate nach der Geburt nachweisbar, also zu

einer Zeit der kritischen Gehirnentwicklung. Erste Neuronen entwickeln sich zeitgleich mit der Androgenproduktion aus den Hoden, die Bildung der Ovarien erfolgt erst später. Man spricht in dem Zusammenhang von organisierenden und aktivierenden Effekten von Hormonen.

Pränatale Hormonbeeinflussung und damit zusammenhängende Geschlechtsdimorphismen im Verhalten sind im Tierversuch zahlreich nachgewiesen, beim Menschen müssen sich Nachweise auf Menschen mit hormonellen Störungen beschränken, so dass ein Nachweis der Effekte nur mit entsprechender Konfundierung anderer Faktoren möglich ist:

So erfolgt bei der kongenitalen adrenergen Hyperplasie (CAH) in den ersten drei Monaten der Embryonalentwicklung eine stark erhöhte Produktion von Androgenen. Männliche Föten sind leicht erhöhten Mengen an Testosteron ausgesetzt, weibliche Föten bereits abnorm erhöhten Mengen. Betroffene Jungen unterscheiden sich in Leistung und Verhalten nicht von gesunden Verwandten. Betroffene Mädchen hingegen zeigen nicht nur ein aggressiveres Spielverhalten und wählen mehr jungen-typische Spielzeuge, sondern verfügen auch über signifikant bessere visuell-räumliche Leistungen, aber geringere verbale Fähigkeiten als ihre gesunden weiblichen Verwandten.

Männer mit Androgen-Insensitivität entwickeln sich bedingt durch eine Unempfindlichkeit des Gewebes gegenüber Androgenen morphologisch zu einer Frau mit weiblichen Genitalien, weisen aber das genetische Muster eines Mannes auf. Die signifikant besseren verbalen Fähigkeiten und niedrigeren visuell-räumlichen Leistungen könnten allerdings sowohl Folge von Hormonen als auch durch Umwelterfahrungen bedingt sein, da die Betroffenen als Mädchen aufgezogen wurden und ggf. sogar glücklich verheiratete aber kinderlos gebliebene Ehefrauen sind (Pinel, 2001).

Weitere Befunde gehen in eine ähnliche Richtung: Es zeigte sich, dass so genannte spät reifende Kinder (unabhängig von Geschlecht) bessere visuell-räumliche Fähigkeiten aufwiesen, früh reifende höhere verbale Fähigkeiten. Mädchen treten statistisch gesehen früher in die Pubertät ein, sind früher geschlechtsreif. Sie sind somit früher höheren Konzentrationen von Sexualhormonen ausgesetzt, was verbale Fähigkeiten stärker als visuell-räumliche begünstigt. Der Vorteil der Mädchen in verbalen Leistungen kann als Folge früherer (hormoneller) Reifung auf Kosten visuell-räumlicher Leistungen interpretiert werden. Da eine frühe körperliche Reife aber gleichzeitig mit dem Erfahrungshorizont konfundiert, existieren auch zahlreiche alternative Erklärungen - dazu in Abschnitt 1.2.2.2 mehr.

Männer mit Androgendefiziten in der Pubertät zeigen im Übrigen in einem direkten Ausmaß geringere räumliche Leistungen: Je geringer der Androgen-Spiegel, desto geringer fällt die Leistung aus. In verbalen Fähigkeiten zeigt sich kein Unterschied zu gesunden Probanden.
Widersprüchlich zum vorhergehenden deuten einige andere Befunde kongruent im Sinne Nyborgs darauf hin, dass ein höherer Spiegel männlicher Hormone bei Männern räumliche Leistungen senken, bei Frauen hingegen steigert.

Schwankende Hormonmengen im Zyklus einer Frau bringen körperliche Veränderungen mit sich, nachweisbar z. B. über eine Veränderung der Schmerzschwelle, die im übrigen bei Frauen generell höher ist, d. h. Frauen empfinden Schmerz erst bei stärkeren Schmerzreizen als Männer (Litscher, 2004).
Während der Menstruation sind die zirkulierenden Hormonmengen, die Konzentrationen von Östrogen und Progesteron gering, es liegen im Körper der Frau nunmehr nur dreimal so hohe Mengen wie beim Mann vor. Und tatsächlich ist ein besseres Abschneiden in räumlichen Tests bei Frauen genau dann zu beobachten, wenn diese menstruieren.
Während der Zyklusmitte, um den Eisprung herum, liegt die Hormonkonzentration des Östrogens etwa 50mal höher als beim Mann und ein besseres Abschneiden der Frauen in verbalen Tests ist messbar. Unklar ist allerdings, wie hoch hier die Effektstärke ist.
Der Anstieg verbaler Leistungen bei Frauen im Klimakterium den „Wechseljahren“ lässt auf einen modulierenden Effekt von Östrogengaben schließen.

Hormonschwankungen bei Männern drücken sich in einem höheren Testosteronspiegel morgens und (saisonal betrachtet) im Herbst aus, der Zusammenhang zu kognitiven Leistungen ist hier noch ungeklärt.
Hormone beeinflussen schon in der frühen Entwicklung in einem gewissen Ausmaß die kognitiven Strukturen. Hohe Level an Androgenen sind assoziiert mit hoher räumlicher Leistung bei Mädchen. Aber auch in Pubertät und Adoleszenz sind vielfältige Wirkungen von Hormonen auf kognitive Leistungen nachweisbar. Es ist bei der Untersuchung von hormonellen Einflüssen stets zu bedenken, dass Hormonsekretionen immer auch von Erfahrungen und externen Stimuli moduliert werden.

Manifestierung des Geschlechts im Gehirn

Die Entwicklung des menschlichen Gehirns beginnt pränatal und dauert bis ins hohe Alter an. Dabei fußt sie auf komplexen Mechanismen. Zahlreiche Befunde weisen auf eine hormonelle Beeinflussung von Sexualhormonen im Mutterleib hin. Von Ovarien und Hoden produzierte Hormone maskulinisieren oder feminisieren die Gehirnentwicklung. Damit besteht die Möglichkeit eines Einflusses von Hormonen auf geschlechtstypisches Verhalten und Neuroanatomie.
Wichtig ist hierbei, dass eine Rückwirkung von Umwelterfahrungen auf die Hirnanatomie inzwischen nachgewiesen ist (Roffman, Marci, Glick, Dougherty, & Rauch, 2005). Neuronale Strukturen sind also Resultat genetischer, hormoneller und Umwelteinflüsse.
Die Untersuchung neuroanatomischer Unterschiede folgt keineswegs einem deterministischen Ansatz, wie oft vermutet. So werden Erkenntnisse über statistische Mittelungstechniken und meist aus eher kleinen Stichproben gewonnen.
Wenn im Folgenden die Rede von geschlechtsspezifischen Diskrepanzen ist, so muss betont werden, dass die Ähnlichkeiten in den Gehirnen von Männern und Frauen größer sind als die Unterschiede.
Sollten sich die Gehirne von Männern und Frauen unterscheiden, stellt sich die Frage, in welcher Weise und welchem Umfang, und vor allem: Wie wirken sich die Unterschiede aus?

Es wurde vermutet, dass die Gehirne von Frauen kleiner als Männergehirne sind. Da aber Gehirn und Körpergröße korrelieren, muss die Körpergröße kontrolliert werden. Je nachdem, welche Methoden zur statistischen Kontrolle verwendet werden, bleiben die Resultate gleich, verringern sich oder heben sich sogar ganz auf. Aber selbst wenn das Ergebnis beibehalten wird, lässt Größe nicht gleichsam auf Leistung schließen. Wie bereits erwähnt, ist die Intelligenz bei Männern und Frauen vergleichbar.
Bei näherer Betrachtung stellt sich aber nicht die Gehirngröße als wichtig heraus, sondern die Dichte der Neuronen und der Grad der Vernetzung. Im Alter wird im Übrigen ein stärkerer und früherer Verlust an Gehirnmasse bei Männern beobachtet, als Erklärung ein schützender Einfluss von Östrogenen vermutet.
Abgesehen von der Gehirnmasse konnten noch weitere neuroanatomische Ungleichheiten zwischen Männern und Frauen nachgewiesen werden.

Diverse Unterschiede betreffen spezifische Hirnareale (z.B. den Hippocampus, Hypothalamus und den so genannten sexuell-dimorphe Nucleus (SDN) des präoptischen Areals (POA)).
Hinzu kommt, dass metabolische Aktivitäten im Gehirn je nach Geschlecht anders verteilt zu sein scheinen, worauf Geschlechterunterschiede im EEG (Elektroenzephalogramm) hinweisen. Die zerebrale Durchblutung scheint bei Frauen höher als bei Männern zu sein, selbst bei identischen Aufgaben.

Der wahrscheinlich größte neuroanatomische Unterschied, der aber gleichzeitig in der Fachliteratur stark umstritten ist, betrifft die Lateralisation. Demnach sollen Frauen das Gehirn eher beidhemisphärisch, also symmetrischer nutzen, Männer hingegen asymmetrischer oder stärker lateralisiert (lateral = seitlich). Einige Befunde stützen diese Annahme.
Kongruent zu dieser Theorie konnte nachgewiesen werden, dass Testosteron die Ausprägung des Corpus Callosum beeinflusst. Hierbei handelt es sich um Nervenfasern, welche die Kommunikation zwischen beiden Hirnhälften herstellen. Frauen besitzen i. A. ein größeres und anders geformtes Corpus Callosum, allerdings scheint die Befundlage uneindeutig.
Weiterhin wirken unilaterale Hirnschädigungen je nach Geschlecht anders: Eine Schädigung der linken Hemisphäre bei Männern bewirkt Defizite bei verbalen Aufgaben, bei Frauen Defizite bei verbalen und räumlichen Aufgaben. Ist nur die rechte Hemisphäre betroffen, treten bei Männern Defizite bei visuell-räumlichen Aufgaben zutage, bei Frauen hingegen weder visuell-räumliche noch verbale Defizite (Kolb & Wishaw, 1993).

Das Gehirn von Frauen könnte also mehr bilateral organisiert sein, es erfolgt mehr Kommunikation zwischen beiden Hemisphären, während bei Männern jede Hemisphäre unabhängiger arbeitet, so dass unilaterale Schädigungen stärkere Auswirkungen zeigen.
Es gibt auch Hinweise darauf, dass bei den meisten Frauen die Sprache bilateral organisiert ist, d.h. die linke Hemisphäre steuert die Verarbeitung von Sprachprozessen, die rechte Hemisphäre die Verarbeitung von Sprache *und* visuell-räumlichen Leistungen. Bei den meisten Männern liegen die Fähigkeiten unilateral organisiert vor, also dominiert die rechte Hemisphäre für räumliche Verarbeitungsprozesse, die linke für verbale Anforderungen.
In welcher Weise erklärt eine solche Lateralisation unterschiedliche kognitive Leistungen bei Personen ohne Hirnschädigungen?
Dazu wird vermutet, dass visuell-räumliche Prozesse empfindlicher gegen andere, parallel arbeitenden Prozesse in derselben Hemisphäre sind. Frauen verwenden verbale Strategien bei räumlichen Problemen, die Sprachorganisation ist bilateral organisiert. Die Leistung in

räumlichen Tests ist aber verringert, da die Hirnstrukturen für räumliche Prozesse durch die Teilung mit verbalen Prozessen überfordert sind, Halpern nennt es „crowded out“. Für die Didaktik würde diese Theorie bedeuten, dass eine geringere Verbalisierung möglicherweise räumliche Leistungen bei Schülerinnen erhöhen könnte.
Eine hormonelle Beeinflussung der Lateralisation wurde bereits unter den hormonellen Erklärungsansätzen erwähnt: Durch pränatale Androgeneinwirkung auf die linke Hemisphäre entwickeln sich Asymmetrien des Gehirns. Es zeigt sich zudem, dass eine frühere Reifung in der Pubertät weniger Lateralisation zur Folge hat. Da Mädchen im Schnitt früher als Jungen reifen, ist bei ihnen die Lateralisation auch weniger stark ausgeprägt.
Dies würde aber auch folgenden Schluss implizieren: Frauen reifen immer früher in der modernen Gesellschaft. Wirkt sich diese frühere Reifung kontraproduktiv auf eine Nivellierung der Geschlechterunterschiede aus, weil Mädchen demnach immer schlechter in räumlichen Fähigkeiten werden müssten? Das Gegenteil scheint der Fall zu sein. Mädchen werden vermutlich besser in solchen Fähigkeiten, die Geschlechterunterschiede gehen zurück. Halpern gibt auch zu bedenken, dass eine stärkere Lateralisation und höhere räumliche Fähigkeiten nicht zwangsläufig zusammenhängen müssen, selbst wenn sie kovariieren.

Die Evolution der Geschlechterrollen – Fakt oder Mythos?

Evolutionäre Erklärungsansätze haben den Vorteil, dass sie sehr plausibel das Verhalten des heutigen Menschen im Licht des evolutionären Druckes erklären. Aber genau hierin liegt auch ihr Nachteil: Jedes Ergebnis kann im Licht der Evolution interpretiert werden, Beweise sind nur schwer zu führen, wenn sie den Menschen und vor allem die menschliche Psyche betreffen. Diskussionen um die Evolution des menschlichen Sehapparates werden nur selten kontrovers geführt. Geht es aber um menschliche Verhaltensweisen, wird stets befürchtet, dass eine evolutionäre Begründung eine Art Rechtfertigung enthält. Auf einige Argumentationsweisen, egal ob evolutionär oder psychosozial, trifft dies sicherlich zu. Dennoch muss ein evolutionärer Ansatz nicht zwangsläufig so benutzt werden.
Stellt sich heraus, dass es sich im Laufe der Evolution als günstig erwiesen hat, dass der Mann sich mit möglichst vielen Frauen verpaart und Kinder zeugt, so bedeutet dies nicht, Ehebruch in der modernen Gesellschaft zu rechtfertigen. An diesem Beispiel wird deutlich, dass Normen, Konventionen und andere gesellschaftliche Mechanismen das Verhalten moderieren. Wichtig zu betonen ist, dass evolutionäre (ebenso wie hormonelle oder soziologische) Ansätze mögliche Ursachen nur erklären und nicht rechtfertigen. In Kürze soll hier gezeigt

werden, in welcher Weise evolutionsbiologische Ansätze die geschlechtsspezifischen Unterschiede in kognitiven Leistungen erklären kann.

Es wird angenommen, dass der Mensch aus einer Jäger-Sammler-Gesellschaft hervorgegangen ist. Vor allem die optimal angepassten Individuen konnten ihre Gene weitergeben, so dass eine Auslese stattfand. Hierbei wird häufig irrtümlicherweise angenommen, dass sich vorteilhafte Eigenschaften zur Perfektion entwickelt hätten, was aber nicht zutrifft. Weitergegeben wurden jene Eigenschaften, die am wenigsten Nachteile mit sich brachten bzw. die entweder vorteilhaft für das Überleben und eine Weitergabe des Erbguts waren und/oder diejenigen, die keinen Nachteil dafür darstellten.

In der Forschung zu den Geschlechterunterschieden ist die Theorie verbreitet, dass Männer jagten, während Frauen fürs Gebären, die Kinderversorgung und für das Sammeln von Früchten zuständig waren. Kurzum: Männer und Frauen haben sich auf unterschiedliche Aufgaben spezialisiert.

So benötigten Männer für die Jagd eine gute räumliche Orientierungsfähigkeit, während sich Frauen vom Lagerplatz nicht weit entfernten. Zum Sammeln benötigten sie ggf. noch ein gutes Gedächtnis für Orte.

Diese Art von Erklärung ist verlockend, jedoch existieren nur wenige Belege dafür. Dass ausschließlich oder vorwiegend Männer Jäger und Ernährer waren, kann nicht als sicher belegt gelten. Die prähistorische Gesellschaftsform kann höchstens rekonstruiert werden. Bei heute zu beobachtenden Jäger-Sammler-Gesellschaften sind zwar geschlechtsspezifische Arbeitsteilungen zu beobachten, allerdings sind dort auch Frauen als Jägerinnen tätig und entfernen sich auch zur Kleinwildjagd vom Lagerplatz, beteiligen sich sogar an Treibjagden auf Großwild. Die Kinderaufzucht und -pflege wird häufig von älteren Sippenmitgliedern mit übernommen. Einige Wissenschaftler gehen sogar davon aus, dass es in der prähistorischen Gesellschaft sogar eine Gleichberechtigung von Mann und Frau gegeben hat (Weniger, 2003). Offenbar nicht eindeutig zu beantworten, wie groß der Einfluss der Evolution bei der Entwicklung von Geschlechterrollen zu beurteilen ist.

Natürlich kann auch gelten, dass es sich vielleicht einfach nur nicht als nachteilig erwiesen hat, dass gewisse kognitive Unterschiede zwischen den Geschlechtern bestehen. Selbst wenn sie per se keinen Vorteil brachten (und so keinen „Sinn“ darstellen), wirkten sie sich nur nicht nachteilig auf das Weitergeben der Gene aus.

An dieser Stelle sei noch darauf hingewiesen, dass trotz allem das Geschlecht im evolutionären Sinn *die* bedeutende Variable zur Fortpflanzung ist. Eine Kategorisierung des Gegenübers in männlich/weiblich ist eine Voraussetzung für sexuelle Attraktion. Eine solche

Dichotomisierung kann keine kognitiven Unterschiede erklären, aber aufzeigen, warum sich die Wichtigkeit dieser Kategorie bis in die moderne Gesellschaft tradiert hat.

Hinsichtlich der biologischen Erklärungsmechanismen ist zu beachten, dass komplexe Interaktionen zwischen Erfahrungen und biologischen Anlagen vorliegen: Wenn Jungen über leicht bessere visuell-räumliche Fähigkeiten verfügen, praktizieren diese auch mit großer Wahrscheinlichkeit öfter Tätigkeiten mit solchen Anforderungen (wie Ballspiele etc.) und werden von ihrer Umwelt weiterhin darin bestärkt. In welcher Weise psychosoziale Mechanismen wirken, wird im Folgenden erläutert.

1.2.2.2 Gender statt Sex – psychosoziale Betrachtungsweisen des Geschlechts

Psychosoziale Erklärungsansätze sehen kognitive Geschlechterunterschiede als Resultat sozialer, gesellschaftlicher Konstrukte. In diesem Zusammenhang wird deshalb oftmals von gender statt Geschlecht gesprochen.
In der englischen Sprache wird mit zwei Begriffen eine Unterscheidung zwischen der biologischen Geschlechterkategorie *(sex)* und dem gesellschaftlich, sozial und kulturell entwickelten Geschlecht *(gender)* angestrebt. Nach Stewart und McDermott (Stewart & McDermott, 2004) stellt *gender* ein gutes Instrument für die psychologische Forschung dar, speziell um Ergebnisse zu analysieren und zu modellieren.
Im Folgenden soll erläutert werden, wie kognitive Unterschiede mittels psychosozialer Ansätze zu erklärt werden können.

Das Geschlecht wird zu einem großen Teil - von seiner biologischen Bedeutung abgelöst - gesellschaftlich konstruiert und beeinflusst auf diese Weise menschliches Verhalten (und kognitive Leistungen). Geschlechterunterschiede im Sinne von gender sind allgegenwärtig. Die meisten geschlechtsstereotypen Praktiken sind uns dabei im Alltag nicht bewusst. Einige Fragen (an Halpern (2000) angelehnt und ergänzt) verdeutlichen Ungleichheiten, die mit gender zu tun haben:

- *Haben Sie sich schon einmal die Beine rasiert?*
- *Benutzen Sie farbigen Lippenstift?*
- *Lackieren Sie Ihre Fingernägel?*
- *Ziehen Sie gelegentlich hochhackige Schuhe an?*
- *Tragen Sie Röcke?*
- *Tragen Sie ihr Haar so lang, dass Sie einen Zopf flechten könnten?*
- *Wenn bei Ihnen zuhause gegrillt wird, stehen Sie am Grill?*
- *Interessieren Sie sich für Autos?*
- *Würden Sie sich schämen, wenn Sie in der Öffentlichkeit weinen müssten?*
- *Denken Sie, dass sie mit mehr Muskeln attraktiver wirken würden?*
- *Versuchen Sie, technische Geräte selbst zu reparieren?*

An den Fragen wird deutlich, dass Unterschiede zwischen den Geschlechtern sozial konstruiert sein können, ohne biologisch determiniert zu sein.

Verschiedene psychologische Theorien können die Geschlechterunterschiede aus jeweils anderer Perspektive erklären. So sehen Lerntheorien mittels differentieller oder stellvertretender Verstärkung den Ursprung von Diskrepanzen in Leistungen etwas anders gelagert als Erwartungswertmodelle oder Theorien kognitiver Schemata. Initiale biologisch unterschiedliche Neigungen werden unter dem Einfluss der Umwelterfahrungen intensiviert, Unterschiede vergrößert.

Geschlechtstypische Sozialisierung in Kindheit und Jugend

Die geschlechtstypische Sozialisierung beginnt in der frühen Kindheit und dauert bis ins hohe Alter an.

Bereits in den ersten Lebenstagen eines Kindes erfährt dieses eine unterschiedliche Behandlung je nach Geschlecht: Je nachdem, ob es als Junge oder Mädchen geboren wird, bekommt es einen anderen Namen (nur sehr wenige Vornamen sind für beide Geschlechter verwendbar), anders gefärbte Kleidung und später auch andere Spielzeuge.

Die soziale Umwelt stellt also frühzeitig unterschiedliche Bedingungen für Jungen und Mädchen her (siehe dazu auch Trautner, 1994). Es wäre sogar in den ersten Lebensjahren des Kindes undenkbar, das Geschlecht eines Kindes nicht zu berücksichtigen: Nachdem ein Kind geboren ist, lautet in den meisten Fällen die erste Frage: „Junge oder Mädchen?“, und nicht etwa „Groß oder klein?“ usw.

Bereits in frühem Alter werden Mädchen häufiger für Schönheit belohnt, was sich erst später auswirkt: Wenn Frauen mehr Zeit in solche Dinge wie Frisieren, Auswahl farblich abgestimmter Kleidung etc. investieren, fließt effektiv mehr Zeit in nicht kognitive bzw. nicht

(schul-) leistungsassoziierte Tätigkeiten (Halpern, 2000). Frauen geht Zeit verloren, die für andere Tätigkeiten genutzt werden könnte.

Untersuchungen mit Kleinkindern zeigen, dass bereits im Alter von drei bis fünf Jahren die meisten Kinder einem anderen Kind geschlechtstypische Spielzeuge schenken würden (Lobel & Menashri, 1993). Eltern reagieren je nach Geschlecht anders auf ihr Baby: Männlichen Säuglingen wird mehr körperliche Aufmerksamkeit gewidmet (wie z.B. auf den Arm nehmen), mit weiblichen Säuglingen wird mehr gesprochen (Stewart, 1976). Ob eine Studie jüngeren Datums die Befunde replizieren konnte, ist unklar. Im Übrigen stellten sich die Erwartungen der Eltern an die Kinder als guter Prädiktor für die Unterschiede in kognitiven Leistungen heraus (nach Stewart & McDermott, 2004): Erwartungen von Eltern beeinflussen das Verhalten der Kinder.

Bis zur Pubertät dürfen sich Jungen weiter vom Elternhaus entfernen als Mädchen. Welche Rolle der elterlichen Beeinflussung bei der Übernahme von Geschlechtsstereotypen zukommt, ist noch nicht geklärt. Mehr Bedeutung wird dagegen den peer groups beigemessen. Die in der Sozialpsychologie aufgedeckten Effekte von in-group und out-group sind im Jugendalter auf Gruppen von Jungen und Mädchen anwendbar. Während Gruppenmitglieder Normen erfüllen, wird die jeweilige out-group extremisiert, stilisiert und abgelehnt.

Jugendliche scheinen dabei eine Art geschlechtsspezifische Selbst-Segregation zu betreiben: Jugendliche haben zumeist gleichgeschlechtliche Freundeskreise und gehen unterschiedlichen Freizeitaktivitäten nach, Halpern (2000) nennt hier Videospiele und Billardspielen als Beispiele, beides eher Hobbys, denen Jungen nachgehen und in denen räumliche Fähigkeiten gefordert sind:

Heranwachsende Mädchen in den USA lesen und telefonieren mehr. Sie verrichten häufiger Haushaltstätigkeiten und konsumieren mehr Fernsehsendungen. Jungen engagieren sich sportlich mehr und spielen häufiger Computer- und Videospiele.

Im Durchschnitt bevorzugen Jungen Ballspiele stärker als Mädchen; generell erfordern jungenspezifische Spielzeuge und Aktivitäten mehr aktive Handlungen. Es ist dabei unklar, ob Jungen solche Alltagsspiele stärker bevorzugen, weil sie über bessere visuell-räumliche Fähigkeiten verfügen oder ob durch die Beschäftigung diese lediglich stärker geschult und dementsprechend präferiert werden. Derselbe unklare Zusammenhang gilt auch für jene Mädchen, die sich von sich aus mehr mit jungentypischen Spielen beschäftigen und die auch bessere räumliche Fähigkeiten aufweisen als solche Mädchen, die weniger jungentypischen Aktivitäten nachgehen.

Es konnte im Experiment gezeigt werden, dass bestimmte Videospiele räumliche Fähigkeiten trainieren (Subrahmanyam & Greenfield, 1994). Vom einem solchen Training profitierten Jungen und Mädchen zwar gleichermaßen, danach konnten aber noch immer geschlechtstypische Unterschiede beobachtet werden. Übung und angemessenes Feedback zu räumlichen Leistungen fördern also gleichermaßen Mädchen und Jungen, die Unterschiede bleiben aber erhalten. Durch unterschiedliche Förderung wird das Niveau, aber nicht der Unterschied verändert.

Es wurde bereits bei den biologischen Erklärungsansätzen auf den Zusammenhang zwischen dem Alter bei körperlicher Reifung und den kognitiven Leistungen eingegangen. In diesem Zusammenhang sind unterschiedliche Erfahrungen von Bedeutung. Früh reifende Jungen haben Vorteile in den meisten Sportarten, in denen es auf Körpergröße oder Kraft ankommt; durch ihre Körpergröße genießen sie unter Gleichaltrigen ein höheres Ansehen. Umgekehrt genießen körperlich früh gereifte Mädchen mehr geschlechtsspezifische Aufmerksamkeit, speziell von älteren Jungen.

Medien und Geschlecht

Beim Medienkonsum setzen sich Jugendliche vermehrt geschlechtstypischen Rollen aus (Allan & Coltrane, 1996). Männliche Rollen werden als aktiver, dominanter, aggressiver, autonomer und zielorientierter skizziert, stereotype Männer im Fernsehen arbeiten meist in angesehenen Berufen. Frauen hingegen werden viel öfter als passiv oder in stereotypen femininen Rollen gezeigt: Sie kümmern sich um Kinder oder Haushalt, kaufen ein oder kochen. Ergänzend könnte hier ein neuer Trend berichtet werden: Einige Filme, vor allem im Science Fiction-Genre (aber auch in anderen Sparten) haben starke Frauen als Protagonisten, die auch Männern körperlich überlegen sind. In manchen Fällen sind aber auch diese Rollen mit traditionellen Werten wie Kinderpflege verknüpft. Beispiele sind „Kill Bill“, „Aliens- Die Rückkehr“, „Resident Evil“, „Tomb Raider“, „Underworld“ - erfolgreiche Kinofilme mit starken weiblichen Protagonisten, die aber trotzdem weibliche Stereotype erfüllen. Obwohl sich diese Rollen bereits in Film, Fernsehen und Videospielen abgeschwächt haben, bleiben sie immer noch stark geschlechtstypisiert.

Ob sich nun Jugendliche hauptsächlich an gleichgeschlechtlichen Modellen in den Medien orientieren, konnte nicht zweifelsfrei nachgewiesen werden. Es bleibt aber die Frage, ob durch die stetige Aussetzung von Stereotypen diese nicht doch in irgendeiner Weise verstärkt werden.

Geschlechterrollen in der Schule

Auch der Schule wird eine richtungweisende Position bei der Vermittlung von Geschlechtsrollenstereotypen zugeschrieben. In einer Fülle von Studien ist belegt, dass Lehrer mit Jungen und Mädchen unterschiedlich umgehen. Jungen erhalten mehr Aufmerksamkeit und Lob, aber auch Tadel sowie detailliertere Rückmeldung über ihre Leistungen. Sie werden häufiger aufgerufen und dominieren Diskussionen (vgl. dazu Hoffmann, 2002; und Frasch & Wagner, 1982). Dieser Effekt scheint nach Frasch & Wagner etwas stärker bei männlichen Lehrern aufzutreten. Besonders ungünstig scheint der Umstand, dass dies sowohl für „weiblich" geltende Fächer, aber in noch stärkerem Maße für die als „männlich" klassifizierten Fächer gilt. Erstaunlicherweise wurden Parallelen dazu bei Studenten gefunden, wie die Autorinnen berichten: In Diskussionen wurden Beiträge von weiblichen Kommilitonen häufiger ignoriert oder überhört. Trotzdem erhalten Mädchen in der Schule im Durchschnitt bessere Noten als Jungen (z.B. Weinert & Helmke, 1997).

Eine Studie von Noble (1987) zeigt, dass Begabung oder Talent bei den Geschlechtern unterschiedlich „beliebt" sind. Von talentierten Jungen, talentierten Mädchen, durchschnittlichen Jungen und durchschnittlichen Mädchen sind talentierte Mädchen am wenigsten beliebt.

Der Zusammenhang zwischen interessenspezifischem Unterricht und geschlechtsspezifischer Physikleistung konnte direkt im Schulalltag nachgewiesen werden (Hoffmann, 2002; Hoffmann, Häußler, & Peters-Haft, 1997). So profitierten Jungen *und* Mädchen sogar noch nach einem Jahr von einem einführenden Physikunterricht, der stärker an Interessen der Schulkinder ausgerichtet war. Allerdings zeigten Mädchen beste Leistungen, wenn Sie beim Treatment monoedukativ, also nach Geschlechtern getrennt, unterrichtet wurden. Nur die Jungen und Mädchen, die in der Studie monoedukativ beschult wurden, verloren im Erhebungszeitraum kein Interesse am Fach Physik. Vor der Applikation der interessenskongruenten Physikunits berichteten Mädchen deutlich weniger Interesse an Physik als Jungen; je mehr Units davon unterrichtet wurden, desto stärker stieg das Interesse der Mädchen. Das Selbstkonzept für das Fach Physik beeinflusste dabei das Interesse der Mädchen stärker als das der Jungen.

Es zeigte sich weiterhin, dass die Physikleistung der Mädchen deutlich vom Umgang der Lehrer mit dem Verhalten der Jungen im Physikunterricht abhing: Die Leistungen der Mädchen sanken, wenn lautes und störendes Verhalten der Jungen toleriert wurde. Hoffmann sieht die Aufgabe eines geschlechterfairen Physikunterrichts darin, ein positives Selbstkonzept der Mädchen in Physik zu fördern, dies kann durch interessantes

Aufgabenmaterial für Mädchen und monoedukativen Unterricht gefördert werden. Da die Untersuchung nur an Gymnasien durchgeführt wurde, ist eine Übertragbarkeit auf andere Leistungsgruppen schwierig.

Einen monoedukativen Unterricht in Physik empfehlen aber auch Hannover & Kessels (2001). In ihrer Studie wurden Jungen und Mädchen an sieben Berliner Gesamtschulen zu Beginn des achten Schuljahres in folgende Experimentalgruppen eingeteilt: Monoedukativer Physikunterricht in einer Mädchengruppe sowie in einer Jungengruppe und koedukativer Unterricht.

Es zeigte sich nach nur einem Schuljahr experimenteller Beschulung im Fach Physik, dass durch monoedukativen Unterricht eine geschlechtsspezifische Polarisierung reduziert werden konnte. So zeigten sich Mädchen aus monoedukativ unterrichteten Gruppen in Physik motivierter und besaßen ein positiveres physikbezogenes Selbstkonzept als Mädchen aus koedukativ unterrichteten Gruppen; sie unterschieden sich statistisch nicht mehr von den männlichen Mitschülern.

Das Plus an Selbstkonzept und Motivation schlug sich auch im Kurswahlverhalten nieder: In reinen Mädchengruppen unterrichtete Mädchen wählten nicht nur häufiger als gemischtgeschlechtlich unterrichtete Mädchen einen Fortgeschrittenenkurs, sondern wurden auch nach Lehrerurteilen häufiger für einen solchen vorgeschlagen. Ein Effekt auf die Physiknoten konnte jedoch nicht nachgewiesen werden. Die Ursache liegt dabei vermutlich in der Notengebung selbst, da diese orientiert am sozialen Bezugsrahmen erfolgt.

Interessant sind die Ergebnisse bezüglich der monoedukativen Jungengruppen: Während keine steigernden Effekte auf Leistung, Selbstkonzept und Motivation zu beobachten waren, wählten monoedukativ unterrichtete Jungen seltener Fortgeschrittenenkurse als koedukativ unterrichtete. Ein nach Geschlechtern getrennter Unterricht hatte also zur Folge, dass Mädchen sich weniger stark auf mädchenspezifische und Jungen sich weniger stark auf jungenspezifische Fächer verteilten. Dies ist insofern bemerkenswert, als dass durch die Trennung das Geschlecht als segregierendes Merkmal eher betont wird und dies trotzdem zu einer Vernachlässigung genau dieses Merkmals bei den Schülern führt. Hannover (1992) und Kessels (2002; 2004) begründen das damit, dass in geschlechtshomogenen Klassen weniger stark die eigene Geschlechtsidentität als Rollenbild aktiviert wird, das wiederum mit bestimmten Fächern konform oder nonkonform geht. So wecken Mathematik und naturwissenschaftliche Fächer die Assoziationen von Männlichkeit. In einer koedukativen Klasse wird durch die Anwesenheit der Jungen bei Mädchen stärker das eigene Geschlecht

aktiviert, das dann dem Rollenbild der mathematisch-naturwissenschaftlichen Fächer widerspricht.

Es ist zu beachten, dass die Effektstärken der erzielten Veränderungen nur kleine bis mittlere Ausmaße erreichen. Dennoch kann monoedukativer Unterricht offenbar wesentlich dazu beitragen, Geschlechterunterschiede in der Schule zu nivellieren, auch wenn nach außen hin der Öffentlichkeit durch Segregation die Geschlechtsvariable weiterhin als wichtig bewusst gemacht wird. Zu dieser Schlussfolgerung gibt es auch Gegenstimmen (Horstkemper, 2004).

Eine Studie mit Erwachsenen (Lount, Messé & Kerr, 2000) weist im Kontext einer Leistungsveränderung durch Aktivierung von Geschlechtsstereotypen darauf hin, dass sich Männer bei maskulin geltenden Tätigkeiten stärker anstrengen, wenn sie mit einer leistungsfähigen Frau zusammenarbeiten. Bei Frauen konnte jedoch kein so eindeutiges Muster gefunden werden. Bei einer „maskulinen Tätigkeit" zeigten sie unter Zusammenarbeit mit einem leistungsfähigen Mann entweder gleiche Motivation wie unter alleiniger Arbeit, strengten sich mehr an oder verloren sogar an Motivation. Sollte dies zutreffen, so müsste in der reinen Jungengruppe ein leichter Leistungsrückgang in Physik messbar gewesen sein. In der Tat wählten weniger monoedukativ beschulte Jungen einen Fortgeschrittenenkurs.

Einen reflektierenden Beitrag über die Vor- und Nachteile monoedukativen Unterrichts, v.a. mit Bezug auf Mädchenschulen, liefern Herwartz-Emden, Schurt & Waburg (2005). Die Autorinnen zeigen, dass sich ein einheitliches Fazit zu Mädchenschulen nicht ziehen lässt.

Interessanterweise zeigte sich in einer Befragung von Schülern von Faulstich-Wieland & Horstkemper (1992) der ausgeprägte Wunsch nach Koedukation bei immerhin fast 70% der Schüler. Von den positiven Aspekten monoedukativen Unterrichts zeigten sich, wenn überhaupt, mehr Mädchen als Jungen überzeugt. Ähnliche Häufigkeiten berichten auch Herwartz-Emden et al. (2005).

Ob das schulische Umfeld insgesamt eher Mädchen oder Jungen fördert, ist eine Frage, zu der es keine klare Antwort gibt.

Neben deutlichen Anzeichen für ein an den Interessen von Jungen orientiertes Curriculum, lassen sich Verhaltensweisen von Lehrern nachweisen, die Jungen einen Vorteil versprechen (Hoffmann, 2002). Aber es zeigt sich auch, dass Mädchen bessere Noten erhalten und bei gleicher Leistung öfter eine Übergangsempfehlung für das Gymnasium erhalten als Jungen (Schnitzer, Isserstedt, Müßig-Trapp & Schreiber, 1998). Buschmann (1994) weist darauf hin, dass durch die Geschlechterrollenpolarisierung in Gesellschaft und Pädagogik neben der Mädchenproblematik auch eine ausgeprägte Benachteiligung für Jungen existiert, siehe dazu

auch Hanna (2003): Jungen sind stärker unter den Sitzenbleibern und den Sonder-/Förderschülern vertreten.

Horstkemper (2002) zieht folgendes Fazit: Mädchen und Frauen werden im Bildungswesen nicht mehr aktiv benachteiligt. Vielmehr hat eine geschlechtstypische Interessensausbildung zur Folge, dass Frauen weniger zukunftsträchtige und hoch bezahlte Berufe und Studienfächer wählen. Dadurch sind Männer noch immer in solchen Berufen überrepräsentiert. Darauf wird nachfolgend noch einmal eingegangen.

Unterschiede in Mathematik können also folgendermaßen erklärt werden: Frauen sind gegenüber Mathematik negativer eingestellt, sie schätzen Mathematik als weniger wichtig für ihre beruflichen Ziele ein, sie besitzen weniger Vertrauen in ihre mathematischen Fähigkeiten und erhalten weniger positive Verstärkung, Unterstützung und Lehrerfeedback bei der Beschäftigung mit Mathematik. Mathematik besitzt eher einen männlichen (also unfemininen) Ruf und so entscheiden sich Mädchen auch weniger für Mathematik-Leistungskurse. Diskrepanzen in der Motivation zu Mathematik lassen sich sogar bei gleichen mathematischen Leistungen beobachten (Zimmer et al., 2004). Es ist also durchaus erklärbar, wenn Frauen/Mädchen schlechter in manchen mathematischen Tests abschneiden. Ein ähnliches Modell lässt sich für verbale Fähigkeiten konstruieren.

Geschlechterkonstruktion im Erwachsenenalter

Unterschiede zwischen den Geschlechtern dauern auch im Erwachsenenalter an, dies zeigt sich u.a. im Bereich der Berufswahl. Auch wenn Frauen und Männer im Verlauf der Zeit immer mehr Gleichberechtigung erfahren, gibt es einige Hinweise darauf, dass die Gleichbehandlung in einigen Berufsfeldern noch nicht vollständig erreicht worden ist.

Z.B. sind im Top-Management von Großunternehmen nur 3% Frauen beschäftigt (Krell, 1999).

Während der Anteil der Geschlechter in Jura bzw. den Wirtschaftswissenschaften in etwa gleich ist, beträgt der Anteil weiblicher Studierender in den Ingenieurswissenschaften unter 30% (Schnitzer et al., 1998). Beim Studium der Geisteswissenschaften an deutschen Hochschulen liegt der Frauenanteil bei 68%, Männer sind dort also unterrepräsentiert (Georg & Bargel, 2001).

In den USA (und auch in Deutschland) schließen deutlich weniger Frauen als Physikerinnen ab. In Russland ist der Beruf des Physikers mit wenig Prestige verbunden und wird dort (deshalb?) hauptsächlich von Frauen ausgeübt (Halpern, 2000).

In der Tendenz üben Frauen in wissenschaftlichen Bereichen mehr untergeordnete Tätigkeiten oder Dienstleistungen für Betreuer und/oder Vorgesetzte aus (beispielsweise Protokoll führen). Die höhere Bereitschaft zu Kooperation und Teamarbeit, auch die manchmal mangelnde Fähigkeit von Frauen, sich im Wissenschaftsbetrieb wie männliche Kollegen durchzusetzen, tragen dazu bei (Mohr, 1987). Auch Breitenbach (1994) sieht das Problem, dass die kommunikativen und sozialen Fähigkeiten die Benachteiligung von Frauen eher festigen.

Der Anteil habilitierender Frauen hat sich in den Jahren 1988-1998 zwar fast verdoppelt, er betrug aber 1998 über alle Fächer gerechnet trotzdem nur gute 15%. In den Sprach- und Kulturwissenschaften liegt der Anteil bei fast einem Drittel, in den anderen Fächern deutlich darunter (Horstkemper, 2002).

Berufstätige Mütter sind noch immer anderen Erwartungen und Abwertungen ausgesetzt als berufstätige Väter, selbst wenn die Kinder aus dem Säuglingsalter heraus sind und die Kinderpflege de facto von beiden Geschlechtern übernommen werden kann: Mitte der 1990er Jahre nahmen in Deutschland weniger als 2% aller Männer Erziehungsurlaub wahr (Krell, 1999).

Bestimmte berufliche Tätigkeiten sind demnach geschlechtsspezifisch besetzt. Frauen sind weniger in hoch bezahlten, angesehenen Berufen tätig.

Dabei scheint es so zu sein, dass Männer in ihrer Berufswahl eingeschränkter sind als Frauen: Frauen können eher in Männerdomänen arbeiten als Männer in typischen Frauenberufen, welche dann mehr Abwertungen erfahren.

Halpern (2000) erklärt es damit, dass Frauenberufe wie Sekretärin, Krankenschwester etc. mit weiblichen Attributen besetzt sind, die wiederum eine Abwertung darstellen, wohingegen traditionelle Männerberufe wie Arzt, Rechtsanwalt, Physiker mit mehr Prestige verbunden sind. Ein Wechsel in das „gegengeschlechtliche“ Berufsfeld würde also für Männer einen Abstieg im Prestige, für Frauen einen Aufstieg bedeuten.

Als Ursache für geschlechtsspezifische Berufswahlen können weniger die Fähigkeiten per se herangezogen werden, sondern mehr die geschlechtstypischen Unterschiede in Motivation, persönlichen Zielen, Werten und Normen.

Viele Berufe erfordern Eigenschaften, die eher Männern zugeschrieben werden, wie ein gewisses Potenzial an Durchsetzungsfähigkeit, Dominanz oder Zielstrebigkeit. Entweder wird dies Frauen nicht zugetraut oder Frauen verfolgen andere Ziele bzw. leben solche Wertanschauungen, die ihnen seit frühester Kindheit vermittelt wurden.

In diesem Zusammenhang scheint folgender Befund von Bedeutung: Es ist nachgewiesen, dass Frauen länger leben, sich mehr um ihre eigene Gesundheit kümmern (Klotz, 1998). Möglicherweise ist auch genau das der Grund, warum sich Frauen häufig nicht für „harte" Männerberufe entscheiden, die mit 80 Arbeitsstunden pro Woche, starkem psychischen Druck und evtl. gesundheitlichen Beeinträchtigungen verknüpft sind. Verbunden mit der Frage, ob solche hoch angesehenen Berufe überhaupt erstrebenswert sind, wenn sie einen hohen körperlichen und psychischen Tribut fordern, drängt sich die Vermutung auf, dass Frauen evtl. weniger benachteiligt werden, sondern sich selbst benachteiligen, weil sie sich seltener für solche berufliche Laufbahnen entscheiden. Möglicherweise entscheiden sie sich aufgrund von Erziehung, Rollen und Normen sogar für gesündere Lebenseinstellungen, in denen Privatleben und Freizeit stärker geschätzt werden. Es ist eine Frage der persönlichen Wertigkeit, warum denn eine solche Orientierung weniger Ansehen genießt als eine Arbeitsstelle mit hohem Gehalt, eine steile berufliche Karriere und ob nun Frauen oder Männer die „klügere" Entscheidung treffen.
Wichtig ist hierbei, dass Männern und Frauen die gleichen Rechte zustehen, sich für den Beruf zu entscheiden, den sie ausüben möchten. Wenn aber eine Beeinflussung über Werte und Normen bereits im frühen Kindesalter die Interessen beeinflusst und darüber Fähigkeiten formt, ist die Frage der Zugangsberechtigung notwendig, aber nicht hinreichend.

Stereotype und geschlechtsspezifisches Verhalten

Das Wissen um Stereotype interagiert mit dem Wissen um das eigene Geschlecht. Eine nach Geschlecht unterschiedliche Nutzung kognitiver Strategien kann dabei in der Sozialisation begründet sein. Es ist belegt, dass Stereotypen Einfluss auf die Geschlechtsidentität und daran gekoppeltes Verhalten nehmen. Dies trifft zu, wenn sie aktiviert werden, aber auch wenn sie nicht bewusst gemacht werden und sogar dann, wenn sie nicht für wahr oder richtig gehalten werden. Fiedler definiert ein Stereotyp als „die erwartete Korrelation zwischen bestimmten Eigenschaften und Gruppenmitgliedschaft" (Fiedler, 1996; S.162).
Obwohl häufig angenommen, sind Geschlechtsrollen und Geschlechterstereotype selten binär, sondern vielfältig abgestuft und überlappen sich. Auf verschiedenste Weise nehmen sie Einfluss auf Erwartungen, Emotion, Motivation und Verhalten. Dabei wird eher ein wahrscheinlichkeitsbasierter Zusammenhang angenommen: In einer Studie von Deaux (1984) werden Frauen mit einer Wahrscheinlichkeit von 0,58 mit Unabhängigkeit und von 0,64 mit Wetteifer assoziiert, Männer hingegen mit einer Wahrscheinlichkeit von 0,78 bzw. 0,84, obwohl es sich um Individuen handelt, die man im Einzelfall gar nicht kennt. Es ging also um

„die Frau“ oder *„den Mann“*. Die Annahme eines Wahrscheinlichkeitszusammenhangs wird in Abschnitt 3.5 von Bedeutung sein.

Im Rahmen einer Studie von Intons-Peterson (1988) wird von großen Schnittmengen von zugeschriebenen Eigenschaften berichtetet, aber es gibt auch distinkte Assoziationen: Während Männer eher als selbstbewusste Menschen gesehen werden, die besser unter Druck arbeiten, verbindet man mit Frauen eher einfühlsame Menschen, die verletzte Gefühle mildern wollen.

Fiedler (1996) zufolge vereinfachen Stereotype Wahrnehmung und Denken, und verändern– was noch wichtiger ist – auch das Verhalten. So werden in sozialen Interaktionen Erwartungen des Gegenübers eher bestätigt als ihnen zuwider gehandelt. Werden (Alltags-) Hypothesen über Stereotype geprüft, so besteht die Tendenz, diese bestätigt zu finden.

An dieser Stelle sei nochmals auf die Studie von Spencer, Steele und Quinn (1999) hingewiesen. Die gefundenen Diskrepanzen in den Mathematikleistungen zwischen männlichen und weiblichen College-Studenten verschwanden, als man den Probanden vor dem Test mitteilte, dass keine Geschlechterunterschiede erwartet werden, während sie ohne spezielle Instruktionen durchaus auftraten. Stereotype haben also die Tendenz, sich selbst zu bestätigen.

Zu den Bedingungen, unter denen Stereotype die Leistung beeinflussen, gehören folgende (Halpern, 2000): Das Stereotyp

1. muss für die Leistung relevant sein – es ist für die Mathematikleistung irrelevant, Buddhist zu sein.

2. muss der Person etwas bedeuten – es könnte z.B. einer Hebamme egal sein, dass Frauen in Mathematik schlechter abschneiden.

3. braucht nicht für zutreffend gehalten werden - allein die Annahme, dass viele an das Stereotyp glauben, kann ausreichen.

4. verringert die Leistung bei schwierigen Aufgaben stärker als bei leichten Aufgaben.

Die Auswirkungen der Bewusstwerdung von Stereotypen sind z.T. nachgewiesen, aber in manchen Studien nicht replizierbar gewesen. Einige Ergebnisse sollen nachfolgende vorgestellt werden:

Generell wird stereotypenkonformes Verhalten verstärkt, nonkonformes Verhalten (im Sinne der Lerntheorie) bestraft. Wie in der Lerntheorie nachgewiesen, beeinflussen Verstärkung und Bestrafung die Auftretenswahrscheinlichkeiten. Diese Mechanismen wirken auch auf das geschlechtstypisierte Verhalten im Sinne von Stereotypen.

Androgynie – das psychologische Geschlecht

Ein weiterer Erklärungsansatz betreffend *sex* und *gender* besteht in dem Konzept der (psychologischen) Androgynität (folgende Darstellungen dazu orientieren sich v.a. an Bock & Alfermann, 1999). Hierbei wird von der Auffassung der bipolaren Geschlechtsidentität Abstand genommen und stattdessen die geschlechtliche Identität mittels zweier unabhängiger Dimensionen erklärt. Da einem Individuum gleichzeitig maskuline wie feminine Eigenschaften zugeschrieben werden können, wird jeder Person auf den Skalen Maskulinität und Femininität jeweils ein Wert zugeordnet. Diskretisiert man die Bereiche dieser beiden Kontinua, so lassen sich daraus vier verschiedene Typen unabhängig vom biologischen Geschlecht konstruieren (z.B. Strauss & Möller, 1999):

1. ***Die Maskulinen*** besitzen eine hohe Ausprägung auf der maskulinen Skala und eine niedrige Ausprägung auf der femininen Skala.
2. ***Die Femininen*** besitzen eine niedrige Ausprägung auf der maskulinen Skala und eine hohe Ausprägung auf der femininen Skala.
3. ***Die Androgynen*** besitzen eine hohe Ausprägung auf beiden Skalen.
4. ***Die Undifferenzierten*** besitzen eine niedrige Ausprägung auf beiden Skalen.

In einigen Studien konnte die praktische Relevanz des Konzepts belegt werden: So zeigten geschlechtstypisierte Personen (also Maskuline und Feminine) bessere Leistungen in Aufgaben mit geschlechtstypischen Anforderungen, die Androgynen hingegen flexiblere Handlungsstrategien. Androgynität wird generell mit einer Fülle positiver Eigenschaften in Verbindung gebracht (Alfermann, Reigber & Turan, 1999): So sollen Androgyne über ein höheres Selbstwertgefühl und psychisches Wohlbefinden besitzen und über ein differenzierteres Verhaltensrepertoire verfügen. Damit sollen sie bessere zwischenmenschliche Beziehungen führen und sich in sozialen Situationen angemessener verhalten. Mit Androgynität soll auch ein höheres Ausmaß an Leistungsorientierung und Kreativität einhergehen. Im Grunde wird Androgynität als ein Ideal beschrieben (Krell, 1999). Fasst man Androgynität als neues psychologisches Geschlecht auf, können die üblichen Geschlechterrollen bei der Erklärung von Verhaltensmustern in den Hintergrund treten. Alfermann et al. (1999) geben als Beispiel an, dass sich Arbeitgeber für Stellenbewerber nicht aufgrund ihres Geschlechts entscheiden, sondern wegen der Erfüllung der Anforderungen der Stelle, die dann auf der Skala der Androgynität zu verorten sind, z.B. „sozial und kompetent mit Kunden umgehen“ (S.143).

Insgesamt erklärt das biologische Geschlecht für die Verarbeitung von Informationen weitaus weniger Varianz auf als das psychologische Geschlecht oder die Geschlechterrollenidentität (Hannover, 1999). Da das psychologische Geschlecht – anders als das biologische – stärker und leichter veränderbar ist, stellt Androgynie ein gutes Ziel bei erzieherischen oder politischen Maßnahmen dar.
Einziger Nachteil des Konzepts ist, dass Androgynität im engeren Sinne als Identität aufgefasst wird, also in der Regel über Selbst- anstatt über Fremdbeschreibungen erhoben wird, z.B. über das Bem Sex-Role-Inventory (BSRI; Bem, 1974).

In der Zusammenschau wird verständlich, dass sich Mädchen weniger für Mathematik interessieren (und demnach dort auch weniger Leistungen erbringen), dagegen bessere verbale Leistungen als Jungen zeigen. Dennoch können solche Erklärungsmuster die Geschlechterunterschiede nicht vollständig begründen. Erst eine Kombination aus biologischen und psychosozialen Faktoren ergibt ein vollständiges Bild. Dies wird im folgenden Abschnitt verdeutlicht.

1.2.2.3 Verknüpfung von Sex und Gender – Ein Psychobiosozialer Ansatz

Möchte man biologische Ursachen für kognitive Geschlechterunterschiede nachweisen, stößt man schnell an Grenzen, da jede Versuchsperson einer Fülle von Umwelterfahrungen ausgesetzt ist, die sich auf das Ergebnis auswirken können. Es fällt mitunter aber ebenso schwer, psychosoziale Konstruktionen von Geschlecht abgelöst von den biologischen Unterschieden zu betrachten.
Menschen wachsen als biologische Lebewesen in einer kulturspezifischen Umwelt auf. Das biologische und das psychosoziale Geschlecht – also sex und gender – sind untrennbar miteinander verflochten und meistens im Versuchsdesign konfundiert. Den unabhängigen Effekt der einzelnen Komponenten transparent zu machen erscheint wenig aussichtsreich. Der psychobiosoziale Ansatz greift genau diese Problematik auf. Ihm zufolge liegt die Schwierigkeit gerade in der Dichotomie von Anlage und Umwelt, von biologisch und psychosozial. Diese stellt eine künstliche Einteilung dar, da nicht entschieden werden kann, wo biologische Ursachen enden und psychosoziale Faktoren beginnen. Daher wird im psychobiosozialen Ansatz auf diese Unterteilung bewusst verzichtet (Halpern, 2000). Anlage und Umwelt werden stattdessen als kombinierte Effekte gesehen, die sich gegenseitig verstärken.

Diese Überlegung fußt auf Studien, in denen nachgewiesen werden konnte, dass sich die Erfahrung aktiv modellierend auf das Gehirn auswirkt. Die Auswirkung kognitiver Verhaltenstherapie lässt sich z. B. mit Hilfe bildgebender Verfahren objektiv nachweisen (Roffman et al., 2005). Somit wurde gezeigt, dass sich eine Änderung im Verhalten aktiv im Gehirn niederschlägt. Es wirkt demnach nicht nur die Hirnanatomie auf das Verhalten, sondern auch der umgekehrte Weg ist nachgewiesen.

Der psychobiosoziale Ansatz besitzt den Vorteil, dass keine Forschungsrichtung zur Erklärung diskrepanter Leistungen von Jungen und Mädchen bzw. Männern und Frauen unberücksichtigt bleiben muss.

1.2.3 Wohin geht die Geschlechterforschung?

Wie oben deutlich gemacht, ist die Geschlechterforschung noch immer von aktuellem Interesse. Neben verschiedenen Diskussionen über die Bedeutung von tatsächlichen Diskrepanzen zwischen Mädchen und Jungen oder Frauen und Männern existieren verschiedene Erklärungsansätze. Es stellt sich die Frage, in welche Richtung die Forschung in diesem Bereich überhaupt gehen sollte. Vielleicht sollte davon abgesehen werden, immer detailliertere deskriptive Befunde erheben zu wollen und stattdessen eher auf die Herstellung von Geschlechterfairness zu fokussieren. Dies scheint dann allerdings weniger eine Aufgabe der Forschung zu sein.

Betrachtet man die Erklärungsansätze für geschlechtsspezifische Unterschiede in kognitiven Leistungen, so wird deutlich, dass es sich um einen Wahrscheinlichkeitszusammenhang handeln muss. Geschlechterunterschiede sind nur unter bestimmten Bedingungen, in speziellen Tests in einer gewissen Stichprobe zu beobachten, und auch dort sind die Zusammenhänge niemals deterministisch zu sehen. Für den Einzelfall gelten die Aussagen ohnehin nicht, sondern für den Durchschnitt, da sich die Verteilungen von Leistungen meist stark überlappen. Bei Leistungsvergleichen wird dazu noch oftmals die Einteilungen nach dem biologischen Geschlecht, also nach *sex* anstatt nach *gender* vorgenommen, wohl wissend, dass ein gewisser Anteil des jeweils anderen Geschlechts ähnlicher abschneiden wird als ein Teil des selben Geschlechts (siehe Abbildung 2).
Gerade im Hinblick auf die probabilistische Qualität der Geschlechterunterschiede scheint das eher ungünstig. Zielt man auf die Unterschiede im Sinne von *gender* ab, so sollten in ihren

Leistungen eher „feminine“ mit „maskulinen“ Jugendlichen verglichen werden. Auf diesem Wege sollten die Unterschiede verdeutlicht werden, da sich mehr Individuen in einer Gruppe befinden, die gleiche Leistungen erbringen.

Eine Möglichkeit dazu besteht darin, Geschlecht im Licht der so genannten Item-Response-Theory (IRT, vgl. Kapitel 2) zu untersuchen. Hierbei wird gerade von einem probabilistischen Zusammenhang ausgegangen, wie man ihn bei den Geschlechterunterschieden erwartet. Den Vorteilen, Möglichkeiten und Grenzen einer solchen Methode im Rahmen der Geschlechterforschung soll in der vorliegenden Arbeit nachgegangen werden.

Ein Beispiel der Analyse mit latenten Modellen bezüglich *gender* findet sich bei Strauss & Möller (1999). Allerdings wurde hier nicht die Leistung mit einbezogen, es ging vielmehr um die Erfassung des psychologischen Geschlechts selbst. Es stellt eine neue Herausforderung dar, Geschlechterunterschiede auf latenter Ebene mit den Modellen aus der Rasch-Familie zu untersuchen. Dafür scheinen insbesondere die so genannten large scale assessment - Studien eine gute Datenbasis zu liefern, da die dort erhobenen Daten große Vorteile bergen. Darauf soll im nächsten Abschnitt eingegangen werden.

1.3 Geschlechterforschung in large scale assessment Studien

PISA, TIMSS und PIRLS gehören zu den wichtigsten internationalen Schulleistungsstudien und finden beträchtliche Resonanz in der Öffentlichkeit. Sie zählen zu den so genannten large scale assessment - Studien, definiert als Tests, die an einer großen Anzahl von Personen appliziert werden, um administrativen Instanzen Entscheidungshilfen bereit zustellen (z.B. DePascale, 2005).

Dabei machen internationale Komitees und Qualitätsüberwachungsinstanzen internationale Vergleiche möglich. Gerade diese internationale Konzeption der Studien bietet gute Möglichkeiten, Geschlechterunterschiede über Staats- und Kulturgrenzen hinweg zu analysieren. Umfangreiche repräsentativ gezogene Stichproben sichern weitreichende Interpretationen ab. Die parallele Erhebung vieler Inhaltsbereiche ermöglicht eine breite Erfassung von Diskrepanzen.

Der sicherlich größte Nachteil dieser Studien ist, dass eine Trennung in unabhängige und abhängige Variable kaum möglich ist und sie somit „nur“ den Status von Korrelationsstudien haben (vgl. zur Untersuchungsklassifikation z. B. Musahl & Schwennen, 2001). Kausalaussagen sind also nicht möglich. Ein Nachweis von Kausaleffekten kann nur in Laborexperimenten gelingen, dabei ist jedoch die externe Validität stark eingeschränkt.

Von der breiten Bevölkerung werden die gewonnenen Informationen von Schulleistungsstudien oftmals nur oberflächlich wahrgenommen oder außerhalb des Zusammenhangs interpretiert, dabei gehen jene über den reinen Wettbewerbscharakter hinaus, dessen Ruf ihnen oft anhaftet.

Biologische Ansätze sind für vergleichende Schulleistungsuntersuchungen nicht so wichtig, da es dort stets um eine bildungspolitische Diskussion geht. Wenn es biologische Ursachen gibt, dann werden diese mit hoher Wahrscheinlichkeit in jedem Land als gleich betrachtet und als durch kulturelle Erfahrungen bzw. verschiedene System der Beschulung modifiziert gesehen. Denn selbst bei biologischen Vorteilen kann sich unterschiedliche Förderung anders auswirken und genau um diese Förderung geht es bei dieser Art von Studien. Die Unterschiede zwischen den Geschlechtern sind in der Tat nicht in jedem Land in gleicher Weise ausgeprägt.

Aufbau, Ziele und Ergebnisse wichtiger internationaler Schulleistungsstudien (PISA, TIMSS, PIRLS/IGLU) im Rahmen der Geschlechterforschung werden im Folgenden vorgestellt, Befunde verglichen, um daran anknüpfend neue Analysen zu PISA-Daten vorzustellen (siehe zum Aufbau Abbildung 1 oben).

1.3.1 Projekt PISA

Um den Mitgliedsstaaten der OECD (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung) Indikatoren für ihr Bildungssystem bereit zu stellen, wurde eine zentrale internationale Schulleistungsstudie entwickelt: PISA (Programme for international student assessment). Im Folgenden wird ein kurzer Überblick über PISA gegeben, detaillierte Darstellungen finden sich sowohl in den nationalen wie internationalen Berichten bei (Baumert et al., 2001; OECD, 2000; 2003; 2004; Prenzel et al., 2004a). Da sich die empirischen Arbeiten in der vorliegenden Veröffentlichung auf die Datenbasis von PISA 2003 beziehen, wird (sofern nicht anders erwähnt) die Umsetzung von PISA 2003 geschildert.

Zielsetzung

Ziel von PISA ist eine standardisierte vergleichende Messung in allen Teilnehmerstaaten. Es handelt sich um eine politisch konzipierte und gestaltete Studie, die vor allem den Regierungen der Teilnehmerstaaten Entscheidungen zur Verbesserung der Bildungssysteme ermöglichen soll. Die Politikorientierung PISAs wird von diversen Quellen hervorgehoben (Baumert et al., 2001; OECD, 2000; 2003; 2004; Prenzel et al., 2004a). Bereits hier wird deutlich, dass die Zielsetzung eine andere ist als bei einer für rein wissenschaftliche Zwecke entwickelten Studie und sich so auch Umsetzung und Implikationen von einer solchen unterscheiden.

Koordination der Studie

Im Auftrag der OECD durchgeführt, liegt die internationale Projektkoordination beim Australian Council for Educational Research (ACER), federführend für das internationale PISA-Konsortium. Herr Prof. Dr. Prenzel fungiert in Deutschland auf nationaler Ebene als Projektmanager. Die deutsche Projektkoordination hat das Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN) in Kiel inne.

Zielgruppe

Wie groß die geographische Reichweite von PISA ist, verdeutlicht die OECD anhand des folgenden Rechenbeispiels: An PISA 2003 nahmen 41 Staaten teil, davon alle 30 OECD-Länder, insgesamt haben 49 Länder mindestens einmal an PISA teilgenommen und weitere elf sind bei PISA 2006 in die Erhebung eingeschlossen, was ungefähr ein Drittel der

fünfzehnjährigen Weltbevölkerung ausmacht. Getestet wurden allein in der internationalen Stichprobe, also ohne nationale Ergänzungen, über eine Viertelmillion Schüler.

Zielpopulation bei PISA sind die fünfzehnjährigen Jugendlichen in Bildungseinrichtungen. In diesem Alter befinden sie sich am Ende der Pflichtschulzeit und sollten über die wichtigsten Kompetenzen für den Übergang in das Erwachsenenleben verfügen.

Zu beachten ist hierbei, dass die Schüler unabhängig von Klassenstufen in die Studie eingeschlossen wurden. Das bedeutet, dass die Schüler in Ländern mit einem frühen Einschulungsalter zum Zeitpunkt des Testens bereits eine längere Zeit als in anderen Ländern beschult wurden. Ebenso spielen Zurückstellungen und Klassenwiederholungen bei einer Altersstichprobe eine große Rolle. Inhaltlich gesehen hat diese Stichprobe die Bedeutung, dass das Benchmarking zu einem bestimmten Lebensalter stattfindet, es geht also auch um den Umgang mit Lebenszeit und Bildung. Näheres zur Stichprobe und Stichprobenziehung finden sich unter 3.2 und im Technical Report der OECD (2005).

Auf nationaler Ebene wurde durch die Erweiterung der Stichprobe ein deutschlandinterner Ländervergleich ermöglicht. So wuchs die Anzahl der getesteten Schüler in Deutschland von 4660 Schülern der internationalen Stichprobe auf insgesamt 44.580 Schüler. Um Vergleiche in einem Klassenzusammenhang treffen zu können, wurden zusätzlich zur Altersstichprobe zwei ganze neunte Klassen pro Schule getestet. Um weitere Analysen durchführen zu können, wurde bei einer Anzahl von Schulen nach einem Jahr eine Messwiederholung in den zehnten Klassen durchgeführt.

Basiskompetenzen

Erhoben werden bei PISA die Basiskompetenzen in den Bereichen Lesen, Mathematik und Naturwissenschaften sowie fächerübergreifende Kompetenzen, die als Voraussetzungen nicht nur für den aktuellen und nachfolgenden Bildungsprozess, sondern auch für eine aktive Teilnahme am gesellschaftlichen Leben angenommen werden. Die Anforderungen sind hierbei gerade nicht zwangsläufig als am Lehrplan orientiert einzuschätzen, Näheres zur curricularen Validität führen Blum et al. (2004a) und Rost, Walter, Carstensen, Senkbeil, & Prenzel (2004) aus. Die Erhebung zahlreicher Hintergrundinformationen ergänzt die Messung der o.g. Basiskompetenzen.

Da für die Entwickler von PISA ein innovatives Konzept der Bildungsindikatoren im Vordergrund stand (OECD, 2004), wurden die verwendeten Erhebungsinstrumente durch internationale Expertengruppen entwickelt. Die Rahmenkonzepte sind sehr ausführlich bei der OECD (2003) zu finden.

Von entscheidender Bedeutung ist die konzeptionelle Auffassung der Kompetenzen als „literacy“. Dieser Begriff meint damit eine Grundbildung, die mehr für allgemeine Fähigkeiten steht, die für das Leben in der heutigen (westlichen Industrie-)Kultur und Gesellschaft als notwendig erachtet werden. Diese Auffassung erklärt auch, warum andere Fachbereiche aus dem Bildungssektor nicht erhoben werden. Als zentrale Schlüsselkompetenzen werden bei der OECD also die Kompetenzen im Lesen, in Mathematik und den Naturwissenschaften sowie in den so genannten cross-curricular competencies gesehen. Bei PISA 2003 stellt Problemlösen diese bereichsübergreifende Kompetenz dar.
Die Erhebung von Voraussetzungen für eine Teilnahme am gesellschaftlichen Leben steht dabei im Vordergrund.
Im Zusammenhang des Vergleichs zwischen Schülergruppen spielen bei PISA die Kompetenzstufen eine wichtige Rolle. Aufzufassen als Aufgabengruppen mit ansteigender Anforderung sind diese über Antwortwahrscheinlichkeiten definiert. Items höherer Kompetenzstufen lassen sich dabei schwerer bewältigen als solche auf niedrigeren Stufen. Eine Zuteilung der Aufgaben folgte dabei über die zu bewältigenden kognitiven Anforderungen der jeweiligen Stufe, der auch jeweils ein Punkte-Intervall auf der Skala zugewiesen ist. Ein Beispiel verdeutlicht die inhaltliche Interpretation:
Schüler der höchsten Kompetenzstufe in Mathematik (Kompetenzstufe 6, über 669 Kompetenzpunkte) sind in der Lage, anspruchsvolle mathematische Fähigkeiten flexibel auch auf neue komplexe Probleme anzuwenden und diese auch adäquat zu kommunizieren. Im Kontrast dazu sind Schüler auf der niedrigsten Kompetenzstufe (Stufe 1, zwischen 358 und 420 Kompetenzpunkten) lediglich dazu in der Lage, mathematische Routine-Operationen in klar umrissenen und gewohnten Aufgabenstellungen anzuwenden. Ein gewisser Anteil von Schülern bewältigt nicht einmal die Anforderungen der niedrigsten Kompetenzstufe, ihre mathematischen Fähigkeiten sind dann als „unter Stufe 1“ zu klassifizieren.
Wird ein Schüler einer bestimmten Kompetenzstufe zugeordnet, so bedeutet dies, dass er dieser Stufe zugehörige Items mit einer bestimmten Wahrscheinlichkeit löst. Aufgaben, die leichter sind und sich demnach am unteren Ende der Kompetenzstufe befinden, werden mit einer größeren Wahrscheinlichkeit gelöst als solche, die sich am oberen Ende befinden und demnach schwerer sind. Es unterscheiden sich aber auch Schüler auf derselben Kompetenzstufe hinsichtlich ihrer Antwortwahrscheinlichkeiten auf ein Item, wenn ihre Fähigkeit entweder eher am oberen oder am unteren Ende der Kompetenzstufe liegt.

Selbstverständlich ist ein Schüler auch mit einer gewissen (eher geringeren) Wahrscheinlichkeit in der Lage, Aufgaben höherer Stufen korrekt zu beantworten, dagegen steigt die Wahrscheinlichkeit für die Lösung von Items niedrigerer Kompetenzstufen stark an. Über die Kompetenzstufen werden die Schüler für die Interpretation von Extremgruppen inhaltlich einer Risikogruppe (auf oder unterhalb der ersten Kompetenzstufe) oder einer Spitzengruppe (höchste Kompetenzstufe) zugesprochen.

Zu beachten ist, dass bei jedem Inhaltsbereich eine andere Anzahl von Kompetenzstufen festgelegt ist und jeweils andere Anforderungen definiert sind.

Inhaltsbeschreibungen sowie Entwicklung und Bildung der Kompetenzstufen zeigt die OECD (2005).

Um welchen Preis die Basiskompetenzen erlangt werden, wird in Schulleistungsstudien vernachlässigt. So sollte auch interessieren, mithilfe welcher moralischer Grundprinzipien Schüler und Schulleiter welche Leistungen erzielen. In einer Kultur mit zunehmenden gesellschaftlichen Problemen wie mangelnder Toleranz, Aggressionspotenzial und Egoismus sollte Bildung nicht nur auf den zentralen Leistungsaspekt fokussieren, sondern auch Kompetenzen im moralischen Urteilen umfassen (z. B. Lind, 2003). Die moralische Urteilskompetenz als wichtige gesellschaftliche Fähigkeit wird weder von Politik noch der Gesellschaft üblicherweise als Bildungsziel anerkannt. Im Vorwort des nationalen PISA-Bandes (Prenzel et al., 2004a) benennt die Präsidentin der Ständigen Konferenz der Kultusminister der Länder Deutschlands Doris Ahnen es als Aufgabe, zu eruieren, wie gut die Schüler auf die Herausforderungen der Zukunft vorbereitet sind. Zu den Herausforderungen der Zukunft gehören aber ebenso die Leistungen im moralischen Urteilen und Handeln Diese könnten sich als wichtig für die Zukunft unserer Gesellschaft herausstellen. Aktuelle Bildungsstudien (darunter auch PISA) lassen solche Aspekte unberücksichtigt.

Erhebungswellen

Um den politischen Entscheidungsträgern Fortschritte und Entwicklungen in den von der OECD definierten Basiskompetenzen nachvollziehbar zu machen, ist PISA zyklisch angelegt, zunächst mit den Erhebungswellen 2000, 2003 und 2006. Dabei wird in jedem Zyklus ein anderer Schwerpunkt gesetzt, d.h. ausführlicher und umfangreicher untersucht. PISA 2000 fokussiert auf die Lesekompetenz, drei Jahre später ist bei PISA 2003 die mathematische Kompetenz von zentralem Interesse und PISA 2006 untersucht die Leistungen in den Naturwissenschaften genauer. PISA 2009 sieht wieder die Lesekompetenz im Zentrum der Studie (Prenzel, Drechsel, Carstensen, & Ramm, 2004b). Bei einer genaueren Erhebung einer

Leistungsdomäne werden verschiedene Subskalen als Ergänzung zum Gesamtwert ermittelt, da mehr Aufgaben zur Verfügung stehen. Bei PISA 2003 dienten die mathematischen Subskalen „Quantität“, „Raum und Form“, „Veränderung und Beziehungen“ sowie „Unsicherheit“ zur detaillierten Analyse der mathematischen Kompetenz.

Aufgabenmaterial und Durchführung der Studie

Um die internationale Vergleichbarkeit gewährleisten zu können, wurde eine Reihe von Standards entwickelt und deren Einhaltung überwacht. So beziehen sich die internationalen Richtlinien nicht nur auf Stichprobenziehung und Vorgabe standardisierten Testmaterials, sondern auch auf die Testdurchführung durch geschulte Testleiter. Der international vorgeschriebene Test wurde an einem Testtag jedem Schüler als paper & pencil - Test mit einem Nettoumfang von zwei Testzeitstunden vorgegeben. Durch ein Multi-Matrix-Testdesign konnte ein gesamter Umfang an Aufgaben von mehr als 6,5 Stunden erzielt werden, wovon bei PISA 2003 etwa 3,5 Stunden auf Mathematikaufgaben entfielen. Danach füllten die Schüler einen etwa 30-minütigen Fragebogen zu Hintergrundinformationen wie Motivation, Lernumfeld und persönlichen Angaben aus. Ein weiterer Fragebogen wurde von den Schulleitern der PISA-Testschulen bearbeitet, um Schulmerkmale erfassen zu können.

Nähere Angaben zu den technischen Grundlagen, z.B. zum Multi-Matrix-Design machen Carstensen, Knoll, Rost, & Prenzel (2004) und die OECD (2005). Die Aufgaben erfordern sowohl freie Antworten als auch Antworten in Form von Multiple Choice.

Weil viele Aufgaben über die verschiedenen Zyklen beibehalten werden, gibt es nur eine begrenzte Menge an freigegebenen Aufgaben, die zur Einsicht zur Verfügung stehen.

Da eine komplette Einsicht der Aufgagen demnach nicht möglich ist, können Kritiker der Studie nur wenig inhaltliche Kritik zu den Aufgaben anbringen, Außenstehende die Qualität der Aufgaben nur schwer beurteilen. Damit ist die übliche Transparenz von Forschung nicht wirklich gegeben.

Eine Auswahl freigegebener Items mitsamt ihren Lösungen findet sich in den oben zitierten Quellen oder im Internet unter http://pisa.ipn.uni-kiel.de/ sowie http://www.mpib-berlin.mpg.de/pisa/.

Während einwandfreie Übersetzungen in die Landessprachen streng kontrolliert wurden, bleibt das ungelöstes Problem die Vergleichbarkeit international unterschiedlicher Konnotationen von z. B. Hintergrundfragen. Gerade Fragen zu Interesse, Anstrengung, Emotion und Motivation sind nicht unbedingt vergleichbar. So kann eine gleiche Angabe Verschiedenes bedeuten oder mit verschiedenen Angaben das Gleiche gemeint sein. Gibt also

ein finnischer Schüler an, dass er sich nicht besonders für Lesen interessiert, kann dies immer noch mehr Interesse ausdrücken als die Antwort eines japanischen Schülers, der viel Interesse fürs Lesen angibt. Dies kann dann den Zusammenhang zwischen Interesse und Leseleistung verfälschen. Dennoch eröffnet ein Vergleich über verschiedene Staaten hinweg einen breiteren Blickwinkel auf verschiedene Bildungssysteme.
Neben den strengen internationalen Vorgaben bietet PISA eine Reihe von nationalen Zusatzoptionen, von denen in Deutschland intensiv Gebrauch gemacht wurde. So wurde hier bei PISA 2003 umfangreiches, z. T. stärker am Curriculum orientiertes nationales Testmaterial von Expertengruppen entwickelt, das den Schülern an einem zusätzlichem Testtag vorgelegt wurde. Das Instrumentarium wurde durch weitere Fragebögen ergänzt, zusätzlich zu Schülern und Schulleitern auch Eltern und Lehrer befragt.

Auswertungen und Ergebnisse

Die Auswertung der internationalen Daten erfolgte international standardisiert durch das ACER, das nationale Zusatzmaterial wurde vorwiegend am IPN ausgewertet.
Um Schülerleistungen vergleichbar zu machen, wurden die Daten mit Hilfe der Item-Response-Theory (IRT) skaliert. Auf diese Weise kann man sowohl Schülerleistungen wie auch Aufgabenschwierigkeiten genauer bestimmen. Allgemeines zur IRT findet sich in Abschnitt 2, die spezielle Anwendung innerhalb von PISA in den oben genannten technischen Veröffentlichungen.
Will man die Ergebnisse von PISA einordnen, so muss beachtet werden, dass Mittelwert und Standardabweichung nur auf die OECD-Staaten normiert sind. Bei der Definition der internationalen Kompetenzskalen wurden ein Mittelwert von 500 und eine Standardabweichung von 100 gewählt. Um die nationalen Kompetenzskalen kenntlich zu machen, wurden hier ein Mittelwert von 50 und eine Standardabweichung von 10 gewählt.
Innerhalb der internationalen Metrik wird vorgeschlagen, einen Unterschied von 40 Kompetenzpunkten zwischen Schülergruppen als Leistungszuwachs von ungefähr einem Schuljahr entsprechend zu interpretieren (Prenzel et al., 2004b).

Es ist deutlich geworden, dass sich PISA einigen der Kritikpunkte stellen kann, die zu Studien zu Geschlechterunterschieden angeführt werden: Einbezogen wurden internationale Stichproben mit verschiedenen Bildungssystemen in verschiedene Kulturen (allerdings auf Industriestaaten begrenzt) und umfasste eine große durch internationale Standards gesicherte repräsentative Stichprobe. Allerdings wird nur eine kleine Altersspanne zur Untersuchung

herangezogen. Es wird zu einem Zeitpunkt getestet, wo die geschlechtliche (auch sexuelle) Orientierung der Schüler eine große Rolle spielt, Erziehung und Stereotypen über einen langen Zeitraum wirksam werden konnten. Ergebnisse zu Geschlechterunterschieden im Rahmen von PISA werden im nächsten Abschnitt dargestellt.

1.3.1.1 PISA 2003 - Geschlechterunterschiede im internationalen Vergleich

Die umfangreichen Ergebnisse aus PISA 2003 sind bei der OECD (2004) sowie Prenzel et al. (2004a) zu finden, in diesem Zusammenhang wird nur auf die Ergebnisse zu den Geschlechterunterschieden eingegangen, um die Analysen in Kapitel 3 nachvollziehbar zu machen.

Bei PISA wurde nur das biologische Geschlecht erfasst, die subjektiven, selbst wahrgenommenen Facetten der Geschlechtszugehörigkeit im Sinne einer Identität (z.B. nach dem Ansatz der Androgynie) finden keine Berücksichtigung.

Eine Beschreibung eventueller geschlechtsspezifischer Defizite wie bei PISA ist für den pädagogischen Auftrag unerlässlich. Denn eine detaillierte Beschreibung kann aufzeigen, in welchem Bereich Förderungsbedarf besteht.

Die internationalen Leistungsdomänen bei PISA 2003 liefern ein detailliertes Bild der Geschlechterunterschiede der verschiedenen PISA-Teilnehmerstaaten (Burba & Rost, 2006; OECD, 2004; Zimmer et al., 2004). Es gelingt ihnen in unterschiedlicher Weise, dem politischen Bestreben, Geschlechterunterschiede zu minimieren, gerecht zu werden. Einige der hier vorgestellten Ergebnisse sind in Tabelle 2 als Übersicht gezeigt.

Kompetenzbereich Lesen

Die einzige Inhaltsdomäne, in der die Unterschiede zwischen Mädchen und Jungen über alle OECD-Staaten hinweg gleich gerichtet sind, ist die Lesekompetenz. Es gehören wesentlich mehr Jungen der Risikogruppe im Lesen an.

Die Effektstärke der mittleren Leseunterschiede variiert zwischen minimal 0,21 in Korea und maximal 0,56 in Finnland, stets zu Gunsten der Mädchen. In allen Staaten liegen also Jungen mit ihren Leseleistungen im Mittel hinter den Mädchen, es handelt sich hierbei nach der Einteilung von Cohen (1988) um kleine bis mittlere Effekte, siehe auch Tabelle 1.

Auffällig ist der Befund, dass es unter den OECD-Staaten also keinem einzigen Land gelingt, die Geschlechterunterschiede im Lesen zu beseitigen, in allen Staaten stets die Mädchen statistisch bedeutsam besser als die Jungen abschneiden. Hierbei muss zwar beachtet werden,

dass dieser Wert nur für den mittleren Durchschnitt gilt und sich die Verteilungen der Leseleistungen von Jungen und Mädchen noch immer stark überschneiden: Im OECD-Mittel liegt der Unterschied bei 34 Kompetenzpunkten, also einer drittel Standardabweichung, einer Effektstärke von 0,35.

Tabelle 2 Geschlechterunterschiede in den kognitiven Leistungsdomänen der OECD-Staaten, statistisch signifikante Differenzen sind fettgedruckt. Die Länder sind nach dem mittleren Kompetenzwert in Mathematik sortiert. Tabelle nach Zimmer et al. (2004). MW steht dabei für den Mittelwert, d für die Effektstärke des Unterschieds, J-M zeigt die Differenz zwischen der Leistung der Jungen (J) und der Mädchen (M). Ein positiver Wert indiziert höhere Leistungen von Jungen.

OECD-Staaten	Mathematik			Lesen			Naturwissenschaften			Problemlösen		
	MW	J - M	d	MW	J - M	d	MW	J - M	d	MW	J - M	d
Finnland	544	**7**	0.09	543	**-44**	-0.56	548	**-6**	-0.07	548	**-10**	-0.12
Korea	542	**23**	0.26	534	**-21**	-0.26	538	**18**	0.18	550	8	0.09
Niederlande	538	5	0.06	513	**-21**	-0.25	524	5	0.05	520	4	0.05
Japan	534	8	0.08	498	**-22**	-0.21	548	4	0.04	547	-2	-0.02
Kanada	532	**11**	0.13	528	**-32**	-0.36	519	**11**	0.11	529	0	0.01
Belgien	529	8	0.07	507	**-37**	-0.34	509	0	0.00	525	-3	-0.03
Schweiz	527	**17**	0.17	499	**-35**	-0.38	513	**10**	0.10	521	-2	-0.03
Australien	524	5	0.06	525	**-39**	-0.41	525	0	0.00	530	-6	-0.07
Neuseeland	523	**14**	0.15	522	**-28**	-0.27	521	**16**	0.15	533	-3	-0.03
Tsch. Rep.	516	**15**	0.16	489	**-31**	-0.33	523	6	0.06	516	7	0.07
Island	515	**-15**	-0.17	492	**-58**	-0.61	495	**-10**	-0.11	505	**-30**	-0.36
Dänemark	514	**17**	0.18	492	**-25**	-0.29	475	**17**	0.17	517	5	0.06
Frankreich	511	**9**	0.09	496	**-38**	-0.40	511	0	0.00	519	-1	-0.01
Schweden	509	**7**	0.07	514	**-37**	-0.39	506	5	0.05	509	**-10**	-0.11
Österreich	506	8	0.08	491	**-47**	-0.47	491	-3	-0.03	506	-3	-0.03
Deutschland	503	**9**	0.09	491	**-42**	-0.39	502	6	0.05	513	-6	-0.06
Irland	503	**15**	0.17	515	**-29**	-0.34	505	2	0.02	498	1	0.01
Slow. Rep.	498	**19**	0.20	469	**-33**	-0.36	495	**15**	0.15	492	7	0.07
Norwegen	495	6	0.07	500	**-49**	-0.49	484	2	0.02	490	**-8**	-0.09
Luxemburg	493	**17**	0.19	479	**-33**	-0.34	483	**13**	0.12	494	2	0.03
Polen	490	6	0.06	497	**-40**	-0.42	498	**7**	0.07	487	-1	-0.01
Ungarn	490	**8**	0.08	482	**-31**	-0.34	503	-1	-0.01	501	-4	-0.04
Spanien	485	**9**	0.10	481	**-39**	-0.42	487	4	0.04	482	-6	-0.06
USA	483	**6**	0.07	495	**-32**	-0.32	491	5	0.05	477	-1	-0.01
Portugal	466	**12**	0.14	478	**-36**	-0.40	468	**6**	0.07	470	0	0.00
Italien	466	**18**	0.19	476	**-39**	-0.40	486	6	0.05	470	-4	-0.04
Griechenland	445	**19**	0.21	472	**-37**	-0.36	481	**12**	0.12	449	2	0.02
Türkei	423	**15**	0.14	441	**-33**	-0.36	434	0	0.01	408	2	0.02
Mexiko	385	**11**	0.13	400	**-21**	-0.23	405	**9**	0.11	384	5	0.05
OECD-∅	500	**11**	0.11	494	**-34**	-0.35	500	**6**	0.06	500	-2	-0.01

Dennoch passt gerade die Einheitlichkeit über die Staaten hinweg zu den Befunden, wie sie in den Abschnitten weiter oben erläutert sind. Gerade bei starken Prädispositionen würde man erwarten, dass in allen Ländern die Unterschiede gleich gerichtet sind. Und dies ist auch

tatsächlich der Fall was die Lesekompetenz betrifft. Auf der anderen Seite kann ebenso argumentiert werden, dass in allen Ländern die gleiche Kultur hinsichtlich Geschlechterstereotype vorherrscht.

Wie unter 1.2.2.2 näher ausgeführt, wird durch unterschiedliche Förderung das Niveau, aber nicht der Unterschied an sich verändert: In allen anderen Ländern gilt derselbe Zusammenhang, jedoch ist das Niveau ein anderes. Während in einem Land Jungen zwar ebenfalls schlechter als Mädchen abschneiden, sind die Jungen in diesem Land (z.B. Finnland) im Durchschnitt noch immer besser als der Durchschnitt der Mädchen in einem anderen Land, das bei PISA schlechter abgeschnitten hat (z.B. Spanien).

Kompetenzbereich Mathematik

Im Bereich der Mathematik sind die Aussagen nicht mehr so eindeutig. In immerhin sieben der 31 OECD-Staaten sind im Mittel keine signifikanten Unterschiede mehr zwischen Mädchen und Jungen in ihrer mathematischen Kompetenz zu beobachten. In Island schneiden sogar die Mädchen besser als Jungen ab. In den übrigen 21 Staaten (darunter auch Deutschland) zeigen die Jungen bessere mathematische Leistungen, allerdings nur mit einer geringen Effektstärke. Ein Unterschied von 11 Punkten und einer Effektstärke von 0,11 stellt den Durchschnitt über alle OECD-Länder dar.

Im Gegensatz zu der Lesekompetenz sind also die Ergebnisse nicht mehr über alle Staaten homogen, in der Tendenz bestätigen sie aber dennoch die üblichen Geschlechterstereotype, lediglich Island fällt heraus. Dass die Unterschiede in vielen Ländern nicht signifikant sind, kann auch die Ursache darin haben, dass die mathematische Kompetenz in Subskalen untergliedert ist, die in einer Gesamtskala zusammengefasst sind. Je nach Subskala divergieren die Ergebnisse: So sind bei der mathematischen Subskala „Raum und Form“ in wesentlich mehr Ländern signifikante Unterschiede zu Gunsten der Jungen zu beobachten als in der Subskala „Quantitatives Denken“, bei der in wenigen Ländern überhaupt noch signifikante Unterschiede zu beobachten sind OECD (2004). Island bleibt aber auch auf Subskalenebene weiterhin das einzige Land, in dem Mädchen signifikant bessere Leistungen im mathematischen PISA-Test zeigen.

Die Befunde der Subskalen stützen die Ausführungen aus Abschnitt 1.2.1: Jungen zeigen tendenziell bessere Leistungen bei visuell-räumlichen Aufgaben. Dass in einigen wenigen Ländern keine signifikanten Unterschiede auf der Subskala „Raum und Form“ zu beobachten sind und in Island Mädchen sogar besser abschneiden, kann darauf hindeuten, dass es hier gelungen ist, gegen bestehende Stereotype und eventuelle biologisch-psychosoziale Einflüsse

zu arbeiten, Geschlechterunterschiede in diesem Bereich also nicht zwangsläufig auftreten müssen.

Kompetenzbereich Naturwissenschaften

Bei der naturwissenschaftlichen Kompetenz des internationalen Tests, zentral für Leistungen in einer von Naturwissenschaften geprägten Kultur, ist eine noch uneinheitlichere Bilanz zu verzeichnen. In etwa der Hälfte (16 von 31) aller OECD-Staaten gibt es keine statistisch bedeutsamen Unterschiede, dazu gehört auch Deutschland. In elf Staaten zeigen Jungen signifikant bessere Leistungen, in zweien gilt dies umgekehrt für Mädchen. Selbst die signifikanten Unterschiede bewegen sich alle im Bereich kleiner Effektstärken und liegen mit der Ausnahme von Island unter 0,1. Im OECD-Mittel liegen die Unterschiede nur noch bei 6 Punkten und einer Effektstärke von 0,06. Letztlich ist bei so kleinen, wenn auch z. T. noch signifikanten Effekten fraglich, ob diese Unterschiede noch praktisch relevant sind.

Auch wenn dieser noch immer nicht die Grenze zur Signifikanz überschreitet, zeigt sich im Vergleich zu PISA 2000 bei deutschen Schülern im internationalen Naturwissenschaftstest allerdings eine leichte Vergrößerung des Geschlechterunterschiedes, der mit einer generellen Leistungssteigerung einhergeht. Bei genaueren Analysen zeigt sich zudem, dass die Geschlechterunterschiede stark vom Leistungsniveau abhängig sind. Vergleicht man die Leistung von Jungen und Mädchen in den Leistungsquartilen, so zeigt sich dort das folgende Bild: Im unteren Quartil beträgt der Geschlechterunterschied nur einen Kompetenzpunkt, im obersten Quartil hingegen zeigen Jungen im Mittel eine um 14 Punkte bessere Leistung als Mädchen. Mit höherer Kompetenz gehen also größere Geschlechterunterschiede einher, allerdings liegen sie selbst im obersten Quartil zwischen einer und zwei zehntel Standardabweichungen.

Kompetenzbereich Problemlösen

Im Hinblick auf die Leistungen im Problemlösen, der bereichsübergreifenden Kompetenz bei PISA 2003, zeichnet sich ein insgesamt recht homogenes und geschlechterfaires Bild ab. In fast allen Staaten sind keine Unterschiede zwischen den Leistungen von Jungen und Mädchen auszumachen. In Island, Schweden und Norwegen zeigt sich die Problemlösekompetenz der Mädchen höher ausgeprägt als die der Jungen, und nur in Island wird annähernd ein mittlerer Effekt erreicht. Über alle OECD-Länder hinweg liegt der Unterschied bei nur zwei Punkten zu Gunsten der Mädchen mit einer Effektstärke von praktisch Null.

Da die Problemlösekompetenz das kognitive Potenzial eines Schülers anzeigt, das für die komplexen Problemlöseprozesse außerhalb von Unterricht und Schule, aber auch für die mathematische Kompetenz im PISA-Test benötigt wird (Leutner, Klieme, Meyer, & Wirth, 2004), kann eine Differenz zwischen den Kompetenzbereichen Problemlösen und Mathematik darauf hinweisen, dass das allgemeine kognitive Potenzial aus verschiedenen Gründen nicht ausgeschöpft werden kann.

In den meisten Staaten sind keine Geschlechterunterschiede im Problemlösen, zugleich aber statistisch bedeutsame Unterschiede in der Mathematikkompetenz zu Gunsten der Jungen zu verzeichnen. Das spricht dafür, dass in diesen Ländern das allgemeine kognitive Potenzial der Mädchen nicht hinreichend ausgeschöpft wird (siehe dazu Zimmer et al., 2004). Mädchen könnten also bessere Leistungen in Mathematik erzielen, wenn sie ausreichend darin gefördert würden. In Deutschland ist genau dieses Muster zu erkennen.

Zusammenhang von Geschlechterunterschieden und Leistungsniveau

Bei der Einordnung der Größe der Geschlechterunterschiede ist gleichzeitig zu berücksichtigen, auf welchem Niveau sich die Leistung der Schüler bewegt. Dass hohe Gesamtleistungen mit hohen Geschlechterunterschieden einhergehen, konnte bei PISA 2003 nur zum Teil festgestellt werden. Finnland und Korea bestätigen dieses Bild, während sich immerhin drei Staaten ausmachen lassen, die insgesamt überdurchschnittliche Leistungen in allen Kompetenzbereichen und gleichzeitig keine Geschlechterunterschiede in Mathematik, Naturwissenschaften und Problemlösen aufweisen.

Dies betrifft die Niederlande, Belgien und Australien. Jedoch zeigen auch in diesen Ländern die Unterschiede im Lesen kleine bis mittlere Effekte zu Gunsten der Mädchen. Lässt man die durchschnittlichen Leseleistungen unberücksichtigt, reiht sich Japan in diese Befundlage mit ein.

Es wird also deutlich, dass in einigen Ländern sowohl Unterschiede zwischen den Geschlechtern ausgeglichen und gleichzeitig ein hohes Leistungsniveau erreicht werden konnte. Eine Minimierung der geschlechtlichen Heterogenität muss beim Lesen allerdings auch hier erst noch gebahnt werden.

Die Betrachtung von Geschlechterunterschieden in Extremgruppen zeigen Zimmer et al. (2004). Teilt man die Schüler in Leistungsgruppen nach Kompetenzstufen ein, so zeigt jeweils der Anteil von Geschlechtern innerhalb der unteren bzw. oberen Kompetenzstufen, wie stark die Geschlechter überhaupt in diesen Leistungsgruppen vertreten sind. In Mathematik, den Naturwissenschaften und dem Problemlösen gibt es nur leichte

Anteilsunterschiede. Bei der Lesekompetenz passen die Anteilsverschiebungen zu den deutlichen Unterschieden in der Leistung: Auf den oberen Kompetenzstufen sind deutlich mehr Mädchen vertreten (ca. 16%), auf den unteren wesentlich mehr Jungen (ca. 12%).
Hinsichtlich der Leistungen ist ein Unterschied von neun Kompetenzpunkten zu Gunsten der Jungen in Mathematik in der Gruppe kompetenzstarker Jugendlicher zu beobachten, aber kein Unterschied in der Gruppe der Risikoschüler. Bei den Naturwissenschaften ist ein vergleichbares Bild erkennbar.
Im Lesen ist dieser Zusammenhang umgekehrt. Hier ist der Vorsprung der Mädchen von 17 Punkten in der unteren Leistungsgruppe zu beobachten, bei den kompetenteren Schülern ist er nicht mehr vorhanden. Beim Problemlösen zeigt sich innerhalb der Extremgruppen dieses speziellen Vergleichs kein Kompetenzunterschied zwischen den Geschlechtern.
Diese Befunde reihen sich in die weiter oben genannten Ergebnisse ein und zeigen eher geringe Unterschiede an.

Emotion und Motivation

Neben den kognitiven Leistungen spielen bei PISA auch motivationale und emotionale Aspekte eine große Rolle. Auch hier sind wesentliche Unterschiede zwischen Mädchen und Jungen beobachtbar.
In Bezug auf Mathematik schätzen Mädchen ihre Fähigkeit geringer als Jungen ein, diese Unterschiede sind auch in der Gruppe der Leistungsstarken ausgeprägt (Zimmer et al., 2004), wenn auch nicht mehr so deutlich wie in der gesamten Stichprobe. Sogar bei vergleichbarer Leistung differieren die Selbsteinschätzungen noch immer in geschlechtsstereotyper Weise.
In einer großen Anzahl von Staaten zeigt sich, dass Jungen stärker an Mathematik interessiert sind (OECD, 2004). Da Emotion und Motivation meist eng mit der Leistung gekoppelt sind (Pekrun et al., 2004; Pekrun & Zirngibl, 2004) wirken sie sich auch auf Kurs- und Berufswahl aus. Dabei sind noch immer Geschlechterunterschiede in stereotyper Richtung beobachtbar. Daher ist es nach wie vor eine zentrale gesellschaftspolitische Aufgabe, mit geeigneten Maßnahmen die Geschlechterfairness besonders im Bildungsbereich zu fördern.

1.3.1.2 PISA 2003 - Geschlechterunterschiede im nationalen Naturwissenschaftstest

Von den in Deutschland entwickelten nationalen Zusatztests zu PISA 2003 spielt bei den Analysen unter Kapitel 2 vor allem der Naturwissenschaftstest eine größere Rolle. Nachfolgend werden dessen Konzeption und Charakteristika näher beschrieben.

Appliziert wurde der nationale Naturwissenschaftstest am zweiten Testtag an der internationalen Stichprobe (216 Schulen, insgesamt 4660 Schüler).

Von einer Reihe nationaler Experten entwickelt, um stärker am deutschen Curriculum orientiert zu sein und die naturwissenschaftlichen Leistungen weiter zu differenzieren, bilden sieben so genannte kognitive Teilkompetenzen den Kern des Tests. Sie bedingen die Lösungswahrscheinlichkeiten der Aufgaben stärker als Fächer und Inhalte. Dabei repräsentieren sie kognitive Teilprozesse, die zur Lösung naturwissenschaftlicher Problemstellungen durchlaufen werden.

Dazu zählen nach Rost et al. (2004) der Umgang mit Grafiken und Zahlen, die Verwendung mentaler Modelle, konvergentes und divergentes Denken sowie die Verbalisierung von Sachverhalten und das Bewerten, siehe Abbildung 3. Bei Prenzel et al. (2004a) finden sich auch Aufgabenbeispiele dazu.

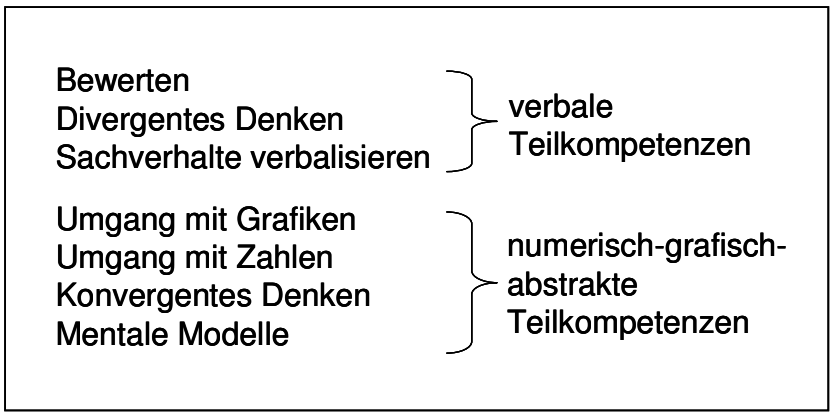

Abbildung 3 Kognitive Teilkompetenzen des nationalen Naturwissenschaftstests bei PISA 2003

Der große Vorteil des nationalen Naturwissenschaftstests besteht darin, dass diese sieben Teilkompetenzen mit allen neun vorgegebenen Inhalten der Aufgaben vollständig gekreuzt sind und der Test deshalb auch als Facettentest bezeichnet werden kann. Nur so kann ohne Konfundierung mit einem speziellen Thema der Aufgabe die Teilkompetenz methodisch sauber erfasst werden. Die Auswahl der vorgelegten Inhaltsbereiche ist in Abbildung 4 mit einer Zuordnung zu Schulfächern gezeigt.

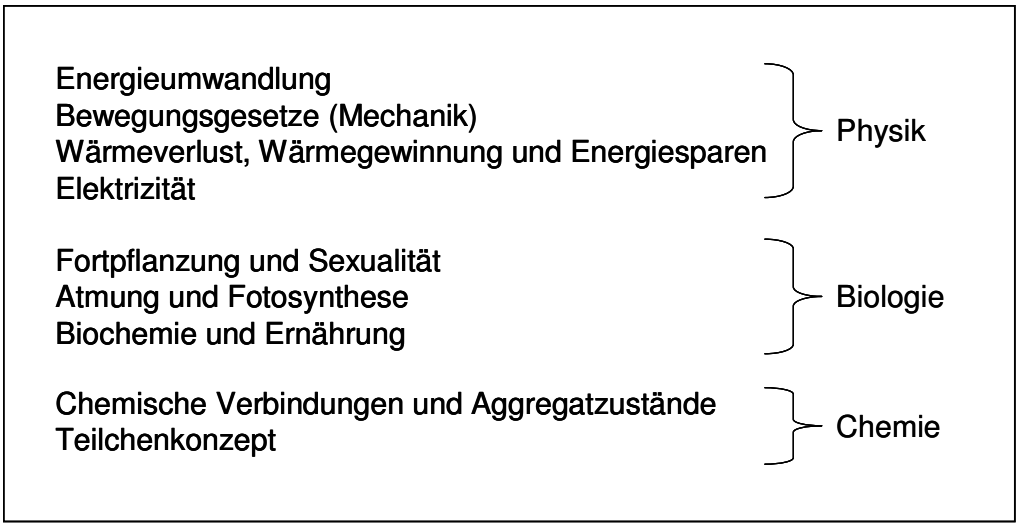

Abbildung 4 Im nationalen Naturwissenschaftstest von PISA 2003 vorgelegte Inhaltsbereiche mit zugeordneten Schulfächern

Pro Inhalt wurde jede der sieben Teilkompetenzen über genau ein Item erfasst, bei neun Inhalten ergibt das genau 63 Items für den nationalen Naturwissenschaftstest.

Ein Umgang mit Grafiken erfordert die Verarbeitung bildlich-grafischer Informationen, ein Umgang mit Zahlen und die Umsetzung numerisch repräsentierter naturwissenschaftlicher Informationen. Während es beim konvergenten Denken darauf ankommt, aus gegebenen Sachverhalten zu genau einer Lösung zu kommen, geht es beim divergenten Denken gerade darum, möglichst verschiedene Lösungsansätze für naturwissenschaftlicher Probleme zu entwickeln. Bei der Nutzung mentaler Modelle macht sich der Schüler räumlich-geometrische Symbole zu Nutze, um komplexere naturwissenschaftliche Sachverhalte kognitiv zu repräsentieren und so Informationen zusammenzufassen und zu reduzieren. Der Prozess des Bewertens umfasst das Heranziehen geeigneter Informationen zu Entscheidungen; dabei zählt nicht, welche Entscheidung getroffen wird, sondern die Auswahl und Darlegung der Begründungen. Unter dem Verbalisieren von Sachverhalten wird die Fähigkeit verstanden, naturwissenschaftliche Sachverhalte adäquat verbal auszudrücken. Für Beispiele und weiterführende Beschreibungen der kognitiven Teilkompetenzen sei hier auf Rost et al. (2004) verwiesen.

Bei der Beschreibung fällt auf, dass die Anforderungen in verbale und eher abstrakt-analytische Prozesse unterteilbar sind: Bewerten, divergentes Denken und Sachverhalte verbalisieren auf der einen Seite, konvergentes Denken, Umgang mit Grafiken, Umgang mit Zahlen und Mentale Modelle auf der anderen Seite, siehe Abbildung 3.

Bei einer Analyse der Aufgabenschwierigkeiten (Rost et al., 2004) zeigte sich zudem, dass die Aufgaben, die auf verbale Kompetenzen zugeschnitten sind, im Mittel leichter zu lösen waren

als die abstrakt-analytisch ausgerichteten Items. Hinsichtlich der Schulfächer waren keine wesentlichen Unterschiede in der Schwierigkeit zu verzeichnen.

Im Gesamtmittel des Tests gibt es keine nennenswerten Unterschiede zwischen Mädchen und Jungen, auch für die Schulfächer sind die Geschlechterunterschiede gering. So unterscheiden sich Mädchen und Jungen in ihren Leistungen in Physik und Chemie nicht signifikant und die signifikant besseren Leistungen von Mädchen in Biologie sind mit 1,35 Punkten, also etwas über einer zehntel Standardabweichung, gering. Die geringen Unterschiede zwischen den Fächern können auch als Folge des Facettendesign auftreten, da sich gerade spezifische kognitive Anforderungen, wie sie im Schulalltag mit bestimmten Schulfächern assoziiert sind, im Test nun auf alle Schulfächer verteilen.

Es zeigt sich allerdings, dass Jungen und Mädchen über ein spezielles Profil der kognitiven Teilkompetenzen verfügen, dargestellt in Abbildung 5.

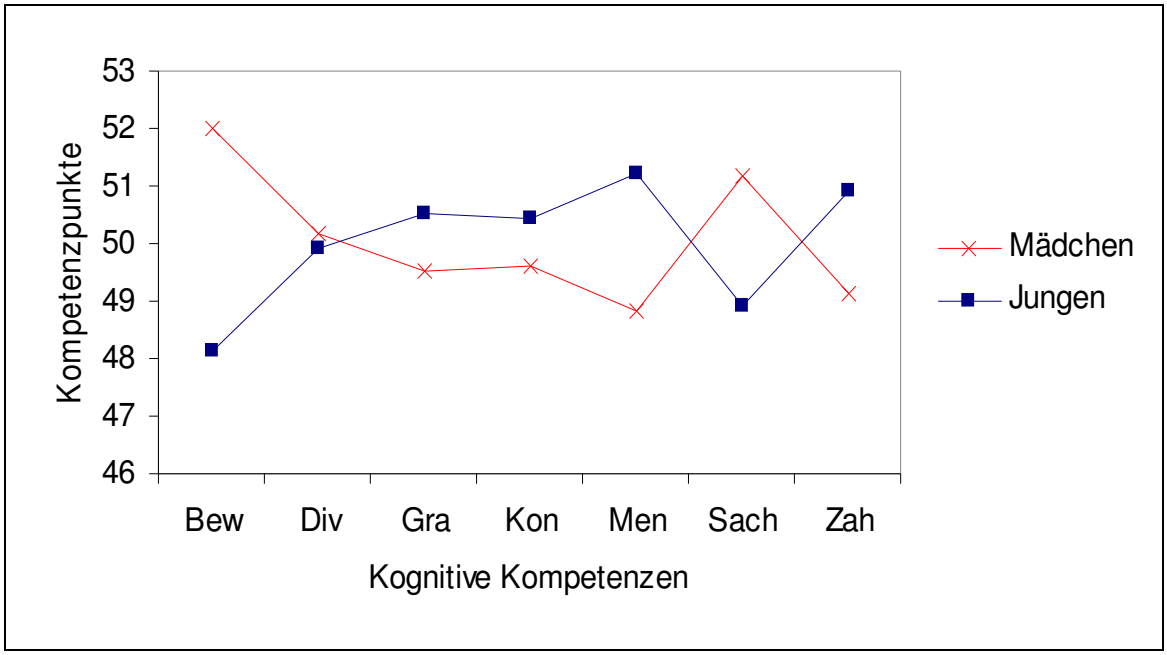

Abbildung 5 Geschlechterunterschiede in den kognitiven Teilkompetenzen des nationalen Naturwissenschaftstest bei PISA 2003, nach Rost et al. (2004). Bew = Bewerten, Div = Divergentes Denken, Gra = Umgang mit Grafiken, Kon = Konvergentes Denken, Men = Mentale Modelle, Sach = Sachverhalte verbalisieren, Zah = Umgang mit Zahlen

Demnach zeigen sich im Durchschnitt Mädchen besser den verbalen Teilkompetenzen gewachsen, Jungen den numerisch-grafisch-abstrakten Anforderungen, insbesondere bei der Verwendung mentaler Modelle. Dabei haben allerdings die Unterschiede eher geringere Ausmaße, nur im extremsten Fall (Teilkompetenz Bewerten) beträgt die Differenz fast vier Kompetenzpunkte, also nicht einmal eine halbe Standardabweichung. Im Übrigen sind diese geschlechtsspezifischen Kompetenzprofile über die Leistungsquartile hinweg stabil.

Dieses Ergebnis bestärkt die Befunde gängiger Untersuchungen zu Geschlechterunterschieden in kognitiven Leistungen (vgl. Abschnitt 1.2.1), dass Mädchen eher verbal, Jungen dagegen abstrakt-analytisch orientiert sind. Wie in Abbildung 2 schematisch dargestellt, überlappen sich die geschlechtsspezifischen Verteilungen dabei stark, die Unterschiede gelten nur für den Durchschnitt.

Das heißt, ggf. würden diese Profile noch deutlicher, wenn man Schülergruppen nach Teilkompetenzen einteilt, dann könnten die Anteile von Mädchen und Jungen die Stärke der Geschlechterassoziation ausdrücken. Dieser Frage wird weiter unten nachgegangen.

Geschlechterunterschiede in den verschiedenen Schulformen werden in diesem Rahmen nicht berichtet, da die Bildungsbeteiligung von Jungen und Mädchen variiert und so die Ergebnisse verzerrt.

1.3.1.3 PISA 2000 – Geschlechterunterschiede im Überblick

Um die Darstellung der Geschlechterunterschiede bei PISA zu vervollständigen und einen Vergleich zu anderen Studien wie TIMSS (vgl. 1.3.2) zu ermöglichen, werden im Folgenden rückblickend Befunde von PISA 2000 vorgestellt (Stanat & Kunter, 2001).

Auch bei PISA 2000 schnitten Mädchen in ihrer durchschnittlichen Leseleistung in allen OECD-Staaten besser ab als Jungen. Im OECD-Mittel betrug dieser Unterschied 32 Punkte, Deutschland lag bei einer Differenz von 35 Punkten. Die Erfassung der Lesegeschwindigkeit zeigte in Deutschland ebenfalls einen klaren Vorsprung der Mädchen.

Bei einer genaueren Analyse der Leseleistungen konnte eine Überlegenheit der Mädchen konsistent über die verschiedenen Subskalen zum Lesen hinweg beobachtet werden. Obwohl beim Lesen nicht-kontinuierlicher Texte eine relative Stärke der Jungen auf diesem Teilgebiet sichtbar wurde, wiesen Mädchen auch hierbei noch immer signifikant bessere Leistungen auf, wenn auch mit einer Effektstärke von etwa einer zehntel Standardabweichung nicht mehr so deutlich wie in den anderen Bereichen des Lesens.

Die Einstellung der Jungen zum Lesen, die bei PISA 2000 im Rahmen der Erhebung von Hintergrundmerkmalen, Motivation und Emotion erfasst wurde, war deutlich negativer als die der Mädchen.

Die Verschiedenheit der mathematischen Leistung von Jungen und Mädchen war über die PISA-Länder hinweg weniger konsistent als im Lesen und lag mit einem mittleren statistisch bedeutsamen Wert von 11 Punkten der OECD-Länder zu Gunsten der Jungen in einem

deutlich kleineren Bereich als beim Lesen. Deutschland übertraf diesen Wert mit 15 Kompetenzpunkten signifikanter Differenz nur unwesentlich.

Während in etwa der Hälfte aller Länder Jungen signifikant besser abschnitten, gab es im Gegensatz zu PISA 2003 kein einziges Land, in dem bei Mädchen signifikant bessere Leistungen als bei Jungen beobachtbar waren. Tendenziell (aber ohne statistische Bedeutsamkeit) zeigte sich ein solcher Zusammenhang allenfalls in Island, Neuseeland und der Russischen Föderation. In Island sind also die nicht signifikanten Unterschiede zwischen Mädchen und Jungen in der mathematischen Kompetenz bei PISA 2000 zu signifikant besseren Leistungen der Mädchen in PISA 2003 angewachsen. Dies ist insofern als eher ungünstig zu beurteilen, als dass die Förderung eines Geschlechts nicht zu Lasten des anderen gehen sollte.

Der internationale Naturwissenschaftstest bei PISA 2000 förderte im Schnitt keine signifikanten Leistungsunterschiede zwischen Mädchen und Jungen zu Tage. Auf drei Länder mit signifikanten Unterschieden mit einem Vorsprung von Jungen kamen drei Länder mit solchen von Mädchen, wohingegen sich die Unterschiede in Deutschland nicht statistisch absichern ließen.

Auch bei PISA 2000 wurden Staaten ausgemacht, die ein geschlechterfaires Bild von Leistungen zeigen, dazu zählen die Vereinigten Staaten und Großbritannien (Stanat & Kunter, 2003).

Insgesamt veränderte sich das Bild von PISA 2000 zu PISA 2003 lediglich marginal (Tabelle 3), die Geschlechterunterschiede sind also über die Zeit hinweg stabil geblieben bzw. haben sich in den Naturwissenschaften sogar vergrößert. Dabei ist allerdings zu beachten, dass sich die Auswahl der einbezogenen Länder verändert hat.

Tabelle 3 Geschlechterunterschiede in Kompetenzpunkten für die Leistungsdomänen Lesen, Mathematik und Naturwissenschaften bei PISA. Angegeben ist *das internationale Mittel* über alle getesteten OECD-Länder für zwei Messzeitpunkte. Positive Werte bedeuten einen Kompetenzvorteil der Jungen, ein negativer zu Gunsten der Mädchen. Signifikante Unterschiede sind fett gedruckt.

Leistungsdomäne	2000	2003
Lesen	**-32**	**-34**
Mathematik	**11**	**11**
Naturwissenschaften	0*	**6**

*in den einschlägigen Werken wird stets von einem nicht signifikanten Mittelwert oder nicht vorhandenem Unterschied berichtet, daher wird der entsprechende Wert als Null angegeben.

1.3.1.4 Zusammenfassung der Geschlechterunterschiede bei PISA 2003

Die Geschlechterunterschiede bei PISA lassen sich wie folgt zusammenfassen:
Die größten Unterschiede bestehen im Lesen zu Lasten der Jungen. In Deutschland zeigen die 15-jährigen Jungen gegenüber Mädchen Defizite im Lesen, schneiden dafür nur leicht besser in Mathematik ab. In den Naturwissenschaften und im Problemlösen gibt es auf globaler Testebene keine nennenswerten Unterschiede. Im internationalen Naturwissenschaftstest nehmen die Geschlechterunterschiede mit größerer Kompetenz zu.
Beim nationalen Test zeigen sich geschlechtsspezifische Stereotypen bestätigende Muster bezüglich kognitiver Teilkompetenzen. Erhebungen zu Emotion und Motivation bekräftigen ebenfalls Rollenstereotype.
Da die Leseleistung als zentrale Voraussetzung für Lernleistungen in anderen Domänen angenommen wird, sind Jungen im Durchschnitt stärker im Nachteil als Mädchen.
Insgesamt bestätigen die Befunde aus PISA 2003 die Ergebnisse aus anderen Studien; sie zeigen aber auch, dass in einigen Staaten eine Minimierung der Unterschiede zwischen Mädchen und Jungen möglich ist.
Wie weiter bereits angedeutet, bietet die Bezugnahme auf das biologische Geschlecht nur wenig neue Erkenntnisse. Obgleich die Berücksichtigung des biologischen Geschlechts in Alltag und Forschung einfacher ist, kann die psychologische Zugehörigkeit für die Unterschiede von Subpopulationen bedeutender sein. In den meisten Studien wird aber das rein biologische Geschlecht, also *sex* statt *gender* erhoben. Analysen mit dem Geschlecht finden also stets auf erstens deterministischer und zweitens nicht-latenter Ebene statt.

1.3.2 TIMSS

TIMSS ist die Abkürzung für „Third International Mathematics and Science Study". Diese von der IEA (International Association for the Evaluation of Educational Achievement) initiierte international angelegte Studie vereint die Testung von Mathematik und Naturwissenschaften nach einer einzelnen Testung der Leistungsdomänen in der ersten und zweiten Erhebungswelle. So wurden zu Beginn die beiden Studien First International Mathematics Study (FIMS) und First International Science Study (FISS) vorgelegt, später erfolgten die Second International Mathematics Study (SIMS) und Second International Science Study (SISS) noch immer getrennt. Die Darstellungen der TIMS-Studie halten sich an die internationalen und nationalen Berichtsbände (Baumert, Bos, & Lehmann, 2000;

Mullis, Martin, Fierros, Goldberg, & Stemler, 2000a). Wie PISA nutzt TIMSS ein Multimatrix-Testdesign, die Auswertungen erfolgten ebenfalls mit Hilfe der Item-Response-Theory.

Im Zentrum steht bei TIMSS die Erhebung von „literacy“ oder Grundbildung, also die Erfassung von Basisqualifikationen unserer (westlichen Industrie-) Gesellschaft. Bei TIMSS werden im Gegensatz zu anderen large scale assessment - Studien parallel drei Altersgruppen untersucht. Population I umfasst Grundschüler der beiden Klassenstufen, die den größten Anteil an Neunjährigen aufweisen, in den meisten teilnehmenden Ländern entspricht dies der dritten und vierten Klasse. Deutschland beteiligte sich nicht an Untersuchungen zur Population I.

Schüler der Sekundarstufe I bilden die Zielgruppe der Population II. Als internationale Vorgabe sollten die beiden Klassen aufgenommen werden, welche den größten Anteil an 13-jährigen beinhalteten. Größtenteils entsprach das der 7. oder 8. Klasse, in Deutschland aber häufiger niedrigeren Klassenstufen. Deutschland suchte allerdings eher Anschluss an die Klassenvorgabe anderer Länder, daher wurden ältere Schüler als in anderen Ländern in den Test aufgenommen, also solche aus den 7. und 8. Klassen. Diese Gruppe ist am stärksten mit PISA vergleichbar.

Eine breitere Rahmendefinition bestimmte die Population III. Dazu zählte der Abschlussjahrgang der Sekundarstufe II in allgemein bildenden und beruflichen Schulen.

Als regelmäßiges international übergreifendes Instrumentarium eines Bildungssystemvergleichs stellte TIMSS zugleich die Weichen für PISA: Nach der Nutzung von TIMSS-Ergebnissen entwickelte die OECD eine eigene Studie, um dieses eigenen Wünschen anzupassen.

Wie PISA verfügt auch TIMSS über ein periodisches Erhebungssystem. Bei TIMSS 1995 wurden alle drei oben ausgeführten Stichproben untersucht, der zweite Zyklus von TIMSS im Jahr 1999 (nun als Trends in International Mathematics and Science Study bezeichnet) sah nur noch die Testung von Achtklässlern (Population II) vor, der darauf folgende Zyklus TIMSS 2003 erweiterte die Testung so, dass wieder Viert- und Achtklässler (Population I und II) gestestet wurden. Die nächste Runde von TIMSS ist 2007 mit den gleichen Klassenstufen vorgesehen. Deutschland beteiligte sich an TIMSS 1995 und wird auch 2007 wieder teilnehmen, wohingegen die Zyklen 1999 und 2003 ausgelassen wurden.

Der Mathematiktest bei TIMSS orientierte sich am Curriculum der teilnehmenden Staaten. In einem Multimatrix-Testdesign vorgelegt, umfasste er verschiedene Inhaltsbereichen. TIMSS 1995 umfasste: 1. whole numbers, 2. fractions and proportionality, 3. measurement,

estimation and number sense, 4. data representation, analysis and probability, 5. geometry, 6. patterns, relations and functions. Etwa ein Viertel der Aufgaben mussten die Schüler in freiem Antwortformat bearbeiten.
Der 1995 an Population I angewendete Naturwissenschaftstest umfasste die Stoffgebiete Geowissenschaften, Lebenswissenschaften (life sciences), Physik/Chemie sowie eine Restkategorie. Analog zum Mathematiktest wurde jener im Multimatrix-Testdesign vorgegeben und erforderte bei etwa einem Viertel der Aufgaben Lösungen in freiem Antwortformat.

Obgleich Deutschland innerhalb von TIMSS nicht an Untersuchungen der Population I teilgenommen hat, sollen hier auch die Ergebnisse für Geschlechterunterschiede bei Grundschülern dargestellt werden. Geschlechterunterschiede in der Primarstufe sind zwar nicht mit der Zielstichprobe bei PISA vergleichbar, aber insofern von Interesse, als dass jene mit den Befunden bei IGLU-E vergleichbar sind (siehe dort). Zudem bilden Geschlechterunterschiede in Schulsystemen im Grundschulalter die Grundlage für spätere Ungleichheiten zwischen Mädchen und Jungen. Bestehen bereits in jungem Alter größere Leistungsunterschiede, werden sich diese in einer unterschiedlichen Wissensgrundlage bemerkbar machen, auf denen spätere Beschulung ansetzt. Durch geschlechtsspezifische Leistungen werden stets auch Überzeugungen, Motivation und Emotionen vermittelt, sowohl bei Schülern als auch bei Lehrern.

1.3.2.1 Leistungsunterschiede von Jungen und Mädchen bei TIMSS

Die Geschlechterunterschiede über alle Zyklen sind in Tabelle 4 dargestellt, die in der Tabelle und im Text enthaltenen Kennzahlen und Zusammenhänge können bei Köller & Klieme (2000), Martin, Mullis, Gonzalez, & Chrostowski (2004), Martin et al. (2000), Mullis et al. (2000a), Mullis, Martin, Gonzalez, & Chrostowski (2004), Mullis et al. (2000b) nachgelesen werden.
Analog zu PISA sind die geschlechtsspezifischen Ergebnisse der mathematischen und naturwissenschaftlichen Kompetenz in TIMSS-Population I eher von geringerem Ausmaß. In nur wenigen Ländern sind signifikante Unterschiede zu Gunsten der Jungen im Mathematiktest nachweisbar, dabei liegen die größten Differenzwerte 15 Punkte zwischen einer zehntel und einer fünftel Standardabweichung. Im Naturwissenschaftstest zeigen sich in

mehr Ländern signifikante Leistungsunterschiede - stets zu Gunsten der Jungen (Mullis et al., 2000a).

Bei der verwendeten TIMSS Skala mit einem Mittelwert von 500 und einer Streuung von 100 entsprechen die maximalen Unterschiede beinahe einer drittel Standardabweichung, was gerade im Vergleich zu PISA und in Anbetracht des geringen Alters der Schüler relativ hoch ist.

Bei Population II, also Schülern der Sekundarstufe I, sind nun in deutlich mehr Ländern ungleiche Leistungen zwischen Mädchen und Jungen zu beobachten, die Unterschiede zwischen Jungen und Mädchen in der mathematischen Kompetenz bereits etwas deutlicher, in der naturwissenschaftlichen Kompetenz wesentlich deutlicher ausgeprägt.

Die größten Differenzen werden bei TIMSS 1995 in Israel mit fast 30 Punkten, also einer knappen drittel Standardabweichung erreicht.

Die größten Unterschiede lassen sich in der Stichprobe von Population III beobachten. Bei Schülern gegen Ende der Sekundarstufe II besteht in fast allen Ländern ein Leistungsvorsprung der Jungen, in Norwegen beträgt dieser im Zyklus von 1995 sogar 54 Punkte, also eine halbe Standardabweichung. Bei Testung der „advanced mathematical literacy“ erreichen die Unterschiede eine Größe von 92 Punkten in der Tschechischen Republik und im Länderdurchschnitt 37 Punkte. Für Deutschland berichten Mullis et al. (2000a) einen Wert von 29 Skalenpunktunterschieden in Mathematiktest und 32 Punkten im Test der „advanced mathematical literacy“.

Nationale Analysen der Subpopulation der gymnasialen Oberstufe (Köller & Klieme, 2000) zeigten geschlechtsstereotypes Kurswahlverhalten sowohl in Mathematik als auch in Physik. Dort wurden die Geschlechterunterschiede in den Leistungen v. a. durch Aufgaben aus dem Bereich Zahlen, Gleichungen und Funktionen in Mathematik und in Physik hauptsächlich durch solche über physikalische Alltagsprobleme produziert. Die geringsten Geschlechterdifferenzen sind erstaunlicherweise in der Geometrie zu finden; dies ist ein leichter Widerspruch zu den Ergebnissen aus PISA 2003, wo die Subskala „Raum und Form“ einen deutlicheren Leistungsvorsprung der Jungen zu Tage treten ließ. Im Grundkurs des Faches Mathematik sind die Leistungsunterschiede zwischen Mädchen und Jungen praktisch nicht relevant, im Leistungskurs sehr viel stärker ausgeprägt. Anders ausgedrückt gilt hier wie im internationalen Naturwissenschaftstest bei PISA 2003, dass die Unterschiede mit zunehmender Kompetenz steigen. Interessanterweise konnte im Fach Physik in TIMSS dieser Effekt nicht nachgewiesen werden, denn in diesem Fall sind die Geschlechterunterschiede durchaus vergleichbar hoch ausgeprägt.

Tabelle 4 Geschlechterunterschiede in Kompetenzpunkten (Skala mit Mittelwert 500 und einer Standardabweichung von 100) für die Leistungsdomänen Mathematik und Naturwissenschaften. Angegeben ist das internationale Mittel über alle getesteten TIMSS-Länder für drei Messzeitpunkte. Alle Werte sind positiv, drücken also einen Vorsprung zu Gunsten der Jungen aus.

Mathematik	1995	1999	2003
Population I	2	-	1
Population II	8	4	1
Population III	33	-	-
Naturwissenschaften	**1995**	**1999**	**2003**
Population I	9	-	1
Population II	17	15	6
Population III	39	-	-

Betrachtet man den Geschlechterunterschied über die getesteten Populationen und die Messzeitpunkte hinweg (Tabelle 4) lassen sich mit Vorsicht folgende Aussagen treffen: Die Kluft zwischen Jungen und Mädchen in Mathematik und den Naturwissenschaften nimmt bei TIMSS mit dem Alter zu, aber mit den Messzeitpunkten ab. Dieser Befund reiht sich durchaus in andere Studien ein (z.B. Fennema, 1996). Möglicherweise gibt es eine generelle Tendenz, dass die Geschlechterunterschiede abnehmen, sei es, weil in Bildungssystemen geschlechterfair unterrichtet wird, sei es, weil kulturelle und gesellschaftliche Unternehmungen zur Gleichberechtigung fruchten. Da sich bekanntlich Sozialisation und Biologie bedingen und gegenseitig beeinflussen, kann eine solche Tendenz Ursache oder Wirkung sein.

Generell zeigt sich das Muster, dass mit höheren Klassenstufen die Geschlechterunterschiede größer werden und in deutlich mehr Ländern auftreten. Dabei scheint es so zu sein, dass in den Naturwissenschaften deutlichere Kompetenzdifferenzen auftreten als in Mathematik. Weiterhin ist eine Abhängigkeit vom Stoffgebiet beobachtbar, Mädchen sind besser in Algebra und den life sciences sowie bei Umweltfragestellungen, Jungen profitieren von Aufgaben in measurement, Physik, Chemie und Geowissenschaften.

Eine solche Schlussfolgerung kann nur mit einigem Vorbehalt gezogen werden. So veränderte sich in TIMSS die Länderzusammensetzung mit jedem Messzeitpunkt, was sich selbstverständlich im Mittelwert niederschlägt. Der Alterseffekt enthält, wenn er über die Erhebung 1995 interpretiert wird, genau genommen einen Kohorteneffekt. Allerdings kann er prinzipiell auch als Längsschnitteffekt beobachtet werden, wenn die Population I aus dem

Jahr 1995 mit der Population II aus dem Jahr 1999 verglichen wird. Ein weiterer TIMSS-Zyklus im Jahr 2007 wird zeigen, ob obige Interpretationen stabil bleiben.

1.3.2.2 PISA und TIMSS – Ein Vergleich

Tabelle 5 Vergleich der geschlechtsspezifischen Ergebnisse für PISA und TIMSS für die Leistungsdomänen Mathematik und Naturwissenschaften bei den verschiedenen Erhebungswellen. Dargestellt ist der über alle teilnehmenden Länder gemittelte Kompetenzunterschied zwischen Mädchen und Jungen in Skalenpunkten.

	TIMSS 1999*	PISA 2000	TIMSS 2003*	PISA 2003
Mathematik	4	11	1	11
Naturwissenschaften	15	0**	6	6

* angegeben ist der Wert für Population II, weil dieser mit der Altersgruppe von PISA vergleichbar ist.
**in den einschlägigen Werken wird stets von einem nicht signifikanten Mittelwert oder nicht vorhandenem Unterschied berichtet, daher wird der entsprechende Wert als Null angegeben.

Vergleicht man die geschlechtsspezifischen Ergebnisse für PISA und TIMSS (Tabelle 5), so lässt sich folgendes Muster beobachten: In der Erhebungswelle 1999 (TIMSS) bzw. 2000 (PISA) zeigt PISA geringere Unterschiede in den Naturwissenschaften, aber größere in Mathematik. Beim darauf folgenden Messzeitpunkt im Jahr 2003 repliziert sich der Befund für Mathematik, aber die Werte für Naturwissenschaften sind nun zwischen PISA und TIMSS vergleichbar.

Betrachtet man nicht die internationalen Durchschnittsebene, sondern deutschlandspezifischen Ergebnisse (Tabelle 6), kann für TIMSS nur noch der Wert aus der Erhebungswelle in 1995 herangezogen werden, da Deutschland an TIMSS weder 1999 noch 2003 beteiligt war. Prinzipiell zeigt sich ein ähnliches Ergebnis: Die Unterschiede in Mathematik sind bei PISA höher, dafür in den Naturwissenschaften niedriger ausgeprägt als bei TIMSS.

Tabelle 6 Vergleich der geschlechtsspezifischen Ergebnisse für PISA und TIMSS für die Leistungsdomänen Mathematik und Naturwissenschaften. Dargestellt ist der Kompetenzunterschied zwischen deutschen Mädchen und Jungen sin Skalenpunkten. Da Deutschland weder an TIMSS 1999 noch an TIMSS 2003 teilgenommen hat, kann hier lediglich auf den Wert in TIMSS 1995 zurückgegriffen werden.

nur für Deutschland	TIMSS 1995*	PISA 2000	PISA 2003
Mathematik	3	15	9
Naturwissenschaften	18	3	6

* angegeben ist der Wert für Population II, weil dieser mit der Altersgruppe von PISA vergleichbar ist.

Dass Unterschiede zwischen Jungen und Mädchen in der TIMSS-Population anders als bei PISA ausfallen, kann mehrere Ursachen haben. Zum einen ist das Aufgabenmaterial bei TIMSS stärker am Curriculum orientiert und in den Naturwissenschaften möglicherweise weniger sprachlastig. Bei mehr verbal gestaltetem Material können Mädchen eventuelle Defizite eher ausgleichen. Eine Variation der Anteile von Biologie, Chemie und Physik in naturwissenschaftlichen Tests modifiziert ebenfalls Leistungsunterschiede zwischen Mädchen und Jungen.

1.3.3 IGLU/PIRLS

PIRLS steht für „Progress in International Reading Literacy Study“ und wird in Deutschland als IGLU (Internationale Grundschul-Lese-Untersuchung) bezeichnet. Als internationales Projekt mit einer Teilnahme von 35 Ländern wurde IGLU von der IEA (International Association for the Evaluation of Educational Achievement) entwickelt. Die Ausführungen an dieser Stelle halten sich an den nationalen Berichtsband zu IGLU (Bos et al., 2003a).

Ähnlich wie bei PISA und TIMSS geht es um eine Erfassung wichtiger Basiskompetenzen und die Erhebung von Grundbildung (Literacy), diesmal bei vorwiegend Neunjährigen, in Deutschland bei Viertklässlern. Informationen aus Fragebögen für Eltern, Lehrer, Schulleiter und der getesteten Kinder ergänzte die Erhebung zentraler Voraussetzungen zur Bewältigung aktueller Lebensanforderungen und nachfolgender Bildungsprozesse im Leben hauptsächlich Neunjähriger einer klassenbasierten Stichprobe.
International standardisiert war die Vorgabe von Tests zum Leseverständnis. Im Rahmen einer nationalen Option innerhalb der als IGLU-E bezeichneten deutschen Erweiterung erfolgte die Messung der mathematischen und naturwissenschaftlichen Kompetenz an einem zweiten Testtag, an dem sich aber nicht alle Länder Deutschlands beteiligten.
Um die Ergebnisse in den internationalen Kontext von TIMSS Population I einzuordnen, wurden Items aus TIMSS verwendet.

Der Stichprobenumfang von IGLU umfasste in Deutschland 7.633 Schüler an 211 Schulen, bei IGLU-E immerhin noch 5.943 Schüler an 168 Schulen. An IGLU-E nahmen nur noch 12 Länder Deutschlands teil, während IGLU mit dem ersten Testtag alle 16 Länder einschloss. Wie bei PISA wurden den Schülern in einem Multi-Matrix-Design rotierte Aufgabenblöcke von einer etwa zweistündigen Nettobearbeitungszeit vorgegeben.
Bevor die Ergebnisse von IGLU mit PISA verglichen werden, muss betont werden, dass also sowohl eine andere Altersstufe als auch mit anderen Verfahren in einer anderen Auswahl von Teilnehmerstaaten getestet wurde.
Da Neunjährige unter einem anderem sowohl kulturellen wie hormonellen Einfluss stehen als Fünfzehnjährige sind geringere Geschlechterunterschiede als bei PISA zu erwarten.
Innerhalb von IGLU-E ist eine internationale Einordnung nicht möglich, da die Testbereiche Mathematik und Naturwissenschaften eine nationale Ergänzung darstellen. Aus diesem Grund werden bei IGLU-E einige Befunde auf die TIMSS-Skala bezogen skaliert. Hierbei wurde

jeweils nur eine Teilstichprobe der Population I herangezogen, da TIMSS 3. und 4. Klassen, IGLU-E hingegen nur 4. Klassen umfasste. Dazu wurde in einem aufwändigen Verfahren ein Teil der TIMSS-Daten reskaliert.

1.3.3.1 Geschlechterunterschiede bei IGLU

Wie auch bei PISA schneiden auch im Lesetest IGLUs in allen Ländern Mädchen besser ab als Jungen. Die Geschlechterunterschiede sind in allen Ländern auf dem Niveau von 0,05 signifikant von Null verschieden. Wie erwartete sind die Unterschiede nicht so deutlich wie bei der Stichprobe der 15-Jährigen aus PISA 2003: Die höchsten Unterschiede bei IGLU betragen 27 Skalenpunkte, dies entspricht knapp einer drittel Standardabweichung, während bei PISA der maximale Unterschied im Lesen die Hälfte einer Standardabweichung überschreitet.

Deutschland liegt mit einem Geschlechterunterschied von nur 13 Punkten klar im unteren Bereich, nur Italien zeigt einen nennenswert niedrigeren Differenzwert (Bos et al., 2003b).

Der Test zur Mathematik umfasste Aufgaben zu den inhaltlichen Bereichen Arithmetik, Geometrie, Größen und Sachrechnen. Bezogen auf die nationale Metrik bei IGLU schnitten Jungen im Durchschnitt 24 Skalenpunkte besser als Mädchen ab, dies entspricht etwa einer Viertel Standardabweichung (Walther, Geiser, Langeheine, & Lobemeier, 2003). Dies ist im Vergleich zu einem Geschlechterunterschied bei PISA 2003 von etwa einer zehntel Standardabweichung in Deutschland etwas höher.

Genauere Analysen des Tests zeigten keine Unterschiede zwischen Mädchen und Jungen bezüglich ihrer Leistungen hinsichtlich der Aufgaben zur Größenthematik (Lobemeier, 2005).

Bei der Erhebung des naturwissenschaftlichen Verständnisses wurden Aufgaben zu den Themen Physik/Chemie, Biologie sowie Erde/Umwelt vorgelegt. Die Kompetenzunterschiede zwischen Mädchen und Jungen in Deutschland betragen hier auf der TIMSS-Skala 15 Punkte (Prenzel, Geiser, Langeheine, & Lobemeier, 2003). Die Skalenunterschiede beziehen sich auf eine reskalierte Teilstichprobe von TIMSS (nur 4. Klassen), die Autoren geben hier weder geänderte Standardabweichung noch Effektgrößen an. Geht man in etwa von einer Standardabweichung in der Größe um 100 aus, wie sie in der Gesamtpopulation von TIMSS (also 3. und 4. Klasse) vertreten war, so bewegen sich die Unterschiede also zwischen einer zehntel und einer fünftel Standardabweichung.

Wie in der Mathematik ist der Anteil von Mädchen auf den unteren Kompetenzstufen höher, auf den oberen Stufen niedriger.

Insgesamt zeigen sich also auch bei Viertklässlern im Rahmen von IGLU/IGLU-E bereits deutliche Geschlechterunterschiede.

1.3.4 PISA, TIMSS, IGLU – Gemeinsamkeiten und Unterschiede

Will man die Geschlechterunterschiede für IGLU mit denen bei TIMSS und PISA vergleichen, so bietet sich der Vergleich der national erzielten Befunde an (Tabelle 7). Dies liegt darin begründet, dass Naturwissenschaften und Mathematik in IGLU nicht international, sondern im Rahmen der nationalen Ergänzung appliziert wurden, in TIMSS hingegen war die Erhebung der Lesekompetenz nicht Teil der Untersuchung.
Der Vergleich für Population I in TIMSS kann (obgleich dieser adäquater wäre) nicht gezogen werden, da sich Deutschland in 1995 nur mit Population II und III beteiligt war.

Tabelle 7 Vergleich der geschlechtsspezifischen Ergebnisse für IGLU, PISA und TIMSS in den Leistungsdomänen Mathematik und Naturwissenschaften. Dargestellt ist der Kompetenzunterschied zwischen Mädchen und Jungen in Skalenpunkten für Deutschland. Da Deutschland weder bei TIMSS 1999 noch bei TIMSS 2003 teilgenommen hat, kann hier lediglich auf den Wert in TIMSS 1995 zurückgegriffen werden.

nur für D	IGLU 2001	TIMSS 1995*	PISA 2000	PISA 2003
Lesen	-13	-	-35	-42
Mathematik	24**	3	15	9
Naturwissenschaften	15**	18	3	6

* angegeben ist der Wert für Population II, in Deutschland wurde Population I im Rahmen von TIMSS 1995 nicht getestet.
** Werte aus der nationalen Ergänzungsstudie IGLU-E.

Während die Leseunterschiede zwischen Jungen und Mädchen bei IGLU im Vergleich zu PISA eher gering einzuschätzen sind, werden im Bereich mathematischer und naturwissenschaftlicher Kompetenz sogar stärkere Ausmaße als bei den wesentlich älteren Schülern von PISA und TIMSS erreicht. Die Leistungsdiskrepanz in den Naturwissenschaften ist bei IGLU mit den älteren Schülern von TIMSS vergleichbar und sogar höher als in der PISA-Stichprobe.
Dies ist insofern bemerkenswert als dass Geschlechterunterschiede mit zunehmendem Alter wachsen. Demnach sollten bei der Stichprobe von IGLU/IGLU-E die Unterschiede zwischen Mädchen und Jungen durchgehend geringer sein als bei PISA, was sich aber nur für die Lesekompetenz nachweisen lässt, bei den Naturwissenschaften und Mathematik zeigt sich sogar der entgegengesetzte Zusammenhang.

Die Unterschiede zwischen den Untersuchungen lassen sich u.a. auf anders gestaltetes Aufgabenmaterial zurückführen. Da andere Länder miteinbezogen wurden und sich die Auswahl der Länder auf die Skalierung der Ergebnisse auswirkt, kann dies auch die Unterschiede mit begründen. Weiterhin ist zu berücksichtigen, dass IGLU auf einer klassenbasierten Stichprobe basiert, was prinzipiell auch für TIMSS gilt, während PISA gerade unabhängig von Klassen testet. Auch wenn ähnliche Konzepte bei der Kompetenzdefiniton und –messung zugrunde gelegt werden, so sind doch die Befunde nicht zwangsläufig vergleichbar.

Für sich genommen zeigen aber die Befunde aus IGLU, dass bereits im Grundschulalter sehr deutliche Geschlechterunterschiede vorhanden sind. Wie schon PISA geht aber auch IGLU nicht über eine Erfassung des biologischen Geschlechts hinaus, so dass Unterschiede zwischen Mädchen und Jungen rein an der Biologie und nicht am psychologischen Geschlecht orientiert sind. Man kann also auch hier weniger von gender differences als mehr von sex differences sprechen.

1.3.5 Zusammenfassung

Eine Betrachtung der Geschlechterunterschiede in international angelegten Studien verdeutlicht, dass die Diskrepanzen zwischen Mädchen und Jungen mit Lebensalter, Kohorte, Testverfahren, Leistungsdomäne und kognitiven Anforderungen variieren.

Homogen über Staaten hinweg sind die Befunde zum Lesen. Man ist geneigt, bei so einheitlichen Ergebnissen biologische Erklärungsmodelle dafür heranzuziehen. Aber es ist festzuhalten, dass sich die Kulturen in Industrienationen hinsichtlich ihrer Rollenbilder allenfalls in ihrem Ausmaß, aber nicht in der Richtung unterscheiden. Die Daten aus PISA lassen die Vermutung zu, dass Jungen stärker benachteiligt sind als Mädchen, da die Leseunterschiede erstens stärker ausgeprägt und zweitens über alle Staaten gleichgerichtet sind, was weder auf Mathematik noch Naturwissenschaften zutrifft.

Gerade was letztere Domänen angeht, sind verschiedene Muster zu finden. Einige Staaten verstehen es besser als andere, eine geschlechterfaire Bildungsbasis zu schaffen. Da auch stereotypenkonträre Muster zu finden sind, wird belegt, dass rollenkonforme Leistungen weniger naturgegeben sind bzw. durch Förderung, Erwartungen und Selbstbild durchaus beeinflussbar sind.

Auch wenn teils noch recht starke Ungleichheiten zu finden sind, muss doch bedacht werden, in welchem Rahmen sich diese bewegen, dass diese häufig überinterpretiert werden und hinter anderen weitaus größeren Verschiedenheiten zurücktreten.

Bedauerlicherweise wurde in allen genannten Studien auf eine Erhebung des psychologischen Geschlechts verzichtet, das Geschlecht nur auf den biologischen Unterschied reduziert, geschlechtsspezifische Rollenbilder weder von Lehrern, Eltern noch den Schülern selbst erhoben. So sind Leistungsprofile weiterhin stets an der biologischen Einteilung des Geschlechts festgemacht, obwohl sich bereits gezeigt hat, dass die Biologie nur ein Einflussfaktor neben Sozialisation und psychologischen Einflüssen ist.

Ebenso interessant wäre zu sehen, welche Items im Test Geschlechterunterschiede produzieren, und ob diese tatsächlich rollenkonform Stereotype bedienen. In den folgenden Abschnitten wird versucht, diesen Fragen innerhalb der PISA-Daten nachzugehen.

2 Die Item Response Theory (IRT)

Neben den allgemein verwendeten statistischen Verfahren werden im Folgenden die Modelle der so genannten probabilistischen Testtheorie Verwendung finden.
Nachstehend sollen Eignung und Vorzüge der Item-Response-Theory (IRT) in Hinblick auf die Fragestellungen und Ergebnisse der vorliegenden Publikation gezeigt werden, der Aufbau des Kapitels 2 ist in Abbildung 6 veranschaulicht. Die Darstellung der verwendeten testtheoretischen Verfahren kann kaum erschöpfend sein, sondern nur als Überblick dienen. Um die Arbeit in allen Facetten nachvollziehen zu können, sei auf die zahlreiche Literatur zum Thema IRT verwiesen, einen guten Überblick geben Rost (2004), Baker & Kim (2004) sowie Schuchmann (1999).

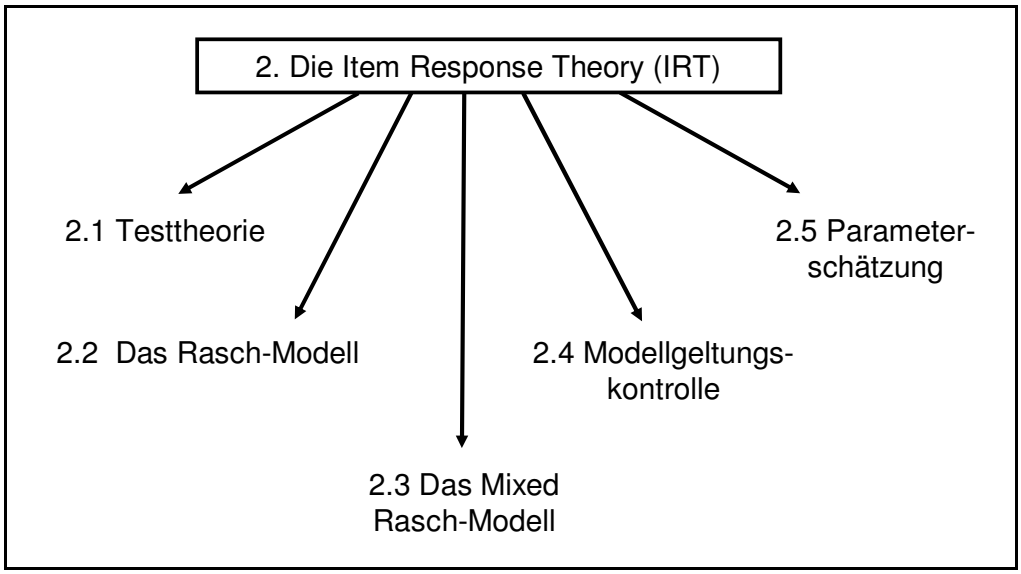

Abbildung 6 Gliederung von Kapitel 2.

2.1 Allgemeines zur Testtheorie

Definitionen, was unter Testtheorie zu verstehen ist, finden sich bei Rost (2004) und Lienert & Raatz (1998). Im Folgenden soll nur auf die Theorie psychologischer Tests eingegangen werden, also auf die Beziehung zwischen Testverhalten und der Erfassung des psychologischen Merkmals, dabei wird, sofern nicht anders angegeben, vor allem auf Rost (2004) Bezug genommen.
Bei der Testtheorie wird weniger von einer speziellen allgemeingültigen Theorie ausgegangen, die auf alle Stichproben, Tests, Daten und psychologischen Merkmale

anwendbar ist, es muss stattdessen zwischen verschiedenen formalen Modellen unterschieden werden, die aber alle unter dem Begriff der Testtheorie subsumiert werden können.

Allen Modellen gemeinsam ist die Annahme, dass ein latentes, also nicht direkt beobachtbares (psychologisches) Merkmal das Testverhalten beeinflusst, also das Testergebnis produziert. Obwohl nur das Testergebnis direkt beobachtbar ist, wird versucht, über das zu messende Merkmal eine Aussage zu treffen. Dieses Merkmal kann mittels des Testergebnisses, das über das Testverhalten gleichsam konstruiert wird, quantifiziert, qualifiziert oder zugleich quantifiziert und qualifiziert werden. Es wird beispielsweise versucht, über Antworten auf Fragebogenitems (Testverhalten) Rückschlüsse über die Intelligenz einer Person (quantitativ) oder mit einem anderen Test die Klassifizierung kognitiver Stile (qualitativ) zu ziehen. Dabei ist natürlich zu beachten, dass die Konzepte wie Intelligenz oder kognitive Stile „nur“ Konstrukte sind. Man hat es in der Psychologie vor allem mit nicht beobachtbaren Variablen zu tun, aber das Modell ist rein prinzipiell auch auf manifeste Variablen anwendbar: So kann angenommen werden, dass das biologische Geschlecht (direkt beobachtbar, also manifest) einen maßgeblichen Einfluss auf das Verhalten in einem Test nimmt. Was allerdings den psychologischen Anteil an dieser manifesten Variable ausmacht, ist die Annahme, dass diese mit einer Reihe von latenten Variablen zusammenhängt, die das Testergebnis sehr wahrscheinlich beeinflussen (Rollenstereotype, kognitive Strukturen, Einstellungen etc., siehe Abschnitt 1.2.2.2). Auch konfundieren manifeste wie latente Variablen in der Regel mit einer Reihe anderer ebenfalls manifester oder latenter Variablen. Weiterhin sind die erhobenen Daten oftmals von Störfaktoren verrauscht.

Aus Gründen der Vereinfachung kann man sich deshalb entweder auf ein zu messenden Konstrukt festlegen oder es wird ein allgemeines zentrales Konstrukt angenommen, das hinter vielen Variablen steht und ggf. aussagekräftiger ist, als wenn einzelne Konstrukte vermessen werden.

Psychologische Theorien haben das Ziel, zu vereinfachen und allgemeingültige Aussagen zu machen (wenn dazu selbstverständlich auch oft Ausnahmen oder Gültigkeitseinschränkungen gehören), außerdem modellieren sie stets in irgendeiner Form empirisch beobachtbare Ergebnisse, ansonsten wären sie für die empirische Wissenschaft unbrauchbar.

Während man bei der klassischen Testtheorie (siehe z.B. Krauth, 1995) davon ausgeht, dass Testergebnis und Merkmal in deterministischer Weise zusammenhängen, wird in der probabilistischen Testtheorie (IRT) von einem Wahrscheinlichkeitszusammenhang ausgegangen.

Bereits an dieser Stelle wird deutlich, warum sich die IRT besonders gut für die Untersuchung von Geschlechterunterschieden eignet:
Das biologische Geschlecht ist manifest, aber da es mit so vielen latenten Merkmalen zusammenhängt, ist ein deterministischer Zusammenhang zu den meisten erfassbaren Reaktionen nicht anzunehmen. Die meisten Verhaltensunterschiede zeigen sich überlappende Verteilungen, und nur wenige Verhaltensweisen sind nach Geschlechtern disjunkt:
Männer können kein Kind gebären oder säugen, diese Verhaltensweisen überlappen sich nicht. Aber bereits auf dem Gebiet der Kriminalität gibt es schon gewisse Überlappungsbereiche (Heinz, 2002), Frauen werden deutlich seltener kriminell als Männer. Bei emotionalem und kognitivem Verhalten sind die Überlappungsbereiche dann wiederum so groß, dass die Unterschiede z.T. nur mit komplizierten Verfahren überhaupt sichtbar werden (z.B. räumliche Leistungen, vgl. auch Abbildung 2).
Das geschlechtsspezifische Verhaltensspektrum besteht also zum größten Teil aus einer Schnittmenge zwischen Männern und Frauen und es existieren nur wenige disjunkte Mengen von Verhalten zwischen Mann und Frau.
Es muss davon ausgegangen werden, dass bei psychologisch interessantem Verhalten zum Geschlecht also nur ein Wahrscheinlichkeitszusammenhang besteht. Daher scheint eine Verwendung der IRT auf diesem Gebiet nur folgerichtig und geradezu geboten.
Die speziellen Annahmen und Voraussetzungen der IRT sollen in den folgenden Abschnitten unter den einzelnen Testmodellen vorgestellt werden.

Grob unterschieden werden Testmodelle danach, ob sie dichotome (zweikategorielle), nominale (qualitativ unterschiedliche kategoriale) oder ordinale (quantitativ unterschiedliche kategoriale) Itemantworten modellieren. Weiterhin können Itemkomponenten- und mehrdimensionale Modelle herangezogen werden sowie solche, die der Erfassung der Veränderungsmessung dienen. Die hier verwendeten wichtigsten Modelle sollen weiterhin näher beschrieben werden.

2.2 Das Rasch-Modell für dichotome Itemantworten (RM)

Das Rasch-Modell ist nach dem Dänen Georg Rasch benannt, der das Rasch-Modell in seiner Grundform für quantitative Variablen entwickelt hat (vgl. Fischer & Molenaar, 1995). Zentrale Größen im Rasch-Modell sind die Item- und Personenparameter, die Modellgleichung lautet wie folgt:

$$p(X_{vi} = 1) = \frac{\exp(\Theta_v - \sigma_i)}{1 + \exp(\Theta_v - \sigma_i)}$$

mit Θ_v = Personenparameter der Person v

und σ_i = Itemschwierigkeit von Item i.

Die Anwendung des Rasch-Modells setzt voraus, dass die Itemcharakteristiken der Daten monoton steigend sind. Die Wahrscheinlichkeit dass eine Person ein Item löst (bzw. im Sinne der Frage zustimmt) muss also mit steigender Fähigkeit größer werden. Die Items müssen lokal stochastisch unabhängig sein, das heißt, die Lösung eines Items darf nicht von der Lösung eines anderen abhängen. Weiterhin muss die Anzahl der gelösten Items eine erschöpfende Statistik (auch als suffiziente Statistik bezeichnet) für den Leistungsparameter Θ sein. D.h., die Summe aller gelösten Items enthält alle Informationen über die zu testende Person (hinsichtlich des latenten Merkmals), die benötigt werden, um die Personenfähigkeit zu schätzen.

Neben diesen Voraussetzungen gibt es einige Eigenschaften des Rasch-Modells bzw. eines Datensatzes, auf den das Rasch-Modell gilt:

Der Zusammenhang zwischen Personenmerkmal und Antwortwahrscheinlichkeit (der Lösung) eines Items wird in einer so genannten Itemcharakteristikkurve (kurz oft auch einfach: Itemcharakteristik oder ICC) dargestellt, die Lösungswahrscheinlichkeit wird also über alle Personen (bzw. über den gesamten Bereich der Personenvariable) aufgetragen, für Abbildungen siehe Rost (2004). In einer solchen Kurve kann man bereits etwas über die Itemschwierigkeit aussagen: Definiert ist diese durch den Wert, der bei der 50%-Lösungswahrscheinlichkeit die Kurve schneidet, abgelesen auf der Skala der Personenvariable. Hieran wird deutlich, dass Personen- und Itemparameter auf einer Skala liegen, was auch latente Additivität genannt wird. Die Itemschwierigkeit kann auch als Lokation eines Items bezeichnet werden, je höher die Itemschwierigkeit ist, desto weiter nach

rechts ist die ICC verschoben oder anders formuliert umso höher muss die Personenfähigkeit sein, um eine 50%-ige Lösungswahrscheinlichkeit zu erreichen. Dieser Zusammenhang ist auch in der Modellgleichung des Rasch-Modells nachzuvollziehen: Je größer die Itemschwierigkeit, desto kleiner die Lösungswahrscheinlichkeit (bei gleicher Personenfähigkeit).

Weiterhin ist die Trennschärfe des Items an einer solchen Kurve zu erkennen: Bei monotonen Itemfunktionen ist der Anstieg ein Ausdruck der Trennschärfe, da kleine Unterschiede in der Personenfähigkeit stark unterschiedlichen Lösungswahrscheinlichkeiten zugeordnet sind. Grafisch lässt sich also schnell ermitteln, wie schwer und wie trennscharf ein Item ist.

Eine wichtige Eigenschaft stellt die Beschränkung der Parameter dar. So wird eine homogene Population angenommen, für alle Personen gelten die gleichen Itemschwierigkeiten, die sich aber itemabhängig unterscheiden dürfen. Die Trennschärfe wird jedoch für alle Items als gleich angenommen, man spricht von parallelen Itemfunktionen: Die ICCs können verschoben sein (unterschiedliche Schwierigkeiten), müssen aber parallel verlaufen (gleiche Trennschärfe). Da es sich also um ein Personenmerkmal handelt, das von allen Items erfasst wird und bei allen Items in gleicher Weise „trennt“ oder quantifiziert spricht man beim Rasch-Modell von Eindimensionalität. Item- und Personenparameter werden dabei auf derselben Dimension abgebildet.

Aufgrund der mathematischen Entwicklung des Modells, der Logarithmierung der Parameter, liegen diese auf einer Differenzenskala. Um den Nullpunkt festzulegen, muss eine Form der Normierung angewandt werden. Man kann entweder die Summe der Itemparameter, der Personenparameter oder einen einzelnen Itemparameter auf Null festlegen. Konventionell wird die erste Möglichkeit gewählt:

$$\sum_i \sigma_i = 0$$

mit σ_i = Itemschwierigkeit von Item i.

Je nach Normierung können und müssen die berechneten Parameter in spezieller Weise interpretiert werden. In nachfolgenden Rechnungen werden stets die Itemparameter summennormiert, so dass einem Personenparameter von Null die Lösungswahrscheinlichkeit von 0,5 entspricht.

Wenn man Personen in ihrer Fähigkeit vergleichen will, so lassen sich Unterschiede in den Personenparametern erst nach Delogarithmierung direkt interpretieren. Dies kann allerdings unabhängig von der Art der Normierung, der Schwierigkeit der Testitems und der Fähigkeit

der anderen getesteten Personen vorgenommen werden. Diese Eigenschaft nennt sich Stichprobenunabhängigkeit oder Invarianz der Parameterwerte und ist selbstverständlich nur dann gültig, wenn das Rasch-Modell auf die Daten passt.
Wie die Schätzung der Modellparameter und eine Prüfung der Modellgeltung erfolgen, wird weiter unten erläutert. Für das Rasch-Modell existieren weiterhin Adaptationen für andere Daten als dichotome Items, die aber in der vorliegenden Arbeit unberücksichtigt bleiben.
Die Verwendung des Rasch-Modells hat also einige Vorzüge. Manchmal jedoch geht man von vornherein davon aus, dass es sich bei der gestesteten Population *gerade nicht* um eine homogene Gruppe handelt, in der jeweils dieselben Itemparameter gelten. So ist es durchaus berechtigt anzunehmen, dass gewisse Items leichter von Mädchen als von Jungen gelöst werden und umgekehrt. In diesem Fall scheint es angemessener, mit dem so genannten Mixed Rasch-Modell den Datensatz zu modellieren. Im folgenden Abschnitt soll näher darauf eingegangen werden.

2.3 Das Mixed Rasch-Modell (MRM)

Beim Mixed Rasch-Modell (im Folgenden oft kurz als MRM geschrieben) handelt es sich um das gemeinsame Obermodell des Rasch-Modells und der latenten Klassenanalyse, sichtbar an der folgenden Modellgleichung:

$$p(X_{vi}=1)=\sum_{g=1}^{G}\pi_g\,\frac{\exp(x_{vi}(\Theta_{vg}-\sigma_{ig})}{1+\exp(\Theta_{vg}-\sigma_{ig})}$$

mit π_g = Wahrscheinlichkeit der Zugehörigkeit zu Klasse g,

Θ_{vg} = Personenparameter der Person v in der latenten Klasse g,

und σ_{vi} = Itemschwierigkeit von Item i in der latenten Klasse g.

Während im Rasch-Modell von nur einer Klasse mit konstanten Itemparametern, aber unterschiedlichen Fähigkeitsausprägungen (quantitative Personenvariable) ausgegangen wird, werden in der latenten Klassenanalyse zwar unterschiedliche Klassen angenommen, allerdings dürfen sich die Personen innerhalb der Klassen nicht in ihren Antwortwahrscheinlichkeiten unterscheiden, es gibt also *eine* klassenspezifische Fähigkeitsausprägung (qualitative Personenvariable).

Beim MRM sind nun beide Vorzüge vereint: Der große Vorteil besteht also darin, dass gleichzeitig qualifiziert *und* quantifiziert werden kann. Das heißt, es identifiziert (in ihrem Antwortverhalten) qualitativ verschiedene Gruppen, in denen jeweils das Rasch-Modell gilt und die Fähigkeit quantifiziert werden kann. Dabei trennt es die Personengruppen gerade so, dass die Itemparameter maximal unterschiedlich sind und diese Aufteilung kann, muss aber nicht mit einem manifesten Teilungskriterium korrespondieren.

Der Vorteil dieses Modells liegt also darin, dass eine latente Heterogenität in den Daten aufgedeckt werden kann, die im probabilistischen Sinn für Unterschiede in den Daten verantwortlich ist.

Ausgehend von der Annahme, dass das Geschlecht als eine latent wirkende Struktur mehr Erklärungswert oder Vorhersagekraft besitzt als das biologische Geschlecht, sollte ein solches latentes Geschlecht mithilfe des MRM identifiziert werden können. Dem wird in weiter unten nachgegangen.

Die Eigenschaften des MRM sollen im Folgenden kurz umrissen werden.

Eine Summennormierung erfolgt jetzt innerhalb von Gruppen bzw. Klassen und die ermittelten Itemschwierigkeiten oder mittleren Lösungswahrscheinlichkeiten können in Itemprofilen so dargestellt werden, dass die Unterschiede zwischen Klassen sehr gut an Items festgemacht werden können. Hierbei ist zu beachten, dass aufgrund der Summennormierung *innerhalb* der Klassen die Itemprofile der Schwierigkeiten *zwischen* Klassen nicht im Niveau interpretiert werden können. Dies kann ausschließlich über die Profile der Lösungswahrscheinlichkeiten erfolgen. Allerdings gibt der relative Unterschied zwischen den Profilen Aufschluss über den Grad der Heterogenität der Daten. Je weiter die Profile auseinander liegen, desto größer ist der qualitative Unterschied zwischen den Klassen. Diese Eigenschaft wird sich für die Auswertungen in der vorliegenden Studie als nützlich erweisen.

Mittels einer Auswertung mit dem MRM wird pro Person als qualitativer Kennwert die wahrscheinlichste Klassenzugehörigkeit (sowie die Wahrscheinlichkeit dieser Zuordnung) und als quantitativer Kennwert ein Fähigkeitsparameter innerhalb der zugeordneten Klasse ermittelt. Dabei wird die Klassenzugehörigkeit nach dem Antwortmuster bestimmt, es wird die Klasse zugeordnet, in welcher die Wahrscheinlichkeit des Antwortpatterns am größten ist. Die Treffsicherheit der Klassenzuordnung wird über den Mittelwert der Zugehörigkeitswahrscheinlichkeiten über alle Personen für jede Klasse angegeben.

Die Anzahl der Klassen beim Mixed Rasch-Modell kann a priori festgelegt werden, wenn man genauere Hypothesen über die untersuchte Population hat. Dies ist sicherlich die elegantere Lösung als post hoc verschiedene Lösungen mit unterschiedlichen

Klassenanzahlen zu vergleichen, je nach Modellfit einer Lösung den Vorzug zu geben und post hoc zu erklären, warum gerade dieses Modell auf die Daten besser passt. Aber auch diese Methode kann dazu benutzt werden, die Daten in adäquater Weise zu modellieren und ein plausibles Erklärungsmodell zu entwerfen.

2.4 Modellgeltungskontrolle in der IRT

Die Modellpassung auf einen gegebenen Datensatz ist stets eine Frage, *wie gut* das Modell passt und nicht *ob* das Modell passt. Hierbei kann man sowohl statistische Maße als auch inhaltliche Kriterien zur Beurteilung heranziehen. Im Idealfall ergänzen sich beide Arten der Modellgeltungskontrolle, meist aber ist zwischen beiden abzuwägen. Rost (2004) stellt folgende Kriterien bei einer Prüfung der Modellpassung in den Vordergrund: Ein brauchbares Modell vereint einen hohen Erklärungswert und geringe Komplexität (wenige zu schätzende Modellparameter) mit einem gewissen Geltungsbereich innerhalb der speziellen Forschungsrichtung. Die folgenden Abschnitte beziehen sich nur auf die statistischen Modellgeltungsmaße.

Man unterscheidet zwischen inferenzstatistischen Maßen, informationstheoretischen Maßen und anderen spezielleren Modellgeltungstests.

Als inferenzstatistische Methoden zur Modellgeltungskontrolle können Likelihoodquotiententest, die Chi-^{2} Statistik und das so genannte Bootstrapping verwendet. Da in der vorliegenden Veröffentlichung weder die Chi-^{2} Statistik noch das Bootstrapping Verwendung finden, sei hierfür auf Rost (2004) verwiesen. Der Likelihoodquotiententest soll im Anschluss kurz umrissen werden.

Das Produkt der Patternwahrscheinlichkeit über alle Personen definiert die Likelihood (=Wahrscheinlichkeit), hierbei wird das Produkt nicht durch unbeobachtete Pattern beeinflusst. Die Likelihoodfunktion beschreibt die Wahrscheinlichkeit der Daten unter Annahme eines Modells:

$$L=P(Daten|Modell)$$

Plausibel ist, dass das Modell umso besser passt, je höher dieser Kennwert ist:

Wenn die Wahrscheinlichkeit der beobachteten Daten unter Annahme eines Modells höher ist als unter Annahme eines anderen Modells, wird das erste besser passen. Je höher also die Auftretenswahrscheinlichkeit der Daten, desto besser passt das Modell.

Der Likelihoodquotiententest (likelihood ratio test) vergleicht nun die Likelihoods unter zwei Modellen, die aus denselben Daten errechnet werden. Jede Veränderung der Daten, z. B. eine Kategorienzusammenlegung, ist nicht zulässig. Zum Vergleich wird ein Quotient (der likelihood ratio LR) gebildet, das Modell im Nenner ist ein Obermodell zum Modell im Zähler, das Zählermodell muss sich also durch Restriktion aus dem Nennermodell ergeben, welches für die Daten gültig sein muss und daher L_0 genannt wird:

$$LR = \frac{L_0}{L_1}$$

Unter diesen Voraussetzungen gilt:

$$-2\log(LR) \rightarrow \chi^2 \text{ mit } df = n_p(L_1) - n_p(L_0)$$

Anders formuliert:

$$\chi^2_{emp} = -2\log(LR)$$

$$\chi^2_{krit} = \chi^2 \text{ mit } df = n_p(L_1) - n_p(L_0)$$

Die Entscheidungsregel lautet:

Verwirf H_0, wenn $\chi^2_{emp} \geq \chi^2_{krit}$.

Neben den inferenzstatistischen Methoden zur Modellgeltungskontrolle und insbesondere dann, wenn die Voraussetzungen derselben nicht erfüllt sind, können informationstheoretische Maße zum Vergleich von Modellen herangezogen werden. Diese Maße haben den Nachteil, dass es keine Grenzwerte gibt, die eine eindeutige Entscheidung herbeiführen können, es muss stets abgewogen werden, ab wann man mit einem Modell zufrieden ist oder welcher Unterschied als bedeutsam zu interpretieren ist.

Die informationstheoretischen Maße sind in als Überblick in Tabelle 8 dargestellt.

Wie aus der Tabelle hervorgeht, fließen in die Informationskriterien dieselben Maße ein wie beim Likelihoodquotiententest. Das Vorgehen ist jedoch ein anderes: Für verschiedene Modelle werden die Kennzahlen verglichen und *je geringer* der Wert ist, desto besser passt das Modell auf die Daten. Dabei gibt es aber keine Richtlinien, um wie viel geringer dieser Wert sein muss, damit man sich für das Modell bzw. gegen ein anderes entscheidet. Theoretisch denkbar ließe sich großer Umfang an Modellen für die Daten verwenden, aber ob diese Modelle dann noch mit einer inhaltlichen Theorie harmonieren, sei dahingestellt. Das

heißt also, man wägt bei verschiedenen Modellen zwischen inhaltlicher und informationstheoretischer Modellpassung ab. In Tabelle 8 finden sich zwar Richtlinien zur Benutzung der Kriterien, doch die Auswahl folgt keinen festen Vorschriften. Insgesamt sind die informationstheoretischen Maße zur Modellgeltungskontrolle eher als grobe Auswahlkriterien zu betrachten, die aber eine hohe praktische Relevanz haben, da sie sehr schnell erste interpretierbare Hinweise liefern.

Tabelle 8 Eine Übersicht über informationstheoretische Maße zur Modellgeltungskontrolle. n_p gibt die Anzahl der Modellparameter an, N die Stichprobengröße, L steht für die Likelihood des Modells

Indexbezeichnung	Formel	Anwendungsgebiet
AIC Akaikes information criterion	**= -2 log L + 2n_p**	wenig Items große Patternhäufigkeiten
BIC Bayes information criterion / Schwartz criterion	**= -2 log L + (log N) n_p**	viele Items kleine Patternhäufigkeiten
CAIC constistent AIC	**=-2 log L+ (log N) n_p+ n_p**	große Stichprobe

Spezielle Tests für Rasch-Modelle finden sich überblicksartig sehr gut bei Rost (2004) dargestellt. Den meisten von Ihnen ist gemeinsam, dass sie zentrale Annahmen und Eigenschaften des Rasch-Modells prüfen. Als bekannte Vertreter seien hier der Andersen-Test zur Prüfung auf Personenhomogenität und der Martin-Löf-Test zur Prüfung der Itemhomogenität genannt. Interessanterweise lassen sich spezielle Tests häufig durch einfachere Verfahren ersetzen. So kann eine Modellpassung des Mixed Rasch-Modells, ausgehend von a priori Annahme über verschiedene Gruppen, bereits die Gültigkeit des Rasch-Modells hinreichend in Frage stellen. Dies ist insofern lohnenswert, da das Mixed Rasch-Modell ja gerade diejenigen Gruppen trennt, für welche die Itemparameter *maximal* unterschiedlich sind, wobei diese Trennung nicht mit einem manifesten Trennungsmerkmal korrespondieren muss.

Sind keine a priori Annahmen vorhanden, erscheinen spezielle Tests folgerichtig und zweckentsprechend.

Ein Hilfsmittel zur Veranschaulichung der Personenhomogenität ist der grafische Modelltest. Hierbei werden die Itemparameter von zwei Personengruppen in einem Kreuzdiagramm verglichen. Wenn die Itemparameter der beiden Gruppen perfekt übereinstimmen, liegen alle Werte auf einer Linie von 45 Grad. Je größer die Heterogenität zwischen den Gruppen ausgeprägt ist, desto größer sind die Abweichungen von einer 45 Grad-Linie. Gleichzeitig ist auch erkennbar, *welche* Items die Unterschiede zwischen den Personengruppen produzieren

und welche Gruppe diese ggf. leichter oder schwerer löst. Diese Methode ist gut geeignet, um einen ersten Eindruck vom Ausmaß einer möglichen Personenheterogenität zu bekommen. Allerdings muss bereits eine Hypothese darüber vorliegen, nach *welchem Kriterium* die Personen eingeteilt werden. Ein grafischer Modelltest wird weiter unten bei den Datenanalysen verwendet.

Obwohl es sich beim grafischen Modelltest mehr um eine Art deskriptive, veranschaulichende Methode handelt als einem Test, wäre es durchaus denkbar, daraus einen Test zu konstruieren: Mittels Distanzmaßen, ähnlich zu Regressionsmodellen könnte ein Kennwert entwickelt werden, mit dem die Abweichung quantifiziert werden könnte, dies ist in Abbildung 7 veranschaulicht.

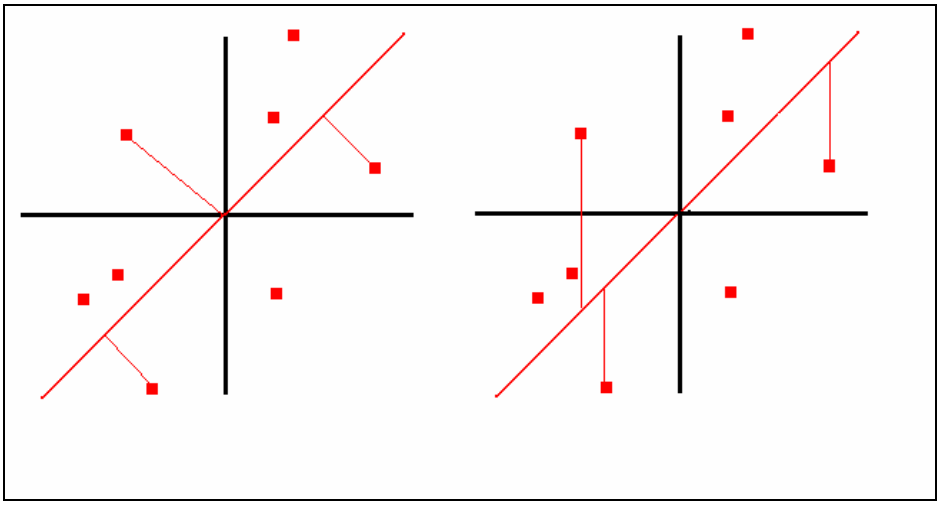

Abbildung 7 In einem solchen Sinne könnte eine Quantifizierung eines grafischen Modelltests entwickelt werden. Die Berechnung von Distanzmaßen könnte orthogonal zum Achsensystem oder orthogonal zur 45 Grad-Linie erfolgen.

2.5 Parameterschätzung bei Rasch-Modellen

Da die Parameterschätzung per se in folgenden Berechnungen nur eine untergeordnete Rolle spielt, soll diese hier nur in groben Zügen dargestellt werden.

Die Parameter werden geschätzt und nicht berechnet, Stichprobendaten können stets nur eine Schätzung von Populationsdaten liefern (vgl. z.B. Hays, 1994), die Aussagen sollen jedoch über Populationen gemacht werden.

Ausgangspunkt für die Parameterschätzung bei Rasch-Modellen ist die Likelihoodfunktion aller beobachteten Itemantworten.

Die Funktion weist dabei so viele Veränderliche auf wie es Parameter im Modell gibt. Alle Modellparameter werden auf den Wert festgelegt, an dem die Likelihoodfunktion ihr Maximum hat, so werden ihnen die wahrscheinlichsten Werte unter Annahme des Modells zuordnet. Diese so genannte Maximum Likelihood-Schätzung bestimmt die Modellparameter so, dass das *gemeinsame* Maximum für alle Modellparameter verwendet wird. Errechnet wird dies über Maxima mittels partiellen Differenzierens über Gleichungssysteme mit mehreren Unbekannten.

Die Auflösungen der Gleichungen erfolgen demnach in einem multidimensionalen Raum, der so viele Dimensionen hat wie es Modellparameter gibt. Ermittelt werden die Koordinaten des Maximums, das sind dann die Schätzwerte der Modellparameter.

Das ganze wird in einem iterativen Verfahren (= iterativ für schrittweise) errechnet, in dem immer wieder dieselben Rechenschritte durchgeführt werden bis man sich einer Lösung angenähert hat, die Lösung konvergiert. Dabei gibt es verschiedene Abbruchkriterien, z.B. Anzahl der Iterationen oder wenn die Werte sich von Iteration zu Iteration nur noch in einem bestimmten Betrag unterscheiden.

Hierbei existieren gewisse Konventionen für Abbruchkriterien, zu beachten ist aber auch der Verlauf der Funktion, die Sattelpunkte aufweisen oder mehrgipflig sein kann. In beiden Fällen könnte das Abbruchkriterium der Betragsänderung greifen und die Lösung das Maximum danach nicht erreicht haben.

Generell gibt es bei Rasch-Modellen drei Likelihoodfunktionen: UML (unbedingte/unconditonal Maximum Likelihood), CML (bedingte/conditional Maximum Likelihood), MML (marginal Maximum Likelihood). Weiterhin gibt es noch die Scoreverteilungslikelihood (SL), die sich aber aus der CML und den Scoreparametern multiplikativ zusammensetzt, genauer: Die SL ergibt sich aus conditonal Likelihood

multipliziert mit allen Scorewahrscheinlichkeiten, was in einem späteren Kapitel noch eine Rolle spielen wird.

Für die Schätzung der Itemparameter kann neben der CML auch eine explizite, nicht-iterative Methode verwendet werden.

Die Schätzung der Personenparameter ist mit größerem Aufwand verbunden als bei Itemparametern. Schätzer für die Personenparameter sind die MLE (maximum likelihood estimators), die WLE (weighted likelihood estimators) und die EAP (Expected a posteriori Schätzer), die auf der Funktion der MML basieren. Die so genannten plausible values sind Schätzer, die ebenfalls auf der Funktion der MML basieren, aber per Zufall aus der Funktion ermittelt werden und eine bessere Varianzschätzung als die EAPs liefern. Die Vor- und Nachteile der jeweiligen Methoden sowie eine detaillierte Beschreibung und Anwendungsmöglichkeiten sind bei Walter (2005) zu finden.

Im Rahmen von large scale assessment Studien ist die Parameterschätzung unter dem Gesichtspunkt von missing data zu betrachten. Zu den nicht bearbeiteten (non response) Aufgaben kommen noch designbedingte missings dazu, also Aufgaben, die nur einem Teil der Personenstichprobe vorgelegt wurden und dadurch nicht von allen Schülern bearbeitet wurden. Da in den meisten Fällen diese missings ausgewertet werden müssen, gibt es aufwändige Methoden der Imputation (sprachlich so viel wie: „Rechtfertigung“) von Daten, also Ersetzung fehlender Werte. Wie dies im Rahmen von PISA konkret erfolgt, ist bei der OECD (2005) nachzulesen.

3 Analysen zur Rolle des Geschlechts bei PISA 2003

In diesem Kapitel werden die vertiefenden empirischen Analysen zur Rolle der Geschlechterunterschiede bei PISA 2003, Fragestellungen und Hypothesen sowie Schlussfolgerungen vorgestellt. Der Aufbau des Kapitels ist in Abbildung 8 dargestellt.

Zunächst werden die Fragestellungen erläutert und die zugrunde liegende Stichprobe beschrieben. Daran schließen sich ein Geschlechtervergleich auf Itemebene, die Konstruktion geschlechtsspezifischer Skalen, die Bildung und Interpretation inhaltsspezifischer latenter Klassen sowie die Analyse inhaltsübergreifender kognitiver Strukturen an. Am Ende des Kapitels werden die Fragestellungen beantwortet.

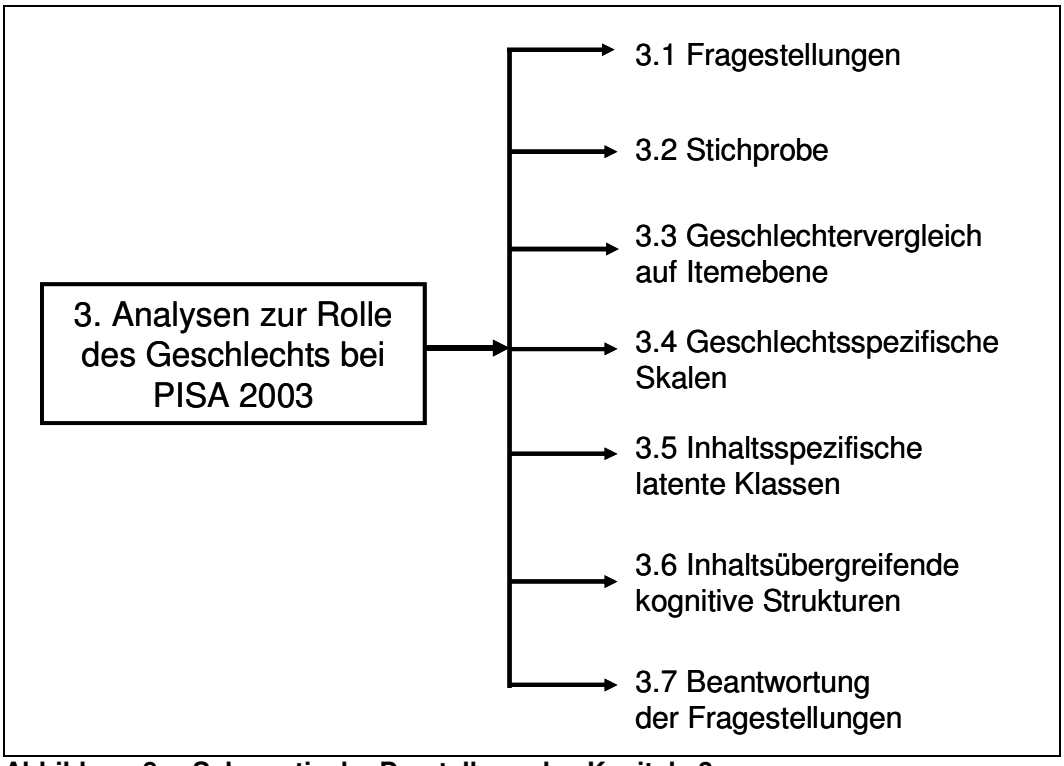

Abbildung 8 Schematische Darstellung des Kapitels 3.

3.1 Fragestellungen

Wie weiter oben bereits ausgeführt, bietet PISA zur Untersuchung von Geschlechterunterschieden die Vorteile einer hohen repräsentativen Stichprobe, die je nach Fragestellung auch international und kulturübergreifend verwendet oder auch auf die nationale Ebene beschränkt werden kann. Nachteile ergeben sich aus der sehr festgelegten Altersspanne (v. a. im Vergleich mit TIMSS) sowie einer engen Erfassung von Geschlecht, da lediglich das biologische (manifeste) Geschlecht, aber weder Einstellungen zu Geschlechterstereotypen noch eine Geschlechtsidentität im Sinne eines psychologischen Geschlechts erfasst wurde, das einige Zusammenhänge mit psychologischen Variablen nachweislich besser erklärt als das biologische Geschlecht (vgl. 1.2.2.2).

In den nachfolgenden empirischen Abschnitten sollen Geschlechterunterschiede bei PISA näher beleuchtet werden.

Im Zentrum des Interesses steht dabei die Frage, wie bedeutsam und welcher Art die Geschlechterunterschiede im PISA-Test sind. Bei den mittleren Kompetenzunterschieden (1.3.1.1) ist unklar, wodurch die Geschlechterunterschiede produziert werden. Liegt der Leistungsunterschied an unterschiedlich ansprechendem Aufgabenmaterial oder ist er eher vorwiegend in den kognitiven Anforderungen (z.B. verbal vs. visuell-räumlich) begründet, die den einzelnen Aufgaben zugrunde liegen?

Sollten sich Inhalte für Jungen und Mädchen als unterschiedlich ansprechend erweisen, bleibt zu klären, ob sich die postulierten stereotypen Muster darin wiederfinden oder ob sich Interessen- und Fähigkeitsschwerpunkte verändert haben. Es stellt sich die Frage, welche Items die Geschlechter trennen und was diese Aufgaben inhaltlich und anforderungsspezifisch charakterisiert. Diese Analyse interessiert nicht nur für die deutschen Schüler, sondern auch aus einer internationalen Perspektive. Anhand der länderübergreifenden mittleren Kompetenzunterschiede zwischen Mädchen und Jungen (Tabelle 2) sind je nach Vergleichsland verschiedene Ergebnisse zu erwarten.

Die Geschlechterunterschiede im nationalen Naturwissenschaftstest wurden bereits auf verschiedene kognitive Stärken bezogen (Rost et al., 2004). Mädchen zeigen bessere Leistungen in den verbal-bewertenden Teilkompetenzen, Jungen in den grafisch-numerisch-abstrakten. Die geschlechtsspezifischen Profile (Abbildung 5) unterscheiden sich aber nicht sehr deutlich von einander: Die größten Unterschiede betragen etwa 4 Kompetenzpunkte (nationale Metrik mit M=50 und SD=10).

Da sich die Verteilungen von Jungen und Mädchen stark überlappen, scheint das dichotome biologische Geschlecht anhand der in 1.2.2.2 diskutierten Befunde wenig Erklärungswert zu besitzen. Es erscheint sinnvoll, das biologische Geschlecht durch eine kontinuierliche Geschlechtsvariable im Sinne von *gender* zu ersetzen. Dabei stellt sich die Frage, ob sich die verbal-bewertenden und grafisch-numerisch-abstrakten Teilkompetenzen zu einer Skala zusammenfassen lassen, die sich post hoc als geschlechterkorreliert erweist und die Leistungsunterschiede bei PISA besser erklären kann als das biologische Geschlecht.

Obwohl das Geschlecht im Sinne von *gender* in einem Wahrscheinlichkeitszusammenhang zu Leistungsvariablen steht, wurden Geschlechterunterschiede nur selten auf latenter Ebene mittels probabilistischer Testmodelle erfasst. Hinsichtlich der kognitiven Teilkompetenzen im nationalen Naturwissenschaftstest ist zu klären, ob sich eine Trennung von grafisch-numerisch-abstrakten („männlichen“) und verbal-bewertenden („weiblichen“) Teilkompetenzen auch durch latente Testmodelle nachweisen lässt. Das Mixed Rasch-Modell identifiziert bedeutsame Klassen, sofern Unterschiede in den Antwortmustern der Schüler vorhanden sind. Sollten heterogene (z.B. geschlechtsspezifische) Personengruppen Unterschiede in den Antworten produzieren, müssten sie mit dem Mixed Rasch-Modell identifiziert werden. Die zugrunde liegende latente Trennungsvariable kann inhaltlich beschrieben und auf ihren Bezug zum Geschlecht und zu anderen zentralen PISA-Variablen geprüft werden.

Da sich die Inhaltsbereiche thematisch stark unterscheiden, kann davon ausgegangen werden, dass sie Jungen und Mädchen unterschiedlich ansprechen. Die Analysen müssen deshalb zunächst für jeden Inhaltsbereich getrennt, dann global über alle Inhaltsbereiche erfolgen.

Sollte sich mit latenten Modellen kein oder nur ein geringer Bezug zum Geschlecht herausstellen, scheint das (biologische) Geschlecht für die Teilkompetenzen im nationalen Naturwissenschaftstest nur von untergeordneter Bedeutung zu sein.

In den folgenden Abschnitten 3.3 bis 3.6 werden die genannten Fragestellungen aufgegriffen und unter Formulierung von Hypothesen mittels empirischer Methoden beantwortet.

3.2 Stichprobe (Verortung in PISA)

Die folgenden Ausführungen zur Stichprobenziehung sind an die nationalen und internationalen Berichte (OECD, 2004; OECD, 2005; Prenzel et al., 2004a) angelehnt und können dort im Detail nachgelesen werden.

Um eine internationale Vergleichbarkeit zu gewährleisten, wurden strenge Kriterien bei der Auswahl der Schüler von der OECD vorgeschrieben und deren Einhaltung auf internationaler Ebene überwacht.

Im Rahmen von PISA wurden allein für die international vorgeschriebene Stichprobe in allen Staaten insgesamt ca. 250 000 Schüler getestet, in Deutschland waren es 4660.

Definiert wurde die Testpopulation über das Lebensalter, so dass *unabhängig* von Klassen oder Leistungsstufen Fünfzehnjährige in Bildungseinrichtungen, zum Testzeitpunkt im Alter zwischen 15;3[1] und 16;2 Jahren, einbezogen wurden.

Es muss unterschieden werden zwischen Aufnahme in die Stichprobe (Erfassung der Zielpopulation), Stichprobenziehung und Teilnahme, dies gilt jeweils für Schulen und Schüler.

Bei der Erfassung und Stichprobenziehung wurde Wert auf einen hohen Ausschöpfungsgrad gelegt, so dass auf internationaler Schulebene bis auf wenige Ausnahmen weniger als 3%, auf Schülerebene weniger als 6% von der Aufnahme in die Stichprobenziehung ausgeschlossen wurden. Verzerrrungen der Leistungsmessungen durch den Ausschlussanteil werden nach der OECD auf weniger als 5 Kompetenzpunkte des nationalen Durchschnitts geschätzt, sind also als eher gering zu beurteilen.

Die Stichprobenziehung selbst erfolgte über eine mehrfach stratifizierte Wahrscheinlichkeitsstichprobe. Stratifikation bedeutet dabei Schichtung, stratifizieren heißt, in eine Schichtenfolge einordnen. Hierbei wurden verschiedene Stratifizierungsvariablen je nach Land berücksichtigt (OECD, 2005). Einzelne Schüler wurden nur dann ausgeschlossen, wenn eine geistige Behinderung vorlag, eine körperliche Behinderung die Testteilnahme unmöglich machte oder die Testsprache nicht genügend beherrscht wurde, also die Testsprache nicht Muttersprache war und der Schüler weniger als ein Jahr in dieser Sprache beschult wurde. Andere Gründe bedurften der Abstimmung durch das internationale Konsortium.

[1] Verwendet wird an dieser Stelle die u.a. in der psychologischen Diagnostik verwendete Nomenklatur, bei der die Anzahl der Lebensmonate mit einem Semikolon vom Lebensjahr abgetrennt ist.

In Deutschland beteiligten sich 216 Schulen, es gingen nach Abschluss der Testung die Daten von 92% aller gezogener Schüler ein, eine Aufschlüsselung dieser endgültigen Stichprobe nach Schularten findet sich bei Prenzel et al. (2004b). Der Einwand einer möglichen selektiven Beteiligung leistungsstärkerer Schüler, welche eine Verzerrung der Ergebnisse zu besseren Leistungen zur Folge hätte, kann zumindest in Deutschland ausgeschlossen werden. Diesbezüglich zeigt die Arbeit von Kienzl (2005), dass die Noten der teilgenommenen Schüler sich nur geringfügig von denen aller gezogenen Schüler unterscheiden.

In der vorliegenden Publikation wird ausschließlich auf die internationale Schülerstichprobe Bezug genommen, eine Übersicht zu den nationalen Erweiterungen der Stichprobe geben Prenzel et al. (2004b).

3.3 Welche Items trennen die Geschlechter? – Ein Geschlechtervergleich auf Itemebene

Mit Blick auf die Ausführungen in Abschnitt 1.2 und die bei PISA 2003 vorgelegten Inhaltsbereiche (vgl. 1.3.1 und Prenzel et al., 2004a), lassen sich folgende Fragestellungen formulieren:

- *Bestätigen sich die Annahmen aus dem theoretischem Teil auch bei den PISA-Aufgaben?*
- *Lassen sich in den PISA-Aufgaben die geschlechtstypischen Muster wiederfinden?*
- *Gilt dies sowohl für die kognitiven Anforderungen (verbal vs. visuell-räumlich) als auch für die Aufgabeninhalte (Thema Autos, Gesundheit etc.)?*
- *Welche Items trennen die Geschlechter und was charakterisiert diese Items inhaltlich?*
- *Sind geschlechtsspezifische Itemunterschiede über verschiedene Länder hinweg stabil?*

Folgendes Muster ist zu erwarten: Mädchen sollten solche Aufgaben leichter lösen, welche die Interessen von Mädchen beinhalten (Umweltthemen, Ernährung und Fortpflanzung etc.), insbesondere solche, die verbal präsentiert sind.
Jungen sollten solche Aufgaben leichter lösen, die grafisch-räumliche Problemlösestrategien erfordern und die für sie ansprechende jungenstereotype Inhalte (z.B. Autos, Abenteuer etc.) beinhalten.
Eine geschlechtsspezifische Schwierigkeit müsste sich in den Unterschieden der Itemparameter von Jungen und Mädchen zeigen. Ziel der folgenden Analysen ist demnach die Identifikation und inhaltliche Charakterisierung von geschlechtsspezifischen Einzelitems im Haupttest von PISA 2003. Mittels des grafischen Modelltests lässt sich die geschlechterbezogene Heterogenität der untersuchten Population bei PISA 2003 klären, zunächst in der Stichprobe der Jugendlichen in Deutschland, danach in einem Vergleich mit mehreren Ländern.

3.3.1 Geschlechtsspezifische Items bei deutschen Schülern

Die Items, die in den Haupttest eingingen, wurden im Feldtest erprobt und bereits auf eine gender-by-item interaction geprüft, allerdings ist im technical report der OECD nicht beschrieben, nach welchen Kriterien die Items letztendlich selektiert wurden bzw. ab welchem Wert Items entfernt wurden (OECD, 2005). Die Selektion von Items richtet sich bei

einem solchen Vorgehen danach, welche Items sich als ungewöhnlich schwer oder ungewöhnlich leicht für eine Teilgruppe gezeigt haben. Dies wird innerhalb der IRT auch als differential item functioning (DIF) bezeichnet.

Für PISA 2000 findet sich eine international übergreifende Darstellung des differential item functioning nach Geschlecht (Adams & Carstensen, 2002). Allerdings wird hier den Unterschieden in den Einzelitems nicht inhaltlich auf Itemebene nachgegangen.

Innerhalb der Berichtsbände von PISA 2003 sind Geschlechterunterschiede bisher vor allem auf Gesamtskalenebene untersucht worden.

Im Folgenden soll gezeigt werden, wie für beide Geschlechter *getrennt* Itemparameter über das Rasch-Modell ermittelt werden, um einen grafischen Modelltest zwischen Jungen und Mädchen vorzunehmen. Die Leistungswerte in den internationalen Domänen Mathematik, Naturwissenschaften, Problemlösen und Lesen sowie für das nationale Zusatzmaterial Naturwissenschaften und Mathematik der PISA-I Stichprobe (15-jährige) wurden für deutsche Schülerinnen und Schüler ohne Verwendung eines Hintergrundmodells nach Geschlechtern getrennt skaliert. Dies veranschaulicht Abbildung 9.

Verwendet wurde die Software Conquest mit denselben Spezifikationen, mit denen die Daten bei PISA 2003 skaliert wurden: Die Items wurden summennormiert, d.h. die Summe der Items auf Null gesetzt (vgl. dazu die Abschnitte über IRT). Die Iterationen wurden bei einem Maximum von 1000 oder einer Änderung der Parameter von weniger als 0,01 abgebrochen.

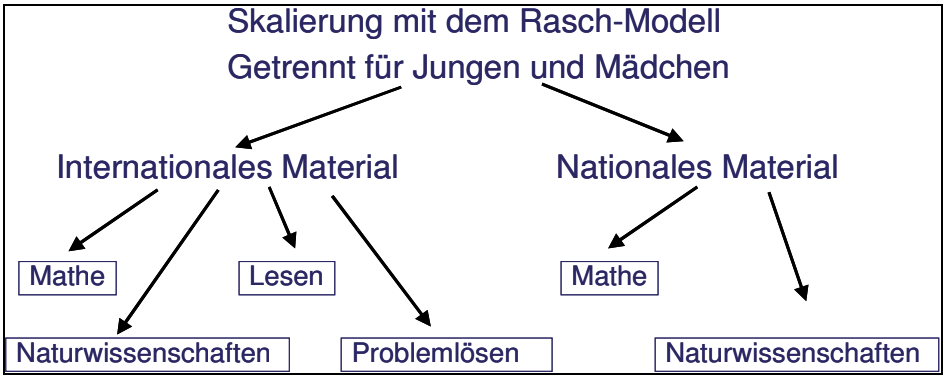

Abbildung 9 Darstellung der geschlechtsspezifischen Skalierung für die Jungen und Mädchen in Deutschland. Das internationale Material wurde komplett skaliert, beim nationalen Material nur die Bereiche Naturwissenschaften und Mathematik.

Die grafischen Modelltests für die einzelnen Domänen im Rahmen einer geschlechtsspezifischen Skalierung mit dem Rasch-Modell zeigen eine gewisse Heterogenität zwischen den Geschlechtern auf. Sie sind in Abbildung 10 bis Abbildung 15 dargestellt. Die Gerade ist die Winkelhalbierende, auf der alle Items liegen müssten, wenn die Schwierigkeiten gleich sind (vgl. dazu die Bedeutung eines grafischen Modelltests bei Rost,

2004). Wie deutlich zu sehen ist, trifft dies auf die Mehrzahl aller Items zu. Dennoch können einige Items ausgemacht werden, bei denen je nach Geschlecht Unterschiede in ihrer Schwierigkeit auftreten.

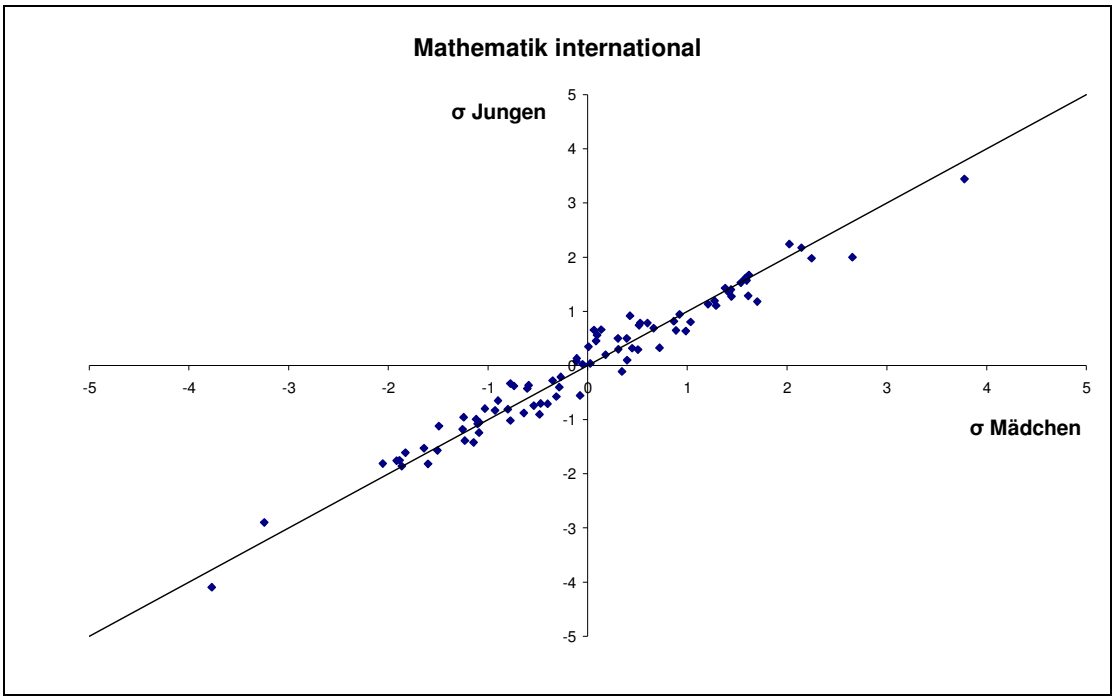

Abbildung 10 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Mathematiktests (Stichprobe: deutsche Schüler).

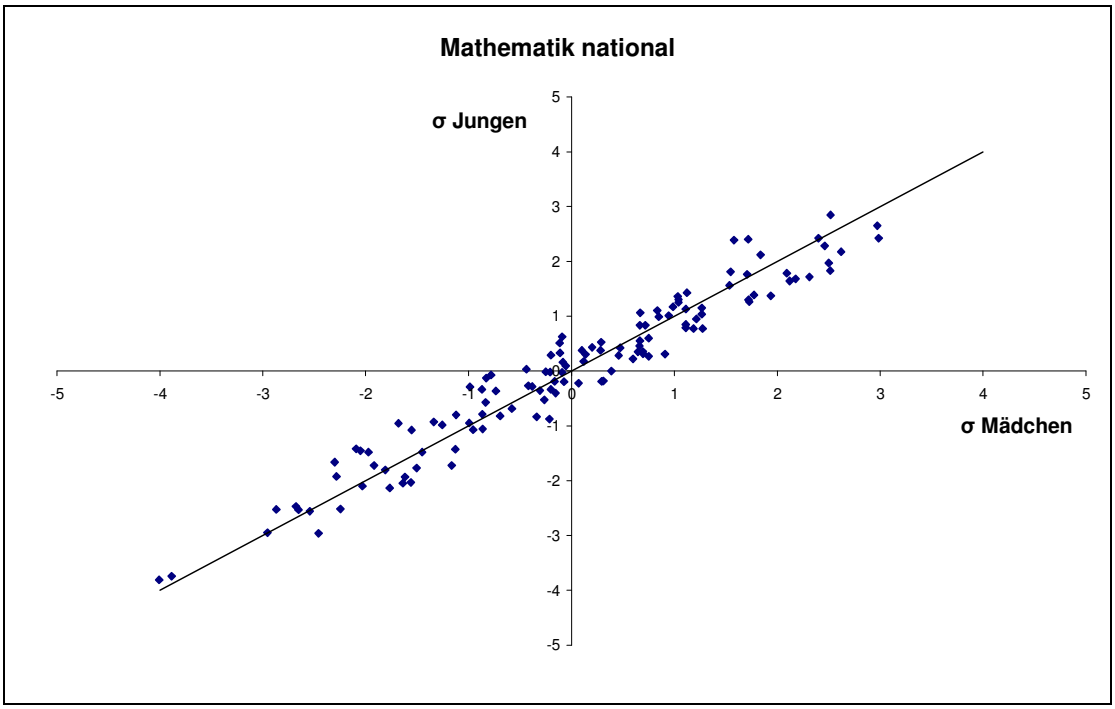

Abbildung 11 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des nationalen Mathematiktests (Stichprobe: deutsche Schüler).

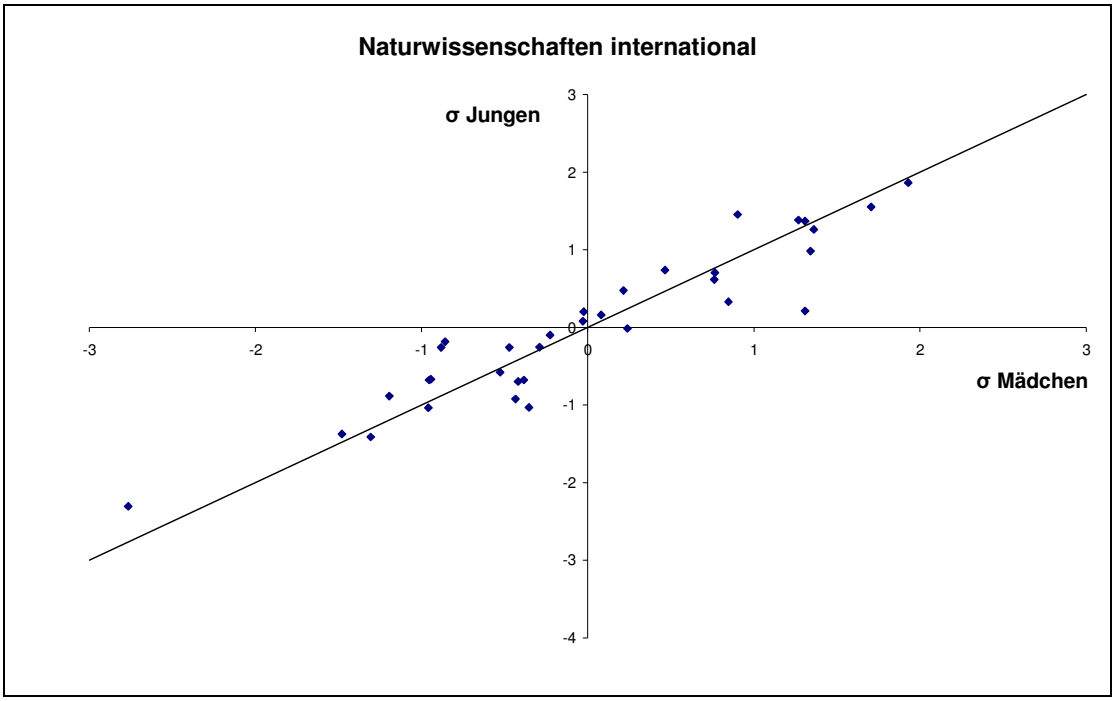

Abbildung 12 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Naturwissenschaftstests (Stichprobe: deutsche Schüler).

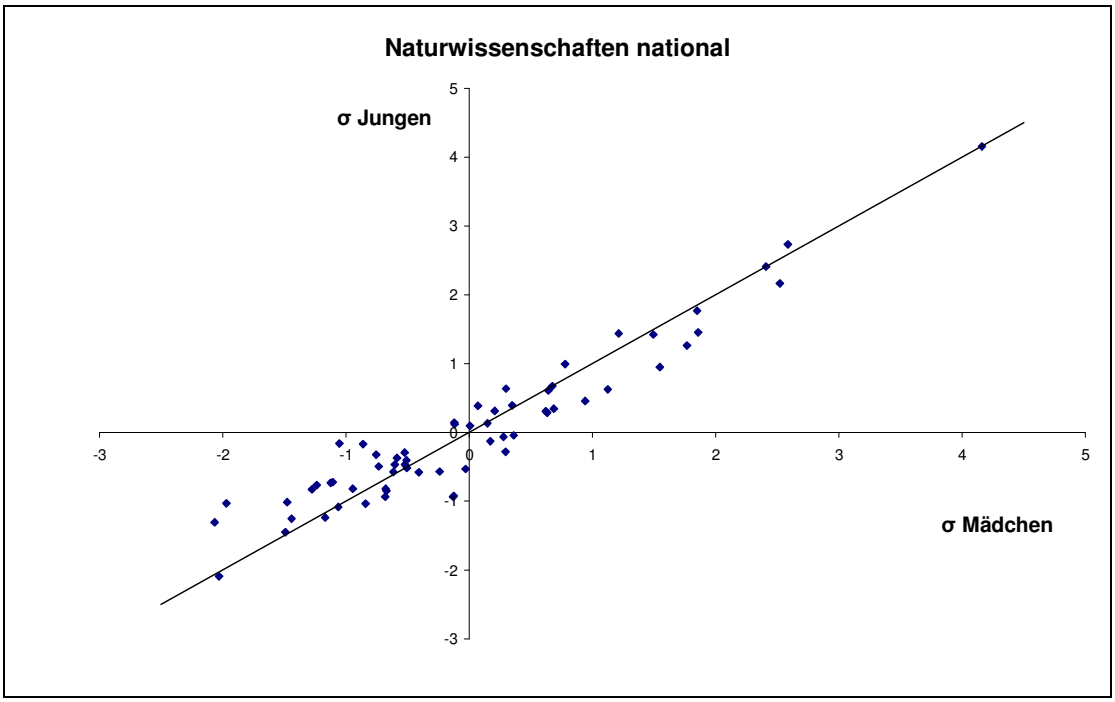

Abbildung 13 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des nationalen Naturwissenschaftstests (Stichprobe: deutsche Schüler).

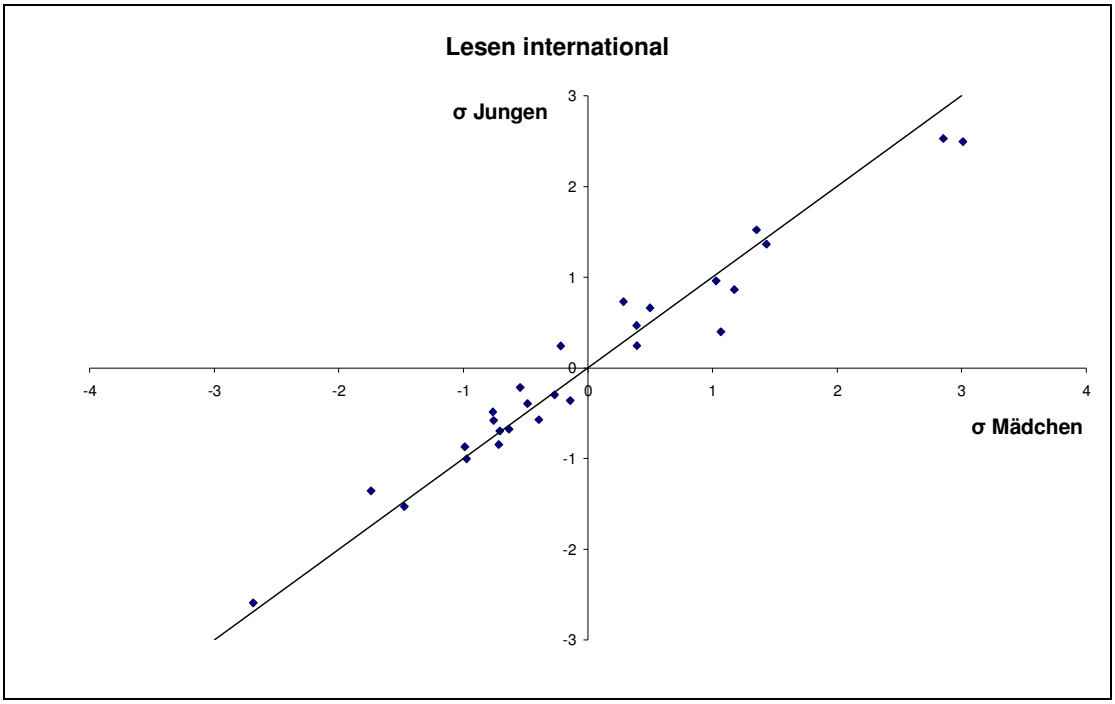

Abbildung 14 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Lesetests (Stichprobe: deutsche Schüler).

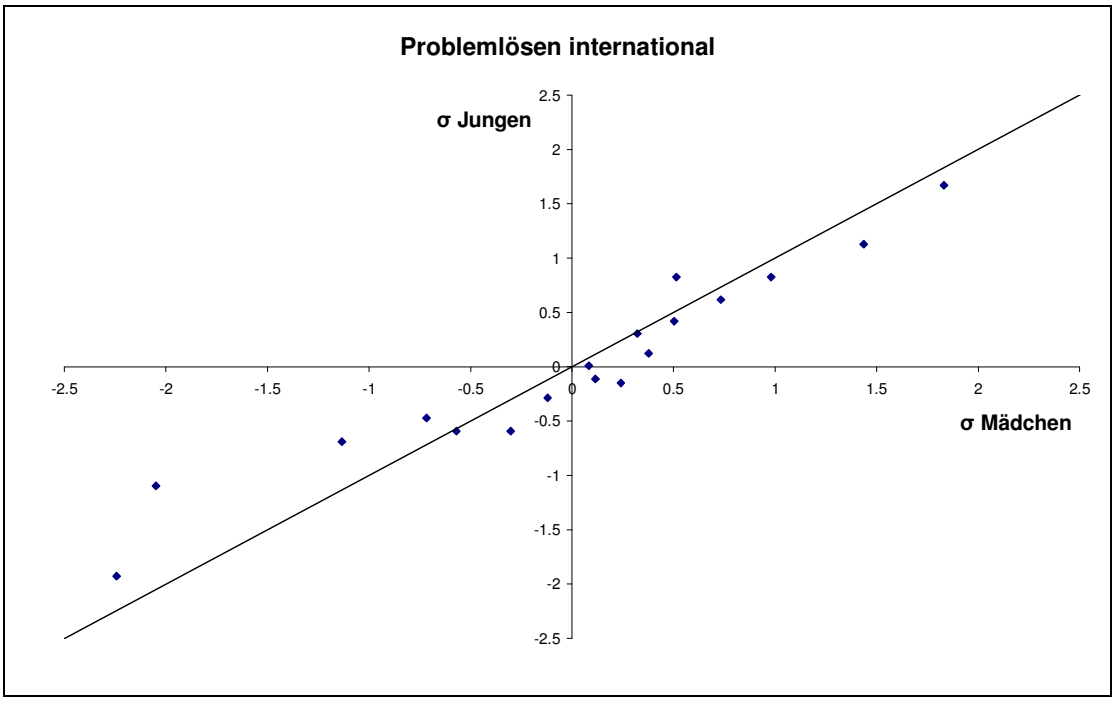

Abbildung 15 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Problemlösetests (Stichprobe: deutsche Schüler).

Über den Abstand von der 45°-Linie lassen sich die Items identifizieren, die von jeweils einem Geschlecht leichter gelöst werden können. Als Kriterium, um ein Item als geschlechtsspezifisch zu bezeichnen, wird ein Unterschied in der Itemschwierigkeit von mind. 0,5 auf der logit-Skala angesetzt, was in etwa 35,5 Skalenpunkten auf der internationalen Skala gleichkommt: Nach Knoche & Lind (2004) entsprechen 0,1 logits etwa 7,1

Skalenpunkte innerhalb der internationalen Metrik. Allerdings gilt dieser Wert streng genommen nur für den Fall, dass ein Hintergrundmodell berechnet wurde, was bei vorliegenden Analysen nicht zutrifft.

Dieser willkürliche Wert von 0,5 logits kann in der Tat als substanziell bezeichnet werden, wie bei Draba (1977) beschrieben.

Die Items, bei denen Mädchen eine geringere Itemschwierigkeit zeigen (also diejenigen Items, die leichter von Mädchen gelöst werden) sollen hier als mädchenspezifische Items bezeichnet werden, dasselbe soll umgekehrt für jungenspezifische Items gelten.

Tabelle 9 Anzahl der Items, bei denen sich nach geschlechtsspezifischer Skalierung die Itemparameter zwischen Jungen und Mädchen in Deutschland um mindestens 0,5 logits unterscheiden sowie die Anzahl aller Items. Das Item wird dem Geschlecht zugeordnet, das einen geringeren Schwierigkeitsparameter aufweist (=das Item leichter löst).

Inhaltsdomäne	Anzahl jungenspezifischer Items	Anzahl mädchenspezifischer Items	Anzahl aller Items
Mathematik international	2	2	84
Mathematik national	10	12	124
Naturwissenschaften international	3	3	34
Naturwissenschaften national	6	4	63
Lesen international	2	0	28
Problemlösen international	0	1	27
$\sum$	**23**	**22**	**360**

Tabelle 9 zeigt wie viele der Items aus welchen inhaltlichen Domänen geschlechtsspezifisch sind, Tabelle 10 die zugehörigen Aufgabenkürzel.

Wie aus Tabelle 10 hervorgeht, sind 45 von insgesamt 360 hier untersuchten Items als geschlechtsspezifisch zu betrachten, also etwa ein Achtel aller Items. Dies ist einerseits ein relativ geringer Anteil, andererseits ein beträchtlicher, wenn man bedenkt, dass nur solche Items für den Haupttest verwendet wurden, die sich nach dem Feldtest als geschlechterfair erwiesen haben.

Die Ergebnisse stützen die Befunde von Knoche & Lind (2004), allerdings wird hier eine detaillierte inhaltliche Itemanalyse angeschlossen.

Mit einem Blick auf den Inhalt der Items lassen die eingangs postulierten Hypothesen größtenteils bestätigen:

Tabelle 10 Aufgeführt sind die Items, bei denen sich nach geschlechtsspezifischer Skalierung die Itemparameter zwischen Jungen und Mädchen in Deutschland um mindestens 0,5 logits unterscheiden. Zur Veröffentlichung international freigegebene Aufgaben sind unterstrichen. In deutschen Quellen publizierte Aufgaben sind mit * versehen, vgl. Tabelle 35 im Anhang.

Itemlabel	Inhaltsdomäne	leichter gelöst von
M421Q02, M803Q01	Mathematik international	Jungen
MPROZ1, MSPAR1, MTISC2*, MVIDE2, MONLI1, MONLI3 MSWIM1, MREEC1, MVERS1, MDACH2	Mathematik national	Jungen
M510Q01, M603Q02	Mathematik international	Mädchen
MBRUC1*, MFUNK1, MFUNK3, MKAUF1, MSUSA1, MQUAD1, MRECH1, MTINT1, MSECD1, MSECP2, MSECT3, MLIGL1	Mathematik national	Mädchen
S269Q03, S269Q04, S327Q01	Naturwissenschaften international	Jungen
NATMM, NAUTOK*, NAUTOM*, NBADZ, NENRGD, NMULLG	Naturwissenschaften national	Jungen
S128Q03*, S131Q04, S133Q01	Naturwissenschaften international	Mädchen
NAUTOB, NBADB, NSEXB, NWEIZB	Naturwissenschaften national	Mädchen
R104Q05, R220Q01	Lesen international	Jungen
X402Q01*	Problemlösen international	Mädchen

Im Bereich der Naturwissenschaften fallen mädchenspezifische Items vor allem durch den Kontext zu Ernährung, Gesundheit und Verhütung auf, weiterhin werden die Ergebnisse gestützt, die bereits bei PISA 2003 (Rost et al., 2004) berichtet sind: So erfordern mädchenspezifische Items oftmals komplexe Bewertungs- und Entscheidungsprozesse (kognitive Teilkompetenz „Bewerten“) oder sind durch Aufgabenstellungen mit Informationsentnahme aus Textpassagen charakterisiert. Ein Beispiel für ein Item, das Mädchen leichter lösen, ist „Das geklonte Schaf Dolly“ (internationales Testmaterial).

Jungen lösen Aufgaben leichter, in denen Physikkenntnisse von Vorteil sind oder/und räumliches Vorstellungsvermögen angesprochen wird oder mentale Modelle benutzt werden müssen. Auch scheint der Kontext - die inhaltliche Einkleidung - bei diesen Aufgaben eher die Jungen angesprochen zu haben (z.B. Autos). Ein Beispiel für ein jungenspezifisches Item ist „Im Straßenverkehr: Vorsicht bei vereister Fahrbahn!“.

Die hier genannten Aufgaben finden sich bei Prenzel et al. (2004a)[2], vgl. dazu auch Tabelle 35 im Anhang.
Insgesamt werden also in der Inhaltsdomäne der Naturwissenschaften die Geschlechterstereotype eher bekräftigt.

In Mathematik können die Hypothesen nur teilweise bestätigt werden. Bekräftigt werden die Hypothesen durch folgende Befunde:
Jungen lösen solche Aufgaben leichter, in denen räumliches Vorstellungsvermögen gefordert ist, jungenspezifische Aufgaben sind v.a. den Stoffgebieten Geometrie und Stochastik zuzuordnen. Ein Beispiel für ein jungenspezifisches Mathematik-Item ist das der Flächenberechnung einer Tischdecke (publiziert in Blum et al., 2004a). Bei Bearbeitung dieser Aufgabenstellung steht weniger ein für Jungen besonders ansprechender Inhalt, sondern mehr die kognitive geometrisch-räumliche Repräsentationen des Aufgabenmaterials im Vordergrund.
Ebenso erwartungskongruent verhält sich eine Aufgabe zu Lebensmitteln, wegen seiner geschlechtsspezifische inhaltliche Einkleidung als mädchenspezifisches Item interpretiert werden kann.
Erwartungskonträr verhalten sich zwei Aufgaben aus der Geometrie, die Mädchen besser lösen, obwohl sie das räumliche Vorstellungsvermögen ansprechen und über keine besondere inhaltliche mädchentypische Einkleidung verfügen.
Technische Aufgaben treten in etwa gleichen Anteilen sowohl mädchenspezifisch wie auch jungenspezifisch auf. Ein freigegebenes und publiziertes technisches Mädchenitem ist „Bruchrechnung“ (veröffentlicht bei Blum et al., 2004a).
Insgesamt lässt sich zur Domäne Mathematik also sagen, dass Stereotype z. T. bestätigt werden, aber kein eindeutiges Muster bei technischen Aufgaben vorliegt und auch geometrische Aufgaben mädchenspezifisch auftreten.

In der Inhaltsdomäne Lesen gibt es trotz des hohen Lesevorsprungs der Mädchen (vergleiche dazu Zimmer et al., 2004) keine mädchenspezifischen Aufgaben, dafür zwei jungenspezifische Aufgaben. Die beiden Items zeichnen sich dadurch aus, dass zur Bearbeitung der Aufgaben über die Lesekompetenz hinaus entweder visuell-räumliches

[2] Nicht freigegebene PISA-Aufgaben können aufgrund der Verschwiegenheitspflicht an dieser Stelle nicht genannt oder explizit inhaltlich beschrieben werden. Dies hat den Grund, dass die Aufgaben bei PISA 2006 nochmals zum Einsatz kommen und eine Verbreitung in der Öffentlichkeit die Ergebnisse verfälschen könnte. Eine Anzahl veröffentlichter Aufgaben mit ihren Lösungen findet sich auch unter http://pisa.ipn.uni-kiel.de/., vgl. dazu auch Tabelle 35 im Anhang.

Vorstellungsvermögen oder der Umgang mit Tabellen gefordert ist, um alle Informationen der Aufgaben verarbeiten zu können. Innerhalb dieser Anforderungen treten im Durchschnitt die Stärken von Jungen besser hervor (siehe 1.3.1.3). Ein Item befasst sich mit einem spannenden Abenteuer einer Exkursion in fremde Länder, auch das andere Item hat Reisen zum Thema. Bei beiden Aufgaben scheint das Thema der Aufgabe besonders Jungen anzusprechen und die kognitive Anforderung zur Bewältigung der Aufgaben kommt den Stärken der Jungen entgegen.

Die Ergebnisse zum Problemlösen weisen auf ein geschlechtsspezifisches Item zu Gunsten der Mädchen hin; bei 27 Items fällt dies kaum ins Gewicht. Allerdings erfordert gerade diese Aufgabe bei der Bearbeitung den Umgang mit Grafiken bzw. einem mentalen Modell, verhält sich demnach eher erwartungskonträr. Möglicherweise kommt hier aber auch der Vorsprung der Mädchen im Lesen stärker zum Tragen, da die Aufgabe hohe verbale Fähigkeiten voraussetzt.

Die Geschlechtsspezifität zeigt sich nach obigen Befunden oft in einer Kombination aus ansprechendem geschlechterstereotypen Inhalt und geschlechtsspezifischer kognitiver Anforderung. Die oben genannten Hypothesen wurden also teilweise bestätigt.

Die Anzahl der geschlechtsspezifischen Items reicht aufgrund des Multimatrix-Testdesigns (nicht jeder Schüler hat jede Aufgabe bearbeitet, siehe dazu Aufbau und Organisation von PISA, 1.3.1) nicht aus, um daraus eine geschlechterkorrelierte Skala zu bilden. D.h. die geschlechterspezifischen Items lassen sich nicht zu einem aussagekräftigen Wert zusammenzufassen. Außerdem besteht die Möglichkeit, dass die Geschlechtsspezifität der Aufgaben ein Charakteristikum der deutschen Stichprobe ist. Dem soll im nächsten Abschnitt näher nachgegangen werden.

3.3.2 Geschlechtsspezifische Items im internationalen Vergleich

Um zu klären, ob die geschlechtsspezifischen Itemunterschiede über verschiedene Länder hinweg stabil sind, wurden die Leistungswerte in den Domänen Mathematik, Naturwissenschaften, Problemlösen und Lesen der PISA-I Stichprobe für die Schüler aus insgesamt sieben Ländern (Deutschland, Österreich, Finnland, Island, Niederlande, USA, Japan) analog zum vorherigen Abschnitt nach Geschlechtern getrennt skaliert. Die Auswahl an Ländern erfolgte nach folgenden auf Deutschland bezogenen Gesichtspunkten: Österreich als deutschsprachiges Vergleichsland, Niederlande mit einem ähnlichen Schulsystem und relativ geringen Geschlechterunterschieden (vergleiche Burba & Rost, 2006; Zimmer et al., 2004), Finnland als Spitzenreiter mit eher großen Geschlechterunterschieden, Island als dem einzigen Land, in dem Mädchen signifikant höhere Mathematikleistungen als Jungen erbringen sowie USA und Japan als Länder mit kontrastierenden Schulsystemen im Vergleich zu Deutschland.

Hierbei konnte nur das internationale und nicht das nationale Testmaterial berücksichtigt werden, da nur die deutschen Schüler das in Deutschland entwickelte nationale Instrumentarium bearbeitet haben.

Es ergeben sich für jedes Land vier grafische Modelltest (je einer für Mathematik, Lesen, Naturwissenschaften und Problemlösen), also insgesamt 28 Grafiken. Da dies den Rahmen dieses Abschnitts sprengen würde, sind in Tabelle 11 die Items aufgeführt, die (analog zu den Rechnungen im vorhergehenden Abschnitt) einen Unterschied in der Itemschwierigkeit um 0,5 logits zwischen den Geschlechtern aufweisen und zwar in der Mehrzahl der Länder, also in mindestens vier der sieben ausgewählten Länder. Die Abbildungen für die deutschen Schüler finden sich in Abbildung 10 bis Abbildung 15, Abbildung 30 bis Abbildung 53 im Anhang zeigen die grafischen Modelltests der anderen Vergleichsländer.

Obwohl nun verschiedene Länder miteinbezogen werden, gibt es noch immer eine Anzahl von Items, die von jeweils einem Geschlecht leichter gelöst werden. Übereinstimmungen in der Geschlechterspezifität existieren auch über Länder hinweg.

Das Item s269r04t aus dem Bereich der Naturwissenschaften ist sogar bei allen sieben ausgewählten Ländern ein jungenspezifisches Item. Bei der Lösung spielen Kenntnisse in Physik und Chemie eine Rolle, die Einkleidung ist an Technologie orientiert. Bei diesem Item scheint also gerade die Kombination aus inhaltlicher Einkleidung und kognitiven Anforderungen Jungen anzusprechen.

Tabelle 11 Aufgeführt sind die Items, bei denen sich nach geschlechtsspezifischer Skalierung die Itemparameter zwischen Jungen und Mädchen um mindestens 0,5 Logits unterscheiden und zwar in mindestens vier der sieben ausgewählten Länder. Zur Veröffentlichung international freigegebene Aufgaben sind unterstrichen. Die in deutschen Quellen publizierte Aufgabe ist mit * versehen, vgl. Tabelle 35 im Anhang.

Itemlabel	Inhalts-domäne	leichter gelöst von	Länder						
			DEU	AUT	FIN	ISL	JPN	NLD	USA
M192Q01	Mathematik	Jungen		✕	✕	✕		✕	✕
M421Q02	Mathematik	Jungen	✕		✕	✕	✕		
M510Q01	Mathematik	Mädchen	✕	✕	✕	✕			✕
M598Q01	Mathematik	Mädchen		✕	✕		✕	✕	✕
M603Q02	Mathematik	Mädchen	✕	✕	✕	✕		✕	
R220Q01	Lesen	Jungen	✕	✕	✕		✕	✕	
S269Q04	Naturwissensch.	Jungen	✕	✕	✕	✕	✕	✕	✕
S327Q01	Naturwissensch.	Jungen	✕		✕	✕			✕
S128Q03*	Naturwissensch.	Mädchen	✕	✕	✕			✕	✕

Eine auffällige Besonderheit stellt Item R220q01 dar. Dies ist das einzige Item in der Inhaltsdomäne Lesen, das über mindestens 4 Länder hinweg von einem Geschlecht leichter gelöst werden kann. Neben der Lesekompetenz erfordert die Aufgabe den Umgang mit Grafiken und Zahlen sowie eine grafisch-räumliche Vorstellung bereits bei der Aufgabenstellung. Wie bereits oben in der deutschen Stichprobe diskutiert, spricht auch der Inhalt eher Jungen an: Es geht um das Abenteuer einer Exkursion in fremde Länder.

Bei jedem der in Tabelle 11 aufgeführten Items sind entweder inhaltliche Einkleidung oder kognitive Anforderungen für die Geschlechterspezifität verantwortlich zu machen. Sie stützen die oben formulierten Hypothesen.

Es zeigen jedoch nur 5 von 84 internationalen Mathematikaufgaben, 3 von 34 Naturwissenschaftsaufgaben, nur eine von 28 Leseaufgaben und *keines* der 27 Problemlöse-Items über Länder hinweg stabile Geschlechtereffekte in der Itemschwierigkeit, ein vergleichsweise geringer Anteil. Die Ergebnisse deuten insgesamt darauf hin, dass die Geschlechterunterschiede im Aufgabenmaterial von PISA als eher gering zu beurteilen sind.

Die Annahmen aus Abschnitt 1.2 wurden teilweise durch die PISA-Aufgaben auf Itemebene bestätigt: Geschlechtstypische Muster lassen sich sowohl hinsichtlich der kognitiven Anforderungen als auch der inhaltlichen Einkleidung von Items kongruent zu den eingangs formulierten Hypothesen finden. Wie oben ausgeführt, gibt es auch Items, die sich erwartungskonträr verhalten.
Es bleibt zu beachten, dass sich die Geschlechtsspezifität nur bei einem geringen Anteil von Items gezeigt hat, für einige Items allerdings über eine Auswahl von Ländern hinweg stabil.
Die geringe Anzahl geschlechtsspezifischer Items ist möglicherweise ein Effekt der geschlechterfairen Auswahl von Items über Erprobungen im Feldtest. Dennoch treten Geschlechterunterschiede in den Leistungen auf, wie vielfach gezeigt (Burba & Rost, 2006; Zimmer et al., 2004).

3.4 Welchen Erklärungswert haben geschlechtsspezifische Skalen für die Leistungsvariablen bei PISA?

Wie weiter oben (vgl. Abschnitt 1.3.1.2) beschrieben, unterscheiden sich die Geschlechter in den kognitiven Teilkompetenzen des nationalen Naturwissenschaftstests in einem auffälligen Profil (siehe Abbildung 5).
Ausgehend vom Konzept der psychologischen Androgynie (vgl. 1.2.2.2), stellen sich folgende Fragen:

- *Lässt sich eine geschlechtskorrelierte Skala aus den Teilkompetenzen konstruieren, welche die Leistungsvariablen bei PISA besser erklärt als das biologische Geschlecht?*
- *Hängt eine nicht dichotome geschlechtskorrelierte Variable mit den Leistungsunterschieden stärker zusammen als das biologische Geschlecht?*

Werden die kognitiven Teilkompetenzen geschlechtsassoziiert zusammengefasst, so dass jeder Schüler eine Ausprägung auf einer Weiblichkeitsdimension, einer Männlichkeitsdimension und einer Differenzenskala besitzt, sollte dieser Wert die Leistungsvariablen besser erklären, da die zugrundeliegende Variable nicht - wie das biologische Geschlecht - dichotom ist, sondern ein breiteres Spektrum an geschlechtsspezifischen Ausprägungen abbildet.
Dies sollte sich darin zeigen, dass die Leistungswerte mit der um den Leistungsanteil bereinigten Differenzenskala stärker korrelieren als mit dem Geschlecht.

Um die Fragestellungen zu beantworten, wurden die Teilkompetenzen „Sachverhalten verbalisieren“, „Divergentes Denken“ und „Bewerten“ zu einer weiblichen Dimension (nachfolgend **W-Skala** genannt) zusammengefasst, „Konvergentes Denken“, „Umgang mit Grafiken“, „Umgang mit Mentalen Modellen“ und „Umgang mit Zahlen“ zu einer männlichen Dimension (nachfolgend **M-Skala** genannt). Die Werte wurden addiert und gemittelt. Die Differenz der beiden Skalenwerte bildet die Differenzen-Skala (oder **D-Skala**). Tabelle 12 zeigt die zugeordneten kognitiven Teilkompetenzen.

Tabelle 12 Zugeordnete kognitive Teilkompetenzen für die Weiblichkeits- , Männlichkeits- und Differenzenskala.

W Skala = Weiblichkeitsskala	Bewerten, Divergentes Denken, Sachverhalte verbalisieren
M-Skala = Männlichkeitsskala	Umgang mit Zahlen, Mentale Modelle, Konvergentes Denken, Umgang mit Grafiken
D-Skala = Differenzenskala	M-Skala – W-Skala

Die Berechnungen wurden mit den bei PISA berechneten Plausible Values (PVs) durchgeführt (vgl. Carstensen et al., 2004; Walter, 2005).

Die neu gebildete D-Skala reflektiert die Geschlechterunterschiede in den kognitiven Teilkompetenzen und ist– im Gegensatz zum biologischen Geschlecht – nicht dichotomisiert, sondern kontinuierlich. Dies entspricht den Überlegungen aus Kapitel 1.2: Die (psychologische) Geschlechtszugehörigkeit ist gerade nicht als dichotom aufzufassen.

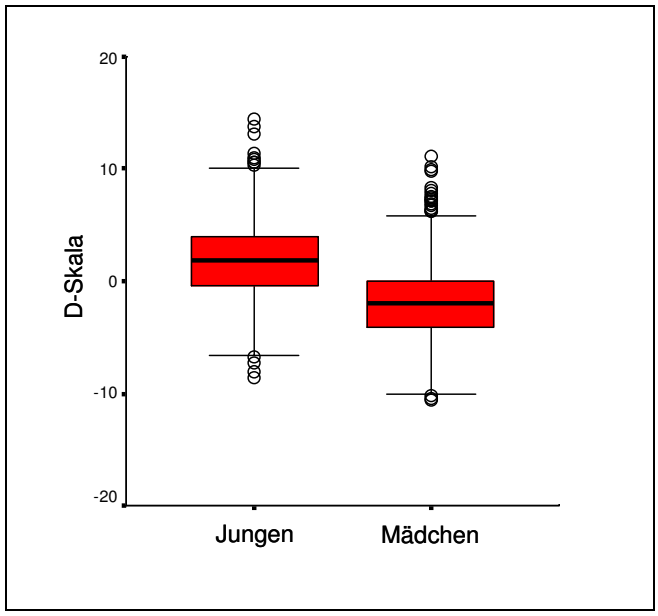

Abbildung 16 Dargestellt ist die Verteilung der Jungen und Mädchen auf die D-Skala. Die Box zeigt den Interquartilabstand, d.h. sie enthält alle Werte zwischen dem 1. und 3. Quartil. Die fett gedruckte Linie in der Box bildet den Median ab. Die Kreise ober- und unterhalb zeigen Ausreißerwerte, die mindestens 1,5 Boxlängen vom 1. bzw. 3. Quartil entfernt liegen.

Wie sich Jungen und Mädchen auf die D-Skala verteilen, ist in Abbildung 16 dargestellt. Die mittleren D-Werte für Jungen und Mädchen finden sich Tabelle 13.

Es zeigt sich, dass Jungen - wie zu erwarten - ein höherer mittlerer D-Wert als Mädchen zugeordnet wird, die Verteilungen von Jungen und Mädchen auf der D-Skala überschneiden sich.

Tabelle 13 Mittlerer Wert auf der D-Skala für Jungen und Mädchen

	mittlerer Wert der D-Skala
Jungen	1,77
Mädchen	-1,88

Um die Konstruktvalidität der Skalen zu prüfen, werden sie auf die Kriterien der konvergenten und diskriminanten Validität in Bezug zur manifesten Geschlechtsvariable untersucht. Die Korrelationen der Skalen untereinander und mit dem Geschlecht finden sich in Tabelle 14.

Tabelle 14 Dargestellt sind Korrelationen zwischen dem biologischen Geschlecht und aus den kognitiven Teilkompetenzen des nationalen Naturwissenschaftstests konstruierten Skalen sowie Partialkorrelationen und Interkorrelationen zwischen den Skalen. Die Erklärung der Abkürzungen findet sich in Tabelle 12 und dem Text.

Korrelationen		pv
r (Geschlecht - W-Skala)	r_{GW}	-0.114
r (Geschlecht - M-Skala)	r_{GM}	0.083
r (Geschlecht - D-Skala)	r_{GD}	0.387
r^2(Geschlecht - D-Skala)	r_{GD}^2	0,150
r (W-Skala - M-Skala)	r_{WM}	-0.872
partial r (Geschlecht - W-Skala ohne M-Skala)	$r_{GW.\ M}$	-0.381
partial r (Geschlecht - M-Skala ohne W-Skala)	$r_{GM.\ W}$	0.374

Die Korrelationen zwischen Geschlecht und W- bzw. M-Skala fallen gering aus, wie den ersten beiden Zeilen der Tabelle zu entnehmen ist, die Korrelation zwischen Geschlecht und Differenzenskala liegt bereits etwas höher (dritte Zeile). Betrachtet man die Korrelation zwischen M- und W-Skala (vierte Zeile), liegt der Schluss nahe, dass ein großer Anteil der M-

bzw. W-Skala allein leistungsbedingt und nicht geschlechtsspezifisch ist. Es existiert also ein großer gemeinsamer Anteil der beiden Skalen, der in die Korrelation mit dem Geschlecht einfließt. Nimmt man diesen gemeinsamen Anteil nach dem Prinzip der Partialkorrelation heraus, vergrößert sich die Korrelation zwischen Geschlecht und (um den Einfluss der M-Skala bereinigter) W-Skala bzw. (um den Einfluss der W-Skala bereinigter) M-Skala, was in den letzten beiden Zeilen der Tabelle 14 abzulesen ist.

Nach Auspartialisierung der jeweils anderen Skala entspricht die Partialkorrelation derjenigen zwischen Geschlecht und D-Skala. Wegen der Differenzbildung kann man die D-Skala bereits als um den Anteil der Leistungsunterschiede bereinigt auffassen. Quadriert man die Korrelation (r_{GD}) von –0,387, ergibt dies einen Anteil von ca. 15% an der Differenzenskala ($r_{GD}{}^2$), der allein durch das Geschlecht erklärt wird.

Die Korrelation zwischen D-Skala und Geschlecht ist zufrieden stellend; somit ist bei der geschlechterkorrelierten (aber nicht -determinierten) D-Skala die Konstruktvalidität gegeben. Um abzuschätzen, ob eine Reduktion der W- bzw. M-Skala zu einer bedeutenden Verbesserung der Korrelation führt, wurden nur die stark nach Geschlechtern differenzierenden Teilkompetenzen in die Skalen mit eingeschlossen, also nur „Sachverhalte verbalisieren“ und „Bewerten“ für die W-Skala sowie „Mentale Modelle“ und „Umgang mit Zahlen“ für die M-Skala. Dies brachte lediglich einen Gewinn von 5% in der aufgeklärten Varianz ($r_{GD}{}^2$), die Korrelationen sind in Tabelle 15 gezeigt. Dieser Gewinn geht aufgrund der Verringerung der Skalenlänge jedoch zu Lasten der Reliabilität. Daher wird von einer Reduktion der Skalen abgesehen und weiterhin die in Tabelle 12 beschriebenen Skalen verwendet (Korrelationen aus Tabelle 14).

Tabelle 15 Korrelationen zwischen dem biologischen Geschlecht und den konstruierten Skalen sowie Partialkorrelationen und Interkorrelationen zwischen den Skalen. Die Erklärung der Abkürzungen findet sich in Tabelle 12 und dem Text.

Korrelationen		pv
r (Geschlecht - W-Skala)	r_{GW}	-0,159
r (Geschlecht - M-Skala)	r_{GM}	0,114
r (Geschlecht - D-Skala)	r_{GD}	0,436
r ^{2}(Geschlecht - D-Skala)	$r_{GD}{}^2$	0,190
r (W-Skala - M-Skala)	r_{WM}	-0,804
partial r (Geschlecht - W-Skala ohne M-Skala)	$r_{GW.\ M}$	-0,425
partial r (Geschlecht - M-Skala ohne W-Skala)	$r_{GM.\ W}$	0,413

Die hohe Interkorrelation zwischen der W- und M-Skala von 0,87 zeigt mit umgerechnet 76% (ergibt sich aus $0,87^2$) den Anteil der gemeinsamen Varianz an, ein solcher Anteil wird durch die gemeinsame Leistung erklärt.

Die Vorhersagekraft für Leistungsunterschiede bei PISA zeigt sich an den Korrelationen zwischen D-Skala und den Leistungsskalen bei PISA (Tabelle 16), zum Vergleich sind dort die Korrelationen des manifesten Geschlechts mit den Leistungsvariablen angegeben. Die Werte unterscheiden sich nur geringfügig: Die D-Skala kann die Leistungsunterschiede nicht besser erklären als das biologische Geschlecht, hinsichtlich der Lesekompetenz sogar schlechter.

Tabelle 16 Angegeben sind die Korrelation zwischen Geschlecht, D-, M-, W-Skala und den internationalen Leistungsdomänen Mathematik, Lesen, Naturwissenschaften und Problemlösen.

	Mathematik	Lesen	Naturwissenschaften	Problemlösen
Geschlecht	0,046	-0,184	0,017	-0,009
D-SKALA	0.061	-0.091	0.015	0.018
W-SKALA	0.807	0.785	0.796	0.781
M-SKALA	0.851	0.752	0.817	0.803

Es zeigen sich geringe Korrelationen zwischen D-Skala und den Leistungsskalen, während die M-Skala und W-Skala mit den Leistungsdomänen eine hohe Korrelation aufweisen. Dies legt die Schlussfolgerung nahe, dass die (um den allgemeinen Leistungsanteil bereinigte) D-Skala zwar die auf den kognitiven Teilkompetenzen im nationalen Naturwissenschaftstest beruhenden Geschlechterunterschiede enthält, diese selbst aber nicht groß genug sind, um für den Leistungsunterschied in den internationalen Inhaltsdomänen eine Rolle zu spielen.
Diese Befunde verhalten sich konträr zur eingangs formulierten Hypothese. Die aus den kognitiven Teilkompetenzen des nationalen Naturwissenschaftstests konstruierte Differenzenskala kann die Leistungsunterschiede bei PISA nicht besser erklären als das manifeste Geschlecht.

3.5 Inhaltsspezifische latente Klassen

In den vorherigen Abschnitten wurde deutlich, dass die Leistungsunterschiede bei PISA nur in geringem Ausmaß mit dem biologischen (manifesten) Geschlecht korrelieren (siehe Tabelle 16).

Die in den geschlechtsspezifischen Profilen der kognitiven Teilkompetenzen des nationalen Naturwissenschaftstests dargestellten Unterschiede sind so gering, dass sich keine geschlechterspezifische Skala daraus bilden lässt. Aus Abbildung 5 geht hervor, dass (innerhalb der nationalen Metrik mit M=50 und SD=10) die größten geschlechtsspezifischen Kompetenzwertunterschiede in den kognitiven Teilkompetenzen bei etwa 4 Kompetenzpunkten liegen.

Bei einem Vergleich der Teilkompetenzen überlappen sich die Verteilungen von Jungen und Mädchen stark (s. Abbildung 2). Die Ausführungen aus Abschnitt 1.2.2.2 legen nahe, dass die biologische Erfassung des Geschlechts Unterschiede in kognitiven Leistungen weniger Erklärungswert besitzt als das psychologische Geschlecht, z.B. *gender* oder androgyne Klassen. Das biologische Geschlecht definiert nicht per se die Leistungsunterschiede. Es bedingt in einem Wahrscheinlichkeitszusammenhang die kognitiven Strukturen, die zu Unterschieden in der Aufgabenbearbeitung führen. Das psychologische Geschlecht ist ein Korrelat der kognitiven Strukturen.

Folglich sind die Antwortmuster im nationalen Naturwissenschaftstest (Abbildung 5) ein „Produkt“ der Unterschiede in kognitiven Strukturen.

Es erscheint nahe liegend, dass ein auf den Antwortmustern basierendes latentes Merkmal zur Erklärung der Daten herangezogen wird und auf den Bezug zum biologischen Geschlecht untersucht wird.

Unter Annahme einer latenten zugrunde liegenden Variable sowie eines Wahrscheinlichkeitszusammenhangs zwischen dieser Variable und der naturwissenschaftlichen Leistung bietet sich die Untersuchung mittels probabilistischer Testmodelle an. Das Mixed Rasch-Modell analysiert Personenunterschiede auf latenter Ebene. Es trennt über die Unterschiede in den Antwortmustern die Klassen so, dass die Itemparameter zwischen den Klassen maximal unterschiedlich sind.

Eine solche Trennung von Personengruppen auf Basis der Antwortmuster bietet den Vorteil, dass genau das Teilungskriterium herangezogen wird, das zur Erklärung der Daten bedeutsam ist. Zeigt sich das Geschlecht mit dem Teilungskriterium assoziiert, kann das Geschlecht als

relevant zur Erklärung der Leistung interpretiert werden. Spielt es für die Trennung von Personen keine bedeutende Rolle, sind die im nationalen Naturwissenschaftstest gefundenen Geschlechterunterschiede zu vernachlässigen,

Folgenden Fragestellungen soll nachgegangen werden:

- *Ist eine Trennung der Personen in verschiedene Klassen auf Basis der kognitiven Teilkompetenzen im nationalen Naturwissenschaftstest von PISA 2003 sinnvoll? Passt das Mixed Rasch-Modell auf die Daten?*
- *Lässt sich auf Basis der kognitiven Teilkompetenzen im nationalen Naturwissenschaftstest eine latente Variable identifizieren, welche die Daten in verschiedene Klassen trennt? Welches Merkmal trennt die Klassen? Kann das latente Merkmal als psychologisches Geschlecht aufgefasst werden?*
- *Unterscheiden sich die geschlechtsspezifischen Profile der kognitiven Teilkompetenzen stärker, wenn man als Teilungskriterium das latente Merkmal anstelle des biologischen Geschlechts heranzieht?*
- *Welche Klassenlösung passt am besten auf die Daten?*
- *Ist die Klassenlösung mit dem Geschlecht assoziiert?*

Den Fragestellungen liegen folgende Hypothesen zugrunde:
Es ist anzunehmen, dass sich die kognitiven Teilkompetenzen in verbal-bewertend und grafisch-numerisch-abstrakt einteilen lassen und diese Kategorien auch für die geschlechtsspezifische Unterschiede relevant sind (vgl. Rost et al., 2004), und 1.3.1.2).
Eine latente Klasse sollte sich durch hohe Lösungswahrscheinlichkeiten in den verbal-bewertenden Teilkompetenzen (Bewerten, Divergentes Denken, Sachverhalte verbalisieren) auszeichnen, die andere latente Klasse durch hohe Lösungswahrscheinlichkeiten in den grafisch-numerisch-abstrakten Teilkompetenzen (Umgang mit Zahlen, Umgang mit Grafiken, Konvergentes Denken, Mentale Modelle).

Im Profil der Teilkompetenzen sollten deutlichere Unterschiede auftreten, wenn man nicht nach dem manifesten Geschlecht teilt, sondern nach dem latenten Merkmal.
Laut dieser Hypothese sollte die Zweiklassenlösung nach dem Mixed Rasch-Modell am besten auf die Daten passen. Das Mixed Rasch-Modell mit zwei latenten Klassen sollte sich gegenüber anderen Modellen durchsetzen und der Modellfit sich durch

Likelihoodquotiententests und informationstheoretische Maße (s. Punkt 2.4) in allen Inhaltsbereichen des Tests nachweisen lassen.

Konkurrierende Modelle sind: **Das Rasch-Modell für dichotome Daten** (Einklassenlösung) als Vergleichsmodell mit der Annahme einer homogenen Personengruppe dies entspricht (unter Annahme einer Personenhomogenität) dem Standardverfahren bei PISA, die **Einteilung nach dem Geschlecht** (zwei manifeste Klassen), um einen Vergleich mit der üblicherweise verwendeten, manifesten Geschlechtsvariable zu ermöglichen, die **Einteilung nach Leistung** (Hoch- und Niedrigscorer), um Leistungsklassen in den Vergleich mit einzubeziehen und d**as Mixed Rasch-Modell mit drei latenten Klassen**, um das Zweiklassen Mixed Rasch-Modell gegen mehr Klassen abzusichern. Bei der Dreiklassenlösung könnte eine dritte Klasse eine androgyne Klasse von Schülern repräsentieren, die in beiden Teilkompetenzgebieten (verbal-bewertend und grafisch-numerisch-abstrakt) gut abschneiden.

Im nationalen Naturwissenschaftstest wurden unterschiedliche Inhaltsbereiche vorgelegt (vgl. Abbildung 4). Es ist zu klären, ob die genannten Hypothesen für jeden Inhaltsbereich gelten. Dabei ist davon auszugehen, dass die Themenwahl des jeweiligen Inhalts in unterschiedlicher Weise die Schüler anspricht und auf diesem Weg die Itemparameter und Lösungswahrscheinlichkeiten bedingt (vgl. Abschnitt 3.3).

Die einzelnen Schritte zur Beantwortung der Fragestellungen werden im Folgenden dargestellt.

3.5.1 Berechnung von Klassenlösungen

Mittels des vollständig gekreuzten Facettendesigns wurde jede Teilkompetenz in jedem Inhalt erhoben, pro Inhaltsbereich sieben Items, ein Item entspricht einer kognitiven Teilkompetenz.

Da die Inhaltsbereiche des nationalen Naturwissenschaftstests bereits durch die inhaltliche Einkleidung für die Geschlechter unterschiedlich ansprechend sind, werden die Analysen für jeden Inhalt getrennt durchgeführt.

Aufgrund des Multimatrix-Testdesigns schwanken die Schülerzahlen pro Inhaltsbereich. Tabelle 17 zeigt, wie viele Schüler welchen Inhaltsbereich bearbeitet haben.

Tabelle 17 Anzahl der Schüler pro Inhalt. Aufgrund des Multimatrix-Testdesigns wurden nicht allen Schülern alle Inhalte vorgelegt.

Inhaltsbereich	Anzahl der Schüler
Atmung und Fotosynthese	1321
Bewegungsgesetze (Mechanik)	995
Biochemie und Ernährung	986
Chemische Verbindungen und Aggregatzustände	976
Elektrizität	1322
Energieumwandlung	1318
Fortpflanzung und Sexualität	1301
Teilchenkonzept	980
Wärmeverlust, Wärmegewinnung und Energiesparen	1311

Tabelle 18 Den Schülern im Test vorgelegte Blöcke von Inhaltsbereichen, die in verschiedenen Kombinationen erfasst worden sind.

Block 1	Atmung	Teilchenkonzept	Energieumwandlung	Ernährung
Block 2	Atmung	Elektrizität	Wärmeverlust	Chem. Verbindungen
Block 3	Bewegungsgesetze	Elektrizität	Wärmeverlust	Ernährung
Block 4	Bewegungsgesetze	Elektrizität	Energieumwandlung	Atmung
Block 5	Teilchenkonzept	Elektrizität	Wärmeverlust	Fortpflanzung
Block 6	Fortpflanzung	Energieumwandlung	Chem. Verbindungen	Ernährung
Block 7	Fortpflanzung	Bewegungsgesetze	Chem. Verbindungen	Teilchenkonzept
Block 8	Fortpflanzung	Atmung	Energieumwandlung	Wärmeverlust

Das Mixed Rasch-Modell mit zwei latenten Klassen wurde mit der von Matthias von Davier entwickelten Software „Winmira“ (www.winmira.von-davier.de) auf die dichotomen Itemantworten pro Inhaltsbereich angewendet. Folgende Abbruchkriterien wurden verwendet: maximal 250 Iterationen, Parameterveränderung in der Schätzung von 0,0005.

Bei drei der Inhaltsbereiche (Atmung und Fotosynthese, Energieumwandlung, Teilchenkonzept) wurden mehr als 250 Iterationen benötigt, in diesen Fällen wurde in Nachberechnungen ein Iterationsmaximum von 1000 festgelegt, das dann nicht mehr überschritten wurde. Nicht bearbeitete oder nicht gelöste Aufgaben wurden als missing behandelt.

Um die Ergebnisse des Mixed Rasch-Modells einordnen zu können und zu validieren, wurden vier zusätzliche Modelle für jeden Inhaltsbereich berechnet:

- **Das Rasch-Modell für dichotome Daten (Einklassenlösung)**
- **Einteilung nach dem Geschlecht (zwei manifeste Klassen)**
- **Einteilung nach Leistung (Hoch- und Niedrigscorer)**
- **Das Mixed Rasch-Modell mit drei latenten Klassen**

Diese Modelle wurden nicht eigens ausgewertet, sondern sollen eine Einordnung der Ergebnisse des Mixed Rasch-Modell mit zwei latenten Klassen ermöglichen. Abbildung 17 zeigt die berechneten Modelle nach Art ihrer Klasseneinteilung.

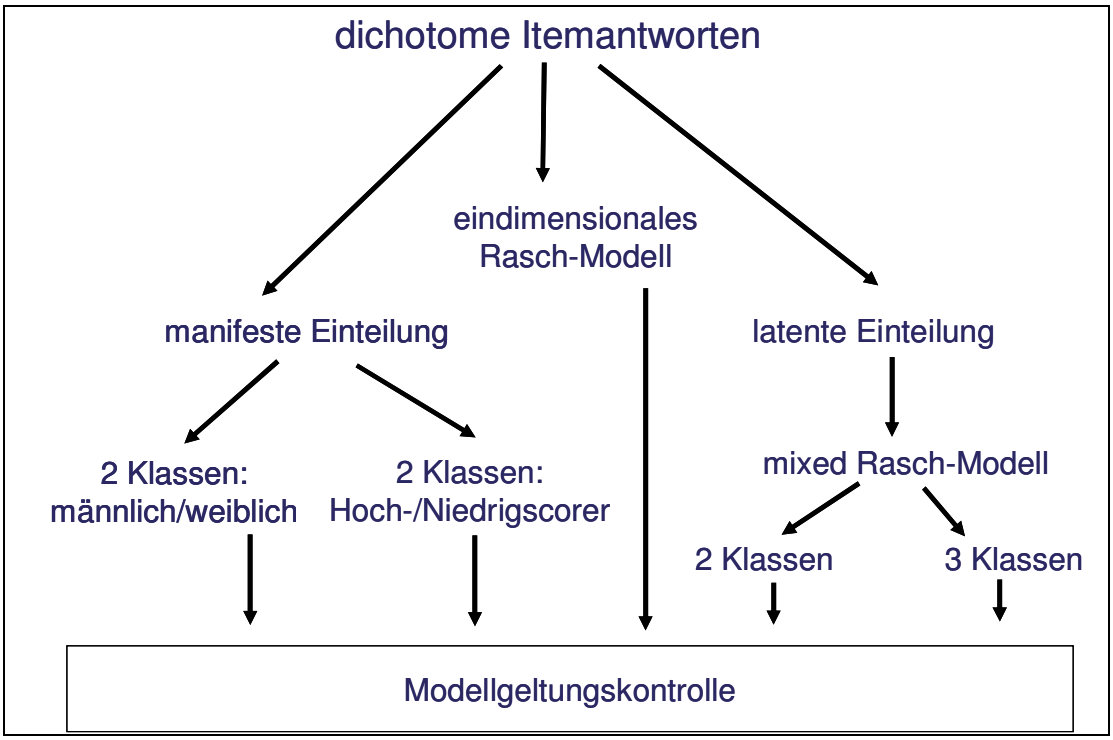

Abbildung 17 Verwendete Modelle der manifesten und latenten Klasseneinteilung auf Basis der Itemantworten des nationalen Naturwissenschaftstests.

Geschlecht und Leistung wurden als manifeste Teilungskriterien verwendet. Bei der manifesten Einteilung nach Leistung wurden die Schüler in die beiden Gruppen Hoch- und Niedrigscorer eingeteilt. Dazu wurde derjenige Trennscore verwendet, bei dem für das zugehörige Modell in jedem Inhaltsbereich die Likelihood maximal ist. Eine maximale Likelihood indiziert, dass die Wahrscheinlichkeit der beobachteten Daten bei Wahl genau dieses Trennscore am höchsten ist. Tabelle 19 zeigt, welcher Trennscore bei welchem Inhaltsbereich gewählt wurde. Der Trennscore ist in allen Inhaltsbereich relativ niedrig, was bedeutet, dass der Test relativ schwer ist.

Tabelle 19 Je nach Inhaltsbereich ist der Trennscore zwischen Hoch- und Niedrigscorer angegeben sowie die Spannweite des Scores der beiden Gruppen.

Inhaltsbereich	Trennscore	Score Niedrigscorer	Score Hochscorer
Atmung und Fotosynthese	2	0-2	3-7
Bewegungsgesetze (Mechanik)	3	0-3	4-7
Biochemie und Ernährung	3	0-3	4-7
Chemische Verbindungen und Aggregatzustände	1	0-1	2-7
Elektrizität	1	0-1	2-7
Energieumwandlung	2	0-2	3-7
Fortpflanzung und Sexualität	2	0-2	3-7
Teilchenkonzept	2	0-2	3-7
Wärmeverlust, Wärmegewinnung und Energiesparen	3	0-3	4-7

Im Gegensatz zu den anderen Modellen wurden die Berechnungen zu den manifesten Leistungsklassen mit dem von Gerhard Fischer entwickelten Programm LpcM-Win durchgeführt. Der Grund dafür liegt in der Schätzung der Likelihoods: Bei der Trennung eines Datensatzes in niedrig- bzw. hochscorende Personen wird die Schätzung der Scoreverteilunglikelihood (sL) verfälscht, die sich aus der conditioned Likelihood (cL) multipliziert mit den Scorewahrscheinlichkeiten ergibt (vgl. dazu das Rechenbeispiel zur Veranschaulichung in Exkurs 1 unten sowie Rost, 2004 zur Erläuterung der Likelihoods). Daher muss in diesem Fall die conditioned Likelihood (cL) verwendet werden. Bei der conditioned Likelihood cL gehen die Scorewahrscheinlichkeiten nicht in die Likelihoodschätzung mit ein.

Die Ergebnisse zu den genannten Vergleichsmodellen finden sich – sofern sie benötigt werden – bei den Darstellungen zur Zweiklassenlösung mit dem Mixed Rasch-Modell, da sie für die oben genannten Fragestellungen und Hypothesen nur in als Bezugsgrößen von Interesse sind.

Exkurs 1 Unterschiede in der Schätzung der Likelihood

Vergleich von conditioned Likelihood (cL) mit Scoreverteilungslikelihood (sL) bei manifester Klasseneinteilung nach Hoch-/Niedrigscorern

Hypothetischer Datensatz sei folgender:

Antwortpattern	Score	Scorewahrscheinlichkeit
1 1 1	3	3/10
1 1 1	3	3/10
1 1 1	3	3/10
1 1 0	2	1/10
0 0 1	1	6/10
1 0 0	1	6/10
1 0 0	1	6/10
0 1 0	1	6/10
0 0 1	1	6/10
0 1 0	1	6/10

Trennt man nun den Datensatz in Hoch- und Niedrigscorer beim Score 2, so beinhaltet der Datensatz der Hochscorer folgende Daten:

Antwortpattern	Score	Scorewahrscheinlichkeit
1 1 1	3	3/3 = 1
1 1 1	3	3/3 = 1
1 1 1	3	3/3 = 1

In diesem Extrembeispiel würde bei einer Trennung der Datensätze die Information über die Scorewahrscheinlichkeit wegfallen, da in der Scorelikelihoodfunktion mit 1 multipliziert würde. Eine Berechnung der Scoreverteilungslikelihood (sL) würde also bei nach Score getrennten Datensätzen zu einem verfälschten Wert führen.
Mit dem Programm LpcM-Win erfolgt eine softwaregesteuerte Trennung ohne die Datensätze separieren zu müssen, weiterhin wird die conditioned Likelihood (cL) anstelle der Scoreverteilungslikelihood (sL) berechnet.

3.5.2 Kennwerte der Zweiklassenlösungen mit dem Mixed Rasch-Modell: Treffsicherheiten und Klassengrößen

Bei der Betrachtung der Klassengrößen (Tabelle 20) zeigt sich in drei Inhaltsbereichen (Atmung und Fotosynthese, Energieumwandlung, Teilchenkonzept), dass die Klassengrößen unausgewogen sind, d.h. jeweils eine latente Klasse sehr klein ist. Wenn die Klassengrößen zu klein werden, ist die jeweilige latente Klasse von untergeordneter Bedeutung und die Anwendung des Mixed Rasch-Modells auf die Daten fragwürdig.
Die mittleren Treffsicherheiten (Tabelle 21) sind als zufrieden stellend zu beurteilen.

Tabelle 20 Klassengrößen der ermittelten latenten Klassen nach Inhaltsbereich

Inhaltsbereich	Klassengröße	
	Klasse 1	Klasse 2
Atmung und Fotosynthese	0.77	0.23
Bewegungsgesetze	0.57	0.43
Biochemie und Ernährung	0.69	0.31
Chemische Verbindungen	0.61	0.39
Elektrizität	0.61	0.39
Energieumwandlung	0.96	0.04
Fortpflanzung	0.53	0.47
Teilchenkonzept	0.92	0.08
Wärmeverlust	0.50	0.50

Tabelle 21 Treffsicherheiten beim Mixed Rasch-Modell mit zwei latenten Klassen.

Inhaltsbereich	Treffsicherheit	
	Klasse 1	Klasse 2
Atmung und Fotosynthese	0.832	0.729
Bewegungsgesetze	0.787	0.751
Biochemie und Ernährung	0.913	0.866
Chemische Verbindungen	0.815	0.785
Elektrizität	0.857	0.798
Energieumwandlung	0.991	0.745
Fortpflanzung	0.839	0.805
Teilchenkonzept	0.929	0.711
Wärmeverlust	0.828	0.841
Mittlere Treffsicherheit über alle Inhalte und Klassen:		0.820

3.5.3 Verteilung von Jungen und Mädchen auf die latenten Klassen des Mixed Rasch-Modells

Die bedingten Verteilungen von Jungen und Mädchen (Tabelle 22) auf die Mixed Rasch-Klassen zeigen in etwa gleiche Anteile. Daraus folgt, dass die Klassen nur wenig Bezug zum Geschlecht haben. Leichte Abweichungen ergeben sich für die drei Inhaltsbereiche Bewegungsgesetze, Energieumwandlung, Fortpflanzung und Sexualität. Erklärbar wird dieser Effekt einerseits durch die Thematik, die stark geschlechtsspezifisch ist, andererseits durch die naturwissenschaftliche (Schul-)Fächerzuordnung der Aufgaben (Physik bzw. Biologie).

Tabelle 22 Bedingte Verteilungen von Jungen und Mädchen auf die Mixed Rasch-Klassen. Je nach Inhaltsbereich variieren die Anteile, stärkere geschlechtsspezifische Variationen sind unterlegt.

Inhaltsbereich		Jungen	Mädchen	Gesamt
Atmung und Fotosynthese	Klasse 1	82.3	81.4	81.8
	Klasse 2	17.7	18.6	18.2
Bewegungsgesetze	Klasse 1	65.0	50.6	58.0
	Klasse 2	35.0	49.4	42.0
Biochemie und Ernährung	Klasse 1	71.8	70.4	71.1
	Klasse 2	28.2	29.6	28.9
Chemische Verbindungen	Klasse 1	62.8	64.4	63.6
	Klasse 2	37.2	35.6	36.4
Elektrizität	Klasse 1	67.4	63.0	65.2
	Klasse 2	32.6	37.0	34.8
Energieumwandlung	Klasse 1	80.0	87.8	83.9
	Klasse 2	20.0	12.2	16.1
Fortpflanzung und Sexualität	Klasse 1	46.9	56.2	51.5
	Klasse 2	53.1	43.8	48.5
Teilchenkonzept	Klasse 1	91.5	96.9	94.3
	Klasse 2	8.5	3.2	5.7
Wärmeverlust und Energiesparen	Klasse 1	52.9	50.5	51.7
	Klasse 2	47.1	49.5	48.3

3.5.4 Was kennzeichnet die Mixed Rasch-Klassen? Parameterprofile und Lösungswahrscheinlichkeiten

Im Rahmen der Auswertung des Mixed Rasch-Modells ist bei Parameterprofilen, in denen die Itemschwierigkeiten dargestellt werden, allgemein zu beachten, dass der Abstand zwischen den Profilen zweier latenten Klassen die Unterschiedlichkeit reflektiert. Mit einem Blick auf die Itemparameter wird also deutlich, wie stark die qualitativen Unterschiede in den Klassen ausgeprägt sind. Das Niveau einzelner Items ist aufgrund der konventionell verwendeten Summennormierung der Items auf Null nicht interpretierbar. Damit zeigt ein solches Profil nur qualitative und keine quantitativen Informationen. Letztere sind wiederum in den Profilen der Lösungswahrscheinlichkeiten abzulesen.

Abbildung 19 bis Abbildung 28 zeigen die Profilverläufe der Itemparameter und Lösungswahrscheinlichkeiten für jeden Inhaltsbereich. Nicht dargestellt sind die Profile für die Inhalte Atmung und Fotosynthese, Energieumwandlung, Teilchenkonzept. In diesen Fällen scheint eine Anwendung des Mixed Rasch-Modells unangemessen, da sowohl die Klassengrößen sehr klein werden, als auch für die Itemparameter zu extreme Werte geschätzt wurden. Viele Itemparameter liegen höher als 4 bzw. niedriger als –4 auf der logit-Skala, z. T. sogar noch deutlich extremer. Eine solche Lösung deutet darauf hin, dass die Personen dieser Klasse unskalierbar sind, während in der anderen Klasse das Rasch-Modell gilt. Diese Vermutungen werden noch dadurch gestützt, dass genau für diese drei Inhaltsbereiche die zunächst angesetzten 250 Iterationen zur Parameterschätzung mit Winmira nicht ausreichten, sondern deutlich mehr Iterationen als bei den anderen Inhaltsbereichen benötigt wurden. Der Interpretationswert wird dadurch für die Inhaltsbereiche stark eingeschränkt, so dass sie im Folgenden von der Darstellung ausgeschlossen werden. Es kann davon ausgegangen werden, dass das Mixed Rasch-Modell mit zwei latenten Klassen nicht auf diese drei genannten Inhaltsbereiche anzuwenden ist.

Im Anschluss werden die Itemprofile der Mixed Rasch-Klassen genauer analysiert, um zu erklären, welches Merkmal die Klassen trennt.

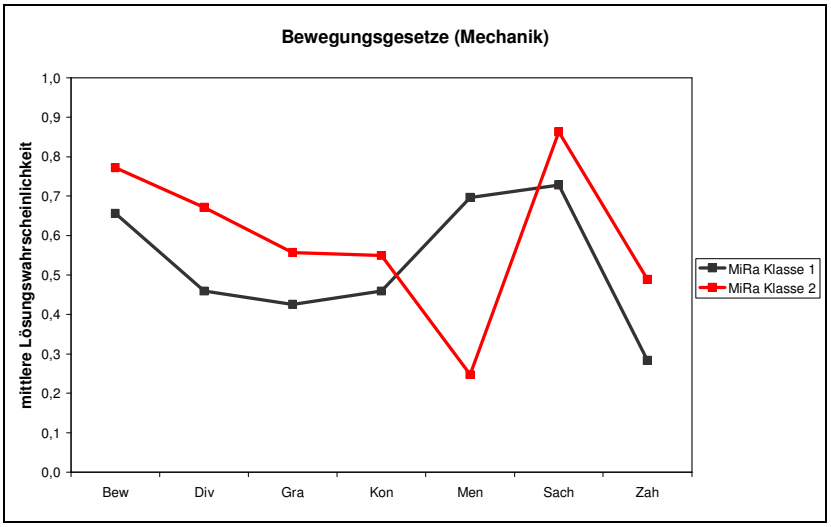

Abbildung 18 Mittlere Lösungswahrscheinlichkeiten der zwei mit dem Mixed Rasch-Modell ermittelten latenten Klassen für den Inhaltsbereich Bewegungsgesetze (Mechanik).

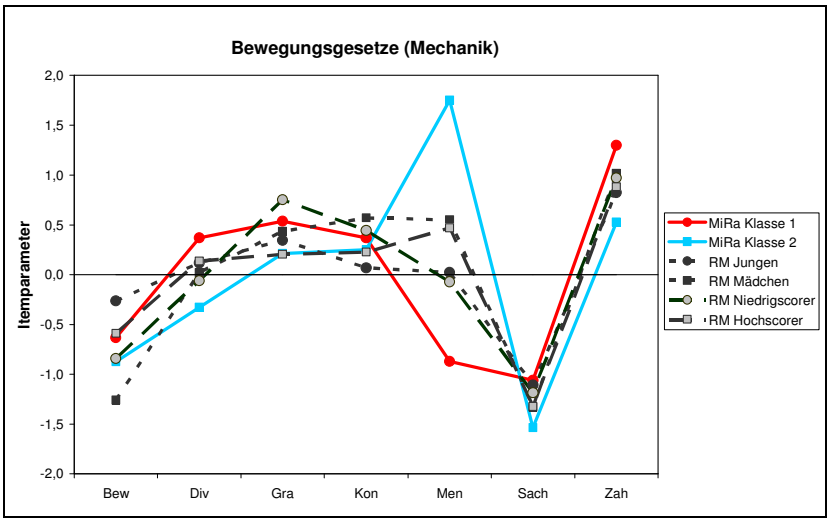

Abbildung 19 Itemschwierigkeitsprofile der Zweiklassenlösungen im Inhaltsbereich Bewegungsgesetze (Mechanik). Farbig hervorgehoben sind die Profile der beiden latenten Klassen aus dem Mixed Rasch-Modell (für die Abkürzungen s. Abbildung 5).

An den Itemschwierigkeitsprofilen im Inhaltsbereich Bewegungsgesetze (Mechanik) (Abbildung 19) ist zu erkennen, dass die latente Klasseneinteilung mit dem Mixed Rasch-Modell stärkere Unterschiede aufdeckt als die übrigen Klassenlösungen– die Abstände zwischen den Klassen sind hier am deutlichsten. Die mit dem Mixed Rasch-Modell ermittelten latenten Klassen unterscheiden sich stärker als die Einteilung nach dem Geschlecht oder der Leistung nach dem Trennscore: Die qualitativen Unterschiede sind hier größer.

Im Profil der Lösungswahrscheinlichkeiten (Abbildung 18) zeigt sich auch, dass es sich bei den zwei latenten Klassen des Mixed Rasch-Modells im Wesentlichen um Leistungsklassen handelt. Würde sich nur das Leistungsspektrum abbilden, würde dazu auch das Rasch-Modell

genügen. Offenbar scheint aber das Item zu mentalen Modellen eine latente Einteilung in zwei Klassen zu rechtfertigen und die Personen in Klassen zu trennen.

Besonders deutlich wird dieser Zusammenhang beim Item zur kognitiven Teilkompetenz „Mentale Modelle“, das sich für die Klasse 2 mit Blick auf die mittleren Lösungswahrscheinlichkeiten als zu schwer zeigt, umgekehrt wird von Personen in Klasse 1 (die von den Leistungen her unter der Klasse 2 liegt) das Item besser gelöst als in der eigentlich leistungsfähigeren Klasse.

Mit Blick auf den Inhalt kann vermutet werden, dass es sich bei dieser Aufgabe weniger um definiertes Schulwissen handelt, sondern mehr um Alltagswissen. Zur Lösung der Aufgabe muss das Verhalten von Fahrzeugen im Straßenverkehr, die Bewegung im Raum abgeschätzt werden.

Klasse 1, in der das Item mit höherer Wahrscheinlichkeit gelöst wird, schneidet bei Schulwissen im Sinne des Curriculums in Fach Physik zwar weniger gut ab, dafür besser in dieser Aufgabe, in der praktische Erfahrungen und Alltagserlebnisse von Vorteil sind. Wie aus Tabelle 22 hervorgeht, gehören dieser Klasse etwas weniger Mädchen an. Mädchen sind offenbar zwar in der generellen Leistungsklasse (Klasse 2) stärker vertreten, aber seltener in der Klasse, die gerade über einen praktischen Zugang zu Physik im Straßenverkehr zu verfügen scheint (Klasse 1). Dies ist durch geschlechtsspezifische Interessensschwerpunkte und Erfahrungen erklärbar.

Klasse 2 weist vermutlich deshalb eine niedrigere Lösungswahrscheinlichkeit auf, weil die Aufgabe deutlich schwerer zu lösen ist, wenn nur das in Physik vermittelte Wissen verwendet wird. Kontrastierend dazu verhält sich das Item zum Umgang mit Zahlen: Für die Lösung sind curricular verankerte Wissensstrukturen, die in Klasse 2 besonders gut ausgeprägt zu sein scheinen, förderlich, Alltagswissen ist weniger nützlich bzw. wird hier nicht benötigt.

Dieser Befund zeigt sich auch in den Parameterprofilen: Bei dem Item zum Mentalen Modell ist der qualitative Unterschiede zwischen den Klassen am größten, erkennbar an dem Abstand der beiden latenten Klassenprofile. Weiterhin dreht sich das Verhältnis der Itemschwierigkeiten bei dem Item zu mentalen Modellen zwischen den Klassen um: In Klasse 2 ist der Schwierigkeitsparameter höher als in Klasse 1, während bei dem Item zum Umgang mit Zahlen der umgekehrte Zusammenhang gilt: Mentale Modelle sind in Klasse 2 schwieriger, der Umgang mit Zahlen ist in Klasse 1 schwieriger. D.h. diese beiden Items machen mit obiger inhaltlicher Interpretation den Unterschied zwischen den Klassen aus.

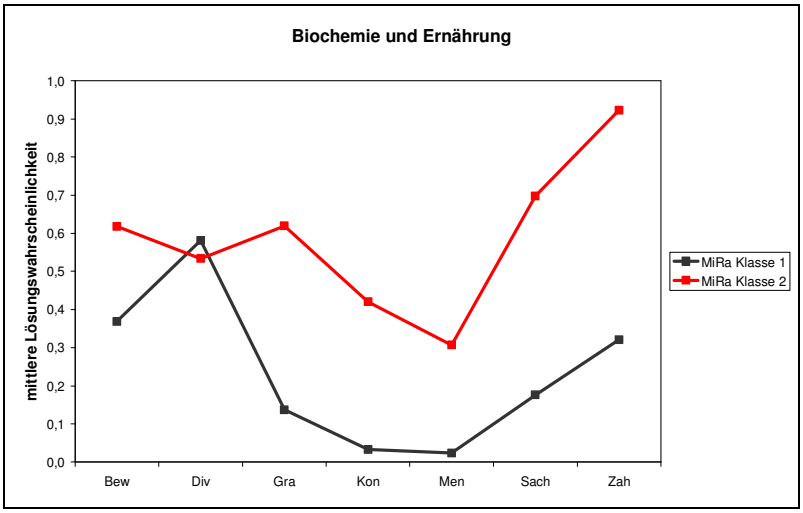

Abbildung 20 Mittlere Lösungswahrscheinlichkeiten der zwei mit dem Mixed Rasch-Modell ermittelten latenten Klassen für den Inhaltsbereich Biochemie und Ernährung für alle sieben Items, die jeweils eine kognitive Teilkompetenz erfassen (für die Abkürzungen s. Abbildung 5).

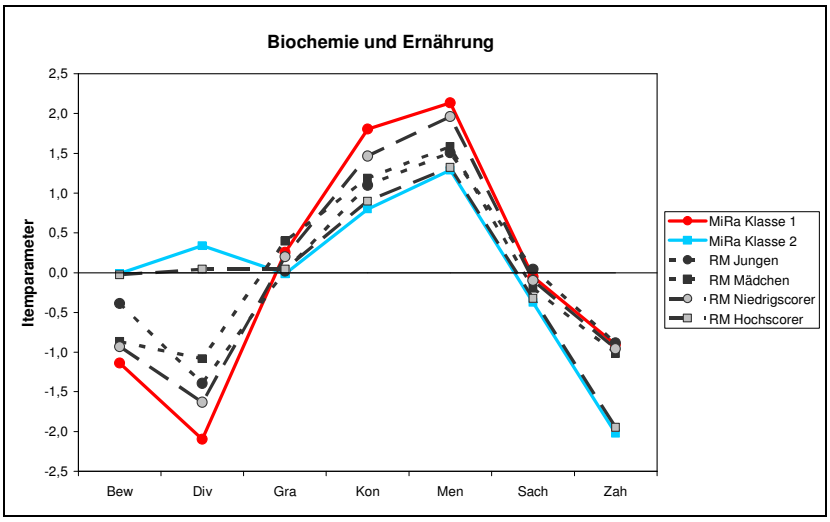

Abbildung 21 Profilverläufe der Itemschwierigkeiten in allen Zweiklassenlösungen im Inhaltsbereich Biochemie und Ernährung. Farbig hervorgehoben sind die Profile der beiden latenten Klassen aus dem Mixed Rasch-Modell (s. Text, für die Abkürzungen s. Abbildung 5).

In den Profilverläufen der Itemschwierigkeiten im Inhaltsbereich Biochemie und Ernährung (Abbildung 21) zeigt sich der stärkste qualitative Unterschied zwischen den Klassen im Item zum divergenten Denken. Gemeinsam mit dem Item zum Bewerten drehen sich die Itemschwierigkeiten der Klassen um:

Divergentes Denken (und Bewerten) ist in Klasse 2 schwieriger, konvergentes Denken ist in Klasse 1 schwieriger, ebenso wie die übrigen Items.

Dies ist auch im Profil der Lösungswahrscheinlichkeiten (Abbildung 20) zu sehen: Während Klasse 2 bei allen anderen Items deutlich höhere Wahrscheinlichkeiten aufweist, fallen bei diesem Item die Lösungswahrscheinlichkeiten der beiden Klassen fast gleich aus, Klasse 1,

die sonst durchweg geringere Lösungswahrscheinlichkeiten zeigt, löst die Aufgabe sogar mit einer leicht höheren Wahrscheinlichkeit.

Thematisch geht es bei der Aufgabe zum divergenten Denken um praktische Ernährungstipps, die von den Schülern hinsichtlich spezieller Kriterien generiert werden sollen. Die anderen Items zum gleichen Thema sind praxisferner gestaltet. Wie beim Inhaltsbereich Bewegungsgesetze trennt offenbar gerade die Alltagsnähe des Items die Personen in zwei Gruppen. Zur Lösung des Items zum konvergenten Denken kommt in diesem Inhaltsbereich praktischen Erfahrungen keine entscheidende Bedeutung zu.

Bei diesem Inhaltsbereich sind keine auffälligen Schwankungen der Geschlechterverteilungen zu beobachten. Das zeigt, dass in diesem Inhaltsbereich keine Kontingenz zwischen Alltagsnähe und Geschlecht angenommen werden kann.

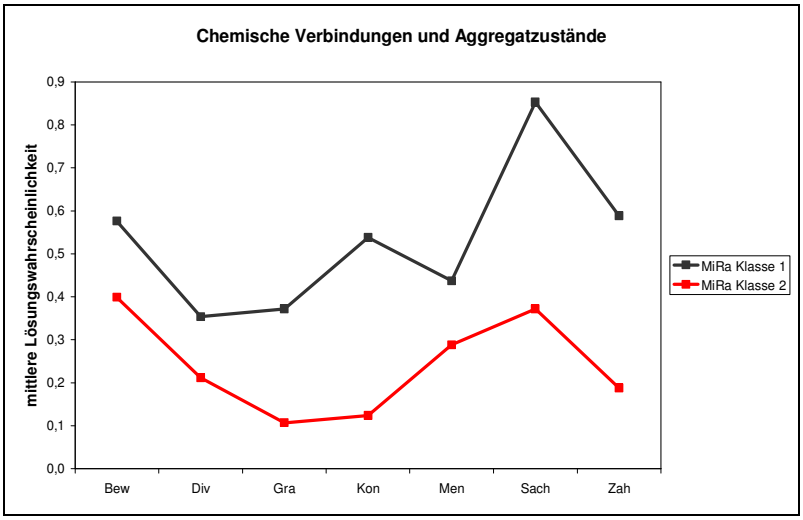

Abbildung 22 Mittlere Lösungswahrscheinlichkeiten der zwei mit dem Mixed Rasch-Modell ermittelten latenten Klassen für den Inhaltsbereich Chemische Verbindungen und Aggregatzustände für alle sieben Items, die jeweils eine kognitive Teilkompetenz erfassen (für die Abkürzungen s. Abbildung 5).

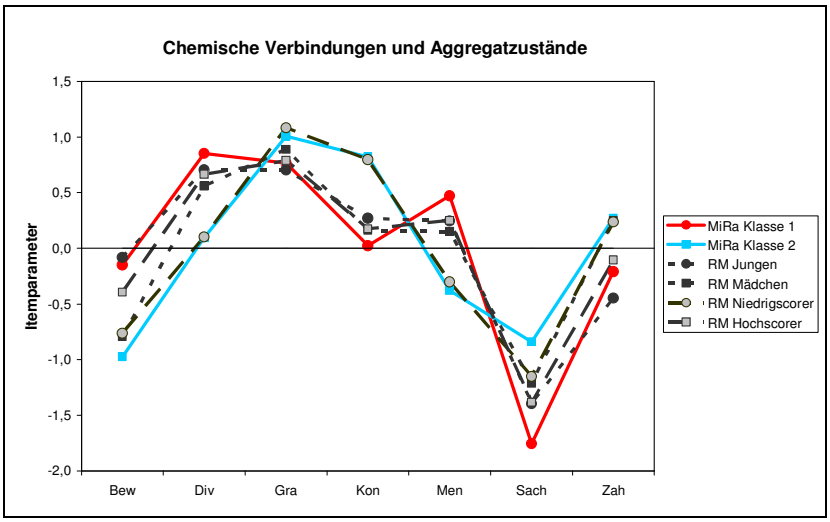

Abbildung 23 Profilverläufe der Itemschwierigkeiten in allen Zweiklassenlösungen im Inhaltsbereich Chemische Verbindungen und Aggregatzustände. Farbig hervorgehoben sind die Profile der beiden latenten Klassen aus dem Mixed Rasch-Modell (für die Abkürzungen s. Abbildung 5).

Im Inhaltsbereich Chemische Verbindungen und Aggregatzustände zeigen sich in den Klassenprofilen der Lösungswahrscheinlichkeiten keine Überschneidungen (Abbildung 22). In den Itemprofilen (Abbildung 23) überschneiden sich mehrere Items in ihren Schwierigkeiten. So weisen die Items zu Bewerten, divergentem Denken und mentalen Modellen in Klasse 1 höhere Schwierigkeitsparameter auf als in Klasse 2, während der Umgang mit Grafiken und Zahlen sowie Konvergentes Denken und Sachverhalte verbalisieren in Klasse 2 schwieriger als in Klasse 1 ist. Vergleicht man daraus z.B. Mentale Modelle zum Umgang mit Zahlen, so bietet sich zum Iteminhalt kein eindeutiges Erklärungsmuster an, wie innerhalb der anderen Inhaltsbereiche diskutiert. Beide Items

beziehen sich mehr auf vermittelte naturwissenschaftliche Kenntnisse und Schlussfolgerungen als auf Alltagserfahrungen.

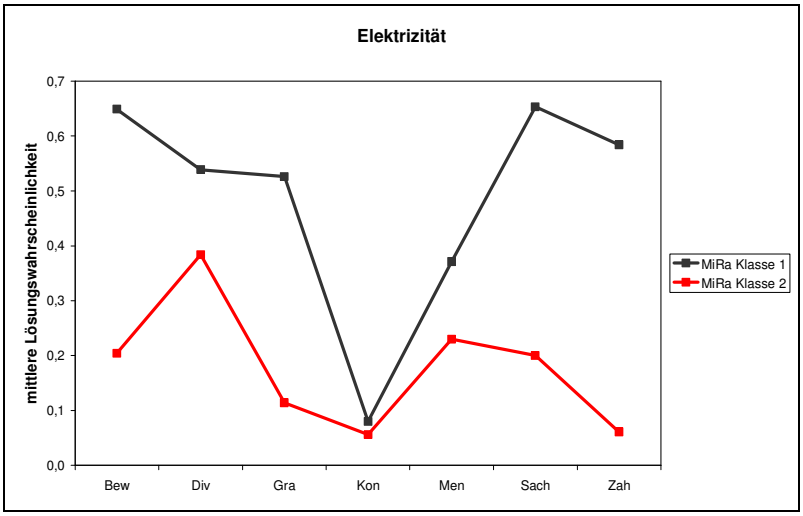

Abbildung 24 Mittlere Lösungswahrscheinlichkeiten der zwei mit dem Mixed Rasch-Modell ermittelten latenten Klassen für den Inhaltsbereich Elektrizität für alle sieben Items, die jeweils eine kognitive Teilkompetenz erfassen (s. Text, für die Abkürzungen s. Abbildung 5).

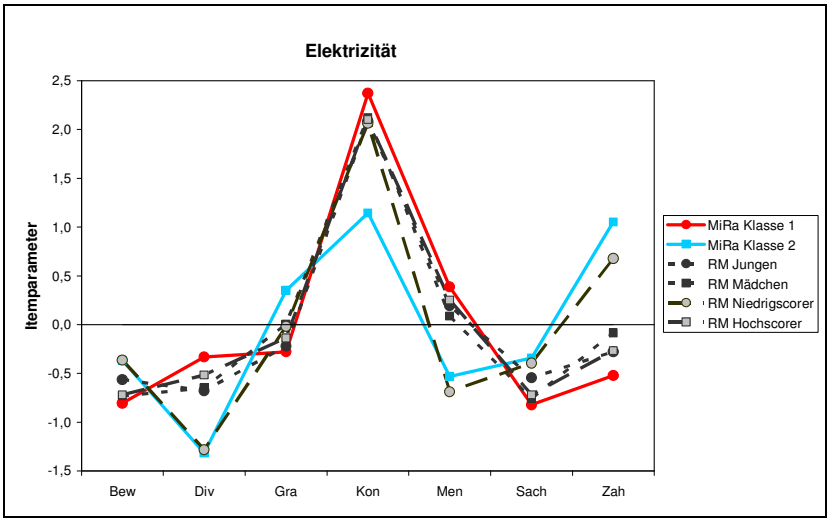

Abbildung 25 Profilverläufe der Itemschwierigkeiten in allen Zweiklassenlösungen im Inhaltsbereich Elektrizität. Farbig hervorgehoben sind die Profile der beiden latenten Klassen aus dem Mixed Rasch-Modell (s. Text, für die Abkürzungen s. Abbildung 5).

Im Inhaltsbereich Elektrizität zeigen die mittleren Lösungswahrscheinlichkeiten (Abbildung 24) einen Klassenunterschied beim Item zum konvergenten Denken. Während Klasse 1 bei allen anderen Items eine deutlich höhere Lösungswahrscheinlichkeit aufweist, entspricht diese hier etwa derjenigen der Klasse 2: Das Item zum konvergenten Denken zeigte sich in Klasse 1 zu schwer. Leichte Verringerungen der Lösungswahrscheinlichkeiten sind auch beim Item zum mentalen Modell sichtbar. Beim Item zum divergenten Denken wird die relative Stärke von Klasse 2 sichtbar.

Dies zeigt sich auch in den Itemparameterprofilen (Abbildung 27). Die genannten Items sind in Klasse 1 schwerer, während die übrigen Items in Klasse 2 schwerer sind.

Die Items konvergentes Denken und mentale Modelle sind in diesem Inhaltsbereich zwar nicht mit Alltagserfahrungen lösbar, praktische Experimente im Physikunterricht (z.B. selbst elektrische Schaltungen bauen) können aber zum Finden der Lösung ausschlaggebend sein. Divergentes Denken hängt in diesem Inhaltsbereich wieder mit einer praktischen Anwendung zusammen: Die Jugendlichen in Klasse 2 scheinen hier ihre Stärken zu zeigen.

Es lässt sich also auch hier schlussfolgern, dass in Klasse 1 Probleme mit praktischen Aspekten eher schlechter zu lösen sind. Klasse 2 schneidet zwar in den genannten Items nicht besser ab als Klasse 1 (geringere Lösungswahrscheinlichkeiten), zeigt aber relative Stärken im Praxisbezug, wenn es um Alltagsprobleme geht (Item divergentes Denken).

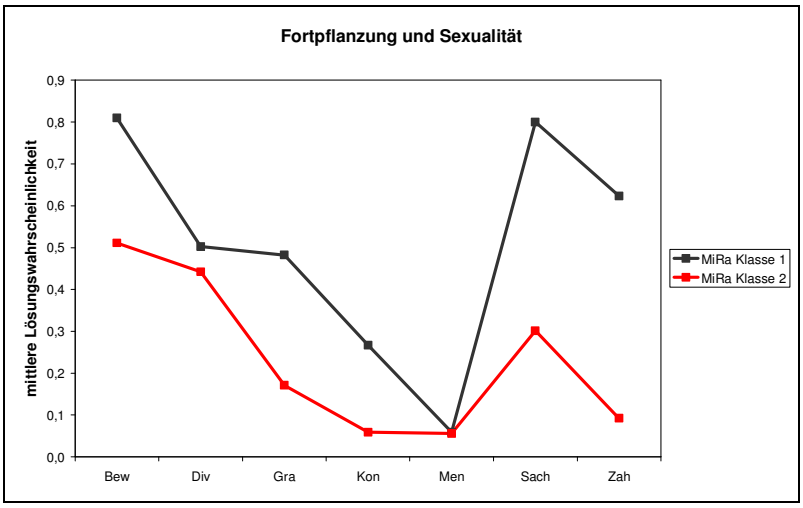

Abbildung 26 Mittlere Lösungswahrscheinlichkeiten der zwei mit dem Mixed Rasch-Modell ermittelten latenten Klassen für den Inhaltsbereich Fortpflanzung und Sexualität für alle sieben Items, die jeweils eine kognitive Teilkompetenz erfassen (s. Text, für die Abkürzungen s. Abbildung 5).

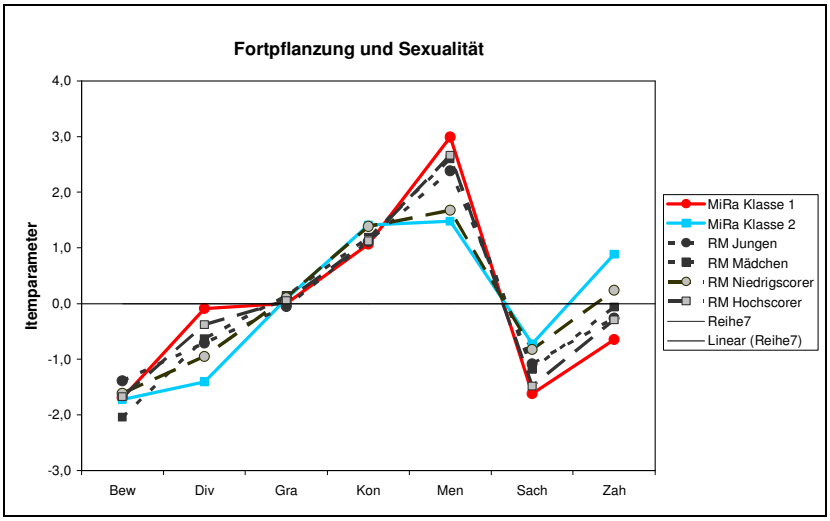

Abbildung 27 Profilverläufe der Itemschwierigkeiten in allen Zweiklassenlösungen im Inhaltsbereich Fortpflanzung und Sexualität. Farbig hervorgehoben sind die Profile der beiden latenten Klassen aus dem Mixed Rasch-Modell (s. Text, für die Abkürzungen s. Abbildung 5).

Die Profile der mittleren Lösungswahrscheinlichkeiten zeigen relative Schwächen der Klasse 1 im Item zum mentalen Modell, bei dem die Lösungswahrscheinlichkeit der beiden Klassen gleich ist. Beim Item zum divergenten Denken nähern sich die Profile ebenfalls an, verlaufen ansonsten etwa parallel. Bei den Parameterprofilen zeigen sich die Items zum mentalen Modell und divergenten Denken in Klasse 1 schwerer, in Klasse 2 leichter. (Die Unterschiede im Niveau sind nicht interpretierbar, aber die Überschneidungen bzw. Vertauschung der Parameterschwierigkeiten zwischen den Klassen.) Der umgekehrte Zusammenhang gilt beim Sachverhalte verbalisieren und Umgang mit Zahlen.

Betrachtet man die Items inhaltlich, so fällt auf, dass – abgesehen von den Items zum Sachverhalte verbalisieren und dem Umgang mit Zahlen – in diesem Inhaltsbereich viele

Items einen praktischen Bezug haben (Verhütung, Schwangerschaft, Gesundheit). Gerade das Item zu mentalen Modellen sollte von fünfzehnjährigen Jugendlichen lösbar sein, da es das Wissen um eine Vermeidung von Schwangerschaft in Zusammenhang mit dem weiblichen Zyklus erfordert. Dennoch ist wird in beiden Klassen nur mit einer geringen Wahrscheinlichkeit gelöst.

Das Item zum divergenten Denken erfordert naturwissenschaftliches Alltagswissen (Gesundheit in der Schwangerschaft), dies spricht offenbar die Stärken von Klasse 2 an, so dass sich die Lösungswahrscheinlichkeiten der aus Klasse 1 annähern können.

Dass die Items zum Sachverhalte verbalisieren und Umgang mit Zahlen in Klasse 1 geringere Schwierigkeitsparameter aufweist, ist kongruent zu obigen Interpretationen: Bei diesen Items wird die Anwendung naturwissenschaftlichen Verständnisses auf nicht primär alltagsbezogene, praktische, sondern mehr abstrakte Inhalte benötigt.

Dass sich mehr Mädchen in Klasse 1 (der Klasse mit den höheren Lösungswahrscheinlichkeiten) befinden scheint plausibel: Das Thema Fortpflanzung, gerade in Kombination mit Gesundheit sowie die biologische Inhaltskomponente scheint Mädchen stärker anzusprechen.

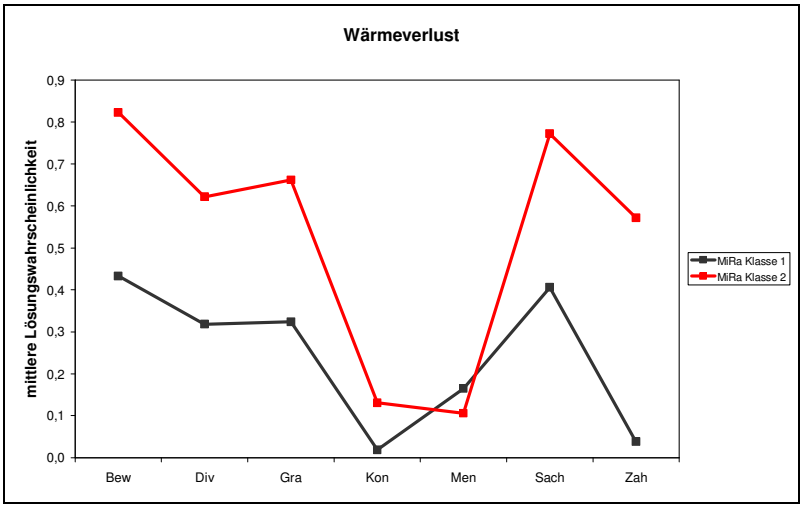

Abbildung 28 Mittlere Lösungswahrscheinlichkeiten der zwei mit dem Mixed Rasch-Modell ermittelten latenten Klassen für den Inhaltsbereich Wärmeverlust für alle sieben Items, die jeweils eine kognitive Teilkompetenz erfassen (s. Text, für die Abkürzungen s. Abbildung 5).

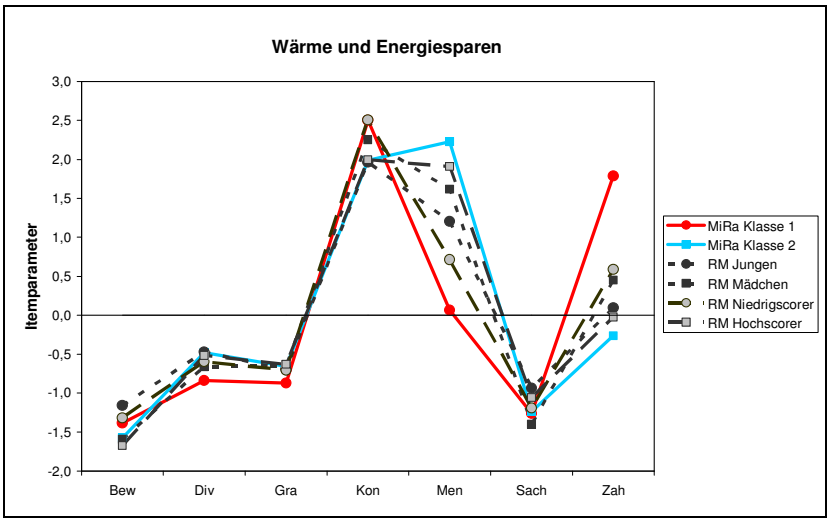

Abbildung 29 Profilverläufe der Itemschwierigkeiten in allen Zweiklassenlösungen im Inhaltsbereich Wärme und Energiesparen. Farbig hervorgehoben sind die Profile der beiden latenten Klassen aus dem Mixed Rasch-Modell (s. Text, für die Abkürzungen s. Abbildung 5).

Im Inhaltsbereich Wärmeverlust und Energiesparen überschneiden sich die Profile der mittleren Lösungswahrscheinlichkeiten. Hier fällt besonders das Item zu mentalen Modellen auf, das in Klasse 2 (die Klasse der „Könner“) zu schwer ist und mit einer geringeren Wahrscheinlichkeit gelöst wird als in der Klasse 2 (die Klasse der „Nicht-Könner“).

In den Schwierigkeitsprofilen zeigt sich der qualitativ stärkste Unterschied zwischen den Klassen im Item zu mentalen Modellen und zum Umgang mit Zahlen, die Schwierigkeitsprofile überkreuzen sich bei den beiden Items.

Das Item zu mentalen Modellen erfordert physikalisches Verständnis im Zusammenhang mit thermischer Ausdehnung von Festkörpern. Dieses Verständnis wird im Physikunterricht häufig zu vermitteln versucht, indem es praktisch gezeigt wird. Das Item zum Umgang mit

Zahlen beinhaltet eine eher ungewöhnliche Kombination von Physik und Alltag (Physik zum Thema Kleidung).
Auch hier kann der Unterschied zwischen den Klassen ggf. auf Unterschiede im Praxisbezug der Items zurückgeführt werden.

Über alle Inhaltsbereiche lässt sich zusammenfassen, dass es sich bei den latenten Klassen aus dem Mixed Rasch-Modell um Leistungsklassen handelt. Da zur Modellierung reiner Leistungsklassen aber auch das Rasch-Modell ausreicht, muss noch ein Unterschied zwischen den beiden latenten Klassen aus dem Mixed Rasch-Modell im Antwortmuster vorhanden sein. Wie oben ausgeführt, liegt der Unterschied zwischen den latenten Klassen möglicherweise im Alltagswissen oder dem Praxisbezug: Die Klasse der „Könner“ zeigt in einigen Inhaltsbereichen geringere Lösungswahrscheinlichkeiten als die Klasse der „Nicht-Könner“.
Mit Vorsicht ist folgende Interpretation möglich: Jugendliche in der Klasse der „Könner“ sind in der Lage, ihr naturwissenschaftliches Verständnis flexibler auf neue Aufgabenstellungen anwenden, profitieren aber nicht so sehr von praxisnahen Aufgaben oder es fehlt ihnen an entsprechenden naturwissenschaftlichen Erfahrungen. Bei solchen Aufgaben werden sie von der Klasse der „Nicht-Könner“ in der Lösungswahrscheinlichkeit überholt oder liegen etwa gleich.
Andersherum lässt sich schließen, dass die Gruppe der „Nicht-Könner“ von praxisnahen Veranschaulichungen z.B. im Unterricht stark profitiert und in alltagsnahen Aufgaben Stärken zeigt, allerdings starke Schwierigkeiten hat, das naturwissenschaftliche Alltagsverständnis auf neue Bereiche zu übertragen.
Dieser qualitative Unterschied zwischen den Klassen ist nur in Ausnahmefällen (z.B. Inhaltsbereich Bewegungsgesetze) mit dem Geschlecht assoziiert.
Die Erwartung, dass das Mixed Rasch-Modell die geschlechtsspezifischen Profile in den kognitiven Teilkompetenzen verdeutlicht, konnte nicht bestätigt werden: Betrachtet man den Zusammenhang zwischen Itemschwierigkeit und Klassenzugehörigkeit, so wird deutlich, dass die Profile nicht mit den geschlechterspezifischen Mustern korrespondieren, d.h. die Klassen können nicht jeweils einer verbal-bewertenden und einer grafisch-numerisch-abstrakten Klasse zugeordnet werden.
Es ist weiterhin auffällig, dass die Klassenlösung nach Geschlecht einen geringeren qualitativen Unterschied macht als die Einteilung nach dem optimalen Trennscore, was die

Aussage stützt, dass die Geschlechtsunterschiede eher als gering im Vergleich zur allgemeinen Leistung zu beurteilen sind.
Die mit dem Mixed Rasch-Modell ermittelten latenten Klassen zeigen größere qualitative Unterschiede als die übrigen Klassenlösungen (Einteilung nach dem Geschlecht oder der Leistung nach dem Trennscore) – die Abstände zwischen den Klassen sind hier am größten.

Auf die Darstellung der Profile des Mixed Rasch-Modells mit drei latenten Klassen wurde in diesem Abschnitt verzichtet. Es hat sich gezeigt, dass bei sieben von neun Inhaltsbereichen Klassengrößen gegen Null gehen und die geschätzten Parameter zu extreme Werte auf der logit-Skala annehmen, so dass kaum Interpretationswert für die Befunde gegeben ist. Das Mixed Rasch-Modell mit drei latenten Klassen passt offenkundig nicht auf die Daten, was wiederum die Anwendung der Zweiklassenlösung des Mixed Rasch-Modells stützt.
Da sich das Mixed Rasch-Modell mit zwei latenten Klassen bei den Inhaltsbereiche Atmung/Photosynthese, Energieumwandlung und Teilchenkonzept als nicht passend gezeigt hat, wird bei folgenden Auswertungen auf diese Inhaltsbereiche verzichtet.

3.5.5 Wie gut passt das Mixed Rasch-Modell auf die Daten? Informationstheoretische Maße

Tabelle 23 gibt die informationstheoretischen Maße für die Klassenlösungen wieder, zur Berechnung und Beschreibung sei auf Tabelle 8 verwiesen. Es fällt auf, dass die Zahlen dicht beieinander liegen, was den Nachteil dieser Methode der Modellprüfung nochmals deutlich werden lässt: Es gibt keinen Grenzwert, ab dem ein Modell nicht mehr auf die Daten anwendbar ist oder um wie viel sich ein Modell von einem anderen unterscheiden muss, um wesentlich besser oder schlechter zu passen.
Da die Unterschiede zwischen den Modellen klein sind, kann zwischen den Modellen wenig differenziert werden.
Die Zweiklassenlösung mit dem Mixed Rasch-Modell passt am besten auf die Inhalte Biochemie und Ernährung (gemäß BIC und CAIC), Elektrizität (AIC, BIC, CAIC), Fortpflanzung und Sexualität (AIC, BIC), Wärmeverlust (AIC, BIC, CAIC).
Bei den Inhalten Bewegungsgesetze und Chemische Verbindungen konnte sich das Mixed Rasch-Modell mit zwei latenten Klassen nach den informationstheoretischen Maßen gegen die anderen Modelle nicht deutlich durchsetzen.

Tabelle 23 Informationstheoretische Maße AIC, BIC und CAIC (vgl. Tabelle 8) für jeden Inhaltsbereich und jede Klassenlösung, abgesehen von der manifesten Einteilung nach Leistung. Unterlegt ist jeweils der geringste Wert pro Spalte pro Inhaltsbereich, der dabei die beste Modellpassung indiziert.

Bewegungsgesetze (Mechanik)	**AIC**	**BIC**	**CAIC**
Rasch-Modell	8810.80	8874.53	8887.53
Rasch-Modell (2 Klassen nach Geschlecht)	8750.15	8859.59	8885.59
Mixed Rasch-Modell (2 latente Klassen)	8791.67	8914.24	8939.24
Mixed Rasch-Modell (3 latente Klassen)	8781.43	8962.84	8999.84
Biochemie und Ernährung	**AIC**	**BIC**	**CAIC**
Rasch-Modell	7730.71	7794.32	7807.32
Rasch-Modell (2 Klassen nach Geschlecht)	7712.89	7822.10	7848.10
Mixed Rasch-Modell (2 latente Klassen)	7542.49	7664.83	7689.83
Mixed Rasch-Modell (3 latente Klassen)	7523.60	7704.67	7741.67
Chemische Verbindungen und Aggregatzustände	**AIC**	**BIC**	**CAIC**
Rasch-Modell	8411.62	8475.10	8488.10
Rasch-Modell (2 Klassen nach Geschlecht)	8369.05	8478.00	8504.00
Mixed Rasch-Modell (2 latente Klassen)	8406.33	8528.42	8553.42
Mixed Rasch-Modell (3 latente Klassen)	8418.68	8599.37	8636.37
Elektrizität	**AIC**	**BIC**	**CAIC**
Rasch-Modell	10971.46	11038.89	11051.89
Rasch-Modell (2 Klassen nach Geschlecht)	10970.76	11087.60	11113.60
Mixed Rasch-Modell (2 latente Klassen)	10895.95	11025.62	11050.62
Mixed Rasch-Modell (3 latente Klassen)	10901.07	11092.99	11129.99
Fortpflanzung und Sexualität	**AIC**	**BIC**	**CAIC**
Rasch-Modell	9886.55	9953.77	9966.77
Rasch-Modell (2 Klassen nach Geschlecht)	9840.16	9956.58	9982.58
Mixed Rasch-Modell (2 latente Klassen)	9819.22	9948.49	9973.49
Mixed Rasch-Modell (3 latente Klassen)	9830.46	10021.78	10058.78
Wärmeverlust, Wärmegewinnung und Energiesparen	**AIC**	**BIC**	**CAIC**
Rasch-Modell	10108.26	10175.58	10188.58
Rasch-Modell (2 Klassen nach Geschlecht)	10086.38	10203.00	10229.00
Mixed Rasch-Modell (2 latente Klassen)	10028.03	10157.49	10182.49
Mixed Rasch-Modell (3 latente Klassen)	10036.43	10228.03	10265.03

3.5.6 Wie gut passt das Mixed Rasch-Modell auf die Daten? Likelihoodquotiententests

Tabelle 24 zeigt als Übersicht die Likelihood der berechneten Modelle sowie die zu einem Likelihoodquotiententest benötigten Kennwerte. Die Likelihood des Rasch-Modells ist unter L_0 angegeben, da es im Likelihoodquotiententest als Referenz dient. Bei der manifesten Zweiklassenlösung nach Leistung ist als Referenz ebenfalls die Likelihood des Rasch-Modells angegeben, allerdings berechnet mit LpcM-Win, da beide Kennwerte über die conditioned Likelihood cL anstatt wie bei Winmira über die Scoreverteilungslikelihood sL geschätzt wurden (vgl. Exkurs 1). An dieser Stelle wird das Signifikanzniveau nicht adjustiert und der kritische Chi^2-Wert für 5% verwendet, da jeder Test als einzelne Fragestellung betrachtet wird. Bei Betrachtung der empirischen Werte (-2 log LR) wird aber deutlich, dass

die Kennzahlen deutlich größer sind als der kritische Wert, so dass die Ergebnisse nach Alpha-Adjustierung nichts anders ausfallen.

Tabelle 24 Ergebnisse der Likelihoodquotiententests (letzte Spalte) sowie die die dazu erforderlichen mit LpcM-Win (mit * gekennzeichnet) oder Winmira berechneten Kennwerte. Als L_0 wird jeweils die Likelihood des Rasch-Modells verwendet. RM steht hierbei für Rasch-Modell, MRM für Mixed Rasch-Modell, df bezeichnet die freien Modellparameter.

Bew.gesetze	**L_1**	**L_0**	**-2 log LR**	**Chi^2_{krit}**	**df L_1**	**df L_0**	**df (L_1-L_0)**	**Sig (5%)**
RM Geschlecht	-4349.08	-4392.40	86.64	22.36	26	13	13	ja
RM Leistung	-2480.96*	-2498.81*	35.70*	12.57	12	6	6	ja
MRM 2 Klassen	-4370.84	-4392.40	43.12	21.03	25	13	12	ja
Biochemie	**L_1**	**L_0**	**-2 log LR**	**Chi^2_{krit}**	**df L_1**	**df L_0**	**df (L_1-L_0)**	**Sig (5%)**
RM Geschlecht	-3830.45	-3852.35	43.80	22,36	26	13	13	ja
RM Leistung	-1963.58*	-2040.85*	154.54*	12.57	12	6	6	ja
MRM 2 Klassen	-3746.24	-3852.35	212.22	21.03	25	13	12	ja
Chem. Verbind.	**L_1**	**L_0**	**-2 log LR**	**Chi^2_{krit}**	**df L_1**	**df L_0**	**df (L_1-L_0)**	**Sig (5%)**
RM Geschlecht	-4158.52	-4192.81	68.58	22.36	26	13	13	ja
RM Leistung	-2299.59*	-2310.15*	21.13*	12.57	12	6	6	ja
MRM 2 Klassen	-4178.17	-4192.81	29.28	19.68	24	13	11	ja
Elektrizität	**L_1**	**L_0**	**-2 log LR**	**Chi^2_{krit}**	**df L_1**	**df L_0**	**df (L_1-L_0)**	**Sig (5%)**
RM Geschlecht	-5459.38	-5472.73	26.70	22.36	26	13	13	ja
RM Leistung	-2993.99*	-3034.80*	81.62*	12.57	12	6	6	ja
MRM 2 Klassen	-5422.97	-5472.73	99.52	21.03	25	13	12	ja
Fortpflanzung	**L_1**	**L_0**	**-2 log LR**	**Chi^2_{krit}**	**df L_1**	**df L_0**	**df (L_1-L_0)**	**Sig (5%)**
RM Geschlecht	-4894.08	-4930.27	72.38	22.36	26	13	13	ja
RM Leistung	-2480.79*	-2515.06*	68.53*	12.57	12	6	6	ja
MRM 2 Klassen	-4884.61	-4930.27	91.32	21.03	25	13	12	ja
Wärmeverlust	**L_1**	**L_0**	**-2 log LR**	**Chi^2_{krit}**	**df L_1**	**df L_0**	**df (L_1-L_0)**	**Sig (5%)**
RM Geschlecht	-5017.19	-5041.13	47.88	22.36	26	13	13	ja
RM Leistung	-2629.74*	-2670.12*	80.76*	12.57	12	6	6	ja
MRM 2 Klassen	-4989.02	-5041.13	104.22	21.03	25	13	12	ja

Das Mixed Rasch-Modell mit zwei latenten Klassen passt in allen o.g. Inhaltsbereichen besser auf die Daten als die Einklassenlösung nach dem Rasch-Modell. Dies zeigt sich an den signifikanten Likelihoodquotiententests (Tabelle 24, zum Testverfahren vgl. 2.4).

Der Tabelle ist auch zu entnehmen, dass sich nach dem Likelihoodquotiententest alle hier berechneten Zweiklassenlösungen gegenüber der Einklassenlösung mit dem Rasch-Modell durchsetzen konnten. Modelle mit zwei (oder mehr) Klassen auf die Daten passen dann auf die Daten, wenn Heterogenität in den Antwortmustern zu finden ist.

3.5.7 Interpretation inhaltsspezifischer latenter Klassen

Das Mixed Rasch-Modell mit zwei latenten Klassen lässt sich auf die Daten der naturwissenschaftlichen Inhaltsbereiche anwenden und gegen andere Klassenlösungen absichern. Dieser Befund wird durch die informationstheoretischen Maße und Likelihoodquotiententests zum Modellfit gestützt. Nicht bestätigt haben sich diese Ergebnisse in den drei Inhaltsbereichen Atmung/Fotosynthese, Energieumwandlung und Teilchenkonzept.

Die mit dem Mixed Rasch-Modell identifizierten Klassen zeigen einen größeren qualitativen Unterschied in den Profilen als die Einteilung nach manifestem Geschlecht oder in Niedrig- und Hochscorer, die Abstände zwischen den Profilen sind hier am größten. Eine Teilung in Klassen mit verbal-bewertenden bzw. grafisch-numerisch-abstrakten Stärken konnte dabei aber nicht bestätigt werden.

Inhaltlich sind die Klassen durch ihr Leistungsniveau gekennzeichnet, mit Vorsicht lassen sich die Klassen mit Bezug zu verschiedenen Leistungsaspekten interpretieren. Der Unterschied zwischen den latenten Klassen liegt im Alltagswissen oder dem Praxisbezug: Die Klasse der „Könner“ wendet ihr naturwissenschaftliches Verständnis flexibler auf neue Aufgabenstellungen an, profitiert aber nicht von praxisnahen Aufgaben und wird bei ihnen von der Klasse der „Nicht-Könner“ in der Lösungswahrscheinlichkeit überholt oder liegt etwa gleich. Die Gruppe der „Nicht-Könner“ zeigt Stärken in alltags- und praxisnahen Aufgaben, hat aber starke Schwierigkeiten, das naturwissenschaftliche Alltagsverständnis auf neue Bereiche zu übertragen. Für die drei Inhaltsbereiche Atmung/Fotosynthese, Energieumwandlung sowie Teilchenkonzept konnte ein solcher Zusammenhang nicht nachgewiesen werden: Bei diesen Inhalten kann das Mixed Rasch-Modell nicht auf die Daten angewendet werden.

Dieser qualitative Unterschied zwischen den Klassen ist - mit Ausnahme des Inhaltsbereichs Bewegungsgesetze - nicht mit dem Geschlecht assoziiert: Die latente Variable, die für die Unterschiede in den Daten bedeutsam ist, hängt mit dem (biologischen) Geschlecht nur in geringem Ausmaß zusammen. Die geschlechtsspezifischen Unterschiede sind auf latenter Ebene für nationalen Naturwissenschaftstest und die Profile in den kognitiven Teilkompetenzen für die meisten Inhaltsbereiche zu vernachlässigen.

Ob das durch das Mixed Rasch-Modell identifizierte latente Merkmal mit anderen Leistungsvariablen als dem nationalen Naturwissenschaftstest bei PISA korreliert wird im nächsten Abschnitt zu klären versucht.

3.6 Inhaltsübergreifende kognitive Strukturen

Im vorhergegangen Abschnitt hat sich gezeigt, dass bei Einzelbetrachtung der Inhaltsbereiche aus dem nationalen Naturwissenschaftstest das Mixed Rasch-Modell zwei Klassen von Schülern nach einem bedeutsamen latenten Merkmal getrennt hat. Das Geschlecht scheint für die latente Variable für die meisten Inhaltsbereiche nur wenig relevant zu sein.
Es bleibt zu zeigen, ob die mit dem latenten Teilungskriterium assoziierte kognitive Struktur sich nicht nur inhaltlich, sondern auch empirisch über alle Inhaltsbereiche konsistent zeigt. Um die kognitive Struktur als zentral zur Erklärung von Leistungsunterschieden einzuordnen, muss ein inhaltsübergreifender Bezug zu anderen zentralen PISA-Variablen nachgewiesen werden.

Die Fragestellungen lassen sich wie folgt zusammenfassen:

- *Erweist sich das latente Merkmal nach einer inhaltsübergreifenden Aggregation als stabil?*
- *In welchem Bezug steht das den latenten Klassen zugrundeliegende inhaltsübergreifende Merkmal zu zentralen PISA-Variablen, insbesondere zum Geschlecht und zu den internationalen Leistungstests?*

Das im Mixed Rasch Modell relevante latente Merkmal reflektiert eine Leistungsvariable mit unterschiedlichem Praxiswissen (Abschnitt 3.5.7). Da sie als grundlegender kognitiver Prozess verstanden werden kann, ist zu erwarten, dass eine solche Leistungskomponente in allen Inhaltsbereichen beobachten werden kann.
Weiterhin wird davon ausgegangen, dass das latente Merkmal nur gering mit dem Geschlecht zusammenhängt: Die kognitive Leistungsfähigkeit mit unterschiedlichem Praxiswissen wird als nicht geschlechtsspezifisch betrachtet. Alltagswissen und Erfahrungen können geschlechtsspezifisch gesammelt werden, sollten sich aber in einem Test mit unterschiedlichen Themengebieten und Interessenschwerpunkten insgesamt nicht auswirken.
Der Zusammenhang zu zentralen Leistungsvariablen aus dem PISA-Test sollte hoch sein, da die Leistungsvariablen bei PISA insgesamt hohe Korrelationen aufweisen (Leutner et al., 2004). Auch der Zusammenhang zwischen dem latenten Merkmal und Problemlösekompetenz sollte nicht höher als zu den anderen Leistungsdomänen ausfallen: Die Problemlösekompetenz ist fächerübergreifend angelegt, beinhaltet aber nicht

zwangsläufig solche Problemstellungen, die das Alltagswissen oder praktische Erfahrungen besonders ansprechen.

Unklar ist, in welcher Weise die latente Variable in Beziehung zu den Motivationsskalen aus PISA Pekrun & Zirngibl (2004) steht und ob sich hier deutlichere Zusammenhänge zeigen.

3.6.1 Inhaltsübergreifende Aggregation von Klassenzugehörigkeiten

In Abschnitt 3.5 wurden die Klassenlösungen des Mixed Rasch-Modells inhaltlich interpretiert. Dies wurde für jeden Inhaltsbereich separat vorgenommen.

Die Annahme eines gemeinsamen zugrundeliegenden Merkmals begründet eine inhaltsübergreifende Aggregation der Klassenzuteilung durch das Mixed Rasch-Modell.

Dazu wurden die Klassen über alle Inhalte sortiert, da bei Winmira die Klassennummer nach Größe zugewiesen wird. Beim Mixed Rasch-Modell ist also die größere Klasse immer die erste Klasse. Es ist nicht anzunehmen, dass das latente Merkmal in Klasse 1 im Inhaltsbereich Elektrizität dieselbe Ausprägung annimmt wie das Merkmal in Klasse 1 im Inhaltsbereich Biochemie/Ernährung.

Um das Merkmal über die Inhalte gleich zu richten, wird als Kriterium das Item herangezogen, das den Umgang mit mentalen Modellen erfordert, da es in vielen Inhaltsbereichen die geringsten Lösungswahrscheinlichkeiten aufweist und sich als relevant zur Interpretation der Klassenprofile herausgestellt hat (vgl. 3.5.4).

Zu den nachfolgenden Analysen wurden die Klassenzugehörigkeit so umgepolt, dass der Mixed Rasch-Klasse mit einem höheren Schwierigkeitsparameter im Parameterprofil Null zugeordnet wurde, der anderen Eins. Diese Zuteilung ist willkürlich und hätte auch umgekehrt erfolgen können. Tabelle 25 gibt wieder, wie die Zuteilungen vorgenommen wurden.

Tabelle 25 Dargestellt ist die Umpolung von Klassen.

	Klasse im Mixed RM	Zuteilung nach Itemprofilen
Bewegungsgesetze	1	1
Biochemie und Ernährung	1	2
Chemische Verbindungen	1	2
Elektrizität	1	2
Fortpflanzung	1	2
Wärmeverlust	1	1

Wie aus Tabelle 26 hervorgeht, liegen durch den Ausschluss von Inhaltsbereichen für die Schüler Klassenzugehörigkeiten aus ein bis vier Inhaltsbereichen vor: Die möglichen Klassenpattern bestehen aus Tupeln, Tripeln und Quadrupeln und wurden von einer unterschiedlichen Anzahl von Schülern bearbeitet (Tabelle 27).

Tabelle 26 Schülern im Test vorgelegte Blöcke von Inhaltsbereichen, die in verschiedenen Kombinationen erfasst worden sind. Unterlegt sind die Inhaltsbereiche, die aus den Analysen ausgeschlossen wurden (vgl. 3.5.4).

Block 1	Atmung	Teilchenkonzept	Energieumwandlung	Ernährung
Block 2	Atmung	Elektrizität	Wärmeverlust	Chem. Verbindungen
Block 3	Bewegungsgesetze	Elektrizität	Wärmeverlust	Ernährung
Block 4	Bewegungsgesetze	Elektrizität	Energieumwandlung	Atmung
Block 5	Teilchenkonzept	Elektrizität	Wärmeverlust	Fortpflanzung
Block 6	Fortpflanzung	Energieumwandlung	Chem. Verbindungen	Ernährung
Block 7	Fortpflanzung	Bewegungsgesetze	Chem. Verbindungen	Teilchenkonzept
Block 8	Fortpflanzung	Atmung	Energieumwandlung	Wärmeverlust

Tabelle 27 Prozentuale Verteilung der Schüler auf die Anzahl bearbeiteter Domänen

Anzahl bearbeiteter Inhaltsbereiche	Anteil der Schüler in %
1	12.6
2	25.3
3	49.5
4	12.6

Die Klassenzugehörigkeiten der einzelnen Inhaltsbereiche können nur dann aggregiert werden, wenn die Zuordnung in den meisten Fällen über die Inhalte konsistent erfolgte.

Die Häufigkeiten der Klassenpattern je nach Blockkombination (Schüler in %) sowie die mittlere Trefferwahrscheinlichkeit zeigen Tabelle 28 bis Tabelle 30. Letztere wurde als Mittelwert der Zuordnungswahrscheinlichkeiten zu einer der beiden Mixed Rasch-Klassen berechnet. Den Tabellen ist zu entnehmen, dass sowohl die mittleren Trefferwahrscheinlichkeiten als auch die Häufigkeiten auf eine übereinstimmende Klassenzugehörigkeit über die Inhaltsbereiche hinweg (z.B. (1 1) im Tupel, (1 1 1) im Tripel und (1 1 1 1) im Quadrupel) hinweisen.

Deshalb wird für jeden Schüler eine individuelle aggregierte Klasse über die Inhaltsbereiche hinweg gebildet, die zugeordnete Klasse je nach Klassenpattern ist Tabelle 28 bis Tabelle 30 zu entnehmen. Bei uneindeutigen Kombinationen (z.B. (1 2) im Tupel oder (1 1 2 2) im

Quadrupel) wird die wahrscheinlichste Klasse zugeordnet: diejenige, bei der die Zuordnungswahrscheinlichkeit der Mixed Rasch-Klassen maximal ist.

Tabelle 28 Relative Häufigkeit (in %) und mittlere Trefferwahrscheinlichkeit (p) der Klassenpattern (Tupel) in Block 4 und Block 8 (vgl. Tabelle 26).

Block 4			
Klassenpattern Bewegungs-gesetze	Elektrizität	% Schüler	p
1	1	31,1	0,82
1	2	13,0	0,77
2	1	34,3	0,82
2	2	21,6	0,86
Block 8			
Klassenpattern Fortpflanzung und Sexualität	Wärme-umwandlung	% Schüler	p
1	1	34,4	0,85
1	2	16,9	0,77
2	1	16,0	0,77
2	2	32,8	0,82

Tabelle 29 Relative Häufigkeit (in %) und mittlere Trefferwahrscheinlichkeit (p) der Klassenpattern (Tripel) in den Blöcken 2, 5, 6 und 7 (vgl. Tabelle 26). Die letzte Spalte zeigt die zugewiesene aggregierte Klasse. Zur besseren Lesbarkeit sind einige Zeilen unterlegt.

Block 2					
Klassenpattern					
Chemische Verbindungen	Elektrizität	Wärme-umwandlung	% Schüler	p	aggregierte Klasse
1	1	1	33,6	0,88	1
1	1	2	16,8	0,81	1
1	2	1	3,7	0,79	1
1	2	2	9,5	0,79	2
2	1	1	3,4	0,79	1
2	1	2	8,3	0,74	2
2	2	1	1,2	0,65	2
2	2	2	23,5	0,87	2
Block 5					
Klassenpattern					
Fortpflanzung und Sexualität	Elektrizität	Wärme-umwandlung	% Schüler	p	aggregierte Klasse
1	1	1	34,5	0,89	1
1	1	2	7,7	0,80	1
1	2	1	5,5	0,83	1
1	2	2	6,2	0,81	2
2	1	1	5,8	0,83	1
2	1	2	9,5	0,79	2
2	2	1	4,3	0,76	2
2	2	2	26,5	0,87	2
Block 6					
Klassenpattern					
Fortpflanzung und Sexualität	Chemische Verbindungen	Ernährung	% Schüler	p	aggregierte Klasse
1	1	1	21,6	0,88	1
1	1	2	17,3	0,80	1
1	2	1	1,9	0,80	1
1	2	2	3,4	0,80	2
2	1	1	6,2	0,78	1
2	1	2	21,6	0,82	2
2	2	1	1,2	0,66	2
2	2	2	26,9	0,87	2
Block 7					
Klassenpattern					
Fortpflanzung und Sexualität	Chemische Verbindungen	Bewegungs-gesetze	% Schüler	p	aggregierte Klasse
1	1	1	20,0	0,80	1
1	1	2	25,8	0,81	1
1	2	1	6,5	0,76	1
1	2	2	4,6	0,74	2
2	1	1	6,8	0,78	1
2	1	2	8,0	0,76	2
2	2	1	12,0	0,75	2
2	2	2	16,3	0,83	2

Tabelle 30 Relative Häufigkeit (in %) und mittlere Trefferwahrscheinlichkeit (p) der Klassenpattern (Quadrupel) in Block 3 (vgl. Tabelle 26). Die letzte Spalte zeigt die zugewiesene aggregierte Klasse. Bei uneindeutigen Pattern wurde nach Wahrscheinlichkeit zugeordnet. Zur besseren Lesbarkeit sind einige Zeilen unterlegt.

Block 3						
Klassenpattern						
Ernährung	Wärme-umwandlung	Bewegungs-gesetze	Elektrizität	% Schüler	p	aggregierte Klasse
1	1	1	1	9,6	0,90	1
1	1	2	1	13,9	0,85	1
1	1	1	2	0		1
1	1	2	2	0,3	0,66	nach Wk
1	2	1	1	0,6	0,70	1
1	2	2	1	3,9	0,81	nach Wk
1	2	1	2	0,3	0,85	nach Wk
1	2	2	2	0,3	0,87	2
2	1	1	1	6,9	0,83	1
2	1	2	1	16,0	0,82	nach Wk
2	1	1	2	1,8	0,78	nach Wk
2	1	2	2	2,4	0,83	2
2	2	1	1	10,5	0,80	nach Wk
2	2	2	1	14,2	0,84	2
2	2	1	2	6,9	0,82	2
2	2	2	2	12,3	0,90	2

3.6.2 Bezug zu Variablen aus dem PISA-Test

Die für den Schüler vorliegende aggregierte Klassenzugehörigkeit kann in Beziehung zu anderen Variablen gesetzt werden. Da sich die Klassen als leistungsassoziiert erwiesen haben, wird im Folgenden von L+ (höhere Leistung) und L- (niedrigere Leistung) gesprochen.

In Tabelle 31 sind die Anteile von Jungen und Mädchen in den Klassen wiedergegeben. Die Anteilsunterschiede fallen gering aus, in L+ sind etwas mehr Mädchen vertreten, in L- etwas mehr Jungen.

Tabelle 31 Anteile von Mädchen und Jungen in den aggregierten latenten Klassen. Klasse L+ steht hierbei für die obere aggregierte Leistungsklasse, L- für die untere aggregierte Leistungsklasse. Die Zahlen addieren sich zeilenweise zu 100%.

Klasse	Anteil Mädchen	Anteil Jungen
Klasse L+	51,89	48,11
Klasse L-	46,62	53,38

Tabelle 32 stellt die mittleren Kompetenzwerte für die zwei aggregierten Klassen L+ und L- dar. Die Kompetenzunterschiede in den Klassen betragen bei den internationalen

Kompetenzbereichen etwa 110-130 Kompetenzpunkte (etwas mehr als eine Standardabweichung). Diese Zahlen drücken aus, wie stark der Leistungsunterschied ist.

Tabelle 32 Angegeben sind die mittleren Kompetenzpunkte insgesamt, sowie nach Geschlecht getrennt und als Kompetenzunterschied zwischen Jungen und Mädchen (J-M) für die beiden aggregierten Klassen. Gezeigt sind die vier internationalen Leistungsdomänen Mathematik, Lesen, Naturwissenschaften und Problemlösen (internationale Metrik mit MW=500 und SD=100) sowie nationalen Naturwissenschaften (nationale Metrik mit MW= 50 und SD=10).

int. Mathematik	L+	L-	Klassendifferenz
mittlere Kompetenz	569,92	451,11	118,80
mittlere Kompetenz Mädchen	562,56	444,20	
mittlere Kompetenz Jungen	577,86	457,16	
mittlere Kompetenzdifferenz (J-M)	15,30	12,96	
int. Lesen	**L+**	**L-**	**Klassendifferenz**
mittlere Kompetenz	557,06	432,91	124,15
mittlere Kompetenz Mädchen	570,13	454,91	
mittlere Kompetenz Jungen	542,96	413,70	
mittlere Kompetenzdifferenz (J-M)	-27,17	-41,21	
int. Naturwissenschaften	**L+**	**L-**	**Klassendifferenz**
mittlere Kompetenz	580,74	452,33	128,41
mittlere Kompetenz Mädchen	574,71	447,57	
mittlere Kompetenz Jungen	587,25	456,49	
mittlere Kompetenzdifferenz (J-M)	12,54	8,93	
int. Problemlösen	**L+**	**L-**	**Klassendifferenz**
mittlere Kompetenz	571,60	459,67	111,93
mittlere Kompetenz Mädchen	567,19	462,04	
mittlere Kompetenz Jungen	576,36	457,61	
mittlere Kompetenzdifferenz (J-M)	9,17	-4,43	

Zu den Geschlechterunterschieden in den Leistungsklassen ist zu bemerken, dass bezüglich Mathematik und Naturwissenschaften (internationale Tests) in der Tendenz stärkere Geschlechterunterschiede in der leistungsstärkeren Gruppe (L+) zu verzeichnen sind. Allerdings liegen diese nur bei etwa einer bis anderthalb zehntel Standardabweichungen also zwischen 12 und 15 Kompetenzpunkten zu Gunsten der Jungen.

Die mit Abstand größten Kompetenzunterschiede mit bis zu 41 Kompetenzpunkten (fast einer halben Standardabweichung) zeigen sich im Lesen, allerdings sind hier die Unterschiede in der leistungsschwächeren Klasse deutlich größer.

Diese Ergebnisse reihen sich in die Befunde des nationalen PISA 2003 Berichts ein. So berichten Rost et al. (2004) über größere Unterschiede im oberen Kompetenzbereich, dem vierten Leistungsquartil. Dort betragen die Geschlechterunterschiede im internationalen

Naturwissenschaftstest 14 Kompetenzpunkte, dies entspricht ziemlich genau dem Leistungsunterschied zwischen Mädchen und Jungen in der leistungsstarken Mixed Rasch-Klasse von etwa13 Punkten in den Naturwissenschaften.

Ein anderer Ansatz bei PISA 2003 zeigt Geschlechterunterschiede in Extremgruppen, die nach Kompetenzstufen gebildet wurden Zimmer et al. (2004). Hierbei ist zu beachten, dass die Kompetenzstufen im Gegensatz zu Leistungsquartilen problematisch sind, dass die Anteile von Jungen und Mädchen auf den Kompetenzstufen variieren, wohingegen die Anteile in den Quartilen fix sind, da jeweils die oberen 25% der Daten eingeschlossen werden und so die Kompetenzen von 25% der Mädchen mit denen von 25% der Jungen verglichen werden können. Aus diesen Gründen kann das Mischungsverhältnis bei Kompetenzstufen die Geschlechterunterschiede verzerren.

Trotz einer möglichen Maskierung der Unterschiede zwischen Mädchen und Jungen durch unterschiedliche Anteile passen die dargestellten Befunde der Autoren zu dem Bild, das die Leistungsklassen nach dem Mixed Rasch-Modell zeigen:

In der internationalen Mathematik- und Naturwissenschaftskompetenz gibt es größere Geschlechterunterschiede im oberen Leistungsbereich, während beim Lesen die größeren Unterschiede im unteren Leistungsbereich zu beobachten sind, die Unterschiede im Lesen sind von allen Kompetenzbereichen am größten. Lediglich im Problemlösen gibt es in beiden Leistungsgruppen keine nennenswerten Geschlechterunterschiede.

Insgesamt zeigt sich der Befund aus dem Mixed Rasch-Klassen kongruent zu den Ergebnissen aus PISA 2003.

Korrelationen geben einen ersten Eindruck über den linearen Zusammenhang zwischen aggregierter Klasse und zentralen PISA-Variablen. Hierbei ist zu bedenken, dass nichtlineare Zusammenhänge nicht erfasst werden; Drittvariablen können die Korrelationen zweier Variablen beeinflussen. Wird eine Drittvariable vermutet, so sollte diese auspartialisiert werden. Mit der Partialkorrelation wird rechnerisch eine Konstanthaltung der Drittvariablen erreicht, um die Korrelation zwischen zwei Variablen zu erhalten, die um den Einfluss der Drittvariablen bereinigt ist.

Bei der Korrelation zwischen einer dichotomen und intervallskalierten Variablen wird der punktbiseriale Korrelationskoeffizient r_{pbis} verwendet. Er ergibt sich durch Umformung der Produkt-Moment-Korrelation r_{xy} zweier intervallskalierter Merkmale. Sollte ein nicht-linearer Zusammenhang vorliegen bzw. vermutet werden, ist es nützlich, das Zusammenhangsmaß Eta η zu berechnen. Der Koeffizient Eta erfasst sowohl lineare wie nicht-lineare Zusammenhänge.

Liegt ein linearer Zusammenhang vor, entspricht Eta dem Wert des Korrelationskoeffizienten, bei nicht-linearen Beziehungen übertrifft Eta die Korrelation, so dass die Differenz zwischen der Eta und der Korrelation anzeigt, welches Ausmaß dem nicht-linearen Zusammenhang zuzusprechen ist. Analog zu R^2 gibt η^2 den Anteil aufgeklärter Varianz an. Ursprünglich verwendet wird Eta in der Varianzanalyse.

Genau genommen erlaubt das Skalenniveau vieler Variablen eine Berechnung einfacher Korrelationen nicht, so dass die berichteten Kennwerte verzerrt sein könnten. Deshalb wurden andere Maße des Zusammenhangs herangezogen. Tabelle 33 zeigt die verwendeten Maße und zugehörige Anwendungsfelder. Weiterführende Informationen dazu finden sich z.B. bei Hofstätter & Wendt (1974) oder Siegel (1956).

Tabelle 33 Neben Korrelationen verwendete Zusammenhangsmaße

Koeffizient	Beschreibung
Phi Φ	Zusammenhangsmaß für zwei dichotome Variablen
Cramers V	Zusammenhangsmaß für zwei nominalskalierte Variablen, bei dem mindestens eine Variable mehr als zwei Ausprägungen hat
Eta η	lineares und nicht-lineares Zusammenhangsmaß zwischen kategorialer und intervallskalierter Variable

Tabelle 34 Zusammenhangsmaße zwischen aggregierter Klasse und zentralen PISA-Variablen. Berechnet wurden Korrelationen (r), Phi, Cramers V oder Eta sowie die Partialkorrelation, bei der der Summenscore des nationalen Naturwissenschaftstests auspartialisiert wurde.

Variable	Korrelation mit aggregierter Klasse	Φ / V / η mit aggregierter Klasse	partial r
Mathematikkompetenz im internationalen Test	-0,62	0,62 (η)	-0,15
Naturwissenschaftskompetenz im internationalen Test	-0,62	0,62 (η)	-0,16
Lesekompetenz im internationalen Test	-0,59	0,59 (η)	-0,15
Problemlösekompetenz im internationalen Test	-0,60	0,60 (η)	-0,16
kognitive Grundfähigkeit	-0,51	0,51 (η)	-0,11
Geschlecht	0,05	0,05(Φ)	0,04
Motivationsskala Selbstkonzept Mathematik	-0,08	0,08 (η)	-0,01
Motivationsskala Angst in Mathematik	0,17	0,17 (η)	0,02
Motivationsskala Interesse an Mathematik	-0,02	0,02 (η)	0,03
Motivationsskala Kompetitives Lernen	0,08	0,08 (η)	0,02
Motivationsskala Instrumentelle Motivation	0,03	0,03 (η)	0,03
Index of economic, social and cultural status (ESCS)	-0,35	0,35 (η)	-0,05
Migrationshintergrund	0,20	0,23(V)	0,03

Tabelle 34 gibt zunächst die einfachen Korrelationen zwischen aggregierter Klasse und einer Auswahl zentraler PISA-Variablen an. Für eine genauere Beschreibung sei auf den nationalen PISA-Berichtsband verwiesen (Pekrun & Zirngibl, 2004; Prenzel et al., 2004a).

An den Kennzahlen wird deutlich, dass die Korrelationen zu Leistungsvariablen in den internationalen Kompetenzbereichen (Mathematik, Lesen Naturwissenschaften und Problemlösen) sowie zur kognitiven Grundfähigkeit wie erwartet hoch sind. Der Zusammenhang zwischen aggregierter Klasse und Problemlösetest unterscheidet sich nicht von den anderen Leistungsdomänen. Ein Zusammenhang zu Motivationsvariablen konnte sich nicht bestätigen.

Wird das Leistungsniveau (über den Summenscore im nationalen Naturwissenschaftstest) auspartialisiert (Tabelle 34, partial r) fallen die Zusammenhänge gering aus. Dies wird durch eine Analyse über multiple Regression (Auswahl der Prädiktoren wie in Tabelle 34) bestätigt: Wenn man den Summenscore aus dem Regressionsmodell herausnimmt, zeigt sich eine Abnahme in dem Anteil aufgeklärter Varianz von 0,3.

Um zu stützen, dass sich die Klassen in ihrem Antwortverhalten bzgl. praxisbezogener Aufgaben unterscheiden, müssten detailliertere Analysen mit neuen Testverfahren angeschlossen werden.

Die Werte für Korrelation und Eta sind identisch bzw. nahezu identisch, so dass keine nicht-linearen Zusammenhänge beobachtbar sind: Es gibt bezogen auf den Vergleich mit Eta keinen nennenswerten nicht-linearen Zusammenhang zwischen den ausgewählten Variablen und der Klassenteilung im Mixed Rasch-Modell.

Im Vergleich zwischen Korrelationskoeffizienten und Phi sowie Cramers V ergeben sich nur geringe bzw. keine Unterschiede zur einfachen Korrelation.

Das latente Merkmal hat sich als stärker leistungs- und weniger geschlechterkorreliert erwiesen. Würden stärkere Unterschiede zwischen Jungen und Mädchen in den kognitiven Teilkompetenzen vorhanden sein und diese in bedeutsamer Weise die Population teilen, hätte die Klassenlösung nach dem Mixed Rasch-Modell stärker geschlechtskorreliert sein müssen. Wenn die Gruppenunterschiede im Mixed Rasch-Modell ohne Berücksichtigung der manifesten Geschlechtsvariable nicht mit dem Geschlecht korrelieren, dann hat das manifeste (biologische) Geschlecht nur wenig Einfluss auf die Unterschiede in den Daten. Auf latenter Ebene sind die geschlechtsspezifischen Unterschiede für nationalen Naturwissenschaftstest inhaltsübergreifend zu vernachlässigen. Für die Profile der kognitiven Teilkompetenzen ist das Geschlecht nicht so aussagekräftig wie die Leistung.

Die eingangs formulierten Fragestellungen lassen sich folgendermaßen beantworten:
Das latente Merkmal hat sich auch inhaltsübergreifend als bedeutsam herausgestellt. Ein Zusammenhang zu den internationalen Leistungsdomänen bei PISA konnte nachgewiesen werden. Die der Teilung im Mixed Rasch-Modell zugrunde liegende kognitive Struktur hängt weder mit den bei PISA erfassten Motivationsskalen noch mit dem Geschlecht zusammen.

3.7 Beantwortung der Fragestellungen

In Abschnitt 3.3 konnte gezeigt werden, dass für die überwiegende Anzahl der im PISA-Test eingesetzten Items keine bedeutsamen Geschlechterunterschiede in inhaltlicher Einkleidung und kognitiver Anforderung bestehen. Bei einer kleinen Anzahl von Items zeigt sich eine Geschlechtsspezifität in einer Kombination aus ansprechendem Inhalt und kognitiver Anforderung, die überwiegend durch Stereotype Interessens- und Fähigkeitsmuster erklärt werden können. Bei einem geringen Teil der Items zeigt sich ein solcher Effekt über eine Auswahl von PISA-Teilnehmerländern hinweg.
Eine aus den kognitiven Teilkompetenzen des nationalen Naturwissenschaftstests (verbal-bewertend vs. grafisch-numerisch-abstrakt) konstruierte Skala (vgl. 3.4) sollte die Leistungsunterschiede bei PISA besser erklären als das biologische Geschlecht. Die Skala zeigte sich nach dem Konzept der Konstruktvalidität mit dem Geschlecht assoziiert, konnte sich aber in ihrer Brauchbarkeit nicht gegenüber dem biologischen Geschlecht durchsetzen.

Um Unterschiede qualitativer und quantitativer Art zwischen Personengruppen aufzudecken, lässt sich das Mixed Rasch-Modell mit zwei latenten Klassen auf die Daten der naturwissenschaftlichen Inhaltsbereiche im nationalen Test anwenden und gegen andere Klassenlösungen absichern. Davon ausgenommen sind die drei Inhaltsbereiche Atmung/Fotosynthese, Energieumwandlung und Teilchenkonzept.
Die mit dem Mixed Rasch-Modell identifizierten Klassen zeigen in den Profilen der Teilkompetenzen einen qualitativen Unterschied, der größer ist als bei einer Einteilung nach manifestem Geschlecht. Eine Teilung nach verbal-bewertenden bzw. grafisch-numerisch-abstrakten Kompetenzen konnte dabei aber nicht bestätigt werden.
Die zugrunde liegende latente Trennungsvariable im Mixed Rasch-Modell ist inhaltlich durch das Leistungsniveau zu beschreiben, gehen aber darüber hinaus: Mit Vorsicht lassen sich die Klassen mit Bezug zu verschiedenen Leistungsaspekten interpretieren. Dabei machen die Aufgaben zum Alltagswissen oder Praxisbezug den Unterschied zwischen den Klassen aus.
Dieser Unterschied zwischen den Klassen ist - mit Ausnahme des Inhaltsbereichs Bewegungsgesetze - nicht mit dem (biologischen) Geschlecht assoziiert. Die geschlechtsspezifischen Unterschiede auf latenter Ebene sind für den nationalen Naturwissenschaftstest und die Profile in den kognitiven Teilkompetenzen für die meisten Inhaltsbereiche zu vernachlässigen.

Das latente Merkmal hat sich auch über die Inhalte hinweg als bedeutsames Merkmal herausgestellt, das stärker leistungs- und weniger geschlechterkorreliert ist. Der Zusammenhang zu den PISA-Leistungsvariablen bestätigt diesen Zusammenhang. Das manifeste Geschlecht hat nur wenig Einfluss auf die Unterschiede in den Daten des nationalen Naturwissenschaftstests. Für die Profile der kognitiven Teilkompetenzen ist das Geschlecht nicht so aussagekräftig wie die qualitativen Unterschiede in den Leistungsklassen. Das Geschlecht tritt also nicht nur hinter den allgemeinen Niveauunterschied, sondern auch hinter die qualitativen Unterschiede zwischen Leistungsklassen zurück.

Die Unterschiede wurden auf latenter Ebene für das Geschlecht gesucht, konnten aber „nur“ für Leistung identifiziert werden. Dies lässt den Schluss zu, dass die Geschlechterunterschiede bei PISA, insbesondere im nationalen Naturwissenschaftstest von untergeordneter Bedeutung sind.

4 Diskussion

Bei der Skalierung der internationalen PISA-Items in Mathematik, Naturwissenschaften, Lesen und Problemlösen und der nationalen Items in Mathematik und Naturwissenschaften ließen sich nur wenige Items als geschlechtsspezifisch identifizieren. Diese bestätigten aber geschlechtsstereotype Unterschiede hinsichtlich Interesse und kognitiver Anforderungen.
Die Bedeutsamkeit des Geschlechts für die Leistungen im PISA-Test konnte in anderen Analysen nicht nachgewiesen werden. Aus den für den nationalen Naturwissenschaftstest relevanten kognitiven Teilkompetenzen konnte keine für die Leistungsvorhersage valide geschlechtskorrelierte Skala konstruiert werden. Außerdem konnten die Teilkompetenzen nicht zweifelsfrei geschlechtsspezifisch (verbal-bewertend vs. grafisch-numerisch-abstrakt) zugeordnet werden. Dies ließ sich über die Identifizierung und Interpretation latenter Klassen zeigen. Das Geschlecht tritt hinter einem anderen, wesentlichen bedeutenderen Aspekt von Heterogenität zurück:
Obwohl qualitative Unterschiede zwischen Jungen und Mädchen Im Zentrum des Interesses standen, wurden qualitative Unterschiede zwischen Leistungsklassen gefunden. Dieser Unterschied besteht in der Anwendung von Alltagswissen oder praxisnahen Fähigkeiten. Während die leistungsschwächeren Schüler hier ihre relativen Stärken einbringen können, lassen sich bei leistungsstarken Schülern Schwächen beobachten.
Um die Ergebnisse einordnen zu können, werden im Folgenden die verwendeten Methoden diskutiert, Schlussfolgerungen und daraus abzuleitende Perspektiven skizziert.

4.1 Diskussion der Methoden

Verwendete Erhebungsinstrumente
Die verwendeten nationalen und internationalen Erhebungsinstrumente wurden im Rahmen von PISA 2003 appliziert. Für eine hohe Aufgabenqualität garantierten die Aufgabenentwicklung durch internationale und nationale Expertengruppen sowie Erprobungen im Feldtest. Standardisierte Übersetzungen in mehrschrittiger Überprüfung gewährleisten eine grundsätzliche internationale Vergleichbarkeit – wenn sich auch ein landesspezifischer Erfahrungshintergrund in unterschiedlicher Bearbeitung niederschlagen kann. Internationale Vorgaben vereinheitlichte Testsitzungen, Testleiterschulungen, Auswerter-Trainings sowie ein externes Qualitätsmonitoring sicherten die Erfüllung hoher

Standards von Vergleichbarkeit über Länder hinweg. Durch nationale Ergänzungen konnte ein Bezug zum deutschen Curriculum hergestellt werden. Reliabilität, Validität und Objektivität der Instrumente, insbesondere der Kompetenztests, sind mehrfach nachgewiesen (vgl. Carstensen et al., 2004; OECD, 2004; OECD, 2005; Prenzel et al., 2004a).
Die Erhebungsinstrumente wurden auch auf Geschlechterfairness geprüft (OECD, 2005). Limitationen in Bezug auf Fragen nach Geschlecht und *gender* bestehen aber darin, dass die erhobenen Hintergrundvariablen eine Erfassung von Geschlechtsrollen oder Geschlechtsidentität unmöglich machen. So konnten die Ergebnisse nicht zu diesen Merkmalen in Beziehung gesetzt werden, was eine Veränderung der Befundlage vermuten lässt.

Stichprobenziehung und Repräsentativität

Die den Daten zugrunde liegende Stichprobenziehung genügt durch internationale Vorgaben den hohen Standards bei Schulleistungsstudien. Eine mehrfach stratifizierte Wahrscheinlichkeitsstichprobe aus der Zielpopulation sichert die Repräsentativität der Stichprobe. Durch eine genaue Erfassung und eine hinreichende Ausschöpfung der Zielpopulation (in Deutschland 96,2%) wurde einer Leistungsverzerrung der Ergebnisse entgegengewirkt. Eine selektive Schülerbeteiligung konnte in Deutschland ausgeschlossen werden (Prenzel et al., 2004b).
Durch die Altersstichprobe (15-jährige in Bildungssystemen) ergeben sich ggf. Konsequenzen für die Aussagen über Geschlechterrollen. Die Jungendlichen in diesem Alter befinden sich in einer Phase, wo die eigene Geschlechtsidentität eine große Rolle spielt. D.h. Geschlechtsrollenstereotype sind in der getesteten Population womöglich besonders präsent. Dieser Aspekt kann dazu geführt haben, dass die Inhalte der Aufgaben die geschlechtsspezifischen Interessen betonend wahrgenommen wurden (Ergebnisse aus 3.3), dies gilt besonders, da die Testsitzungen gemischtgeschlechtlich stattfanden. Dabei wird die Geschlechtsidentität stärker aktiviert (vgl. dazu Kessels, 2002; 2004).
Eine Untersuchung auf geschlechtsspezifische Itemparameter und die Bedeutsamkeit geschlechtskorrelierter Klassen sollte demnach auch im monoedukativen Kontext geprüft werden.

Auswertungsverfahren und technische Grundlagen

Die zu den Analysen herangezogenen Auswertungsverfahren stammen aus der Item Response Theory (IRT) und entsprechen in der Anwendung den internationalen und nationalen

Auswertungsprozeduren im Rahmen von PISA 2003. Sofern nicht selbst berechnet, wurden Leistungsschätzwerte und Antworten auf Ebene der Einzelitems internationalen und nationalen Datensätzen entnommen. Die Rasch-Skalierbarkeit der Items wurde bei PISA in Feldtest und Haupttest überprüft.

In die Analysen wurde Schularten oder anderen Merkmale der deutschen Bildungslandschaft (wie Stadt/Land, etc.) nicht einbezogen. Unter Umständen ergeben sich (sowohl hinsichtlich der Leistungsklassen als auch der Geschlechtsspezifität) andere Befunde, wenn nur Teile der Stichprobe betrachtet oder Kontrollvariablen verwendet werden.

Verwendet wurden das Rasch-Modell für dichotome Daten sowie das Mixed Rasch-Modell. Beide Methoden stellten sich als erfolgreich für die Beantwortung der Fragen nach bedeutsamen Geschlechterunterschieden heraus. Dabei zeigen gerade die Annahmen der IRT eine gute Anknüpfbarkeit an die Zusammenhänge zwischen Geschlecht und Leistung: Das dem Testverhalten zugrunde liegende latente Merkmal steht „nur“ in einem Wahrscheinlichkeitszusammenhang zum Testergebnis.

Für die vorliegenden Analysen mit dem Mixed Rasch-Modell ist von Bedeutung, dass die Zuordnung zu Klassen stets mit einer bestimmten Wahrscheinlichkeit erfolgt. Mit anderen Worten: Eine Person wird der wahrscheinlichsten Klasse zugeordnet. Mit einer geringeren Wahrscheinlichkeit gehört sie aber auch der anderen Klasse an. In dieser Arbeit gingen die Zuordnungswahrscheinlichkeiten in die Betrachtungen zwar mit ein (vgl. Abschnitt 3.5 und 3.6) und wurden bei der inhaltsübergreifenden Analyse bei uneindeutiger Zuordnung mit einberechnet (s. 3.6.1), es erfolgte allerdings keine durchgehende Bezugnahme in Form einer stärkeren Verrechung der Trefferwahrscheinlichkeiten bei der Aggregation über Inhaltsbereiche. Dies sollte in weiteren Analysen berücksichtigt werden.

4.2 Schlussfolgerungen und Ausblick

Perspektiven der Geschlechterforschung

Positiv formuliert bedeutet die Kenntnis von geschlechtsspezifischen Items, dass durch entsprechende Anforderungen und inhaltliche Einkleidung Jungen zur korrekten Bearbeitung von Leseaufgaben und Mädchen zu einer erfolgreichen Lösung von Naturwissenschafts- und Mathematikaufgaben motiviert werden können. Ein ansprechendes (geschlechtsspezifisches) Material im Unterricht kann also hilfreich sein, um Lernstoff für Schüler stärker zugänglich

zu machen. Diese Erkenntnis ist nicht neu. Schade ist, dass sich die Inhalte noch immer in stereotyper Weise unterscheiden.

Für die Geschlechterforschung ergibt sich aus den Analysen eine methodische Botschaft: Ein neuer Schwerpunkt könnte darin bestehen, ein latentes Geschlecht zu konstruieren oder das Geschlecht in einem Wahrscheinlichkeitszusammenhang zur Leistung zu betrachten. Hierbei sollte der Einsatz latenter Testmodelle in Zukunft genauer geprüft werden. Wie in der vorliegenden Arbeit gezeigt werden konnte, bietet gerade die Trennung auf latenter Ebene den Vorteil, Personendaten quantitativ und qualitativ zu analysieren: Die Bedeutsamkeit einer trennenden Variable ergibt sich dabei aus den Daten selbst, quasi vorurteilsfrei. Post hoc kann dann der Bezug zu inhaltlichen Konzepten hergestellt werden. Dabei wird vermieden, Unterschiede stärker zu betonen, als aufgrund der Unterschiede in den Daten gegeben. Möglicherweise eröffnen sich so gänzlich neue Perspektiven und Schlussfolgerungen.

Es ist aber auch unter Einsatz neuer Methoden unklar, wohin aktuelle und zukünftige Geschlechterforschung gehen soll, und ob psychologische Geschlechterforschung noch politisch korrekt ist. Eine genauere Erklärung von Geschlechterunterschieden ist nicht zwangsläufig nötig, wenn das Ziel darin besteht, Geschlechterunterschiede in kognitiven Leistungen, der Schullaufbahn und der Berufswahl zu reduzieren.
In diesem Zusammenhang erscheint es sinnvoller, von der Erhebung und Erklärung der Geschlechterunterschiede abzusehen. Ziel wäre stattdessen die Entwicklung empirisch fundierter Maßnahmen und Interventionen zur Herstellung von Geschlechterfairness, auf denen politische Maßnahmen fußen könnten.
Ob und wie sich dazu grundlegende gesellschaftliche Strukturen, die von Anbeginn auf jeden Menschen ein- und auswirken, ändern müssen, bleibt zu klären.

Horstkemper (2002) sieht es trotzdem als wichtige Aufgabe an, das Geschlecht künftig weiterhin in alle Fragen der Bildungsforschung mit einzubeziehen. Dabei sollen sowohl Inhalte als auch die Methoden von Bildung sowie die Institutionen selbst ins Zentrum gerückt werden. Konkret würde das bedeuten, Geschlechterforschungsgruppen in die Bildungsstudien stärker einzubeziehen. Wenn es auch in Zukunft im Rahmen jeder Schulleistungsstudie Thema sein wird, Geschlechterunterschiede im Sinne von Chancengleichheit zu erheben, müssen Erhebungen zur Geschlechtsspezifik des Schulalltags die Leistungserhebungen ergänzen.

Um das realisieren zu können, würde sich für PISA und andere Schulleistungsstudien in Zukunft zusätzlich zum biologischen Geschlecht die Erfassung der Geschlechtsidentität anbieten, z.B. nach dem Konzept der psychologischen Androgynität. Da die Geschlechtsrollenorientierung über mehr Aussagekraft als das biologische Geschlecht verfügt, ergeben sich möglicherweise ganz neue Zusammenhänge zwischen dem Geschlecht (*gender*) und der Leistung, auch in Hinblick auf die Lehrer- und Elternvariablen. Interessant wäre der Leistungsvergleich der Androgynen mit den Femininen und Maskulinen – je nach Leistungsdomäne.

Fokussierung auf das allgemeine Leistungsniveau

Nach der Analyse auf latenter Ebene sind die geschlechtsspezifischen Unterschiede in den kognitiven Teilkompetenzen für sechs von neun Inhaltsbereichen zu vernachlässigen. Ein im Mixed Rasch-Modell identifiziertes latentes Merkmal hat sich über die Inhalte hinweg als bedeutsames Merkmal herausgestellt, das stärker leistungs- und weniger geschlechterkorreliert ist. Es konnte also weder eine valide Geschlechtsskala ermittelt werden, noch entsprechen die Klassen einem verbal-bewertenden oder räumlich-grafisch-abstrakten Profil.

Obwohl der nationale Naturwissenschaftstest intensiv auf Geschlechterunterschiede untersucht wurde, konnten diese nicht nachgewiesen werden, selbst mit verschiedenen methodischen Ansätzen konnten die Geschlechterdifferenzen nicht vergrößert werden. Hätte das (biologische) Geschlecht einen starken Einfluss auf die Daten, hätte sich das in den Analysen zeigen müssen. Dies lässt den Schluss zu, dass die Geschlechterunterschiede bei PISA, insbesondere im nationalen Naturwissenschaftstest nur von untergeordneter Bedeutung sind. Sie sind weniger wichtig als die qualitativen Leistungsunterschiede.

Daraus ergibt sich, dass die Teilkompetenzen möglicherweise im pädagogischen Zusammenhang mit dem Geschlecht nicht so stark assoziiert sind wie bei Rost et al. (2004) angenommen. Es erscheint daher sinnvoller, eher auf die Förderung des allgemeinen Leistungsniveaus anstelle auf Geschlechtsunterschiede zu fokussieren. Die Implikation für Schulleistungsstudien besteht also darin, nicht primär nach geschlechterspezifischen Klassen zu suchen, sondern zunächst Leistungsklassen genauer zu betrachten.

Um solche Zusammenhänge auch für die Antwortmuster in den Domänen Lesen, Mathematik und Problemlösen sowie im internationalen Naturwissenschaftstest zu replizieren, sollten bei PISA relevante Leistungsklassen auf Basis der Antwortmuster gebildet werden. Der Einsatz

latenter Testmodelle sollte bei PISA und in anderen Schulleistungsstudien ermöglichen, relevante Leistungsklassen näher, stärker qualitativ zu untersuchen.

Es konnte in der vorliegenden Arbeit gezeigt werden, dass die Population leistungsstarker Schüler bei Aufgaben, die Praxisbezug oder Alltagswissen erfordern, deutlich in ihren Leistungen absinkt, während die eher leistungsschwachen Schüler gerade von solchen Aufgaben profitieren und dort ihre relativen Stärken zeigen. Es bestehen zwischen verschiedenen Leistungsklassen demnach nicht nur quantitative, sondern auch erhebliche qualitative Unterschiede.
Ziel ist es, solche qualitativen Unterschiede – auch außerhalb von PISA – genauer zu erfassen und Implikationen für unterschiedliche Förderungsmaßnahmen im Unterricht zu entwickeln, denn dieser Aspekt von Heterogenität ist im Unterricht nur schwer zu berücksichtigen. Da es sich in Schulklassen nur selten um leistungshomogene Gruppen handelt, muss dennoch ein Weg gefunden werden, mit diesem qualitativen Unterschied umzugehen. Eine Möglichkeit besteht darin, Sachverhalte mit unterschiedlich starkem Praxisbezug im Unterricht zu bearbeiten, so dass jede Leistungsgruppe berücksichtigt wird. Eine andere Möglichkeit ist die Förderung leistungsstarker Schüler in praxisnahen Problemstellungen, also umgekehrt der vermehrte Einsatz solcher Inhalte. Für die eher leistungsschwachen Schüler gilt es, die Stärken im Unterricht zu nutzbar zu machen und so auch das allgemeine Leistungsniveau zu heben. Durch eine Vermittlung alltagsnaher Problemfelder können so generelle Fähigkeiten geschult werden.

Bei der Vorbereitung von Maßnahmen zur Verbesserung von Schule und Unterricht scheint von zentraler Bedeutung, genau zu identifizieren, auf welche Art von Heterogenität in Schülergruppen das Hauptaugenmerk gerichtet werden muss. Unter der Vielzahl von Aspekten verdient das Geschlecht offenkundig weniger Aufmerksamkeit als das allgemeine Leistungsniveau, denn (psychologische) Unterschiede zwischen Jungen und Mädchen sind klein - wenn auch groß in den Köpfen von Menschen.

5 Literatur

Adams, R., & Carstensen, C. H. (2002). Scaling outcomes. In R. Adams & M. Wu (Hrsg.), *PISA 2000 Technical Report* (S. 149-162). Paris: OECD.

Alfermann, D., Reigber, D., & Turan, J. (1999). Androgynie, soziale Einstellungen und psychische Gesundheit: Zwei Untersuchungen an Frauen im Zeitvergleich. In U. Bock & D. Alfermann (Hrsg.), *Androgynie: Vielfalt der Möglichkeiten* (Vol. 4, S. 142-155). Stuttgart: Metzler.

Allan, K., & Coltrane, S. (1996). Gender display in television commercials: A comparative study of television commercials in the 1950s and 1980s. *Sex Roles, 35*, 185-203.

Baker, F. B., & Kim, S.-H. (2004). *Item Response Theory Parameter Estimation Techniques* (Vol. Second Edition). New York: Marcel Dekker.

Baker, M. A. (1987). Sensory functioning. In M. A. Baker (Hrsg.), *Sex differences in human performance* (S. 5-36). New York: Wiley.

Baumert, J., Bos, W., & Lehmann, R. H. (2000). *Dritte Internationale Mathematik- und Naturwissenschaftsstudie: Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn*. Opladen: Leske + Budrich.

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

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*(2), 155-162.

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F., & Carstensen, C. H. (2004a). Mathematische Kompetenz. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 47-92). Münster: Waxmann.

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F., & Carstensen, C. H. (2004b). Mathematische Kompetenz. In P. M. & al. (Hrsg.), *PISA 2003: Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 47-92). Münster: Waxmann.

Bock, U., & Alfermann, D. (1999). *Androgynie. Vielfalt der Möglichkeiten* (Vol. 4). Stuttgart: Metzler.

Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Walther, G., & Valtin, R. (Hrsg.). (2003a). *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich*. Münster: Waxmann.

Bos, W., Lankes, E.-M., Schwippert, K., Valtin, R., Voss, A., Badel, I., & Plaßmeier, N. (2003b). Lesekompetenzen deutscher Grundschülerinnen und Grundschüler am Ende der vierten Jahrgangsstufe im internationalen Vergleich. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, G. Walther, & R. Valtin (Hrsg.), *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich* (S. 69-134). Münster: Waxmann.

Breitenbach, E. (1994). Geschlechtsspezifische Interaktion in der Schule. *Die Deutsche Schule, 86*, 170-191.

Burba, D., & Rost, J. (2006). Mädchen und Jungen – unterschiedliche Fertigkeiten trotz gleicher Fähigkeiten? Ergebnisse aus PISA 2003. *in Vorbereitung*.

Buschmann, M. (1994). Jungen und Koedukation. Zur Polarisierung der Geschlechterrollen. *Die Deutsche Schule, 86*, 192-214.

Carstensen, C. H., Knoll, S., Rost, J., & Prenzel, M. (2004). Technische Grundlagen. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 371-387). Münster: Waxmann.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Vol. 2nd). Hillsdale, NJ: Lawrence Earlbaum Associates.

Deaux, K. (1984). From individual differences to social categories: Analysis of a decade's research on gender. *American Psychologist, 39*, 105-116.

DePascale, C. A. (2005). *The Ideal Role of Large-Scale testing in a Comprehensive Assessment System*: www.testpublishers.org/Documents/Large_Scale_Assessment_v3.0.pdf.

Draba, R. E. (1977). *The identification and interpretatation of item bias*. Chicago: University of Chicago (Research Memorandum 25).

Faulstich-Wieland, H., & Horstkemper, M. (1992). "Ohne Jungs fehlt der Klasse der Pep!" Koedukation aus der Sicht von Schülerinnen und Schülern. *Die Deutsche Schule, 84*, 348-360.

Fennema, E. (1996). Mathematics, Gender and Research. In G. Hanna (Hrsg.), *Towards Gender Equity in Mathematics Education* (S. 9-26). Dordrecht: Kluwer.

Fiedler, K. (1996). Die Verarbeitung sozialer Informationen für Urteilsbildung und Entscheidungen. In W. Stroebe, M. Hewstone, & G. M. Stephenson (Hrsg.), *Sozialpsychologie*. Berlin: Springer.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models - Foundations, recent developments, and applications*. New York: Springer.

Frasch, H., & Wagner, A. C. (1982). "Auf Jungen achtet man einfach mehr ..." Eine empirische Untersuchung zu geschlechtsspezifischen Unterschieden im Lehrer/innenverhalten gegenüber Jungen und Mädchen in der Grundschule. In I. Brehmer (Hrsg.), *Sexismus in der Schule* (S. 260-278). Weinheim: Beltz.

Gage, N. L., & Berliner, D. C. (1996). *Pädagogische Psychologie*. Weinheim: Beltz.

Georg, W., & Bargel, T. (2001). *Das Studium der Geisteswissenschaften. Eine Fachmonographie aus studentischer Sicht*. Bonn: BMBF.

Gittler, G., & Vitouch, O. (1994). Empirical contribution to the question of sex-dependent inheritance of spatial ability. *Perceptual and motor skills, 78*, 407-417.

Halpern, D. (2000). *Sex differences in cognitive abilities* (3rd edition. Aufl.). Mahwah, NJ: Erlbaum.

Hanna, G. (2003). Reaching Gender Equity in Mathematics Education. *The Educational Forum, 67 (3)*, 204-214.

Hannover, B. (1992). Spontanes Selbstkonzept und Pubertät. Zur Interessenentwicklung von Mädchen koedukativer und geschlechtshomogener Schulklassen. *Bildung und Erziehung, 45*, 31-46.

Hannover, B. (1999). Androgynie: Die Kontextabhängigkeit der Geschlechtsrollenidentität. In U. Bock & D. Alfermann (Hrsg.), *Androgynie: Vielfalt der Möglichkeiten* (Vol. 4, S. 131-141). Stuttgart: Metzler.

Hannover, B., & Kessels, U. (2001). Monoedukativer Anfangsunterricht in Physik in der Gesamtschule. Auswirkungen auf Motivation, Selbstkonzept und Einteilung in Grund- oder Fortgeschrittenenkurse. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 33*(4), 201-215.

Hays, W. L. (1994). *Statistics* (5. Aufl.). Fort Worth: Harcourt Brace College Publishers.

Heinz, M., Kranz, M., & Kuster, F. (2004). Stichwort "weiblich/männlich". In J. Ritter, K. Gründer, & G. Gabriel (Hrsg.), *Historisches Wörterbuch der Philosophie* (Vol. Band 12: W-Z). Basel: Schwabe AG Verlag.

Heinz, W. (2002). Frauenkriminalität. *Bewährungshilfe*, 131-152.

Herrmann, B., Grupe, G., Hummel, S., Piepenbrink, H., & Schutkowski, H. (1990). *Prähistorische Anthropologie. Leitfaden der Feld- und Labormethoden*. Berlin: Springer.

Herwartz-Emden, L., Schurt, V., & Waburg, W. (2005). Mädchenschulen zwischen Traditionalismus und Emanzipationsanspruch. Forschungsstand und Forschungsdesiderata. *Zeitschrift für Pädagogik, 51*(3), 342-362.

Hoffmann, L. (2002). Promoting girls' interest and achievement in physics classes for beginners. *Learning and Instruction, 12*, 447-465.

Hoffmann, L., Häußler, P., & Peters-Haft, S. (1997). *An den Interessen von Mädchen und Jungen orientierter Physikunterricht. Ergebnisse eines BLK-Modellversuchs* (Vol. 155). Kiel: Institut für die Pädagogik der Naturwissenschaften (IPN).

Hofstätter, P. R., & Wendt, D. (1974). *Quantitative Methoden der Psychologie*. Frankfurt/Main: Barth.

Horstkemper, M. (2002). Bildungsforschung aus der Sicht pädagogischer Frauen- und Geschlechterforschung. In R. Tippelt (Hrsg.), *Handbuch Bildungsforschung* (S. 409-423). Opladen: Leske + Budrich.

Horstkemper, M. (2004). Trennen gegen Typisierung? - Eine Gegenrede. *Friedrich Jahresheft, 12,* 94.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender Differences in Mathematics Performance: A Meta-Analysis. *Psychological Bulletin, 107*(2), 139-155.

Hyde, J. S., & Linn, M. C. (1988). Gender Differences in Verbal Ability: A Meta-Analysis. *Psychological Bulletin, Vol. 104*(1), 53-69.

Intons-Peterson, M. G. (1988). *Gender concepts of Swedish and American Youth*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson, G. B., & Raven, P. H. (1996). *Biology: Fourth edition*. Boston, MA: The McGraw-Hill Companies, Inc.

Johnson, S. (1996). The contribution of large-sclae assessment programmes to research on gender differences. *Educational Research and Evaluation, 2*(1), 25-49.

Kessels, U. (2002). *Undoing Gender in der Schule. Eine empirische Studie über Koedukation und Geschlechtsidentität im Physikunterricht*. Weinheim, München: Juventa.

Kessels, U. (2004). Mädchenfächer - Jungenfächer? Geschlechtertrennung im Unterricht. *Friedrich Jahresheft, 12,* 90-94.

Kienzl, A. (2005). *Überprüfung der Schulnotenverteilungen im Rahmen von PISA 2003 (Diplomarbeit)*. Universität Kiel: Psychologisches Institut.

Klotz, T. (1998). *Der frühe Tod des starken Geschlechts. Forum Männergesundheit*. Göttingen: Cuvillier Verlag.

Knoche, N., & Lind, D. (2004). Eine differenzielle Itemanalyse zu den Faktoren Bildungsgang und Geschlecht. In M. Neubrand (Hrsg.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland. Vertiefende Analysen im Rahmen von PISA 2000* (S. 73-86). Wiesbaden: VS Verlag.

Kolb, B., & Wishaw, I. Q. (1993). *Neuropsychologie*. Heidelberg: Spektrum Akademischer Verlag.

Köller, O., & Klieme, E. (2000). Geschlechtsdifferenzen in den mathematisch-naturwissenschaftlichen Leistungen. In J. Baumert, W. Bos, & R. H. Lehmann (Hrsg.), *Dritte Internationale Mathematik- und Naturwissenschaftsstudie: Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* (Vol. II: Mathematische und physikalische Kompetenzen am Ende der Schullaufbahn, S. 373-404). Opladen: Leske + Budrich.

Krauth, J. (1995). *Testkonstruktion und Testtheorie*. Weinheim: Psychologie Verlags Union.

Krell, G. (1999). Androgynie, Management, Personalpolitik: Androgyne Führungskräfte oder/ und Organisationen als Erfolgsfaktor? In U. Bock & D. Alfermann (Hrsg.), *Androgynie: Vielfalt der Möglichkeiten* (Vol. 4, S. 173-181). Stuttgart: Stuttgart.

Lehmann, R. H. (1994). Lesen Mädchen wirklich besser? Ergebnisse aus der internationalen IEA-Lesestudie. In S. Richter & H. Brügelmann (Hrsg.), *Mädchen lernen ANDERS lernen Jungen*. Bottighofen: Libelle.

Leutner, D., Klieme, E., Meyer, K., & Wirth, J. (2004). Problemlösen. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 147-175). Münster: Waxmann.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz.

Lind, G. (2003). *Moral ist lehrbar. Handbuch zur Theorie und Praxis moralischer und demokratischer Bildung*. München: Oldenbourg.

Litscher, G. (2004). Quantitative Bestimmung geschlechtsspezifischer thermischer Empfindungs- und Schmerzwellen vor und nach Lasernadelstimulation. *Biomedizinische Technik, 49*, 106-110.

Lobel, T. E., & Menashri, J. (1993). Relations of conceptions of gender-role transgressions and gender constancy to gender-typed toy preferences. *Developmental Psychology, 29*, 150-155.

Lobemeier, K. (2005). *Welche Leistungen erbringen Viertklässler bei Aufgaben zum Thema Größen? - Untersuchungen zur mathematisch-naturwissenschaftlichen Kompetenz im Grundschulalter im Rahmen von IGLU*. Christian-Albrechts-Unversität, Kiel.

Loring-Meier, S., & Halpern, D. F. (1999). Sex differences in visual-spatial working memory: Components of cognitive processing. *Psychonomic Bulletin & Review, 6*, 464-471.

Lount, R. B., Messé, J. L. A., & Kerr, N. L. (2000). Trying Harder for Different Reasons. Conjunctivity and Sex Composition as Bases for Motivation Gains in Performing Groups. *Zeitschrift für Sozialpsychologie, 31*(4), 221-230.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford: Stanford University Press.

Maier, P. H. (1999). Empirische Studien zu geschlechtsspezifischen Differenzen. In P. H. Maier (Hrsg.), *Räumliches Vorstellungsvermögen. Ein theoretischer Abriss des Phänomens räumliches Vorstellungsvermögen* (S. 169-209). Donauwörth: Auer.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 International Science Report*. Chestnut Hill: TIMSS & PIRLS International Study Center, Lynch School of education, Boston College.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S. J., Garden, R. A., & O'Connor, K. M. (2000). *TIMSS 1999 International Science Report.* Chestnut Hill: International Study Center, Lynch School of education, Boston College.

Meinz, E. J., & Salthouse, T. A. (1998). Is age kinder to females than to males? *Psychonomic Bulletin & Review, 5*, 56-70.

Mohr, W. (1987). *Frauen in der Wissenschaft. Ein Bericht zur sozialen Lage von Studentinnen und Wissenschaftlerinnen im Hochschulbereich*. Freiburg i.B.: Dreisam-Verlag.

Morrell, C. H., Gordon-Salant, S., Pearson, J. D., Brant, L. J., & Fozart, J. L. (1996). Age- and gender-specific reference ranges for hearing level and longitudinal changes in hearing level. *Journal for the Acoustical Society of America, 100*, 1949-1967.

Mullis, I. V. S., Martin, M. O., Fierros, E. G., Goldberg, A. L., & Stemler, S. E. (2000a). *Gender Differences in Achievement, IEA's Third International Mathematics and*

Science Study. Chestnut Hill: International Study Center, Lynch School of education, Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 International Mathematics Report*. Chestnut Hill: TIMSS & PIRLS International Study Center, Lynch School of education, Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J., & Smith, T. A. (2000b). *TIMSS 1999 International Mathematics Report*. Chestnut Hill: International Study Center, Lynch School of education, Boston College.

Musahl, H.-P., & Schwennen, C. (2001). Stichwort "Versuchsplanung". In G. Wenninger (Hrsg.), *Lexikon der Psychologie*. Heidelberg: Spektrum Akademischer Verlag.

Nicholson, K., & Kimura, D. (1996). Sex differences for speech and manual skill. *Perceptual and motor skills, 82*, 3-13.

Noble, K. D. (1987). The dilemma of the gifted woman. *Psychology of Women Quarterly, 5*, 89-140.

Nyborg, H. (1983). Spatial ability in men and women: Review and new theory. *Advances in behaviour research and therapy, 5*, 89-140.

OECD. (2000). *Measuring Student Knowledge and Skills: The PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy*. Paris: OECD.

OECD. (2003). *The PISA 2003 Assessment Framework - Mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.

OECD. (2004). *Lernen für die Welt von morgen. Erste Ergebnisse von PISA 2003*. Paris: OECD.

OECD. (2005). *PISA 2003 Technical Report*. Paris: OECD.

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

Pekrun, R., & Zirngibl, A. (2004). Schülermerkmale im Fach Mathematik. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 191-210). Münster: Waxmann.

Pinel, J. P. J. (2001). *Biopsychologie.* Heidelberg: Spektrum Akademischer Verlag.

Pinker, S. (2004). *Why nature & nurture won't go away*: http://pinker.wjh.harvard.edu/articles/papers/nature_nurture.pdf.

Plomin, R., DeFries, J. C., McClearn, G. E., & McGuffin, P. (2001). *Behavioral Genetics* (4. Aufl.). New York: Freeman.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U. (2004a). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs*. Münster: Waxmann.

Prenzel, M., Drechsel, B., Carstensen, C. H., & Ramm, G. (2004b). PISA 2003 - eine Einführung. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 13-46). Münster: Waxmann.

Prenzel, M., Geiser, H., Langeheine, R., & Lobemeier, K. (2003). Das naturwissenschaftliche Verständnis am Ende der Grundschule. In W. Bos, E.-M. Lankes, M. Prenzel, K.

Schwippert, G. Walther, & R. Valtin (Hrsg.), *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich* (S. 143-180). Münster: Waxmann.

Pschyrembel, W. (Hrsg.). (1994). *Pschyrembel Medizinisches Wörterbuch - Sonderausgabe* (257. Aufl.). Hamburg: Nikol.

Rebok, G. W. (1987). *Life-span cognitive development*. New York: Holt, Rinehart & Winston.

Richardson, J. T. E. (1991). Gender differnces in imagery, cognition and memory. In R. H. Logie & M. Denis (Hrsg.), *Mental images in human cognition* (S. 271-303). New York: Elsevier.

Richter, S. (1996). *Unterschiede in den Schulleistungen von Mädchen und Jungen. Geschlechterspezifische Aspekte des Schriftsprachenerwerbs und ihre Berücksichtigung im Unterricht* (Vol. 30). Regensburg: Roderer.

Robinson, N. M., Abbott, R. D., Berninger, V. W., & Busse, J. (1996). The structure of abilities in math-precoscious young children: gender similarities and differences. *Journal of Educational Psychology, 88*, 341-352.

Roffman, J. L., Marci, C. D., Glick, D. M., Dougherty, D. D., & Rauch, S. L. (2005). Neuroimaging and the functional neuroanatomy of psychotherapy. *Psychological Medicine, 35*(10), 1385-1398.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2., vollst. überarb. u. erw. Aufl.). Bern: Hans Huber.

Rost, J., Walter, O., Carstensen, C. H., Senkbeil, M., & Prenzel, M. (2004). Naturwissenschaftliche Kompetenz. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 111-146). Münster: Waxmann.

Schnitzer, K., Isserstedt, W., Müßig-Trapp, P., & Schreiber, J. (1998). *Das soziale Bild der Studentenschaft in der Bundesrepublik Deutschland. 15. Sozialerhebung des Deutschen Studentenwerks*. Bonn: BMBF.

Schuchmann, M. (1999). *Probabilistische Testtheorie*. München: Oldenbourg Wissenschaftsverlag.

Siegel, S. (1956). *Nonparametric Statistics For The Behavioral Sciences*. New York: McGraw-Hill.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 4-28.

Stanat, P., & Kunter, M. (2001). Geschlechterunterschiede in Basiskompetenzen. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann, & M. Weiß (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 249-269). Opladen: Leske + Budrich.

Stanat, P., & Kunter, M. (2003). Kompetenzerwerb, Bildungsbeteiligung und Schullaufbahn von Mädchen und Jungen im Ländervergleich. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann, & M. Weiß (Hrsg.), *PISA 2000 - Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 211-242). Opladen: Leske + Budrich.

Stewart, A. J., & McDermott, C. (2004). Gender in Psychology. *Annual Review of Psychology, 55*, 519-544.

Stewart, V. (1976). Social influences on sex differences in behavior. In M. S. Teitelbaum (Hrsg.), *Sex differences: Social and biological perspectives* (S. 138-174). New York: Anchor Books.

Stones, I., Beckmann, M., & Stephens, L. (1982). Sex-related differences in mathematical competencies of pre-calculus college students. *School science and mathematics, 82*, 295-299.

Strauss, B., & Möller, J. (1999). Androgynie: Typ oer Trait? Zur Struktur und Messung des psychologischen Geschlechts. In U. Bock & D. Alfermann (Hrsg.), *Androgynie: Vielfalt der Möglichkeiten* (Vol. 4, S. 200-209). Stuttgart: Metzler.

Subrahmanyam, K., & Greenfield, P. M. (1994). Effect of video game practice on spatial skills in girls and boys. *Journal of Applied Developmental Psychology, 15*, 13-32.

Trautner, H. M. (1994). Geschlechtsspezifische Erziehung und Sozialisation. In K. A. Schneewind (Hrsg.), *Psychologie der Erziehung und Sozialisation*. Göttingen: Hogrefe.

Vandenberg, G. (1987). Sex differences in mental retardation and their implications for sex differences in ability. In L. A. Reinisch, L. A. Rosenblum, & S. A. Sanders (Hrsg.), *Masculinity/femininity: Basic perspectives*. New York: Oxford University Press.

Voland, E. (1993). *Grundriß der Soziobiologie*. Stuttgart, Jena: G. Fischer.

Walter, O. (2005). *Kompetenzmessung in den PISA-Studien. Simulationen zur Schätzung von Verteilungsparametern und Reliabilitäten.* Lengerich: Pabst Science Publishers.

Walther, G., Geiser, H., Langeheine, R., & Lobemeier, K. (2003). Mathematische Kompetenzen am Ende der vierten Jahrgangsstufe. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, G. Walther, & R. Valtin (Hrsg.), *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich* (S. 189-222). Münster: Waxmann.

Weinert, F. E., & Helmke, A. (1997). *Entwicklung im Grundschulalter*. Weinheim: Psychologie Verlags Union.

Weniger, G.-C. (2003). *Projekt Menschwerdung. Streifzüge durch die Geschichte der Menschheit*. Heidelberg/Berlin: Spektrum der Wissenschaft.

Zimmer, K., Burba, D., & Rost, J. (2004). Kompetenzen von Jungen und Mädchen. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 211-223). Münster: Waxmann.

6 Anhang

6.1 Quellen freigegebener und publizierter Aufgaben

Bezugnehmend auf die Aufgaben aus dem Differential Item functioning findet sich unten stehend eine Liste von geschlechtsspezifischen Items, die freigegeben *und* publiziert oder veröffentlicht sind, ggf. mit Quellenangabe.

Tabelle 35 Liste von *freigegebenen* Items, die im DIF als geschlechtsspezifisch interpretiert worden sind, mit Quellenangabe, sofern sie publiziert worden sind.

Itemlabel	Inhaltsdomäne	national/ international	leichter gelöst von	Ort der Veröffentlichung
M510Q01	**Mathematik**	international	Mädchen	bisher in deutschen Quellen unveröffentlicht
S128Q03*	**Naturwissenschaften**	international	Mädchen	Prenzel et al. (2004a) http://pisa.ipn.uni-kiel.de/
X402Q01*	**Problemlösen**	international	Mädchen	Prenzel et al. (2004a) http://pisa.ipn.uni-kiel.de/
MTISC2*	**Mathematik**	national	Jungen	Blum et al. (2004b)
MBRUC1*	**Mathematik**	national	Mädchen	Blum et al. (2004b)
NAUTOK*	**Naturwissenschaften**	national	Jungen	Prenzel et al. (2004a)
NAUTOM*	**Naturwissenschaften**	national	Jungen	Prenzel et al. (2004a)

6.2 Grafiken zum Differential Item Functioning nach Geschlecht im internationalen Vergleich

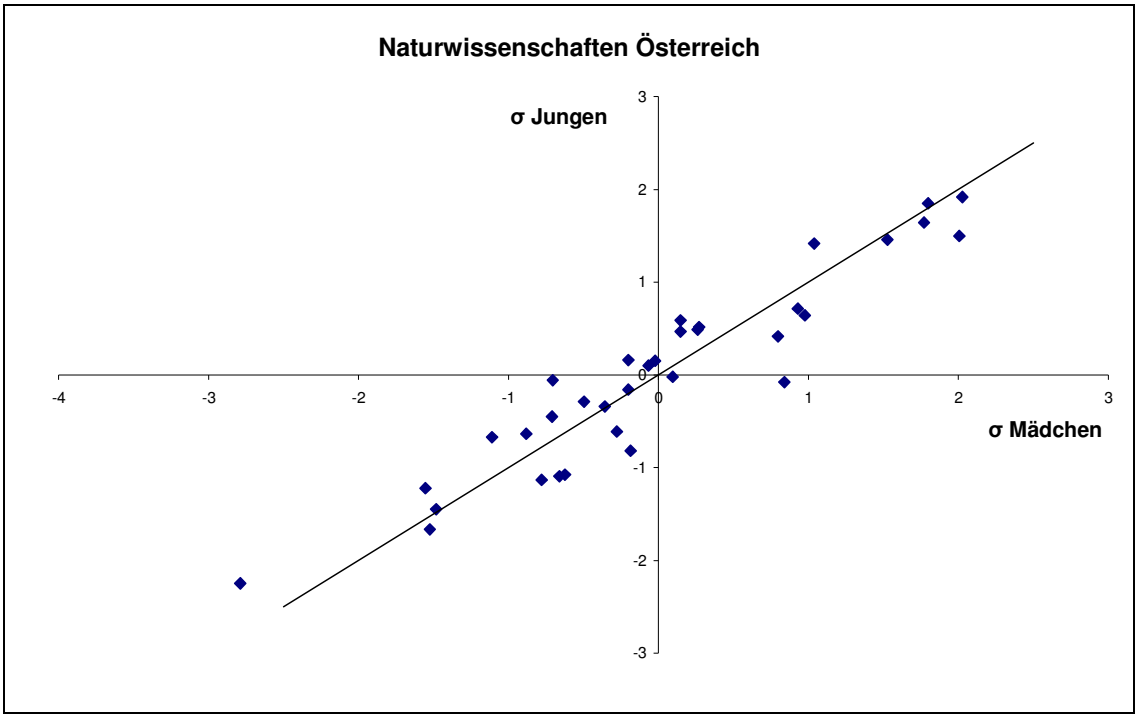

Abbildung 30 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Naturwissenschaftstests. (Stichprobe: österreichische Schüler)

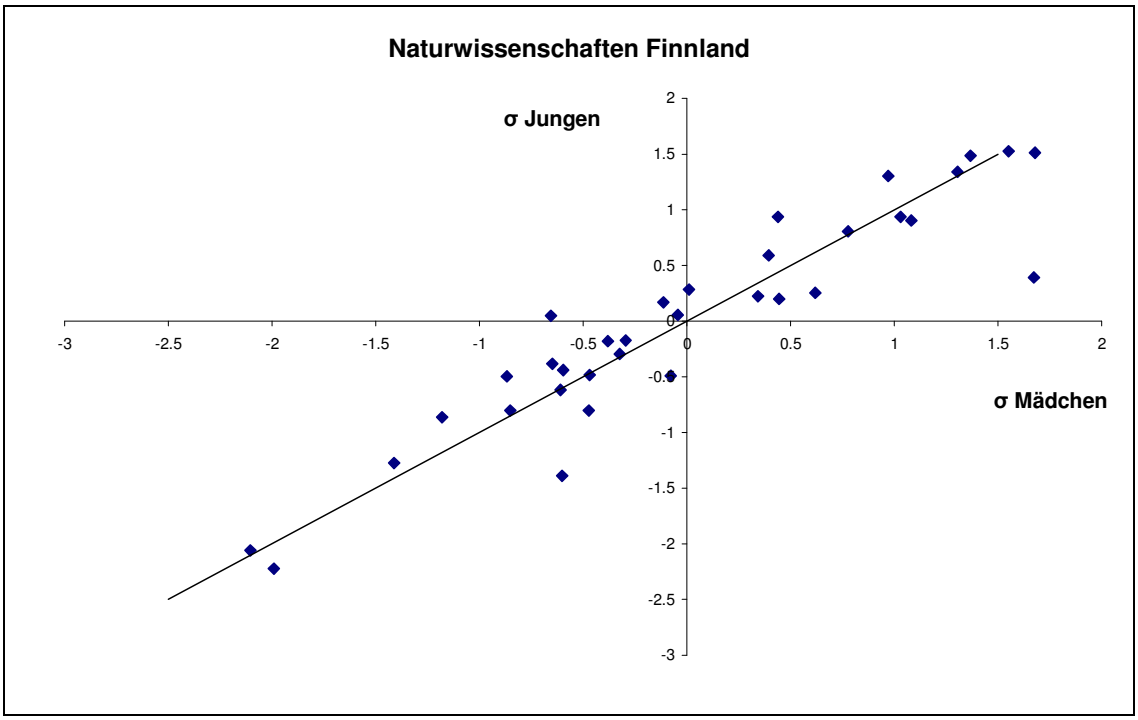

Abbildung 31 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Naturwissenschaftstests. (Stichprobe: finnische Schüler)

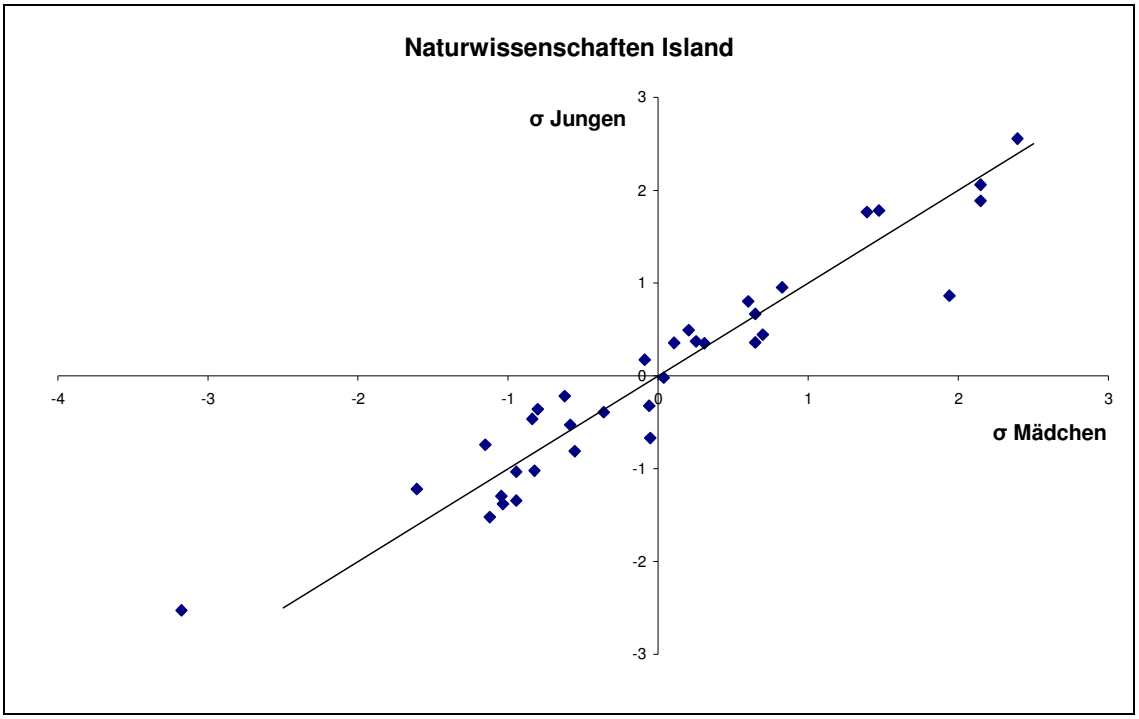

Abbildung 32 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Naturwissenschaftstests. (Stichprobe: isländische Schüler)

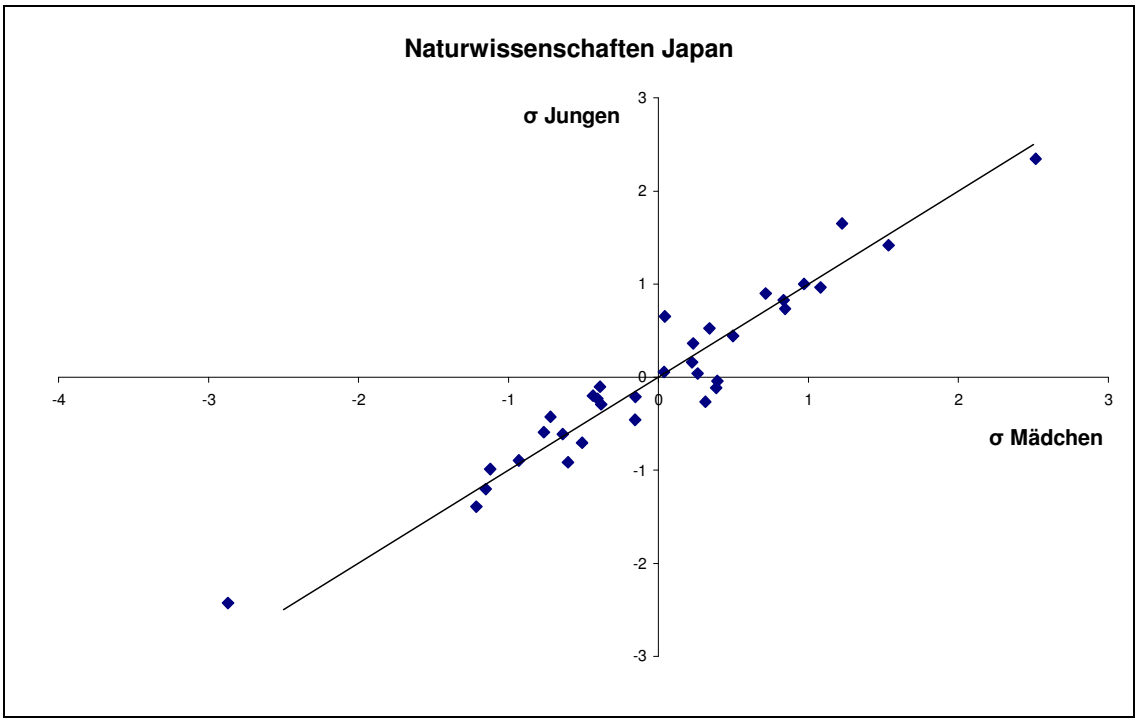

Abbildung 33 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Naturwissenschaftstests. (Stichprobe: japanische Schüler)

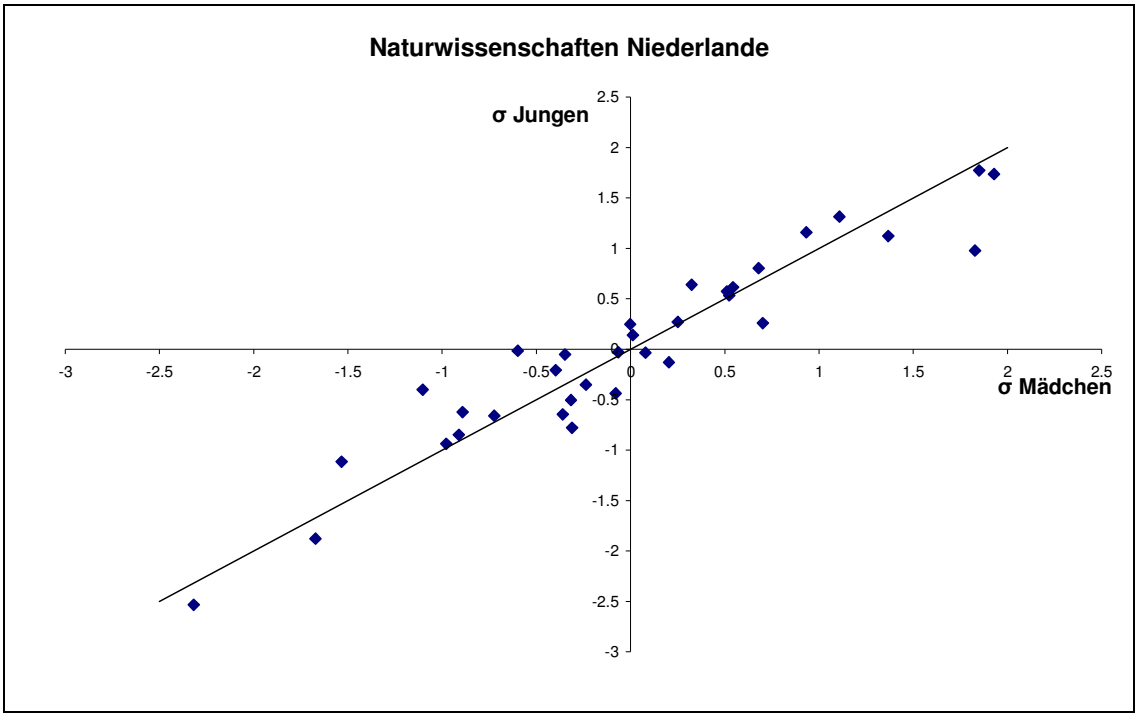

Abbildung 34 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Naturwissenschaftstests. (Stichprobe: niederländische Schüler)

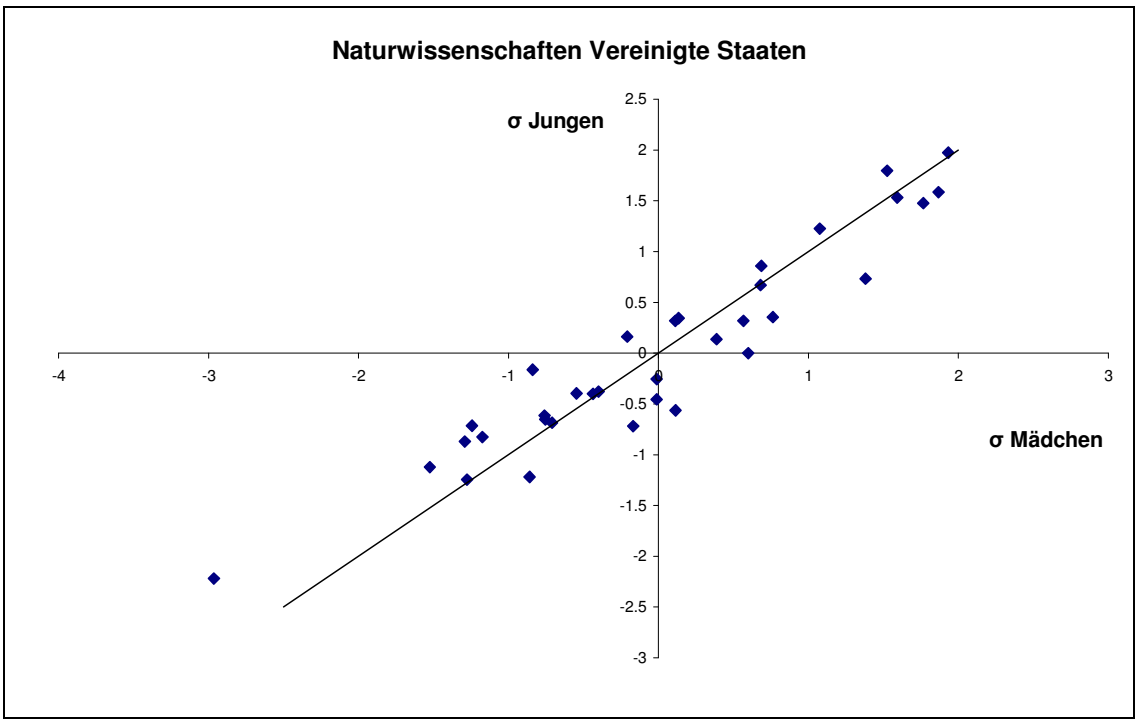

Abbildung 35 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Naturwissenschaftstests. (Stichprobe: Schüler aus den Vereinigten Staaten)

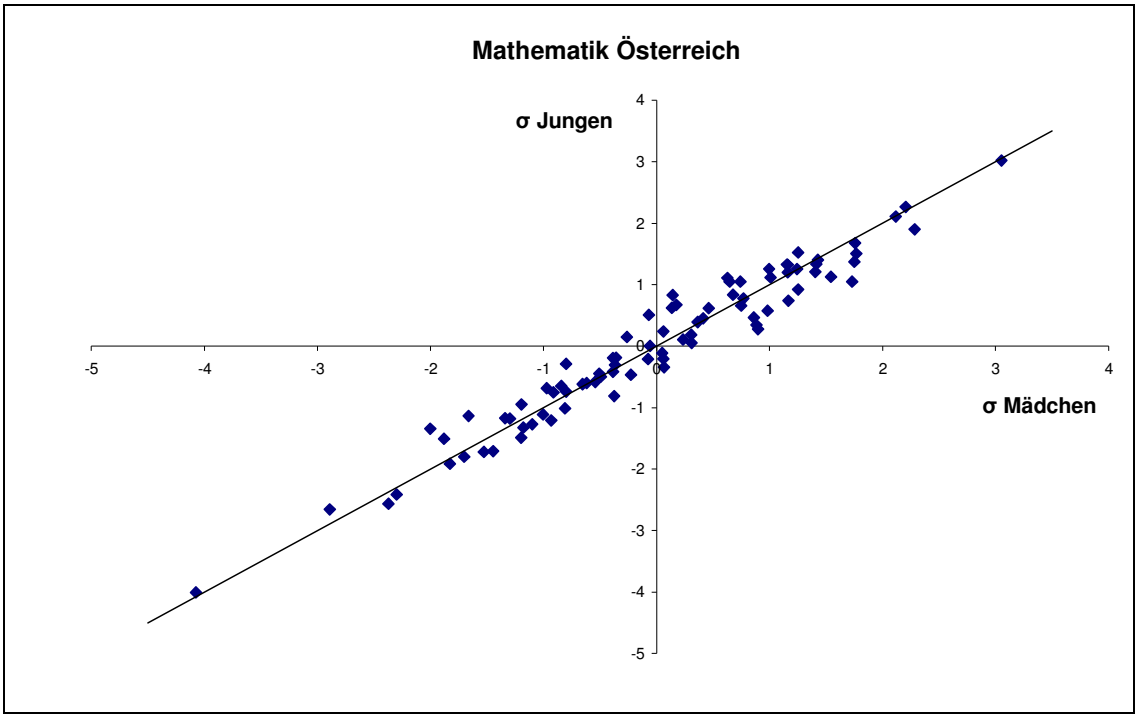

Abbildung 36 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Mathematiktests. (Stichprobe: österreichische Schüler)

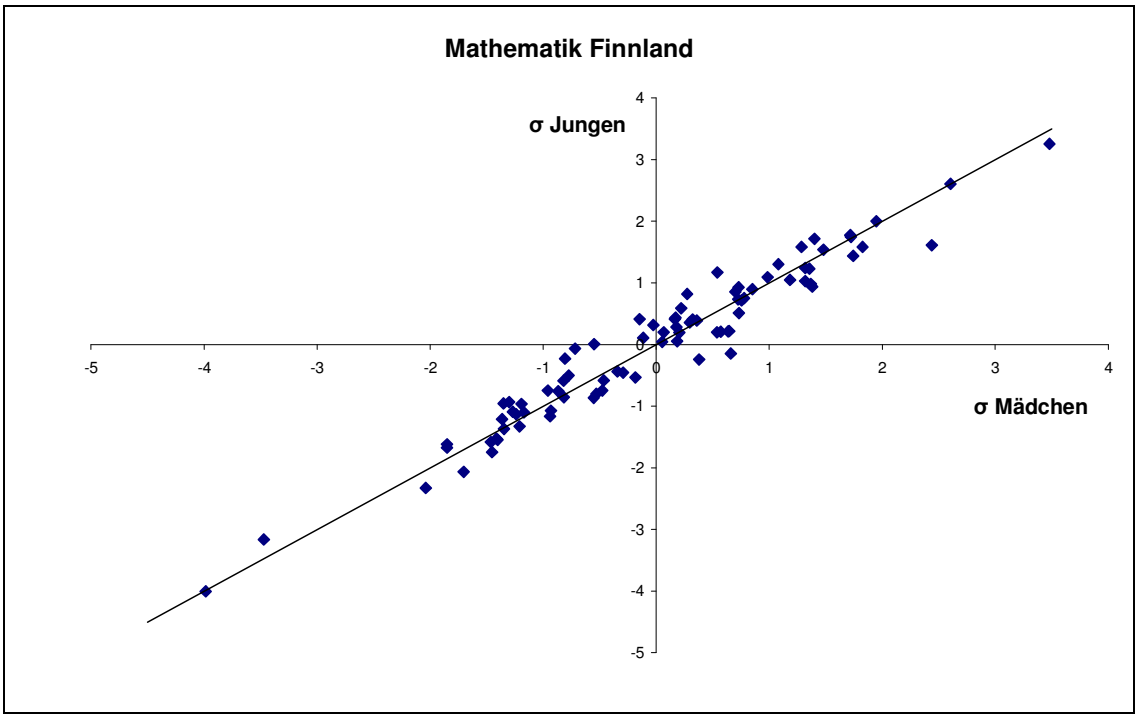

Abbildung 37 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Mathematiktests. (Stichprobe: finnische Schüler)

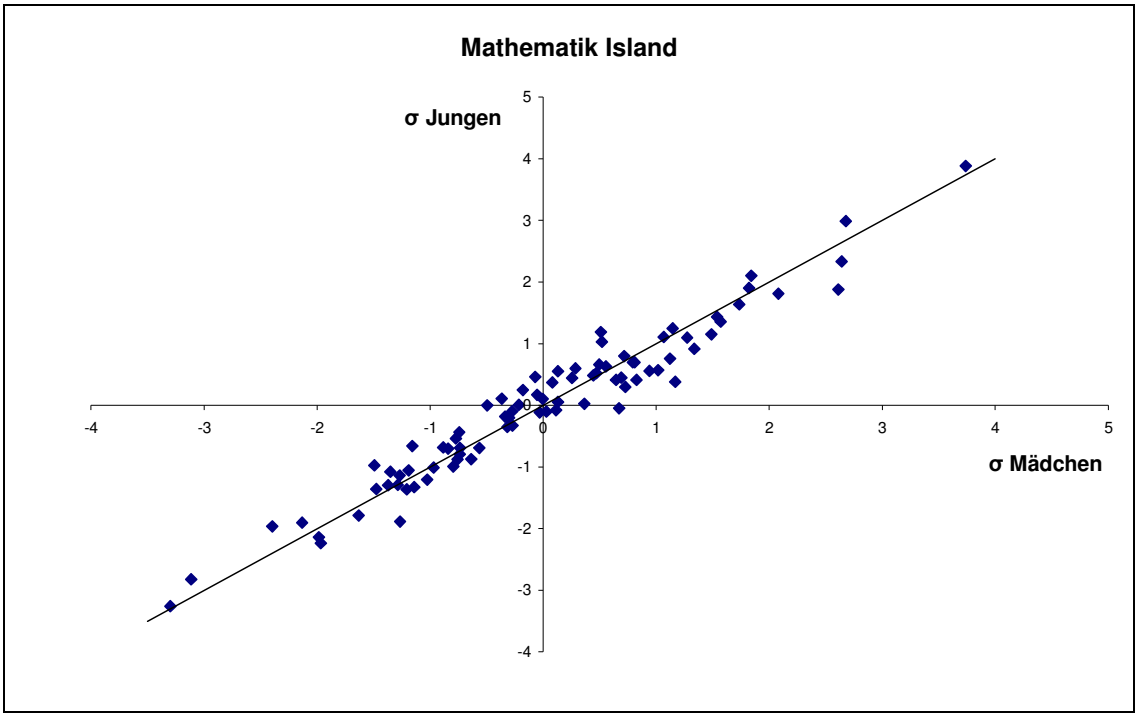

Abbildung 38 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Mathematiktests. (Stichprobe: isländische Schüler)

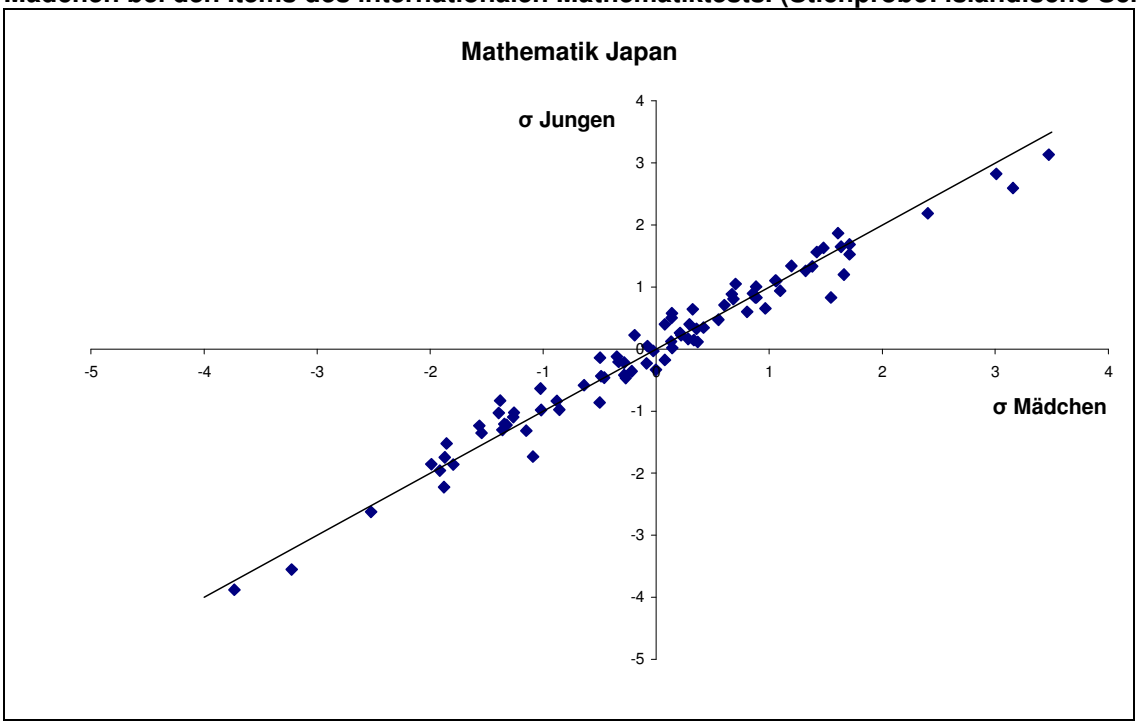

Abbildung 39 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Mathematiktests. (Stichprobe: japanische Schüler)

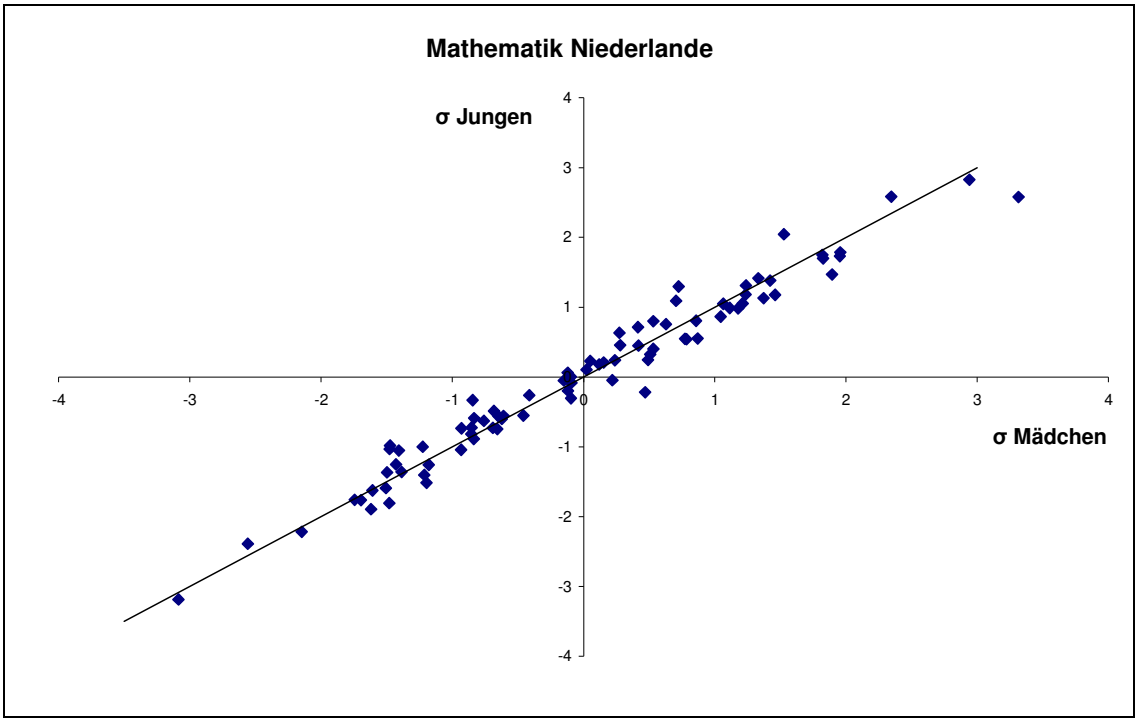

Abbildung 40 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Mathematiktests. (Stichprobe: niederländische Schüler)

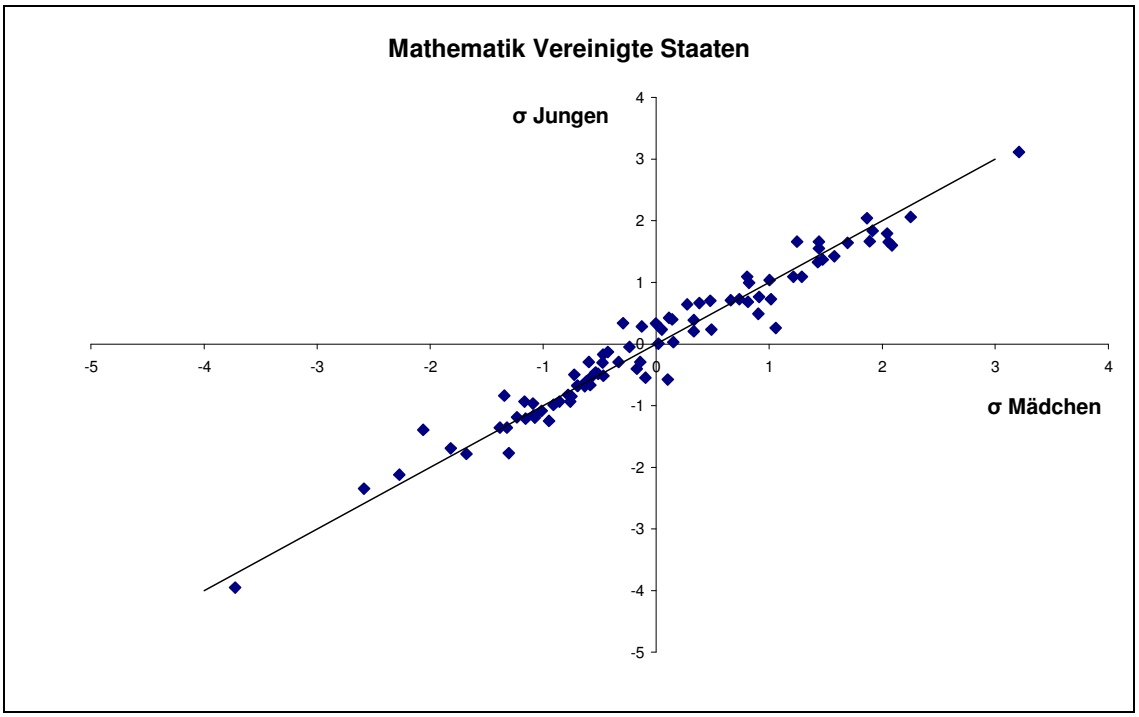

Abbildung 41 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Mathematiktests. (Stichprobe: Schüler aus den Vereinigten Staaten)

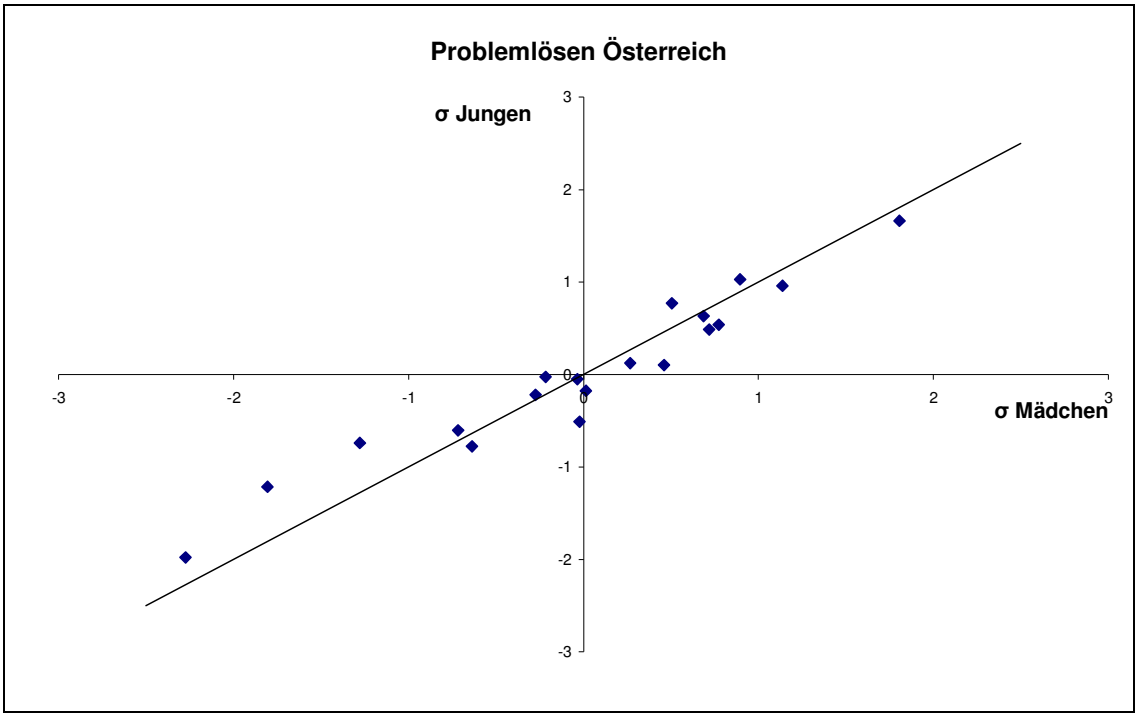

Abbildung 42 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Problemlösetests. (Stichprobe: österreichische Schüler)

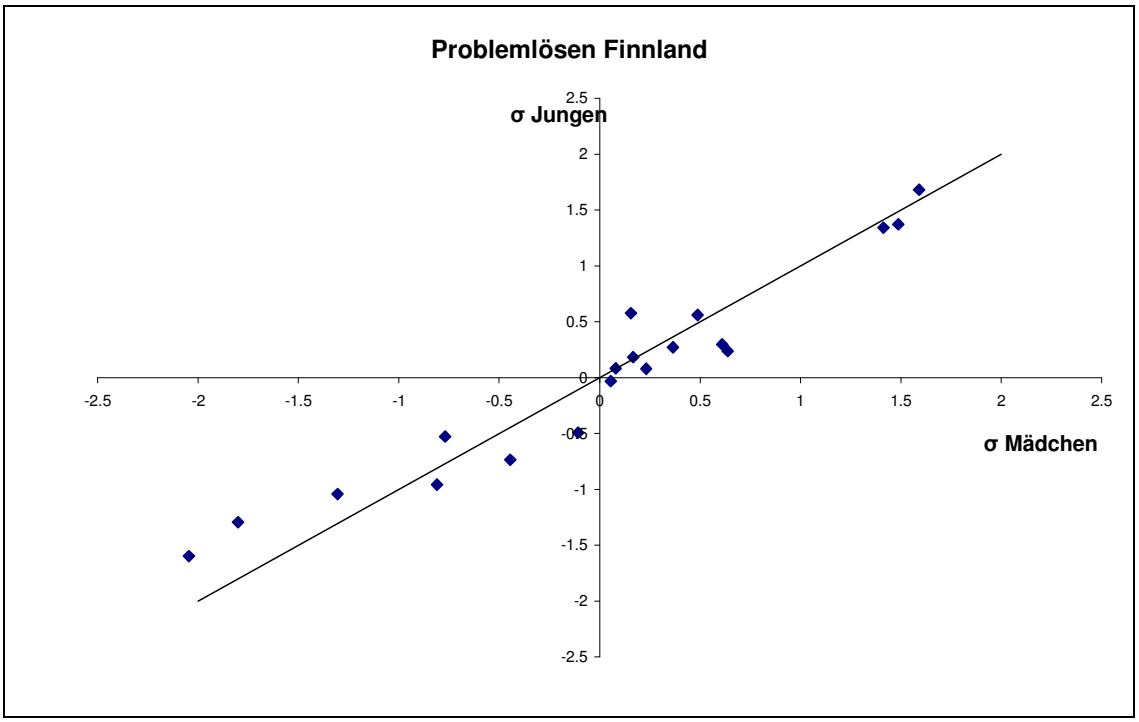

Abbildung 43 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Problemlösetests. (Stichprobe: finnische Schüler)

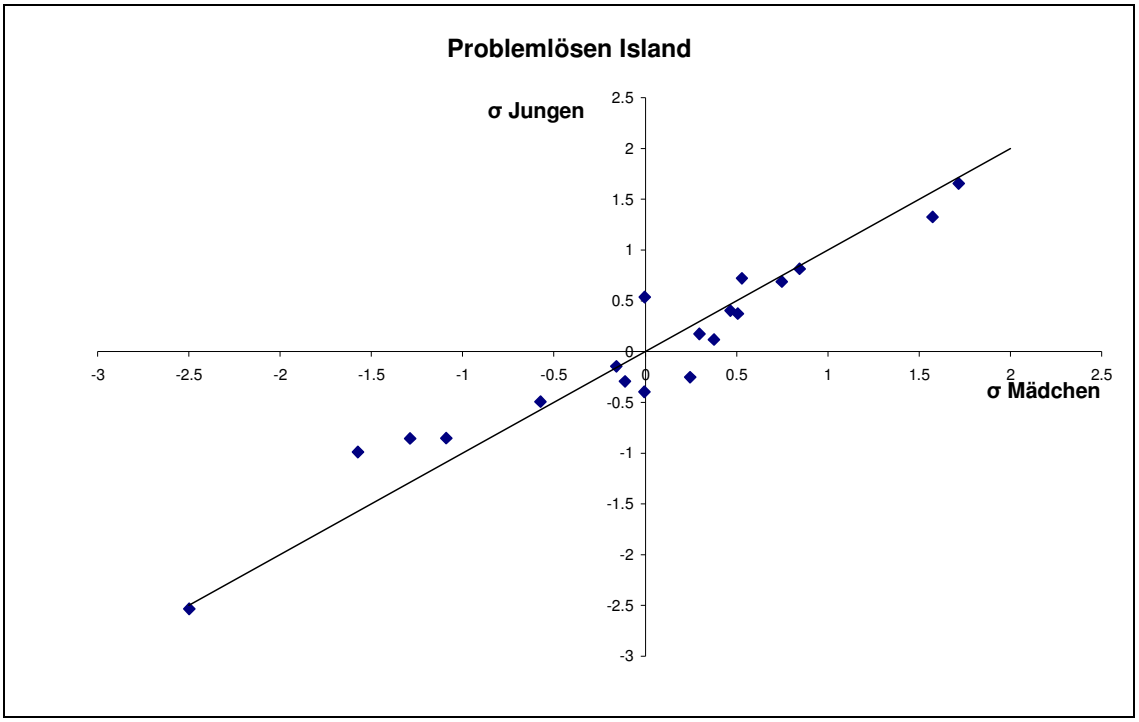

Abbildung 44 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Problemlösetests. (Stichprobe: isländische Schüler)

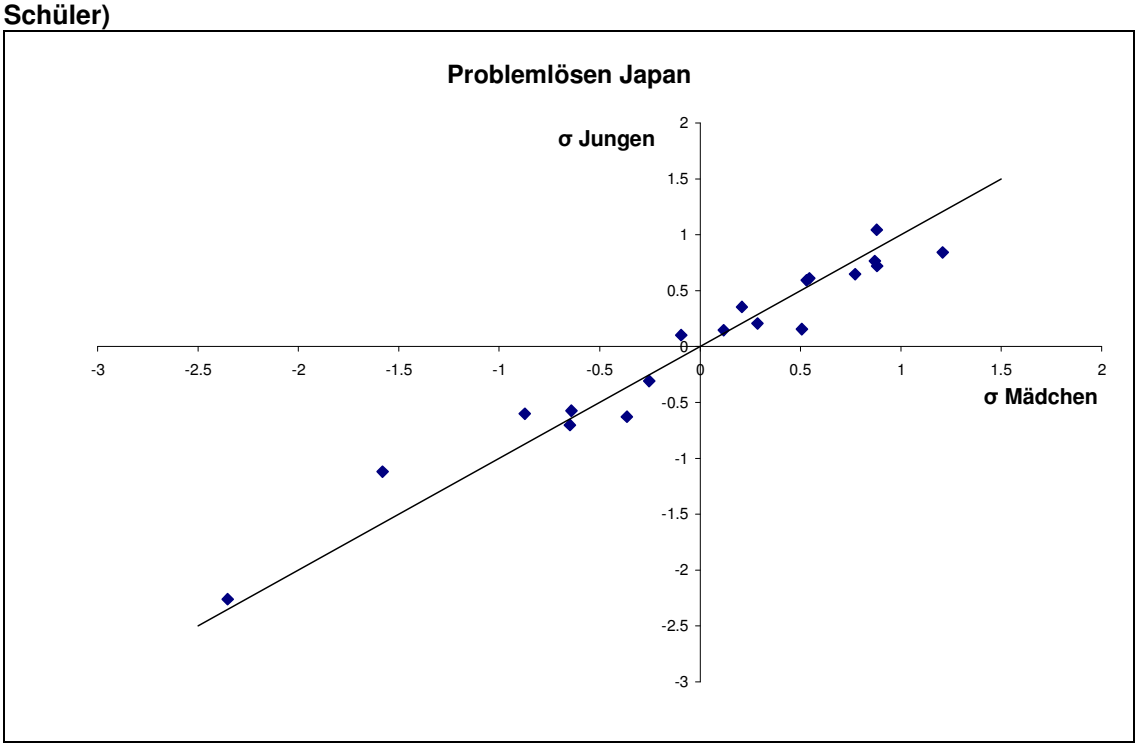

Abbildung 45 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Problemlösetests. (Stichprobe: japanische Schüler)

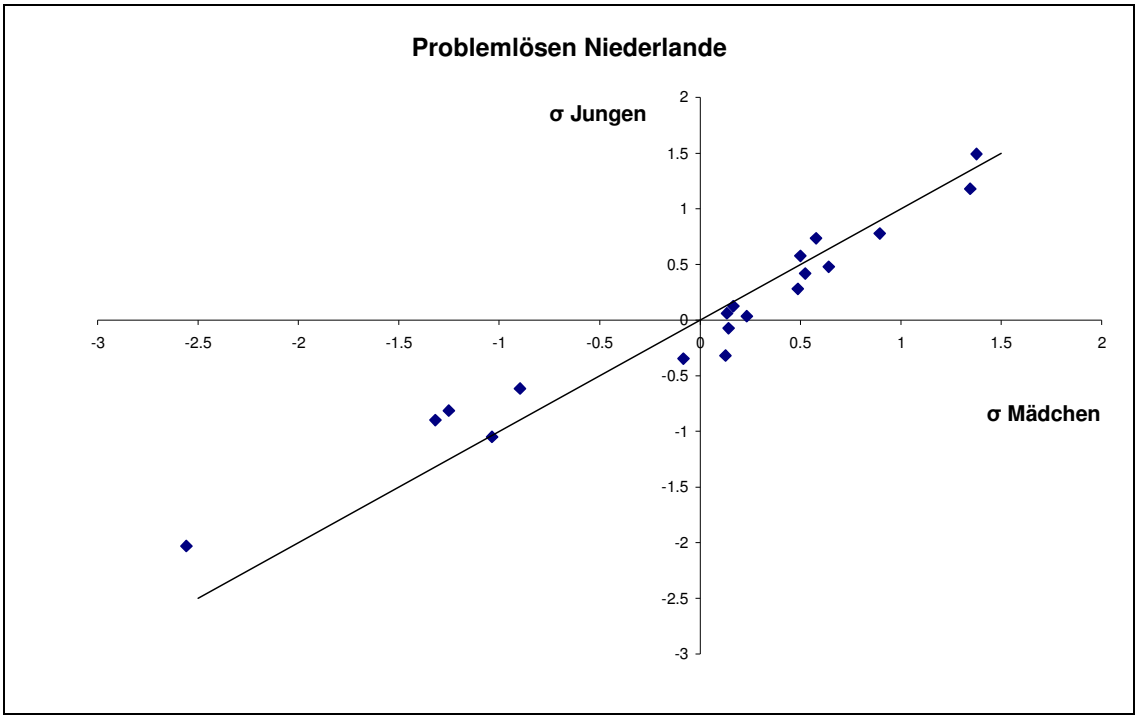

Abbildung 46 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Problemlösetests. (Stichprobe: niederländische Schüler)

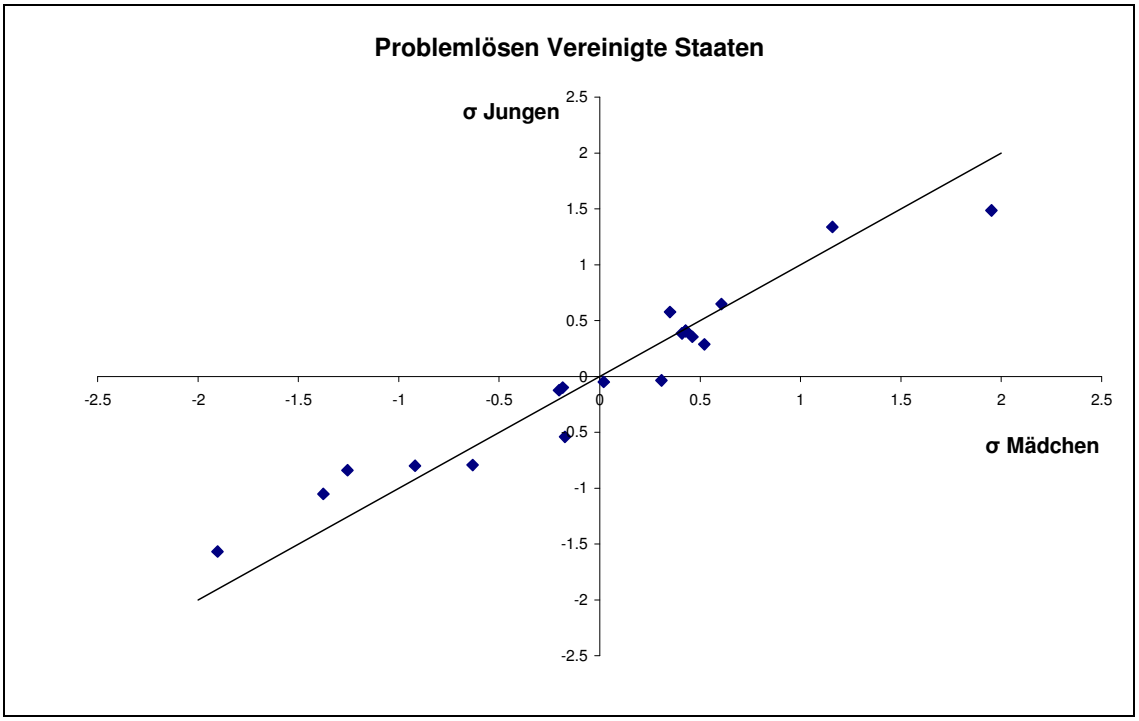

Abbildung 47 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Problemlösetests. (Stichprobe: Schüler aus den Vereinigten Staaten)

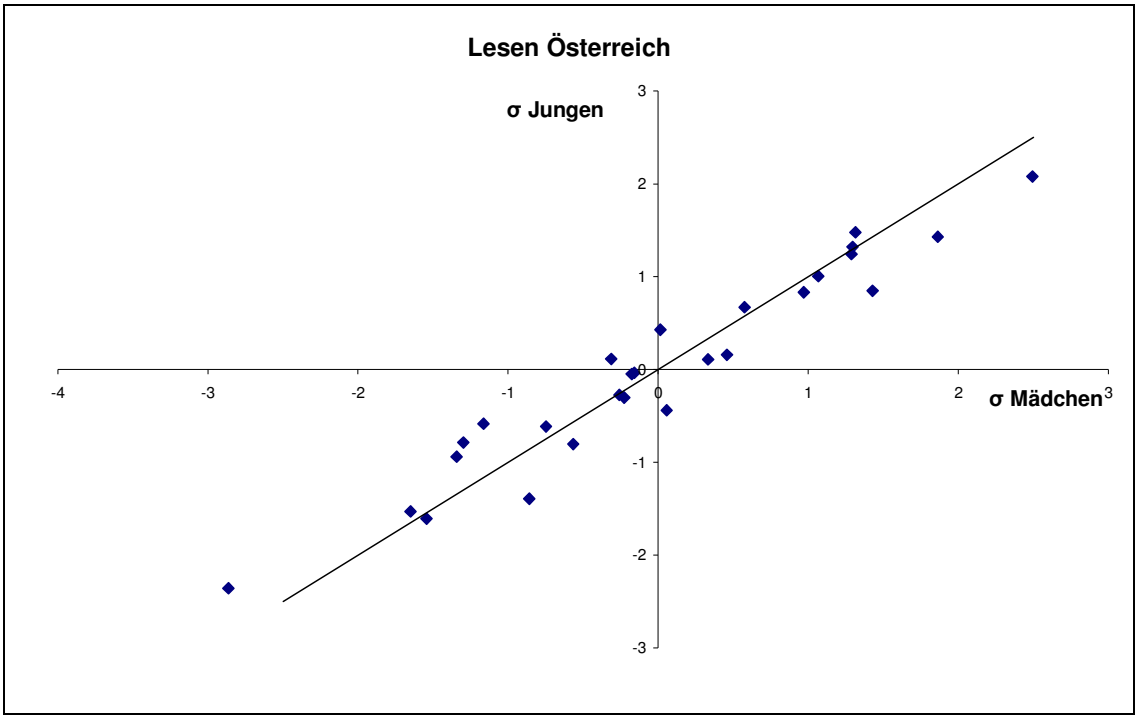

Abbildung 48 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Lesetests. (Stichprobe: österreichische Schüler)

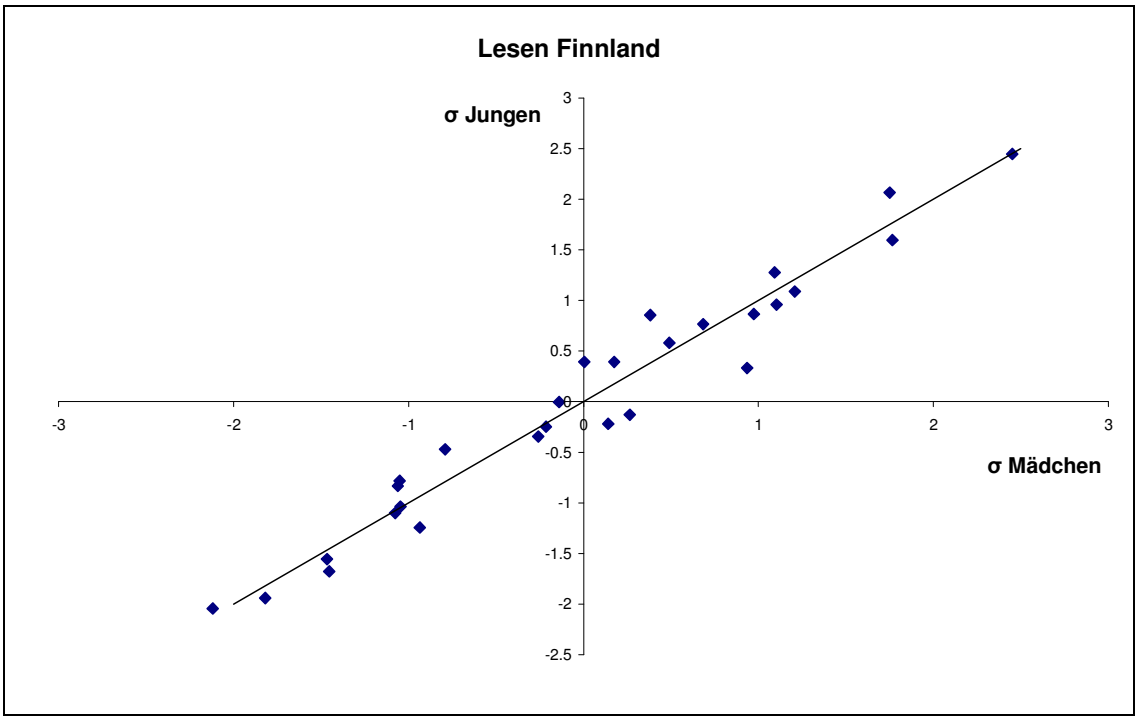

Abbildung 49 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Lesetests. (Stichprobe: finnische Schüler)

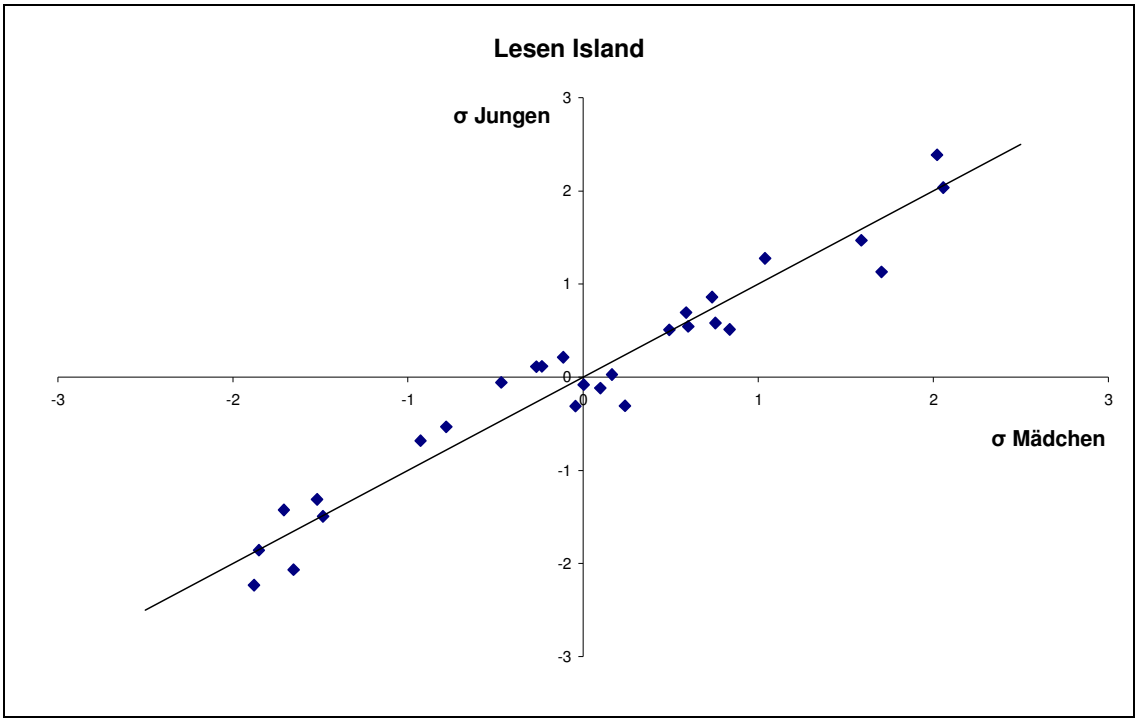

Abbildung 50 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Lesetests. (Stichprobe: isländische Schüler)

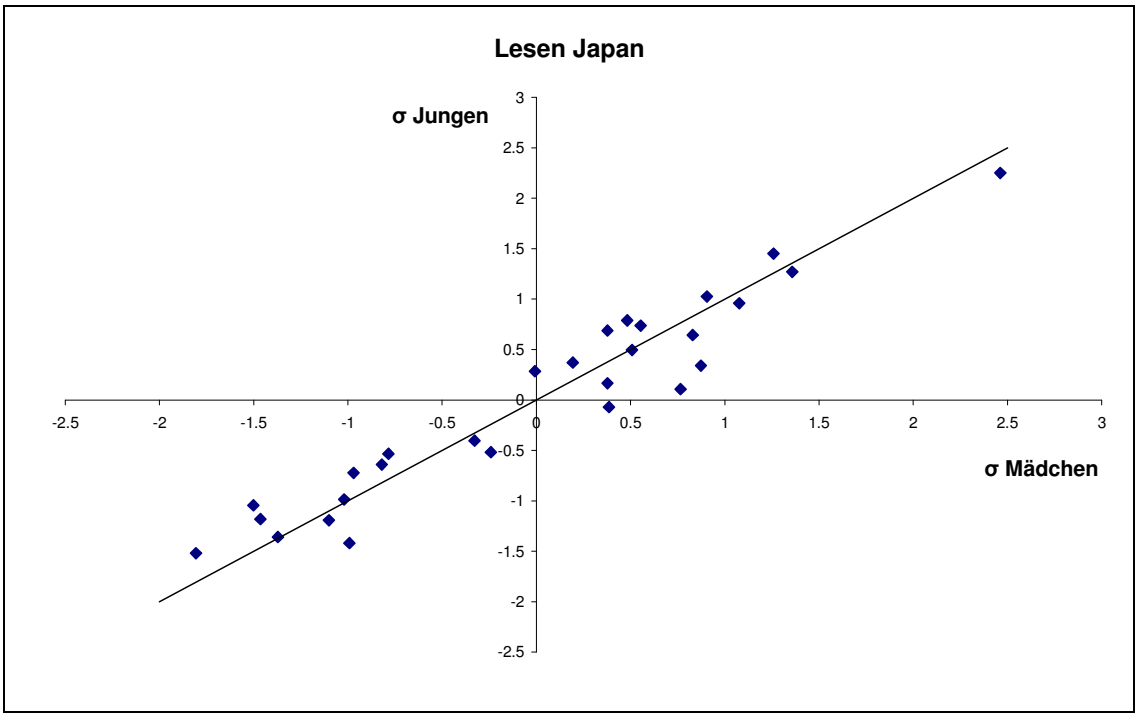

Abbildung 51 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Lesetests. (Stichprobe: japanische Schüler)

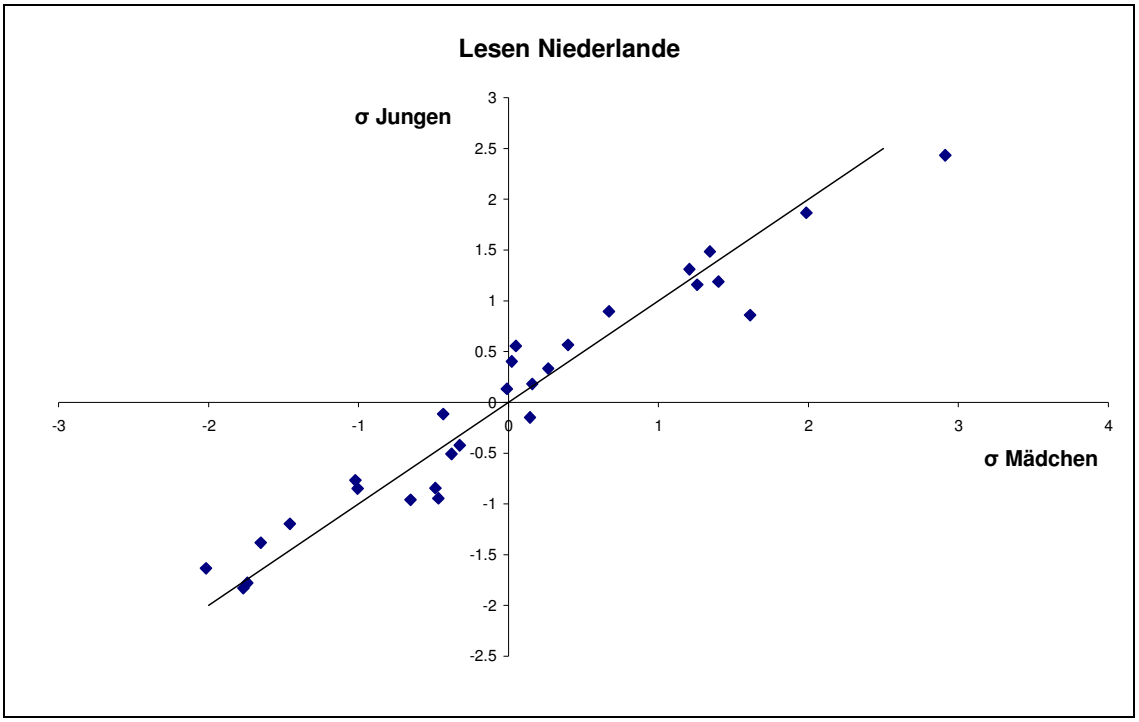

Abbildung 52 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Lesetests. (Stichprobe: niederländische Schüler)

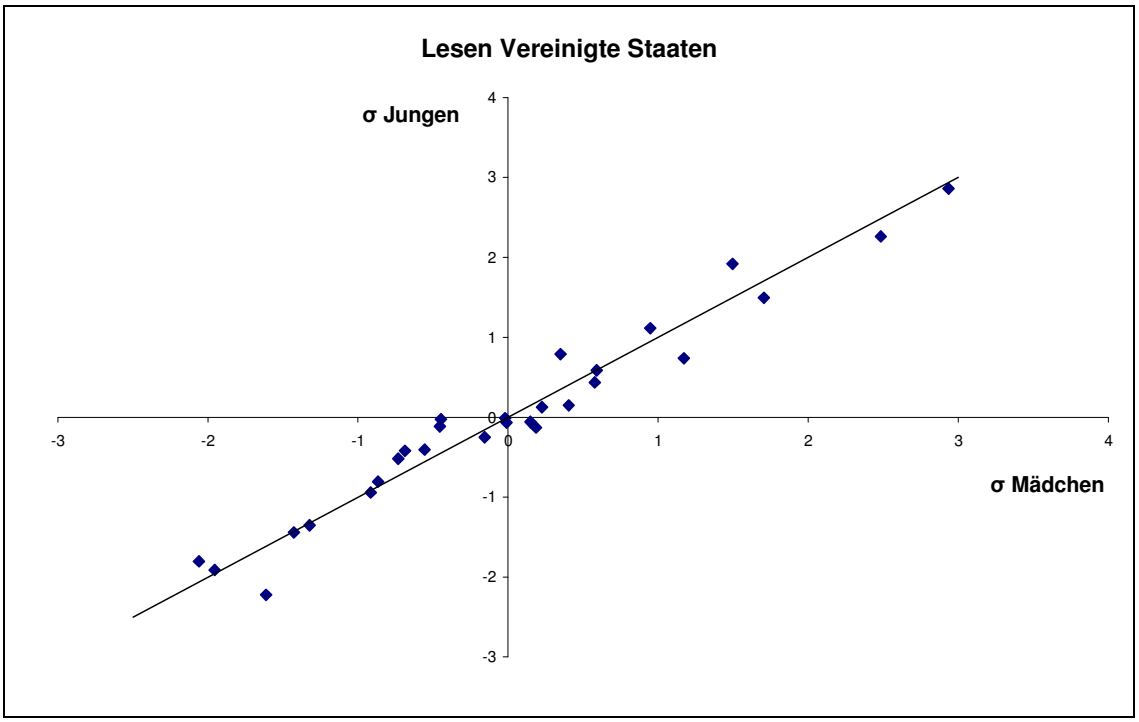

Abbildung 53 Itemschwierigkeiten σ nach einer getrennten Skalierung von Jungen und Mädchen bei den Items des internationalen Lesetests. (Stichprobe: Schüler aus den Vereinigten Staaten)

7 Verzeichnis der Abbildungen und Tabellen

7.1 Abbildungen

7.2 Tabellen und Exkurse

Lebenslauf

Name:	Désirée Sylvia Burba
Geburtsdatum:	12.02.1977
Geburtsort:	Oldenburg i.H.
Staatsangehörigkeit:	deutsch
Familienstand:	verheiratet
Schulbildung	
1983-1987	Grundschule Puttgarden
1987-1996	Gymnasium in Burg a.F. Abschluß Abitur
Studium	
1996-2002	Studium der **Psychologie** an der Christian-Albrechts-Universität zu Kiel, Diplomarbeitsthema: „Hemisphärendominanz bei der Verarbeitung emotionaler Reize“ Abschluß mit Diplom
2004-2006	Interdisziplinärer Promotionsstudiengang **„gender studies“** an der Christian-Albrechts-Universität zu Kiel
Berufliche Tätigkeiten	
Oktober 1998 bis März 1999	**Tutorin** für das Empirische Praktikum bei PD Dr. Ellen Aschermann
Oktober 1998 bis Juli 2000	**Tutorin** für Statistik bei PD Dr. Johannes Andres
Oktober 2000 bis September 2001	**Wissenschaftliche Hilfskraft** bei Prof. Dr. Thomas Bliesener
Juni 2002 bis Dezember 2002	**Diplom-Psychologin** in Klinik Schwedeneck (Mutter-Kind-Kur-Einrichtung)
November 2002 bis Oktober 2005	**Wissenschaftliche Mitarbeiterin** der PISA-Koordinierungsstelle für Deutschland am Leibniz-Institut für die Pädagogik der Naturwissenschaften (IPN), Doktorandin im Projekt "Programme for International Student Assessment" (PISA)
seit April 2006	**Empirikerin** in der Qualitätsagentur des Instituts für Qualitätsentwicklung an Schulen Schleswig-Holstein (IQSH)

Veröffentlichungen

Burba, D., & Rost, J. (2006). Mädchen und Jungen – unterschiedliche Fertigkeiten trotz gleicher Fähigkeiten? Ergebnisse aus PISA 2003. *in Vorbereitung*.

Prenzel, M., & Burba, D. (2006). PISA-Befunde zum Umgang mit Heterogenität. In G. Opp, T. Hellbrügge, & L. Stevens (Hrsg.), *Kindern gerecht werden. Kontroverse Perspektiven auf Lernen in der Kindheit*. Bad Heilbrunn: Klinkhardt.

Zimmer, K., Burba, D., & Rost, J. (2004). Kompetenzen von Jungen und Mädchen. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost, & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs* (S. 211-223). Münster: Waxmann.

Zimmer, K., Stick, A., Burba, D., & Prenzel, P. (2006). PISA 2003 - Kompetenzmuster von Jungen und Mädchen in den deutschen Ländern. *in Vorbereitung*.

Burba, D., Rost, J.& Draxler, C. (2005) Vortrag *„Modellierung geschlechtskorrelierter kognitiver Strukturen“* auf der 67. Tagung der Arbeitsgruppe für empirische pädagogische Forschung (AEPF), Salzburg, 20.09.2005