Dissertation

Zur Erforschung von Mathematikleistung

Theoretische Studie und empirische Untersuchung des Einflussfaktors Raumvorstellung

zur Erlangung des akademischen Grades

Doktor der Pädagogik (Dr. paed.)

im Fach Mathematik

Fakultät für Mathematik

der Technischen Universität Dortmund

vorgelegt von

Dipl.-Math. Andreas Büchter

Erstgutachter: Prof. Dr. Hans-Wolfgang Henn, TU Dortmund

Zweitgutachterin: Prof.in Dr. Regina Bruder, TU Darmstadt

Einreichung: 23. Juli 2010

Mündliche Prüfung: 12. Oktober 2010

Zusammenfassung

Seit Mitte der 1990er Jahre widmet sich die empirische Bildungsforschung verstärkt der quantitativen Erforschung von *Mathematikleistung*. Dabei werden in Deutschland relativ stabile Geschlechterunterschiede in der *Mathematikleistung* zugunsten männlicher Versuchspersonen festgestellt. Inhaltliche Erklärungsversuche bringen regelmäßig *Raumvorstellung* als möglichen Mediator für diese Geschlechterunterschiede ins Spiel, ohne dass hierfür inhaltlich passende und empirisch hinreichend abgesicherte Befunde vorliegen. Vor diesem Hintergrund ist die inhaltliche Kernfrage der vorliegenden Arbeit entstanden:

„Inwieweit lassen sich Geschlechterunterschiede in der Mathematikleistung durch Geschlechterunterschiede in der Raumvorstellung erklären?“

In einer umfassenden theoretischen Studie werden zunächst aktuelle Grundlagen und Befunde der quantitativ-empirischen Erforschung von *Mathematikleistung* zusammengefasst und aus inhaltlicher und methodischer Perspektive diskutiert. Anschließend wird der vornehmlich durch psychologische Forschungsansätze geprägte Gegenstand *Raumvorstellung* in seiner historischen Entwicklung und mit aktuellen Befunden dargestellt.

Auf dieser Basis wird im empirischen Teil der Arbeit zunächst ein Instrument entwickelt, mit dem *Raumvorstellung* ausdifferenziert und effizient erfasst werden kann. Mithilfe dieses Instruments wird der Zusammenhang von *Raumvorstellung* und *Mathematikleistung* unter besonderer Berücksichtigung etwaiger Geschlechterunterschiede untersucht. Als Instrument für die Erfassung von *Mathematikleistung* wird dabei die nordrhein-westfälische Lernstandserhebung in der Jahrgangsstufe 9 (LSE 9) verwendet.

Die erhobenen Daten werden mit einem breiten Methodeninventar ausgewertet. Neben klassischen Verfahren der multivariaten Statistik finden vor allem ein- und mehrdimensionale Rasch-Modelle sowie Strukturgleichungsmodelle Anwendung, wobei sich die Methodenauswahl eng an der inhaltlichen Fragestellung orientiert.

Mit einer inhaltlich und empirisch tragfähigen Ausdifferenzierung der beteiligten Konstrukte gelingt es, Geschlechterunterschiede in der *Mathematikleistung* statistisch vollständig durch entsprechende Geschlechterunterschiede in der *Raumvorstellung* zu erklären. Dabei spielt die Raumvorstellungskomponente *mentale Rotation* eine zentrale Rolle.

Insgesamt zeigen die Ergebnisse der empirischen Untersuchung, dass (a) *Raumvorstellung* ein wesentlicher Bestandteil in Rahmenmodellen für die Erforschung von *Mathematikleistung* sein sollte, (b) *Raumvorstellung* dabei in theoretisch und empirisch abgesicherte Komponenten ausdifferenziert betrachtet werden muss und (c) mehrdimensionale Modellierungen von *Mathematikleistung* für mathematikdidaktische Fragestellungen in der Regel ergiebiger sind als eindimensionale Modellierungen.

Inhaltsverzeichnis

Vorwort

Forschungsberichte – und dies gilt insbesondere für Dissertationen – erwecken häufig den Eindruck, dass sich die Themen geradezu zwangsläufig aus dem aktuellen Stand der Wissenschaft und noch offenen Fragen ergeben. Dies ist kaum verwunderlich, da wissenschaftliche Veröffentlichungen gewöhnlich nicht den (manchmal verworrenen) Prozess der Gewinnung von Ergebnissen darstellen, sondern die Ergebnisse selbst und den im Nachhinein geglätteten Weg dorthin. Irrwege werden in den seltensten Fällen beschrieben, subjektive Setzungen oder Entscheidungen im Rahmen des „objektiven Erkenntnisgewinns“ nicht immer erwähnt. Natürlich zeichnet sich wissenschaftliches Arbeiten gerade durch kontrollierte Subjektivität und einen möglichst hohen Grad an Systematik und intersubjektiver Nachvollziehbarkeit aus, aber zumindest die Themenfindung bei Einzelvorhaben (wie Dissertationen) dürfte doch allzu oft von individuellen Präferenzen und aktuellen Rahmenbedingungen, unter denen die Forschenden arbeiten, geprägt sein. Auch hieraus entsteht am Ende häufig ein echter Beitrag zum Stand der jeweiligen Disziplin. Der Weg zu „meinem Thema“ ist durch die folgenden – mehr oder weniger unsystematisch zustande gekommenen – Ereignisse und Erfahrungen entstanden:

- Als 1997 die Ergebnisse der *TIMS*-Studie veröffentlicht und diskutiert wurden, habe ich im Rahmen des „Qualifikations- und Forschungskolloquiums“ am Dortmunder *Institut für Schulentwicklungsforschung* (als Mathematikstudent im Kreis von Schulpädagogen) intensiv die Befunde miterörtert. Damals wurde mein Interesse für Schulleistungsstudien im Allgemeinen und solche im Fach Mathematik im Besonderen geweckt.
- Im Herbst 1998 habe ich mich im Rahmen eines Kontakts zum späteren *Institut für Kognitive Mathematik* (Osnabrück) erstmalig intensiv mit „prädikativem“ und „funktionalem“ Denken auseinandergesetzt. Dabei sind mir die Aufgaben zur Diagnostik, bei denen es vor allem um mentale Manipulationen von Figuren geht, und die Geschlechterunterschiede bezüglich der Denkstile besonders in Erinnerung geblieben.
- Einige Zeit später, nach Lehrtätigkeiten in der Jugendberufshilfe und der beruflichen Qualifizierung, habe ich als Mitarbeiter am Dortmunder *Institut für Schulentwicklungsforschung* den Modellversuch „Selbstständiges Lernen in der gymnasialen Oberstufe – Mathematik (SelMa)“ evaluiert, mich also forschend mit Mathematikunterricht befasst. Mein Büro habe ich mit einem Kollegen geteilt, der überwiegend mit Arbeiten für *PISA 2003* beschäftigt war, was zu einer intensiven methodischen wie inhaltlichen Diskussion dieser Schulleistungsstudie geführt hat.
- Zur gleichen Zeit wurden am *Institut für Schulentwicklungsforschung* zentrale Befunde von *PISA 2000* intensiv diskutiert. Im entsprechenden Bericht wurden Geschlechterunterschiede in der Mathematikleistung u. a. auf unterschiedlich ausgeprägte Raumvorstellung zurückgeführt.

- Nach meinem Wechsel ans Dortmunder *Institut für Entwicklung und Erforschung des Mathematikunterrichts* im Jahr 2002 habe ich im Rahmen des mathematikdidaktischen Kolloquiums einen Vortrag von Cornelie Leopold mit dem Titel „Fähigkeit der Raumvorstellung – Genderaspekte und Förderung“ gehört. Die Aufgabenbeispiele aus Tests zur Raumvorstellung und die Geschlechterspezifik haben mich zum Teil an die Thematik der Denkstile erinnert.

Über diesen längeren Zeitraum hinweg ist dadurch eine Fragestellung entstanden, die ich im Folgenden ergebnisoffen (sic!) untersucht habe. Der Zusammenhang von Raumvorstellung und Mathematikleistung, insbesondere ein möglicher Beitrag zur Erklärung geschlechtsspezifischer Leistungsunterschiede, standen dabei im Vordergrund. Darüber hinaus vermutete ich auch eine inhaltliche Nähe zur Thematik prädikativen vs. funktionalen Denkens.

Die vorliegend Arbeit gehört folglich nicht zur Kategorie „mathematikdidaktische Entwicklungsforschung“, aus der in der Regel direkt umsetzbare oder zumindest einfach adaptierbare Konzepte für den Mathematikunterricht stammen, sondern eher zur Kategorie „Grundlagenforschung“ (in großer Nähe zur Bezugsdisziplin *Psychologie* – und zum Teil auf deren Gebiet). Methodisch orientiert sie sich an den aktuellen Studien der empirischen Bildungsforschung. Insgesamt fokussiert diese Arbeit stark auf die in schriftlichen Tests messbaren Leistungen, was bei mir auch zu Unbehagen führt, da spätestens seit der Berichtslegung zu *PISA 2000* Leistungsvergleiche und die in ihnen verwendeten Aufgabenformate die Diskussion über die Entwicklung des Fachunterrichts in Teilen dominieren.

Nach diesen persönlichen Anmerkungen zu „meinem Thema“ ist es mir ein äußerst wichtiges Anliegen, Danke zu sagen. Dieser Dank gebührt all denen, die mich in den vergangenen Jahren bei dieser Arbeit unterstützt und mich zuweilen angetriebenen haben. Allen voran gebührt dieser Dank meiner Familie, die mich neben meiner hauptberuflichen Tätigkeit und anderen „mathematikdidaktischen Hobbys“ des Öfteren auch für die vorliegende Arbeit entbehren musste und mich trotzdem selbstlos unterstützt hat. Des Weiteren gilt mein Dank den Schülerinnen und Schülern, die meine Vor- und Hauptuntersuchung als Versuchspersonen ertragen und getragen haben, und den Kollegen, die mir als „Türöffner“ Zugang zu diesen Schülerinnen und Schülern verschafft haben. Bedanken möchte ich mich auch bei allen Kolleginnen und Kollegen, die in fachlichen Diskussionen zur Ausschärfung der einen oder anderen Stelle der vorliegenden Arbeit beigetragen haben – aus Furcht einzelne Personen zu vergessen, versuche ich erst gar nicht, alle aufzuzählen; alle, denen dieser Dank gebührt, wissen, dass sie eingeschlossen sind. Schließlich möchte ich mich für die Betreuung und Begutachtung dieser Arbeit ganz herzlich bei Prof. Dr. Hans-Wolfgang Henn und bei Prof.in Dr. Regina Bruder bedanken, insbesondere auch für ihre Geduld …

Dortmund im Juli 2010
Andreas Büchter

1 Einleitung: Mathematikleistung im Fokus

Die Veröffentlichung und Diskussion der Ergebnisse der „Third International Mathematics and Science Study (TIMSS[1])“ haben das bundesdeutsche Schulsystem in der zweiten Hälfte der 1990er Jahre in seinen Grundfesten erschüttert. In der Sekundarstufe I waren die *Mathematikleistungen* deutscher Schülerinnen und Schüler im internationalen Vergleich lediglich durchschnittlich (vgl. Baumert & Lehmann, 1997) – und passten somit nicht ansatzweise zum damaligen Selbstbild der Bildungspolitik, -administration und -praxis. Für die an der bildungspolitischen Diskussion Beteiligten war dabei nicht nur der Vergleich zu anderen Nationen erschütternd, sondern auch Lösungshäufigkeiten[2] zu einzelnen Aufgaben bzw. Aufgabenbereichen. Über die reinen Testergebnisse hinaus offenbarte die Unterrichtsstudie *TIMSS-Video* Problembereiche der Unterrichtsgestaltung und der längerfristigen Unterrichtsplanung (vgl. Knoll, 1998), die aber zumindest in der Mathematikdidaktik auch schon vorher diskutiert wurden.

Eine bildungspolitische Konsequenz, die aus den *TIMSS*-Ergebnissen gezogen wurde, war die Einrichtung des zunächst auf fünf Jahre angelegten BLK-Modellversuchsprogramms „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts (SINUS)“ (vgl. BLK, 1997), das anschließend mit „SINUS-Transfer“ um insgesamt vier Jahre verlängert wurde. Darüber hinaus stellten die Diskussionen über die *TIMSS*-Ergebnisse und die ersten Konsequenzen, die daraus gezogen wurden, den Einstieg in die sogenannte „empirische Wende“ (Lange, 1999) dar. Dieser Prozess wurde durch die zyklische Teilnahme an den *PISA*-Studien noch verstärkt. Fortan wurde die „Leistung der Schule“ (Weinert, 2001) vor allem anhand der Fachleistungen von Schülerinnen und Schülern gemessen und zum Kristallisationspunkt der Diskussionen über Schule (vgl. Bonsen et al., 2004). Fast alle Bundesländer führten in den 2000er Jahren zentrale Vergleichsarbeiten[3] und – soweit diese noch keine längere Tradition hatten – zentrale Prüfungen am Ende der Sekundarstufen ein. Auf curricularer Ebene manifestierte sich die „empirische Wende“ vor allem in Fachleistungsstandards[4], die seitdem vorgeben, was Schülerinnen und Schüler am Ende bestimmter Bildungsabschnitte können sollen (vgl. Büchter et al., 2005).

---

[1] Alle in dieser Arbeit verwendeten Abkürzungen sind im Abkürzungsverzeichnis (S. 278 f.) erläutert.

[2] Präziser müsste es eigentlich „die geschätzten Lösungswahrscheinlichkeiten“ heißen, da die Testergebnisse mithilfe der „Item Response Theory (IRT)“, die häufig auch als „Probabilistische Testtheorie (PTT)“ bezeichnet wird, ausgewertet wurden. Das dabei hauptsächlich verwendete Testmodell, das „Rasch-Modell (RM)“ wird auch im Rahmen der vorliegenden Arbeit ein zentrales Analyseinstrument sein.

[3] Diese zentralen Verfahren werden zum Teil anders bezeichnet, in NRW z. B. als „Lernstandserhebungen“.

[4] Länderübergreifend sind dies die Standards der Kultusministerkonferenz (KMK, 2004, 2005a, 2005b); in den meisten Bundesländern gibt auf der Basis der KMK-Standards – und mit dem Anspruch, diese auf Landesebene umzusetzen, – „Kernlehrpläne“, „Kerncurricula“ oder ähnlich bezeichnete curriculare Vorgaben.

Schulleistungsforschung im Aufwind …

Mit dieser starken Orientierung an Fachleistungen, der Teilnahme an nationalen wie internationalen Schulleistungsstudien sowie der Einführung zentraler Vergleichsarbeiten wurde auch die empirische Bildungsforschung gestärkt. Damit begann eine engere Kooperation zwischen der Psychologie, der Schulpädagogik und den Fachdidaktiken, wobei die Mathematikdidaktik in vielen Bereichen eine Vorreiterrolle übernommen hat. Mit einer verstärkten Forschungsförderung in diesem Bereich, u. a. über DFG-Schwerpunktprogramme, soll dazu beigetragen werden, dass Schülerleistungen und Leistungsunterschiede nicht nur erfasst und verglichen, sondern – über geeignete Rahmenmodelle und darauf basierenden Untersuchungen – auch zunehmend besser „erklärt“[5] werden können.

… und in der Kritik

So grundsätzliche Veränderungsprozesse wie der oben skizzierte rufen natürlich auch Kritik hervor. Aus Teilen der Mathematikdidaktik wurde insbesondere die *PISA*-Studie und daran die aktuelle methodologische Grundlegung der Erforschung von *Mathematikleistung* hinterfragt (vgl. Meyerhöfer, 2005; Jahnke & Meyerhöfer, 2007; fachübergreifend Hopmann, et al. 2007). Neben der generellen Problematisierung interkultureller Schulleistungsvergleiche sowie bildungs- und wissenschaftspolitischer Aspekte von „*PISA & Co.*“ wurden u. a. die folgenden Punkte kritisiert, die für die Erforschung von *Mathematikleistung* generell von Interesse (und hier auf „Large Scale Assessments“ bezogen) sind:

- Tests, die ausschließlich schriftlich und mit ökonomisch auswertbaren Aufgabenformaten (mit einem großen Anteil von „Multiple-Choice-Items“) gestellt werden, können höchstens einen Ausschnitt von curricular intendierter mathematischer Bildung erfassen. Umfassendere bildungstheoretische Konstrukte, wie z. B. „Mathematische Grundbildung“ bzw. „Mathematical Literacy“ (vgl. Klieme et al., 2001, S. 141 ff.; Jablonka, 2007), können nicht als Ganzes Gegenstand solcher Untersuchungen sein. Die Aufmerksamkeit wird infolge der Studien und der Diskussion ihrer Ergebnisse aber vor allem auf die messbare Leistung und die zugrundeliegenden Testaufgaben gelenkt.
- Die Selektion der Aufgaben, die statistische Verdichtung der Testdaten und darauf aufbauende Analysen und Interpretationen können immer nur unter der Voraussetzung der Gültigkeit des zugrundegelegten Testmodells stattfinden. Eine (auch nur leicht) eingeschränkte Modellgeltung und bestimmte forschungspragmatische Konventionen[6] kön-

[5] Dabei geht es zunächst um eine statistische „Erklärung“ im Sinne der Vorhersage von Leistungsdaten mithilfe statistischer Modelle. Für die Mathematikdidaktik ist darüber hinaus die theoretisch-inhaltliche Erklärung des gezeigten Leistungsverhaltens und von Lernprozessen von besonderer Bedeutung.

[6] In Analogie zur (willkürlichen) Festlegung des Signnifikanzniveaus bei Hypothesentests, die häufig nur eine Konvention unreflektiert fortschreibt, trifft dies z. B. auf Kriterien für die Aufgabenselektion oder auf Kennwerte für die Modellgüte zu (z. B. Gewichtung des *Kriteriums der Sparsamkeit* eines Modells).

nen zu Artefakten führen, die anschließend sowohl zur Grundlage von bildungspolitischen Konsequenzen werden als auch konzeptionell prägend für nachfolgenden Leistungsuntersuchungen sein können. Ein Beispiel hierfür ist die Frage der Dimensionalität von Fachleistungen, insbesondere von *Mathematikleistung*. Auf der Basis zirkulär anmutender Begründungszusammenhänge wurde *Mathematikleistung* in der empirischen Bildungsforschung lange Zeit überwiegend eindimensional modelliert (vgl. Kap. 2.3.3).

- Die Studien sollen jeweils einen bestimmten Zweck erfüllen. *PISA* soll z. B. zur Generierung von Indikatoren für die Leistungsfähigkeit von Bildungssystem im internationalen Vergleich beitragen. Bei der Interpretation der Ergebnisse solcher Studien wird dieser relativ enge Rahmen häufig verlassen und stattdessen ein breiterer Rahmen unterstellt. Dies gilt interessanter Weise nicht nur für die Bildungspolitik, sondern auch für beteiligte Wissenschaftlerinnen und Wissenschaftler – und häufig auch für die Kritikerinnen und Kritiker dieser Studien.

Bei aller Kritik an bisherigen Schulleistungsstudien, deren methodischen Grundlagen oder „der" empirischen Bildungsforschung insgesamt, wird in der Breite der beteiligten Wissenschaften kaum bezweifelt, dass die verwendeten Tests zumindest hinreichend breite Ausschnitte vieler Komponenten von *Mathematikleistung* erfassen. Zentrale Befunde zur *Mathematikleistung*, wie z. B. eine lokale Stärke deutscher Schülerinnen und Schüler beim Kalkül, eine lokale Schwäche im Bereich Stochastik, Geschlechterunterschiede[7] oder die enge Kopplung an die soziale Herkunft, wiederholen sich von Studie zu Studie und stehen auch im Einklang mit qualitativen Befunden zum deutschen Mathematikunterricht. Im Sinne einer Methodentriangulation und einer heuristischen Argumentation können solche qualitativen Befunde die Testergebnisse plausibilisieren und damit unterstützen.

Differenzierte Blicke auf Schulleistung: Erklärungsansätze für Unterschiede

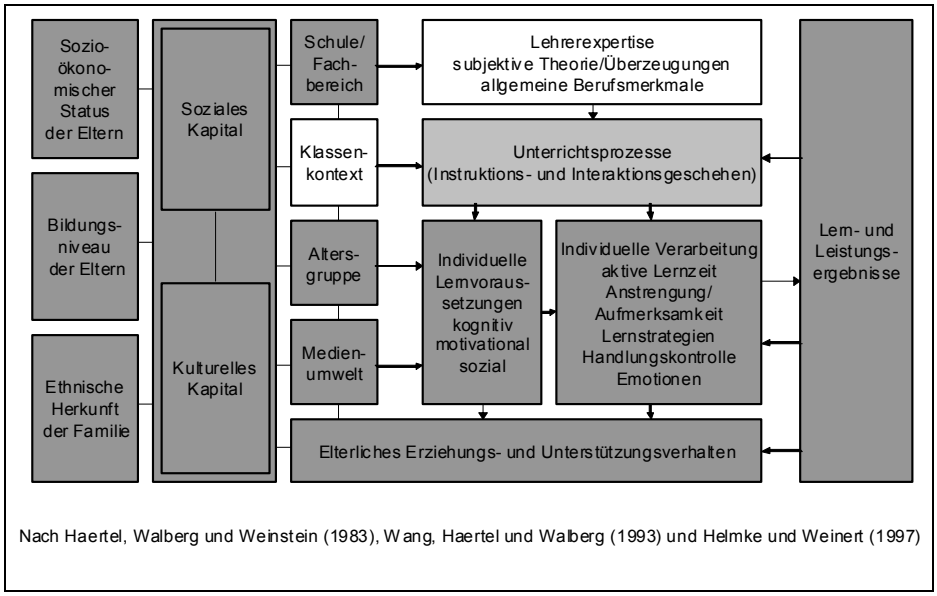
Die großen Schulleistungsstudien erheben über die Fachleistungen hinaus viele Variablen zum Hintergrund der getesteten Schülerinnen und Schüler. Dies sind neben demographischen Angaben z. B. Einstellungen zum Lernen, allgemeine kognitive Fähigkeiten oder der

[7] In der vorliegenden Arbeit werden durchgängig die Begriffe „Geschlecht" bzw. „Geschlechterunterschiede" verwendet. Eine mögliche Differenzierung der Kategorie „Geschlecht" in die beiden Kategorien „Sex" und „Gender" spielt in dieser Arbeit nur implizit bei den möglichen Erklärungsansätzen für Geschlechterunterschiede in der *Mathematikleistung* bzw. in der *Raumvorstellung* eine Rolle. Dort werden sowohl die gängigen biologischen Modelle (passend zur Kategorie „Sex") als auch die gängigen sozialisationstheoretischen Modelle (passend zur Kategorie „Gender") skizziert.

Eine weitere Bemerkung zum Thema „Geschlechterunterschiede" ist an dieser Stelle wichtig: Entsprechende Aussagen über Unterschiede in der *Mathematikleistung*, in der *Raumvorstellung* oder in anderen Komponenten kognitiver Leistung beziehen sich stets auf Mittelwerte der betrachteten Gruppen. In der Regel sind bei entsprechenden Test unter den Versuchspersonen mit den besten bzw. den schlechtesten Testleistungen jeweils sowohl männliche als auch weibliche Versuchspersonen.

sozioökonomische Status. In Abbildung 1.1 ist das „*PISA*-Rahmenmodell“ dargestellt, das eine konzeptionelle Grundlage für die Datenerhebungen der Untersuchung darstellt.[8]

Abbildung 1.1: „Bedingungen schulischer Leistungen – Allgemeines Rahmenmodell“ (Quelle: Baumert et al., 2001, S. 33)



Bei der vertiefenden Auswertung der Untersuchung können dann Fachleistungen und Hintergrundvariablen mit Verfahren der multivariaten Statistik zueinander in Beziehung gesetzt werden (z. B. *Mathematikleistung* und Einstellungen zum Lernen unter rechnerischer Kontrolle der allgemeinen kognitiven Fähigkeiten). Auf entsprechenden Analysen beruhen dann z. B. Befunde zur Kopplung von sozialer Herkunft und Fachleistungen oder zu Geschlechterunterschieden in der *Mathematikleistung*.

Da viele manifeste Variablen wie *Geschlecht* oder *Familieneinkommen* nicht direkt auf Fachleistungen wirken, ist ein Rahmenmodell wie das obige umso hilfreicher, je mehr mögliche Mediatorvariablen enthalten sind, über die interindividuelle Unterschiede entstehen können. Eine solche Mediatorvariable kann im Falle des Zusammenhangs von *Geschlecht* und *Mathematikleistung* zum Beispiel das *bereichsspezifische Fähigkeitsselbstkonzept* (vgl. Moschner & Dickhäuser, 2006) sein. Wenn potenzielle Mediatorvariablen in der Untersuchung nicht erhoben wurden, müssen differenzielle Befunde ggf. unter Rück-

[8] Bei den internationalen *PISA*-Erhebungen wird der Altersjahrgang der 15-Jährigen untersucht, der in allen Bildungssystemen über mindestens zwei Jahrgangsstufen verteilt ist. Umgekehrt gibt es in der Regel keine Lerngruppen, die nur aus 15-Jährigen bestehen. Daher werden potenziell relevante Merkmale, die nur auf Ebene der Lerngruppe erfasst werden können, bei *PISA* nicht erhoben – im Schaubild sind sie durch den weißen Hintergrund gekennzeichnet.

griff auf die Ergebnisse anderer Untersuchungen interpretiert werden. Ein Beispiel hierfür ist die Rückführung von Geschlechterunterschieden in der *Mathematikleistung* auf entsprechende Unterschiede in der *Raumvorstellung* im ersten ausführlichen Ergebnisbericht zu *PISA 2000* (Deutsches PISA-Konsortium, 2001):

> „In den Bereichen Mathematik und Naturwissenschaften sind nach wie vor die Mädchen benachteiligt. […] Darüber hinaus konnten sowohl in den Naturwissenschaften als auch in der Mathematik geschlechtsspezifische Stärken und Schwächen bei verschiedenen Anforderungen identifiziert werden. Diese Ergebnisse weisen darauf hin, dass Leistungsnachteile für Mädchen insbesondere bei Aktivitäten zu beobachten sind, die sich auf Modellierungen beziehen (Heranziehen eines mentalen Modells in den Naturwissenschaften, rechnerisches Modellieren sowie Mathematisierung von Situationen in der Mathematik). Dies wiederum dürfte zumindest teilweise auf die in der Literatur beschriebene relative Schwäche von Mädchen im räumlichen Vorstellungsvermögen zurückzuführen sein. So konnte beispielsweise Klieme (1986) zeigen, dass der Geschlechterunterschied bei mathematischen Modellierungs- bzw. Anwendungsaufgaben auf das bei anderen Arten von Aufgabenstelllungen beobachtete Niveau reduziert wird, wenn man die Fähigkeit zum bildlichen Denken kontrolliert (vgl. auch Maier, 1999[b])“ (Stanat & Kunter, 2001, S. 267; Erg. d. d. Verf.).

Im wiedergegebenen Zitat wird auf Ergebnisse einer Untersuchung verwiesen, die im Jahr 1986 von Klieme veröffentlicht wurde und die mit Studienanfängern durchgeführt wurde. Da sich die Sozialisationsbedingungen seit den frühen 1980er Jahren erheblich verändert haben, kann nicht ausgeschlossen werden, dass sich die Ergebnisse mittlerweile anders darstellen. Bei einer Erklärung von Leistungsunterschieden, die im Rahmen einer Schulleistungsuntersuchung in der Sekundarstufe I festgestellt werden, muss zusätzlich berücksichtigt werden, dass Studienanfänger eine „ausgelesene“ Stichprobe darstellen.

Erkenntnisleitendes Interesse der vorliegenden Arbeit

Das Hauptanliegen der vorliegenden Arbeit ist es, aktuelle Ergebnisse zum Zusammenhang von *Raumvorstellung* und *Mathematikleistung* unter Berücksichtigung möglicher Geschlechterunterschiede zu liefern, die etwa für die „*PISA*-Population“[9] gültig sind. Damit wird auch ein Beitrag zur Ausdifferenzierung von Rahmenmodellen geleistet, die der Untersuchung von *Mathematikleistung* zugrunde liegen.

Unter der zuvor dargestellten Zielsetzung sollen im Bereich der individuellen kognitiven Voraussetzungen Variablen erfasst werden, die möglicherweise als Mediatorvariablen für Geschlechterunterschiede bei der *Mathematikleistung* fungieren. Neben der *Raumvorstellung* kommen hier z. B. unterschiedliche *Denkstile*[10] infrage, wobei auch eine Interaktion

[9] Dieser Population der Schülerinnen und Schüler, die sich kurz vor dem Ende der Sekundarstufe I befinden, kommt eine besondere Bedeutung zu. Die 15-Jährigen sind in Deutschland (nahezu ohne Ausnahme) vollzeitschulpflichtig im allgemeinbildenden Schulsystem und befinden sich an der Schwelle zur beruflichen Ausbildung oder vertieften schulischen Bildung.

[10] Im Rahmen dieser Arbeit werden „Denkstile“ im Sinne des Konstrukts „prädikativen vs. funktionalen Denkens“ der Osnabrücker „Kognitiven Mathematik“ verstanden (vgl. Schwank, 2003a).

dieser beiden Variablen theoretisch plausibel ist. Für die Mathematikdidaktik stellt die befriedigende Erklärung von Leistungsunterschieden (hier zwischen den Geschlechtern) eine Voraussetzung zur Klärung der Frage bei, ob und ggf. wie solche Leistungsunterschiede im Mathematikunterricht bearbeitet werden können. Diese Zielsetzung ist zunächst in den folgenden drei Arbeitsschritten umgesetzt worden:

- Der aktuelle Stand der Erforschung von *Mathematikleistung* sowie eine Klärung des Konstrukts *Raumvorstellung* waren Gegenstand einer umfassenderen theoretischen Studie, wobei ein Schwerpunkt auf die Systematisierung der konzeptionellen Entwürfe und der Befunde zur *Raumvorstellung* gelegt wurde.
- Im Rahmen einer Voruntersuchung wurde ein Instrument, bestehend aus mehreren zuverlässigen und inhaltlich klar konturierten Tests zur Messung von *Raumvorstellung*, entwickelt und erprobt. Darüber hinaus wurde untersucht, ob sich die Erfassung von *Denkstilen* mit *Paper and Pencil Tests* realisieren lässt.[11]
- Der Einsatz des so entwickelten Instruments erfolgte dann im Rahmen der Hauptuntersuchung in zeitlicher Nähe zu den nordrhein-westfälischen „Lernstandserhebungen in der Jahrgangsstufe 9 (LSE 9)“. Dabei wurde darauf geachtet, dass die Datensätze des eigenen Instruments und die der *LSE 9* (als Test für *Mathematikleistung*) für jeden Schüler und jede Schülerin zusammengeführt und somit im Zusammenhang ausgewertet werden können.

Für die zuvor genannte Zielsetzung ist es dabei wichtig, dass die betrachteten Konstrukte ohne zu große Substanzverluste, also inhaltlich hinreichend breit und hinreichend differenziert, durch Tests operationalisiert und mit diesen Tests gemessen werden.

Im Sinne eines pragmatischen Vorgehens kann die theoretische Grundlegung auf dem aktuellen Forschungsstand stattfinden. Eine über diesen Stand hinausgehende Klärung der fraglichen Konstrukte ist nicht erforderlich. Die eigene Arbeit hat daher einen Schwerpunkt in der Instrumentenentwicklung zum Konstrukt *Raumvorstellung*. Die *Mathematikleistung* wurde mit zentral zur Verfügung gestellten *LSE 9* erfasst, auf die inhaltlich kein Einfluss genommen werden konnte. Die vorliegende Arbeit dokumentiert die oben genannten Arbeitsschritte, stellt die Auswertung der Voruntersuchung und der Hauptuntersuchung dar und diskutiert die Ergebnisse sowie mögliche Konsequenzen aus mathematikdidaktischer Perspektive.

[11] Ergänzend wurde das „Bereichsspezifische Fähigkeitsselbstkonzept – Mathematik“ berücksichtigt, da diese Variable aus dem Bereich *Selbstbezogene Kognition* ebenfalls eine potenzielle Mediatorvariable für Geschlechterunterschiede in der *Mathematikleistung* ist. Schließlich wurden bestimmte Fachnoten als externe Kriteriumsvariablen erhoben.

2 Grundlagen und Befunde der Erforschung von Mathematikleistung

Schulleistungsforschung wird – wie bereits in der Einleitung geschehen – im Folgenden vor allem im Sinne der großen nationalen und internationalen Schulleistungsstudien („Large Scale Assessments“) verstanden. Hierunter fallen neben den bereits erwähnten Studien *TIMSS* und *PISA* z. B. auch die Normierungsstudien zu den KMK-Bildungsstandards (vgl. z. B. Blum et al., 2006; Granzer et al., 2009) oder Forschungsprojekte wie PALMA (vgl. Pekrun et al., 2006), das die Entwicklungsverläufe von *Mathematikleistung* in der Sekundarstufe I untersucht. Auch die Vergleichsarbeiten, die mittlerweile in fast allen Bundesländern geschrieben werden, basieren auf der gleichen Forschungslogik und -methodik. Die Sichtweise dieser Studien auf Schulleistungen und deren Erforschung ist natürlich nicht die einzige und je nach Zielsetzung können andere Ansätze, insbesondere auch qualitative Ansätze, angemessener sein. Für die vorliegende Arbeit stellt diese Sichtweise aber keine unangemessene Einschränkung dar, da die Arbeit (a) einen Beitrag innerhalb dieses Forschungsparadigmas leisten soll und (b) die *Mathematikleistung*[12] im empirischen Teil mit den *LSE 9* erfasst wird. Im Bereich der psychometrisch abgesicherten Erfassung von Fachleistungen stellen die oben genannten Studien zurzeit sicherlich noch den „State of the Art“ dar.

Die Erforschung von *Mathematikleistung* unterscheidet sich etwa von der Erforschung von Naturwissenschaftsleistung nicht in der zugrundeliegenden Forschungslogik und -methodik, sondern vor allem durch die unterrichtsfachspezifische Begründung des jeweiligen Konstrukts *Fachleistung* und die konkret eingesetzten Aufgabenformate und Aufgaben. Die oben genannten Studien sind deshalb in der Regel kooperative Vorhaben von Psychometrie, pädagogischer Psychologie und Fachdidaktik (sowie je nach Anlage der Studie auch Schulpädagogik) im Rahmen der empirischen Bildungsforschung.

Für die Fragestellung der vorliegenden Arbeit ist zunächst von Interesse, wie differenzielle Befunde zur *Mathematikleistung* zustande kommen und wie sie erklärt werden. Daher werden im Folgenden bildungstheoretische Grundlagen der konkreten Fachleistungskonstrukte und Fachleistungstests, die zugrundeliegende Forschungsmethodik, die aus dieser Forschung resultierenden Kompetenzmodelle sowie Rahmenmodelle zur Erfassung von

[12] In dieser Arbeit werden überwiegend die Bezeichnungen „Schulleistung“, „Schulleistungsforschung“, „Schulleistungsstudien“, „Mathematikleistung“ etc. verwendet, da sie den jeweils fraglichen Gegenstand knapp und in einer üblichen Weise benennen. Brunner (2006) verwendet in Anlehnung an Weinert (2001) stattdessen die Bezeichnungsweise „mathematische Schülerleistung“, da die Leistungen zunächst von Schülerinnen und Schülern erbracht werden und die Frage, welchen Anteil die (einzelne) Schule an diesen Leistungen hat, noch weiterer Forschung bedarf (vgl. Weinert, 2001). Zu Missverständnissen dürfte allerdings weder die eine noch die andere Bezeichnung führen.

Schulleistungen dargestellt. Anschließend wird exemplarisch und eher kontrastierend aufgezeigt, welche anderen mathematikdidaktischen Perspektiven auf *Mathematikleistung* typisch sind. Zur Vorbereitung des empirischen Teils der vorliegenden Arbeit werden dann ausgewählte Befunde der Erfassung von *Mathematikleistung* zusammengestellt.

2.1 Mathematikleistung als Gegenstand der empirischen Bildungsforschung

Bei der Rezeption und Diskussion der Ergebnisse von bekannten Schulleistungsstudien – in jüngerer Vergangenheit vor allem *PISA 2000*, *2003* und *2006* – lässt sich gut beobachten, dass verschiedene „Öffentlichkeiten" an unterschiedlichen Aspekten der Studien und der Ergebnisse interessiert sind.

In der nicht-fachlichen medialen Verarbeitung und der Diskussion in der Bevölkerung insgesamt sind vor allem die „Rankings" von großem Interesse. Zwar wird immer wieder betont, dass Rankings nicht das Ziel dieser Studien sind, zumal sich die genauen Reihenfolgen verschiedener Teilnehmer(-staaten) nicht zufallskritisch absichern lassen[13]. Dennoch werden im Rahmen der Berichtslegung die geschätzten Leistungsmittelwerte in eine Reihenfolge gebracht und als solche diskutiert – ohne z. B. zu berücksichtigen, ob die Unterschiede der Mittelwerte überhaupt praktisch relevant sind. Dem „Erkenntnisinteresse" dieser Zielgruppe würde tatsächlich eine reine Erfassung der Fachleistungen ohne zusätzlich erhobene Variablen genügen. Als Grundlage muss hierfür neben einem bildungstheoretisch und fachdidaktisch tragfähigen Fachleistungskonstrukt im Wesentlichen ein geeignetes Testmodell zur Verfügung stehen, mit dem solche Ergebnisse generiert werden können.

Die Bildungsadministration ist hingegen an zusätzlichem „Steuerungswissen" interessiert, dass vor allem mithilfe der Hintergrundvariablen in entsprechenden Rahmenmodellen generiert wird. Für die Gestaltung des Bildungssystems sind Fragen wie die Kopplung von sozialer Herkunft und Bildungserfolg, die (mit Blick auf Fachleistungen beurteilte) Übergangsgerechtigkeit an „Scharnierstellen" des Bildungssystems oder die etwaige Abhängigkeit der Fachleistungen von institutionellen Merkmalen (z. B. Klassengröße) von zentralem Interesse.

---

[13] Da es sich bei den fraglichen Studien nicht um Vollerhebungen handelt, wird ein stets ein Rückschluss von einer Stichprobe auf eine Grundgesamtheit gezogen. Bei der Schätzung der Leistungsverteilung in der Grundgesamtheit können also Schätzfehler entstehen, die sich lediglich in Form von Wahrscheinlichkeitsabschätzungen (mit Konfidenzintervallen) kontrollieren lassen. Außerdem können die Messungen selbst fehlerbehaftet sein. Eine Konsequenz hieraus ist, dass bei ähnlich großen Leistungsmittelwerten nicht zuverlässig beurteilt werden kann, wer wirklich „besser" ist. Größere Unterschiede können „zufallskritisch abgesichert" werden, d. h. es wird auf der Basis von Hypothesentests oder Parameterschätzungen beurteilt, ob ein zufälliges Zustandekommen der unterschiedlichen Mess- und Schätzergebnisse unter bestimmten Voraussetzungen sehr unwahrscheinlich ist (zu Grundkonzepten der Beurteilenden Statistik vgl. z. B. Büchter & Henn, 2007, Kap. 4).

Für die betroffenen Fachdidaktiken hingegen ist u. a. von Interesse, wie das Konstrukt Fachleistung grundgelegt und begründet wird, welche konkreten Testaufgaben verwendet werden und wie die Ergebnisse ausgewertet und fachlich interpretiert werden. Da die Fachdidaktiken in der Regel eng bezogen auf Lehr-Lernprozesse arbeiten, besteht auch ein großes Interesse daran, Rückschlüsse von Leistungsdaten auf Prozessqualität zu ziehen. Darüber hinaus ist die Rückführung unterschiedlicher Leistungsergebnisse auf ausgewählte Hintergrundvariablen von großem Interesse, da dies zur Klärung der Frage beitragen kann, welche Faktoren die Ausprägung von Fachleistung maßgeblich beeinflussen.

Außer den hier genannten Gruppen gibt es natürlich viele weitere, die spezifische Interessen an den Schulleistungsstudien haben; auf diese weiteren Gruppen und ihre Interessen wird hier aber nicht weiter eingegangen.

Im Folgenden werden zunächst aktuelle bildungstheoretische und fachdidaktische Grundlagen der Erforschung von *Mathematikleistung* dargestellt, bevor Grundzüge aktueller Testmodelle, vor allem des „Rasch-Modells (RM)“, das in verschiedenen Varianten den aktuellen Schulleistungsstudien zugrunde liegt, skizziert werden. Anschließend werden Kompetenzmodelle, die im Rahmen dieser Studien relevant sind, betrachtet – und methodenkritisch diskutiert. Schließlich werden Rahmenmodelle für die Entstehung und die differenzierte Untersuchung von Schulleistungen, vor allem das *PISA*-Rahmenmodell, dargestellt.

2.1.1 Bildungstheoretische Grundlagen der Erforschung von Mathematikleistung

Eine wichtige theoretische Grundlage für die Erforschung von *Mathematikleistung* stellt, gewissermaßen als erster Schritt der Leistungsmessung, die Klärung und Präzisierung des Konstrukts „Mathematikleistung“ dar. Hierauf aufbauend können dann Aufgaben ausgewählt, angepasst oder neu entwickelt werden, die das fragliche Konstrukt bestmöglich operationalisieren. Die heute üblichen Schulleistungstests sind – von *TIMSS* und *PISA* bis hin zu Vergleichsarbeiten – schriftliche Tests mit zeitlich überschaubaren Items[14]. Dieses Format bedingt, dass es Bereiche mathematischen Arbeitens gibt – wie z. B. das geeignete Explorieren komplexerer Problemsituationen –, die nicht (vollständig) erfasst werden. Wenn entsprechende Studien dennoch ohne nennenswerte Einschränkung beanspruchen, *Mathematikleistung* zu erfassen, dann wird davon ausgegangen, dass das zugrundeliegende Konstrukt zumindest in wesentlicher Substanz operationalisiert wurde.

[14] In den üblichen Mathematiktests können sowohl kurze Aufgaben als auch Teilaufgaben ein „Item“ sein, das die kleinste Beobachtungseinheit darstellt und dessen erfolgreiche Bearbeitung logisch nicht von der erfolgreichen Bearbeitung anderer Beobachtungseinheiten abhängig sein soll.

Aus bildungstheoretischer Sicht können grundsätzlich zwei unterschiedliche Konzepte von *Mathematikleistung*[15] identifiziert werden, die sich mit den englischen Ausdrücken „achievement“ bzw. „proficiency“ charakterisieren lassen. Zwar können beide Wörter mit „Leistung“ übersetzt werden, sie betonen aber unterschiedliche zugrundegelegte Maßstäbe. So fokussiert *achievement* im Sinne von (schulischem) „Erfolg“ oder (curricularem) „Erreichen“ darauf, wie sich eine Schülerin oder ein Schüler bezüglich der institutionell-fachlichen Vorgaben entwickelt („*Ist er ein guter Lerner?*“). Demgegenüber wird mit *proficiency* im Sinne von „Befähigung“ betrachtet, welche Leistungen jemand in einem Bereich erbringt („*Verfügt sie über die Fähigkeit?*“). Diese vor allem im Fremdsprachenbereichen wichtige Unterscheidung trennt also die Bewährung innerhalb der Bildungsinstitution (*achievement*) von der Bewährung auch im außerinstitutionellen Bereich (*proficiency*).

Achievement-Tests im obigen Sinne zeichnen sich durch ihre *curriculare Validität* aus. Testaufgaben müssen einen direkten Bezug zum jeweiligen Curriculum aufweisen bzw. von diesem aus begründbar sein. Da die Curricula als gegeben vorausgesetzt werden können, besteht die bildungstheoretische Grundlegung in diesem Fall also zunächst aus der Entscheidung für diesen Test-Ansatz und dann aus der Analyse der jeweiligen Curricula. Dieser curriculumorientierte Ansatz liegt z. B. der *TIMSS*-Mittelstufenstudie (vgl. Baumert & Lehmann, 1997), einem Teil der *TIMSS*-Oberstufenstudie (vgl. Baumert, Bos & Lehmann, 2000b) und auch Vergleichsarbeiten (vgl. z. B. Heymann & Pallack, 2007) zugrunde. Für internationale Vergleichsstudien wie *TIMSS* ist dabei von zentraler Bedeutung, dass es eine hinreichend große Schnittmenge der verschiedenen nationalen Curricula gibt. Für Mathematik stellt Baumert (2002, S. 106 ff.) fest, dass es ein „internationales Kerncurriculum“ (S. 106) gibt, in dem sich „eine kulturübergreifende Verständigung und Kanonisierung“ bei der „Auswahl und Sequenzierung der Stoffe“ (S. 107) widerspiegelt.

Proficiency-Tests bedürfen dagegen eines Konzeptes, das die Anforderungssituationen festlegt, die ein „Befähigter“ bewältigen können sollte. Während Achievement-Tests eine Deskription von Curricula vorausgeht, bedürfen Proficiency-Tests über die Entscheidung für diesen Test-Ansatz hinaus also noch weiterer normativer Schritte zur Präzisierung des Konstrukts. Sowohl ein Teil der *TIMSS*-Oberstufenstudie (vgl. Baumert, Bos & Lehmann, 2000a) als auch die *PISA*-Studie (vgl. Deutsches PISA-Konsortium, 2001) basieren auf entsprechenden Konzepten, die dort jeweils „Grundbildungskonzept“ genannt werden. Die deutsche *PISA*-Expertengruppe Mathematik hat dabei das rein nützlichkeitsorientierte Konzept „mathematical literacy“, das dem internationalen Teil der *PISA*-Studie zugrunde liegt (vgl. OECD, 1999), zum Konzept „mathematische Grundbildung“ erweitert (vgl. Neubrand, 2001), das der deutschen Ergänzungsstudie zugrunde liegt. Dabei wurde insbesondere die deutsche Diskussion um allgemeinbildenden Mathematikunterricht (vgl. Heymann, 1996; Winter, 1995) konzeptionell berücksichtigt.

[15] Diese grundlegenden Betrachtungen gelten analog für andere Unterrichtsfächer bzw. Testdomänen.

Da das oben genannte Konzept „mathematische Grundbildung“ auch grundlegend für die Entwicklung der KMK-Bildungsstandards (KMK 2004, 2005a, 2005b) und der auf diese Standards bezogenen Kerncurricula der Bundesländer war, ist der Unterschied zwischen curricular validen Tests und solchen, die auf einem Grundbildungskonzept basieren, nahezu verschwunden. Ein alleiniger Blick in die Testhefte dürfte kaum noch Rückschlüsse auf den gewählten Ansatz zulassen. Lediglich bei der Aufgabenentwicklung schränkt der Anspruch curricularer Validität mögliche Aufgabenstellungen ein wenig ein, da diese eng an die Curricula (ggf. verschiedener Bundesländer oder Nationen) angebunden sein müssen.

Dass die konkreten Aufgaben kaum noch Rückschlüsse auf den gewählten Ansatz zulassen, ist aber möglicherweise auch auf Probleme bei der Operationalisierung des Grundbildungskonzepts zurückzuführen. Kritikerinnen und Kritiker der Schulleistungsstudien verweisen immer wieder darauf, dass – ihres Erachtens – die zunächst konzeptionell breit angelegte *Mathematikleistung* im realen Test nicht hinreichend in ihrer Substanz erfasst werden kann (vgl. z. B. Jablonka, 2007). Übrig bliebe dann (fast unabhängig vom bildungstheoretischen Ansatz) das deutlich schmalere Konstrukt „testbare Mathematikleistung“.

2.1.2 Aktuelle Testmodelle und deren Implikationen

Die aktuellen Schulleistungsstudien und Vergleicharbeiten generieren Ergebnisse auf der Basis von elaborierten Forschungsmethoden. Dabei spielen Testmodelle aus der „Item Response Theory (IRT)“ eine zentrale Rolle. Mit ihrer Hilfe wird von beobachtbarem (manifestem) Testverhalten, das aus Antworten auf Testaufgaben besteht, auf prinzipiell nicht beobachtbare (latente) Personeneigenschaften wie *Mathematikleistung*[16] geschlossen. Diesem Rückschluss liegt die Annahme zugrunde, dass das Testverhalten durch diese Personeneigenschaft bedingt ist und durch diese erklärt werden kann. Mithilfe des Testmodells erhält man so aus dem Testverhalten einen Messwert für diese spezielle Personeneigenschaft. *Mathematikleistung* wird dabei als quantitative Eigenschaft modelliert, die mit geeigneten *IRT*-Modellen intervallskaliert auf einem unbeschränkten Kontinuum gemessen werden kann. Übliche Annahmen solcher Modelle sind:

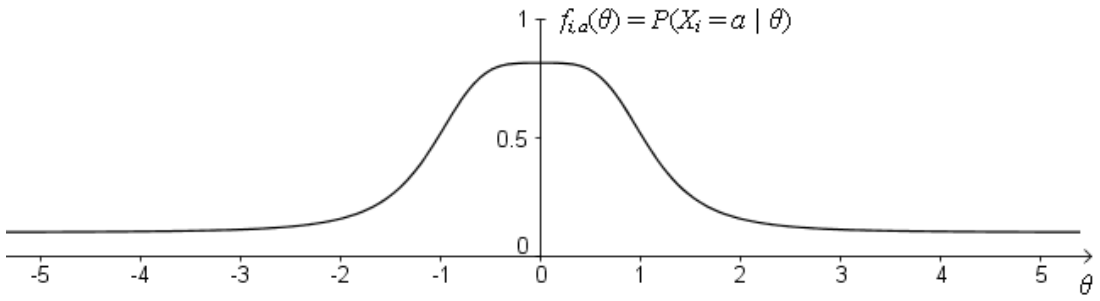
- *Itemhomogenität*: Alle Items messen dieselbe Personeneigenschaft.

[16] Insbesondere in der psychologischen Literatur wird diese latente Personeneigenschaft heute überwiegend als „(mathematische) Kompetenz“ bezeichnet. Dabei wird begrifflich zwischen nicht beobachtbarer *Kompetenz*, im Test beobachteter *Performanz* und *Leistung* (als aufgrund von Wertmaßstäben bewerteter Performanz) unterschieden (vgl. von Saldern, 1997, S. 30 ff.; Bonsen et al., 2004, S. 195 ff.). In der vorliegenden Arbeit wird die zugrunde liegende latente Eigenschaft als Mathematik*leistung* einer Person im Sinne des Vermögens bezeichnet, in einem Mathematik-Leistungstest erfolgreich abzuschneiden. Damit soll berücksichtigt werden, dass in Schulleistungstests ein Konstrukt erfasst wird, das zwar sicherlich relevant mit mathematischer Kompetenz korrespondiert, das aber durch die speziellen Testformate eine eigenen Zuschnitt erhält. Zusätzlich sei bemerkt, dass die Bezeichnungsweise „Leistung“ auch in der Bildungsforschung bis in die späten 1990er Jahre üblich war, so z. B. im Bericht zu *TIMSS/II* (Baumert & Lehmann, 1997).

- *Personenhomogenität*: Alle Versuchspersonen bearbeiten die Items aufgrund derselben zugrundeliegenden Personeneigenschaft.
- *Lokale stochastische Unabhängigkeit der Items*: Bei Versuchspersonen mit derselben Ausprägung der zugrundeliegenden Personeneigenschaft hängt die Wahrscheinlichkeit für die richtige Bearbeitung eines Items nicht von der Bearbeitung anderer Items ab.

Ein Vorteil von *IRT*-Modellen liegt in der expliziten (und theoretisch fundierten) Annahme über den Zusammenhang zwischen Personeneigenschaft und Testverhalten. Im Rahmen der Analyse der Testdaten kann dann überprüft werden, wie gut diese Annahme zu den erhobenen Daten passt. Der in einem Testmodell formulierte Zusammenhang zwischen einer quantitativen Personeneigenschaft und dem Testverhalten kann für jedes Item sehr anschaulich durch die zugehörigen „Itemfunktionen" (auch „Itemcharakteristiken" oder engl. „Item Characteristic Curves (ICCs)") für die jeweils möglichen Antwortkategorien dargestellt werden. Diese Funktionen $f_{i,a}$ geben die Wahrscheinlichkeit dafür, dass bei Item i die Antwortkategorie a gewählt wird, in Abhängigkeit von der Ausprägung θ der Personeneigenschaft an: $f_{i,a}(\theta) = P(X_i = a \mid \theta)$. Die folgende Abbildung stellt eine Itemfunktion für das Item i dar, bei der die Wahrscheinlichkeit für die Antwortkategorie a bis zur Ausprägung $\theta = 0$ der Personeneigenschaft monoton wächst und anschließend monoton fällt.

Abbildung 2.1: *ICC* für Item *i*, Antwortkategorie *a* und Eigenschaftsausprägung *θ*



Ein möglicher Nachteil von *IRT*-Modellen wird sichtbar, wenn einzelne Items oder Personen mit ihrem jeweils zugehörigen Testverhalten[17] nicht gut in das Testmodell passen. Entsprechende Personen können dann in einem erweiterten Testmodell als Klasse der „Unskalierbaren" (vgl. Rost, 2004, S 180) berücksichtigt oder aus dem Datensatz entfernt werden. Bei Items, die nicht gut in das Testmodell passen besteht fast nur die Möglichkeit, sie aus

[17] Unter „Testverhalten eines Items" wird hier die Gesamtheit der Antworten aller Versuchspersonen auf das Item verstanden. Das „Testverhalten einer Versuchsperson" wird wie oben verstanden als Gesamtheit der Antworten der Versuchsperson auf alle Items.

dem Test zu entfernen. Sowohl das Entfernen von Items aus dem Test als auch das Entfernen von Personen aus dem Datensatz kann aber inhaltlich äußerst problematisch sein kann, da das Konstrukt bzw. sein Geltungsbereich hierdurch möglicherweise eingeschränkt wird.

Über die hier genannten Aspekte hinaus gibt es viele weitere Vor- und Nachteile von *IRT*-Modellen, die in einschlägigen Fachbüchern (z. B. Fischer & Molenaar, 1995; J. Rost, 2004) diskutiert werden. Ein Vergleich zur sogenannten „Klassischen Testtheorie (KTT)" weist deutliche Unterschiede dieser Ansätze auf: Während die *IRT* vor allem Testmodelle bereitstellt, die – ausgehend vom angenommenen Zusammenhang zwischen Personeneigenschaft und Testverhalten – Messwerte liefern, geht die *KTT* von der Existenz von Messwerten aus und trifft lediglich Annahmen über die möglichen Messfehler (deswegen wird die *KTT* auch „allgemeine Messfehlertheorie" genannt, vgl. J. Rost, 2004, S. 12).[18] In diesem Sinne ergänzen sich *IRT* und *KTT*, da *IRT*-Modelle u. a. intervallskalierte Messwerte für latente Personeneigenschaften liefern, mit denen die *KTT* weiterarbeiten kann.

Aufgrund seiner spezifischen Vorteile wird aus der Familie der quantifizierenden *IRT*-Modelle vor allem das „Rasch-Modell (RM)" in zahlreichen Varianten bzw. mit zahlreichen Verallgemeinerungen in Schulleistungsstudien verwendet (vgl. Carstensen et al., 2007). Im empirischen Teil der vorliegenden Arbeit werden die Tests ebenfalls mithilfe des *RMs* ausgewertet. Grundzüge dieses Modells und übliche Varianten bzw. Verallgemeinerungen werden im Folgenden dargestellt und diskutiert.[19]

Grundzüge des Rasch-Modells

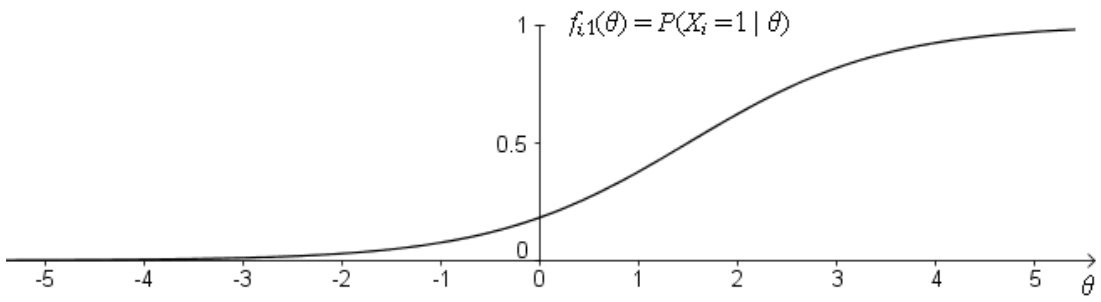
Viele psychometrische Tests sollen eine Personeneigenschaft mithilfe dichotomer Items quantifizieren. Dabei können dichotome Items aus Sicht der Versuchspersonen durchaus mehrere Antwortalternativen anbieten oder auch eine offene Antwort einfordern. Für die Analyse der Testdaten werden die Itemantworten jedoch mit lediglich zwei verschiedenen Auswertungskategorien kodiert, die im Falle von Leistungstests „richtig" und „falsch" bedeuten. Wenn bei einem Test theoretisch plausibel ist, dass die Wahrscheinlichkeit für richtige Antworten in Abhängigkeit von der quantitativen Personeneigenschaft (streng) monoton wächst, dann kann das *RM* ein geeignetes Testmodell für die Analyse der Testda-

[18] Diese vorausgesetzte Existenz von Messwerten auf einem bestimmten Skalenniveau ist bei manifesten Variablen, wie physikalischen Größen o. Ä., selten problematisch. Für latente Variablen, wie Fachleistung oder *Raumvorstellung*, die nicht direkt beobachtbar sind, sondern z. B. mittels Antwortverhalten auf Testitems erschlossen werden müssen, kann a priori nicht von existierenden Messwerten auf einem bestimmten Niveau ausgegangen werden. Eine ausführliche Darstellung der *KTT* leisten z. B. Lienert & Raatz (1998).

[19] Die Darstellung orientiert sich, wenn nicht ausdrücklich auf andere Quellen hingewiesen wird, an den entsprechenden Ausführungen in J. Rost (2004) und wird hier nicht im Detail mit konkreten Textstellen belegt.

ten sein (konkreter: das eindimensionale zweikategorielle Rasch-Modell[20]). Die *ICCs* für richtige Antworten haben in diesem Modell den folgenden Verlauf und unterscheiden sich nur durch horizontale Verschiebungen.

Abbildung 2.2: *ICC* für die richtige Antwort (*a* = 1) auf ein Item im *RM*



Das *RM* verbindet eine psychologisch plausible Annahme („Wahrscheinlichkeit für richtige Antwort wächst streng monoton mit der Ausprägung der Personeneigenschaft") mit mathematisch wünschenswerten Eigenschaften, die aus der zugrundeliegenden Modellgleichung folgen. Diese Modellgleichung wird im Folgenden hergeleitet:

Als Ausgangspunkt für die Herleitung des *RM* kann man das Anliegen betrachten, die Wahrscheinlichkeit für richtige Antworten möglichst einfach darzustellen, z. B. durch additive Zerlegung in die Ausprägung θ der Personeneigenschaft und einen Itemparameter σ_i, der als „Itemschwierigkeit" bezeichnet wird,. Der Ansatz $P(X_i = 1 \mid \theta) = \theta - \sigma_i$ impliziert, dass Personeneigenschaft und Itemparameter auf derselben Dimension liegen und dass die Wahrscheinlichkeit für die richtige Antwort steigt, wenn die Differenz $\theta - \sigma_i$ größer wird, die Ausprägung der Personeneigenschaft also gegenüber der Itemschwierigkeit steigt. Da die Wahrscheinlichkeit aber per definitionem nur Werte aus dem Intervall $[0; 1]$ annehmen kann, ist dieser einfache additive Ansatz zu einfach. Denn θ und σ_i können sich grundsätzlich an beliebigen Stellen eines unbeschränkten Kontinuums befinden, wobei sie für verschiedene Items bzw. für verschiedene Personen unabhängig voneinander variieren können. Daher kann auch die betrachtete Differenz beliebig groß oder beliebig klein werden.

Eine Lösung für dieses Problem stellt eine Transformation der Wahrscheinlichkeiten $P(X_i = 1 \mid \theta)$ in zwei Schritten auf ein ebenfalls unbeschränktes Kontinuum dar. Im ersten

[20] Dieses Testmodell wurde 1960 von Rasch publiziert und stellt die einfachste Form von Rasch-Modellen dar. Wenn in der Literatur von *dem* Rasch-Modell geschrieben wird, ist in der Regel diese Variante gemeint.

Schritt wird der „Wettquotient“ der Wahrscheinlichkeiten für die richtige und die falsche Antwort gebildet, der dann im zweiten Schritt logarithmiert wird (mit dem natürlichen Logarithmus ln). Für eine einfachere Notation werden die beiden Wahrscheinlichkeiten für richtige bzw. falsche Bearbeitungen des Items i wie folgt abgekürzt: $p_{i,1} = P(X_i = 1 \mid \theta)$ und $p_{i,0} = P(X_i = 0 \mid \theta)$. Für die Bildung des Wettquotienten muss noch der Fall $p_{i,0} = 0$ ausgeschlossen werden. Dies stellt inhaltlich keine Einschränkung dar, wenn man ein probabilistisches Testmodell haben möchte, bei dem richtige Antworten je nach Ausprägung der Personeneigenschaft zwar beliebig wahrscheinlich oder beliebig unwahrscheinlich, aber nie sicher oder unmöglich werden. Mit dem Wettquotienten

$$\frac{p_{i,1}}{p_{i,0}} = \frac{p_{i,1}}{1-p_{i,1}}$$

werden die betrachteten Werte vom Intervall $]0; 1[$ auf das Intervall $]0; \infty[$ transformiert, das immer noch nach unten beschränkt ist. Durch Logarithmieren des Wettquotienten erhält man die „Logits der Wahrscheinlichkeiten“ und damit die Transformation vom beschränkten Intervall $]0; 1[$ auf das unbeschränkte Intervall $]-\infty; \infty[$, sodass die Logits nun – wie gewünscht – additiv zerlegt werden können:

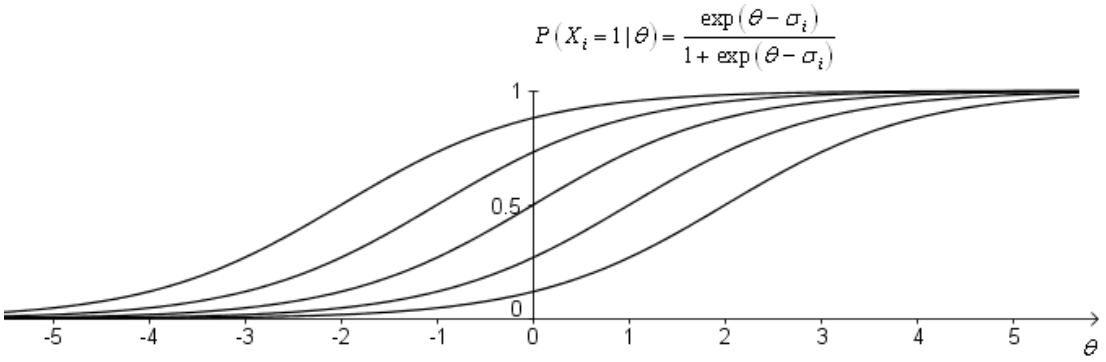
$$\ln\left(\frac{p_{i,1}}{p_{i,0}}\right) = \ln\left(\frac{p_{i,1}}{1-p_{i,1}}\right) = \theta - \sigma_i \,.$$

Die oben betrachteten *ICCs* geben sehr anschaulich jeweils die Wahrscheinlichkeiten für richtige Antworten an. Zur Funktionsgleichung für die *ICCs* im *RM* gelangt man, wenn man die Gleichung mit der additiven Zerlegung der Logits nach $p_{i,1}$ auflöst:

$$\begin{aligned}
& \ln\left(\frac{p_{i,1}}{1-p_{i,1}}\right) = \theta - \sigma_i \\
\Leftrightarrow \quad & \frac{p_{i,1}}{1-p_{i,1}} = \exp\left(\theta - \sigma_i\right) \\
\Leftrightarrow \quad & p_{i,1} = \left(1-p_{i,1}\right)\cdot\exp\left(\theta - \sigma_i\right) = \exp\left(\theta - \sigma_i\right) - p_{i,1}\cdot\exp\left(\theta - \sigma_i\right) \\
\Leftrightarrow \quad & p_{i,1} + p_{i,1}\cdot\exp\left(\theta - \sigma_i\right) = \exp\left(\theta - \sigma_i\right) \\
\Leftrightarrow \quad & p_{i,1}\cdot\left(1+\exp\left(\theta - \sigma_i\right)\right) = \exp\left(\theta - \sigma_i\right) \\
\Leftrightarrow \quad & p_{i,1} = \frac{\exp\left(\theta - \sigma_i\right)}{1+\exp\left(\theta - \sigma_i\right)} \,.
\end{aligned}$$

Die letzte Gleichung ist (bei passend gewählter Itemschwierigkeit σ_i) die Funktionsgleichung der *ICC* in Abb. 2.2 (S. 21). Die Gleichung zeigt, dass die *ICC* für $\theta = \sigma_i$ den Wert 0,5 annimmt. Die Itemschwierigkeit gibt also auch an, für welche Ausprägung der Personenfähigkeit die Wahrscheinlichkeit für eine richtige Antwort 0,5 beträgt. Für die ICC in Abb. 2.2 gilt $\sigma_i = 1{,}5$. Die folgende Abbildung 2.3 stellt *ICCs* für $\sigma_i = -2, -1, 0, 1, 2$ dar.

Abbildung 2.3: *ICC* im *RM* für die richtige Antwort und für $\sigma_i = -2, -1, 0, 1, 2$



Die Wahrscheinlichkeit für eine falsche Antwort ergibt sich analog zur Wahrscheinlichkeit für eine richtige Antwort zu $p_{i,0} = \frac{1}{1+\exp(\theta-\sigma_i)}$, was über den obigen Rechenweg bestätigt oder als Gegenwahrscheinlichkeit berechnet werden kann.

Die beiden Gleichungen für richtige und falsche Antworten können mithilfe der Variable *a* für die Antwortkategorien, die die Werte 0 (für eine/die falsche Antwort) und 1 (für eine/die richtige Antwort) annehmen kann, zur Modellgleichung des (eindimensionalen zweikategoriellen) *RMs* zusammengefasst werden:

$$p_{i,a} = \frac{\exp\left(a \cdot (\theta - \sigma_i)\right)}{1 + \exp(\theta - \sigma_i)}.$$

Insbesondere wenn man die Verortung des *RMs* innerhalb der *IRT* betonen möchte, wird auch die Bezeichnung „1-parametriges, logistisches Modell (1-PL)“ verwendet. Hierin kommt zum Ausdruck, dass (a) sich die Kernidee dieses Testmodells auch als logistische Regression der beiden Antwortkategorien „0 (= falsch)“ und „1 (= richtig)“ auf die fragliche Personeneigenschaft darstellen lässt und dass (b) das Testmodell nur einen Itemparameter, nämlich der Itemschwierigkeit σ_i, enthält. Auf nahe liegende Verallgemeinerungen des *RM* mit mehr als einem Itemparameter wird weiter unten noch eingegangen. Zuvor werden aber einige typische Eigenschaften des *RM*/*1-PL* dargestellt und diskutiert.

Ein Testmodell, das sehr wenige Parameter enthält, ist in der Regel gröber als ein Modell, das mehr Parameter verwendet. Dies kann mit dem Nachteil verbunden sein, dass das beobachtete Testverhalten nicht besonders präzise durch das Testmodell erklärt werden kann. Zugleich kann aber auch der Vorteil bestehen, dass ein „sparsames“ Testmodell theoretisch besser begründet und empirisch (mathematisch) einfacher gehandhabt werden kann:

- Das *RM* ist restriktiv bezüglich des Verlaufs der *ICCs*. Der einzige Itemparameter im Modell, die Itemschwierigkeit, nimmt keinen Einfluss auf die Form der *ICCs*, sondern

nur auf deren (horizontale) Lage. Die *ICCs* verlaufen also parallel, ihre horizontale Lage wird durch die Itemschwierigkeit σ_i bestimmt. Der parallele Verlauf der *ICCs* impliziert, dass alle Items dieselbe Trennschärfe[21] haben müssen – eine durchaus restriktive Annahme[22]. Aus Sicht des *RM* kann eine abweichende Trennschärfe, die dazu führt, dass das betrachtete Item empirisch nicht gut zum *RM* passt, darauf hindeuten, dass für die erfolgreiche Bearbeitung dieses Items noch andere Personeneigenschaften als die eigentlich betrachtete relevant sind. Bei der Entwicklung von Fachleistungstest geben solche Items immer wieder Anlass für intensive Diskussionen: Weicht ein Item in einer Voruntersuchung beim Verlauf der *ICC* bzw. bei der Trennschärfe so stark vom Testmodell ab, dass es bei der Testzusammenstellung aus psychometrischer Sicht außen vor gelassen werden soll, dann stellt sich die Frage, ob (a) die erfolgreiche Bearbeitung des Items andere Fähigkeiten als nur die Fachleistung erfordert oder (b) das zugrundeliegende Fachleistungskonstrukt selbst schon mehrdimensional ist oder (c) das Item nur schlecht formuliert ist oder … Diese Fragen lassen sich quantitativ-empirisch nur selten befriedigend klären. Werden entsprechende Items pragmatisch außen vor gelassen, stellt sich die Frage, ob das eigentlich zugrundeliegende Fachleistungskonstrukt nur eingeschränkt operationalisiert wird.

- Auf der anderen Seite – sozusagen als Gegenleistung für die starke Restriktion – hat das *RM* statistische Eigenschaften, die bei der Testauswertung sehr vorteilhaft sind. Einige dieser Eigenschaft werden hier kurz benannt:
  - Mit dem *RM* werden aus höchstens ordinal interpretierbaren Testrohdaten (mehr oder weniger richtig gelöste Testaufgaben) intervallskalierte Messwerte für die latenten Personeneigenschaften. Diese befinden sich sogar auf einer logarithmierten Verhältnisskala, also einer Differenzskala (durch die Skalierung nach dem *RM* wird die Einheit festgelegt, nicht aber der Nullpunkt). Da viele Verfahren der multivariaten Statistik mindestens Intervallskalenniveau voraussetzen, ist diese Eigenschaft besonders wichtig.
  - Wenn das *RM* hinreichend gut auf eine (große) Itemmenge und eine Population von Versuchspersonen passt, dann ist die Differenz zweier Personenparameter (= Ausprägung der Personeneigenschaft) unabhängig von den konkret für den Vergleich ausgewählten Items. Diese Eigenschaft des RM wird als „spezifische Objektivität“

[21] Die Trennschärfe eines Items gibt – inhaltlich dargestellt – an, wie gut das Item zwischen Personen mit hohen und Personen mit niedrigen Testergebnissen unterscheidet. Rechnerisch wird ein entsprechender Itemkennwert in der Regel durch eine Korrelation zwischen Itemergebnissen und (Gesamt-)Testergebnis ermittelt (vgl. Lienert & Raatz, 1998, S. 78 ff.).

[22] Dass diese Annahme restriktiv ist, wird schon daran deutlich, dass die *KTT* viele Vorschläge und Verfahren bereitstellt, wie mit unterschiedlichen Itemtrennschärfen innerhalb eines Tests bzw. bei der Itemauswahl für einen Test umzugehen ist.

bezeichnet. Diese Eigenschaft ist für die Testpraxis, vor allem auch für die Schulleistungsstudien, von großer Bedeutung.

- Die Anzahl der richtig gelösten Items enthält alle im *RM* relevanten Informationen über das Testverhalten einer Testperson. Weitere Daten, wie z. B. das konkrete Bearbeitungsmuster („Welche Items wurden richtig / falsch gelöst?“), liefern innerhalb des *RM* keine weitere Information. Die Statistik „Anzahl richtig gelöster Items für jede Versuchsperson“ ist in diesem Sinne eine „erschöpfende“ oder „suffiziente Statistik“.
- Das *RM* bietet – u. a. aufgrund dieser suffizienten Statistik – gute Möglichkeiten, die Parameter des Modells (Itemschwierigkeiten und Personenparameter) im Rahmen der Skalierung zu schätzen.

Varianten und Verallgemeinerungen des Rasch-Modells

Wenn das ebenso restriktive wie vorteilhafte (eindimensionale zweikategorielle) *RM* nicht zu den Itemformaten passt (keine Richtig-Falsch-Auswertung) oder für die Erklärung der erhobenen Testdaten nicht ausreicht, dann gibt es verschiedene Möglichkeiten der Variation bzw. Verallgemeinerung des Testmodells, die im Folgenden skizziert werden.

- Neben psychometrischen Tests, bei denen die Itemantworten dichotom (in der Regel mit *richtig* bzw. *falsch*) kodiert werden, gibt es auch Beispiele für Tests und Fragebögen, bei denen mehr als zwei (geordnete) Kategorien für die Kodierung verwendet werden. Bei einer Anpassung des *RM* auf diese Formate bleiben seine statistisch vorteilhaften Eigenschaften im Wesentlichen erhalten.
  - Bei Leistungstests ist dies etwa der Fall, wenn ein komplexeres Item außer *richtigen* und *falschen* Bearbeitungen auch noch *teilweise richtige* Bearbeitungen (auch „partial credits“ genannt) ermöglicht, die mit Blick auf die zugrundeliegende Personeneigenschaft schwieriger sind als falsche Bearbeitung, aber einfacher als richtige Bearbeitungen. Ggf. lassen sich auch noch weitere Zwischenkategorien bestimmen, die zusammen eine sinnvolle (ordinale) Abstufung zwischen *richtig* und *falsch* ergeben.[23] Der logistische Ansatz des *RMs* lässt sich (a) auf „ordinale Items“ anpassen und (b) auf Tests anwenden, die sowohl *ordinale Items* als auch *dichotome Items* enthalten.
  - Insbesondere Fragebögen, die Personeneigenschaften über Selbstauskünfte erfassen, oder Tests, die Einstellungen zu bestimmten Themen erheben, arbeiten häufig mit sogenannten „Ratingskalen“. Bei diesem Itemformat sollen die Versuchspersonen

[23] Die Möglichkeit, teilweise richtige Bearbeitungen mit im Testmodell zu verarbeiten, kann entscheidend zur Akzeptanz von Leistungstests in der Schulpraxis beitragen: Lehrkräfte kritisieren an Tests mit ausschließlich *dichotomen Items* häufig (aus pädagogischen Gründen), dass Leistungen, die hinter teilweise richtigen Bearbeitungen stecken, nicht honoriert werden können.

auf einer *n*-stufigen Skala[24] das Ausmaß der Zustimmung oder Ablehnung zu bestimmten Aussagen angeben oder einschätzen, inwieweit ein Statement auf sie zutrifft. Bei Ratingskalen kreuzen die Versuchspersonen also bereits die (geordnet vorliegenden) Antwortkategorien an, die zu Kodierung verwendet werden. Das *RM* für Ratingskalen lässt sich aus dem Modell für *ordinale Items* (s. o.) durch Restriktionen ableiten. Diese Restriktionen formalisieren Annahmen über das Antwortverhalten bei Ratingskalen bzw. die inhaltlich Bedeutung der Antwortkategorien von Ratingskalen. Im empirischen Teil der vorliegenden Arbeit ist das *RM* für Ratingskalen von Bedeutung, da das *bereichsspezifische Fähigkeitsselbstkonzept Mathematik* mit einer vierstufigen Ratingskala erfasst wird.

- Bei der Anwendung des eindimensionalen zweikategoriellen *RM* wird häufig kritisiert, dass komplexere Strukturen, die in den Items stecken oder die Personenfähigkeit charakterisieren, unberücksichtigt bleiben. Wenn tatsächlich hinreichend konkrete theoretische Vorannahmen für die Strukturen existieren, lässt sich die Kernidee des *RM* hierauf anpassen, wobei wesentliche statistisch vorteilhafte Eigenschaften erhalten bleiben.
  - Geht man davon aus, dass die Personeneigenschaft ein eindimensionales latentes Konstrukt ist und die Itemschwierigkeit sich (linear) aus *m* verschiedenen Itemkomponenten zusammensetzt, so kann der Itemparameter σ_i linear in entsprechende „Basisparameter“ (der Itemschwierigkeit) η_j für die fraglichen Komponenten zerlegt werden: $\sigma_i = \sum_{j=1}^{m} q_{ij} \cdot \eta_j$. Setzt man die Zerlegung in die Modellgleichung des (eindimensionalen zweikategoriellen) *RMs* ein, so ergibt sich die Gleichung des (zweikategoriellen) „Linear-logistischen Testmodells (LLTM)“: $p_{i,a} = \frac{\exp\left(a \cdot \left(\theta - \sum_{j=1}^{m} q_{ij} \cdot \eta_j\right)\right)}{1 + \exp\left(\theta - \sum_{j=1}^{m} q_{ij} \cdot \eta_j\right)}$.

    Die theoretischen Vorannahmen müssen beim *LLTM* für alle Items eines Tests festlegen, mit welchem Gewicht q_{ij} die Komponenten η_j in das Item *i* eingehen. Augenscheinlich stellt dieses Testmodell hohe Ansprüche an die theoretische Modellbildung. Gelingt diese Modellbildung allerdings, so hat sie den Vorteil, dass anstelle der möglicherweise zahlreichen verschiedenen Itemparameter σ_i nur noch die (in der Regel wenigen) Basisparameter η_j aus den Testdaten geschätzt werden müssen. Eine Verallgemeinerung dieses Testmodells für mehr als zwei (geordnete) Antwortkategorien ist mit den gleichen Überlegungen und Rechnungen wie beim (eindimensionalen) *RM* für *ordinale Items* möglich.

[24] Häufig werden vier Kategorien verwendet, weil bei größerer Kategorienanzahl die Entscheidung für eine Kategorie und gegen die Nachbarkategorien immer unschärfer wird. Bei ungeraden Kategorienanzahlen wird häufig die Mitte bevorzugt. Letztlich muss die Kategorienzahl aber bei jedem Fragebogen bzw. Test neu mit Blick auf Forschungsgegenstand und die zu untersuchende Population festgelegt werden.

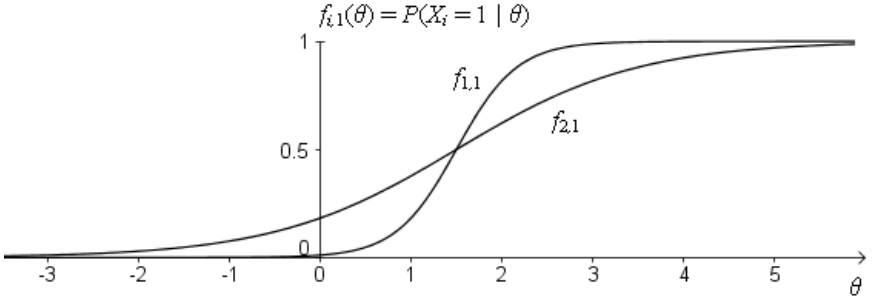
- Eine ähnliche Überlegung wie bei der Zerlegung der Itemschwierigkeiten in Komponenten führt bei den Personenparametern zum mehrdimensionalen (zweikategoriellen) *RM*. Der Personenparameter θ wird dabei linear in „Basisparameter“ θ_j zerlegt: $\theta = \sum_{j=1}^{m} q_{ij} \cdot \theta_j$. So kann die Personeneigenschaft mehrdimensional modelliert werden.

  Geht man z. B. davon aus, dass *Mathematikleistung* im Wesentlichen auf die drei Komponenten „Rechenfertigkeit“ (mit Personenparameter θ_1), „Begriffliches Wissen“ (θ_2) und „Problemlösefähigkeit“ (θ_3) zurückgeführt werden kann, dann müssen für jedes Item *i* die Gewichte q_{i1} bis q_{i3} (theoretisch) festgelegt werden, mit denen die drei Komponenten bei diesem Item vermutlich aktiviert werden müssen. Die Modellgleichung für dieses mehrdimensionale (zweikategorielle) *RM* ergibt sich wieder durch Einsetzen: $p_{i,a} = \frac{\exp\left(a \cdot \left(\sum_{j=1}^{m} q_{ij} \cdot \theta_j - \sigma_i\right)\right)}{1 + \exp\left(\sum_{j=1}^{m} q_{ij} \cdot \theta_j - \sigma_i\right)}$. Auch für dieses Testmodell ist eine Verallgemeinerung auf mehr als zwei (geordnete) Antwort möglich.

- Während bei den zuvor dargestellten Varianten des *RM* die Testmodelle im Wesentlichen mit einem (ggf. linear zerlegten) Itemparameter ausgekommen sind, es sich also weiterhin um *1-parametrige, logistische Modelle* handelt, berücksichtigen Verallgemeinerungen des *RM* weitere Itemparameter. Diese Testmodelle können dann z. B. unterschiedliche Formen von *ICCs* berücksichtigen, indem sie „Trennschärfeparameter“ oder „Rateparameter“ einführen. Im Folgenden werden die Verallgemeinerungen des *RM*/*1-PL* auf logistische Modelle mit zwei bzw. drei Parametern (*2-PL* bzw. *3-PL*) skizziert. Dabei werden jeweils die Modelle für *dichotome Items* betrachtet – eine Verallgemeinerung auf mehr als zwei (geordnete) Antwortkategorien ist wie oben möglich.

  - Eine übliche Kritik am *RM* bezieht sich auf den parallelen Verlauf der *ICCs* mit dergleichen Trennschärfe für alle Items. Mathematisch lässt sich diese Restriktion leicht aufheben, wenn man einen *Trennschärfeparameter* in das *RM* einführt. Betrachtet man zwei *ICCs* für die Kategorie „1“ mit gleicher Itemschwierigkeit, aber freigegebenen Trennschärfen, so ergibt sich graphisch z. B. das folgende Bild (Abb. 2.4):

Abbildung 2.4: *ICCs* mit unterschiedlichen Trennschärfen



Augenscheinlich muss der fragliche Trennschärfeparameter eine Stauchung bzw. Streckung in horizontaler Richtung bewirken können, ohne die Verortung des Items auf dem gemeinsamen Kontinuum der Itemschwierigkeit und der Ausprägung der Personeneigenschaft zu verändern. Diese Lokalisierung erfolgt über die Stelle, an der die Wahrscheinlichkeit für eine richtige Antwort 0,5 beträgt. Rechnerisch führen diese Überlegungen zum folgenden (zweikategoriellen) *2-parametrigen, logistischen Testmodell* (2-PL oder auch „Birnbaum-Modell" genannt): $p_{i,a} = \frac{\exp\left(a \cdot \beta_i \cdot (\theta - \sigma_i)\right)}{1 + \exp\left(\beta_i \cdot (\theta - \sigma_i)\right)}$.

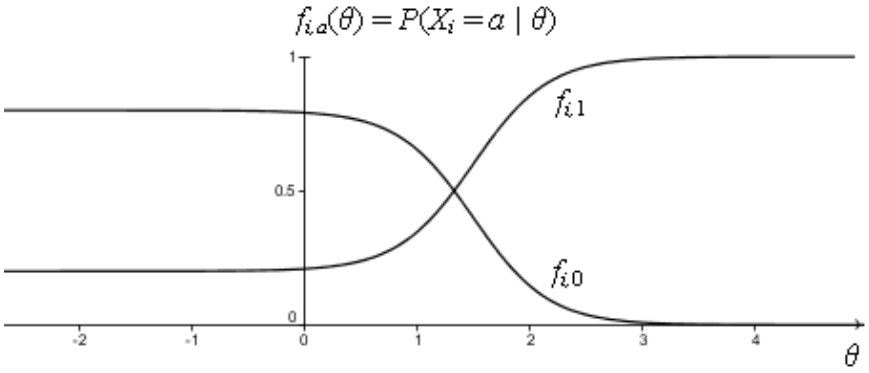
Innerhalb dieses Modells lässt sich nicht mehr global sagen, ob ein Item einfacher oder schwieriger ist als ein anderes, da dies von der konkreten Ausprägung der Personeneigenschaft abhängt.

- Möchte man über unterschiedliche Trennschärfen hinaus auch die mögliche Ratewahrscheinlichkeit für Items berücksichtigen, gelingt dies mathematisch einfach durch Einführung eines *Rateparameters als* dritten Itemparameter. Die Frage, ob die Modellierung des Testverhaltens mit einem *Rateparameter* angemessen ist, lässt sich kaum theoretisch, sondern meistens eher empirisch beantworten.[25] Wird ein solcher Parameter im Testmodell berücksichtigt kann er aus den Testdaten mitgeschätzt werden – und, falls er sich nicht signifikant[26] von Null unterscheidet, wieder eliminiert werden. Graphisch drückt sich die Berücksichtigung eines Rateparameters darin aus, dass die *ICCs* für richtige Antworten sich für geringe Ausprägungen der Personenfähigkeit nicht der horizontalen Koordinatenachse, sondern einer hierzu parallelen (nach oben verschobenen) Geraden annähern. Die *ICC*s für falsche Antworten nähern sich dementsprechend nicht der horizontalen Geraden durch den Punkt (0 | 1), sondern einer nach unter verschobenen parallelen Geraden. Die Abbildung 2.5 veranschaulicht auch den schlichten Sachverhalt, dass sich bei einem zweikategoriellen Testmodell die Wahrscheinlichkeiten für die richtige Antwort und für die falsche Antwort für ein Item und eine Ausprägung der Personeneigenschaft immer zu Eins addieren, da sie Gegenwahrscheinlichkeiten zueinander sind.

[25] Dies hängt jeweils vom konkreten Test und den Versuchspersonen ab. So gibt es zahlreiche Tests bei denen schwierige Aufgaben Lösungswahrscheinlichkeiten unterhalb der Ratewahrscheinlichkeit aufweisen. Haben die Testergebnisse keine Konsequenzen für die Versuchsperson und findet der Test nicht unter Zeitdruck, also als „Power-Test" statt, dann gibt es kaum Anlass zum Raten. Darüber hinaus lassen sich selbst bei Multiple-Choice-Items oft derart „attraktive" Falschantworten finden, die z. B. typische Fehlvorstellungen abbilden, dass diese mehr oder weniger zielgerichtet gewählt und nicht geraten werden. Theoretisch-inhaltliche und empirisch-technische Fragen zum Raten bei *PISA*-Tests werden in den aufeinander bezogenen Diskussionsbeiträgen von Meyerhöfer (2004), Lind (2004), Woschek (2004) und Meißner (2004) erörtert.

[26] Das Konzept der Signifikanz wird in der vorliegenden Arbeit so verwendet, wie es in der empirischen Bildungsforschung überwiegend üblich ist, d. h. im Sinne eines einfaches Hypothesentests (vgl. Büchter & Henn, 2007, Kap. 4.2) mit einem Signifikanzniveau von 5 %.

Abbildung 2.5: *ICCs* für die richtige bzw. falsche Antwort mit Berücksichtigung der Ratewahrscheinlichkeit



Rechnerisch führt die Berücksichtigung eines Rateparameters zum (zweikategoriellen) *3-PL*: $p_{i,a} = a \cdot \gamma + (1-\gamma) \cdot \frac{\exp\left(a \cdot \beta_i \cdot (\theta - \sigma_i)\right)}{1 + \exp\left(\beta_i \cdot (\theta - \sigma_i)\right)}$.

Auf den ersten Blick scheinen die verallgemeinerten Testmodelle *2-PL* und *3-PL* attraktiver und leistungsfähiger zu sein als das *RM*/*1-PL*, weil sie sich bestimmten – auch empirisch feststellbaren – Eigenarten von Testitems besser anpassen. Allerdings „erkauft" man sich die größere Flexibilität durch doppelt bzw. dreimal so viele Itemparameter, die aus den Testdaten geschätzt werden müssen, und durch den Verlust wünschenswerter statistischer Eigenschaften des *RM*/*1-PL* (wie der *spezifischen Objektivität* und der *suffizienten Statistiken*). J. Rost (2004) führt aus, dass dieses Verhältnis zwischen *RM*/*1-PL* einerseits und *2-PL* bzw. *3-PL* andererseits zu einer Art Schulenstreit geführt hat:

> „Die Anhänger des Rasch-Modells betonen die vorteilhaften statistischen Eigenschaften des 1-pl Modells, die bei den mehrparametrigen Modellen verloren gehen, und argumentieren, dass mit dem Rasch-Modell mehr als nur irgendein *Testmodell* definiert ist, sondern eine eigene *Messtheorie* begründet wird – die Theorie spezifisch objektiver Messungen" (S. 133; Herv. i. O).

Grundsätzlich kann man davon ausgehen, dass bei aktuellen Schulleistungsstudien *spezifisch objektive Messungen* die testtheoretische Grundlage bilden.

Die Frage der Schätzung der Modellparameter im Rahmen der *IRT*-Skalierung wird an dieser Stelle nicht vertieft, sondern nur erwähnt: Ausgehend vom jeweiligen Testmodell werden Itemparameter und Personenparameter anhand der beobachteten Daten geschätzt.[27] Darüber hinaus muss überprüft werden, ob die Voraussetzungen der Testmodelle erfüllt

[27] Diese Parameter sind prinzipiell unbekannt, können aber mithilfe der beobachteten Daten geschätzt werden. Im Sinne der mathematischen Statistik handelt es sich um Varianten von *Maximum-Likelihood-Schätzungen* (vgl. Büchter & Henn, 2007, Kap. 4.1), die als iterative Rechenalgorithmen implementiert werden. Die Bestimmung der Modellparameter ist als *Schätzung* stets mit (kontrollierter) Unsicherheit behaftet.

sind. Konkurrierende Modelle (z. B. eindimensional vs. mehrdimensional) lassen sich, wenn ein Modell durch Restriktionen aus dem anderen Modell hervorgeht, durch bestimmte Varianten einfacher Hypothesentests oder, andernfalls, auf der Basis von informationstheoretischen Maßen miteinander vergleichen (vgl. Kap. 4.2.3). Auf diese und weitere Aspekte der konkreten Modellierung von Testdaten geht z. B. Rost (2004) ausführlich ein.

2.1.3 Kompetenzmodelle als Grundlagen und als Befunde der Bildungsforschung

Spätestens seit Beginn der Berichtslegung zu *PISA 2000* (Deutsches PISA-Konsortium, 2001) werden die primär interessierenden Konstrukte von Schulleistungsstudien als Kompetenzen bezeichnet. Noch bis in die späten 1990er-Jahre war die übliche Bezeichnungsweise Fachleistungen (z. B. bei der Berichtslegung zu *TIMSS/II*, vgl. Baumert & Lehmann, 1997). Tatsächlich ist, wie im Folgenden dargestellt wird, mit dem Bezeichnungswechsel auch ein begrifflicher Wandel verbunden. Der Begriff *Kompetenz* fokussiert eher auf die Bewährung von Individuen in Problemsituationen (vgl. Bruder et al., 2008, S. 10 ff.), während *Fachleistung* stark an schulische oder schulähnliche Leistungssituationen gebunden ist. Sowohl die verschiedenen Fachdidaktiken als auch die anderen an der empirischen Bildungsforschung beteiligten Disziplinen beziehen sich dabei überwiegend auf den Kompetenzbegriff, den Weinert (2001) – basierend auf seinen umfangreichen Arbeiten im pädagogischen Bereich – formuliert hat. Er definiert Kompetenzen als

> „die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können“ (S. 27 f.).

So verstanden umfassen Kompetenzen deutlich mehr als den Dreiklang „Kenntnisse, Fertigkeiten und Fähigkeiten“. Neben der erweiterten Sichtweise ist bei diesem Kompetenzbegriff besonders wichtig, dass er die Bewährung des Individuums in konkreten Anforderungssituationen als Zielsetzung für (schulisches) Lernen impliziert. Mit Blick auf die Ausführungen in Kap. 2.1.1 zu „proficiency“ und „achievement“ betont der Kompetenzbegriff den *Proficiency-Ansatz*, während *Fachleistung* eher zum *Achievement-Ansatz* gehört.[28] Dementsprechend ist für die auf dem obigen Kompetenzbegriff fußenden *PISA*-Studien nicht curriculare Validität das Leitprinzip bei der Testentwicklung, sondern es werden fachliche Grundbildungskonzepte zugrunde gelegt. Über die bildungstheoretischen und fachdidaktischen Grundlagen der *PISA*-Studien ist der Kompetenzbegriff auch zur fachübergreifenden Grundlage der Entwicklung der KMK-Bildungsstandards geworden (vgl.

[28] Ohnehin ist der Begriff „Fachleistung“ an einen institutionellen Kontext gebunden, da nur innerhalb von entsprechenden Bildungseinrichtungen „Fächer“ existieren und dort inhaltliche und organisatorische Zwecke erfüllen. Inhaltlich werden Fächer durch Curricula konstituiert, sodass die Nähe von „Fachleistung“ zum „Achievement-Ansatz“ fast zwangsläufig ist.

Klieme et al., 2003). Dies wird in den Bildungsstandards Mathematik (KMK, 2004, 2005a, 2005b) vor allem an der expliziten Berücksichtigung und Beschreibung von sechs sogenannten „allgemeinen mathematischen Kompetenzen“ sichtbar, die implizit die zugehörigen mathematischen Tätigkeiten betonen.

Der Anspruch von aktuellen Schulleistungsstudien wie *PISA* ist also, Kompetenzen auf der Basis von Grundbildungskonzepten zu messen. Ob die zurzeit verwendeten *Paper and Pencil Tests* tatsächlich in der Lage sind, entsprechend weit gefasst Konstrukte zu erfassen, bleibt dabei umstritten (vgl. z. B. Jablonka, 2007). Dass dieser große Anspruch womöglich nur etwas bescheidener eingelöst werden kann, ist allerdings auch den Protagonisten der Schulleistungsstudien klar. So schreibt die „Deutsche PISA-Expertengruppe Mathematik, PISA 2000“ bei Ausführungen zu „Kompetenzstufen“, dass „eigentlich … die Bezeichnung *Leistungsstufe* angebracht [wäre]“ (Knoche et al., 2002, S. 171; Herv. i. O; Erg. d. d. Verf.). Analog dazu wird in der fachdidaktischen und bildungspolitischen Diskussion auch immer wieder angemerkt, dass es sich bei den Bildungsstandards eigentlich um *Fachleistungsstandards* handelt.

Kompetenzbereichsmodelle als Grundlage der Messung von Mathematikleistung

Ein Ausgangspunkt für die Erfassung von Fachleistungen ist die Festlegung eines entsprechenden Konstrukts. Aktuelle Schulleistungstests sind – auch wenn dies zuweilen anders behauptet wird – kein Sammelsurium von Aufgaben zur Leistungsüberprüfung, die zufällig ausgewählt wurden und in einer ungeklärten Beziehung zum Fach stehen, sondern basieren auf systematischen Überlegungen. Dies wird im Folgenden exemplarisch für die internationale Erhebung und die nationale Ergänzung bei *PISA 2000* dargestellt. Auf abstrakter Ebene resultieren aus diesen Überlegungen sehr verdichtete Beschreibungen des Konstrukts:

> „Nach der PISA-Konzeption gehört zur mathematischen Grundbildung ein Verständnis der Rolle, die Mathematik in der sozialen, kulturellen und technischen Welt spielt, und die Fähigkeit, Sachverhalte unter mathematischen Gesichtspunkten angemessen zu beurteilen. Mathematische Grundbildung schließt aber auch die Fähigkeit ein, Mathematik aktiv zu nutzen, um Anforderungen des Alltags zu bewältigen“ (Baumert et al., 2001, S. 15).

Diese Zielvorstellung soll umgesetzt werden, indem ein Kompetenzbereichsmodell entwickelt wird, das die wesentlichen Bereiche mathematischer Kompetenz berücksichtigt. Basierend auf einem solchen Modell kann die fachliche Ausgewogenheit des Tests (durch eine Zuordnung von Aufgaben zu den Kompetenzbereichen) überprüft und (durch zielgerichtetes Entwickeln von Aufgaben aus den Kompetenzbereichen heraus) sichergestellt werden. Das Kompetenzbereichsmodell der internationalen Erhebung der *Mathematikleistung* bei *PISA 2000* berücksichtigt vor allem die typischen mathematischen Tätigkeiten als „competencies“ und zentrale mathematische Inhalte als „big ideas“ (vgl. Neubrand et al., 2001). Im internationalen Konzept von *PISA 2000* sind die *competencies* besonders wich-

tig.[29] Ausgehend von einer ausführlichen Kompetenzliste werden drei stark verdichtete „Kompetenzklassen“ gebildet:

- Class 1: reproductions, definitions, and computations
- Class 2: connections and integration for problem solving
- Class 3: mathematical thinking, generalisation and insight

Die mathematischen Inhalte werden im internationalen Konzept von *PISA 2000* mit zwei *big ideas*, nämlich „change and growth“ und „shape and space“, nur exemplarisch berücksichtigt, da Mathematik erst bei *PISA 2003* Untersuchungsschwerpunkt ist. Das internationale Konstrukt bei *PISA 2000* kann also wie folgt charakterisiert werden: Primär für die Erfassung mathematischer Kompetenz ist möglichst vollständige Berücksichtigung der typischen mathematischen Tätigkeiten, die sich an exemplarisch ausgewählten inhaltlichen Schwerpunkte konkretisieren. Bei den Testaufgaben werden über *competencies* und *big ideas* hinaus noch weitere Merkmale berücksichtigt, so z. B. Kontexte (innermathematisch vs. außermathematisch) und Situationen („Nähe“ der Aufgabenstellung zu den Schülerinnen und Schülern).

Die nationale Ergänzungsstudie im Rahmen von *PISA 2000* strebt interessanterweise zusätzlich – zumindest implizit – curriculare Validität an. Die deutsche Expertengruppe Mathematik schreibt hierzu: „Unterschiede [zwischen internationalem und nationalem Konzept] ergeben sich jedoch, wenn Differenzierungen so vorgenommen werden, dass Gegebenheiten des deutschen Mathematikunterrichts abgebildet werden können“ (ebd., S. 51; Erg. d. d. Verf.). So werden bei der Ausdifferenzierung der Kompetenzklassen z. B. explizit „Technische Fertigkeiten“ – eine lokale Stärke deutscher Schülerinnen und Schüler (vgl. BLK, 1997, S. 15) – berücksichtigt. Bei den mathematischen Inhalten orientiert sich das Konzept der nationalen Ergänzungsstudie explizit an den Themengebieten der deutschen Curricula und unterscheidet für den Test *Arithmetik*, *Proportionalität*, *Algebra*, *Geometrie* sowie *Stochastik und Umgehen mit Daten* (vgl. Neubrand et al., 2001, S. 55). Auch in der nationalen Ergänzungsstudie werden bei der Aufgabenentwicklung weitere (zum Teil schwierigkeitsbestimmende) Merkmale explizit berücksichtigt.

Auf der Grundlage von Kompetenzbereichsmodellen und zusätzlichen Aufgabenmerkmalen können, wie oben dargestellt wurde, fachdidaktisch klar strukturierte Tests entwickelt werden. Wenn die zugrundegelegten Kompetenzbereiche und Aufgabenmerkmale dabei relevante Dimensionen des Fachs möglichst vollständig klassifizieren, sind solche Tests nicht nur klar strukturiert, sondern auch fachlich hinreichend repräsentativ.

[29] Anders als bei den *big ideas*, die nur einen Ausschnitt der Schulmathematik der Sekundarstufe I darstellen, wird bei den *competencies* eine möglichst vollständige Berücksichtigung der typischen mathematischen Tätigkeiten angestrebt.

Kompetenzstufenmodelle als Ergebnis der Messung von Mathematikleistung

Während Kompetenzbereichsmodelle eine theoretische Grundlage der Erfassung von *Mathematikleistung* darstellen, sind Kompetenzstufenmodelle[30] ein Ergebnis von Schulleistungsstudien. Bei einem Kompetenzstufenmodell für ein Fach oder einen Bereich eines Fachs wird eine eindimensionale Leistungsskala in verschiedene Abschnitte, die Kompetenzstufen, unterteilt. Diese Abschnitte sollen inhaltlich möglichst gut beschreibbar sein. Praktisch wird eine inhaltliche Beschreibung aus typischen Merkmalen der Items gewonnen, die „auf dieser Stufe liegen“ (s. u.).

Aus forschungsmethodischer Sicht bieten sich *IRT*-Modelle, vor allem verschiedene Varianten des *RM*s, für die Gewinnung solcher Kompetenzstufenmodelle an. In einem ersten Schritt wird hierfür der fragliche Mathematiktest skaliert, das heißt, mithilfe eines geeigneten Testmodells und geeigneter Schätzverfahren werden aus den Testdaten Messwerte für die *Mathematikleistung* der Versuchspersonen gewonnen. Die für die Kompetenzstufenmodelle erforderliche eindimensionale Leistungsskala kann auf verschiedenen Wegen gewonnen werden:

- Bereits bei der Entwicklung des Mathematiktests wird die Eindimensionalität angestrebt. Dann werden potenzielle Testaufgaben in Voruntersuchungen auf der Grundalge eines eindimensionalen Testmodells (z. B. ein eindimensionales *RM*) analysiert und nicht ins Modell passende Aufgaben anschließend entfernt. Die übrig bleibenden Aufgaben bilden dann einen eindimensionalen Mathematiktest. Dieses pragmatische Vorgehen kann aus fachdidaktischer Sicht dann problematisch sein, wenn Aufgaben auf diesem Weg entfernt werden müssen, die curricular oder bildungstheoretisch wünschenswert erscheinen. Bei sehr heterogenen Populationen scheint ein eindimensionaler Mathematiktest aber auch ohne Verlust von zu vielen wünschenswerten Aufgaben möglich zu sein (vgl. Klieme et al., 2001, S. 156).
- Die Frage der Dimensionalität kann bei der Testentwicklung zunächst aber auch offen gelassen werden. Dann wird auf der Basis eines Kompetenzbereichsmodells ein fachlich ausgewogener Test entwickelt und im Rahmen von Voruntersuchungen oder der eigentlichen Studie die Dimensionalität empirisch untersucht. So können die Testdaten z. B. mit einem eindimensionalen *RM* und parallel mit einem mehrdimensionalen *RM* skaliert werden.[31] Anschließend kann auf der Basis sogenannter „Modellgeltungstests“[32] (vgl. J.

[30] Auch hier lässt sich trefflich streiten, ob es sich tatsächlich um *Kompetenz*stufen oder eher um *Performanz*- oder *Leistung*sstufen handelt. In der Literatur ist es aber üblich, die Bezeichnung „Kompetenzstufenmodelle“ zu verwenden.

[31] Anders als z. B. bei der explorativen Faktorenanalyse der *KTT* müssen hier aber vorab Annahmen getroffen werden, welche Dimensionen existieren und wie stark die einzelnen Items diese Dimensionen „bedienen“. Die Notwendigkeit für diese theoretische Vorarbeit wird bei den Ausführungen zum mehrdimensionalen *RM* in 2.1.2 deutlich.

Rost, 2004, Kap. 5) entschieden werden, ob für *Mathematikleistung* eine eindimensionale Skala erzeugt wird oder ob für verschiedene Bereiche der Mathematik (z. B. verschiedene Arten mathematischen Arbeitens oder verschiedene Inhaltsbereiche) separate, dann aber eindimensionale Leistungsskalen gebildet werden.

Nach der Skalierung einer eindimensionalen Leistungsskala – sei es für die gesamte *Mathematikleistung* oder für Teilbereiche – kann man nutzen, dass im *RM* die Aufgabenschwierigkeit und *Mathematikleistung* auf einer Dimension gemessen werden. Bei *PISA 2000* wurden die einzelnen Items jeweils so an den Stellen der Leistungsskala verankert, dass Versuchspersonen mit dem entsprechenden Leistungswert die fraglichen Items mit einer Wahrscheinlichkeit von 0,62 richtig lösen (d. h. „mit hinreichender Sicherheit", Baumert et al., 2001, S. 52). Wenn man anschließend (a) die Verteilung der Versuchspersonen auf die Leistungsskala und (b) die verorteten Items betrachtet, dann gibt es generell zwei Wege, zu Kompetenzstufen zu gelangen:

- Bei den nordrhein-westfälischen Lernstandserhebungen der Jahre 2004 bis 2008 (*LSE 9*) wurde analysiert, mit welchen Aufgabenmerkmalen die Verortung der Items auf der Leistungsskala erklärt werden kann. Anschließend wurden Abschnitte auf der Leistungsskala so gebildet, dass die Items innerhalb eines Abschnitts inhaltlich möglichst ähnlich zu einander sind und sich inhaltlich möglichst deutlich von den Items in den anderen Abschnitten unterscheiden. Dabei durften die Kompetenzstufen unterschiedlich große Abschnitte der Leistungsskala umfassen (vgl. Fleischer et al., 2007, S. 104 ff.). Dieses Vorgehen lässt sich als „inhaltliche Clusterbildung entlang der Leistungsskala durch Experten" beschreiben. Grundlegend ist hierbei die Analyse der Testaufgaben, die sich auch in der Beschreibung der *Kompetenzstufen* widerspiegelt. Bei den *LSE 9* im Jahr 2004 wurde so ein Kompetenzstufenmodell zum Testschwerpunkt *Modellieren* entwickelt. Dabei wurden im Wesentlichen vier Stufen, im Original als „Niveau" bezeichnet, unterschieden (vgl. Heymann & Pallack, 2007, S. 36 ff.):
  - Kompetenzniveau M 1 (Vorstufe zum Modellieren) – Kontextbezogenes Rechnen
  - Kompetenzniveau M 2 – Elementares Situationsverständnis
  - Kompetenzniveau M 3 – Einschrittiges Modellieren
  - Kompetenzniveau M 4 / M4 plus – Komplexes Modellieren

  Exemplarisch wird für „M 2 – Elementares Situationsverständnis" wiedergegeben, wie die Aufgabenanalyse zur Beschreibung der Kompetenzstufe beiträgt:

[32] Diese Bezeichnung ist insofern irreführend, als Modelle nicht richtig oder falsch, also im eigentlichen Sinne nicht „gültig" oder „ungültig" sein können. Modelle können unter Berücksichtigung eines Verwendungszwecks immer nur mehr oder weniger gut zur Realität, die sie beschreiben sollen, passen. Die sogenannten Modellgeltungstests gehen inhaltlich auch genau so vor: Sie berücksichtigen, wie gut ein Modell bzw. welches Modell die erhobenen Daten unter bestimmten Randbedingungen besser vorhersagt.

> „Aufgaben dieses Niveaus stellen Anforderungen wie: (1) aus einfachen Texten und Abbildungen mathematische Informationen entnehmen, (2) durch die Aufgabenstellung direkt nahe gelegte Lösungsansätze finden und die zugehörigen Rechnungen durchführen. Zur Lösung der Aufgaben müssen die Zahlinformationen lediglich einmalig verarbeitet werden (Verknüpfung von zwei bis drei Zahlen mit Hilfe der Grundrechenarten). Die rechnerischen Lösungen müssen nicht notwendig am Kontext überprüft werden (Validieren)“ (ebd., S. 37 f.)

- Einen anderen Weg der Kompetenzstufenbildung beschreitet *PISA 2000*, *2003*, *2006*. Hier werden die Kompetenzstufen nach statistischen Kriterien gebildet (vgl. Klieme et al., 2001, S. 147). So werden außer einer nach oben und einer nach unten offenen Stufe ausschließlich gleichlange Kompetenzstufen gebildet, wobei die inhaltliche Nähe der dort verankerten Items zueinander zunächst keine Rolle spielt. Außer der äquidistanten Einteilung können weitere statistische Anforderungen an die Kompetenzstufen gestellt werden, z. B. dass alle Versuchspersonen, deren Leistungswerte auf einer bestimmten Kompetenzstufe liegen, die Items, die auf dieser Kompetenzstufe verankert sind, mindesten mit einer gewissen Wahrscheinlichkeit lösen können.[33] Der „technical report“ zu jedem *PISA*-Durchgang (z. B. OECD, 2009) enthält eine Reihe statistischer Kriterien, die bei der Bildung von *PISA*-Kompetenzstufen berücksichtigt werden. Dieses Vorgehen lässt sich als „technische Clusterbildung entlang der Leistungsskala nach statistischen Kriterien“ beschreiben. Für die inhaltliche Beschreibung der Kompetenzstufen werden aber, wie im anderen Modell der Stufenbildung auch, die Aufgaben herangezogen und analysiert, die sich auf der jeweiligen Stufe befinden.

Wenn auf dem einen oder dem anderen Wege ein Kompetenzstufenmodell für *Mathematikleistung* bzw. eine Komponente der *Mathematikleistung* konstruiert wurde, dann ist es in der Regel aufgrund seines Zustandekommens „kumulativ“. Statistisch ist dies aufgrund der Skalierung klar: Versuchspersonen, die aufgrund ihrer Testleistung einer mittleren Stufe zugeordnet werden, können Items von niedrigeren Stufen mit hoher Wahrscheinlichkeit lösen und umgekehrt Items von höheren Stufen mit niedriger Wahrscheinlichkeit lösen. Inhaltlich sollte diese Kumulativität der Leistungsskala dadurch zum Ausdruck kommen, dass die Beschreibung der höheren Stufen erkennen lassen, dass die Anforderungen auf niedrigeren Stufen sicher beherrscht werden.

Auf der Basis solcher Kompetenzstufenmodelle können Ergebnisse von Schulleistungsstudien oder Vergleicharbeiten berichtet und verglichen werden, ohne dass zu viele einzelne numerische Ergebnisse, die erst wieder inhaltlich interpretiert werden müssen, explizit auftauchen. Die Kompetenzstufen liefern durch ihre inhaltlichen Beschreibungen eine kriteriale Interpretation der Testergebnisse mit. Vergleiche von Klassen, Schulen oder Bildungs-

[33] Dies wird nicht die Wahrscheinlichkeit $p = 0{,}62$ sein, die für die Verankerung der Items auf der Leistungsskala verwendet wurde, da eine Kompetenzstufe einen größere Bereich umfasst und Versuchspersonen am unteren Ende der Stufe die Items am oberen Ende der Stufe mit einer Wahrscheinlichkeit von weniger als 0,62 richtig lösen.

systemen können dann stattfinden, indem man die Verteilung der jeweiligen Gruppe auf die Kompetenzstufen berichtet.

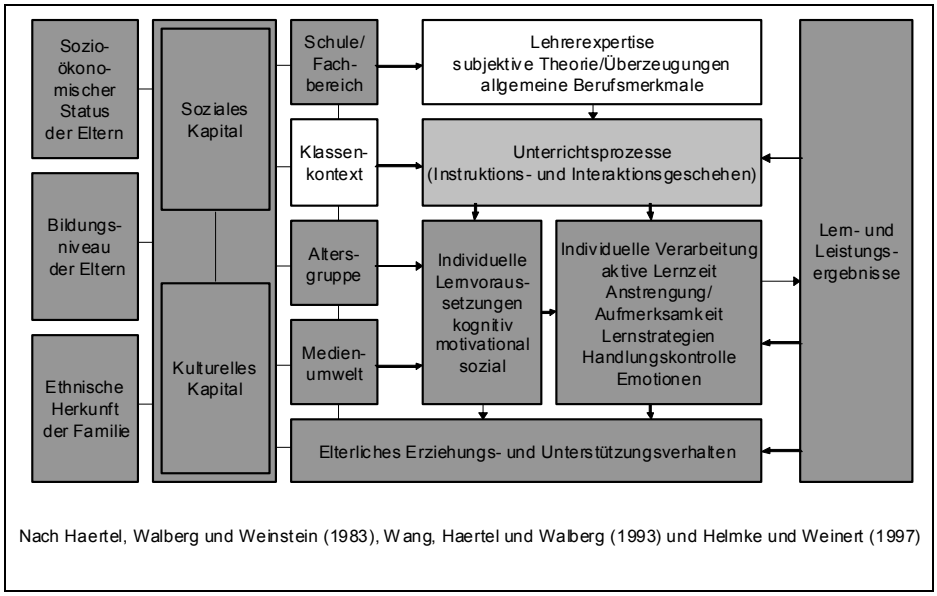
2.1.4 Rahmenmodelle für die Erforschung von Mathematikleistung

Neben dem reinen Vergleich der *Mathematikleistungen* verschiedener Gruppen – seien es Klassen, Schulen oder Bildungssystemen – ist für die Bildungspolitik, die Bildungsforschung, die Fachdidaktiken und die Schulpraxis von großem Interesse, wie solche Leistungen, Leistungsunterschiede und Leistungsentwicklungen zustande kommen, um sachgerechte Konzepte für die Weiterentwicklung der pädagogischen Praxis erarbeiten, implementieren und evaluieren zu können. Dabei bietet sich z. B. das folgende zweistufige Vorgehen an:

- In einem ersten Schritt werden Leistungen, Leistungsunterschiede und Leistungsentwicklungen möglichst gut „statistisch erklärt". Es geht also darum, ein möglichst gutes Vorhersagemodell zu entwickeln, in dem diese Leistungen, Leistungsunterschiede und Leistungsentwicklungen rechnerisch auf andere Faktoren zurückgeführt werden können (z. B. Bildungsalter, sozioökonomischer Status, Unterrichtsgestaltung, …). Auf dieser Basis alleine sollte allerdings keine Interventionsplanung stattfinden, da dieses empirische Vorgehen in der Regel zwar Zusammenhänge zwischen Variablen, nicht aber Wirkungsrichtungen identifizieren kann. Außerdem können auch Wechselwirkungen (z. B. zwischen *Mathematikleistung* der Schülerinnen und Schüler sowie der Gestaltung des Mathematikunterrichts) oder Wirkungen, die über nicht erfasste Variablen vermittelt werden (z. B. *Geschlecht*, *Mathematikleistung* und *Raumvorstellung*), auftreten.
- Im eigentlichen Sinne inhaltlich erklärt werden können Leistungen, Leistungsunterschiede und Leistungsentwicklungen nur, wenn die statistischen Zusammenhänge auch qualitativ und theoretisch abgesichert werden können. Dieser zweite Schritt ist in der Regel erheblich schwieriger als der erste. So kann man im ersten Schritt „ingenieurwissenschaftlich" vorgehen und möglichst viele Variablen und Zusammenhänge erfassen, mit denen die beobachteten Daten insgesamt gut vorhergesagt werden können. Für die inhaltliche Absicherung ist es aber eher wünschenswert, zunächst nur wenige Faktoren zu identifizieren, die nach Möglichkeit nicht nur einen statistisch signifikanten, sondern auch einen praktisch bedeutsamen Einfluss ausüben.

Die Grundlage für die differenzierte Erforschung von Leistungen, Leistungsunterschiedenen und Leistungsentwicklungen ist also in jedem Fall ein theoretisches Rahmenmodell, das potenzielle relevante Variablen und mögliche Wirkungszusammenhänge berücksichtigt, aber noch nicht vollständig spezifiziert sein muss. Bei *PISA 2000* ff. stellt das Rahmenmodell aus Abb. 1.1, das hier noch einmal wiedergegeben wird, eine konzeptionelle Basis für die Instrumentenauswahl und Instrumentenentwicklung dar, die weit über Fachleistungstests hinausgeht (Abb. 2.6).

Abbildung 2.6: „Bedingungen schulischer Leistungen – Allgemeines Rahmenmodell“ (Quelle: Baumert et al., 2001, S. 33)



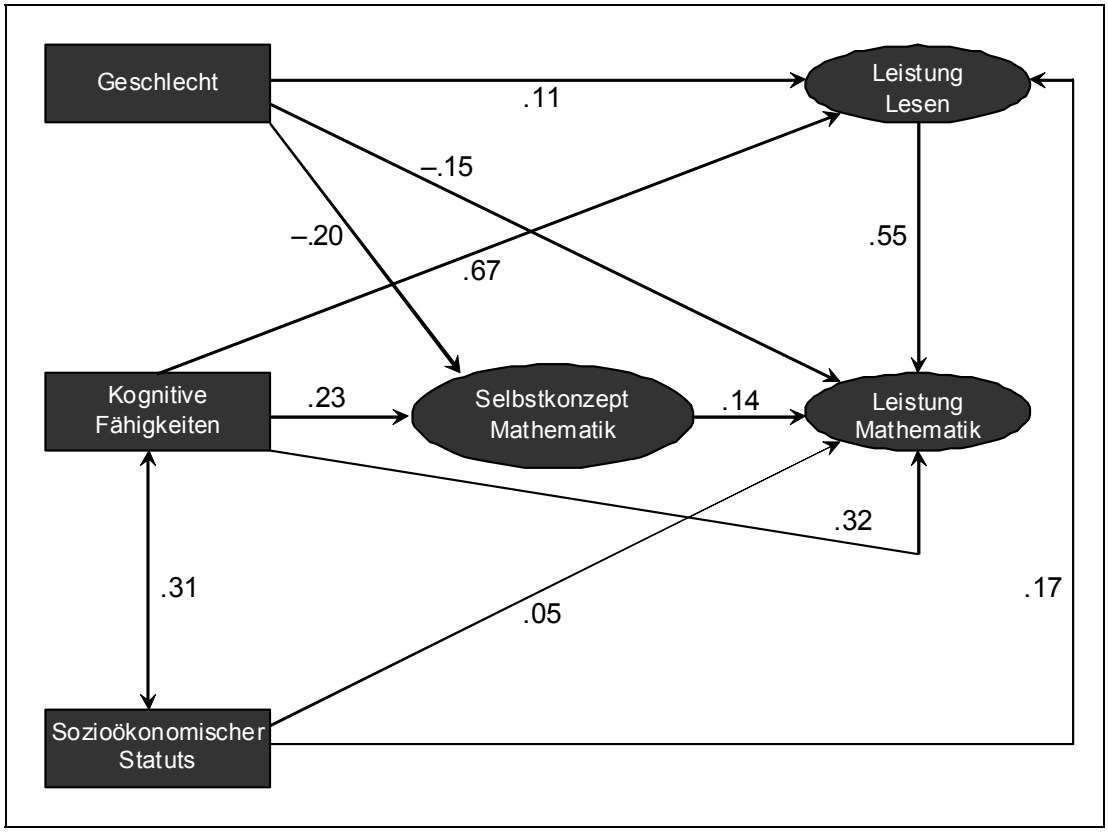
In der Abbildung sind solche Bereiche grau gefärbt, zu denen Daten erhoben werden. Da *PISA 2000* nicht klassen- oder jahrgangsbezogen durchgeführt wurde, sondern die Population der 15-Jährigen im Blick hatte, werden die Bereiche „Klassenkontext“ und „Lehrerexpertise“ nicht berücksichtigt und Schülerauskünfte zu Unterrichtsprozessen nicht auf Klassenebene aggregiert. Gleichwohl stellt das Rahmenmodell dar, dass diese Bereiche prinzipiell Bedeutung für „Lern- und Leistungsergebnisse“ haben. Bei den Erhebungen zu *PISA 2003* mit dem Untersuchungsschwerpunkt *Mathematikleistung* sind diese Bereiche in einer Ergänzungsstudie erfasst worden. Die Ergänzungsstudie hat Aspekte des Klassenkontextes, der Unterrichtsgestaltung und der Lehrerexpertise untersucht (vgl. Baumert et al., 2004; Kunter et al., 2006).

Das Rahmenmodell hilft zwar bei der Klärung potenziell wichtiger Bereiche, stellt aber noch kein empirisch prüfbares Modell dar:

> „Ein Rahmenmodell dient … der ersten Orientierung im Forschungsfeld. Es handelt sich um eine vereinfachende Abbildung der Realität, bei der die Forschungsinteressen der Wissenschaftlerinnen und Wissenschaftler Einfluss auf die Auswahl der als relevant erachteten Kernbereiche und Analyseebenen nehmen. Ein Rahmenmodell ist in der Regel noch relativ komplex, und die zu erfassenden Variablen sind noch nicht operational definiert. Damit ist es in seiner Gesamtheit nicht empirisch überprüfbar“ (Bonsen et al., 2004, S. 218).

Wenn Leistungen, Leistungsunterschiede und Leistungsentwicklungen statistisch erklärt werden sollen, muss eine weiter reduzierende Modellbildung erfolgen, die weniger Konstrukte und konkretere Konstrukte berücksichtigt, die operationalisiert vorliegen oder operationalisiert werden können. Ein Beispiel hierfür ist das Pfadmodell in Abbildung 2.7, mit dem im Rahmen von *PISA 2000* die *Mathematikleistung* erklärt werden sollte.

Abbildung 2.7: „Pfadmodell zur Erklärung der Mathematikleistung“ (Quelle: Klieme et al., 2001, S. 184)



An dieser Stelle ist das Pfadmodell[34] nicht inhaltlich von Interesse, sondern eher formal, da es exemplarisch für reduzierende Modellbildung steht, die betrieben werden muss, wenn Leistungen, Leistungsunterschiede und Leistungsentwicklungen statistisch erklärt werden sollen. An diesem Pfadmodell kann exemplarisch verdeutlicht werden, dass die jeweils geschätzten statistischen Kennwerte davon abhängen, welche Konstrukte berücksichtigt werden. Während im obigen Pfadmodell das standardisierte Pfadgewicht für den Einfluss der kognitiven Fähigkeiten auf die Lesekompetenz 0,67 beträgt, wird in einem anderen Modell, das außer diesen beiden noch drei weitere „leserelevante Konstrukte“ enthält, ein standardisiertes Gewicht von 0,52 geschätzt (vgl. Artelt et al., 2001, S. 129).

Generell sind Modelle, wie das in Abb. 2.7 dargestellte, dann besonders hilfreich, wenn sie möglichst alle relevanten Einflussgrößen berücksichtigen. Welche dies sind, muss im Einzelfall theoretisch und empirisch geklärt werden, wobei es vor allem eine Aufgabe der Fachdidaktiken ist, wesentliche Einflussgröße bzw. Voraussetzungen für fachliches Lernen zu identifizieren und zu operationalisieren. Für die Erklärung von Fremdsprachenleistung

[34] Die Pfadanalyse kann als Spezialfalls von linearen Strukturgleichungsmodellen verstanden werden. Detaillierte Ausführungen findet man z. B. in Backhaus et al. (2008) oder Brachinger & Ost (1996).

etwa dürften andere Konstrukte relevant sein als für die Erklärung von *Mathematikleistung*. Im allgemeinen Rahmenmodell in Abb. 2.6 sind es vor allem die Bereiche, die nah an den „Lern- und Leistungsergebnissen“ stehen, die fachdidaktisch ausdifferenziert und konkretisiert werden müssen.

Bereits oben wurde erwähnt, dass in der nationalen Ergänzungsstudie zu *PISA 2003* Aspekte aus den Bereichen *Unterrichtsprozesse* und *Lehrerexpertise* untersucht wurden. Entsprechende Analysen werden im Rahmen des DFG-Projekts „COACTIV“ vertieft. Die vorliegende Arbeit kann – in einem sehr viel bescheideneren Rahmen und Umfang – möglicherweise entsprechende Beiträge in den Bereichen *individuelle Lernvoraussetzungen* und *individuelle Verarbeitung* leisten.

2.2 Weitere mathematikdidaktische Perspektiven auf Mathematikleistung

Bei den Ausführungen in Kap. 2.1 ist deutlich geworden, dass die Mathematikdidaktik – als *eine* an empirischer Bildungsforschung beteiligte Disziplin – wesentliche konzeptionelle Arbeiten zu dieser Art der Erforschung von *Mathematikleistung* beisteuert. Bei mindestens vier wichtigen Arbeitsschritten sollte die Mathematikdidaktik (im Sinne der Qualität der Schulleistungsstudie) maßgeblich beteiligt sein:

- Bildungstheoretische Grundlegung des Konstrukts *Mathematikleistung*
- Bereitstellung eines geeigneten Kompetenzbereichsmodells
- Entwicklung von geeigneten Testaufgaben
- Inhaltliche Interpretation der Ergebnisse (auch: Beschreibung von Kompetenzstufen)

Die bildungstheoretischen Grundlegung wurde in Kap. 2.1.1, die Bereitstellung eine Kompetenzbereichsmodells und die Beschreibung eine Kompetenzstufenmodells in Kap. 2.1.3 dargestellt und diskutiert. Auf die Entwicklung geeigneter Testaufgaben hingegen wurde bisher höchstens am Rande eingegangen. Aufgabenentwicklung ist in allen Fächern ein ureigener Kern fachdidaktischer Professionalität, da Aufgaben in der Schulpraxis eine herausragende Rolle spielen und Vorschläge für Lernumgebungen häufig erst durch die verwendeten Aufgaben konkret werden. Da sich die Mathematikdidaktik mit dem Lehren und Lernen von Mathematik auseinandersetzt, werden Aufgaben zur Leistungsüberprüfung vor allem mit Blick auf ihre Funktion für das Lehren und Lernen diskutiert.

Aufgaben sind im Unterricht wie im Leistungstest vor allem Anregungen zur fachlichen Tätigkeit. In dieser initiierenden Funktion werden sie von der Mathematikdidaktik intensiv untersucht und optimiert. Die Aufgabenentwicklung für Leistungstests profitiert von dieser langen Auseinandersetzung der Mathematikdidaktik mit Aufgaben, da die fachlichen Gehalte von Aufgaben und die durch sie (idealtypisch) angeregten kognitive Prozesse mit Blick auf das zugrundeliegende Kompetenzbereichsmodell klar erfasst und beschrieben

werden können. Die Beteiligung der Mathematikdidaktik an der Testentwicklung entspricht aber auch einem eigenen Interesse: Aufgaben aus Mathematiktests haben Rückwirkungen auf den Unterricht, entwickeln Unterricht also automatisch mit. Dabei soll möglichst wenig Schaden angerichtet und möglichst viel Nutzen erzeugt werden.

Da *Mathematikleistung* in der vorliegenden Arbeit als empirisches Konstrukt aufgefasst wird und im empirischen Teil der Arbeit die Quantifizierung von *Mathematikleistung* eine zentrale Rolle spielt, ist klar, dass sich die Arbeit selbst im Rahmen der empirischen Bildungsforschung bewegt und daher wesentlich mathematikdidaktische Beiträge aus diesem Bereich berücksichtigt.[35] Da Mathematikdidaktik sich aber im Kern um das Lehren und Lernen kümmert, werden im Folgenden – kontrastierend zur Leistungsmessung nach psychometrischen Gütekriterien – weitere mathematikdidaktische Perspektiven auf *Mathematikleistung* exemplarisch benannt.

Die Bedeutung von Leistung für das Lehren und Lernen von Mathematik

Mögliche Funktionen der Erfassung und Bewertung von Leistung für das Lehren und Lernen von Mathematik systematisieren Büchter & Leuders (2005b) vor allem aus der Perspektive des Arbeitens mit Aufgaben im Mathematikunterricht in den Sekundarstufen. Ähnlich strukturiert, aber unter Berücksichtigung der besonderen Rahmenbedingung und der besonderen Lernkultur in der Grundschule dargestellt, sind die Vorschläge von Sundermann & Selter (2006). „Aufgaben zum Leisten“ werden von Büchter & Leuders (2005b, Kap. 5) wie folgt unterschieden:

- Aufgaben für eine *kompetenzorientierte Diagnose*: Lehr-Lernsituationen sind dann produktiv, wenn möglichst alle Schülerinnen und Schüler – ausgehend von ihren individuellen Vorstellungen und ihren fachlichen Kompetenzen – in individuell zugänglichen Situationen kognitiv herausgefordert werden. Dies kann mit schülerzentrierten Methoden genauso stattfinden wie in einem Unterrichtsgespräch, das Platz für individuelle Denkwege lässt. Voraussetzung für die Gestaltung derart produktiver Lernumgebungen ist, dass die Lehrkraft die individuellen Lernvoraussetzungen der Schülerinnen und Schüler kennt. Testergebnisse z. B. aus Vergleichsarbeiten greifen hier zu kurz, da sie sich eher global zum Leistungsvermögen äußern, nicht aber individuelle Sichtweisen und Schwierigkeiten offenbaren. Wichtig ist hier auch die kompetenzorientierte Sichtweise in dem Sinne, dass identifiziert wird, auf welchen vorhandenen Kompetenzen der Schülerinnen und Schüler die Förderung aufbauen kann, und nicht, welche Defizite be-

[35] Aus der (teilweise vehementen) Kritik an „PISA & Co.“, die oft geisteswissenschaftlichen Ursprungs ist, resultieren übrigens keine Alternativen für die Erfassung und Bewertung von *Mathematikleistung*. Dies ist bemerkenswert, da Leistungsüberprüfungen im deutschen Schulsystem fest verankert sind und Lehrerinnen und Lehrer fast täglich mit dem Erstellen oder Auswerten von Leistungsüberprüfungen beschäftigt sind. Dies soll nicht heißen, dass die Schulpraxis die Formate der Schulleistungsstudien übernehmen soll, sondern dass Mathematikdidaktik insgesamt Beiträge zur sinnvollen Überprüfung von Leistung bereitstellen muss.

stehen (vgl. Scherer, 1999, S. 170 f.). Aufgaben, die eine solche kompetenzorientierte Diagnostik unterstützen, regen vor allem „mathematische Eigenproduktionen von Kindern“ (vgl. Wollring, 1999, S. 272) an.

- Aufgaben für die *Leistungsbewertung*: Die Leistungsüberprüfung und -bewertung ist im deutschen Schulsystem, vor allem in den Sekundarstufen, fest mit schriftlichen Arbeiten und Ziffernnoten verbunden. Mit ihnen wird die „Selektionsfunktion von Schule“ (Fend, 1980, S. 29) bedient. Da das Lernen von Mathematik *das* zentrale Ziel des Mathematikunterrichts ist, sollte die Leistungsbewertung das Lernen nicht behindern, sondern nach Möglichkeit fördern. Für Klassenarbeiten gilt aber dasselbe wie für Vergleichsarbeiten und zentrale Prüfungen: Gelernt wird, was (voraussichtlich) überprüft wird. Da das Lernen neben notwendigen Automatisierungen vor allem verstehensorientiert stattfinden soll, gilt diese Anforderung auch für die schriftlichen Leistungsüberprüfungen im Schulalltag (vgl. Büchter & Leuders, 2008). Darüber hinaus sollten die Aufgaben für die Leistungsbewertung so gestellt sein, dass sie den Lehrkräften, den Schülerinnen und Schülern sowie den Eltern informative Rückmeldungen über die Lernergebnisse (und damit die Lernvoraussetzungen für den weiteren Unterricht) geben können.
- Aufgaben zum *Kompetenzerleben* / zur *Selbsteinschätzung*: Die Bedeutung selbstregulierten Lernens und selbstbezogener Fähigkeitskognitionen wird immer wieder – und empirisch gut abgesichert – betont (vgl. z. B. Artelt et al., 2001). Dass Erfolgserlebnisse das Lernen fördern, ist fast schon eine pädagogische Binsenweisheit. Die pädagogische Psychologie hat aber wesentlich zum Verstehen der Wirkungswege beigetragen. So liefern erfolgreiche Aufgabenbearbeitungen den Schülerinnen und Schülern wichtige Hinweise auf das, was sie können, leisten also einen Beitrag zur Selbsteinschätzung, die eine Grundlage für selbstreguliertes Lernen ist. Zugleich stärken erfolgreiche Aufgabenbearbeitungen das Vertrauen in das eigene Leistungsvermögen (*Fähigkeitsselbstkonzept*), was sich positiv auf weitere Lernprozesse auswirkt.

Über die hier genanten Betrachtungen hinaus gibt es viele weitere Arbeiten – jenseits der empirischen Bildungsforschung –, die sich aus mathematikdidaktischer Sicht mit *Mathematikleistung* befassen. Auffallend ist aber, dass dies im Vergleich zu expliziten Vorschlägen für die Gestaltung des Mathematiklernens nur wenige Arbeiten sind und dass das Leisten immer wieder in seiner Funktion für das Lernen betrachtet wird.

2.3 Befunde zur Mathematikleistung

Die Erforschung von *Mathematikleistung* beginnt historisch natürlich nicht erst mit der Etablierung von *IRT*-Modellen oder der modernen empirischen Bildungsforschung[36]. Als

[36] Unter „moderner Bildungsforschung“ werden die seit den 1990er-Jahren durchgeführten Studien verstanden, die in der Regel (a) kooperativ von Psychologie und Fachdidaktiken durchgeführt werden und (b) empirisch auf *IRT*-Modellen basieren.

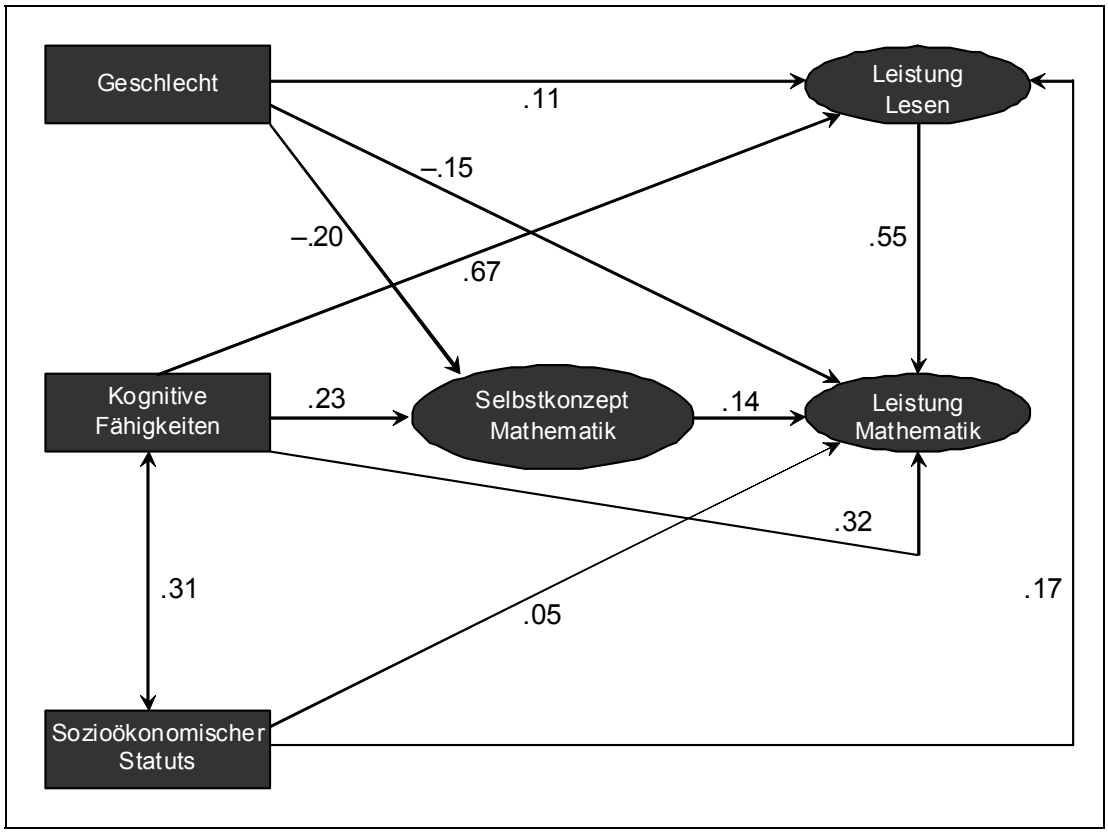
klassisches Fach ist Mathematik seit jeher Gegenstand von Leistungsbewertung und Bestandteil von Eignungsfeststellungsverfahren. Auch im Rahmen von Intelligenztests werden in der Regel Subtests verwendet, die eine große Nähe zu *Mathematikleistung* haben (z. B. Zahlenreihen). So verwundert es nicht, dass es eine kaum zu überschauende Anzahl von Beiträgen gibt, in denen Befunde zur *Mathematikleistung* berichtet werden.

Im Folgenden wird die Darstellung auf solche Befunde reduziert, die der modernen empirischen Bildungsforschung (etwa seit der *TIMSS*-Untersuchung in den 1990er-Jahren) entstammen oder bei denen es sich um klassische Arbeiten handelt, die forschungsmethodisch hinreichend ausgereift sind. Darüber hinaus müssen die Befunde relevant für die Fragestellung der vorliegenden Arbeit sein. Zur ersten Orientierung bezüglich möglicher Einflussfaktoren auf *Mathematikleistung* werden einige ausgewählte Befunde skizziert. Anschließend werden Geschlechterunterschiede in der *Mathematikleistung* und die Dimensionalität von *Mathematikleistung* genauer betrachtet.

2.3.1 Befunde zu ausgewählten Einflussfaktoren

Erste Befunde zur *Mathematikleistung*, die im Rahmen von *PISA 2000* gewonnen wurden, kann man dem bereits weiter oben dargestellten „Pfadmodell zur Erklärung von Mathematikleistung“ entnehmen, das hier noch einmal wiedergegeben wird.

Abbildung 2.8: „Pfadmodell zur Erklärung der Mathematikleistung“ (Quelle: Klieme et al., 2001, S. 184)



An dieser Stelle soll nicht methodisch auf Pfadmodelle eingegangen werden[37], sondern ein erster Eindruck von möglichen Einflussfaktoren auf *Mathematikleistung* vermittelt werden. Die im Pfadmodell wiedergegebenen Effekte („standardisierte Pfadkoeffizienten“) gelten zunächst nur für die Population der *PISA*-Studie (15-jährige Schülerinnen und Schüler) und nur für *dieses* Modell mit *diesen* Konstrukten und *diesen* modellierten Zusammenhängen. Es ist z. B. denkbar, dass ein nicht berücksichtigtes Konstrukt existiert, das sowohl die Leistung im Bereich *Lesen* als auch die Leistung im Bereich *Mathematik* beeinflusst.[38]

Diese Einschränkungen sollen das Modell und seine Bedeutung aber nicht grundsätzlich infrage stellen, denn immerhin hat es eine hohe statistische Erklärungskraft: 76 % der Varianz der *Mathematikleistung* kann mit diesem Pfadmodell statistisch erklärt werden (vgl. Klieme et al., 2001, S. 184). Die geschätzten standardisierten Pfadkoeffizienten im Modell können wie Korrelationskoeffizienten interpretiert werden: Wenn die Leseleistung in der Population um eine Standardabweichung zunimmt, dann wächst die *Mathematikleistung* (im Mittel) um 0,55 Standardabweichungen.

Innerhalb seines Bezugsrahmens deutet dieses Modell also darauf hin, dass sich die Leseleistung, die kognitiven Fähigkeiten („Intelligenz“), das *Fähigkeitsselbstkonzept Mathematik* und der sozioökonomische Status positiv auf *Mathematikleistung* auswirken, wobei der Effekt der Leseleistung am höchsten ist. Der Effekt des soziökonomischen Status ist zwar statistisch signifikant unterschiedlich von Null, aber sehr klein.

Bei der Interpretation der standardisierten Pfadgewichte, die von der Variable „Geschlecht“ ausgehen, muss berücksichtigt werden, dass negative Werte Effekte zugunsten von Jungen und positive Werte Effekte zugunsten von Mädchen darstellen. Das Modell gibt einen direkten Effekt des Geschlechts auf die *Mathematikleistung* zugunsten der Jungen an und darüber hinaus noch zwei vermittelte Effekte. So wirkt das Geschlecht über die Leseleistung zugunsten der Mädchen und über das *Fähigkeitsselbstkonzept* zugunsten der Jungen auf die *Mathematikleistung*. Insgesamt resultiert hieraus ein Gesamteffekt zugunsten der Jungen (vgl. 2.3.2). Der direkte Einfluss des Geschlechts auf die *Mathematikleistung* in diesem Modell wird möglicherweise – so die Aussage der in Kap. 3.3.6 formulierten „spatial mediation hypothesis“ (vgl. Burnett et al., 1979) – über *Raumvorstellung* vermittelt. Da *Raumvorstellung* in *PISA 2000* nicht erfasst wurde, kann dieses Konstrukt im obigen Pfadmodell aber rechnerisch nicht berücksichtigt werden.

[37] vgl. hierzu z. B. Backhaus et al. (2008) oder Brachinger & Ost (1996)

[38] So fehlt im abgebildeten Modell z. B. die Naturwissenschaftsleistung. Knoche et al. (2002, S. 186 ff.) zeigen anhand des Gesamttests *Mathematik* (internationale Erhebung und nationale Ergänzung zusammen), dass die einfache Korrelation zwischen *Mathematikleistung* und *Leseleistung* 0,83 beträgt, diese sich aber auf 0,42 verringert, wenn man die *Naturwissenschaftsleistung* in einer partiellen Korrelation kontrolliert (also „herauspartialisiert“).

Bei den oben berichteten Befunden handelt es sich um die Quantifizierung von Einflüssen auf die *Mathematikleistung* innerhalb des theoretischen *PISA*-Rahmenmodells (Abb. 2.6), konkretisiert durch ein empirisches Pfadmodell (Abb. 2.8). Im Folgenden werden noch Befunde berichtet, die vor allem für die mathematikdidaktische Konstruktion und Diskussion von Leistungstests relevant sind. Zum einen geht es um den Zusammenhang zwischen der curricularen Validität von Mathematiktests und den zugehörigen Testleistungen, zum anderen um schwierigkeitsbestimmende Faktoren bei Testaufgaben zur Mathematik. Beide Aspekte sind für die vorliegende Arbeit von besonderer Bedeutung. So wird im empirischen Teil dieser Arbeit die *Mathematikleistung* mit den nordrhein-westfälischen Lernstandserhebungen in der Jahrgangsstufe 9 (*LSE 9*), also einer zu 100 % curricular validen Vergleicharbeit, erhoben. Da mit der Arbeit insgesamt die „*PISA*-Population" im Blick ist und *PISA* auf einem Grundbildungskonzept basiert, ist von Interesse, wie sich „curricular optimierte" Tests empirisch gegenüber Grundbildungstests verhalten. Die schwierigkeitsbestimmenden Faktoren für Testaufgaben wiederum können bei den Analysen des Zusammenhangs von *Raumvorstellung* und *Mathematikleistung* bedeutsam sein, wenn *Mathematikleistung* nicht global betrachtet, sondern der verwendete Test (*LSE 9*) nach einem Klassifikationsschema in verschiedenen Bereiche aufgeteilt werden soll.

Zusammenhang von curricularer Validität der Tests und Mathematikleistung

Die hier betrachtete Fragestellung lässt sich wie folgt präzisieren: Werden die Testergebnisse bei einer Schulleistungsstudie für eine Nation (bzw. ein Bundesland) besser, wenn statt des ursprünglichen Mathematiktests nur die Aufgaben ausgewertet werden, die im Mathematikunterricht dieses Bildungssystems verankert sind? Für die Beantwortung dieser Frage muss zuvor noch konkretisiert werden, wann eine Aufgabe im Mathematikunterricht eines Bildungssystems „verankert" ist.

Das Untersuchungskonzept von *TIMSS/II* (vgl. Baumert & Lehmann, 1997, S. 46 ff.) unterscheidet drei Stufen eines nationalen Curriculums[39]:

- Das *intendierte Curriculum* wird durch die eigentlichen curricularen Vorgaben, also z. B. Richtlinien und Lehrpläne und ggf. zusätzliche verbindliche Begleitmaterialien, festgelegt. In *TIMSS/II* wird diese Stufe durch Curriculumsanalysen und Expertenbefragungen erfasst.
- Das *implementierte Curriculum* stellt dar, welche fachlichen Gegenstände und Tätigkeiten tatsächlich im Unterricht thematisiert bzw. angeregt wurden. In *TIMSS/II* war das implementierte Curriculum dabei noch sehr „stofflastig" und hat die typischen mathe-

[39] Bei *TIMSS/III* werden sogar vier Stufen des Curriculums unterschieden (vgl. Baumert, Bos et al, 2000a, 2000b). Zusätzlich zu den hier benannten wird (zwischen dem intendierten und dem implementierten Curriculum) noch das potenzielle Curriculum betrachtet, das durch die zugelassenen Schulbücher dokumentiert wird und somit durch eine Schulbuchanalyse erfasst werden kann.

matischen Tätigkeiten noch nicht so berücksichtigt, wie dies heute geschehen würde. Diese Stufe des Curriculums wird durch die Befragung von Schulleitungen und Fachlehrkräften erfasst.

- Schließlich wird das *erreichte Curriculum* durch die Mathematiktests erfasst. Diese Stufe des Curriculums wird also aus den Lern- bzw. Leistungsergebnissen gebildet.

Auf dieser konzeptionellen Basis wurden im Sinne „transkultureller Testfairness […] Tests optimaler nationaler Validität" (ebd., S. 188) berechnet. Hierfür wurde nicht das *intendierte*, sondern das *implementierte Curriculum* zugrunde gelegt, also ein für das jeweilige Bildungssystem unterrichtsvalider Test rechnerisch zusammengestellt. Dieser Test kann insofern deutlich vom Gesamttest abweichen, als sich die bei *TIMSS/II* angestrebte *curriculare Validität* beim Gesamttest auf die *intendierten Curricula* bezieht und die Verbindlichkeit dieser Curricula in den verschiedenen Bildungssystemen unterschiedlich ist.

Tatsächlich schwankt der Anteil der unterrichtsvaliden Aufgaben im Gesamttest zwischen 46 % für Griechenland und 100 % für Ungarn und die USA (vgl. ebd., Tabelle F6, S. 189 f.). Umso erstaunlicher ist es, dass der mittlere Anteil richtig gelöster Aufgaben sich vom Gesamttest zu den jeweils unterrichtsvaliden Test in 28 von 33 Bildungssystemen um maximal 1 % unterscheidet, also praktisch gleich bleibt (darunter auch Griechenland!). Der größte Unterschied ergibt sich für England, wo beim optimierten Test 57 % der Aufgaben im Mittel richtig gelöst wurden, während es beim Gesamttest nur 53 % waren.

Betrachtet man das Ergebnis der deutschen Schülerinnen und Schüler bei den 33 verschiedenen national unterrichtsvaliden Tests, so bleibt der mittlere Anteil richtig gelöster Aufgaben ebenfalls für 28 der 33 dieser Tests praktisch unverändert (54 % ± 1 %). Die größte Abweichung ergibt sich interessanterweise nach oben: Beim dem für England optimierten Test liegt der mittlere Anteil der von deutschen Schülerinnen und Schülern richtig gelösten Aufgaben bei 57 % (+ 3 %).[40] Die Autoren des Berichts fassen die Ergebnisse dieser Betrachtung wie folgt zusammen:

> „Insgesamt kann für das Fach Mathematik sehr zuverlässig gefolgert werden, daß Länderunterschiede in den Mathematikleistungen nicht durch nationale Curriculumbesonderheiten erklärt werden können. Das besagt nicht, daß es keine curriculare Variabilität gäbe. […] Diese Unterschiede [sind] zur Erklärung von internationalen Leistungsdifferenzen [jedoch] ungeeignet" (ebd., S. 191; Erg. d. d. Verf.).

Vergleichbare Betrachtungen zu möglichen Verschiebungen zwischen dem Gesamttest und curricular validen bzw. unterrichtsvaliden Tests gibt es insbesondere bei allen *PISA*-Erhebungen, da die dort verwendeten Leistungstests alle auf Grundbildungskonzepten basieren und ihnen gelegentlich vorgeworfen wird, am implementierten Curriculum vorbei zu

[40] Dies wird von den Autoren so gedeutet, dass der für England optimierte Test überdurchschnittlich viele leichte Aufgaben enthält, also insgesamt ein leichterer Test ist.

zielen (vgl. z. B. Bender, 2005). Im Ergebnis kommen die Autoren der Berichte allerdings immer wieder zum obigen Schluss.

Möglicherweise lassen sich diese Befunde mit der besonderen Anlage der entsprechenden Mathematiktests begründen. Diese sind – anders als Klassenarbeiten – ausgehend von Kompetenzbereichsmodellen (vgl. 2.1.3) auf die Breite des Fachs ausgerichtet. Somit spielt nicht die direkt vorangegangene Unterrichtsreihe und das zielgerichtete Einüben der zugehörigen Inhalte die entscheidende Rolle für die Bearbeitung der Testaufgaben, sondern die mittel- und langfristigen Lernergebnisse der vergangenen Unterrichtsjahre. Möglicherweise sind der Kern des mathematischen Handwerkzeugs und die Flexibilität beim Einsatz dieses Handwerkzeugs – so wie beides über die Jahre ausgebildet wird – weniger speziell als die nationalen Curricula es sind.

Schwierigkeitsbestimmende Faktoren bei Testaufgaben zur Mathematik

Für die Entwicklung von Mathematiktests und für die differenzierte Untersuchung des Zusammenhangs von *Mathematikleistung* mit anderen Personeneigenschaften ist es äußerst nützlich, schwierigkeitsbestimmende Faktoren von Testaufgaben zu kennen. Für die Testentwicklung ist dies von besonderer Bedeutung, da ein Test – passend zur Verteilung der Personenparameter auf die Leistungsskala – in jedem Bereich der Population Items enthalten muss, die hier differenzieren können, die also von einigen Versuchspersonen in diesem Bereich richtig gelöst werden und von anderen nicht. Bei der differenzierten Untersuchung des Zusammenhangs von *Raumvorstellung* und *Mathematikleistung* kann es z. B. sein, dass dieser Zusammenhang für einfache Items anders ausgeprägt ist als für schwierige Items. Vielleicht sind auch nur bestimmte Aufgabenmerkmale „empfänglich“ für einen Einfluss der *Raumvorstellung*.

Neubrand et al. (2002) sind der Frage von schwierigkeitsbestimmenden Faktoren für Mathematikaufgaben anhand der Ergebnisse von *PISA 2000* (internationale Erhebung und nationale Ergänzungsstudie) nachgegangen. Sie unterscheiden zunächst auf der Basis theoretischer Vorarbeiten sechs potenziell schwierigkeitsbestimmende Aufgabenmerkmale:

- Bei der *Komplexität der Modellierung* werden die drei Kompetenzklassen aus dem internationalen *PISA*-Konzept unterschieden (*Reproduktion*, *Verknüpfung* und *Verallgemeinerung*; vgl. 2.1.3, S. 32).
- Die *curriculare Wissensstufe* einer Aufgabe wird ebenfalls dreistufig erfasst (*Grundkenntnisse*, *einfaches Wissen der Sekundarstufe I* und *anspruchsvolles Wissen der Sekundarstufe I*).
- Bei den *Kontexten* von Aufgaben gibt es die Unterscheidung in Aufgaben mit *innermathematischem Kontext*, Aufgaben mit *außermathematischem Kontext* und Aufgaben *ohne Kontext*.

- Unter dem Merkmal *Offenheit* wird dichotom erfasst, ob Aufgaben unterschiedliche (erfolgreiche) Mathematisierungen zulassen.
- Mit dem *Umfang der Verarbeitung* wird die Komplexität einer Aufgabe daran gemessen, „ob und in welchem Umfang neue Größen und Zwischenergebnisse, die noch nicht in der Aufgabenstellung selbst vorgegeben sind, im Verlauf der Lösung der Aufgaben eingeführt werden müssen“ (Neubrand et al., 2002, S. 107). Dieses Merkmal wird ebenfalls dichotom erfasst (hoher vs. niedriger Verarbeitungsumfang).
- Schließlich wird wiederum dichotom erfasst, ob die Aufgaben *Argumentieren* erfordert.

Im Rahmen einer Regressionsanalyse, bei der die *Mathematikleistung* mit den 117 Aufgaben des internationalen Mathematiktests von *PISA 2000* und seiner nationalen Ergänzung gemessen wurde, konnten die sechs genannten Aufgabenmerkmale zusammen 45 % der Leistungsvarianz statistisch erklären. Da nicht alle Aufgabenmerkmale bei allen Aufgaben wirken können (z. B. Offenheit oder Argumentieren bei einer reinen Berechnungsaufgabe), wurden die Aufgaben nach drei Typen des mathematischen Arbeitens in *technische Aufgaben*, *rechnerische Modellierungsaufgaben* und *begriffliche Modellierungsaufgaben* unterschieden, um anschließend für diese drei Gruppen wieder Regressionsanalysen mit den folgenden Ergebnissen durchzuführen:

- Bei *technischen Aufgaben* werden 32 % der Leistungsvarianz alleine durch die *curriculare Wissensstufe* statistisch erklärt.
- Bei *rechnerischen Modellierungsaufgaben* können die *Komplexität der Modellierung*, die *curriculare Wissensstufe*, das Vorhandensein eines *außermathematischen Kontextes* und der *Verarbeitungsumfang* zusammen 49 % der Leistungsvarianz statistisch erklären.
- Bei *begrifflichen Modellierungsaufgaben* wirken vor allem die *Komplexität der Modellierung*, die rein innermathematische Verarbeitung (*Kontext*) und die *Offenheit* der Aufgabe gemeinsam derart, dass 31 % der Leistungsvarianz statistisch erklärt werden können.

Bei den in dieser Untersuchung berücksichtigten Aufgaben und Aufgabenmerkmalen und bei der dargestellten Art der Einteilung von Aufgaben in Aufgabentypen ist bemerkenswert, dass das Aufgabenmerkmal „*Argumentieren erforderlich*“ in gemeinsamen Regressionsmodellen mit den anderen Aufgabenmerkmalen nicht zur statistischen Erklärung von Leistungsvarianz beiträgt. Bei einer isolierten Betrachtung kann dieses Merkmal bei *begrifflichen Modellierungsaufgaben* immerhin 23 % der Leistungsvarianz statistisch erklären, was jedoch offensichtlich bereits durch die drei oben für diesen Aufgabentyp genannten Merkmale berücksichtigt wird.

Für die vorliegende Arbeit ist besonders bedeutsam, dass die Vorhersagekraft der oben betrachteten Aufgabenmerkmale bei *rechnerischen Modellierungsaufgaben* am höchsten

ist. Dieser Aufgabentyp ist in den *LSE 9*, mit denen im empirischen Teil der Arbeit die *Mathematikleistung* erfasst wird, umfassend vertreten. Bei den Analysen im empirischen Teil der Arbeit können also die von Neubrand et al. (2002) identifizierten schwierigkeitsbestimmenden Faktoren für differenzierte Betrachtungen herangezogen werden.

2.3.2 Geschlechterunterschiede

Wer sich mit der Erforschung von *Mathematikleistung* beschäftigt, kommt an der Frage der Geschlechterunterschiede nicht vorbei:

> „Der Leistungsvorsprung, den männliche Probanden gegenüber Mädchen bzw. Frauen – zumindest vom Zeitpunkt der Adoleszenz an – bei der Bewältigung mathematischer Anforderungen gewinnen, gehört zu den am deutlichsten ausgeprägten und am besten dokumentierten Befunden über Geschlechtsunterschiede im Bereich der Psychologie“ (Klieme, 1986, S. 133).

Zwar gibt es immer wieder auch (seriöse) Studien, die in der untersuchten Stichprobe keine signifikanten Leistungsunterschiede zwischen den Geschlechtern feststellen (vgl. z. B. Tartre & Fennema, 1995), in der Regel liegt dies aber an der Stichprobenziehung, an den getesteten Bereichen oder an strukturell verzerrenden Rahmenbedingungen. Strukturell Einfluss nehmen können z. B. institutionelle Rahmenbedingungen, wie die für das Schulsystem typische „Klumpenbildung“: Schülerinnen und Schüler sind in Klassen zusammengefasst, die in Schulen zusammengefasst sind, die ggf. in Schulformen zusammengefasst sind, die ggf. in regionale Bildungssysteme eingebunden sind … Ein Beispiel dafür, wie diese Klumpenbildung Befunde zu Geschlechterunterschieden beeinflussen kann, liefert *TIMSS/II* (vgl. Baumert & Lehmann, 1997):

> „Mädchen erreichen in Mathematik und Physik in allen Schulformen schwächere Leistungen als Jungen. […] Bei der Betrachtung der Leistungsbilanz von Jungen und Mädchen auf der Ebene des gesamten Altersjahrgangs treten im Fach Mathematik keine und im Fach Physik kleinere Leistungsunterschiede zwischen Geschlechtern auf als in den einzelnen Schulformen. Dies ist ausschließlich eine Folge der höheren gymnasialen Bildungsbeteiligung von Mädchen …“ (ebd., S. 26).

Das zunächst scheinbar widersprüchliche Bild, das die *TIMSS/II*-Ergebnisse auf der Ebene der Schulformen und auf der Ebene des Altersjahrgangs ergeben, lässt sich also als „Simpson-Paradoxon“ erklären (vgl. Büchter, 2004). Bei „Klumpenstichproben“ kann es immer wieder passieren, dass sich innerhalb jedes Klumpens das gleiche Bild ergibt, das aus den Klumpen (additiv) erstellte Gesamtbild aber hiervon abweicht. Da Bildungssysteme, wie oben dargestellt wurde, immer in Klumpen organisiert sind, muss bei der Auswertung und Interpretation von Testdaten ein mögliches *Simpson-Paradoxon* immer mitgedacht werden.

Was aber bedeutet dies nun für die Interpretation der *TIMSS/II*-Ergebnisse? Auf der Ebene des gesamten Altersjahrgangs waren die Geschlechterunterschiede verschwunden – ein Widerspruch zum oben zitierten Resultat von Klieme (1986)? Möchte man die Situation erklären, ist es hilfreich, sich zunächst mehr Befunde zu Geschlechterunterschieden anzusehen und die *TIMSS/II*-Ergebnisse dann noch einmal differenziert zu betrachten.

Ausmaß und Stabilität der Geschlechterunterschiede

Die große Zahl belastbarer Studien ergibt zwar global betrachtet kein so einheitliches Bild, wie es das Klieme-Zitat vermuten lässt, schaut man aber etwas differenzierter hin und gruppiert die Befunde angemessen, so lässt sich doch eine stabile Befundlage feststellen. Im Folgenden wird sichtbar, dass die Geschlechterunterschiede zugunsten von Jungen in Deutschland signifikant nachweisbar sind – insbesondere in den großen Schulleistungsstudien, mit anspruchsvoll gezogenen Stichproben und mit zeitgemäßen, breit angelegten Mathematiktests. Für den engeren Zweck der vorliegenden Arbeit genügt eine solche Aussage über Schülerinnen und Schüler in Deutschland. Möchte man jedoch Geschlechterunterschiede inhaltlich erklären, so wird man jenseits nationaler Spezifika international stabile Befunde zur *Mathematikleistung* berücksichtigen.

Gegenläufige Befunde berichten, z. B. Hyde et al. (1990), die in einer Meta-Analyse, vornehmlich für den angelsächsischen Raum feststellen, dass praktisch keine Geschlechterunterschiede bei global betrachteter *Mathematikleistung* bestehen und zuvor oder für einzelnen Komponenten festgestellte Leistungsunterschiede sich im Laufe der Zeit verringern („säkularer Trend"). An dieser Stelle soll nicht differenziert diskutiert werden, warum diese Befunde zum Teil anders ausfallen als aktuelle Befunde für das deutsche Bildungssystem, sondern stattdessen darauf hingewiesen werden, dass internationale Vergleichsstudien auch in Bezug auf Geschlechterunterschiede in der *Mathematikleistung* erhebliche Unterschiede zwischen den Teilnehmerstaaten aufweisen. So berichten Stanat & Kunter (2001, S. 252) für *PISA 2000* die Geschlechterunterschiede in den drei Testbereichen *Lesen*, *Mathematik* und *Naturwissenschaften* mit folgendem Bild:

- Im Bereich *Lesen* gibt es in allen 32 Staaten signifikante Geschlechterunterschiede zugunsten der Mädchen. Diese Unterschiede schwanken allerdings auf der *PISA*-Metrik[41] zwischen über 50 Punkte (Lettland und Finnland) und unter 20 Punkte (Korea und Brasilien). In Deutschland beträgt der Unterschied 35 Punkte.
- Im Bereich *Naturwissenschaften* ist das Bild bezüglich signifikanter Geschlechterunterschiede dagegen recht heterogen. In vier Staaten gibt es signifikante Unterschiede zugunsten der Mädchen und in drei Staaten zugunsten der Jungen. Die anderen Unterschiede sind nicht signifikant. Der größte Unterschied zugunsten der Mädchen ist wiederum in Lettland zu beobachten (über 20 Punkte), der größte Unterschied zugunsten der Jungen in Korea (fast 20 Punkte).
- Im Bereich *Mathematik* ist das Bild bezüglicher signifikanter Unterschiede zwar weniger heterogen als in Naturwissenschaften, aber nicht so homogen wie beim Lesen. Es gibt keine Staaten mit signifikanten Geschlechterunterschieden zugunsten der Mädchen,

[41] Die internationalen Leistungsskalen werden bei *PISA* grundsätzlich so normiert, dass der OECD-Durchschnitt bei 500 Punkten liegt und die Standardabweichung 100 beträgt.

aber 16 Staaten mit Unterschieden zugunsten der Jungen. In Deutschland beträgt der signifikante Unterschied zugunsten der Jungen 11 Punkte, ist auf der *PISA*-Metrik also nur etwa ein Drittel so groß wie der anders gerichtete Unterschied im Lesen.

Dieser Befund bleibt international bei *PISA 2003* und *2006* im Wesentlichen unverändert. Für *PISA 2003* berichten Blum et al. (2004, S. 83) für 29 OECD-Staaten die Geschlechterunterschiede im Bereich Mathematik. Dieses Mal gibt es mit Island einen Staat mit signifikanten Geschlechterunterschieden zugunsten der Mädchen (ca. 15 Punkte) und 21 Staaten mit signifikanten Unterschieden zugunsten der Jungen, darunter Korea mit dem größten Unterschied (über 20 Punkte). In Deutschland beträgt der signifikante Unterschied zugunsten der Jungen in diesem Durchgang 9 Punkte auf der internationalen *PISA*-Metrik. Für *PISA 2006* beträgt der signifikante Unterschied zugunsten der Jungen, bei einem kaum veränderten internationalen Bild, 20 Punkte (vgl. Frey et al., 2007, S. 265). Der von Hyde et al. (1990) berichtete *säkulare Trend* trifft in dieser Form anscheinend nicht auf die *PISA*-Erhebungen zu.[42] Auf der Basis der *PISA*-Erhebungen gibt es also für die Population der 15-Jährigen in Deutschland eine anscheinend belastbare Aussage.

Die beobachteten Unterschiede sind zwar auf der Ebene des Altersjahrgangs nicht besonders groß, aber relativ stabil nachweisbar. Dabei ist allerdings noch nicht die unterschiedliche Bildungsbeteiligung von Mädchen und Jungen berücksichtigt worden, die zum „*TIMSS/II*-Paradoxon“ geführt hat. Betrachtet man die Schulformen separat, so wird der Geschlechterunterschied größer und beträgt in allen Schulformen zwischen 0,25 und 0,30 Standardabweichungen[43] (vgl. Stanat & Kunter, 2001, S. 259).

Das konsistente Bild, das die *PISA*-Erhebungen zur Frage der Geschlechterunterschiede in der *Mathematikleistung* ergeben, wird aber noch durch die eingangs berichteten Befunde im Rahmen von *TIMSS/II* gestört. Wieso gibt es auf der Ebene des Altersjahrgangs bei *PISA* signifikante Unterschiede und bei *TIMSS/II* nicht? Und was legitimiert eigentlich die separaten Schulformbetrachtungen inhaltlich?

Zum Unterschied zwischen den Befunden von *TIMSS/II* und *PISA* geben Stanat & Kunter (2001) eine Erklärung, die im nächsten Abschnitt noch vertieft wird:

[42] Dabei muss natürlich berücksichtigt werden, dass drei Erhebungen, die über sechs Jahre verteilt sind, noch keine besonders belastbare Grundlage für grundsätzliche Trendaussagen darstellt. Trotzdem sind die Daten zumindest nicht geeignet, um die Annahme eines *säkularen Trends* zu unterstützen.

[43] Die Geschlechterunterschiede werden hier nur relativ zur Standardabweichung berichtet, da sie an der entsprechenden Stelle Bericht zu *PISA 2000* auf der nationalen Metrik mit Mittelwert 100 und Standardabweichung 30 mitgeteilt werden. Der Geschlechterunterschied in der Gesamtschule liegt dabei zwar im Bereich der Unterschiede für die anderen Schulformen, wird aber aufgrund der Stichprobengröße in dieser Schulform nicht signifikant.

> „Solche Abweichungen in den Befunden der beiden Studien dürften vor allem auf Unterschiede in den Schwerpunkten der Tests zurückzuführen sein. So stehen die Anwendungen von mathematischen Kenntnissen sowie die konzeptuelle Modellierung im PISA-Mathematiktest noch stärker im Vordergrund als in TIMSS, und es werden somit Teilkompetenzen betont, bei denen sich der Leistungsvorteil der Jungen als vergleichsweise groß erwiesen hat …“ (ebd., S. 253).

Bei *TIMSS/II* hat sich „das gewohnte Bild“ ergeben, wenn man die Frage der Geschlechterunterschiede schulformspezifisch betrachtet. Innerhalb jeder Schulform ergaben sich in der *Mathematikleistung* signifikanten Unterschiede zugunsten der Jungen. Diese schulformspezifische Sicht ist dabei kein „Taschenspielertrick“, um „das gewohnte Bild“ wieder herzustellen, sondern inhaltlich gerechtfertigt, auch wenn man am Ende ein belastbares Ergebnis für den Altersjahrgang bzw. die Jahrgangsstufe erhalten möchte. Denn das „*TIMSS/II*-Paradoxon“ führt bei seiner Aufklärung dazu, dass Schulformen in ihrer Rolle als „differenzielle Entwicklungsmilieus“ berücksichtigt werden:

> „Sowohl Schulformen als auch Einzelschulen innerhalb derselben Schulform stellen institutionell vorgeformte differenzielle Entwicklungsmilieus dar. Schüler und Schülerinnen mit gleicher Begabung, gleicher Fachleistung und gleicher Sozialschichtzugehörigkeit erhalten je nach Schulformzugehörigkeit und je nach besuchter Einzelschule unterschiedliche Entwicklungschancen“ (Baumert et al., 2003, S. 288).

Der hier beschriebene Effekt führt dazu, dass z. B. Schülerinnen und Schüler in Gymnasien – nur aufgrund der dortigen Lernumgebung – deutlich höhere Leistungszuwächse erzielen als Schülerinnen und Schüler in Hauptschulen, die zu Beginn der Sekundarstufe I über gleiche Voraussetzungen (u. a. bzgl. kognitiver Fähigkeiten, Fachleistungen und sozioökonomischem Status) verfügen. Durch die Schulformzugehörigkeit wird also Einfluss auf die Leistungsentwicklung ausgeübt und die Leistungsstände nach mehreren Schuljahren in der Sekundarstufe I sind nicht mehr nur durch individuelle Lernprozesse zu erklären. Berücksichtigt man bei *TIMSS/II* die Schulformzugehörigkeit, das Geschlecht und die *Mathematikleistung* in einem statistischen Modell, so ergibt sich im Rahmen einer Varianzanalyse (vgl. z. B. Backhaus et al., 2008, oder Fahrmeier, Hamerle & Nagl, 1996) ein signifikanter Haupteffekt des Geschlechts auf die *Mathematikleistung*.

Für die Sekundarstufe I ist das Bild der Geschlechterunterschiede für Deutschland nun also hinreichend konsistent. Köller & Klieme (2000) stellen auf der Basis der *TIMSS/III*-Daten eine Fortsetzung dieses Trends mit tendenziell größeren Unterschieden auch für die Sekundarstufe II fest, wobei sie differenzierte Ergebnisse bzgl. verschiedener *Typen mathematischen Arbeitens* und verschiedener Stoffgebiete erhalten. Diese Sicht auf verschiedene Komponenten von *Mathematikleistung* wird ebenfalls im nächsten Abschnitt vertieft.

Zuvor wird aber noch dargestellt, inwieweit Geschlechterunterschiede in der *Mathematikleistung* auch in der Primarstufe in Deutschland auftreten. Winkelmann & van den Heuvel-Panhuizen (2009) gehen dieser Frage anhand der Daten aus der Normierungsstudie für die KMK-Bildungsstandards (KMK, 2005b) nach. Die Ergebnisse dieser Studie sind aufgrund der großen Stichprobe (über 10 000 Schülerinnen und Schüler) und der großen Aufgaben-

zahl (über 500 Items) äußerst belastbar, zumal die Standards in ihrer fachlichen Breite berücksichtigt wurden. Getrennt nach den fünf Leitideen ergeben sich signifikante Geschlechterunterschiede zugunsten der Jungen mit Effektstärken[44] zwischen 0,14 („Raum und Form“) und 0,36 („Größen und Messen“). Die für Deutschland konsistent beobachtbaren Geschlechterunterschiede in der Mathematikleistung treten also anscheinend nicht erst nach dem Übergang in die Sekundarstufe I oder nach Beginn der Pubertät auf.

Geschlechterunterschiede in verschiedenen Komponenten der Mathematikleistung

Studien, die sich dezidiert mit Geschlechterunterschieden in der *Mathematikleistung* auseinandersetzen, berücksichtigen in der Regel, dass *Mathematikleistung* insofern kein homogenes Konstrukt ist, als ganz unterschiedliche kognitive Prozesse bei der Bearbeitung von Mathematikaufgaben involviert sind. Dementsprechend wird *Mathematikleistung* nach unterschiedlichen Modellen in Komponenten aufgeteilt, für die jeweils Geschlechterunterschiede untersucht werden.

Zunächst können Testaufgaben relativ schlicht nach (empirisch) „einfach“ und „schwierig“ sortiert werden, um zu sehen, ob mögliche Geschlechterunterschiede mit der Aufgabenschwierigkeit korrespondieren. Dieses Vorgehen ist letztlich in anderen Fächern ähnlich möglich. Fachspezifisch sind die folgenden Varianten für Komponenten-Modelle:

- Da das Lehren und Lernen von Mathematik traditionell sehr stark an *Stoffgebieten* bzw. *Leitideen* orientiert ist, sind viele Komponenten-Modelle entsprechend strukturiert; dies gilt vor allem für ältere Studien.
- In Deutschland werden allerdings spätestens seit den KMK-Bildungsstandards (KMK 2004, 2005a, 2005b) auch *allgemeine mathematische Kompetenzen* stärker berücksichtigt. Dementsprechend werden z. B. bei Winkelmann & van den Heuvel-Panhuizen (2009) die Testaufgaben u. a. unter dieser Perspektive geclustert.
- Eine weitere Variante der Clusterung berücksichtigt schließlich die *kognitiven Prozesse* bei der Aufgabenbearbeitung und unterscheidet *Typen mathematischen Arbeitens* (vgl. auch „Schwierigkeitsbestimmende Faktoren bei Testaufgaben“, Kap. 2.1.3).

Im Folgenden werden einige differenzierte Befunde zu Geschlechterunterschieden nach den oben genanten Komponenten-Modellen berichtet und anschließend übergreifend diskutiert.

[44] Effektstärken geben in diesem Zusammenhang an, wie groß entsprechende Mittelwertunterschiede im Verhältnis zur Variabilität der Testleistung in der Gesamtpopulation sind. Für die Berechnung von Effektstärken gibt es verschiedene Verfahren, deren Einsatz zum Teil davon abhängig ist, wie die Verteilung der Testleistung in der Gesamtpopulation und in den betrachteten Gruppen gestaltet ist. Die einfachste Variante besteht in der Division der Mittelwertdifferenz durch die Standardabweichung, sodass die Effektstärke dann etwaige Unterschiede als Vielfache der Standardabweichung angibt.

- Mit Blick auf die *empirische Schwierigkeit* der Aufgaben ergibt sich ein relativ einheitliches Bild, nachdem Geschlechterunterschiede zugunsten männlicher Subpopulationen vor allem im Bereich der schwierigen Aufgaben entstehen. So berichten Zimmer et al. (2004, S. 217) für *PISA 2003*, dass die Anteile von Mädchen und die Anteile von Jungen, die auf den unteren Kompetenzstufen der Mathematikskala liegen, gleich groß sind (Mädchen: 21,3 %; Jungen: 21,4 %). Auf den oberen Kompetenzstufen ist der Anteil der Jungen (18,3 %) aber deutlich größer als der Anteil der Mädchen (14,2 %). Dieses Ergebnis bestätigt den Befund von Manger & Eikeland (1998, S. 17): „Boys had significantly higher mean mathematics scores than girls. Significant sex differences favouring boys were found in the subsamples of most difficult tasks, but not in the subsamples of easiest tasks."
- Ein weniger klares Bild ergibt sich, wenn Testaufgaben nach *fachlich-inhaltlichen Bereichen* geclustert werden. Bereits weiter oben wurde berichtet, dass Hyde et al. (1990) bei globaler Betrachtung praktisch keine Geschlechterunterschiede in der *Mathematikleistung* gefunden haben. Allerdings erzielen Jungen bei *Geometrieaufgaben* bessere Ergebnisse (Effektstärke $d = 0{,}13$) und Mädchen bei einfachen *Arithmetikaufgaben* ($d = -0{,}14$). Für den Mathematikunterricht der gymnasialen Oberstufe berichten Köller & Klieme (2000, S. 402) auf der Basis der *TIMSS/III*-Erhebung Geschlechterunterschiede in den Bereichen *Zahlen, Gleichungen und Funktionen*, *Analysis* und *Geometrie* sowohl für die Jahrgangsstufe als auch getrennt nach Grund- und Leistungskursen. Die Effektstärken betragen zwischen 0,02 für *Geometrieaufgaben* in Grundkursen und 0,29 für *Zahlen, Gleichungen und Funktionen* in Leistungskursen bzw. 0,37 für die gesamte Jahrgangsstufe in Bereich *Zahlen, Gleichungen und Funktionen*.[45] Während Hyde et al. (1990) bei Geometrieaufgaben die größten Leistungsunterschiede zugunsten von Jungen identifizieren, ist bei Köller & Klieme (2000) der Unterschied in diesem Bereich am kleinsten (sowohl in den einzelnen Kursarten als auch auf Ebene des Jahrgangs). Interessant ist auch, dass in *Analysis* die Effektstärke für Grundkurse nur bei 0,04 beträgt, für Leistungskurse aber 0,26. Diese Ergebnisse deuten daraufhin, dass die *fachlich-inhaltlichen Bereiche* keine homogenen Cluster bezüglich der Geschlechterunterschiede darstellen.
- Im Rahmen der Normierung der KMK-Bildungsstandards für die Grundschule (KMK, 2005b) untersuchen Winkelmann & van den Heuvel-Panhuizen (2009) Geschlechterunterschiede auch in Aufgabenclustern, die die *allgemeinen mathematischen Kompetenzen* der Standards abbilden (*Problemlösen*, *Argumentieren*, *Kommunizieren*, *Modellieren* und *Darstellen*). Bei den für das Fach charakteristischen Prozessen *Problemlösen*, *Ar-*

[45] Die Maße für die gesamte Jahrgangsstufe sollten allerdings nicht inhaltlich interpretiert werden, da auch hier ein „Simpson-Paradoxon" vorliegt: Die Effektstärken sind auf Ebene der Jahrgangstufe in allen drei Bereichen jeweils größer ist als beide entsprechenden Effektstärken für Grund- bzw. Leistungskurse; hier wirkt sich die höhere Beteilung von Jungen an den Leistungskursen aus.

gumentieren und *Modellieren* beträgt die Effektstärke etwa ein Viertel (jeweils zugunsten der Jungen), beim *Kommunizieren* 0,18 und die Tendenz beim *Darstellen* wird mit 0,10 nicht signifikant. Im Rahmen von *DIF*-Analysen[46] werden allerdings in allen genannten Bereichen jeweils sowohl Items identifiziert, die von Mädchen im Vergleich zu Jungen unerwartet gut gelöst wurden, als auch Items, die von Jungen im Vergleich zu Mädchen unerwartet gut gelöst wurden. Diese Ergebnisse deuten daraufhin, dass auch die *allgemeinen mathematischen Kompetenzen* bezüglich der Geschlechterunterschiede keine homogenen Cluster darstellen.

Von den drei zuvor betrachteten Komponenten-Modellen liefert am ehesten die Clusterung von Testaufgaben nach ihrer *empirischen Schwierigkeit* eindeutige Ergebnisse. Sowohl bei den *fachlich-inhaltlichen Bereichen* als auch bei den *allgemeinen mathematischen Kompetenzen* scheinen innerhalb eines jeden Clusters verschiedene Aufgabenmerkmale auftreten zu können, die zum Teil die Lösungsquoten zugunsten der Jungen und zum Teil die Lösungsquoten zugunsten der Mädchen verschieben können. Dies wird mit Blick auf die Fragestellung der vorliegenden Arbeit exemplarisch für *Geometrie*, also einen *fachlich-inhaltlichen Bereich* ausgeführt: Hier gibt es sowohl Aufgaben, die eine mentale oder reale Transformation figuraler Gegebenheiten erfordern als auch Aufgaben, die direkt auf eine Berechnung führen. An eine Clusterung nach *allgemeinen mathematischen Kompetenzen* lassen sich ebenfalls kritische Fragen stellen, z. B. warum eine Unterscheidung der Kategorien mit Blick auf Geschlechterunterschiede sinnvoll sein sollte – konkret: was unterscheidet *Modellieren* von *Problemlösen* derart, dass in diesen Bereichen unterschiedliche Befunde zu erwarten wären. Es kann zwar curricular sinnvoll sein, beide Kategorien separat auszuweisen, aber aus der Sicht der ablaufenden kognitiven Prozesse sind beide Kategorien doch sehr ähnlich (vgl. Büchter & Leuders, 2005b, S. 30 f.). Und auch die Clusterung nach Aufgabenschwierigkeit ist inhaltlich noch nicht befriedigend, da ganz unterschiedliche Faktoren zur Schwierigkeit einer Aufgabe beitragen können. So können rein reproduktive Aufgaben nahezu beliebig schwer sein.[47]

Der vierte oben genannte Ansatz zur Clusterung berücksichtigt stärker die kognitiven Prozesse, also die geistigen Tätigkeiten bei der Aufgabenlösung. Wenn Geschlechterunterschiede zumindest teilweise kognitiv erklärt werden können, dann führt diese Clusterung möglicherweise zu Aufgabengruppen, die homogen bezüglich der Geschlechterunterschiede sind. Dabei werden zunächst die *Typen mathematischen Arbeitens* betrachtet, die bereits oben unter „Schwierigkeitsbestimmende Faktoren bei Testaufgaben“ untersucht wurden

[46] Mit „Differential Item Functioning (DIF)“ wird für unterschiedliche Gruppen (wie „Jungen vs. Mädchen“) untersucht, ob Items „*unerwartet gut*“ bzw. „*unerwartet schlecht*“ gelöst wurden. Für diesen Zweck wird aus der Gesamttestleistung der beiden Gruppen berechnet, welche Lösungsquoten jeweils auf dieser Basis zu erwarten wären, und das Ergebnis mit den realen Lösungsquoten verglichen.

[47] Bei *PISA 2000* war das Lösen einer quadratischen Gleichung die (empirisch) schwierigste Aufgabe!

(*technische Aufgaben*, *rechnerische Modellierungsaufgaben* und *begriffliche Modellierungsaufgaben*):

- Zwar liegen keine Untersuchungen vor, die explizit auf diese Dreiteilung von Aufgaben Bezug nehmen, aber Köller & Klieme (2000) berichten im Rahmen von *TIMSS/III*, dass die Effektstärken für Geschlechterunterschiede im Bereich *Routineverfahren* im Grundkurs bei –0,11, also zugunsten der Mädchen, und im Leistungskurs bei 0,11, also wieder zugunsten der Jungen, liegen. Bei *komplexen Verfahren* und *Anwenden/Problemlösen* liegen ausschließlich Leistungsunterschiede zugunsten der Jungen vor. Sie betragen im Grundkurs 0,14 (*komplexe Verfahren*) bzw. 0,08 (*Anwenden/Problemlösen*) und im Leistungskurs 0,32 (*komplexe Verfahren*) bzw. 0,30 (*Anwenden/Problemlösen*). Bei dieser Betrachtung sind *Routineverfahren* also die relative Leistungsstärke der Mädchen im Vergleich zu den Jungen. Dieses Ergebnis wird durch analoge Befunde in weiteren Studien unterstützt. So berichtet Klieme (1986) von einem Forschungsstand, der als Leistungsstärke der Mädchen im Vergleich zu den Jungen *algorithmisches Rechnen* identifiziert. Stanat & Kunter (2001) berichten:

  > „Auch in der Mathematik zeigen sich in differenziellen Itemanalysen … anforderungsspezifische Geschlechterdifferenzen. Während Mädchen ihren Leistungsschwerpunkt bei technischen Aufgaben und im innermathematischen Kontext haben, sind bei den Jungen relative Stärken beim rechnerischen Modellieren zu beobachten sowie bei der Mathematisierung von Situationen, wenn mehrere Lösungsansätze denkbar sind“ (ebd., S. 257).

  Möglicherweise lassen sich mit den Betrachtungen zu den *Typen mathematischen Arbeitens* auch die oben genannten erheblichen Unterschiede zwischen den Effektstärken im Bereich Analysis zwischen Grundkursen (d = 0,04) und Leistungskursen (d = 0,26) erklären. Obwohl der deutsche Analysisunterricht in der Breite ohnehin stark kalkülorientiert ist (vgl. Borneleit et al., 2001), dürfte dies im besonderen Maße auf die Grundkurse zutreffen. Über die dort üblichen *Routineverfahren* gehen Leistungskurse häufig insofern hinaus, als sie auch innermathematisches Problemlösen und Argumentieren sowie anspruchsvollere Anwendungen (gelegentlich) berücksichtigen.

Wenn man die Betrachtungen zu Geschlechterunterschieden in unterschiedlichen Komponenten der *Mathematikleistung* zusammenfasst, so scheint für die vorliegende Arbeit am ehesten eine differenzierte Betrachtung der *Mathematikleistung* im Sinne der *Typen mathematischen Arbeitens* weiterzuführen. Anders als bei den Komponenten-Modellen, die curricular oder empirisch bedingt sind, führt die kognitive Perspektive möglicherweise zu konsistenteren Ergebnissen und inhaltlich plausiblen Zusammenhangsvermutungen mit anderen kognitiven Leistungen wie *Raumvorstellung*.

Mögliche Gründe für Geschlechterunterschiede

Eine pädagogisch-didaktisch wichtige Frage bei empirisch festgestellten Leistungsunterschieden zwischen Jungen und Mädchen bzw. Männern und Frauen ist die nach den Hintergründen. Köller & Klieme (2000, S. 376) geben eine Übersicht über typische Erklä-

rungsansätze für Geschlechterunterschiede. Sie unterscheiden *biologische Ansätze*, *kognitive Ansätze*, *psychosoziale Modelle* und *Unterrichtsmodelle* und betonen, dass es zu jedem Ansatz sowohl empirische Unterstützung als auch empirischen Widerspruch gibt. Plausibel scheint zu sein, dass Geschlechterunterschiede nur mit einem komplexen Wirkungsgefüge aus allen genannten Bereichen (empirisch wie theoretisch) erklärt werden können.

Für die vorliegende Arbeit sind vor allem *kognitive Ansätze*, *psychosoziale Modelle* und *Unterrichtsmodelle* interessant, da diese zumindest prinzipiell durch die pädagogisch-didaktische Gestaltung von Schule und Unterricht beeinflusst bzw. berücksichtigt werden können. Im Bereich der *kognitiven Ansätze* verweisen Köller & Klieme vor allem auf die *Raumvorstellung*. Dieser Aspekt wird im Rahmen der „spatial mediation hypothesis“ in Kap. 3.3.6 genauer betrachtet. *Psychosoziale Modelle* berücksichtigen u. a. *Geschlechtsrollenstereotype* („Mathematik ist eher eine ‚männliche' Domäne“), geringeres Interesse am Fach (vermutlich mit den *Geschlechtsrollenstereotypen* zusammenhängend) und das *Fähigkeitsselbstkonzept*. *Unterrichtsmodelle* berücksichtigen Fragen wie die Lehrer-Schüler-Interaktion und die Auswahl von Aufgabenkontexten in Schulbüchern.

Im Rahmen der empirischen Bildungsforschung sind vor allem die *psychosozialen Modelle* als ein Schwerpunkt der pädagogischen Psychologie gut berücksichtigt. Entsprechende Subtests werden bei allen größeren Schulleistungsstudien eingesetzt. An dieser Stelle werden exemplarisch Befunde aus drei Studien genannt, die auf die potenzielle Bedeutung der fraglichen Konstrukte für Geschlechterunterschiede hinweisen.

Zimmer et al. (2004) haben im Rahmen von *PISA 2003* verschiedene Selbsteinschätzungen von Schülerinnen und Schülern in Bezug auf das Fach Mathematik ausgewertet, darunter das *Fähigkeitsselbstkonzept* und die *Selbstwirksamkeitserwartungen*. Es zeigt sich, dass Mädchen sowohl auf den unteren Kompetenzstufen als auch auf den oberen Kompetenzstufen der Mathematikskala jeweils ein geringer ausgeprägtes *Fähigkeitsselbstkonzept* und geringere *Selbstwirksamkeitserwartungen* haben als Jungen auf denselben Kompetenzstufen. Allerdings verringern sich diese Unterschiede beim *Fähigkeitsselbstkonzept* von einer Effektstärke von 0,52 auf den unteren Kompetenzstufen auf 0,18 auf den oberen Kompetenzstufen. Zugleich liegt das *Fähigkeitsselbstkonzept* von Mädchen auf den oberen Kompetenzstufen mit einem z-standardisierten[48] Wert von 0,34 über dem Vergleichswert der Jungen auf den unteren Kompetenzstufen, der 0,04 beträgt. Inhaltlich ist naheliegend, dass das *Fähigkeitsselbstkonzept* in permanenter Wechselwirkung mit der Fachleistung steht, auch wenn in Pfadmodellen manchmal eine einseitige Wirkung postuliert wird bzw. (aus verfahrenstechnischen Gründen) postuliert werden muss.

[48] Bei einer z-Standardisierung wird ein Datensatz so transformiert, dass das arithmetische Mittel 0 und die Standardabweichung 1 beträgt.

Eine relativ komplexe Modellierung psychosozialer Einflüsse liefert Ethington (1992) mit ihrem „Psychological Model of Mathematics Achievement“. Sie analysiert die *Mathematikleistungen* von Jungen und Mädchen in einem Pfadmodell mit 12 Konstrukten, das strukturell auf Eccles et al. (1983) zurückgeht, potenzielle Einflussfaktoren auf *Mathematikleistung*. Dabei nutzt sie Daten der „Second International Mathematics Study (SIMS)“ für die Schätzung der Pfadgewichte. Das Pfadmodell enthält pädagogisch-psychologische Konstrukte, denen ein großes Potenzial für die Erklärung von Leistungsunterschieden im Allgemeinen und von Geschlechterunterschieden im Besonderen zugeschrieben werden, so z. B. das *Fähigkeitsselbstkonzept*, *Selbstwirksamkeitserwartungen* oder *Geschlechtsrollenstereotype*. Berücksichtigt man in diesem Modell nur die Pfade, deren standardisierte Pfadgewichte sich signifikant von Null unterscheiden, so ergibt sich bei Mädchen ein deutlich anderes Modell der Bedingtheit von *Mathematikleistung* als bei Jungen.

In eine ähnliche Richtung weisen Befunde von Tartre & Fennema (1995), die ebenfalls eine Reihe ausgewählter Variablen bezüglich ihres Einflusses auf *Mathematikleistung* im Rahmen einer Längsschnittuntersuchung mit 60 Versuchspersonen über die Jahrgangsstufen 6 bis 12 untersuchen. Über alle Messzeitpunkte und alle möglichen linearen Zusammenhänge hinweg stellen die Autorinnen fest, dass das *Fähigkeitsselbstkonzept* bei Mädchen stärker mit *Mathematikleistung* zusammenhängt als bei Jungen. Fachbezogene *Geschlechtsrollenstereotype* („Mathematik ist eher eine ‚männliche’ Domäne“) hängen nur bei Mädchen signifikant mit *Mathematikleistung* zusammen.[49]

Insgesamt lässt sich also folgern, dass auch psychosoziale Konstrukte eine Rolle im Bedingungsgefüge für *Mathematikleistung* spielen und bei Jungen und Mädchen unterschiedlich wirken. Insbesondere für die Erklärung von Geschlechterunterschieden scheinen sie also von Bedeutung zu sein.

2.3.3 Dimensionalität von Mathematikleistung

Wenn man die Frage nach der Dimensionalität von *Mathematikleistung* zuspitzt auf die Frage, ob *Mathematikleistung* nicht letztlich eindimensional ist, so befindet man sich im Epizentrum vieler Kontroversen zwischen Psychometrie und Mathematikdidaktik, aber auch innerhalb der Psychometrie und innerhalb der Mathematikdidaktik wird hierüber gestritten. An dieser Stelle soll kein Scheinkonflikt zwischen Disziplinen herbeigeredet werden, die innerhalb der empirischen Bildungsforschung produktiv zusammenarbeiten. Es soll aber festgehalten werden, dass diese Frage den Nerv der Mathematikdidaktik trifft. Je nach Zugangsweise, z. B. aus Sicht der Schulpraxis bzw. schulnahen Entwicklungsforschung oder aus der Sicht einer geisteswissenschaftlichen Mathematikdidaktik, ist die komplexe Struktur von *Mathematikleistung* unstrittig. Lehrkräfte kennen genügend Bei-

[49] Je weniger das o. g. Stereotyp ausgeprägt ist, desto besser ist die Leistung.

spiele ganz unterschiedlicher mathematischer Begabungsprofile und müssen im Alltag sehr differenziert wahrnehmen, um individuelle Schülervorstellungen zu verstehen. Und auch die mathematikdidaktisch orientierte empirische Bildungsforschung liefert durchaus empirische Befunde, die mehr als eine Dimension nahe liegen.

Eine Versachlichung der zuweilen emotional bis verbittert geführten Diskussionen kann gelingen, wenn die Fragestellung umformuliert und weiter präzisiert wird. Die Frage, ob Fachleistungen, insbesondere *Mathematikleistung*, hinreichend gut eindimensional gemessen werden können, ergibt sich vor allem aus großen Vergleichsuntersuchungen. Wenn etwa die *Mathematikleistungen* der 15-Jährigen aus verschiedenen Bildungssystemen miteinander verglichen werden soll, dann ist ein Vergleich auf einer Dimension am einfachsten und übersichtlichsten (auch wenn er direkt die Gefahr mit sich bringt, im Sinne eines Rankings verstanden zu werden). Hält man dagegen eine mehrdimensionale Modellierung für angemessener, so müssen Vergleiche für alle Dimensionen einzeln durchgeführt werden. Je nach Zweck einer Untersuchung ist dies möglicherweise nicht sachdienlich. So ist *PISA* z. B. „Teil des Indikatorenprogramms der OECD, dessen Ziel es ist, den OECD-Mitgliedsstaaten vergleichende Daten über die Ressourcenausstattung, individuelle Nutzung sowie Funktions- und Leistungsfähigkeit ihrer Bildungssystem zur Verfügung zu stellen“ (Baumert et al, 2001, S. 15). Für einen solchen Zweck kann *eine* Skala für Mathematik (neben den vielen anderen Skalen) ausreichen.

Aus Sicht der Schulleistungsstudien ist die Frage nach der Dimensionalität in der Regel keine Frage nach der „Natur“ der *Mathematikleistung*, sondern eine Frage nach möglichst einfacher und zugleich möglichst angemessener statistischer Modellierung. „Einfach“ und „angemessen“ sind dabei tendenziell gegenläufige Anforderungen: Einerseits kann bei den meisten statistischen Verfahren ein Modell umso mehr beobachtete Varianz (der Testleistung) erklären, je mehr Parameter es verwendet, also je komplexer ist[50]. Anderseits kann ein Modell in der Regel umso besser theoretisch beschrieben werden, je einfacher es ist, wenn also möglichst sparsam mit Parametern umgegangen wird. Wie bei jeder Modellierung ist hier ein (subjektiver) normativer Akt erforderlich, in dem die Kriterien „Sparsamkeit“ und „Anpassung“ so gewichtet werden, dass es zur Zielsetzung des Vorhabens passt.

„Skalierungspragmatismus“

Von den *TIMSS*-Untersuchungen der 1990er-Jahre über *PISA 2000*, *2003* und *2006* hat sich ein „Skalierungspragmatismus“ (Baumert, Köller et al., 2000, S. 67) etabliert, bei dem Fachleistungen überwiegend eindimensional modelliert werden – wohl wissend, dass viele unterschiedliche kognitive Prozesse bei der Bearbeitung eines Mathematiktests einbezogen

[50] Diese Eigenschaft der meisten Verfahren kann intuitiv gut verstanden werden: Je mehr Parameter man im Rahmen des Verfahrens schätzen kann, desto mehr Anpassungsmöglichkeiten an die gegebenen Daten hat man.

sind. Die eindimensional modellierte Personeneigenschaft wird daher als „mathematisches Fähigkeitssyndrom“ (Köller, 1998) bezeichnet. Baumert, Köller et al. (2000) vergleichen den Skalierungspragmatismus mit der Vergabe von Schulnoten und zeigen auch Grenzen auf:

> „Dieser Pragmatismus korrespondiert sehr gut mit der schulischen Beurteilungspraxis. Schüler erhalten am Ende eines Schuljahres eben eine Mathematiknote und nicht eine Teilnote für Sachgebiet A, eine für Sachgebiet B usw. Dennoch wird es bei spezifischen Fragestellungen immer wieder von Interesse sein, Items nach fachlichen/fachdidaktischen oder psychologischen Kriterien neu zu gruppieren, um mit diesen Subdimensionen zu arbeiten“ (ebd., S. 67).

Sill (2010) zieht auch den Vergleich von eindimensionaler Skalierung und Schulnoten, betont aus mathematikdidaktischer und schulpraktischer Sicht aber vor allem die Grenzen:

> „Die eindimensionale Sichtweise hat den gleichen Wert wie die Mathematiknote auf dem Zeugnis eines Schülers. Für den konkreten Unterricht oder die selbstständige Behebung eigener Defizite ist sie kaum verwendbar. Lehrer und Schüler möchten schon sehr genau wissen, was sie in den einzelnen Themengebieten bei den jeweils konkreten Anforderungen zu erreichen haben“ (ebd., S. 6 f.).

Betrachtet man z. B. die oben dargestellten Befunde zu Geschlechterunterschieden in verschiedenen Komponenten der *Mathematikleistung*, so wird klar, dass eine mehrdimensionale Sichtweise – je nach Fragestellung – auch für die empirische Erforschung von *Mathematikleistung* weiterführend sein kann.

Eindimensionale Ansätze

Innerhalb der Ergebnisberichte zu Schulleistungsstudien wurde die eindimensionale Skalierung und Berichtslegung immer wieder mit hohen latenten Korrelationen[51] zwischen möglichen Subskalen der *Mathematikleistung* begründet (vgl. z. B. Klieme et al., 2001, S. 156 ff.). Am Beispiel der Dimensionalitätsüberlegungen im *TIMSS/III*-Bericht zur „voruniversitären Mathematik“ (Sachgebiete: *Analysis*, *Geometrie* sowie *Zahlen, Gleichungen und Funktionen*) lässt sich ein reflektierter Skalierungspragmatismus beobachten. Zunächst vergleicht Klieme (2000) die Anpassungen eines eindimensionalen und eines dreidimensionalen *RM* unter Berücksichtigung der Parameterzahl und bevorzugt das dreidimensionale Modell mit den o. g. Sachgebieten als Dimensionen. Die zwischen je zwei dieser drei Dimensionen geschätzten latenten Korrelationen betragen zwischen 0,77 und 0,81. Dieses Ergebnis fasst der Autor wie folgt zusammen: „Die Höhe der Korrelationen rechtfertigt die Zusammenfassung der drei Skalen zu einem Gesamtwert. Auf der anderen Seite legen die Befunde auch untertestspezifische Analysen nahe, da diese Korrelationen substantiell unter 1 liegen“ (ebd., S. 64). Dass eine Zusammenfassung der drei Skalen zu einem Gesamtwert

[51] Latente Korrelationen sind messfehlerfreie Korrelationen, deren Werte z. B. im mehrdimensionalen *RM* als Modellparameter gemeinsam mit den Item- und Personenparametern aus den Testdaten geschätzt werden.

trotz substanziell von 1 verschiedenen Korrelationen nicht ausgeschlossen wird, dürfte eine pragmatische Entscheidung im Rahmen der Vergleichsuntersuchung sein.

Interessant ist in diesem Zusammenhang, wie unterschiedlich latente Korrelationen zwischen Subskalen gedeutet werden. So untersucht Klieme (2000, S. 65) auch eine Strukturierung des Tests nach den drei „Verhaltenserwartungen" *Routineverfahren*, *komplexe Verfahren* und *Anwenden/Problemlösen* und kommt aufgrund von latenten Korrelationen zwischen 0,82 und 0,87 zu dem Schluss, dass die Verhaltenserwartungen nicht als qualitativ unterscheidbar angesehen werden sollten. Im Gegensatz dazu sehen Blum et al. (2004, S. 62) bei ihren „Typen mathematischen Arbeitens" (*technische Aufgaben*, *rechnerische Aufgaben* und *begriffliche Aufgaben*), die Kliemes Verhaltenserwartungen konzeptionell entsprechen, dass sich der Typ *technische Aufgaben* aufgrund einer latenten Korrelation von 0,89 zu beiden anderen Aufgabentypen „deutlich" von diesen beiden Typen abhebt – obwohl die latente Korrelation noch höher war als die von Klieme beobachteten. Neben einem Skalierungspragmatismus gibt es anscheinend auch einen Interpretationspragmatismus für solche Parameter.

Von Befunden und Interpretationen wie denen von Klieme (2000) ausgehend werden nicht nur Schulleistungen verglichen, sondern auch pädagogisch-psychologische Grundlagen für weitere Leistungsuntersuchungen gelegt. So kommen Köller & Baumert (2002), insbesondere mit Blick auf die *Mathematikleistung* und Kliemes Befunde, zu folgendem Schluss:

> „Die Befundlage impliziert insgesamt, dass die fachspezifischen Leistungen in Schulleistungstests durch ein komplexeres Fähigkeitssyndrom beeinflusst werden, so dass in empirischen Untersuchungen innerhalb eines einzelnen Faches curriculare oder kognitive Teilkomponenten analytisch kaum trennbar sind, man bei der Untersuchung von schulischen Leistungsverläufen also keinen substanziellen Fehler macht, wenn man die Leistungen innerhalb eines Faches als unidimensional ansieht" (Köller & Baumert, 2002, S. 766).

Auf der Basis von diesem und ähnlichen Schlüssen über die Struktur der Fachleistung wird bei der Testentwicklung häufig schon in dem Sinne „eindimensional gedacht", dass bei der empirischen Erprobung von Aufgaben die Passung in ein eindimensionales *RM* ein Selektionskriterium ist. Passende Aufgaben werden dann in den Hauptuntersuchungen verwendet und die Daten der Hauptuntersuchung deuten wiederum auf Eindimensionalität hin. Auch wenn ein möglicher Zirkelschluss hier stark vereinfacht dargestellt wird, hat seine Logik die empirische Bildungsforschung der jüngeren Vergangenheit in Teilen mitgeprägt.

Mehrdimensionale Ansätze

Die fortgesetzte psychometrische und fachdidaktische Diskussion sowie weitere Befunde haben allerdings auch zu konzeptionellen Weiterentwicklungen geführt. Im Rahmen der Normierungsstudie zu den KMK-Bildungsstandards für die Grundschule (KMK, 2005b) wurden dabei auch die verwendeten Testmodelle und Skalierungsverfahren untersucht. Das seit den 1990er-Jahre übliche Vorgehen war, dass in Schulleistungsstudien jede einzelne Testaufgabe immer genau einer der inhaltlichen Dimension des zugrundeliegenden Kom-

petenzbereichsmodells zugeordnet wird („between-item-dimensionality“). Aufgrund der hochgradigen inneren Vernetzung der Schulmathematik ist es aber durchaus nahe liegend, dass einzelne Aufgaben auch zu mehreren inhaltlichen Dimensionen gehören („within-item-dimensionality“). In Simulationsstudien zeigt Robitzsch (2009, S. 54 ff.), dass die latenten Korrelationen systematisch überschätzt werden, wenn Aufgaben eigentlich zu mehreren Dimensionen gehören, im Modell aber nur einer Dimension zugeordnet werden. Vor diesem Hintergrund ist es möglich, dass viele latente Korrelationen aus den oben berichteten Studien artifiziell (etwas) zu hoch geschätzt wurden.

Dennoch ergeben Modellvergleiche für die fünf Leitideen der Standards, dass ein mehrdimensionales „between-Modell“ die Testdaten etwas besser erklären kann als ein mehrdimensionales „within-Modell“; besonders bemerkenswert ist dabei, dass beide mehrdimensionalen Modelle deutlich besser auf die Daten passen als das alternative eindimensionale Modell (vgl. Winkelmann & Robitzsch, 2009, S. 191 f.). Die latenten Korrelationen zwischen je zwei der fünf Leitideen liegen im „between-Modell“ – vermutlich etwas überschätzt – zwischen 0,73 (*Raum und Form* mit *Zahlen und Operationen*) und 0,85 (*Größen und Messen* mit *Zahlen und Operationen*). Im „within-Modell“ liegen diese Korrelationen – nun etwas unterschätzt – zwischen 0,42 (*Muster und Strukturen* mit *Raum und Form*) und 0,69 (*Größen und Messen* mit *Zahlen und Operationen*).

Mit diesen Befunden wird bei der Normierung der Grundschulstandards wiederum pragmatisch umgegangen. So liefert die Forschergruppe sowohl ein inhaltlich beschriebenes Kompetenzstufenmodell für eine globale Leistungsskala Mathematik, da „alle Dimensionen erhebliche Anteile gemeinsamer Varianz haben“ (ebd., S. 187), als auch inhaltlich beschriebene Kompetenzstufenmodelle für die fünf Leitideen, da zugleich „Evidenz für die analytische Separierbarkeit der inhaltlichen Kompetenzen“ (ebd., S. 187) vorliegt.

Bevor die Befunde der modernen empirischen Bildungsforschung für die vorliegende Arbeit zusammengefasst werden, ist noch eine kritische Betrachtung der Höhe der latenten Korrelationen aus institutioneller Perspektive wichtig. Die berichteten latenten Korrelationen fallen in der Sekundarstufe I wahrscheinlich auch deswegen höher aus als in der Primarstufe, weil – wie schon oben im Zusammenhang mit dem „*TIMSS/II*-Paradoxon“ dargestellt wurde – Schulformen differenzielle Entwicklungsmilieus darstellen, die die Leistungsunterschiede zwischen Schülerinnen und Schülern z. B. aus Hauptschulen und Gymnasien (aus Sicht des Mathematiklernens künstlich) vergrößern. Wenn aber in Hauptschulen, u. a. durch diesen Effekt bedingt, tendenziell auf allen Dimensionen niedrige und in Gymnasien tendenziell auf allen Dimensionen hohe Leistungswerte anzutreffen sind, so fällt die latente Korrelation zwischen diesen Dimensionen zwangsläufig hoch aus.

Insgesamt kann mit Blick auf die Fragestellung der vorliegenden Arbeit festgestellt werden, dass es auch empirisch angemessen sein kann, *Mathematikleistung* mehrdimensional zu modellieren, wobei die Berücksichtigung inhaltlich tragfähiger Dimensionen zu mehr

differenziellen Befunden führen kann als in eindimensionalen Modellen. Der Zusammenhang von *Raumvorstellung* und *Mathematikleistung* sollte daher auf jeden Fall global und nach fachlichen Dimensionen differenziert untersucht werden. Dieses Fazit passt damit auch gut zu einem Ergebnis, dass Treumann im Jahre 1974, also vor der modernen Bildungsforschung, im Rahmen der Untersuchung von „Leistungsdimensionen im Mathematikunterricht“ erzielt hat:

> „Zusammenfassend lässt sich sagen, daß es zwar bei Verwendung angemessener Meßinstrumente an der differentiellen Struktur mathematischer Leistungen kaum einen Zweifel geben kann, aber es scheint nicht möglich zu sein, die ein für allemal verbindliche kognitive Anforderungsstruktur anzugeben, sondern ihre konkrete Beschaffenheit ist immer auch abhängig von den curricularen und institutionellen Rahmenbedingungen, unter denen der Unterricht abläuft“ (Treumann, 1974, S. 221).

3 Grundlagen und Befunde der Erforschung von Raumvorstellung

Wir Menschen leben in einer allgegenwärtigen räumlichen Welt, die wir oft unbewusst wahrnehmen und in der wir – schon vorgeburtlich – ganz unterschiedliche Erfahrungen sammeln. Dabei entwickeln wir Fähigkeiten, die über die Orientierung im direkt wahrnehmbaren Raum weit hinausgehen und sich auch auf mentale Repräsentationen und Manipulationen von räumlichen Gegebenheiten erstrecken. Diese Fähigkeiten prägen unser Leben „mit Raum“ durch unser mehr oder weniger kompetentes Agieren im Raum. Die lebenspraktische Bedeutung solcher Fähigkeiten wird z. B. erkennbar, wenn wir eine Wegbeschreibung „aus dem Kopf“ geben sollen, wenn wir ein sperriges Möbelstück durchs Treppenhaus tragen müssen, wenn wir eine Kerze „gerade“ in ihren Ständer stellen möchten oder wenn wir das erforderliche Briefporto unter Berücksichtigung von Gewicht und Größe aus einer Tabelle erschließen.

Wenn wir hören, dass jemand über eine gute „Raumvorstellung“ verfügt, assoziieren wir intuitiv (und subjektiv) bestimmte Leistungen hiermit. Im Alltäglichen bedarf es auch kaum einer weiteren Begriffspräzisierung; zur Beschreibung der fraglichen Fähigkeit werden hier – teilweise synonym, teilweise mit unterschiedlichen Bedeutungsschwerpunkten – neben „Raumvorstellung“ auch Bezeichnungen wie „räumliches Denken“, „visuelles Denken“, „räumliche Intelligenz“, „Raumorientierung“, „Raumwahrnehmung“, „räumliches Vorstellungsvermögen“ oder „räumliche Vorstellungsfähigkeit“ verwendet. Für die wissenschaftliche Auseinandersetzung mit entsprechenden Phänomenen, insbesondere für die Theoriebildung, ist eine begriffliche Klärung des fraglichen Gegenstands hingegen grundlegend.

Das Konstrukt *Raumvorstellung*[52] wird an dieser Stelle in enger Anlehnung an Linn & Petersen (1985) sowie Leopold (2002) durch eine (zunächst vorläufige) Arbeitsdefinition abgesteckt:

Raumvorstellung umfasst kognitive Fähigkeiten, die für die mentale Repräsentation und Transformation figuraler Informationen benötigt werden.

[52] In der vorliegenden Arbeit wird, außer in einigen wörtlichen Zitaten, die Bezeichnung „Raumvorstellung“ verwendet, da sie in der wissenschaftlichen Diskussion (in verschiedenen Disziplinen) hinreichend üblich ist und den fraglichen Gegenstand im üblichen Sprachverständnis gut beschreibt. Dabei muss berücksichtigt werden, dass Alternativen wie „räumliches Vorstellungsvermögen“ oder „räumliche Vorstellungsfähigkeit“ ebenfalls üblich sind und ihre Berechtigung haben. Auf eine weitergehende philologische Diskussion der Bezeichnung wird an dieser Stelle aus pragmatischen Gründen zugunsten einer intensiveren Klärung des Begriffs verzichtet.

Das Ziel dieses Kapitels ist es, das Konstrukt *Raumvorstellung* basierend auf dem aktuellen Stand der Forschung so zu diskutieren, dass das am Ende der Einleitung formulierte Erkenntnisinteresse durch Forschungsfragen konkretisiert und mit der Auswahl und Entwicklung von geeigneten Instrumenten operationalisiert werden kann. Im Mittelpunkt steht dabei der Zusammenhang von *Raumvorstellung* und *Mathematikleistung* unter Berücksichtigung (und mit dem Ziel der Erklärung) von Geschlechterunterschieden. Bei der Literaturauswahl und -sichtung stehen die folgenden Leitfragen im Vordergrund:

- Wie können die kognitiven Fähigkeiten, die mit „Raumvorstellung" gemeint sind, inhaltlich beschrieben werden? Wie können sie zuverlässig und valide gemessen werden?
- Wie ist dieser Fähigkeitsbereich strukturiert?
- Wie hängt *Raumvorstellung* mit anderen Leistungsbereichen zusammen?
- Welche interindividuellen Unterschiede, insbesondere welche Geschlechterunterschiede, gibt es bei der *Raumvorstellung*? Wie können sie erklärt werden?
- Wie hängen *Mathematikleistung* und *Raumvorstellung* zusammen? Für welche der jeweiligen Teilleistungsbereiche ist der Zusammenhang besonders groß bzw. besonders klein?
- Inwieweit kann *Raumvorstellung* zur Erklärung von Unterschieden, insbesondere von Geschlechterunterschieden, bei der *Mathematikleistung* beitragen?
- Wie kann die *Raumvorstellung* systematisch gefördert werden?

Wesentliche inhaltliche Bereiche, die durch diese Leitfragen berührt werden, sind Gegenstand psychologischer Forschung, sodass die Psychologie die zentrale Bezugsdisziplin in diesem Kapitel ist. *Raumvorstellung* wird in verschiedenen psychologischen Teildisziplinen aus unterschiedlichen, sich letztlich ergänzenden Perspektiven untersucht. Für die vorliegende Arbeit sind insbesondere die folgenden Perspektiven wichtig:

- Psychometrie – die objektive, zuverlässige und valide Messung der *Raumvorstellung*
- Differenzielle Psychologie – interindividuelle Unterschiede bei der *Raumvorstellung*
- Kognitionspsychologie – an der *Raumvorstellung* beteiligte kognitive Prozesse

Neben der Bezugsdisziplin *Psychologie* befasst sich natürlich auch die Mathematikdidaktik selbst mit *Raumvorstellung*, die besonders im Kontext der Geometrie, vor allem der *Raum*geometrie, diskutiert wird. Darüber hinaus ist *Raumvorstellung* immer dann wichtig, wenn die Arbeit mit strukturierten graphischen, symbolischen oder numerischen Darstellungen wichtig ist, wie z. B. die Orientierung in der Hundertertafel im Anfangsunterricht oder die Beobachtung des Ko-Variationsverhaltens einer Funktion an ihrem Graphen.

Im Folgenden wird zunächst untersucht, wie sich das Konstrukt *Raumvorstellung* aus psychologischer Perspektive darstellen lässt (Kap. 3.1). Anschließend wird – eher kontrastierend – dargestellt wie sich die Mathematikdidaktik mit diesem Gegenstand beschäftigt

(Kap. 3.2). Dabei wird sich zeigen, dass für die eigene empirische Untersuchung überwiegend psychologische Ansätze relevant sind. Schließlich werden mit Blick auf die oben genannten Leitfragen ausgewählte Befunde empirischer Forschung zusammengefasst (Kap. 3.3), bevor – als Überleitung zum empirischen Teil dieser Arbeit – eine pragmatische Festlegung des Konstrukts *Raumvorstellung* vorbereitet wird (Kap. 3.4).

3.1 Raumvorstellung als Gegenstand der Psychologie

Raumvorstellung wird in der Psychologie als eine zentrale kognitive Fähigkeit betrachtet und ist ein wesentlicher Bereich bei praktisch allen Modellen von Intelligenz und Begabung. So ist es nicht erstaunlich, dass *Raumvorstellung* ein klassischer Gegenstand und einer der am intensivsten untersuchten Gegenstände der empirisch-psychologischen Forschung ist (vgl. D. H. Rost, 1997, S. 59; Wiedenbauer, 2006, S. 9). In Kap. 3.1.1 wird die Erforschung der *Raumvorstellung* im Kontext der Entwicklung von Intelligenzmodellen dargestellt.

Die empirische Untersuchung der *Raumvorstellung* geht dabei bis ins 19. Jahrhundert zurück. So berichtet Galton (1880; 1883) – bekanntlich in vielen wissenschaftlichen Disziplinen ein Pionier quantitativ-empirischen Arbeitens – von eigenen Untersuchungen zu bildlichen mentalen Vorstellungen („Mental Imagery“), die anfangs auf den Selbstauskünften der Versuchspersonen beruhten. Schon Galton hat bedeutende interindividuelle Unterschiede, insbesondere auch Geschlechterunterschiede, in der Ausprägung der von ihm betrachteten Fähigkeiten festgestellt. Solche interindividuellen Unterschiede und die Bedeutung der *Raumvorstellung* für andere Leistungsbereiche sind es, die das große Interesse z. B. der Differenziellen Psychologie an der *Raumvorstellung* begründen. So beschreibt etwa McGee (1979) seine Motivation zur Arbeit an diesem Thema („Human Spatial Abilities. Sources of Sex Differences“) wie folgt:

> “My major theme is that spatial abilities can be measured reliably and validly, that performance on psychometrically valid and reliable tests of spatial abilities is predictive of success in numerous technical-vocational and perceptual-cognitive tasks, and that individual and group differences, particularly sex differences, whatever their source, have important implications for educators, personnel decision makers, and vocational and career counselors” (Preface, S. vii).

Bei einer weitergehenden begrifflichen Klärung, was *Raumvorstellung* sein soll bzw. welche kognitiven (Teil-)Fähigkeiten *Raumvorstellung* umfassen soll, lässt sich historisch ein Wechselspiel von Theorie und Empirie beobachten, das für die quantitativ-empirische Psychologie typisch ist. Dies lässt sich idealisiert und verkürzt wie folgt darstellen:

- Eine – ggf. aus vorhandenen Theorien abgeleitete – mehr oder weniger intuitive Vorstellung vom Gegenstand ist vorhanden und wird abstrakt[53] oder halbabstrakt[54] formuliert.
- Es werden konkrete Instrumente ausgewählt oder entwickelt, die zunächst „Augenscheinvalidität“ beanspruchen dürfen, also für eine Operationalisierung[55] des Gegenstands infrage kommen.
- In empirischen Untersuchungen werden die empirische Tauglichkeit der einzelnen Instrumente und die „Konstruktvalidität“ überprüft. Für eine Einschätzung der Konstruktvalidität sollten vorab inhaltlich abgeleitete Hypothesen darüber existieren, wie sich die einzelnen Instrumente zu anderen validen Instrumente verhalten[56].
- Die Ergebnisse der empirischen Untersuchungen können in der Regel sowohl zur Ausschärfung des Begriffs als auch zur Weiterentwicklung der Instrumente genutzt werden.

In Kap. 3.1.2 wird gezeigt, wie dieses Wechselspiel von Theorie und Empirie bei der Erforschung der inneren Struktur des Konstruktes *Raumvorstellung* betrieben wurde. Anschließend wird die prädiktive Validität von Raumvorstellungstests, also deren Vorhersagekraft für andere Leistungen, in Kap. 3.1.3 exemplarisch für einige Leistungsbereiche dargestellt.

3.1.1 Raumvorstellung als Bestandteil von Intelligenzmodellen

Intelligenz gehört, wie Klauer (2006b) schreibt, „zu den herausragenden Themen, seit es psychologische Forschung im modernen Sinne gibt“ (S. 275). Da *Raumvorstellung* einen festen Platz in praktisch allen Intelligenzmodellen hat, wurde sie von Anfang an mit zum Gegentand intensiver empirischer Forschung und theoretischer Modellbildung. Viele Kontroversen um das Konstrukt *Raumvorstellung* und dessen Erfassung werden vor dem Hintergrund ähnlicher Auseinandersetzungen um das gesamte Konstrukt *Intelligenz* verständlich.

---

[53] Abstrakt wäre im Fall der *Raumvorstellung* z. B. die Arbeitsdefinition, die „die mentale Repräsentation und Transformation figuraler Informationen“ ins Zentrum rückt.

[54] Halbabstrakt soll hier bedeuten, dass zumindest Teile einer solchen Definition schon in die Nähe einer Operationalisierung gebracht werden. So könnte für die „Transformation figuraler Informationen“ eine Facette als „mentale Bewegung von Objekten“ konkretisiert werden; ähnliche Konkretisierungen für andere Facetten müssten hinzukommen.

[55] Einen Teil einer Operationalisierung der „mentalen Bewegung von Objekten“ könnte der „Mental Rotation Test (MRT)“ (s. u.) darstellen, bei dem „Würfelketten“ nach Rotation wiedererkannt werden sollen; dabei muss wiederum berücksichtigt werden, dass der *MRT* nur eine spezifische Art der „mentalen Bewegung von Objekten darstellt“.

[56] Eine Leitfrage kann hier sein: „Welche Ergebnisse müssen eintreten, damit wir davon ausgehen können, dass die Instrumente den gewünschten Ausschnitt der Realität erfassen?“

Im Folgenden wird die historische Entwicklung der Intelligenzforschung vor allem aus der Sicht faktorenanalytischer Ansätze betrachtet; diese „a-posteriori-Ansätze“[57] bilden eine Hauptrichtung der psychologischen Intelligenzforschung, die z. B. in der Psychometrie und der Differenziellen Psychologie verfolgt wird und die für die vorliegende Arbeit von besonderer Bedeutung ist.

Nachdem sich schon Galton (1883) dem Thema *Intelligenz* zwar durchaus empirisch, aber ohne Tests im heutigen Sinne näherte, begann eine intensive quantitativ-empirische Erforschung von *Intelligenz* spätestens mit Beginn des 20. Jahrhunderts. Im Jahr 1904 formulierte Spearman schon mit dem Titel seiner Arbeit „’General Intelligence’, Objectively Determined and Measured“ Anforderungen an die Intelligenzforschung. Für die Auswertung der in dieser Arbeit verwendeten Tests entwickelt Spearman eine der ersten Varianten der Faktorenanalyse[58], mit der man genau einen Faktor empirisch daraufhin untersuchen kann, wie umfassend er das Testverhalten der Versuchspersonen erklären kann. Auf diesem Auswertungsverfahren basierte Spearmans Befund, dass Testleistungen in ganz unterschiedlichen kognitiven Bereichen zu großen Teilen durch einen einzigen (eindimensionalen) Faktor erklärt werden können. In der Folgezeit baute Spearman (1923) seine Theorie des „Generalfaktors der Intelligenz“, kurz „g“ genannt, aus. Dieser Generalfaktor lässt sich inhaltlich durch Leistungen „wie das Erkennen von Beziehungen und Zusammenhängen“ (Klauer, 2006b, S. 275) beschreiben.[59] Auf der Annahme, dass es einen solchen Generalfaktor gibt, basiert die zusammengefasste Angabe von Intelligenzleistungen durch eine einzige Zahl, den „Intelligenzquotienten (IQ)“.

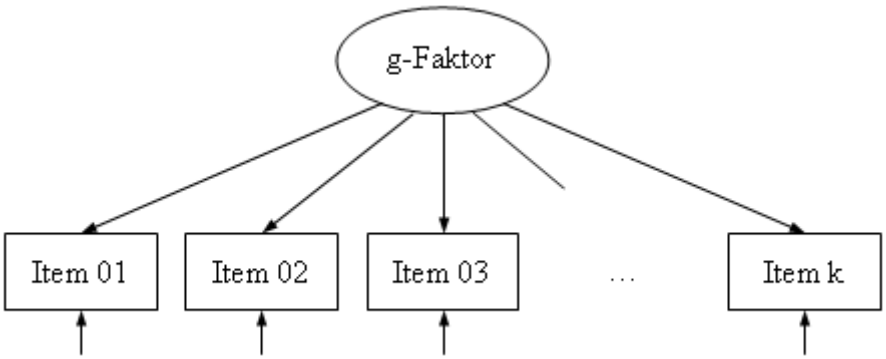
Da schon Spearman viele verschiedene Test mit höchst unterschiedlichen Anforderungsarten (z. B. *sprachliche Analogien*, *numerisches Verständnis*, *Raumvorstellung* oder *logisches Schließen*) verwendete, lag es nahe, dass der Generalfaktor die Testleistungen unterschiedlicher Versuchspersonen nicht vollständig erklären kann. Die verbleibenden (noch nicht erklärten) Reste der Testleistungen führte Spearman auf aufgabenspezifische Faktoren zurück, die er allerdings nicht weiter zu Teilbereichen kognitiver Fähigkeiten zusammenfasste.

[57] Anders als „a-priori-Ansätzen“, die versuchen, das Konstrukt *Intelligenz* vorwiegend theoretisch bzw. geisteswissenschaftlich-erfahrungsbasiert abzuleiten, und die in der Regel kaum empirisch prüfbar sind, basieren „a-posteriori-Ansätze“ vor allem auf der empirischen Auswertung von Tests, die zunächst intuitiv und nach Augenscheinvalidität mit *Intelligenz* in Verbindung gebracht werden. Im Anschluss an eine empirische Klärung der möglichen Struktur wird dann versucht, diese Struktur inhaltlich zu beschreiben.

[58] Eine Darstellung des aktuellen Stands der Methodenentwicklung im Bereich der Faktorenanalyse findet man z. B. bei Backhaus et al. (2008) oder bei Brachinger & Ost (1996).

[59] Eine detaillierte und immer noch aktuelle Darstellung des „g-Faktors“ findet man z. B. bei Jensen (1998).

Abbildung 3.1: Spearmans Intelligenzmodell mit Generalfaktor



Ab Mitte der 1920er Jahre konnten in unterschiedlichen Studien mit veränderter Methodik zwei Faktoren der *Intelligenz* empirisch voneinander getrennt werden: Ein Faktor umfasst vor allem *Raumvorstellung*, der andere vor allem verbale Fähigkeiten (vgl. McGee, 1979, S. 7).[60]

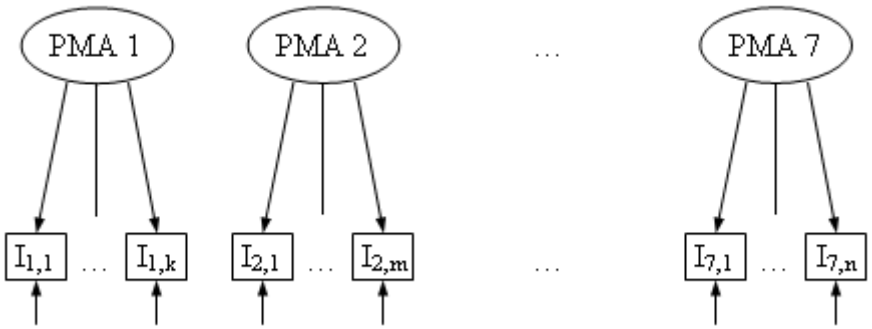
Eine weitere Ausdifferenzierung des Konstrukts *Intelligenz* in mehrere Bereiche kognitiver Leistungen leistet „das weithin akzeptierte Konzept der ‚primary mental abilities' von Thurstone" (D. H. Rost, 1977, S. 15). Thurstone (1934, 1936, 1938) profitierte dabei von den methodischen Weiterentwicklungen, die multiple Faktorenanalysen mit mehreren weitgehend voneinander unabhängigen Faktoren ermöglichten. Mit solchen Verfahren analysierte er eine Vielzahl unterschiedlicher Tests, konnte dabei mehrere Faktoren empirisch voneinander trennen und sieben dieser Faktoren psychologisch plausibel beschreiben: die sieben „primary mental abilities"[61]. Eine dieser primären kognitiven Fähigkeiten war die *Raumvorstellung*, die Thurstone später – wiederum faktorenanalytisch – weiter ausdifferenzierte (Thurstone, 1950). Ein Generalfaktor war in Thurstones Modell nicht vorgesehen, was u. a. an den verwendeten Methoden lag.[62] Dementsprechend konnte bei diesem Ansatz auch nicht ein einziger zusammengefasster IQ-Wert als Maß für die kognitive Fähigkeiten einer Versuchsperson ermittelt werden, sondern es mussten mehrere Fähigkeitswerte betrachtet werden.

[60] Diese Arbeiten konnten Spearman (vermutlich auch aufgrund ihrer methodischen Anlage) noch nicht überzeugen: „Noch 1927 äußerte Spearman Zweifel an der Existenz eines Gruppenfaktors ‚spatial relations' neben bzw. unter ‚g'" (Rost, 1977, S. 59).

[61] Spearman nannte Faktoren, wie die von Thurstone postulierten, später „Gruppenfaktoren", weil ihnen jeweils Erklärungskraft für eine Gruppe von Testaufgaben zugeschrieben wurde.

[62] Thurstones Methodik ließ es noch nicht zu, gleichzeitig mehrere weitgehend voneinander unabhängige Faktoren (Gruppenfaktoren) empirisch zu trennen und einen darüberstehenden Generalfaktor zu isolieren, der das Gemeinsame aller Gruppenfaktoren bündelt.

Abbildung 3.2: Thurstones Intelligenzmodell mit mehreren Gruppenfaktoren



In der Tradition der Arbeiten von Spearman und Thurstone entwickelte sich in der Folgezeit neben weiteren Forschungsparadigmen eine „Englische Schule“ (Spearman) und eine „Amerikanische Schule“ (Thurstone) der Intelligenzforschung, die – bei weiterentwickelten Verfahren – das Primat des Generalfaktors bzw. das Primat der Gruppenfaktoren sowohl theoretisch als auch empirisch betonten. Die *Englische Schule* versuchte dementsprechend, möglichst viel Varianz in den Testleistungen von Versuchspersonen durch einen Generalfaktor zu erklären, während die *Amerikanische Schule* möglichst viel Varianz durch die Gruppenfaktoren erklären wollte. Ganz anschaulich kann man diese Herangehensweisen als „top-down-Ansatz“ bzw. als „bottom-up-Ansatz“ bezeichnen (vgl. Brunner, 2006, S. 37 ff.). Trotz dieser unterschiedlichen Herangehensweisen, eröffneten schon Spearman und Thurstone selbst eine Perspektive zur Integration von Befunden und Konzepten der beiden Schulen:

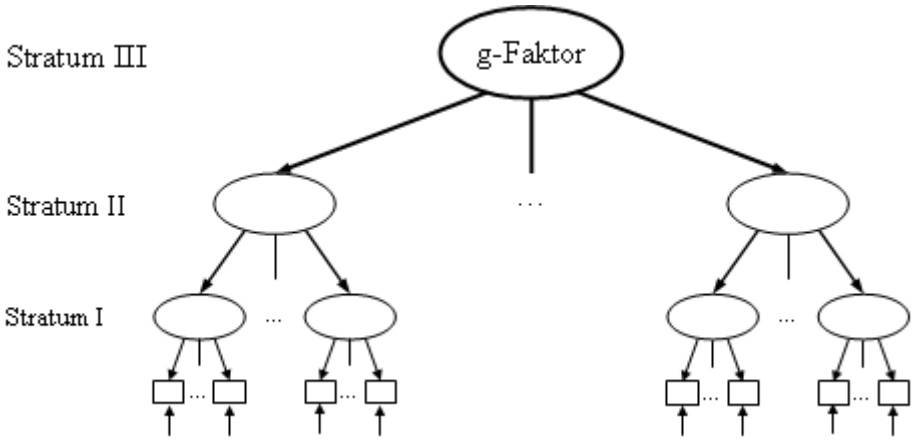
> „Die Kontroverse zwischen Spearman und Thurstone zog sich über viele Jahre hin und endete, was die Gruppenfaktoren und g betrifft, fast mit einem Kompromiß. Spearman … gestand schließlich auch der Annahme von Gruppenfaktoren eine gewisse Berechtigung zu, er nannte und erklärte sie nur anders. Thurstone akzeptierte neben seinen Primärfähigkeiten auch einen allgemeinen Intelligenzfaktor, allerdings in einer besonderen Form …“ (Jäger, 1967, S. 77)

In der zweiten Hälfte des 20. Jahrhunderts gab es grundlegende Ansätze zur Integration von Generalfaktor und Gruppenfaktoren in einem Modell. Vernon (1961) ging als Vertreter der *Englischen Schule* von einem Generalfaktor aus, etablierte unter ihm aber zunächst zwei übergeordnete Gruppenfaktoren („major group factors“), dann mehrere untergeordnete Gruppenfaktoren („minor group factors“) und schließlich auf einer vierten Ebene spezifische Faktoren. Einer der beiden übergeordneten Gruppenfaktoren umfasst dabei als wesentlichen Bestandteil *Raumvorstellung* („practical-mechanical-spatial-physical“).

Carroll (1993, 2005), ein Vertreter der *Amerikanischen Schule*, fand später auch innerhalb eines *bottom-up-Ansatzes* hinreichende empirische Evidenz für einen Generalfaktor. In seiner „3-Stratum-Theorie“, die inhaltlich weit ausdifferenziert und bis heute aktuell ist, berücksichtigt er insbesondere die Arbeiten von Cattel (1963, 1971), der die Unterschei-

dung von „fluiden“ und „kristallinen“ kognitiven Fähigkeiten vorschlug, einen Generalfaktor in seinem Intelligenzmodell aber noch ablehnte. In Carrolls Modell sind acht Gruppenfaktoren, die bei einem etwas anderen inhaltlichen Zuschnitt eine ähnliche Breite wie Thurstones „primary mental abilities“ haben, auf dem mittleren Stratum II angeordnet;[63] einer dieser acht Gruppenfaktoren umfasst wiederum wesentliche Aspekte der *Raumvorstellung* („broad visual perception“).

Abbildung 3.3: Carrolls Intelligenzmodell mit Generalfaktor und Gruppenfaktoren

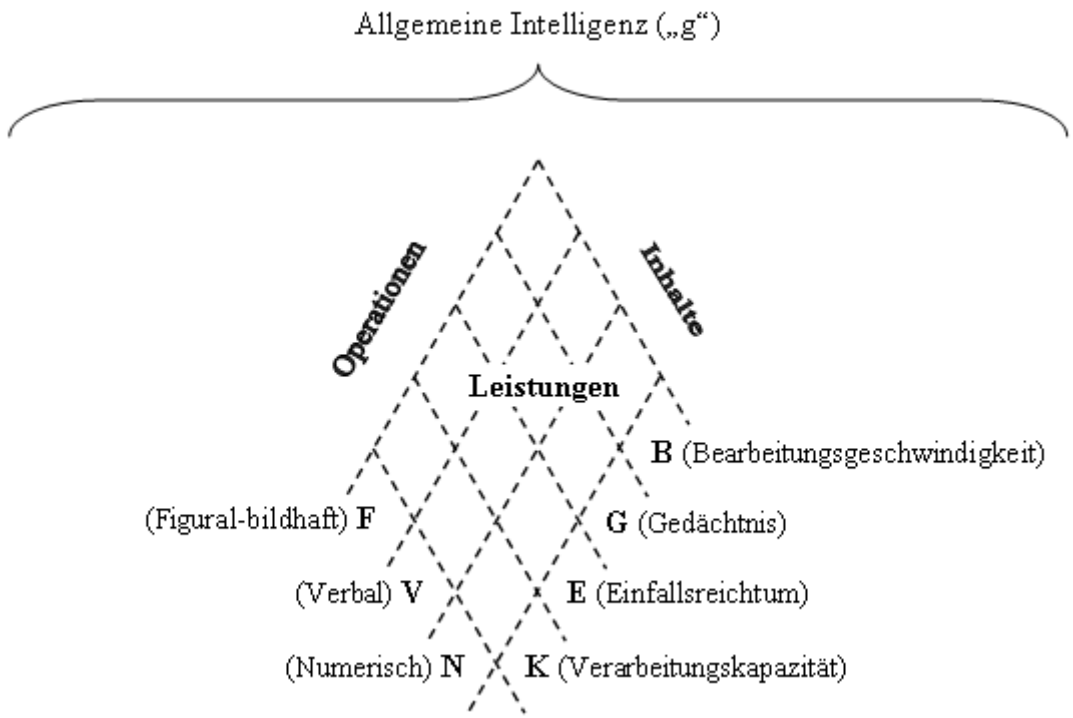


Einen anderen Ansatz zur Integration verschiedener Intelligenzmodelle verfolgte Jäger (1967, 1984), der neben den stärker empirisch-faktorenanalytisch geprägten Modellen aus der *Amerikanischen Schule* und der *Englischen Schule* auch die eher theoretisch abgeleiteten Modelle von Meili (1944, 1964) und Guilford (1956, 1959, 1967) zur Grundlage seiner Arbeit macht. Während Meilis Modell gestaltpsychologisch geprägt ist, orientierte sich Guilford an Prozessen der Informationsverarbeitung. Jäger (1967) machte die Hauptaussagen der insgesamt vier betrachteten Modelle zum Gegenstand einer breit angelegten empirischen Überprüfung und integrierte in der Folgezeit theoretisch und empirisch plausible Elemente im „Berliner Intelligenzstrukturmodell (BIS)“ (Jäger et al., 1996). Dieses Modell erklärt kognitive Leistungen (a) durch einen Generalfaktor und (b) durch eine Kombination von „Operationen“ und „Inhalten“. Formal betrachtet sind dabei neben der Annahme eines Generalfaktors strukturelle Aspekte aus dem Modell von Guilford und Kategorien aus dem

[63] In Vernons Modell gibt es hingegen die etwas breiter gefassten „übergeordneten“ Gruppenfaktoren und die etwas enger gefassten „untergeordneten“ Gruppenfaktoren.

Modell von Meili prägend.[64] *Raumvorstellung* tritt im *BIS* auf der Seite der Inhalte explizit in Erscheinung („figural-bildhaft“).

Abbildung 3.4: Das „Berliner Intelligenzstrukturmodell (BIS)“ nach Jäger et al. (1996)



Von den aktuellen Intelligenzmodellen sehen fast alle einen Generalfaktor und alle (ggf. darunter befindliche) Gruppenfaktoren in hierarchischen Modellen auf mehreren Ebenen vor. Je nach Modell und Verwendungszweck wird als Maß für die Intelligenz eine einzige Zahl (IQ-Wert) oder ein Zahlentupel (für jeden kognitiven Fähigkeitsbereich eine Zahl) angegeben. In praktisch allen Intelligenzmodellen spielt *Raumvorstellung* eine wesentliche Rolle, sei es als Gruppenfaktor oder als Kombination aus Inhalt und Operation. Dabei gilt für alle genannten Strukturmodelle eine gewisse Altersabhängigkeit: „Nach der Differenzierungshypothese der Intelligenz (…) tritt mit zunehmendem Alter (ungefähr ab dem 6. Lebensjahr) bei intellektuellen Leistungen die Bedeutung des Generalfaktors der Intelligenz zugunsten spezifischer Faktoren zurück“ (D. H. Rost, 1977, S. 120). Analog zur Ausdifferenzierung der kognitiven Leistungen bei Heranwachsenden gibt es die Hypothese der Dedifferenzierung im höheren Lebensalter.

[64] Gleichwohl haben sich weder Meilis noch Guilfords Modelle als Ganze empirisch bewährt. Umgekehrt waren Thurstones „primary mental abilities“ empirisch plausibler, lassen sich aber nur reorganisiert inhaltlich wiederfinden.

Neben den hier dargestellten faktorenanalytischen Ansätzen gibt es noch eine Vielzahl weiterer Ansätze der psychologischen Intelligenzforschung, unter den vor allem die kognitivistischen Ansätze noch von größerer Bedeutung sind.[65] Der wissenschaftliche Diskurs über die Struktur der *Raumvorstellung* und Fragen der Messung von *Raumvorstellung* werden vor dem Hintergrund der faktorenanalytischen Intelligenzforschung verständlich. Interessant bei der Entwicklung der Intelligenzmodelle ist u. a., dass die inhaltliche Entwicklung der Modelle, insbesondere deren Ausdifferenzierung, in der Regel erst infolge der methodischen Entwicklung der Verfahren (hier vor allem: Faktorenanalyse) stattfand.

Bei der Darstellung der historischen Entwicklung der Intelligenzforschung in diesem Abschnitt bleibt offen, was *Intelligenz* eigentlich ist. Dies liegt nicht nur daran, dass das Erkenntnisinteresse hier vor allem auf die abstrakte Struktur und das mögliche Auftreten von *Raumvorstellung* innerhalb der Modelle gerichtet war, sondern ist der Tatsache geschuldet, dass es „nach wie vor … keine allgemein akzeptierte Definition von Intelligenz [gibt], wenngleich die Forscher weitgehend darin übereinstimmen, mit welcher Art von Aufgaben Intelligenz zu erfassen und zu messen sei“ (Klauer, 2006b, S. 275; Erg. d. d. Verf.). In Ermangelung einer intensionalen Begriffsklärung bleibt der Intelligenzforschung als kleinster gemeinsamer Nenner also immer noch der fast 90 Jahre alte (selbstironisch formulierte) extensionale Ansatz: „Intelligence is was the tests test“ (Boring, 1923, S. 35).

3.1.2 Modelle der inneren Struktur von Raumvorstellung

Wenn man im Rahmen von Intelligenzmodellen kognitive Fähigkeiten insgesamt betrachtet, kann es durchaus genügen, eine nicht weiter ausdifferenzierte Komponente *Raumvorstellung* zu berücksichtigen. Ist das Erkenntnisinteresse aber – wie in der vorliegenden Arbeit – spezifisch auf *Raumvorstellung*, ihren Zusammenhang mit anderen kognitiven Leistungen und hier insbesondere auf differenzielle Befunde gerichtet, so stellt sich die Frage nach der inneren Struktur von *Raumvorstellung*: Möglicherweise können innerhalb der *Raumvorstellung* Teilbereiche identifiziert werden, für die besonders enge Zusammenhänge mit anderen kognitiven Leistungen, wie etwas *Mathematikleistung*, existieren oder in denen besonders große interindividuelle Unterschiede festgestellt werden können.

Die Ausdifferenzierung von *Raumvorstellung* („innere Struktur“) setzt dabei zunächst voraus, dass sich *Raumvorstellung* als Ganzes von anderen kognitiven Fähigkeiten trennen lässt („äußere Struktur“). In Kap. 3.1.1 wurde dargestellt, dass *Raumvorstellung* sich etwa ab Mitte der 1920er Jahre empirisch stabil von anderen, vornehmlich verbalen kognitiven Fähigkeiten trennen ließ. Auf diesen Befunden basierend konnte fortan innerhalb der fakto-

[65] Darüber hinaus wird im pädagogisch-didaktischen Bereich auch intensiv Bezug auf Konzepte wie Gardners „Multiple Intelligenzen“ (Gardner, 1983, 2006) Bezug genommen. Aus einer stärker empirisch orientierten Sicht hat Rost (2008) diesen Ansatz sehr überzeugend grundlegend kritisiert und dabei dargelegt, warum Gardners Ansatz auch für die Bezugsdisziplinen eigentlich nicht tragfähig ist.

renanalytischen Forschungstradition eine Ausdifferenzierung des Konstrukts *Raumvorstellung* stattfinden. Dabei konnte bald gezeigt werden, dass – in Analogie zur „Differenzierungshypothese der Intelligenz" – eine allgemeine kognitive Fähigkeit im Vorschulalter noch die inneren Strukturen (von Intelligenz und damit auch) von *Raumvorstellung* überlagert:

> „Bisher vorliegende Untersuchungen zum Konzept Raumvorstellung zeigen auf, daß vor dem 7. Lebensjahr nicht mit der Existenz eines eigenständigen Faktors ‚Raumvorstellung' gerechnet werden kann … […] Wie Thurstone (1955) aufweist, hat sich die Dimension Raumvorstellung schon im Alter von 12 Jahren recht weit (ca. 80 % der Erwachsenenleistung sind dann erreicht) entwickelt" (D. H. Rost, 1977, S. 120).

Der aktuelle Stand der Forschung lässt zwar kaum Zweifel daran zu, dass es eine innere Struktur der *Raumvorstellung* mit mehreren empirisch trennbaren Faktoren bzw. Komponenten gibt, ein einheitliches Modell ist gleichwohl nicht in Sicht: „Insgesamt zeigt sich kein Konsens über Anzahl und Art der Faktoren" (Hosenfeld et al., 1997, S. 85). Passend zum Erkenntnisinteresse der vorliegenden Arbeit werden im Folgenden ausschließlich „Small-Scale Fähigkeiten" (Wiedenbauer, 2006, S. 11) betrachtet. Dies sind die klassischen räumlichen Fähigkeiten, zu deren Erfassung Aufgaben mit vollständig überschaubaren räumlichen Informationen verwendet werden.

Innerhalb der psychologischen Forschungsansätze, die für die vorliegende Arbeit relevant sind, konkurrieren seit Mitte des 20. Jahrhunderts vor allem 2-Komponenten-Modelle (vgl. D. H. Rost, 1977; McGee, 1979)[66] mit 3-Komponenten-Modellen (z. B. Thurstone, 1938, 1950; Linn & Petersen, 1985). Da es außerdem sowohl innerhalb der 2-Komponenten-Modelle als auch innerhalb der 3-Komponenten-Modelle Unterschiede in der inhaltlichen Ausgestaltung der Komponenten gibt, existiert nach wie vor eine Vielzahl unterschiedlicher Modelle, die viel Gemeinsames, aber auch spezifische Unterschiede haben.

Betrachtet man die Erforschung der inneren Struktur von *Raumvorstellung* historisch, so fallen Parallelen zur Intelligenzforschung ins Auge: Nachdem es kaum noch Zweifel daran gab, dass ein Faktor *Raumvorstellung* empirisch separiert werden konnte, postulierten Vertreter der *Englischen Schule* einen starken allgemeinen Faktor *Raumvorstellung*, während Vertreter der *Amerikanischen Schule* – allen voran Thurstone – Modelle mit mehreren (Unter-)Faktoren entwickelten. Bei diesen unterschiedlichen Modellen spielte wieder die methodische Ausrichtung der Forschungstraditionen die entscheidende Rolle.

Innerhalb der *Englischen Schule* konnte El Koussy (1935) in einer grundlegenden Arbeit, ausgehend vom typischen *top-down-Ansatz*, einen breiten Faktor *Raumvorstellung* empi-

[66] Neben den klassisch strukturierten 2-Komponenten-Modellen gibt es auch noch die „Kontinuumhypothese" von Zimmermann (1954), der zwei Komponenten der *Raumvorstellung* („Space", „Visualization") in einem bipolaren Kontinuum zwischen „Wahrnehmungsgeschwindigkeit" und „Schlussfolgern" verortet.

risch identifizieren und inhaltlich plausibel deuten („ability to obtain and the facility for utilizing spatial imagery“). Auf El Koussys Arbeit bezogen sich fortan alle weiteren Forschungsarbeiten der *Englischen Schule* bis hin zum Intelligenzmodell von Vernon (s. o.).

Thurstones 3-Faktoren-Modell

Im Gegensatz zu El Koussy verfolgte Thurstone (1938, 1950) den *bottom-up-Ansatz* der auch bei der *Raumvorstellung* zu einer Ausdifferenzierung führte. Thurstones Arbeiten zur *Raumvorstellung* sind dabei ähnlich bedeutend wie seine Arbeiten zur *Intelligenz* insgesamt. Sein 3-Faktoren-Modell ist ein zentraler Bezugspunkt für nahezu alle späteren Forschungsarbeiten in diesem Bereich, die im Ergebnis allerdings überwiegend zu inhaltlich abweichenden Modellen führten. Thurstone (1950) schlug zunächst drei Faktoren vor, die er als Unterfaktoren von „S“ („Space“) mit „S1“ bis „S3“ bezeichnete.[67] Die folgende Beschreibung von „S1“ bis „S3“ orientiert sich an Barratts Integration von Thurstones Arbeit mit den Ergebnissen von Psychologen der US-Luftwaffe (Guilford, 1947), die insbesondere während des 2. Weltkriegs intensive Studien in diesem Bereich durchgeführt und ausgewertet haben (vgl. Barratt, 1953):

Tabelle 3.1: Thurstones 3-Faktoren-Modell der Raumvorstellung

S1 („Spatial Relations“) Die Fähigkeit, ein Objekt[68] zu identifizieren, wenn es in veränderter räumlicher Lage gezeigt wird.	Referenztests: „Figures“, „Flags“ und „Cards“
S2 („Visualization“) Die Fähigkeit, sich Bewegungen oder Veränderungen einzelner Teile eines Objekts vorzustellen.	Referenztest: „Differential Aptitude Test – Spatial Relations Subtest (DAT:SR)”
S3 („Spatial Orientation“) Die Fähigkeit, sich mental in räumlichen Konstellationen zu bewegen, wobei der Betrachter selbst Teil dieser Konstellationen.	Referenztest: „Chair-Window“

[67] Thurstone deutete dabei einen weiteren Faktor an, den er „kinesthetic factor“ nannte und der durch seine Tests „Hands“ und „Bolts“ operationalisiert wurde (vgl. Rost, 1977, S. 72).

[68] „Ein Objekt“ kann dabei auch für eine starre Anordnung mehrerer Objekte stehen.

Eine Hauptkritik an Thurstones 3-Faktoren-Modell stellt die Unterscheidung von *S1* und *S3* infrage. Auf forschungsmethodische Defizite bei der Trennung von *S1* und *S3* wies Pawlik (1968) hin. Er konnte zeigen, dass es noch keine Untersuchung gab, in der *S1* und *S3* simultan identifiziert wurden. Inhaltlich besteht der zentrale Unterschied in Thurstones Beschreibung der Faktoren *S1* und *S3* in der (vorgestellten) eigenen räumlichen Position der Versuchspersonen. Bei *S3* sollen sie gedanklich ein Teil der räumlichen Konfiguration sein und sich gedanklich in ihr bewegen, während sie bei *S1* ein Objekt betrachten und dessen räumliche Lage verändern. Sowohl theoretisch als auch empirisch konnte überzeugend nachgewiesen werden, dass diese kognitiven Prozesse nicht durch die Aufgabenstellungen determiniert sind, sondern von den Präferenzen der Versuchspersonen abhängen.

Bei *S1*-Aufgaben kann man sich gedanklich in eine räumlichen Konfiguration mit dem betrachten Objekt begeben und sich gedanklich um das Objekt herum bewegen, statt nur das Objekt zu betrachten und gedanklich zu drehen. Analog kann man bei *S3*-Aufgaben auch die räumliche Konfiguration betrachten und gedanklich in eine andere räumliche Lage bringen, statt seine eigene Position zu verändern. Diese zunächst theoretisch formulierte Kritik findet sich z. B. bei Michael et al. (1957). Qualitativ-empirisch fundieren kann man sie anhand der kognitionspsychologischen Studie von Barratt (1953), der Studierende aufforderte ihre Lösungsprozesse bei Raumvorstellungsaufgaben zu verbalisieren. Anhand entsprechender Transkripte zum „Guilford-Zimmerman Spatial Orientation Test“ lassen sich tatsächlich die theoretisch erwarteten unterschiedlichen Strategien bei gleichen Aufgabenstellungen identifizieren (vgl. D. H. Rost, 1977, S. 68 f.).[69]

Für die Erforschung der *Raumvorstellung* ist ein anderer Befund von Barratts Studie mindestens genauso bedeutend. Er konnte nachweisen, dass sich Versuchspersonen relativ stabil jeweils einer von zwei Strategien bedienen, die er mit „Part Approach“ und „Whole Approach“ bezeichnet. Während beim „Whole Approach“ Objekte als Ganzes in ihrer räumlichen Lage verändert werden, werden beim „Part Approach“ Teile der Objekte in ihrer räumlichen Beziehung zueinander analysiert. Bei vielen Raumvorstellungsaufgaben können beide Strategien zum Erfolg führen. Versuchspersonen, die sich relativ stabil der einen oder anderen Strategie bedienen werden in der deutschsprachigen Forschung als „Analytiker“ bzw. „Holistiker“ bezeichnet (vgl. Putz-Osterloh, 1977; Hosenfeld et al., 1997). Die Frage der Bearbeitungsstrategien wird in Kap. 3.3.3 genauer betrachtet.

2-Faktoren-Modelle und die Kontinuumshypothese

Aufgrund der Kritik an der fehlenden empirischen Trennung von *S1* und *S3* sowie Befunden zu Bearbeitungsstrategien, wie denen von Barratt, entwickelten sich in der Folgezeit zunächst 2-Faktoren-Modelle. So fassten Michael et al. (1957) die Faktoren *S1* und *S3* zu-

[69] Barrat schlussfolgerte hieraus allerdings noch nicht, dass *S1* und *S3* zusammengeführt werden sollten.

sammen und behielten *S2* bei. Auch in der Bezeichnung lehnten sie sich eng an Thurstone an. Der aus *S1* und *S3* gebildete Faktor wurde mit „SR-O (Spatial Relations and Orientation)“ bezeichnet, der aus S2 gewonnene Faktor mit „Vz (Visualization)“. Zusätzlich deuteten Michael et al. einen weiteren Faktor „K (Kinesthetic Imagery)“ an, was Thurstones Ergänzung seines 3-Faktoren-Modells um einen „Kinesthetic Factor“ entspricht (s. o.). Beim zugehörigen Referenztest „Hands“ geht es um die Fähigkeit der Recht-Links-Unterscheidung im Bezug auf den eigenen Körper.[70] Für diesen möglichen Faktor wurde aber nicht hinreichend empirische Evidenz gefunden.[71]

Für die Faktoren „SR-O“ und „Vz“ gab es hingegen hinreichend empirische Evidenz und die zugehörigen Referenztests unterscheiden sich relativ klar bezüglich ihrer Anforderungen – auch wenn es kognitive Prozesse gibt, die für „SR-O“ charakteristisch sind und die bei „Vz“ eine (untergeordnete) Rolle spielen. Daher kommt D. H. Rost (1977) mit Blick auf die oben genannten Arbeiten von Pawlik, Michael et al. und Barratt zu dem Ergebnis:

> „So können wir auch Thurstones (1950) Vorschlag, S3 als separaten, von S1 losgelösten Faktor anzusprechen, nicht zustimmen und halten die Zusammenfassung in SR-O für gerechtfertigt. Auf der anderen Seite sind die Faktoren SR-O und Vz in ihrer psychologischen Bedeutung deutlich voneinander unterscheidbar, so daß sie als eigenständige, wenn auch nicht voneinander unabhängige Komponenten betrachtet werden können“ (S. 81).

Mit ähnlichen Begründungen wie Michael et al. (1957) und D. H. Rost (1977), aber unter Einbezug weiterer Arbeiten gelangt McGee (1979) in seiner Bestandaufnahme ebenfalls zu dem 2-Faktoren-Modell mit *SR-O* und *Vz*.

Bereits 1954 hat Zimmerman mit seiner „Kontinuumhypothese“ ein Modell mit zwei Komponenten von *Raumvorstellung* vorgelegt, in dem neben idealtypischen Anforderungen der Aufgaben auch individuelle Bearbeitungsstrategien berücksichtigt werden können. Zimmerman, der später auch an der Arbeit Michael et al. (1957) beteiligt war, ging dabei u. a. von Thurstones Faktoren *S1* und *S2* aus. Aufgrund seiner Untersuchung von Lösungsprozessen berücksichtigte er wie Barratt (1953) mögliche Strategieunterschiede bei der Bearbeitung von gleichen Raumvorstellungsaufgaben, die jeweils unterschiedlich intensiv auf verschiedene kognitive Grundfunktionen („Perceptual Speed“, „Reasoning“) zurückgreifen. Zimmerman ordnete zwei Komponenten der *Raumvorstellung* („Space“, „Visualization“)[72] in einem Kontinuum zwischen den beiden Polen *Wahrnehmungsgeschwindigkeit*

---

[70] Bei diesem Test werden Abbildungen von Händen vorgelegt. Die Versuchpersonen sollen dann entscheiden, ob es aus ihrer Sicht (in Bezug zu ihrem Körper) eine rechte Hand oder eine linke Hand ist.

[71] Michael et al. (1957) qualifizieren diesen Faktor als „highly tentative“ (S. 191); dies kann allerdings auch an der Art und der Anzahl der berücksichtigten Aufgaben liegen. Für den vorsichtig vermuteten Faktor „K“ gab Thurstone nur zwei Referenztests an („Hands“ und „Bolts“), während z. B. für „Visualization“ eine Vielzahl von Tests zur Verfügung stand.

[72] Während *Visualization* weitgehend *S2* entspricht, umfasst *Space* neben *S1* auch noch Anteile von *K* (s. o.).

(„Perceptual Speed“) und *Schlussfolgern* („Reasoning“) an. Die Anordnung von Raumvorstellungsaufgaben und Komponenten der *Raumvorstellung* hängt dabei von der konkreten Versuchsperson ab:

> „If the continuum hypothesis is valid we must give very serious consideration to the fact that the difficulty of any item will differ among subjects, and that the points on the continuum where each one of these factors enters in will therefore vary among subjects“ (Zimmerman, 1954, S. 399).

Das 3-Komponenten-Modell von Linn & Petersen

Die oben dargestellten Befunde und Modelle deuten darauf hin, dass eine Klärung der inneren Struktur von *Raumvorstellung* sowohl kognitionspsychologische als auch psychometrische Ergebnisse und Überlegungen berücksichtigen sollte. In ihrer viel beachteten Meta-Analyse zu Geschlechterunterschieden bei der *Raumvorstellung* strukturieren Linn & Petersen (1985) den fraglichen Bereich kognitiver Fähigkeiten, indem sie von kognitionspsychologischen Betrachtungen ausgehen („Welche mentalen Prozesse sind an der Lösung einer Aufgabe beteiligt?“) und die vermutete Struktur psychometrisch plausibilisieren („Sind die gefundenen Maße in den identifizierten Komponenten hinreichend homogen?“). Dabei betonen sie, dass eine allgemeine Antwort auf die Frage nach der Struktur von *Raumvorstellung*, wenn überhaupt, dann ausgehend von den Lösungsprozessen gefunden werden kann (S. 1482). Mit ihrem Vorgehen gelangen Sie zu drei Komponenten der *Raumvorstellung*, die sie kognitionspsychologisch beschreiben und durch die Angabe von Referenztests operationalisieren (S. 1482 ff.):

Tabelle 3.2: Linn & Petersens 3-Komponenten-Modell der Raumvorstellung

„Spatial Perception“ (Räumliche Wahrnehmung) Die Fähigkeit, räumliche Beziehungen unter Bezugnahme auf den eigenen Körper und trotz ablenkender Informationen zu bestimmen.	Referenztest: „Water Level Tasks (WLT)“; „Rod and Frame Test (RFT)“
„Mental Rotation“ (Mentale Rotation) Die Fähigkeit, vorgegebene (zwei- oder dreidimensionale) Objekte (schnell und präzise) mental zu rotieren.	Referenztest: „Mental Rotation Test (MRT)“; „Flags” und „Cards“ (Thurstone)
„Spatial Visualization“ (Räumliche Visualisierung) Die Fähigkeit, komplexere und mehrschrittige Bearbeitung räumlicher Informationen mental durchzuführen. Dabei können Teilprozesse „Spatial Perception“ oder „Mental Rotation“ benötigen.	Referenztest: „Differential Aptitude Test – Spatial Relations Subtest (DAT:SR)”; “Embedded Figures Test (EFT)”

Zur kognitionspsychologischen Begründung der Komponenten stellen Linn & Petersen jeweils die charakteristischen mentalen Leistungen dar, die Versuchspersonen bei der Bewältigung der genannten Referenztests erbringen:

Für *Spatial Perception* ist – ihrer Analyse zufolge – typisch, dass die Versuchspersonen das Referenzsystem der gravitativen Vertikalen (und der orthogonal dazu liegenden Horizontalen) für die richtige Orientierung nutzen. Da die Messung dieser Fähigkeit in der Regel mit *Paper and Pencil Tests* erfolgt, müssen Versuchspersonen ggf. mental eine räumliche Situation herstellen, in der sie in geeigneter Beziehung zur Abbildung stehen. Einschränkend sei hier angemerkt: Die Zusammenfassung der „Vertikalen-Fähigkeiten" und „Horizontalen-Fähigkeiten" bei Linn & Petersen stimmen zwar mit Piagets Entwicklungskonzept des räumlichen Denkens überein (Konzept des euklidischen Koordinatensystems; vgl. Piaget & Inhelder, 1971, „Dritter Teil"), es gibt aber auch empirische Studien, die eher eine Unterscheidung von beidem nahe legen (z. B. Liben, 1978; Quaiser-Pohl et al., 2004).

Die Komponente *Mental Rotation* ist schon in ihrer Bezeichnung durch den Prozess der mentalen Rotation gekennzeichnet. Dieser mentale Prozess verläuft dabei vermutlich in Analogie zu entsprechenden realen räumlichen Tranformationen:[73] „They [Shephard & Cooper (1982)] hypothesized that during a mental rotation the respondent's internal cognitive process have a one-to-one correspondence with the external rotations of the object (…). Thus they infer that a Gestalt-like process governs the rotations of objects" (Linn & Petersen, 1985, S. 1483; Erg. d.d. Verf.). Hier muss einschränkend hinzugefügt werden, dass schon in der oben genannten Studie von Barratt (1953) einige wenige Versuchspersonen enthalten waren, die die Aufgaben eines typischen Referenztests nicht holistisch über die mentale Rotation des vorgegebenen Objekts gelöst haben, sondern analytisch über die räumlichen Beziehungen einzelner Teile des Objekts zueinander.

Für die Komponente *Spatial Visualization* ist im Vergleich zu den beiden zuvor genannten die größere kognitive Komplexität charakteristisch, die sich in den Aufgaben entsprechender Referenztests beobachten lässt. Bei den Lösungsprozessen ist in der Regel ein mehrschrittiges, zumeist analytisches Vorgehen erforderlich, bei dem gegebenenfalls auch auf *Spatial Perception* oder *Mental Rotation* zurückgegriffen wird. Bei oberflächlicher Betrachtung entsprechender Aufgaben kann daher die Unterscheidung zwischen den Komponenten zunächst schwierig sein; entscheidend für *Spatial Visualization* ist dann die größere Komplexität und die in der Aufgabenstellung angelegte Offenheit für mehrere Bearbeitungsstrategien.

[73] Bis heute gibt es in der Psychologie zu dieser Frage allerdings keinen einheitlichen Erkenntnisstand: „Kontrovers wurde und wird diskutiert, ob die mentale Repräsentation ein isomorphes Abbild des zu rotierenden Objektes ist und die mentale Rotation damit einer physikalischen Drehung entspricht" (Wiedenbauer, 2006, S. 26).

Bezieht man das Modell von Linn & Petersen (1985) auf die *Kontinuumhypothese* von Zimmerman (1954), so scheinen *Spatial Perception* und *Mental Rotation* im Kontinuum an der Stelle von *Space* aufzutreten (nahe am Pol *Wahrnehmungsgeschwindigkeit*), während *Spatial Visualization* nahezu deckungsgleich mit Zimmermans *Visualization* ist (nahe am Pol *Schlussfolgern*).

Die gewünschte psychometrische Absicherung der kognitionspsychologisch gewonnenen Komponenten erfolgt bei Linn & Petersen (S. 1485 ff.) über die Betrachtung von Effektstärken für die Geschlechterunterschiede in den Testleistungen. Für jede der drei oben dargestellten Komponenten vergleichen sie alle entsprechenden Effektstärken aus den Studien, die in ihre Meta-Analyse einfließen. Bei hinreichender Homogenität der Effektstärken gehen sie von einer psychometrischen Bestätigung der Komponenten aus. Diese hinreichende Homogenität erreichen sie für die Komponenten *Spatial Perception* und *Mental Rotation* allerdings erst, nachdem sie die einbezogenen Studien unterteilen. Bei *Spatial Perception* werden hierzu Studien nach dem Lebensalter der untersuchten Versuchspersonen in drei Gruppen zusammengefasst. Hingegen werden bei *Mental Rotation* die Studien bezüglich der verwendeten Referenztests unterteilt, wobei sich Homogenität ergibt, wenn Tests in zwei Dimensionen (*Flags*, *Cards*) von Tests in drei Dimensionen (*MRT*) unterschieden werden. Betrachtet man die Tests in zwei Dimensionen, so scheinen hierbei neben der mentalen Rotation auch analytische Strategien eher zum Ziel führen zu können als bei Tests in drei Dimensionen.

Mit den genannten Einschränkungen scheint das Modell von Linn & Petersen mit den drei dargestellten Komponenten also kognitionspsychologisch und psychometrisch plausibel zu sein, wobei größere Klarheit und schärfere Trennung erreicht wird, wenn nur idealtypische Referenztests, wie der *MRT* für *Mental Rotation* verwendet werden. Unter den seit Beginn des 20. Jahrhunderts verwendeten Tests, gerade auch unter den klassischen Tests von Thurstone, scheinen darüber hinaus aber auch solche zu sein, bei denen die Testleistung auf mindestens zwei der drei Komponenten von Linn & Petersen zurückgeführt werden können. Sind interindividuelle Unterschiede bei der genutzten Bearbeitungsstrategie möglich, kann sich auch die relevante Komponente interindividuell unterscheiden.[74] Daher ist es kaum verwunderlich, dass faktorenanalytische Studien bisher nicht zu einem einheitlichen Ergebnis gekommen sind, sondern dass die primär empirisch gewonnene innere Struktur von *Raumvorstellung* zum Teil von den jeweils verwendeten Tests und untersuchten Versuchspersonen abhängen. Dies hat sogar zu einer generellen Kritik an der faktorenanalytischen Forschung geführt:

[74] Beispiele hierfür sind die Tests *Flags* und *Cards*, deren Aufgaben holistisch mit *Mental Rotation* oder analytisch mit *Spatial Visualization* gelöst werden können.

> „Da diese Ergebnisse als Resultat langjähriger, faktorenanalytischer Forschung nicht befriedigen können, wird die Frage aufgeworfen, ob nicht die Faktorenanalyse als die zentrale Forschungsmethode selbst ungeeignet ist, um die anstehenden Probleme zu bewältigen. Für diese Sichtweise sprechen zumindest die bekannten Kritikpunkte, etwa daß faktorenanalytisch gewonnene Ergebnisse sowohl von den in die Analyse einbezogenen Testaufgaben bzw. Tests als auch von den Verteilungseigenschaften der untersuchten Variablen in der jeweiligen Personenstichprobe abhängig sind“ (Gittler, 1990, S. 11).[75]

In Kap. 3.4 wird – basierend auf den obigen Ergebnissen und auf den empirischen Ergebnissen aus Kap. 3.3 – pragmatisch ein eigenes Konstrukt von *Raumvorstellung* diskutiert und theoretisch vorbereitet, das zur Zielsetzung der vorliegenden Arbeit passt.

3.1.3 Vorhersagekraft für andere Leistungsbereiche

Seit Beginn der systematischen Erforschung von *Raumvorstellung* wird den dabei betrachteten kognitiven Fähigkeiten eine zentrale Bedeutung für ein breites Feld von (vor allem beruflichen) Tätigkeiten zugeschrieben, die weit über den technischen Bereich hinausgehen (vgl. D. H. Rost, 1977, S. 82 f.). Auch aktuell finden sich Raumvorstellungsaufgaben in Eignungstests für eine Vielzahl von Ausbildungsberufen bzw. Studiengängen. Wenn *Raumvorstellung* tatsächlich diese weit über die Schulzeit hinausweisende (praktische) Relevanz hat, dann ist dies sicherlich ein gewichtiger Grund, *Raumvorstellung* auch in der Schulzeit gezielt zu fördern. An dieser Stelle werden exemplarisch einige schon etwas ältere Ergebnisse von McGee (1979) und D. H. Rost (1977) dargestellt, die jeweils Befunde aus mehreren bereits damals vorliegenden Studien zusammenfassen und auf eine hohe Vorhersagekraft von Raumvorstellungstests hinweisen:[76]

> „In vielen … empirischen Studien wurde die prediktive Validität von Raumvorstellungstests für den Erfolg in technischen und mechanischen Berufen untersucht. Dabei fand man durchweg mittlere bis hohe Korrelationen ($r = 0{,}30$ bis $r = 0{,}70$) zwischen den Testergebnissen und unterschiedlich definierten Erfolgskriterien (Erfolg bei technischen Prüfungen, Beurteilung durch Arbeitskollegen, Lehrer, Vorarbeiter usw.)“ (D. H. Rost, 1977, S. 84).

Die prädiktive Validität ist dabei nicht auf Schul-, Studien oder Berufserfolg beschränkt. So konnte z. B. gezeigt werden, dass Versuchpersonen, die bereits zwei oder mehr Verkehrsunfälle verursacht haben, in der Tendenz über deutlich geringere dreidimensionale Raumvorstellungsleistungen verfügen als Personen, die noch keinen Verkehrsunfall verursacht haben (vgl. D. H. Rost, 1977, S. 84 f.).

---

[75] Diesen Befund erweiternd stellt Rost (1977) eine Untersuchung dar, in der von zwei unterschiedlichen Autorenteams jeweils ein Test für *Spatial Orientation* und ein Test für *Visualization* (jeweils sensu Thurstone) anhand derselben Stichprobe verglichen wurden: „[Es] ist offensichtlich, daß die Autorenschaft der verschiedenen Tests einen größeren Beitrag zur Interkorrelation der Tests beisteuert als die Gemeinsamkeit der Merkmale“ (Rost, 1977, S. 79; Erg. d. d. Verf.).

[76] Die spezifische Vorhersagekraft von *Raumvorstellung* für *Mathematikleistung* wird in Kap. 3.3.5 dargestellt

Der Zusammenhang zwischen vorab durchgeführten Raumvorstellungstests (z. B. im Rahmen von Aufnahmeprüfungen) und späterem Schul- bzw. Studienerfolg ist intensiv untersucht worden. Einschränkend muss hier berücksichtigt werden, dass die verwendeten Raumvorstellungstests zwar gängige psychometrische Kriterien für eine gute Messung erfüllen, dann aber mit Noten in Zusammenhang gebracht werden, denen ein deutlich komplexerer – aus psychometrischer Sicht „unsauberer" – Bewertungsvorgang zugrunde liegt. Umso beachtlicher ist, dass sowohl D. H. Rost (1997, S. 86 ff.) als auch McGee (1979, S. 24 ff.) jeweils von Studien mit substanziellen Zusammenhängen berichten.

Eine gut ausgeprägte *Raumvorstellung* scheint dabei zumindest bei einschlägigen Ausbildungen auf lange Sicht anderen, globaleren Prognosekriterien überlegen zu sein. McGee (1979, S. 25) zitiert eine Studie, die in einer technisch ausgerichteten Schule durchgeführt wurde (vgl. auch D. H. Rost, 1977, S. 84): Die Raumvorstellungsleistung des Aufnahmetests wurden mit einer umfassenderen Prognose des Schulleiters für anschließenden Schulerfolg verglichen. Für den Schulerfolg am Ende des ersten Jahrgangs waren die Testergebnisse und die Schulleiterprognose gleich gute Prädiktoren (jeweils $r = 0{,}24$). Bis zum Ende des zweiten Jahrgangs erhöhte sich die Vorhersagekraft der Raumvorstellungstests ($r = 0{,}47$), während die Vorhersagekraft der Schulleiterprognose etwas abnahm ($r = 0{,}17$).

Auch wenn, wie hier geschehen, die statistische Vorhersagekraft der *Raumvorstellung* für andere Leistungen bestätigt werden kann, muss natürlich berücksichtigt werden, dass damit noch keine Aussage über Kausalzusammenhänge getroffen wird. Diese Unterscheidung zwischen statistischer Vorhersagekraft und kausalen Wirkungen wird bei der Diskussion des Zusammenhangs von *Raumvorstellung* und *Mathematikleistung* exemplarisch geführt (vgl. Kap. 3.3.5).

3.2 Mathematikdidaktische Perspektiven auf Raumvorstellung

Mathematikdidaktische Perspektiven auf *Raumvorstellung* unterscheiden sich in der Regel deutlich von den zuvor betrachteten, vor allem kognitionspsychologisch und psychometrisch geprägten Perspektiven.[77] Auffallend ist dabei zum einen, dass die Mathematikdidaktik sich dem Thema *Raumvorstellung* – ganz nahe liegend – stärker vom Fach, vor allem von der Geometrie aus nähert:

[77] Eine größere Nähe zur Mathematikdidaktik haben die entwicklungspsychologischen Ansätze. Nicht zuletzt Piagets umfassende Arbeiten in diesem Bereich (vgl. Piaget & Inhelder, 1971) enthalten, trotz aller berechtigten Kritik, die an Piagets Entwicklungspsychologie und deren Rezeption geübt wird, wertvolle Anregungen für die Mathematikdidaktik.

> „Grundsätzlich, und das ist nicht neu, kann man das Analysieren und Messen von Raumvorstellungsvermögen auf (mindestens) zwei verschiedene Sichtweisen gründen. Entweder man betrachtet Raumvorstellungsvermögen und entsprechende Teilfähigkeiten aus der Sicht der Mathematikdidaktiker. Diese interpretieren und klassifizieren die geometrischen Aufgabenstellungen und die dabei beobachteten Antworten geometrisch-inhaltlich und sie erhalten dadurch Aussagen über unterschiedliche Fähigkeitsbereiche […]. Oder, vorzugsweise in der Psychologie, man analysiert zunächst mit Hilfe von Statistik-Methoden die Ergebnisse aus hinreichend vielen Tests, um ‚rechnerische Zusammenhänge' zwischen den Aufgaben festzustellen. Für die so gefundenen ‚zusammengehörigen' Aufgaben lassen sich dann vielleicht auch inhaltliche Gemeinsamkeiten entdecken, z. B. scheinbar gemeinsame Teilfähigkeiten, die zum Lösen dieser Aufgaben erforderlich sind" (Meißner, 2006, S. 32).

Darüber hinaus hat die Mathematikdidaktik vor allem das Lehren und Lernen von Mathematik, also zunächst stärker die Prozesse als deren (Lern-)Ergebnisse, im Blick. Das präzise Messen von *Raumvorstellung* hat dabei höchstens dienende Funktion, z. B. wenn die Wirkung von entsprechenden Unterrichtsreihen evaluiert werden soll. Dabei werden in der Mathematikdidaktik häufig Raumvorstellungstests verwendet, die näher an geometrischen Aufgabenstellungen des Mathematikunterrichts sind. Mathematikdidaktische Modelle der Struktur von *Raumvorstellung* müssen sich nicht primär faktorenanalytisch bewähren, sondern wichtige Kategorien für das Verstehen und Gestalten von Lehr-Lernprozessen bereithalten:

> „Aus didaktischer Perspektive … interessiert nicht allein das Ergebnis eines Lernprozesses, sondern insbesondere der Lernprozess selbst. […] Darüber hinaus macht es aus mathematikdidaktischer Perspektive wenig Sinn, räumlich-visuelle Fähigkeiten isoliert zu betrachten. Geht es auch um Schulung und Förderung dieser Fähigkeiten im Mathematikunterricht, dann sollte auch nach Berührungspunkten mit spezifisch mathematischen Kenntnissen und Fähigkeiten gesucht werden, deren Ausbildung sich positiv auf das Räumliche Vorstellungsvermögen auswirken können" (Pinkernell, 2003, S. 7).

Klassische Beiträge aus der deutschsprachigen Mathematikdidaktik enthalten vor allem bildungstheoretische und pädagogisch-fachliche Begründungen für eine explizite Berücksichtigung der *Raumvorstellung* im Mathematikunterricht sowie konkrete Entwürfe für die Unterrichtsgestaltung (vgl. z. B. Oehl, 1949; Besuden 1973, 1979). In jüngerer Zeit werden aber auch zunehmend Modelle und Befunde aus der Psychologie berücksichtigt und in mathematikdidaktische Arbeiten integriert (vgl. die umfassenden Arbeiten von Maier, 1994, 1996, 1999a, 1999b). Die rege mathematikdidaktische Forschungsaktivität in den vergangenen zehn Jahren deutet darauf hin, dass das Thema bis heute hoch aktuell und relevant ist. Allein im deutschsprachigen Raum finden sich aus dieser Zeit Arbeiten aus drei verschiedenen Arbeitsgruppen (Oldenburg: z. B. Grüßing, 2002, Hartmann & Reiss, 2000; Hellmich & Hartmann, 2002; Münster: z. B. Meißner, 2006; Pinkernell, 2003; Lüneburg: Lüthje, 2008, 2009). Dabei spielt auch die Messung der *Raumvorstellung* eine wichtige Rolle, wenngleich einige Arbeiten aus psychometrischer Sicht methodisch noch nicht ausgereift sind (s. u.).

Die Vielzahl mathematikdidaktischer Beiträge kann und soll hier nicht in ihrer Breite dargestellt werden, zumal für das Erkenntnisinteresse der vorliegenden Arbeit eher der psy-

chologisch-messende Zugang zur *Raumvorstellung* wichtig ist. Zur Kontrastierung dieses psychologischen Zugangs und für eine Diskussion der empirischen Ergebnisse der vorliegenden Arbeit im Hinblick auf den Mathematikunterricht werden an dieser Stelle aber exemplarisch Fragestellungen, Herangehensweisen und Modelle aus der Mathematikdidaktik referiert.

3.2.1 Typische Fragestellungen

Dem oben skizzierten Unterschied zwischen einem mathematikdidaktischen und einem psychometrischen Interesse an *Raumvorstellung* entsprechend, versuchen mathematikdidaktische Untersuchungen stärker Lehr-Lernprozesse, Prozesse der Aufgabenbearbeitung und „geometrienahe“ Leistungen in den Blick zu nehmen. Dies wird an fünf Studien, von denen vier aus den vergangenen zehn Jahren stammen, exemplarisch dargestellt:

Ausgangspunkt für die jüngste hier berücksichtigte Studie von Lüthje (2008, 2009) ist ein Erkenntnisdefizit bezüglich der *Raumvorstellung* von Kindern zum Zeitpunkt der Einschulung. Die Fähigkeiten dieser Population, die am Beginn institutionalisierten Lernens steht, sind für die Mathematikdidaktik besonders interessant, da schulische Lehr-Lernprozesse hierauf aufbauen:

> „[Die] Förderung der Raumvorstellung [ist] eines der zentralen Ziele des Mathematikunterrichts der Grundschule. Doch leider stehen bislang nur unzureichende Informationen über die räumlichen Fähigkeiten, insbesondere von Schulanfängerinnen und Schulanfängern, zur Verfügung. Dies ist auch darauf zurück zu führen, dass nach wie vor keine adäquaten psychometrischen Tests zur Erhebung räumlicher Fähigkeiten für diese Altersgruppen vorliegen“ (Lüthje, 2008, S. 581).

Die Studie von Koops & Sorger (1980; vgl. auch Koops et al., 1981) befasst sich vor allem mit „bildlichen Repräsentationen räumlicher Objekte“, also mit speziellen Anteilen der *Raumvorstellung*, die besonders nah am Geometrieunterricht sind, und berücksichtigt die aus Sicht der Mathematikdidaktik bedeutende Frage nach der Art der Darstellung räumlicher Informationen:

> „Die erstaunlich guten Leistungen der Erstkläßler in den Tests zum räumlichen Vorstellungsvermögen von FREUND/SORGER hinsichtlich der Fähigkeit zur Analyse perspektivischer Darstellungen räumlicher Objekt(e) (-konfigurationen) legten es nahe, diesen Bereich anhand von Aufgabenstellungen, die in stärkerem Maße konstruktive Aktivitäten beinhalten, detaillierter zu untersuchen. Die inhaltliche Strukturierung der Untersuchung orientiert sich dabei an der Theorie OLSONs (…) über die Abhängigkeit einer jeglichen (mathematischen) Aktivität von dem Medium, in dem sie vollzogen wird“ (Koops & Sorger, 1980, S. 1; Herv. i. O.).

Grüßing (2002) schließt mit ihrer Untersuchung an die quantitative Auswertung von Raumvorstellungstests für Schülerinnen und Schüler des 3. und 4. Schuljahres an. Nach dieser Testauswertung blieben nämlich Fragen nach Bearbeitungsstrategien und Erklärungen für unterschiedlich gute Testleistungen offen:

> „Die vorliegende Studie versucht, mit Hilfe von Einzelinterviews, angelehnt an die Methode des Lauten Denkens, das tatsächliche Vorgehen von Kindern im 4. Schuljahr bei räumlich-geometrischen Aufgabenstellungen zu verstehen und zu beschreiben. Tatsächlich zeigen die Kinder bei der Bearbeitung gleicher Aufgabentypen unterschiedliche Strategien“ (ebd., S. 37).

Die Arbeit von Meißner (2006) zeigt, wie Raumvorstellungstests in der Mathematikdidaktik genutzt werden können, um den intendierten Lernerfolg von speziell konzipierten Unterrichtsreihen komplementär zu qualitativen Erhebungen zu überprüfen:

> „Wenn wir also durch unsere Unterrichtsreihe Raumvorstellungen verbessern und anschließend auch einen Erfolg belegen wollen, so bieten sich für die Erfolgsmessung grundsätzlich zwei verschiedene Verfahren an, die sich im optimalen Fall gegenseitig ergänzen. Man kann sowohl möglichst viele Fallstudien und Einzelbeobachtungen zusammentragen und dann versuchen zu verallgemeinern. Oder man entwirft einen Test, der inhaltlich das abfragt, was in der Reihe geschult wurde, und der dann mit testtheoretischen Methoden ausgewertet wird. Wir haben für die Unterrichtsreihe ‚Wir bauen ein Dorf’ beide Wege verfolgt, die Beobachtung von Unterricht mit der Durchführung von Einzelinterviews auf der Basis von speziell dafür entwickelten Interviewaufgaben und die Durchführung von Tests. In diesem Aufsatz müssen wir uns auf das Messen von Raumvorstellungsvermögen durch Tests beschränken“ (ebd., S. 33).

Während der Einsatz von Tests häufig durch die konkret eingesetzten Aufgaben Kategorien für die Erfassung und Interpretation von *Raumvorstellung* mit sich bringt, müssen die zentralen Kategorien für die Interpretation von Lern- und Bearbeitungsprozessen häufig zunächst gewonnen werden. Die qualitative Ergänzung der oben genannten quantitativen Arbeit von Meißner bereitet Pinkernell (2003) mit der Entwicklung eines entsprechenden Modells vor:

> „Zur Auswertung von Unterrichtskonzepten, die im Mathematikunterricht das Räumliche Vorstellungsvermögen fördern sollen, werden in traditionellen Ansätzen in der Regel Tests eingesetzt, die aber über die Auswirkungen des tatsächlichen Unterrichtsgeschehens nur Momentaussagen liefern können. Im Unterricht selbst aber geschieht das eigentlich Spannende aus mathematikdidaktischer Sicht“ (ebd., S. 137).

> „Ziel der Arbeit hier ist die Formulierung und Begründung eines mathematikdidaktisch geprägten Modells des Räumlichen Vorstellungsvermögens. Ein solches Modell erschien notwendig, nachdem eine quantitative Auswertung der Unterrichtsreihe zwar positiv ausfiel, jedoch keine angemessen umfassende Bewertung liefern konnte. Insbesondere waren Aussagen über das Unterrichtsgeschehen selbst unmöglich, dessen Auswertung nach Durchsicht der ebenfalls vorliegenden Videoaufzeichnungen vielversprechend schien“ (ebd., S. 55).

Alle genannten Studien betonen ihren mathematikdidaktischen Ansatz und grenzen sich bewusst ab von „psychometrischen Tests“ (Lüthje, 2008, S. 581) bzw. vom „faktorenanalytische[n] Ansatz“ (Koops & Sorger, 1980, S. 1, Erg. d. d. Verf.). Dabei werden psychometrischen Herangehensweisen allerdings zum Teil Schwächen zugeschrieben, die so nicht existieren: „Aufgaben aus der psychometrischen Forschung setzen implizit eine Strategiehomogenität voraus und erweisen sich damit ohnehin als problematisch zur Erfassung räumlicher Fähigkeiten“ (Lüthje, 2008, S. 581). Wenn Raumvorstellungstests rein faktorenanalytisch ausgewertet werden würden, träfe diese Kritik so zu. Tatsächlich liegen aber auch Arbeiten wie die von Hosenfeld et al. (1997) oder Köller et al. (1994) vor, bei denen

mithilfe des „Mixed Rasch-Modells (MRM)“ unterschiedliche Bearbeitungsstrategien und unterschiedliche Personenfähigkeiten analysiert werden können (vgl. Kap. 3.3.3).

Dennoch haben psychometrische Tests natürlich Grenzen, insbesondere dort, wo es um die Erfassung und Interpretation individueller Denkprozesse geht, die über die Auswahl einer von mehreren idealtypisch identifizierten Strategien hinaus geht. Diese Grenzen können mit Herangehensweisen, die in der Mathematikdidaktik üblich sind, überwunden werden.

3.2.2 Übliche Herangehensweisen

Die Stärken mathematikdidaktischer Arbeiten werden vor allem dort sichtbar, wo es um die Interpretation von individuellen oder kooperativen Aufgabenbearbeitungen bzw. um die Verwendung von Aufgabenformaten geht, die nah am Geometrieunterricht sind. Mit Unterrichtsbeobachtungen, Einzelinterviews nach der Methode des lauten Denkens oder speziell entwickelten Raumvorstellungstests werden die oben formulierten Fragestellungen in den empirischen Studien umgesetzt. Dies wird anhand entsprechender Zitate für vier der oben genannten Arbeiten (Lüthje; Koops & Sorger; Grüßing; Meißner) dargestellt:

> „Daher verfolgt die geplante Untersuchung primär kein psychometrisches Ziel, denn eine rein quantitative Auswertung wird den Fähigkeiten der Schulanfängerinnen und Schulanfänger nicht gerecht. Nicht die Identifikation von Eigenschaften und Eigenschaftsdimensionen steht im Mittelpunkt, sondern die möglichst genaue Erfassung und differenzierte Beschreibung von Lösungsstrategien und damit subjektiver Sichtweisen“ (Lüthje, 2008, S. 581).

> „Der […] durchgeführte faktorenanalytische Ansatz erwies sich – vor allem für die betrachtete Altersstufe – als unbefriedigend; insbesondere liefert er keine Aussagen über Vorgehensweisen und Strategien, die von den Schülern beim Lösen entsprechender Aufgaben vollzogen werden […] . In der hier dargestellten Untersuchung wird daher eine Veränderung des Untersuchungsansatzes vorgenommen. Ihr Ziel besteht nicht in der Identifizierung statistisch abgesicherter Fähigkeiten, vielmehr werden einzelne (durch Klassen von Aufgaben mit inhaltlich homogenen Leistungsanforderungen charakterisierte) Fähigkeitsbereiche einer auch qualitative Aspekte des Lösungsprozesses berücksichtigenden Analyse unterzogen“ (Koops & Sorger, 1980).

> „Als Methode zur Untersuchung dieser Fragestellungen erschien die Methode des Lauten Denkens als besonders geeignet. Schülerinnen und Schüler wurden in Einzelinterviews aufgefordert, alle Gedanken während der Bearbeitung der Testaufgaben auszusprechen. Auf diese Weise ist es möglich, ein Bild des tatsächlichen Lösungsprozesses mit allen angewandten, weitergeführten und wieder verworfenen Strategien zu bekommen“ (Grüßing, 2002, S. 39).

> „Ein weiteres Ziel war es, mit den Testaufgaben sowohl Unterrichtsthemen aufzugreifen (Thema Netze in den Aufgaben 7 bis 9, im Hintergrund vorrangig die Unterrichtsstunden 4 bis 7), als auch Fähigkeitsbereiche zu überprüfen, die im Unterricht nicht explizit betont werden. Hierzu zählen die Aufgaben 1 bis 3, 5 und 6, die Teilfähigkeiten ansprechen, die nach unserer Meinung vor allem in den Unterrichtsstunden 1 bis 3, 6 und 7 implizit benutzt werden“ (Meißner, 2006, S. 35).

Die wiedergegebenen Ausführungen zur Methodik der berücksichtigten Arbeiten zeigen, dass mathematikdidaktische Forschung häufig komplementär zur psychometrischen Forschung stattfindet. In diesem Sinne können sich beide Herangehensweisen ergänzen. Dies gilt insbesondere auch, weil den genannten Stärken in den mathematikdidaktischen Arbei-

ten (Erfassung und Interpretation individueller Denkprozesse und sozialer Aushandlungsprozesse; unterrichtsnahes Testmaterial) manchmal methodische Schwächen bei der quantitativen Auswertung von Tests gegenüber stehen. So fehlen häufig explizite Überlegungen zum Testmodell[78] und – damit zusammenhängend – zum Skalenniveau der erhobenen Daten sowie zu inhaltlich zulässigen Auswertungsverfahren. Gerade bei Studien mit Vor- und Nachtests bietet die psychologische Methodenlehre einen deutlich ausgereifteren Umgang mit den Ergebnissen an (z. B. mit Test-Nachtest-Effekten).

Die vorliegende Arbeit setzt stärker auf die psychometrische Messung der *Raumvorstellung*, um die statistischen Zusammenhänge mit *Mathematikleistung* zu bestimmen. Diese Befunde sollten später aber – dies kann jetzt schon vorweggenommen werden – durch qualitative mathematikdidaktische Studien und Interventionsstudien ergänzt werden, um Wirkungen zwischen *Raumvorstellung* und *Mathematikleistung* inhaltlich zu identifizieren.

3.2.3 Ausgewählte Modelle

Im Folgenden werden exemplarisch die mathematikdidaktischen Modelle von Maier (1994, 1996, 1999a, 1999b) und Pinkernell (2003) dargestellt, um zu zeigen, wie sich solche Modelle von denen aus Kap. 3.1.2 unterscheiden und um zu reflektieren, worin dies begründet liegt.[79] Dabei ist Maiers Modell formal näher an denen aus Kap. 3.1.2; es versucht, Teilfähigkeiten der *Raumvorstellung* sinnvoll zu strukturieren, wobei eine vollständige Überdeckung des fraglichen Bereichs kognitiver Fähigkeiten wichtiger zu sein scheint als Trennschärfe zwischen den Kategorien. Das Modell von Pinkernell ist nicht nur formal anders aufgebaut, sondern auch inhaltlich weiter gefasst; mit ihm sollen Unterrichtsprozesse und individuelle Aufgabenbearbeitungen, bei denen *Raumvorstellung* eine Rolle spielt, für eine Erfassung und Interpretation zugänglich gemacht werden.

Maier (1994) geht bei seinem Modell von der kognitionspsychologisch und psychometrisch orientierten Beforschung des Gegenstandes aus, diskutiert u. a. die in Kap. 3.1.2 dargestellten Modelle von Thurstone und Linn & Petersen und führt beide in einer „‚Landkarte' der räumlichen Intelligenz" (Maier, 1999b, S. 14) pragmatisch (additiv) zusammen:

> „Von epochaler Bedeutung ist die **3-Faktoren-Hypothese nach Thurstone** (1949; 1950), deren Subfaktoren, in umfassender definitorischer Bedeutung, noch heute grundlegende Signifikanz besitzen. Das **Kategoriensystem nach Linn & Petersen** (1985; 1986) stellt eine herausragende Ergänzung dar. Durch Zusammenfassung der beiden Kategoriensysteme manifestieren sich insgesamt **die fünf wesentlichsten Komponenten räumlich-visueller Qualifikationen**" (Maier, 1994, S. 50 f.; Herv. i. O.).

[78] „Wie hängen das beobachtete Testverhalten und die zu messende Fähigkeit miteinander zusammen?"

[79] Die anderen, in Kap. 3.2.1 und Kap. 3.2.2 zitierten Studien arbeiten nicht mit ähnlich elaborierten eigenen Modellen, sondern schließen pragmatisch an vorhandene psychologische und mathematikdidaktische Modelle an. Die Studien von Lüthje und Grüßing beziehen sich dabei auf das hier betrachtete Modell von Maier.

Abbildung 3.5: Maiers ‚Landkarte' der räumlichen Intelligenz (nach Maier, 1994, S. 71)

<table>
<tr><th>Standpunkt der Probanden</th><th>Dynamische Denkvorgänge
Räumliche Relationen am Objekt veränderlich</th><th colspan="2">Statische Denkvorgänge
Räumliche Relationen am Objekt unveränderlich; Relation der Person zum Objekt veränderlich</th><th>Einsatz analytischer Strategien</th></tr>
<tr><td rowspan="2">Person befindet sich außerhalb</td><td>VERANSCHAULICHUNG</td><td colspan="2">RÄUMLICHE BEZIEHUNGEN</td><td>Analytische Strategien zum schlussfolgernden Denken häufig hilfreich</td></tr>
<tr><td colspan="2">VORSTELLUNGSFÄHIGKEIT VON ROTATIONEN</td><td rowspan="3">Analytische Strategien zum schlussfolgernden Denken insbesondere im dynamischen Bereich häufig nicht hilfreich</td></tr>
<tr><td rowspan="2">Person befindet sich innerhalb</td><td>RÄUMLICHE ORIENTIERUNG</td><td>RÄUMLICHE WAHRNEHMUNG</td></tr>
<tr><td>FAKTOR K</td></tr>
</table>

Maier (1994, S. 51 f.) führt aus, dass die Bedeutung der alltagsrelevanten Komponenten „Veranschaulichung", „Räumliche Beziehungen" und „Räumliche Orientierung" durch die Größe der entsprechenden Kästchen ausgedrückt wird. Die hell- und dunkelgrauen Schattierungen unterscheiden Komponenten, bei denen die Versuchsperson Teil des räumlichen Problems ist (dunkelgrau) bzw. das Problem überwiegend „von außen" betrachtet (hellgrau). Darüber hinaus werden Komponenten idealtypisch unterschieden, bei denen dynamische Denkvorgänge bzw. statische Denkvorgänge in Erscheinung treten. Die zusätzlich zu Thurstone und Linn & Petersen aufgenommene Komponente „Faktor K" dürfte dem bei Thurstone und Michael et al. angedeuteten *K* entsprechen (vgl. Kap. 3.1.2). Die gestrichelte Linie zwischen „Faktor K" und „Räumliche Orientierung" deutet an, dass der „Faktor K" eigentlich Teil von „Räumlicher Orientierung" ist. Maiers Studie von „Lösungsstrategien bei räumlichen Teilleistungen" (1994, S. 57 ff.) führt zu der Einteilung in der rechten Spalte.

Für Maiers Zielsetzung sind die Einwände gegen Thurstones Trennung von *S1* und *S3* (vgl. Kap. 3.1.2) anscheinend ebenso wenig relevant wie die deutlichen Überschneidungen in den verwendeten Kategorien. So verwenden Linn & Petersen z. B. den Test *Flags* als Referenztest für *Mental Rotation* (bei Maier „Vorstellungsfähigkeit von Rotationen"), während dieser von Thurstone als Referenztest für *S1* (bei Maier „Räumliche Beziehung") genutzt wird. Insgesamt ist die Strukturierung dieser Landkarte nur vor dem Hintergrund idealtypisch gedachter Anforderungen in entsprechenden Raumvorstellungstests plausibel.

So betonte z. B. die Kritik an Thurstones Trennung von *S1* und *S3* gerade, dass sich Versuchspersonen sowohl bei *S1* wie auch bei *S3* jeweils als Teil des Problems verstehen können – oder auch nicht. Auch können bei Referenztests wie *Flags* eher statische oder eher dynamische Denkvorgänge angewandt und analytische Strategien ebenfalls optional verwendet werden. Die „Landkarte" kann also kein Strukturmodell der *Raumvorstellung* sein, das kognitionspsychologischer oder psychometrischer Überprüfung standhält.

Diese Einwände sollen aber nicht entkräften, dass Maier mit seiner „Landkarte" zentrale Kategorien der *Raumvorstellung*, die vor allem in kognitionspsychologischen und psychometrischen Studien auftreten, für die Mathematikdidaktik erschlossen hat. Durch die Aufnahme des großen Bereichs „Räumliche Orientierung" beschränkt er sich dabei nicht auf *Small-Scale Fähigkeiten* (vgl. S. 73), sondern verweist auf die Bedeutung von Fähigkeiten, die sich nicht in *Paper and Pencil Tests* erfassen lassen. Durch die additiv zustande gekommene größere Kategorienzahl lassen sich vorhandene Studien leichter verorten. Die Betrachtung der Komponenten von *Raumvorstellung* unter den Perspektiven der Bearbeitungsstrategien, der Denkvorgänge und des Involviertseins der Versuchsperson ist für die Gestaltung von Lehr-Lernsituationen fruchtbar, da deutlich wird, welche Anforderungen bei Raumvorstellungsaufgaben variiert werden können. Für den Mathematikunterricht kann das vorgeschlagene Schema also ein hilfreiches Rezeptions-, Analyse- oder Konstruktionswerkzeug sein, wenn die enthaltene Strukturierung nicht zu strikt verstanden wird. Nicht zuletzt diese Nützlichkeit kann erklären, warum Maiers Arbeiten bei allen späteren mathematikdidaktischen Studien (im deutschsprachigen Raum) eine wichtige Grundlage darstellt.

Das Modell von Pinkernell (2003) verfolgt eine andere Zielsetzung und ist auch deutlich anders strukturiert.

> „Ziel der Arbeit hier ist die Formulierung und Begründung eines mathematikdidaktisch geprägten Modells des Räumlichen Vorstellungsvermögens. Ein solches Modell erschien notwendig, nachdem eine quantitative Auswertung der Unterrichtsreihe zwar positiv ausfiel, jedoch keine angemessen umfassende Bewertung liefern konnte. […] … neu ist dagegen eine Neusortierung der räumlich-visuellen Fähigkeiten solchermaßen, dass Berührungspunkte mit Aspekten des Lehrens und Lernens von Mathematik deutlich werden" (ebd., S. 55).

> „Es [das Modell] berücksichtigt neben den rein mentalen räumlich-visuellen Vorstellungsinhalten auch die Handlungen im realen Raum und mit räumlichen Objekten. Es thematisiert Zusammenhänge des Handelns im Raum und mit räumlichen Objekten mit Kenntnissen, Fertigkeiten und Fähigkeiten, die im Geometrieunterricht wie auch in anderen Bereichen der Mathematik vermittelt werden" (ebd., S. 13; Erg. d. d. Verf.).

Pinkernell entwickelt drei Kategorien eines mathematikdidaktischen Modells von *Raumvorstellung*, das offensichtlich viel weiter gefasst ist als die klassischen psychologischen Modelle, die in Kap. 3.1.2 dargestellt sind (Pinkernell, 2003, S. 137 f.):

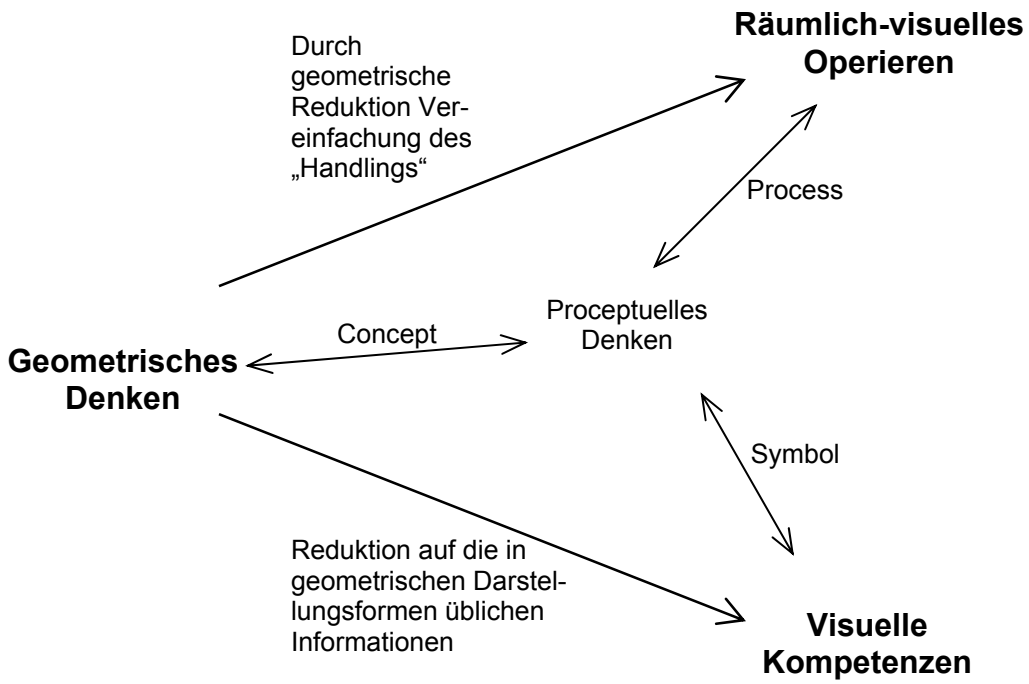
- *Räumlich-visuelles Operieren*: Hierunter fallen alle Handlungen im mentalen Raum und im realen Raum. „In dieser Perspektive richtet sich der Begriff des Räumlichen Vorstel-

lungsvermögens auf die internen sowie externen Repräsentationen von Raumobjekten und ihren Transformationen“ (ebd., S. 137). In diese Kategorie fallen insbesondere die klassischen psychologischen Komponenten der *Raumvorstellung*.

- *Geometrisches Denken*: Dies bedeutet vor allem die Erfassung räumlicher Objekte und Konfigurationen durch geometrische Begriffe (im umfassenden Sinne von „Objekte und ihre Eigenschaften sowie Beziehungen“).
- *Visuelle Kompetenzen* (oder auch Visualisierungskompetenzen): Hierbei geht es um das „Interpretieren und Konstruieren verschiedener in der Geometrie üblicher Darstellungsformen von räumlich-visuellen Objekten“ (ebd., S. 138).

Abbildung 3.6 verdeutlicht die Beziehungen zwischen diesen drei Kategorien, wobei Pinkernell zur Verbindung der Kategorien auf das Konzept der „Procepts“ von Tall & Gray zurückgreift, das – hier etwas verkürzt dargestellt – die Einheit von Begriff, Handlung und Darstellung betont.

Abbildung 3.6: Pinkernells Modell der Raumvorstellung (nach Pinkernell, 2003, S. 139)



Pinkernell hat sein Modell in direkter Auseinandersetzung mit Beobachtungen von Unterricht entwickelt, in dem Schülerinnen und Schüler räumlich-visuelle Anforderungen bewältigen. Das Modell umfasst entsprechend seiner Zielsetzung in größerem Umfang mathematische Kompetenzen („Geometrisches Denken“, „Darstellen“) und über die „Visualisierungskompetenz“ auch den für den Unterricht wichtigen Aspekt der Kommunikation. Dieses Modell scheint einen Mehrwert vor allem für solche Fragestellungen zu haben, die

sich aus mathematikdidaktischer Perspektive nicht nur mit Testleistungen zur *Raumvorstellung* befassen, sondern auch individuelle Vorstellungen und Bearbeitungsstrategien bei der Bewältigung von räumlich-visuellen Anforderungen berücksichtigen. Insofern hätten Studien wie die von Koops & Sorger (1980), Grüßing (2002) oder Lüthje (2008, 2009) mit Pinkernells Modell noch weitergehende Befunde bringen können. Für die reine Messung der *Raumvorstellung* wird man hingegen eher auf die klassischen psychologischen Modelle zurückgreifen oder zumindest in Anlehnung an deren Klassifikationen entsprechende Tests entwickeln.

3.3 Befunde zur Raumvorstellung

Das Feld der Raumvorstellungstests sowie darauf basierender Studien und Befunde ist heute kaum noch überschaubar. Alleine aufgrund der immer wieder substanziell nachgewiesenen Geschlechterunterschiede in spezifischen Komponenten der *Raumvorstellung* ist dieser Gegenstand nicht zuletzt für die Differenzielle Psychologie und die Genderforschung von großem Interesse. Dementsprechend werden solche Komponenten der *Raumvorstellung* besonders intensiv untersucht, in denen große Leistungsunterschiede zwischen den Geschlechtern oder zumindest geschlechtsspezifische Bearbeitungsstrategien auftreten. Dies trifft in besonderer Weise für den engen Bereich *Mental Rotation* und hier auf den *MRT* zu. Für diesen Test lassen sich relativ stabil größere Geschlechterunterschiede identifizieren und dennoch ist es bislang nicht gelungen, die Unterschiede inhaltlich zu erklären.

Im Folgenden werden solche Forschungsergebnisse skizziert, die potenziell für das Erkenntnisinteresse der vorliegenden Arbeit bedeutsam sind. Die vorgestellten Studien und Befunde werden dabei nach den folgenden Leitfragen sortiert dargestellt:

- Wie entwickelt sich die *Raumvorstellung* über die Lebensspanne?
- Welche Geschlechterunterschiede gibt es bei der *Raumvorstellung*?
- Mit welchen Strategien bearbeiten Versuchspersonen Raumvorstellungstests?
- Wie lassen sich interindividuelle Unterschiede in der *Raumvorstellung* erklären?
- Wie hängen *Raumvorstellung* und *Mathematikleistung* zusammen?
- Lassen sich Geschlechterunterschiede in der *Mathematikleistung* durch entsprechende Unterschiede in der *Raumvorstellung* erklären?
- Wie kann die *Raumvorstellung* gefördert werden?

Die Sortierung der Forschungsergebnisse nach diesen sieben Leitfragen soll das Feld ordnen und übersichtlicher gestalten. Gleichwohl gibt es vielfältige Berührungspunkte zwischen den Fragen, die häufig durch Interaktionen zwischen den hauptsächlich fokussierten Konstrukten bedingt sind.

Im Rahmen dieser Arbeit werden fast ausschließlich die klassischen *Small-Scale Fähigkeiten* betrachtet, die durch *Paper and Pencil Tests* erfasst werden können. Dieses Vorgehen kann unter Umständen zu Validitätsproblemen führen, wie die Arbeit von Bishop (1983) zeigt. So ist es keineswegs selbstverständlich, dass Versuchspersonen die ebenen Darstellungen von dreidimensionalen Objekten und Konfigurationen räumlich interpretieren können. Bei allen *Paper and Pencil Tests*, die ein mentales Operieren im dreidimensionalen Raum erfordern, ist dies aber eine notwendige Voraussetzung der Testbearbeitung. Die Bedeutung der genannten Interpretationsleistungen ist für Bishop so groß, dass er „Interpreting Figural Information (IFI)“ als eine von zwei Komponenten in seinem Modell der *Raumvorstellung* vorsieht. Bishop (1983, S. 189) weist in seiner Arbeit darauf hin, dass die fragliche Fähigkeit vor allem kultur- bzw. sozialisationsbedingt ist, also von Lerngelegenheiten abhängt. Diese Lerngelegenheiten dürften aber bei fast allen Jugendlichen und Erwachsenen in Industrienationen hinreichend gegeben sein. Die potenziellen Validitätsprobleme dürften also insbesondere für die Population, die im empirischen Teil der vorliegenden Arbeit untersucht wird, vernachlässigbar sein.

3.3.1 Entwicklung über die Lebensspanne

Entsprechend der Differenzierungshypothese der Intelligenz (vgl. Kap. 3.1.1) kann die innere Struktur der Intelligenz erst ab einem Lebensalter von etwa 6 Jahren empirisch erfasst und beschrieben werden (vgl. D. H. Rost, 2009, S. 244 ff). Demzufolge kann ein eigenständiger Leistungsbereich *Raumvorstellung* frühestens ab diesem Alter differenziert betrachtet werden. Für die vorliegende Arbeit ist von besonderem Interesse, ob die *Raumvorstellung* bei der „*PISA*-Population“ der 15-Jährigen bzw. bei der Population der Schülerinnen und Schüler, die eine neunte Klasse besuchen, schon hinreichend weit entwickelt ist, um potenziell als Mediatorvariable für Geschlechterunterschiede in der *Mathematikleistung* infrage zu kommen.

Mit Blick auf die vorliegende Literatur kann diese Frage eindeutig bejaht werden (vgl. z. B. D. H. Rost, 1977, S. 46 f.; Maier, 1994, S. 78):

- Bereits Thurstone (1955) schätzt im Rahmen der Leistungsentwicklung für seine sieben „Primary Mental Abilities“, dass im Alter von 15 Jahren beim Faktor „Space“ mehr als 80 % der maximalen Leistung erreicht ist.
- Die Normkurven im „Leistungsprüfungssystem (LPS)“ von Horn (1962) zeigen, dass bereits im Alter von 15 Jahren der Leistungshöhepunkt im Bereich der berücksichtigten Raumvorstellungstests nahezu erreicht ist und dass die Leistung schon im jungen Erwachsenenalter wieder abnimmt.
- Dazu passend berichten Herzberg & Lepkin (1954), deren Hauptaugenmerk eigentlich auf Geschlechterunterschieden liegt, Leistungswerte für 16-, 17- und 18-Jährige im Bereich der *Raumvorstellung*, die (bei beiden Geschlechtern) von den 16-Jährigen zu den 18-Jährigen gleichmäßig abnehmen.

- Dagegen sehen Vandenberg & Kuse (1978) bei ihrem „Mental Rotation Test (MRT)“ den Leistungshöhepunkt erst etwa bei einem Lebensalter 23 Jahren bis 28 Jahren, wobei aber schon 16-Jährige auf über 90 % der maximalen Leistung kommen.

Etwas jünger als die zuvor genannten Tests und Befunde sind der „Dreidimensionale Würfeltest (3 DW)“ von Gittler (1990) und der „Bausteine-Test (BST)“ von Birkel et al. (2002). Beide liefern für die Frage der Leistungsentwicklung über die Lebensspanne ähnliche Hinweise:

- Der *3 DW* ist ein Rasch-homogener Test, der als Weiterentwicklung eines noch inhomogenen Subtests mit Würfelaufgaben aus dem „Intelligenz-Struktur-Test (I-S-T 70)“ von Amthauer (1970) entstanden ist (vgl. Kap. 3.3.3). Die Normtabellen des *3 DW* (Gittler, 1990. S. 55 ff.) zeigen, dass die Testleistung im Lebensalter von 13 bis 19 Jahren kontinuierlich steigt. Wie diese Entwicklung nach dem 19. Lebensjahr weitergeht, lässt sich anhand der Normtabellen nicht eindeutig klären. Die vorhandenen Daten deuten allerdings an, dass es sich höchstens um unwesentliche Steigerungen handeln dürfte und dass die Testleistung im mittleren und höheren Erwachsenenalter wieder abnimmt. 15-Jährige dürften – grob geschätzt – etwa 80 % der maximalen Testleistung erreichen.
- Der *BST* ist ein Test, der vor allem „Anteile von logisch schlussfolgerndem und analytischem Denken im dreidimensionalen Raum“ (Birkel et al., 2002, S. 5) erfassen soll. Den Versuchspersonen werden zunächst vier Bausteine präsentiert, die aus jeweils vier Würfeln bestehen. Anschließend sollen sie für Zielfiguren, die jeweils aus acht Würfeln bestehen, entscheiden, aus welchen zwei der vier vorgegebenen Bausteine diese zusammengesetzt sind. Die Normwerte des *BST* zeigen, dass die Testleistung im Lebensalter von 13 bis 17 Jahren kontinuierlich steigt. Wie diese Entwicklung nach dem 17. Lebensjahr weitergeht, bleibt ebenfalls offen. 15-Jährige erreichen etwa 90 % der Testleistung von 19-Jährigen.

Die zitierten Ergebnisse weisen darauf hin, dass sich die Frage, wann genau der Leistungshöhepunkt in der *Raumvorstellung* erreicht ist, vermutlich nicht global beantworten lässt, sondern nach Komponenten der *Raumvorstellung* – vielleicht auch nach konkreten Tests – differenziert betrachtet werden muss. Die Ergebnisse zeigen aber auch, dass bereits die Population der 15-Jährigen der maximalen Testleistung (für die Zwecke der vorliegenden Arbeit) hinreichend nahe kommen und dass eine hinreichende Varianz in den Testleistungen dieser Population vorhanden ist, um Zusammenhänge mit anderen Leistungsbereichen differenziert untersuchen zu können.

3.3.2 Geschlechterunterschiede

Die *Raumvorstellung* gilt als *die* klassische Komponente der Intelligenz, bei der es (die größten) Geschlechterunterschiede zugunsten der Männer gibt. McGee (1979, S. 42 ff.) berichtet von zwei Studien aus der Mitte des 20. Jahrhunderts, die auf der Basis von Thurstones „Primary Mental Abilities“ zwar höhere globale IQ-Werte bei Mädchen,

zugleich aber auch eine signifikant bessere *Raumvorstellung* bei Jungen gemessen haben. Zu einem ähnlichen Resümee gelangt D. H. Rost (1977, S. 29): „Der allgemeinen Überlegenheit des weiblichen Geschlechts im Verbalbereich […] steht eine bessere Leistung des männlichen Geschlechts im logischen Denken (reasoning) und in den Komponenten der Raumvorstellung und Raumorientierung gegenüber."

Derartige Befunde waren schon immer Anlass genauer hinzuschauen: Ein Blick in einschlägige Literaturdatenbanken offenbart, dass Geschlechterunterschiede den am intensivsten untersuchten Teilbereich innerhalb der psychologischen Raumvorstellungsforschung darstellen. Diese intensiven Forschungsaktivitäten führen zu deutlich differenzierteren Befunden als denen, die McGee berichten konnte, und auch zu scheinbaren und tatsächlichen Widersprüchen zwischen verschiedenen Studien. So gibt es Komponenten der *Raumvorstellung*, in denen es anscheinend stabil relativ große Unterschiede zugunsten der Jungen bzw. Männer gibt, während es bei anderen Komponenten vermutlich keine signifikanten Unterschiede gibt. Entsprechende Befunde sind aber wiederum abhängig vom konkreten Testmaterial und der untersuchten Population. Unklar ist dabei auch, ob es im Laufe der vergangenen Jahrzehnte messbare Veränderungen bei den Geschlechterunterschieden gibt.

Ein Versuch, alle Befunde zu Geschlechterunterschieden in der *Raumvorstellung* zu systematisieren und übersichtlich zu berichten, würde den Rahmen der vorliegenden Arbeit sprengen. Die erforderliche Auswahl von Befunden berücksichtigt primär die Fragestellung dieser Arbeit und dient vor allem der Vorbereitung des empirischen Teils. Daher ist es wichtig, dass die Befunde sich auf gut beschreibbare Komponenten der *Raumvorstellung* beziehen, zu denen es bewährte paradigmatische Referenztests gibt. Die Befunde sollten zudem über mehrere Studien hinweg relativ stabil sein.

Die Meta-Analyse von Linn & Petersen

Eine gute Orientierung liefert die Meta-Analyse von Linn & Petersen (1985). Wie in Kap. 3.1.2 dargestellt wurde, basiert diese Arbeit auf drei kognitionspsychologisch identifizierten Komponenten der *Raumvorstellung* (*Spatial Perception*, *Mental Rotation* und *Spatial Visualization*). Die in die Meta-Analyse aufgenommenen 172 Effektstärken für Geschlechterunterschiede wurden zunächst nach diesen drei Komponenten sortiert und anschließend noch für *Spatial Perception* in drei Alterklassen (unter 13 Jahre, 13 bis 18 Jahre und über 18 Jahre) und für *Mental Rotation* in zwei Klassen von Referenztests unterteilt („dreidimensionale" bzw. „zweidimensionale" Tests)[80]. Innerhalb der jeweiligen Klasse sind die Effektstärken weitgehend homogen und ergeben das folgende Bild (S. 1485 ff.):

[80] Hier und im Folgenden werden die Adjektive „zweidimensional" und „dreidimensional" immer dann in Anführungszeichen geschrieben, wenn sie sich auf *Paper and Pencil Tests* beziehen. Dies ist vor allem für den „dreidimensionalen" *MRT* von Bedeutung, da er den Versuchspersonen zweidimensional (sic!) präsen-

- Die größten Geschlechterunterschiede gibt es in der Komponente *mentale Rotation*[81] für den „dreidimensionalen“ *MRT* (Vandenberg & Kuse, 1978). Die aus den einbezogenen Studien geschätzte Effektstärke beträgt 0,94 zugunsten der männlichen Versuchspersonen; sie unterscheidet sich signifikant von Null. Für „zweidimensionale“ Referenztests dieser Komponente ergibt sich eine deutlich kleinere geschätzte Effektstärke von 0,26, die sich aber auch noch signifikant von Null unterscheidet. Bereits in Kap. 3.1.2 wurde bemerkt, dass diese „zweidimensionalen“ Tests vermutlich auch über andere kognitive Prozesse als mentale Rotation gelöst werden können (vgl. auch Heil & Jansen-Osmann, 2008) und daher höchstens unscharf einer Komponente zugeordnet werden können.
- Innerhalb der Komponente *räumliche Wahrnehmung* beträgt die geschätzte Effektstärke bei den über 18-Jährigen 0,64 zugunsten der männlichen Versuchspersonen; sie unterscheidet sich (deutlich) signifikant von Null. In den beiden anderen Altersgruppen (unter 13 Jahre und 13 bis 18 Jahre) beträgt die geschätzte Effektstärke jeweils 0,37, unterscheidet sich aber nicht signifikant von Null.
- Für die Komponente *räumliche Visualisierung* beträgt die geschätzte Effektstärke zwar noch 0,13 zugunsten der männlichen Versuchspersonen, das Konfidenzintervall für die Schätzung (zur Wahrscheinlichkeit 95 %) geht aber von –0,24 bis 0,50, schließt die Null also deutlich ein. Die 81 Effektstärken, die von den Autorinnen einbezogen wurden, liegen zwischen –0,91 und 0,71 (wiederum aus Sicht der männlichen Versuchspersonen). Auch eine nach Altergruppen differenzierte Betrachtung, wie sie bei *räumliche Wahrnehmung* durchgeführt wurde, ergibt keine signifikanten Geschlechterunterschiede.

Im Folgenden werden zunächst weitere (vor allem auch jüngere) Befunde zu den drei Komponenten von Linn & Petersen berichtet, bevor andere Tests berücksichtigt werden, die sich diesen Komponenten nicht so trennscharf zuordnen lassen. Schließlich wird den Fragen der Stabilität von Geschlechterunterschieden über die Lebensspanne und im Wandel der Zeit nachgegangen.

Weitere Befunde zu mentaler Rotation

Im Bereich der („dreidimensionalen“) *mentalen Rotation* liefert der *MRT* von Vandenberg & Kuse (1978), die sich bei der Testentwicklung auf Figuren von Shephard & Metzler (1971) stützten, gut replizierbar signifikante Geschlechterunterschiede zugunsten der

tiert wird, die Versuchspersonen aber mentale Repräsentationen dreidimensionaler Objekte zur Aufgabenbearbeitung heranziehen sollen.

[81] In der vorliegenden Arbeit werden die drei Komponenten nach Linn & Petersen (1985) fortan mit den deutschen Übersetzungen *räumliche Wahrnehmung*, *mentale Rotation* und *räumliche Visualisierung* bezeichnet. Wenn diese Begriffe im Folgenden ohne weiteren Kommentar verwendet werden, so sind sie im Sinne der Festlegung von Linn & Petersen (1985) zu verstehen.

männlichen Versuchspersonen. Mittlerweile liegt eine neu erstellte Version des *MRT* von Peters et al. (1995) vor, die zu denselben Ergebnissen führt. So bestätigen Voyer et al. (1995) in einer jüngeren Meta-Analyse die oben zitierten Befunde von Linn & Petersen (1985) zur *mentalen Rotation*. Aufgrund dieser relativen Stabilität der Geschlechterunterschiede beim *MRT* und der Tatsache, dass diese Unterschiede bis heute nicht befriedigend erklärt werden konnten, gibt es eine fortgesetzte intensive Erforschung der mentalen Rotation und des *MRT* (vgl. z. B. Geiser et al., 2006, 2008; Jansen-Osmann & Heil, 2007; Peters et al., 2006; Voyer & Doyle, 2010; Voyer & Saunders, 2004; Voyer & Hou, 2006).

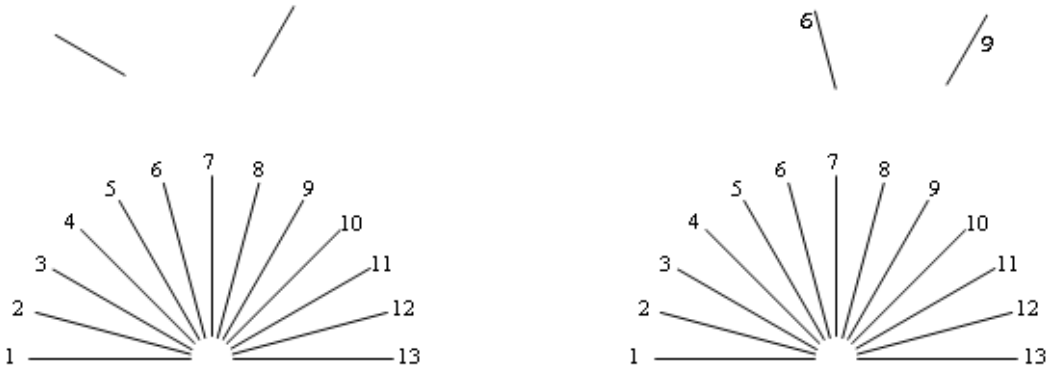
Titze et al. (2008) konnten in einer aktuellen Studie zeigen, dass – anders als zuweilen vermutet (vgl. z. B. Goldstein et al., 1990) – weder die Aufgabenkomplexität des *MRT* noch die Zeitrestriktion aus der originalen Testanweisung ursächlich für die gefundenen Geschlechterunterschiede sind. Die Effektstärke zugunsten der Männer beträgt in dieser Studie 0,42 (Titze et al., 2008, S. 132). In der Studie von Geiser et al. (2008) konnten in einer Stichprobe von 9- bis 23-Jährigen signifikante Geschlechterunterschiede zugunsten der männlichen Versuchpersonen in allen Altersgruppen nachgewiesen werden; die Effektstärken betrugen zwischen 0,52 und 1,49.

Weitere Befunde zu räumlicher Wahrnehmung

Im Bereich der *räumlichen Wahrnehmung* sind nach Linn & Petersen (1985) solche Aufgaben paradigmatisch, die eine Identifikation der Vertikalen bzw. Horizontalen erfordern. Dementsprechend eignen sich vor allem „Water Level Tasks (WLT)“ und der „Rod and Frame Test (RFT)“ als Referenztests. Dabei ist allerdings fraglich, ob sich das Testverhalten von Versuchpersonen in diesen beiden Tests wirklich auf eine zugrundeliegende kognitive Fähigkeit zurückführen lässt, da Korrelationen zwischen beiden Tests in einigen Studien (teilweise geschlechterabhängig) relativ gering ausfielen (z. B. Liben, 1978; Quaiser-Pohl et al., 2004). Dabei ist in der Literatur aber im Wesentlichen unstrittig, dass es in diesem Bereich signifikante Geschlechterunterschiede zugunsten männlicher Versuchspersonen gibt (vgl. Liben, 1991, S. 115). Wie bei der *mentalen Rotation* sind diese Geschlechterunterschiede auch bei der *räumlichen Wahrnehmung* noch nicht befriedigend erklärt (vgl. Liben 1991, S. 126 f.).

Collaer & Nelson (2002) finden in einer jüngeren Studie mit anderem Testmaterial sogar erhebliche Geschlechterunterschiede mit einer Effektstärke von 0,85. Sie verwenden den Test „Judgment of Line Angle and Position (JLAP)“, bei dem die Versuchpersonen pro Item für zwei Linien jeweils entscheiden sollen, welche von 13 vorgegebenen Linien die gleiche Ausrichtung hat (vgl. Abb. 3.7). Dies erfordert augenscheinlich eine differenzierte Orientierung, die mit der Identifikation der Horizontalen und Vertikalen zusammenhängen dürfte. Die Aufgabenschwierigkeit hängt bei beiden Geschlechtern deutlich von der Orientierung der richtigen Lösung ab, wobei die Testleistung von weiblichen Versuchpersonen stärker von der Orientierung abhängt.

Abbildung 3.7: Beispielitems für den Test „Judgment of Line Angle and Position (JLAP)“ (rechtes Item mit Ausweisung der richtigen Lösung)



Weitere Befunde zu räumlichen Visualisierung

Weniger klar sind die Befunde zur *räumlichen Visualisierung*. Schon Linn & Petersen (1985) konnten, wie oben dargestellt wurde, keine konsistenten Geschlechterunterschiede finden: Verschiedene Studien kommen zu unterschiedlichen Ergebnissen, die vom konkret eingesetzten Test und von der untersuchten Population abhängen; so gibt es jeweils nicht zu vernachlässigende Anteile, die (a) keine signifikanten Geschlechterunterschiede, (b) solche zugunsten weiblicher Versuchspersonen oder (c) solche zugunsten männlicher Versuchungspersonen finden. In jüngeren Untersuchungen (z. B. Manger & Eikeland, 1998) wird fast durchgängig berichtet, dass keine signifikanten Geschlechterunterschiede vorliegen. Da *räumliche Visualisierung* sowohl im Sinne von Linn & Petersen (1985) als auch bei Zimmermans (1954) Einordnung im Rahmen seiner Kontinuumhypothese komplexere Aufgabenstellungen umfasst und neben rein räumlichen auch analytische Lösungsstrategien erfordert, können die heterogenen Befunde womöglich auf eine nicht besonders trennscharfe Konzeption dieser Raumvorstellungskomponente zurückgeführt werden.

Befunde zu anderen Raumvorstellungstests

Relativ konsistente (signifikante) Geschlechterunterschiede lassen sich auch bei den beiden Raumvorstellungstests *BST* und *3 DW* finden, deren Normtabellen schon für die Fragestellung von Kap. 3.3.1 ausgewertet wurden. Beide Tests bestehen aus etwas komplexeren Aufgaben, zu deren Lösung mentale Rotation und die Berücksichtigung von räumlichen Beziehungen erforderlich sind. Die trennscharfe Einordnung in die Komponenten nach Linn & Petersen (1985) gelingt somit kaum. Zwar erfüllen die jeweiligen Aufgaben weitgehend die Charakteristika von *räumlicher Visualisierung*, der Anteil von *mentaler Rotation* scheint aber deutlich höher zu sein als bei reinen Aufgaben zu *räumlicher Visualisierung* (wie z. B. dem *DAT:SR*, siehe Kap. 3.1.2).

Im Testmanual für den *BST* (Birkel et al., 2002, S. 33 f.) findet man u. a. die mittleren Testleistungen von Jungen und Mädchen nach Schulformen (Hauptschule, Realschule und Gymnasium) und Jahrgangsstufen bzw. Doppeljahrgangsstufen (insgesamt 8. bis 11. Jahrgang) getrennt ausgewiesen. Hieraus lassen sich – zusammen mit den jeweils für die Gesamtgruppe berichteten Standardabweichungen – die zugehörigen Effektstärken berechnen; sie liegen zwischen 0,51 und 0,72 (jeweils zugunsten der Jungen).

Für den *3 DW* lassen sich auf der Basis der nach unterschiedlichen Fallgruppen getrennt berichteten mittleren Testrohwerte (Gittler, 1990, S. 54 ff.) Effektstärken berechen, die zwischen 0,15 und 0,64 liegen (jeweils zugunsten der männlichen Versuchspersonen).

Die für die beiden Tests berechneten Effektstärken passen gut zum Befund der theoretischen Analyse der jeweiligen Aufgaben: Wenn *mentale Rotation* tatsächlich eine wichtige Rolle bei der Aufgabenlösung spielt, dann lassen sich hiermit gut die signifikanten Geschlechterunterschiede erklären.

Befunde zu Geschlechterunterschieden über die Lebensspanne

Die Frage, in welchem Lebensalter Geschlechterunterschiede zuerst beobachtet werden können, lässt sich auf der Basis der vorliegenden Literatur nicht eindeutig beantworten. Dabei ist klar, dass die Frage nach Raumvorstellungskomponenten oder gar nach verschiedenen Tests getrennt beantwortet werden muss, da z. B. für die Komponente *räumliche Visualisierung* keine signifikanten bzw. nur in einer Minderheit der Studien signifikante Geschlechterunterschiede berichtet werden.

Bei McGee (1979, S. 41) findet man aber auch eher globale Aussagen, dass signifikante Geschlechterunterschiede vor Beginn der Pubertät nicht zuverlässig nachweisbar sind. Diese Einschätzung wird von Maier (1999b, S. 205; i. Orig. m. Herv.) geteilt: „Altersspezifisch ist festzustellen, dass Geschlechtsunterschiede in der *Raumvorstellung* bis zum Eintritt in das Pubertätsalter kaum in Erscheinung treten; sind jedoch Differenzen vorhanden, favorisieren sie die Leistungen der männlichen Probanden."

Halpern (2000, S. 106 f.) gelangt in ihrer Überblicksarbeit allerdings zu der Einschätzung, dass Geschlechterunterschiede in den Komponenten der *Raumvorstellung*, in denen sie signifikant auftreten, nachgewiesen werden können, sobald die entsprechenden Testleistungen zuverlässig erfasst werden können. In welchem Lebensalter dies ist, hängt dabei von der jeweiligen Komponente und dem konkreten Test ab. Für die Komponente *mentale Rotation* berichten Linn & Petersen (1985, S. 1488) beispielsweise, dass ihnen keine Studien vorlagen, in denen der *MRT* von Vandenberg & Kuse (1978) mit unter 13-Jährigen durchgeführt worden ist. In einer Studie zur zweidimensionalen *mentalen Rotation* konnten Heil & Jansen-Osmann (2008a) schon bei 8-jährigen Kindern einen signifikanten Geschlechterunterschied zugunsten der Jungen mit einer Effektstärke von 0,50 feststellen.

Unstrittig dürfte auf jeden Fall sein, dass signifikante Geschlechterunterschiede – sofern überhaupt welche existieren – in der Population der ca. 15-Jährigen, die in der vorliegenden Arbeit genauer untersucht wird, zuverlässig nachgewiesen werden können.

Befunde zu Geschlechterunterschieden im Wandel der Zeit

Auch die Frage, ob sich Geschlechterunterschiede in den vergangenen Jahrzehnten verändert (vor allem: verringert) haben, wird in der Literatur nicht einheitlich beantwortet. So sehen z. B. Stumpf & Klieme (1989) „More Evidence for Convergence". Hosenfeld et al. (1997) ergänzen diesen Befund, indem sie darauf hinweisen, dass Linn & Petersen (1985) und Voyer et al. (1995) zwar im Wesentlichen zu gleichen Befunde gekommen sind, aber die Effektstärken bei Voyer et al., die jüngere Studien in ihre Meta-Analyse einbeziehen, etwas geringer ausfallen.

Die Ergebnisse von Stumpf & Klieme (1989) beruhen auf Auswertungen des „Medizinertests" in den Jahren 1978 bis 1988 (18 Messzeitpunkte), der den Test „Schlauchfiguren" (Stumpf & Fay, 1983) als Subtest zur Erfassung der *Raumvorstellung* enthält. Die Korrelation zwischen Testzeitpunkt und Effektstärke (zugunsten männlicher Versuchspersonen) beträgt dabei –0,926. Die Effektstärke ist im Herbst 1978 mit 0,77 am höchsten und im Herbst 1987 mit 0,38 am niedrigsten. Trotz dieser beeindruckenden Korrelation treten beim zweiten Blick auf die berichteten Daten Fragen auf: So ist die Teilnehmerzahl an den ersten drei Messzeitpunkten dreistellig, an den darauffolgenden elf Messzeitpunkten vierstellig und an den letzten vier Messzeitpunkten sogar fünfstellig; die mittlere Testleistung verändert sich von Messzeitpunkt zu Messzeitpunkt in Anbetracht der am Ende großen Teilnehmerzahlen teilweise signifikant, wobei besonders auffällig ist, dass die Richtung der Veränderung für die mittlere Testleistung der Frauen und die mittlere Testleistung der Männer stets identisch ist. Anscheinend gibt es im (nicht erfassten) „Hintergrund" systematische Veränderungen, die nicht zwingend auf veränderte Sozialisationsbedingungen hinweisen müssen, sondern z. B. auch durch veränderte institutionelle Rahmenbedingungen o. Ä. erklärt werden können. Der identifizierte Trend mit einer Halbierung der Effektstärke von 0,77 auf 0,38 wäre wohl auch zu stark, wenn er nur durch den sozialen Wandel zwischen den Jahre 1978 und 1988 erklärt werden sollte.

Der dargestellten Konvergenzhypothese scheinen auch die jüngeren Resultate zum *MRT* und *JLAP* (s. o.) zu widersprechen, die teilweise Effektstärken in der Nähe von 1,00 oder sogar deutlich darüber bringen. Eine weitergehende Bewertung der Konvergenzhypothese erlauben diese Einzelbefunde aber nicht. Berücksichtigt man die zuvor benannte Kritik an den Interpretationen von Stumpf & Klieme (1989) und die Befunde zu unterschiedlichen Raumvorstellungstests, so deutet alles darauf hin, dass eine solide Bewertung der Konvergenzhypothese eine Längsschnittuntersuchung mit Subtests zu verschiedenen Raumvorstellungskomponenten und mit hinreichend großen Zufallsstichproben in verschiedenen Altersklassen erfordert.

3.3.3 Unterschiedliche Lösungsstrategien bei Testaufgaben

Quaiser-Pohl (1998, S. 27 ff.) diskutiert in ihrer Arbeit zu Geschlechterunterschieden bei der *Raumvorstellung* u. a. methodische Probleme, die bei der Erfassung solcher Unterschiede auftreten können. Dabei arbeitet sie heraus, dass bereits die methodische Herangehensweise an die Frage (z. B. „quantitative“ vs. „qualitative“ Messung von Unterschieden) implizit Werturteile nahe legt (z. B. „besser/schlechter“ vs. „anders“). Die quantitative Sichtweise bei einem Raumvorstellungstest drückt sich in der Frage nach der Anzahl richtig gelöster Items aus. Die qualitative Sichtweise führt vor allem auf die Frage unterschiedlicher Lösungsstrategien innerhalb eines Tests. In der Kombination beider Sichtweisen lassen sich quantitative Unterschiede (unterschiedliche Lösungshäufigkeiten) häufig zumindest zum Teil durch qualitative Unterschiede (andere Lösungsstrategien) erklären.

Für die Frage nach qualitativen Unterschieden bei Raumvorstellungstests ist die Arbeit von Barratt (1953) grundlegend. Resultate dieser Arbeit wurden bereits in Kap. 3.1.2 zur Fundierung der Kritik an Thurstones Trennung von *S1* und *S3* genutzt. Auf der Basis von „verbal reports“ von Versuchspersonen zu Lösungsprozessen bei Raumvorstellungstests zeigte er einen Weg zur Konstruktvalidierung entsprechender Tests auf (vgl. Quaiser-Pohl, 1998, S. 20 ff.):

> „The present investigation was designed primarily to determine if a systematic analysis of introspective data could be of value in defining psychological processes used by subjects in solving problems on paper-and-pencil ‘space’ tests. The general thesis was that one of the first steps in predicting ‘what’ is being measured by tests of ability should be an analysis of the problem-solving processes used by subjects when taking the tests, especially when those tests are used as a basis for defining factors” (Barratt, 1953, S. 17).

Beim Raumvorstellungstest *Figures*, einem „zweidimensionalen“ Test zur *mentalen Rotation*, fand Barratt (1953, S. 22) zwei unterschiedliche Herangehensweisen von Versuchpersonen:

- Beim „Part Approach“ rotieren die Versuchspersonen nur einen Teil der Figur und berücksichtigen anschließend die räumlichen Beziehungen zu anderen Teilen der Figur.
- Beim „Whole Approach“ rotieren die Versuchspersonen die gesamte Figur.

Barratt berichtet auch, dass beim Test *Figures* unter den Versuchspersonen mit den besten Testleistungen (statistisch signifikant) überproportional viele waren, die den *Part Approach* nutzten. Bei ähnlichen Analysen zum *DAT:SR*, der aus komplexeren Aufgaben zur *räumlichen Visualisierung* besteht, konnte Barrat (1952, S. 23) sogar vier verschiedene Bearbeitungsstrategien voneinander trennen.

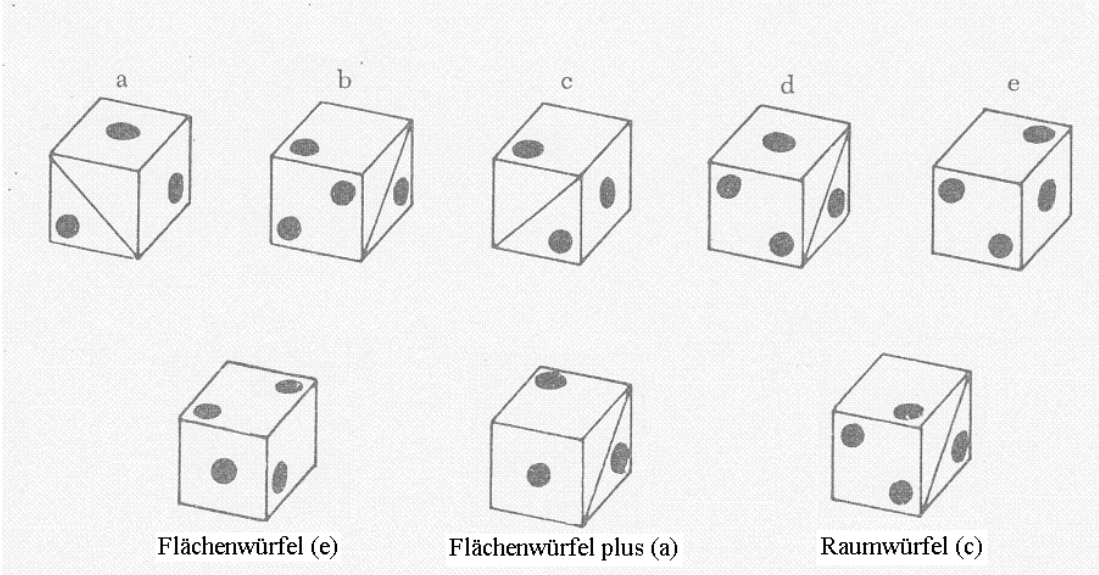
Die „strategische Perspektive“ (Quaiser-Pohl, 1998) wurde im deutschsprachigen Raum zunächst vor allem in der Auseinandersetzung mit den Würfelaufgaben (Untertest WÜ) aus dem *I-S-T* (Amthauer, 1953) bzw. *I-S-T 70* (Amthauer, 1973) produktiv eingesetzt. So konnte z. B. Putz-Osterloh (1977) auf theoriegeleitetem Wege experimentell Evidenz dafür

gewinnen, dass (a) die Aufgaben in drei Gruppen unterteilt werden können („Flächenwürfel“, „Flächenwürfel plus“ und „Raumwürfel“), zu denen es jeweils eigene effektive Lösungsstrategien gibt (vgl. Abb. 3.8, S. 100):

> „Aufgrund unserer nach logischen Gesichtspunkten vorgenommenen Aufgabenanalysen konnten wir zwei bzw. drei Arten von Aufgaben unterscheiden:
>
> 1. Einige Würfel können identifiziert werden, indem nur die Identität zwischen ihren drei sichtbaren Flächen mit den Flächen eines der vorgegebenen Würfel a bis e festgestellt wird. Deshalb bezeichnen wir diese Würfel als ‚Flächenwürfel’.
>
> 1a. Einige Würfel lassen sich identifizieren, indem neben der Flächenidentität die Relation zwischen zwei Flächen überprüft wird (eine Relation muss berücksichtigt werden, da sonst die flächengleichen Würfel a und c miteinander vertaucht würden). Diese Würfel nennen wir ‚Flächenwürfel plus’.
>
> 2. Die übrigen Würfel lassen sich nur richtig identifizieren, wenn neben der Prüfung der Identität von *zwei* sichtbaren Flächen das Erscheinen der dritten angenommen oder *vorgestellt* wird. Hierbei ist die Beachtung der Relation zwischen den Flächen und die richtige Auswahl der zwei relevanten Flächen wichtig. Diese Würfel nennen wir ‚Raumwürfel’“ (ebd., S. 254 f.)

Zu dieser Klassifikation der Würfelaufgaben passend findet die Autorin mindestens zwei Gruppen von Versuchspersonen („Flächenstrategen“ und „Raumstrategen“) gibt, die relativ stabil *eine* Lösungsstrategie wählen.

Abbildung 3.8: Würfelaufgaben aus dem Test IST:WÜ[82] (Amthauer, 1973, Form B 2, S. 18) klassifiziert nach Putz-Osterloh (1977, S. 254 f.)



[82] Beim *IST:WÜ* muss bei jedem Item für einen Würfel entschieden werden, welchem der fünf vorgegebenen Würfel (a bis e) er entspricht. Dabei können die „Itemwürfel“ gegenüber den vorgegebenen Würfeln laut Instruktion gedreht, gekippt oder gedreht und gekippt worden sein.

Auf der Basis der Arbeiten von Barratt (1953) und Putz-Osterloh (1977) untersuchten Köller et al. (1994) die Strategiefrage anhand von Testergebnissen im *IST:WÜ*. Die Autoren berichten von einer Analyse der Testleistungen nach dem „Mixed Rasch-Modell (MRM)“ (vgl. Kap. 4.2.2), die zu einer 2-Klassen-Lösung mit „Holistikern“ („Raumstrategen“ sensu Putz-Osterloh) und „Analytikern“ („Flächenstrategen“ sensu Putz-Osterloh) führt. Während „Holistiker“ bei den „Raumwürfeln“ relativ besser abschneiden, lösen „Analytiker“ die „Flächenwürfel“ relativ besser. Quantitative Unterschiede im *IST:WÜ* lassen sich also zum Teil qualitativ erklären. In einer weiteren Studie haben die Autoren mittels einer „Latent Class Analysis (LCA)“ (vgl. Kap. 4.2.2) der Antwortmuster[83] sogar sechs Klassen von Versuchspersonen identifizieren können. Dabei konnten sie in Analogie zu Putz-Osterlohs Dreiteilung der Aufgaben „Holistiker“ identifizieren und die „Analytiker“ weiter unterscheiden (in „Flächenstrategen“ und „Flächenstrategen plus“ sensu Putz-Osterloh). Darüber hinaus entdeckten sie „Strategieflexible“ und „Abbrecher“ sowie „Rater“. Auf der Basis ihrer Befunde diskutieren Köller et al. Konsequenzen aus der strategischen Perspektive für die psychometrische Raumvorstellungsforschung:

> „Kognitionspsychologische Annahmen über die Prozesse der Aufgabenbearbeitung bei Leistungstests sind nicht immer leicht mit den Voraussetzungen psychometrischer Modelle über das Testverhalten in Einklang zu bringen. Konkret sind die Anforderungen an ein Testmodell zur Messung eindimensionaler Fähigkeiten, wie es das Raschmodell ist, inkompatibel mit der Annahme interindividuell unterschiedlicher Lösungsstrategien bei der Aufgabenbearbeitung. Der Verzicht auf exakte Testmodelle bei der Auswertung solcher Tests kann ebensowenig ein Ausweg sein, wie der Verzicht auf Theorien über die jeweiligen Lösungsprozesse. […] Die Konsequenz aus diesen Ergebnissen besteht darin, daß die Interpretation eines Raumvorstellungstests zunächst in der Identifikation der präferierten Strategie der Personen bestehen muß und erst sekundär in der Quantifizierung der zugehörigen Fähigkeitsausprägung“ (Köller et al., 1994, S. 82 f.)

Hosenfeld et al. (1997) wenden die Ergebnisse und Methoden von Köller et al. (1994) schließlich auf die Frage der Geschlechterunterschiede an.[84] In ihrer Untersuchung waren männliche Versuchspersonen bei einer 2-Klassen-Lösung tendenziell (aber nicht signifikant) überproportional in der Klasse der „Holistiker“ zu finden. Dabei fiel auf, dass die männlichen Versuchspersonen in der Klasse der „Holistiker“ durchschnittlich bessere Testleistungen erzielen als die weiblichen, während es in der Klasse der „Analytiker“ keine entsprechenden Unterschiede gab. Die ermittelten Geschlechterunterschiede in der *Raumvorstellung* führen Hosenfeld et al. (1997, S. 93) dementsprechend vor allem auf die bei männlichen Probanden (durchschnittlich) effizientere Anwendung der holistischen Strategie zurück.

Die dargestellten Befunde zu *IST:WÜ*-Aufgaben aus strategischer Perspektive erschweren die Interpretation von *IST:WÜ*-Testleistungen erheblich (vgl. Putz-Osterloh, S. 262 f.). Da

[83] Die Codierung der Testleistung erfolgte dabei unter Erfassung des jeweils gewählten Distraktors.

[84] Dabei hat O. Köller als Forscher und Autor an beiden Studien und Veröffentlichungen mitgewirkt.

sowohl „zweidimensionale“ („Flächenstrategie“ sensu Putz-Osterloh) als auch „dreidimensionale“ Strategien („Raumstrategien“ sensu Putz-Osterloh) zu einzelnen richtigen Lösungen führen können, ist nicht klar, welches Konstrukt hier gemessen wird. Gittler (1990) entwickelt vor diesem Hintergrund seinen Test *3 DW*, der Rasch-homogene „dreidimensionale“ Würfelaufgaben enthält.

In jüngeren Untersuchungen nähern sich Heil & Jansen-Osmann (2008b) und Geiser et al. (2006) unter Berücksichtigung der strategischen Perspektive der „zweidimensionalen“ bzw. „dreidimensionalen“ *mentalen Rotation*. Heil & Jansen-Osmann (2008b) führen dabei Geschlechterunterschiede in der Bearbeitungszeit von Aufgaben zur „zweidimensionalen“ *mentalen Rotation* auf unterschiedliche Bearbeitungsstrategien zurück. Dabei arbeiten männliche Versuchspersonen tendenziell „holistic“, während weibliche Versuchspersonen tendenziell „analytic, piecemeal“ vorgehen. Geiser et al. (2006) kommen mittels *LCA* zu einer 5-Klassen-Lösung beim *MRT*, wobei eine Klasse „Nonrotators“ darstellt. Bezüglich des zu dieser Klasse gehörigen Anteils konnte zwar kein Unterschied zwischen den männlichen und den weiblichen Versuchspersonen gefunden werden, eine „strategische Frage“ – analog zur Kritik am *IST:WÜ* -Test – stellt sich aber auch für den *MRT*. Die Autoren identifizieren auf der Basis theoretischer Überlegungen und der *LCA* solche *MRT*-Aufgaben, die potenziell ohne mentale Rotation gelöst werden können.

3.3.4 Erklärungsansätze für interindividuelle Unterschiede

Warum haben verschiedene Menschen eine manchmal stark unterschiedlich ausgeprägte *Raumvorstellung*? Warum lassen sich insbesondere relativ konsistent und relativ stabil Gruppenunterschiede (z. B. bzgl. Geschlecht, ethnischer oder sozialer Herkunft) finden? Wie auch bei den anderen kognitiven Fähigkeiten, die in (nahezu) allen wissenschaftlichen Modellen menschlicher *Intelligenz* berücksichtigt werden, gibt es eine Reihe typischer Erklärungsansätze, von denen einige stärker den Aspekt „Anlage“, andere stärker den Aspekt „Umwelt“ und wieder andere vor allem die Interaktion von „Anlage“ und „Umwelt“ betonen. In der Wissenschaft ist mittlerweile unstrittig, dass sowohl „Anlage“ (vor allem genetische Faktoren) als auch „Umwelt“ (Sozialisationsbedingungen) jeweils einen relevanten Beitrag zur Entwicklung kognitiver Fähigkeiten leisten (vgl. Klauer, 2006a).

Befunde aus Studien, die einzelne Beiträge und auch die Interaktion von „Anlage und Umwelt“ präziser quantifizieren, werden hier nicht berichtet, da sie zu weit vom eigentlichen Thema der vorliegenden Arbeit wegführen. Die Frage, welche Faktoren grundsätzlich für die Erklärung der interindividuellen Unterschiede und der Gruppenunterschiede bei der *Raumvorstellung* infrage kommen, ist hingegen für diese Arbeit relevant: Für die pädagogische Psychologie, die Schulpädagogik und die Mathematikdidaktik ist es gleichermaßen bedeutsam, solche Faktoren zu identifizieren, die durch systematisch geplante (institutionelle) Lernangebote beeinflusst werden können. Entsprechende (potenzielle) Einflussfaktoren werden im Folgenden auf der Basis der Arbeiten von Halpern (2000), Lohaus et al.

(1999), Maier (1999b), McGee (1979), Quaiser-Pohl (1998) und D. H. Rost (1977, 2009) skizziert. Speziell für die *Raumvorstellung* liegen dabei über die prinzipielle Annahme der Wirkung hinaus keine belastbaren quantitativen Ergebnisse zur Größe des jeweiligen Einflusses vor; auch sind die Wechselwirkungen unter den (potenziellen) Einflussfaktoren teilweise unklar.

- *Genetischer Einfluss*: Wie bei anderen kognitiven Fähigkeiten wird auch für die *Raumvorstellung* ein genetischer Einfluss angenommen, der – so eine vielfach geäußerte Vermutung – über einen x-gebundenen rezessiven Erbgang funktioniert (vgl. z. B. Quaiser-Pohl, 1998, 51 ff.).
- *Neurologische Einflüsse*: Innerhalb der Theorie der Spezialisierung der beiden Gehirnhälften (Lateralisierung) wird angenommen, dass Leistungen, die mit *Raumvorstellung* assoziiert werden, in der rechten Hemisphäre verortet sind.
- *Hormonelle Einflüsse*: Verschiedenen Studien haben versucht, einen Zusammenhang zwischen der Konzentration verschiedener Hormone bzw. dem Verhältnis dieser Konzentrationen zueinander und der *Raumvorstellung* herzustellen. Besondere Beachtung finden dabei Untersuchungen, in denen die *Raumvorstellung* von Frauen zu unterschiedlichen Phasen im Menstruationszyklus gemessen wurde (vgl. z. B. Maier, 1999b, S. 214 f.).
- *Kultureller Einfluss*: Aufgrund der unterschiedlichen tradierten Lebensgewohnheit in verschiedenen Gesellschaften ließen sich – vor allem vor den Zeiten der Globalisierung – jeweils speziell ausgeprägte Profile der *Raumvorstellung* feststellen, die nicht vom direkten sozialen Umfeld abhängen. Tendenziell lässt sich feststellen, dass das Leistungsprofil mit relativen Stärken und Schwächen stärker kulturell geprägt ist, während das Niveau, auf dem dieses Profil verläuft, stärker vom direkten sozialen Umfeld abhängt (vgl. z. B. D. H. Rost, 1977. S. 24 ff).
- *Einfluss des sozialen Umfelds*: Von frühester Kindheit an interagieren Menschen mit bzw. im „Raum“. Dabei werden sowohl räumlich weiter gefasste Aspekte, wie die Mobilität, als auch räumlich enger gefasste Aspekte, wie z. B. das Spielen mit eher „technischem Spielzeug“, direkt durch das Umfeld beeinflusst. Welche Primärerfahrungen Heranwachsende dabei im Raum machen können, hängt natürlich umso stärker von ihrem sozialen Umfeld ab, je jünger sie sind.
- *Einfluss der Instruktion*: *Raumvorstellung* ist – mal eher implizit, mal auch explizit – in den curricularen Vorgaben vieler Fächer enthalten. Vermutlich leistet sogar jedes einzelne Fach einen Beitrag zur *Raumvorstellung*. Besonders deutlich werden Aspekte der *Raumvorstellung* in Mathematik, in den Naturwissenschaften, in Erdkunde, im technisch oder künstlerisch gestaltenden Unterricht (Technik, Kunst, Darstellen und Gestalten etc.), Sport, aber auch z. B. in Musik (akustische Raumwahrnehmung) oder in den Sprachen (bereitstellen von Kategorien zum Denken und Sprechen über räumliche Konstellationen).

Von den sechs genannten (potenziellen) Einflussfaktoren, ist zunächst natürlich die *Instruktion* direkt der pädagogisch-didaktischen Gestaltung zugänglich. Da Bildungsinstitutionen aber zugleich immer auch einen Teil des *sozialen Umfelds* darstellen und über fachliches Lernen hinaus viele wichtige Primärerfahrungen mitgestalten, dürfte auch in diesem Bereich eingewirkt werden können. Der kulturelle Einfluss kann in einer offenen Gesellschaft wohl kaum unter einer sehr speziellen Perspektive, wie der Förderung der *Raumvorstellung*, zielgerichtet verändert werden.

3.3.5 Zusammenhang mit Mathematikleistung

In nahezu allen vorliegenden Überblicksarbeiten zum Thema *Raumvorstellung* wird auf den vermuteten Zusammenhang mit *Mathematikleistung* hingewiesen (vgl. z. B. Maier, 1999b, 128 ff.; McGee, 1979, S. 24 ff.; Treumann, 1974, 238 ff.). Dabei wurde vor allem in frühen Studien die Mathematiknote als Indikator für *Mathematikleistung* herangezogen, was aufgrund der Komplexität des im institutionellen Rahmen stattfindenden Bewertungsvorgangs sicherlich zu größeren Messfehlern und ggf. auch zu systematischer Verzerrung führt. Dennoch werden fast durchgängig Zusammenhänge zwischen *Raumvorstellung* und *Mathematikleistung* berichtet. Bei den entsprechenden Befunden sollte allerdings nicht nur der Indikator für *Mathematikleistung*, sondern auch der verwendete Raumvorstellungstest kritisch betrachtet werden:

> „Aufgaben, die wie die von Reed (1974) in psychologischen und mathematikdidaktischen Untersuchungen zum Räumlichen Vorstellungsvermögen eingesetzt werden, sind zumeist räumlich-geometrischer Art. Dass hier eine Aufmerksamkeit für die geometrischen Details der in diesen Aufgaben verwendeten Materialien eine positive Rolle spielen kann, liegt also nah“ (Pinkernell, 2003, S. 74).

Ein Zusammenhang von guten Mathematiknoten und guter *Raumvorstellung* wird allerdings auch für die Tests *3 DW* (Gittler, 1990, S. 36) und *BST* (Birkel et al., 2002, S. 49 f.) berichtet, die für Raumvorstellungstests keine außergewöhnliche Nähe zum Mathematikunterricht aufweisen.

In jüngeren Studien wird die *Mathematikleistung* anstelle von Noten auch häufiger durch Tests ermittelt. Lehmann et al. (2002) vergleichen z. B. Schülerinnen und Schüler an mathematisch und sprachlich orientierten Gymnasien anhand einer Reihe von Tests, unter denen neben einem Mathematiktest auch der *MRT* und der *3 DW* sind. Faktorenanalytisch konnten sie aufgrund des engen Zusammenhangs von *Mathematikleistung* und *Raumvorstellung* einen mathematisch-räumlichen und einen fremdsprachlichen Faktor identifizieren. Besonders interessant ist dabei, dass der Zusammenhang von *Mathematikleistung* mit dem *MRT* deutlich enger ist als der mit dem *3 DW*. Die größten Leistungsunterschiede zwischen den Schülerinnen und Schüler der spezialisierten Gymnasien gab es dabei nicht etwa im Mathematiktest, sondern im *MRT*. Der Leistungsvorsprung beim *MRT* betrug trotz vergleichbarer Leistungen in einem allgemeinen Intelligenztest ca. 1,5 Standardabweichungen (zugunsten der mathematisch orientierten Gymnasien).

Die unterschiedlichen Befunde für *MRT* und *3 DW* deuten darauf hin, dass der Zusammenhang von *Raumvorstellung* und *Mathematikleistung* von den jeweils betrachteten Komponenten abhängt. „Dreidimensionale" *mentale Rotation* scheint dabei besonders eng mit allgemeiner *Mathematikleistung* zusammenzuhängen. *Räumliche Visualisierung* hingegen kaum oder gar nicht[85]:

> „Für die Diagnostik und Förderung mathematischer Begabungen sind Raumvorstellungsleistungen unterschiedlich gute Indikatoren. So weist das mentale Rotieren (MRT-Leistungen) mehr auf das Spezifische in der mathematischen Leistungsfähigkeit hin, während das räumliche Visualisieren (3DW-Leistungen) in einem engeren Zusammenhang zur allgemeinen Intelligenz steht und damit im Rahmen einer allgemeinen Intelligenzdiagnostik erfasst werden kann" (Lehmann & Jüling, 2002, S. 42).

> „Boys had significantly higher mean mathematics scores than girls. Significant sex differences favouring boys were found in the subsamples of most difficult tasks, but not in the subsamples of easiest tasks. No significant sex difference in spatial visualization was found" (Manger & Eikeland, 1998, S. 17).

Zusammenhänge zwischen (Komponenten von) *Raumvorstellung* und (Komponenten von) *Mathematikleistung* lassen sich dabei bereits in der frühen Schulzeit nachweisen. Guay & McDaniel (1977) haben Schülerinnen und Schüler aus dem 2. bis 7. Jahrgang vier unterschiedliche Raumvorstellungstests bearbeiten lassen. Die Schülerinnen und Schüler wurden innerhalb jedes Jahrgangs aufgrund ihrer Testleistungen zum Basiswissen und Basiskönnen in Mathematik in „High Achiever" und „Low Achiever" eingeteilt. „High Achiever" erzielten in allen sechs untersuchten Jahrgängen jeweils in allen vier Raumvorstellungstests signifikant bessere Ergebnisse als „Low Achiever". Dies drückt sich auch in konsistenten Korrelationskoeffizienten zwischen den Testergebnissen aus.

In einer jüngeren Studie mit Grundschülerinnen und Grundschülern kann Grüßing (2005) ebenfalls bereits für diese Altersgruppe einen engen Zusammenhang von *Raumvorstellung* und *Mathematikleistung* nachweisen. Ihre Ergebnisse deuten allerdings – anders als die von Guay & McDaniel (1977) und ähnlich wie die von Lehmann & Jüling (2002) – darauf hin, dass dieser Zusammenhang vor allem für anspruchsvollere Mathematikaufgaben besonders eng ist:

> „Für alle drei Bereiche des Tests zu räumlichen Kompetenzen kann ein hochsignifikanter Zusammenhang zur Mathematikleistung nachgewiesen werden. Für den Gesamttest ergibt sich eine Korrelationskoeffizient von r = .52 (Produkt-Moment-Korrelation). […] Tendenziell fällt dieser Zusammenhang für komplexere Anwendungsaufgaben stärker aus als für Aufgaben, die mit Faktenwissen und Routineprozeduren lösbar sind" (Grüßing, 2005, S. 45).

Bei Untersuchungen zum Zusammenhang von *Raumvorstellung* und *Mathematikleistung* wird überwiegend davon ausgegangen, dass *Raumvorstellung* ein Prädiktor für *Mathematikleistung* ist, der *Mathematikleistung* also gewissermaßen als kognitive Grundfähigkeit

[85] Dies wird auch die Befunde von Klieme (1986) unterstützt (vgl. Kap. 3.3.6).

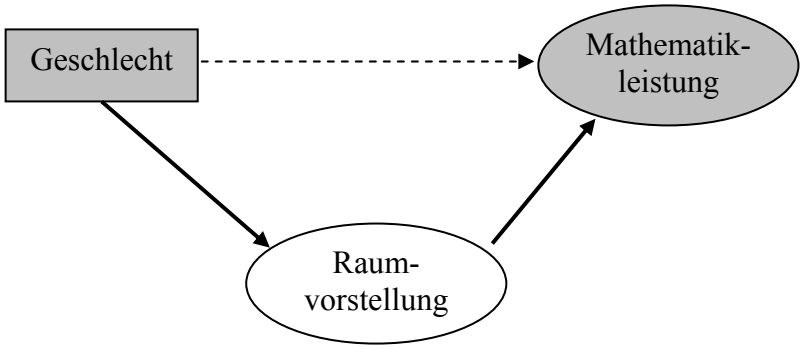
vorausgeht. Diese vermutete Wirkungsrichtung wird auch durch Forschung zur Dyskalkulie gestützt, bei der davon ausgegangen wird, dass eine (zu) schwach ausgeprägte *Raumvorstellung* zu massiven Schwierigkeiten in der Arithmetik führen kann (vgl. Lorenz, 1991). Dabei wird betont, wie wichtig das Erkennen und Nutzen räumlicher Beziehungen zwischen Zahlen, z. B. am Zahlenstrahl oder in der Hundertertafel, ist. Diese Überlegungen lassen sich auch auf das Erkennen und Nutzen von Strukturen und Beziehungen in abstrakten Termen, Gleichungen und Gleichungssystemen oder auf die anschauliche Erfassung des Ko-Variationsverhaltens einer Funktion an ihrem Graphen beziehen.

Umgekehrt lässt sich auch die Vermutung begründen, dass eine gut ausgeprägte *Mathematikleistung* positiv auf die *Raumvorstellung* wirkt. Selbst wenn die Raumvorstellungstests keine auffällige Nähe zum Geometrieunterricht haben, sind (in der Regel zweidimensionale) visuelle Darstellungen von Figuren und Körpern konstitutiver Bestandteil der Aufgaben. Das durch den Mathematikunterricht geschulte Abstraktionsvermögen, aber auch die primär dort erworbene Fähigkeit, solche visuellen Darstellungen zu interpretieren, stellt vermutlich eine Basis für gute Raumvorstellungsleistungen dar. Dies gilt umso mehr, wenn geometrische Begriffe helfen, die visuellen Darstellungen zu strukturieren und begrifflich zu erfassen. Insbesondere die begriffliche Erfassung ist eine Voraussetzung für die Lösungsstrategie „Verbalisieren“ (vgl. Gittler, 1990, S. 18 f.). Wird die Raumvorstellungsleistung komplexer, z. B. bei Aufgaben zu *räumlicher Visualisierung*, dürften sich auch die Beiträge des Mathematikunterrichts zu einer übergreifenden Problemlösekompetenz und zum schlussfolgernden Denken positiv auswirken.

3.3.6 Die „Spatial Mediation Hypothesis“

Betrachtet man die Befunde zu Geschlechterunterschieden in der *Mathematikleistung* (Kap. 2.3.2), zu Geschlechterunterschieden in der *Raumvorstellung* (Kap. 3.3.2) und zum Zusammenhang von *Raumvorstellung* und *Mathematikleistung* (Kap. 3.3.5), so liegt es nahe, diese theoretisch wie empirisch zusammenzuführen. Mit der „Spatial Mediation Hypothesis“ (Burnett et al., 1979) wird angenommen, dass Geschlechterunterschiede in der *Mathematikleistung* im Wesentlichen durch entsprechende Unterschiede in der *Raumvorstellung* erklärt werden können. McGee (1979, S. 4) geht davon aus, dass diese Mediatorfunktion auch für andere Bereiche kognitiver Leistungen gilt: „Furthermore, sex differences in various aspects of perceptual-cognitive functioning (for example, mathematics, field dependence) are interpreted as a secondary consequence of differences with respect to spatial visualization and spatial orientation abilities.“

Abbildung 3.9: „Spatial Mediation Hypothesis“



Die *Spatial Mediation Hypothesis* setzt dabei implizit voraus, dass *Raumvorstellung* auf *Mathematikleistung* wirkt und nicht umgekehrt, da ansonsten vermutet würde, dass Geschlechterunterschiede in der *Mathematikleistung* entsprechende Unterschiede in der *Raumvorstellung* erklären können. Wie in Kap. 3.3.5 dargestellt wurde, ist die Frage der Wirkungsrichtig weder trivial noch theoretisch oder empirisch hinreichend geklärt. In der Literatur findet man aber überwiegend (und in der Regel nicht weiter hinterfragt) die Annahme der Wirkungsrichtung von *Raumvorstellung* auf *Mathematikleistung*.

Berücksichtigt man, dass für verschiedene Komponenten der *Raumvorstellung* (a) der Zusammenhang mit *Mathematikleistung* unterschiedlich groß ist und (b) die Geschlechterunterschiede unterschiedlich groß sind, so liegt es nahe die *Spatial Mediation Hypothesis* entsprechend auszudifferenzieren. Mit Blick auf die in Kap. 3.3.5 berücksichtigten Resultate von Grüßing (2005) und Lehmann & Jüling (2002) muss dabei vermutlich auch die *Mathematikleistung* in verschiedenen Komponenten betrachtet werden.

Aufgrund ihrer Bedeutung für die vorliegende Arbeit wird im Folgenden die Studie von Klieme (1986) ausführlicher dargestellt. Entsprechend der obigen Überlegungen betrachtet Klieme sowohl die *Raumvorstellung* als auch die *Mathematikleistung* jeweils als differenzierte Konstrukte mit unterschiedlichen Komponenten. Die Ergebnisse dieser Studie werden in der deutschsprachigen Literatur zu Geschlechterunterschieden in der *Mathematikleistung* bzw. zum Zusammenhang von *Raumvorstellung* und *Mathematikleistung* regelmäßig zitiert, wobei zurzeit keine widersprechenden Befunde vorzuliegen scheinen.

Klieme stützt seine eigene empirische Studie auf Literaturstudien zu Geschlechterunterschieden in der *Mathematikleistung* und zur *Raumvorstellung*. Auf der Basis der Forschungsliteratur stellt er bezüglich der *Mathematikleistung* fest, „daß Mädchen bzw. Frauen eher algorithmisch/verbal, Jungen bzw. Männer eher vorstellungsmäßig vorgehen und dementsprechend geschlechtsspezifische Stärken wie Schwächen besitzen“ (Klieme, 1986, S. 134). Auf diesem Befund und der Berücksichtigung unterschiedlicher Befunde zur *Raumvorstellung* folgt ein Untersuchungsdesign mit ausdifferenzierten Konstrukten:

> „Es geht uns also nicht um eine globale Falsifikation oder Bestätigung der ‚spatial mediation hypothesis', sondern um die Differenzierung dieser Hypothese für verschiedene Anforderungsbereiche, von der wir insbesondere neue Hinweise auf die Funktionen bildlichen Denkens beim mathematischen Problemlösen erwarten" (ebd., S. 136 f.).

In die empirische Analyse gehen Testergebnisse größerer Gruppen von Versuchspersonen am Übergang von der Sekundarstufe II zum Hochschulstudium ein. Im Bereich der *Mathematikleistung* wurden unterschiedliche „Testbatterien" verwendet, die keine „reinen Rechenfertigkeiten", sondern „mathematische Problemlösefähigkeit" erfassen sollten (ebd., S. 137), wobei einige Tests überwiegend naturwissenschaftliche Aufgabenkontexte enthalten. Im Bereich der *Raumvorstellung* wurden ebenfalls unterschiedliche Tests eingesetzt, die insgesamt eine hinreichende Breite sicherstellen. Die empirische Umsetzung der *Spatial Mediation Hypothesis* erfolgt durch eine Kovarianzanalyse[86], bei der *Geschlecht* als unabhängige Variable, *Mathematikleistung* als abhängige Variable und *Raumvorstellung* als Kovariate berücksichtigt wird (ebd., S. 138). Für die verschiedenen verwendeten Operationalisierungen von *Raumvorstellung* und *Mathematikleistung* wird dann nicht untersucht, „ob die ‚spatial mediation hypothesis' im Einzelfall zutrifft, sondern wie groß der in ihrem Sinne erklärbare Varianzanteil jeweils ausfällt" (ebd., S. 140; i. O. m. Herv.).

Ein wichtiges Resultat von Klieme betrifft zunächst den Zusammenhang von *Mathematikleistung* und verschiedenen Komponenten von *Raumvorstellung*. Er stellt dar, dass der Test „Schlauchfiguren" (Stumpf & Fay, 1983), der hohe Anteile dreidimensionaler *mentaler Rotation* erfordert, erheblich enger mit *Mathematikleistung* zusammenhängt als Tests, die eher im Bereich *räumliche Visualisierung* verortet werden können (Klieme, 1986, S. 143). Dies ist auch für die *Spatial Mediation Hypothesis* von großer Bedeutung: „Grundsätzlich läßt zudem die Regression auf die Leistung im Subtest ‚Schlauchfiguren' die geringste zusätzliche Varianz für solche geschlechtsspezifischen Einflüsse übrig, die nicht auf unterschiedliche Raumvorstellungs-Fähigkeiten zurückführbar sind" (ebd., S. 145).

Anstelle einer generellen Beurteilung der *Spatial Mediation Hypothesis* gibt Klieme in seiner Zusammenfassung eine differenzierte Quantifizierung für die Stärke der Mediatorfunktion der *Raumvorstellung* für Geschlechterunterscheide in der *Mathematikleistung* an. Die Quantifizierung gilt unter Beachtung des obigen Befundes, dass vor allem der Test *Schlauchfiguren* diese Mediatorfunktion wahrnimmt:

> „Die deutlichen Unterschiede zwischen den Geschlechtern bei der Analyse graphischen Materials und bei der Bearbeitung von ‚eingekleideten' arithmetischen Aufgaben, aber auch die geringeren geschlechtsabhängigen Effekte bei algebraischen Anforderungen können unter Umständen zu maximal 60 Prozent auf differentielle Raumvorstellungs-Leistungen zurückgeführt werden. Bis auf 15 bis 30 Prozent geht dieser Anteil zurück, wenn algorithmisierte Operationen oder naturwissenschaftliches Verständnis bei der Lösung eine wichtige Rolle spielen" (ebd., S. 148).

[86] vgl. Kap. 4.2.2.

Auch wenn dieser Befund regelmäßig und ohne Widerspruch zitiert wird, bedeutet dies nicht, dass Einigkeit in der Interpretation des Befundes besteht. So schreibt Quaiser-Pohl (1998, S. 40; Herv. i. O.):

> „In letzter Zeit wurde die *spatial mediation hypothesis* von einigen Autoren in Frage gestellt. So konnten bei Klieme (1985) in mehreren Tests zur mathematischen Problemlösefähigkeit beobachtete geschlechtsspezifische Leistungsunterschiede in keinem Fall allein durch die ebenfalls psychometrisch erfassten Unterschiede in der *spatial ability* erklärt werden."

Dieser Interpretation kann man folgen, man muss es aber sicherlich nicht zwingend. Klieme (1986) konnte für einige Komponenten der *Mathematikleistung* große Anteile der „geschlechtsabhängigen Effekte" (S. 148) durch einen Raumvorstellungstest statistisch erklären. Die gelingt nur für einige Komponenten der *Mathematikleistung* und auch unterschiedlich gut durch verschiedene Raumvorstellungstests. Eine vollständige Erklärung differenzieller Befunde durch eine Mediatorvariable dürfte in der pädagogischen Literatur ohnehin nicht anzutreffen sein. Die Interpretation der Befunde hängt also auch davon ab, ob man geneigt ist „das Glas eher als halbleer oder eher als halbvoll zu betrachten". Der besondere Beitrag von Kliemes Arbeit zur Erforschung der Geschlechterunterschiede in der *Mathematikleistung* dürfte vor allem in der Verwendung differenzierter Konstrukte und in seinen differenzierten Befunden liegen. Eine derart relativierte *Spatial Mediation Hypothesis* wird auch durch Befunde von D. H. Rost (2009, S. 177) unterstützt, der dazu aktuelle Forschungsarbeiten ausgewertet hat und die vermittelnde Funktion der *Raumvorstellung* für Geschlechterunterschiede in der *Mathematikleistung* schon für das Grundschulalter annimmt.

Vom aktuellen Stand der Forschung und Entwicklung im Bereich der Mathematikdidaktik bzw. der empirischen Bildungsforschung ergeben sich aber auch offene Fragen:

- Da die Versuchspersonen in Kliemes Studie sich gerade in der Übergangsphase von der Schule auf die Hochschule befunden haben, liegt den Befunden keine Stichprobe zugrunde, die mit der (nicht ausgelesenen) Gruppe der 15-Jährigen bzw. der Neuntklässler vergleichbar ist. Inwieweit sich die Befunde bei einer derartigen Stichprobe verändern, lässt sich nicht theoretisch abschätzen.
- Eine weitere offene Frage betrifft Komponenten von *Mathematikleistung* wie Rechenfertigkeiten bzw. Basiswissen und -können, die Klieme aufgrund seines Ansatzes („mathematische Problemlösefähigkeit") nicht berücksichtigt hat. Die in Kap. 3.3.5 dargestellten Befunde von Guay & McDaniel (1977) vs. Grüßing (2005) und Lehmann & Jüling (2002) ergeben hier ein widersprüchliches Bild zum Zusammenhang von *Raumvorstellung* und entsprechender *Mathematikleistung*.
- Aktuelle Schulleistungsstudien und auch Vergleichsarbeiten wie die *LSE 9* verwenden Mathematiktests, die – passend zur aktuellen Unterrichtsentwicklung – „grundbildungsorientiert" bzw. „kompetenzorientiert" sind (vgl. z. B. Blum et al., 2006; Bruder et al., 2008). Ob sich durch diese andere Art der Tests bezüglich der hier betrachteten Zu-

sammenhänge relevante Verschiebungen gegenüber den Tests ergeben, die Klieme (1986) eingesetzt hat, lässt sich ebenfalls nicht theoretisch einschätzen.

Im empirischen Teil der vorliegenden Arbeit soll ein Beitrag zur Beantwortung dieser offenen Fragen geleistet werden.

3.3.7 Möglichkeiten zur Förderung der Raumvorstellung

Die Frage, wie die *Raumvorstellung* gefördert werden kann, ist aufgrund der „Relevanz der Raumvorstellung“ (D. H. Rost, 1977, Kap. 6; Maier, 1999b) von großer Bedeutung. Für die Mathematikdidaktik gilt dies besonders aufgrund der oben diskutierten *Spatial Mediation Hypothesis*. Die Befundlage zu dieser Frage ist dabei auf den ersten Blick heterogen bis widersprüchlich. Dies liegt vor allem daran, dass die Förderung der *Raumvorstellung* sehr unterschiedlich gestaltet wird.

In vielen psychologischen Studien besteht die entsprechende Intervention aus einem zeitlich stark begrenzten Training, das in der Regel eng auf eine Komponente oder sogar auf einen Raumvorstellungstest abgestimmt ist und z. B. die gleichen Objekte wie der Test verwendet. Typische Ergebnisse solcher Studien zur *mentalen Rotation* fasst Wiedenbauer (2006, S. 50) zusammen:

> „Zusammenfassend lässt sich feststellen, dass die mentale Rotation trainierbar ist. Die Verbesserung der mentalen Rotationsfähigkeit durch wiederholte Ausführung mentaler Rotationen scheint dabei jedoch auf gedächtnisbasierten Prozessen zu beruhen: Sowohl Erwachsene als auch Kinder profitierten zwar von der Übung, konnten den Übungsgewinn jedoch nicht auf neue, im Training nicht präsentierte Objekte transferieren.“

Wiedenbauers Zusammenfassung kann so interpretiert werden, dass bei einer derart eng zugeschnittenen und relativ kurzen Intervention zwar Trainingseffekte nachgewiesen werden können, diese aber so spezifisch sind, dass nicht von einer allgemeinen Förderung der *Raumvorstellung* gesprochen werden kann (vgl. auch Wiedenbauer, 2006, S. 7). Im Übrigen zeigen nahezu alle Studien mit einem Vortest-Nachtest-Design, dass jeweils alle untersuchten Gruppen, also auch Kontrollgruppen ohne Intervention, im Nachtest besser abschneiden als im Vortest. Der Vortest scheint im Bereich der *Raumvorstellung* schon ein Training für den Nachtest darzustellen. Von einer allgemeinen Förderung der *Raumvorstellung* dürfte dabei nicht auszugehen sein.

Auch stärker mathematikdidaktisch orientierte Studien können zu dem Ergebnis führen, dass inhaltlich enge und zeitlich kurze Fördermaßnahmen nicht zwingend zu bemerkenswerten Verbesserungen der *Raumvorstellung* führen. Dies zeigen z. B. die forschungsmethodisch überzeugenden „Oldenburger Studien“ von Hartmann & Reiss (2000) sowie Hellmich & Hartmann (2002).

In der Studie von Hartmann & Reiss (2000) besteht die Förderung der Experimentalgruppe in einer sechs Unterrichtsstunden lang dauernden Arbeit mit dem Computerprogramm

„Quaderpuzzle“, während die Kontrollgruppe „in dieser Zeit an regulärem, nicht geometriebezogenem Mathematikunterricht“ teilnahm (ebd., S. 87 f.). Anschließend erhielten beide (allgemein eher leistungsstarke) Gruppen Geometrieunterricht. Die vier verschiedenen eingesetzten Raumvorstellungstests zeigten überwiegend keine Gruppenunterschiede. Die Experimentalgruppe erzielte im Nachtest lediglich in einem der vier Raumvorstellungstest („Cube Comparison“) signifikant bessere Ergebnisse als die Kontrollgruppe. Dies wird mit der großen Nähe des Tests zum Trainingsprogramm begründet (ebd., S. 90). Darüber hinaus erzielten die Mädchen der Experimentalgruppe in einem zweiten Raumvorstellungstest signifikant größere Zuwächse gegenüber dem Vortest als die Jungen in der Experimentalgruppe und als die Mädchen in der Kontrollgruppe (ebd., S. 89 ff.). Mittelfristige Auswirkungen der Intervention sind nicht untersucht worden. Bei der Interpretation der Ergebnisse sollte berücksichtigt werden, dass zu zwei berücksichtigten Komponenten der *Raumvorstellung* (*mentale Rotation* und *räumliche Visualisierung*) je zwei Tests verwendet wurden. Da die signifikanten Effekte in ihrer Art aber jeweils auf einen Test beschränkt sind, lassen sich die Befunde nicht zwingend als „Förderung von Raumvorstellung“ interpretieren.

In einer etwas jüngeren Studie von Hellmich & Hartmann (2002) werden Förderschülerinnen und Förderschüler des Förderschwerpunktes „Lernen“ mit einem vergleichbaren Design wie bei Hartmann & Reiss (2000) untersucht; die Kontrollgruppe arbeitete in dieser Studie allerdings lediglich vier Unterrichtsstunden mit dem Computerprogramm „Quaderpuzzle“. Die Ergebnisse fassen die Autoren wie folgt zusammen:

> „Die Untersuchung ist auf ein Training räumlicher Kompetenzen sowie einen Transfer auf das Verständnis in einer Unterrichtseinheit zu ebener und räumlicher Geometrie ausgerichtet. Obwohl die Art des verwendeten Trainingsmaterials wesentliche Komponenten beinhaltet, denen üblich eine Förderung räumlicher Kompetenzen zugeschrieben werden, kann eine Verbesserung in diesem Bereich nicht festgestellt werden. Die Ergebnisse der Untersuchung können zum einen als Hinweis darauf verstanden werden, dass sich räumliche Fähigkeiten nicht durch ein relativ kurzzeitiges Training beeinflussen lassen, zum anderen ist es ein Indiz dafür, dass Mechanismen des Erwerbs oder Trainings räumlicher Fähigkeiten nicht so klar sind, wie sie vielfach erscheinen“ (Hellmich & Hartmann, 2002, S. 60).

Die Förderung von *Raumvorstellung* – so kann man aufgrund der Ergebnisse vermuten – muss also inhaltlich breit genug und über einen hinreichend langen Zeitraum angelegt sein, wenn sie nicht testspezifische Effekte erzielen, sondern tatsächlich kognitive Fähigkeiten fördern soll. Dies ist auch insofern plausibel, als Menschen spätestens von Geburt an Raumerfahrungen machen und *Raumvorstellung* entwickeln. Möchte man dann nach zehn Lebensjahren oder noch später die entsprechenden Fähigkeiten messbar und relevant verbessern, müssen die Versuchspersonen vermutlich wiederholt substanzielle (mentale) Tätigkeiten ausüben. In diese Richtung weisen auch sozialisationstheoretische Argumente, die allerdings für eine prinzipielle Trainierbarkeit der *Raumvorstellung* sprechen:[87]

[87] Sozialisationseffekte dürften generell eher mittel- und langfristig wirken.

> „Auf der anderen Seite belegen diese in einer Vielzahl aufgewiesenen Differenzen die starke Abhängigkeit der qualitativen und quantitativen Raumwahrnehmungs- und Raumvorstellungsleistungen von den jeweiligen sozialisationsspezifischen Einflüssen und lassen durch die aufgewiesene breite Variation die Hypothese von einer Trainierbarkeit der Raumvorstellung wahrscheinlich erscheinen.“ (D. H. Rost, 1977, S. 26)

In einer eigenen Interventionsstudie untersucht D. H. Rost (1977) in einem aufwändigen und forschungsmethodisch vorbildlichen Design die Frage, ob die *Raumvorstellung* von Grundschülerinnen und Grundschülern (3. Klasse) durch geeignete Spiele gefördert werden kann. Zur Förderung der *Raumvorstellung* wurden vier Spiele ausgewählt, die augenscheinlich erfordern, sich (gedanklich) im Raum zu bewegen, die ohne spezielle Lernvoraussetzung und ohne Ermüdungserscheinungen gespielt werden können und die den in der Untersuchung verwendeten Raumvorstellungstests nicht zu ähnlich sind (ebd., S. 145 ff.). Die Fördermaßnahme ist auf sechs Wochen mit drei Unterrichtsstunden pro Woche angelegt. In diesen 18 Unterrichtsstunden durften die Kinder der Experimentalgruppe mit den zur Verfügung gestellten Spielen spielen, wobei darauf geachtet wurde, dass alle Kinder der Experimentalgruppe alle vier Spiele genügend oft spielen (ebd., S. 131 ff). Die Auswertung von Vor- und Nachtests ergaben, dass der gewünschte Fördereffekt in relevantem Maße eingetreten ist:

> „Unter Beachtung der forschungsmethodischen Probleme des Experimentierens im pädagogischen Feld läßt sich feststellen, daß in unserer Stichprobe unter den geschilderten Versuchsbedingungen und den beschriebenen Trainingssituationen eine signifikante und bedeutsame Erhöhung der kombinierten Testleistungen der Experimentalgruppe im Vergleich zu den beiden Kontrollgruppen beobachtet werden konnte“ D. H. Rost (1977, S. 184).

Dieser Befund ist in zweierlei Hinsicht bemerkenswert: Erstens ist die Studie forschungsmethodisch sehr aufwändig und mit Blick auf mögliche Artefakte äußerst reflektiert geplant worden (und in dieser Hinsicht wirklich vorbildlich für vergleichbare Interventionsstudien im pädagogischen Feld). Zweitens sind die Fördermaterialien, also die vier ausgewählten Spiele, den verwendeten Raumvorstellungstest – anders als bei den zuvor dargestellten Studien – nicht zu ähnlich. Daher kann die beobachtete signifikante und bedeutsame Verbesserung der Testleistung der Experimentalgruppe im Vergleich zu den Kontrollgruppen tatsächlich als Verbesserung der *Raumvorstellung* interpretiert werden. Hier hat offensichtlich ein mentaler Transfer der spielbezogenen Trainingseffekte auf anders geartete Raumvorstellungstests stattgefunden.

Empirisch weniger ausgereift, aber bezüglich der inhaltlichen Ausgestaltung äußerst interessant sind die Untersuchungen von Meißner (2006) und Leopold (2002). Meißner berichtet vom Projekt „Wir bauen ein Dorf“, das in sieben bis neun Unterrichtsstunden während des Geometrieunterrichts an Grundschulen durchgeführt wurde und speziell der Förderung der *Raumvorstellung* dienen soll (vgl. auch Pinkernell, 2003, S. 8 ff.). Die Schülerinnen und Schüler arbeiten dabei intensiv und handlungsorientiert mit vielen unterschiedlichen Körpern und Körpernetzen und basteln zum Abschluss der Unterrichtsreihe ein Dorf (vgl. Meißner, 2006, S. 28 ff.). Die Ergebnisse der quantitativen Auswertung der Vor- und

Nachtests zur *Raumvorstellung* bewertet Meißner positiv im Sinne des intendierten Förder-effekts (S. 49). Diese Auswertung weist dabei allerdings forschungsmethodische Mängel auf, die eine Belastbarkeit der Befunde infrage stellen. Leopold (2002) begleitet eigene ingenieurwissenschaftliche Veranstaltungen in Darstellender Geometrie mit Vor- und Nachtests. Die theoretisch erwartbaren positiven Effekte der Lehrveranstaltung auf die *Raumvorstellung* der Studierenden kann mit den eingesetzten Tests auch empirisch gefunden werden. Über eine Verbesserung aller Studierenden hinaus verringern sich zusätzlich die Geschlechterunterschiede vom Vor- zum Nachtest.

Die Befunde von D. H. Rost (1977), Meißner (2006) und Leopold (2002) deuten klar darauf hin, dass inhaltlich breit angelegte und zeitlich nicht zu kurze Fördermaßnahmen die *Raumvorstellung* substanziell fördern können – und das deutlich über Vortest-Nachtest-Effekte und Kopplung an spezielle Tests hinaus. Insbesondere aus mathematikdidaktischer Sicht ist die Frage, wie eine solche Förderung möglichst effektiv gestaltet werden kann, von besonderem Interesse. Für die vorliegende Arbeit genügen zunächst einige vor allem entwicklungs- und lernpsychologisch orientierte Ansätze.[88]

Zur Vorbereitung seiner Interventionsstudie fasst D. H. Rost (1997) unter anderem relevante Befunde zur Entwicklung der *Raumvorstellung* zusammen:

> „Daß es angeborene Tendenzen der Raumauffassung gibt, ist unumstritten. Unabhängig von spezifischen psychologischen ‚Schulen' und ‚Richtungen' herrscht bei den meisten Forschern darüber Einigkeit vor, daß Raumbeziehungen zunächst erst erfahren werden müssen, um dann auch wahrgenommen und vor allem vorstellungsmäßig erfasst werden zu können […] Die psychische Konstruktion des Raumes, ein langzeitiger Vorgang, vorbereitet von ersten Bewegungen des Säuglings … nimmt ihren Ausgang in Lokomotionsversuchen des Babies. […] Es lernt so, aus der Erfahrung des Raumes den Raum psychisch zu konstruieren." (D. H. Rost, 1977, S. 38)

Auch Piaget & Inhelder (1971, S. 21 ff.) unterscheiden wesentlich zwischen Wahrnehmung und Vorstellung: „Die Wahrnehmungs- oder sensomotorischen Strukturen bilden jedoch meist den Ausgangspunkt und dann die Substruktur der gesamten Raumkonstruktion durch die Vorstellung" (S. 23). D. H. Rost (1977, S. 101 ff.) berichtet von mehreren Interventionsstudien, die statistisch nachweisbare Effekte von Fördermaßnahmen gefunden haben. Dabei handelt es sich vor allem um Fördermaßnahmen, die mit Primärerfahrungen im Raum bzw. mit räumlichen Objekten arbeiten und dabei insbesondere auch Verbalisierungen anregen.

Forschungsbedarf gibt es noch in der Frage, welche Erfahrungen den Ausgangspunkt für die Verbesserung der *Raumvorstellung* darstellen können. So berichten Lohaus et al. (1999) von erfolgreichen Fördermaßnahmen, bei denen Computerspiele eine zentrale Rolle spielen. Hier sind also nicht mehr Primärerfahrungen im dreidimensionalen Raum der Aus-

[88] Weitergehende mathematikdidaktische Überlegungen werden in Kap. 7.3 angestellt.

gangspunkt, sondern entsprechende Projektionen auf dem Bildschirm. Vermutlich dürfte es aber bezüglich der Förderung der *Raumvorstellung* einen Unterschied darstellen, ob räumliche Spiele im Raum oder am Bildschirm gespielt werden. Allerdings scheinen bisher keine Studien vorzuliegen, die diesen potenziellen Unterschied untersuchen.

3.4 Zusammenfassung und Diskussion: Unterschiedliche Konstrukte von Raumvorstellung

Die Darstellung der Erforschung der inneren Struktur von *Raumvorstellung* in Kap. 3.1.2 hat offenbart, dass dieser Bereich kognitiver Leistungen zwar seit ungefähr 100 Jahren intensiv theoretisch und empirisch untersucht wird, es aber bis heute (a) keine einheitliche Definition von *Raumvorstellung* gibt und (b) die innere Struktur der *Raumvorstellung* noch nicht abschließend geklärt ist. Der dominante psychometrische Versuch, diese innere Struktur faktorenanalytisch zu klären, scheint vorläufig nicht weiter zu führen; aktuell deutet vieles darauf hin, dass die auf diesem Weg gewonnenen Ergebnisse zu sehr vom konkret berücksichtigten Testmaterial und der konkreten Personenstichprobe abhängen (vgl. Gittler, 1990, S. 11; Zitat auf S. 80 der vorliegenden Arbeit). Dies sollte aber nicht darüber hinweg täuschen, dass mindestens im Bereich der *Small-Scale Fähigkeiten*, die über *Paper and Pencil Tests* erfasst werden können, zwischen den meisten beteiligten Forscherinnen und Forschern Einigkeit darüber erzielt werden kann, ob ein konkreter Test (Anteile von) *Raumvorstellung* erfasst oder nicht – lediglich die etwaige Bezeichnung der fraglichen Anteile kann unterschiedlich aussehen.

Ein Konsens besteht darüber, dass *Raumvorstellung* aus theoretischer wie empirischer Sicht nicht aus einem breiten Faktor besteht, sondern eine innere Struktur aufweist. Wie die Befunde in den Kapiteln 3.3.1 bis 3.3.6 zeigen, hängen Resultate zur *Raumvorstellung* in der Regel von der jeweils betrachten Komponente ab. Unter den aktuell konkurrierenden Modellen von Raumvorstellungskomponenten gibt es keine, die allgemein zu bevorzugen ist, sodass für jede Arbeit einzeln und mit Blick auf die Fragestellungen der Arbeit entschieden werden muss, von welchem Konstrukt ausgegangen wird:

> „Auch wenn einige faktorenanalytische Befunde auf einen relativ starken allgemeinen Faktor räumlicher Fähigkeiten hinweisen, ist insgesamt jedoch eher die Annahme einer Mehrdimensionalität mit der Datenlage in Einklang zu bringen. Eine Strukturierung der Dimensionen, wie sie von Linn und Petersen (1985) vorgelegt wurde, ist dabei eine (von einer Vielzahl möglicher) Klassifikationen, die mehr oder weniger eindeutig auf psychometrische Analysen bezogen sein können“ (Lohaus et al., 1999, S. 22).

Die Fragestellung der vorliegenden Arbeit zielt primär auf die Generierung aktueller Befunde zum Zusammenhang von *Geschlecht*, *Raumvorstellung* und *Mathematikleistung* für Schülerinnen und Schüler kurz vor Ende der Sekundarstufe I. Eine über den berichteten Forschungsstand hinausgehende Klärung des Konstrukts *Raumvorstellung* ist hierfür nicht erforderlich. Es muss aber sichergestellt werden, dass ein hinreichend breites und klares Konstrukt zugrunde gelegt wird, das dabei auch solche Komponenten enthält, die für die

genannte Fragestellung relevant sind und die sich klar durch zur Verfügung stehende Instrumente bzw. adaptierbare Instrumente erfassen lassen. Aus dieser Perspektive werden die in Kap. 3.1.2 dargestellten Modelle der inneren Struktur von *Raumvorstellung* im Folgenden unter Berücksichtigung der Befunde, die in den Kapiteln Kap. 3.3.1 bis 3.3.6 dargestellt wurden, diskutiert.

Für die primär faktorenanalytisch geprägten Modelle der inneren Struktur von *Raumvorstellung* war Thurstones 3-Faktoren-Modell (*S1*: *Spatial Relations*; *S2*: *Visualization*; *S3*: *Spatial Orientation*) der zentrale Bezugspunkt (vgl. Thurstone, 1950). Die Unterscheidung von *S1* und *S3* erwies sich sowohl kognitionspsychologisch als problematisch, da bei *S1* und bei *S3* jeweils die gleichen Lösungsstrategien erfolgreich sein können (vgl. Barratt, 1953), als auch empirisch nicht tragfähig, da beide Komponenten nie innerhalb einer Stichprobe empirisch voneinander getrennt werden konnten (vgl. Pawlik, 1968). Dies führte dazu, dass bereits seit den 1950er-Jahren die Faktoren *S1* und *S3* zusammengefasst wurden und im Folgenden überwiegend ein entsprechendes 2-Faktoren-Modell (*Vz*: *Visualization*; *SR-O*: *Spatial Relations and Orientation*) angenommen wurde, das unter anderem durch die Arbeit von Michael et al. (1957) etabliert und z. B von D. H. Rost (1977) und McGee (1979) unterstützt wurde.[89]

Dagegen legen Linn & Petersen (1985) in ihrer Meta-Analyse zu Geschlechterunterschieden in der *Raumvorstellung* ein 3-Komponenten-Modell (*räumliche Wahrnehmung*, *mentale Rotation* und *räumliche Visualisierung*) vor, das sie zunächst mit Blick auf Lösungsprozesse bei der Bearbeitung der Aufgaben kognitionspsychologisch begründen und anschließend psychometrisch über homogene Effektstärken für Geschlechterunterschiede absichern. Auf eine faktorenanalytische Trennung der Komponenten wurde hingegen verzichtet; entsprechende Versuche wären vermutlich nicht erfolgreich gewesen.[90]

Die meisten jüngeren Arbeiten, die nicht nur eine spezielle Komponente der *Raumvorstellung* oder sogar nur einen konkreten Raumvorstellungstest zum Gegenstand haben, berücksichtigen bei der Klärung ihrer theoretischen Grundlage sowohl das faktorenanalytisch gewonnene 2-Faktoren-Modell (mit den Bezeichnungen nach Thurstone, 1950, oder Michael et al., 1957) als auch das kognitionspsychologisch angereichte 3-Komponenten-Modell (nach Linn & Petersen, 1985). Auch wenn sich einige Arbeiten dann für das eine oder das andere Modell als theoretische Grundlage entscheiden, resultiert aus dem nebeneinander der Modelle – zumindest zwischen einigen Arbeiten – ein gewisses, manchmal

[89] Hierbei ist wichtig zu bemerken, dass sowohl Michael et al. als auch Thurstone noch einen weiteren Faktor „*K*“ (Michael et al.: „Kinesthetioc Imagery“; Thurstone: „Kinesthetic Factor“) andeuten, diesen aber nicht für ähnlich abgesichert wie *Vz* und *SR-O* bzw. *S1* bis *S3* halten und daher auf eine explizite und gleichberechtigte Aufnahme ins jeweilige Modell verzichten.

[90] So weisen die dort genannten „zweidimensionalen“ Referenztests für *mentale Rotation* eine zu große empirische Nähe zu den Referenztests für *räumliche Visualisierung* auf.

auch nur scheinbares „Zuordnungschaos“ auf der Ebene der konkreten Tests. Dies ist bei einer ersten Annäherung an den Gegenstand verwirrend und auf den zweiten Blick erstaunlich, da viele Tests doch recht eindeutig bestimmte mentale Prozesse zur Lösung erfordern (auch wenn nicht alle Tests strategiehomogen sind, siehe Kap. 3.3.3). Dieses „Zuordnungschaos“ soll an einigen Beispielen ohne weitergehende Wertung erläutert werden:[91]

- Beispiel *3 DW* – In der Testbesprechung von Fay (1992, S. 171) steht: „Das Verfahren soll räumliches Vorstellungsvermögen im Sinne von ‚spatial visualization’ und ‚spatial orientation’ erfassen.“ Dabei ist zunächst nicht klar, durch welche Definitionen dieser Komponenten präzisiert werden. In der umfassenden Arbeit zur *Raumvorstellung* von Maier (1999b, S. 64; i. O. m. Herv.) wird der *3 DW* als Test, „der vorwiegend auf die Erfassung der Komponente Räumliche Beziehungen abzielt“, beschrieben. Diese Komponente versteht Maier im Sinne von Thurstones *S1* (ebd., S. 38 ff.). Das Testmanual des *3 DW* (Gittler, 1990, S. 18) offenbart, dass der Testautor „Spatial Visualization“ im Sinne von McGee (1979) versteht, der diese Kategorie im Sinne von *S2* und *Vz* beschreibt. Vermutlich wird „Spatial Orientation“ auch im Sinne McGees verstanden,[92] der unter dieser Bezeichnung Thurstones *S1* und *S3* zusammenfasst und diese Kategorie somit anscheinend im Sinne von *SR-O* nach Michael et al. (1957) versteht. So lässt sich also die nur scheinbar bestehende Differenz zwischen der Zuordnung zu „Spatial Orientation“ (Gittler; Fay) und „Räumliche Beziehung“ (Maier) auflösen[93]. Offen bleibt aber die Frage der Zuordnung zu „Spatial Visualization“, eine Kategorie, die in Maiers Modell vorgesehen ist, auf die er den *3 DW* aber – anders als der Testautor – nicht bezieht.
- Beispiel *IST:WÜ* – Der *IST:WÜ* basiert auf Thurstones Test „Cube Comparison“. Der zuvor betrachtete *3 DW* ist eine „dreidimensionale“ Weiterentwicklung des *IST:WÜ* unter Beachtung der Lösungsstrategien. Während Klieme (1986, S. 135) den *IST:WÜ* der Komponente „räumliches Visualisieren“ zuordnet, verbinden Hartmann & Reiss (2000, S. 87) „Cube Comparison“ mit der Komponente „Mental Rotation“. Klieme (1986) bezieht sich zwar nicht explizit auf eine vorhandene Definition der Komponente „räumliches Visualisieren“, liefert aber eine eigene mit: „mentale Analyse und Veränderung von Objekten unabhängig von der eigenen Position“. Er verwendet außerdem die Kategorie „‚räumliches Orientieren’ (= Beurteilung von Objekten aus verschiedenen in der Vorstellung eingenommenen Perspektiven)“ (ebd., S. 135). Dieser Komponente ordnet

[91] Für die vorliegende Arbeit selbst ist letztlich nur wesentlich, dass die Raumvorstellungstests, die in der eigenen empirischen Untersuchung verwendet werden, theoretisch wie empirisch klar zugeordnet werden.

[92] Dies geht nicht eindeutig aus dem Testmanual hervor.

[93] An diesem Beispiel wird aber deutlich, dass McGees Bezeichnung nicht hilfreich ist, da „Spatial Orientation“ im klassischen Sinne (Thurstones *S1*) verlangt, dass die Versuchsperson im Sinne der Aufgabenstellung Bestandteil der fraglichen räumlichen Konfiguration ist. Beim *3 DW* ist dies augenscheinlich nicht der Fall, sondern vermutlich eher „Spatial Relation“ im klassischen Sinne (Thurstones S3). Daher ist die Bezeichnung *SR-O* von Michael et al. (1957) sicherlich tragfähiger.

er z. B. den Test *Schlauchfiguren* zu, der zumindest sehr hohe Anteile *mentaler Rotation* enthält. Da Hartmann & Reiss (2000) sich auf Linn & Petersen (1985) beziehen, hätten auch sie die Kategorie „Spatial Visualization" für „Cube Comparison" verwenden können, sich aber vermutlich bewusst dagegen entschieden. Hier scheint also keine Bezeichnungsproblematik vorzuliegen, sondern tatsächlich eine grundlegend andere Zuordnung vergleichbarer Tests zu Komponenten der *Raumvorstellung*.

- Beispiel *Schlauchfiguren* – Maier (1999b, S. 42 f.) verwendet *Schlauchfiguren* als Referenztest für „Räumliche Orientierung" im Sinne von Thurstones *S3*, obwohl er in seinem additiv zusammengesetzt Modell auch „Mental Rotation" sensu Linn & Petersen (1985) als Komponente vorsieht. In der Arbeit von Dorst (2007, S. 13) wird *Schlauchfiguren* als „mentale Rotationsaufgabe" verwendet.
- Beispiele *MRT* und *DAT:SR* – Diese beiden Tests werden in der Literatur zwar relativ konsequent zugeordnet (*MRT* → *mentale Rotation*; *DAT:SR* → *räumliche Visualisierung*), dies drückt sich aber nicht in der Bezeichnung der Tests aus: „*Mental Rotations*, a group test of three-dimensional *spatial visualization*" (Vandenberg & Kuse; 1978; Herv. d. d. Verf.); „Differential Aptitude Test – Subtest *Spatial Relations*" (Bennett et al., 1973). Im Falle des *MRT* könnte man zwar versuchen, die Bezeichnung damit zu erklären, dass 1978 in der Regel nur die Komponenten *Vz* und *SR-O* unterschieden wurden; dann wäre aber inhaltlich eine Zuordnung zu *SR-O* geboten. Für den *DAT:SR* ist eigentlich schon mit Thurstones Definition von *S2* die Zuordnung zu *Visualization* klar.

Zwar lässt sich die Zuordnung von Raumvorstellungstests zu einer oder auch mehreren der diskutierten Komponenten sicherlich theoretisch klären, wenn mehrere Expertinnen und Experten die jeweiligen Testmaterialien und Definitionen intensiv studieren und gemeinsam interpretieren, dies klärt aber noch nicht, wie sich die Komponenten der *Raumvorstellung* aus den verschiedenen Modellen zueinander verhalten. Eine knappe Skizze soll dies für die vorgestellten Modelle von Thurstone und Linn & Petersen andeuten:

- Realtiv klar ist die Situation für die Komponente *räumliche Visualisierung* (bzw. *S2* bzw. *Vz*). Ausgehend von Thurstone (1950) findet sich in der Substanz die nahezu identische Beschreibung der Komponente bei Michael et al. (1957), bei McGee (1979) und bei Linn & Petersen (1985). Auch unter Berücksichtigung der Arbeit von Zimmerman (1954; „Kontinuumshypothese", vgl. Kap. 3.1.2) kann man zwei Charakteristika entsprechender Aufgaben festhalten: (a) die relative Komplexität der Aufgaben, die in der Regel mehrschrittiges auch schlussfolgerndes Denken erfordert, und (b) die Beweglichkeit bzw. Möglichkeit der Veränderung einzelner Teile der betrachteten Objekte.
- Für *SR-O* nach Michael et al. (1957), also die Zusammenfassung von *S1* und *S3*, ist hingegen charakteristisch, dass die betrachteten Objekte bzw. Konfigurationen von Objekten (in ihrer internen räumlichen Beziehung zueinander) unverändert bleiben. Entweder sollen die Objekte bzw. Konfigurationen von Objekten als Ganzes mental in ihrer räumlichen Lage verändert werden oder die Versuchsperson soll mental eine andere Position

einnehmen. In diesem Sinne deckt *mentale Rotation* sensu Linn & Petersen (1985) einen Teil von *SR-O* ab.

- Bei Linn & Petersen kommt noch *räumliche Wahrnehmung* hinzu, die am ehesten Anteile von *K* sensu Thurstone (1950) bzw. Michael et al. (1957), also kinästhetische Anteile, beinhaltet. Für die mentale Nutzung der gravitativen Horizontalen oder Vertikalen müssen die Versuchspersonen sich mit ihrer Raumwahrnehmung in Beziehung zum Aufgabenmaterial setzen.

Welche Konsequenzen haben diese Betrachtungen für den empirischen Teil der vorliegenden Arbeit? Aus Sicht der zugrundeliegenden Fragestellung muss zunächst noch einmal festgehalten werden, dass die berücksichtigten mathematikdidaktischen Modelle von *Raumvorstellung* unter anderen Zielsetzungen entwickelt wurden und keine tragfähige Grundlage für die eigene empirische Untersuchung darstellen (vgl. Kap. 3.2.3). Im Rahmen der vorliegenden Arbeit spielt das psychologisch orientierte Konstrukt die wichtigere Rolle. In der Psychologie gibt es eine lange Tradition der Messung von *Raumvorstellung*, die sehr gut zur Zielsetzung und zur Methodik dieser Arbeit passt. Im empirischen Teil soll an diese Tradition angeknüpft werden, da die Zielsetzung ist, aktuelle, aber nicht gänzlich anders geartete Befunde zu generieren. Von den diskutierten Modellen aus der Psychologie scheint das von Linn & Petersen (1985) am besten für die eigene Fragestellung geeignet zu sein. Dies hat mehrere Gründe:

- Die drei vorgeschlagenen Komponenten lassen sich aufgrund der idealtypisch ablaufenden mentalen Prozesse gut beschreiben und klar voneinander trennen. Dabei bleibt allerdings fraglich, ob Versuchspersonen jeweils die intendierten Lösungswege beschreiten. Trotzdem ist das Modell kognitionspsychologisch klar und nachvollziehbar.
- Das Modell wurde unter einer Fragestellung entwickelt, die dem Erkenntnisinteresse der vorliegenden Arbeit sehr nahe kommt. Die psychometrische Absicherung der kognitionspsychologisch identifizierten Komponenten über homogene Effektstärken für Geschlechterunterschiede (ggf. für Teilkomponenten bzw. Subpopulationen) ermöglicht direkt eine Differenzierung der eigenen Untersuchung nach diesen Kategorien. In diesem Bereich muss die eigene Arbeit also nicht mehr explorieren, sondern kann hypothesengeleitet vorgehen.
- Zu den drei fraglichen Komponenten liegen klare Referenztests vor: *WLT* und *RFT* für *räumliche Wahrnehmung*, *MRT* für *mentale Rotation* und *DAT:SR* für *räumliche Visualisierung*. Weitere Tests lassen sich problemlos zuordnen, bei einigen dürften allerdings aufgrund der Definition von *räumlicher Visualisierung* auch die beiden anderen Komponenten stärker berührt werden.
- Linn & Petersen selbst formulieren mit Blick auf ihre drei Komponenten das folgende Forschungsdesiderat: „Sex differences in spatial ability do not generally account for sex differences in mathematics … . Studies that separate each type of spatial ability are needed to clarify these relationships further“ (S. 1493).

Die so getroffene vorläufige Festlegung des Konstrukts *Raumvorstellung* stellt zusammen mit den anderen Betrachtungen und Resultaten dieses Kapitels eine hinreichend tragfähige Grundlage für den empirischen Teil der eigenen Arbeit dar. Die ausführlichen Betrachtungen zur *Raumvorstellung* dienen dem theorie- und hypothesengeleiteten Vorgehen und stellen sicher, dass relevante Komponenten der *Raumvorstellung* in die eigene Untersuchung einbezogen werden. Offen ist noch eine eigene inhaltliche Definition von *Raumvorstellung*. In der Einleitung zu diesem Kapitel steht die folgende Arbeitsdefinition (S. 63):

Raumvorstellung umfasst kognitive Fähigkeiten, die für die mentale Repräsentation und Transformation figuraler Informationen benötigt werden.

Zusammen mit der Festlegung des Konstrukts über die Komponenten von Linn & Petersen und den oben genannten Referenztests reicht diese Arbeitsdefinition als inhaltliche Grundlage aus. Dennoch soll noch eine Beschreibung von D. H. Rost (1977) zitiert werden, da sie mit der Differenz von *Wahrnehmung* und *Vorstellung* eine zusätzliche Perspektive betont: „Raumvorstellung ist nicht mit der Perzeption räumlicher Gegebenheiten gleichzusetzen, da nicht die Wahrnehmung, sondern das über die Wahrnehmung hinausgehende Vorstellen und Bewegen, also das gedankliche Handeln und Hantieren mit räumlichen Objekten, Begriffen und Relationen im Vordergrund stehen“ (ebd., S. 21).

4 Planung einer empirischen Untersuchung des Zusammenhangs von Raumvorstellung und Mathematikleistung

Ausgehend vom Problemaufriss in Kapitel 1 und der dort formulierten Fragestellung sind im Kapitel 2 wesentliche Grundlagen und aktuelle Befunde der Erforschung von *Mathematikleistung* und im Kapitel 3 konzeptionelle Entwürfe und empirische Befunde zur *Raumvorstellung* dargestellt und systematisiert worden. Dabei deuten die Befunde in den Kapiteln 2 und 3 darauf hin, dass sowohl *Raumvorstellung* als auch *Mathematikleistung* in geeignete Komponenten ausdifferenziert betrachtet werden müssen.

Im Folgenden werden die Grundlagen der eigenen empirischen Untersuchung dargestellt. Zunächst werden die zugrundeliegenden Vermutungen und Fragen präzisiert, die betrachteten Konstrukte festgelegt und mögliche Testinstrumente ausgewählt. Anschließend werden die Überlegungen zur Forschungsmethodik dargestellt und schließlich Grundzüge der Untersuchungsplanung begründet.

4.1 Fragestellung, Konstrukte und potenzielle Testinstrumente

Bei der folgenden Darstellung werden Fragestellung, Konstrukte und Instrumente analytisch getrennt dargestellt, obwohl sie inhaltlich auch in ihren Wechselwirkungen betrachtet werden. So sind beispielsweise im Bereich der *Raumvorstellung* einige Fragen nur für ausgewählte Komponenten interessant, die sich wiederum idealtypisch durch einen spezifischen Referenztest charakterisieren lassen: Durch den *MRT* etwa wird eine klar abgegrenzte Komponente festgelegt (*mentale Rotation*), bei der Geschlechterunterschiede stabil mit bedeutenden Effektstärken auftreten und auf die sich viele Arbeiten beziehen.

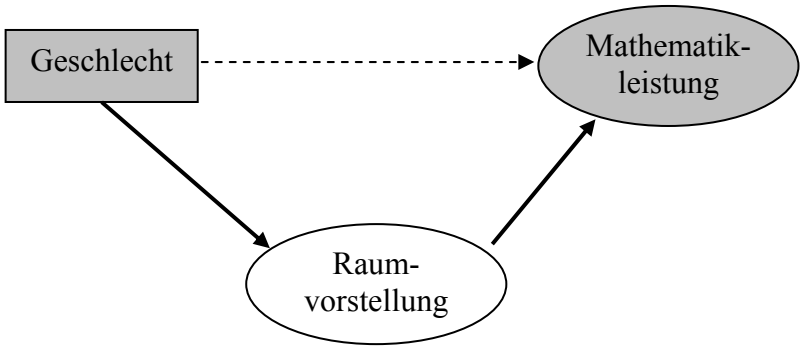
4.1.1 Präzisierung der Fragestellung

Fast alle vorliegenden Studien, die den Zusammenhang von *Raumvorstellung* und *Mathematikleistung* (ggf. unter Berücksichtigung von Geschlechterunterschieden) untersucht haben, erfassen mindestens eines der beiden Konstrukte undifferenziert oder in nur einer Komponente. Eine gewisse Ausnahme stellt die Studie von Klieme (1986) dar, die ein zentraler Bezugspunkt für die vorliegende Arbeit ist. Die Notwendigkeit, beide Konstrukte auch innerhalb *einer* Studie differenziert zu betrachten, betonen verschiedene Autoren, so z. B. Linn & Petersen (1985, S. 1493; Zitat auf S. 118 der vorliegenden Arbeit) oder Manger & Eikeland (1998):

> „Studies should also be conducted in which measures of visual-spatial ability with greater sensitivity to sex differences, such as mental rotation tests, are included in the assessment of the relationship between visual-spatial ability and mathematical achievement. Although research […] indicates that sex differences in these skills are robust, it would be interesting to examine whether the effects of these abilities increase with increasing mathematics task difficulty" (ebd., S. 24).

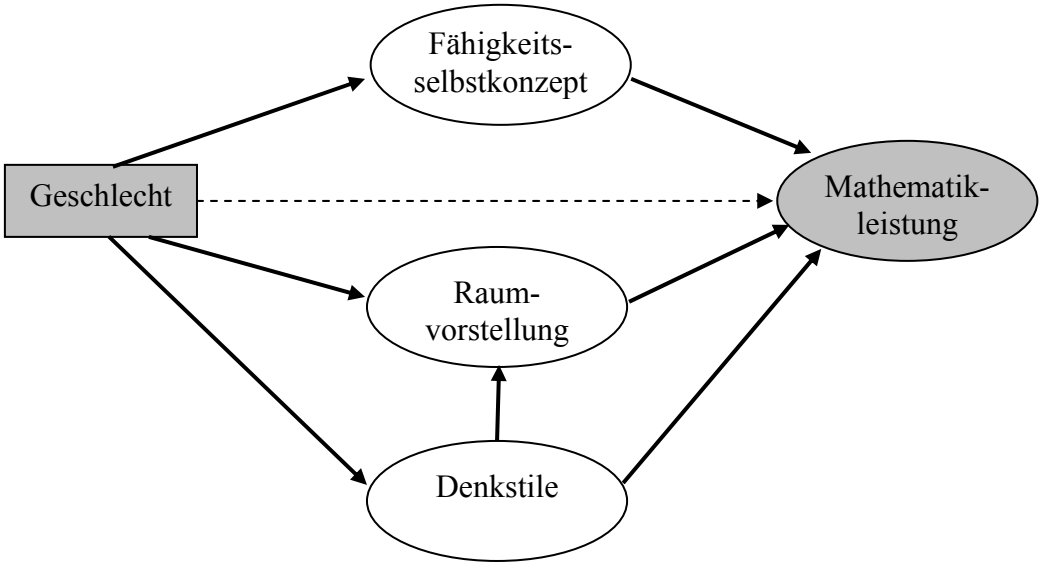
Die Ausgangsfrage der vorliegenden Arbeit ist, ob Geschlechterunterschiede in der *Mathematikleistung* zumindest zum Teil auf entsprechende Unterschiede in der *Raumvorstellung* zurückgeführt werden können. Diese Frage wird durch die *Spatial Mediation Hypothesis* konkretisiert (vgl. Kap. 3.3.6), die schon in Abb. 3.9 schematisch dargestellt worden ist:

Abbildung 4.1 „Spatial Mediation Hypothesis"



In diesem Modell dürfte neben dem über die *Raumvorstellung* vermittelten (indirekten) Effekt ggf. noch ein weiterer Effekt direkt vom *Geschlecht* auf *Mathematikleistung* identifiziert werden können, der aber möglicherweise durch andere Variablen vermittelt wird. Als weitere *Mediatorvariablen* kommen aus dem Bereich der kognitiven Prozesse die *Denkstile* und aus dem Bereich der selbstbezogenen Kognitionen u. a. das *bereichsspezifische Fähigkeitsselbstkonzept* infrage. Beide Konstrukte wurden, da sie in der vorliegenden Arbeit nicht ähnlich zentral wie die *Raumvorstellung* sind, bisher nicht ausführlicher dargestellt. Dies wird im Kap. 4.1.2 geleistet. In Abbildung 4.2 wird die schematische Darstellung der *Spatial Mediation Hypothesis* um diese beiden Konstrukte zu einem erweiterten Modell zur Erklärung von Geschlechterunterschieden in der *Mathematikleistung* ergänzt.

Abbildung 4.2: Erweitertes Modell zur Erklärung von Geschlechterunterschieden in der Mathematikleistung



In diesem erweiterten Modell stellen die *Denkstile* nicht nur einen potenziellen Mediator zwischen *Geschlecht* und *Mathematikleistung*, sondern auch einen potenziellen Mediator zwischen *Geschlecht* und *Raumvorstellung* dar. Dies basiert auf den Betrachtungen in Kap. 3.3.3 zu unterschiedlichen Bearbeitungsstrategien bei Raumvorstellungsaufgaben. Möglicherweise sind die dort betrachteten Bearbeitungsstrategien eng mit den in Kap. 4.1.2 dargestellten Denkstilen verbunden. Die Pfeilrichtungen im obigen Schema sind zunächst aus der Weiterentwicklung der *Spatial Mediation Hypothesis* entstanden. Bereits in Kap. 3.3.5 wurde die überwiegend angenommene Wirkungsrichtung von *Raumvorstellung* auf *Mathematikleistung* kritisch hinterfragt, da auch eine umgekehrte Wirkung oder eine Wechselwirkung denkbar ist. In ähnlicher Weise können weitere im obigen Schema postulierte Richtungen hinterfragt werden. Diese Diskussion wird bei Bedarf später geführt.

Das obige Modell stellt die Grundlage für die eigene empirische Untersuchung dar. Dabei wird nicht angenommen, dass bereits alle potenziellen Mediatorvariablen zwischen *Geschlecht* und *Mathematikleistung* berücksichtigt sind. In Kap. 2.3.2 wurden auch Variablen wie *Interesse am Fach*, *Leistungsangst* oder *Geschlechtsrollenstereotype* als weiteren Kandidatinnen benannt. Die vorliegende Untersuchung nimmt aber primär die Rolle der *Raumvorstellung* in diesem Modell in den Blick. Bevor Testdaten ausgehend von dem obigen Modell, z. B. in einer Pfadanalyse, auswertet werden, müssen bestimmte Voraussetzungen bei den einzelnen Variablen erfüllt sein – ggf. muss das Modell nach ersten empirischen Ergebnissen auch noch modifiziert werden.

So sollten zunächst signifikante Geschlechterunterschiede in der *Mathematikleistung* identifiziert werden, bevor sie möglicherweise auf Mediatorvariablen zurückgeführt werden.

Dabei ist offen, welche innere Struktur von *Mathematikleistung* hier ergiebig ist. Die Befunde aus Kapitel 2 sind in dieser Frage nicht eindeutig. Bei der *Raumvorstellung* stellt sich ebenfalls die Frage, welche konkreten Tests für die vorliegende Arbeit weiterführend sind. Und bei den *Denkstilen* ist ungeklärt, ob sie sich in einem *Paper and Pencil Test* angemessen erfassen lassen.

Mit den obigen Betrachtungen lässt sich die ursprüngliche Fragestellung durch Unterkategorisierungen konkretisieren:

- *Mathematikleistung*
  - Welche Geschlechterunterschiede lassen sich feststellen?
  - Welche Komponenten lassen sich empirisch trennen?
  - Gibt es Komponenten, bei denen Geschlechterunterschiede besonders groß / besonders klein sind?
  - Gibt es inhaltlich beschreibbare Leistungsprofile, die anhand der Lösungsprofile[94] identifiziert werden können?
- *Raumvorstellung*
  - Welche Geschlechterunterschiede lassen sich feststellen?
  - Welche Komponenten lassen sich empirisch trennen?
  - Gibt es Komponenten bzw. Tests, bei denen Geschlechterunterschiede besonders groß / besonders klein sind?
  - Gibt es für die einzelnen Tests Bearbeitungsstrategien, die anhand der Lösungsprofile identifiziert werden können?
- *Zusammenhang von Raumvorstellung und Mathematikleistung*
  - Gibt es jeweils Komponenten, für die der Zusammenhang besonders stark / besonders schwach ist?
  - Welche Geschlechterunterschiede in der *Mathematikleistung* lassen sich statistisch auf entsprechende Unterschiede in der *Raumvorstellung* zurückführen?
- *Denkstile*
  - Können *prädikatives* bzw. *funktionales Denken* in einem *Paper and Pencil Test* angemessen erfasst werden?
  - Welche Geschlechterunterschiede lassen sich ggf. feststellen?

[94] Als Lösungsprofil einer Versuchsperson in einem Test wird der „Antwortvektor“ verstanden, in dem die Antworten zu allen Testitems kodiert sind.

- *Zusammenhang von Denkstilen mit Raumvorstellung und Mathematikleistung*
  - Gibt es ggf. Zusammenhänge zwischen den *Denkstilen* und Bearbeitungsstrategien bei Raumvorstellungsaufgaben?
  - Gibt es ggf. Zusammenhänge zwischen den *Denkstilen* und der Raumvorstellungsleistung?
  - Gibt es ggf. Zusammenhänge zwischen dem jeweiligen *Denkstil* und der *Mathematikleistung* oder Komponenten der *Mathematikleistung*?
- *Bereichsspezifisches Fähigkeitsselbstkonzept – Mathematik*
  - Welche Geschlechterunterschiede lassen sich feststellen?
- *Zusammenhang von Fähigkeitsselbstkonzept mit Mathematikleistung*
  - Welcher Zusammenhang lässt sich feststellen?
- *Gesamtmodell zur statistischen Erklärung der Geschlechterunterschiede in der Mathematikleistung*
  - Welche Konstrukte können bzw. sollen in welcher Konkretisierung und ggf. mit welchen Komponenten in das Modell aufgenommen werden?
  - Wie stark sind die einzelnen Zusammenhänge im Rahmen des Modells?
  - Wie viel Varianz in der *Mathematikleistung* lässt sich mit diesem Modell erklären?
  - Welche Anteile der Geschlechterunterschiede können auf Mediatorvariablen zurückgeführt werden?

4.1.2 Festlegung der Konstrukte

In den Kapiteln 2 und 3 sind solche Grundlagen und Befunde zur *Mathematikleistung* und *Raumvorstellung* ausführlich dargestellt worden, die für die vorliegende Arbeit von Bedeutung sind. Im Folgenden werden die Konstrukte festgelegt, die in der eigenen empirischen Untersuchung erfasst werden sollen. Für die *Raumvorstellung* ist dies in Kapitel 3 umfassend vorbereitet worden. Die *Mathematikleistung* wird mit den *LSE 9* aus dem Jahr 2004 erfasst, sodass kein Einfluss auf die Gestaltung des Tests genommen werden konnte. Das Konzept der Lernstandserhebungen wird hier nur kurz und in Kap. 6.1.3 differenzierter dargestellt, da dieses Instrument erst im zweiten Untersuchungsschritt („Hauptuntersuchung“) eingesetzt wird. Das b*ereichsspezifische Fähigkeitsselbstkonzept* ist in den Kapiteln 1 und 2 am Rande thematisiert worden und wird noch einmal systematisch dargestellt. Schließlich muss das Konstrukt *Denkstile* noch so diskutiert werden, dass es im Rahmen der eigenen Untersuchung berücksichtigt werden kann.

Raumvorstellung

Im empirischen Teil der vorliegenden Arbeit wird *Raumvorstellung* im Sinne der Arbeitsdefinition aus Kapitel 3 verstanden:

Raumvorstellung umfasst kognitive Fähigkeiten, die für die mentale Repräsentation und Transformation figuraler Informationen benötigt werden.

Eine so verstandene *Raumvorstellung* umfasst natürlich auch „Large-Scale Fähigkeiten" wie etwa Orientierungsleistungen in einer Großstadt oder im (ggf. unübersichtlichen) Gelände. Eine pragmatische Fokussierung auf *Small-Scale Fähigkeiten*, die in *Paper and Pencil Tests* erfasst werden können, stellt jedoch für die Fragestellung der eigenen Untersuchung keine Einschränkung dar, da beim Lehren und Lernen von Mathematik in der Regel *Small-Scale Fähigkeiten* eine Rolle spielen. Dies liegt schon darin begründet, dass schriftliche Lernmaterialien, aber auch konkrete Objekte, mit denen handlungsorientiert gelernt werden kann, in der Regel keinen großen Platzbedarf haben.

In Kap. 3.4 wurde ausführlich begründet, warum das 3-Komponenten-Modell der *Raumvorstellung* nach Linn & Petersen (1985) für die vorliegende Arbeit vermutlich am ergiebigsten ist. Für dieses Modell mit den Komponenten *räumliche Wahrnehmung*, *mentale Rotation* und *räumliche Visualisierung* sprachen vor allem die folgenden Punkte:

- Die mentalen Prozesse, die bei den drei Komponenten idealtypisch beteiligt sind, lassen sich klar und trennscharf beschreiben.
- Es liegen gut erprobte und häufig verwendete Referenztests vor, die sich klar und trennscharf auf diese Komponenten beziehen.
- Innerhalb der Komponenten konnten Linn & Petersen (zum Teil nach weiterer Unterteilung) homogene Effektstärken für Geschlechterunterschiede finden.
- Die Autorinnen selbst fordern Forschung zur komponentenspezifischen Untersuchung des Zusammenhangs von *Raumvorstellung* und *Mathematikleistung* ein.

Hiermit liegt eine theoretisch abgeleitete und potenziell ergiebige konzeptionelle Grundlage für die empirische Untersuchung vor, auf deren Basis die Auswahl, Zusammenstellung und ggf. Adaption vorliegender Raumvorstellungstests erfolgen kann.

Mathematikleistung

Die *Mathematikleistung* wird mit den nordrhein-westfälischen Lernstandserhebungen in der Jahrgangsstufe 9 aus dem Jahr 2004 (LSE 9) erfasst. Wie andere Vergleichsarbeiten auch sollen die *LSE 9* den Schulen Leistungsdaten zur Verfügung stellen, die innerhalb des Prozesses der ergebnisorientierten Unterrichtsentwicklung genutzt werden können (vgl. Büchter & Leuders, 2005a). Über die Leistungsdaten hinaus, anhand derer die Lehrkräfte ihre Lerngruppen mit anderen Lerngruppen innerhalb der Schule und landesweit vergleichen können, erhalten die Schule umfassende didaktische Hinweise zu den eingesetzten Testaufgaben und zur möglichen Weiterarbeit im Unterricht. Konzeptionell sind die *LSE 9* eng auf die nordrhein-westfälischen Kernlehrpläne (MSJK, 2004a-d) abgestimmt, stellen also einen curricular validen Test dar. Allerdings wird nicht die Breite des Curriculums

erfasst. Zwar werden alle vier Inhaltsbereiche der Kernlehrpläne abgedeckt, innerhalb der „prozessbezogenen Kompetenzen“ wird bei den *LSE 9* im Jahr 2004 aber auf den Bereich *Modellieren* fokussiert.

Die Testentwicklung findet bei den *LSE 9* wie bei den großen Schulleistungsstudien mit einer Pilotierung der Aufgaben und *IRT*-Skalierung der Ergebnisse statt (vgl. Heymann & Pallack, 2007). Das Kompetenzstufenmodell aus den *LSE 9* im Jahr 2004, das eine Modellieren-Skala darstellt, ist in Kap. 2.1.3 kurz vorgestellt worden. Mit Blick auf mögliche Komponenten-Modelle der *Mathematikleistung* ist aufgrund der o. g. Beschränkung auf den Prozess des Modellierens eine Strukturierung nach den typischen mathematischen Prozessen (*Modellieren*, *Problemlösen*, *Argumentieren*) nicht möglich. Allerdings hatten die Ergebnisse in Kap. 2.3.2 auch nicht darauf hingedeutet, dass eine solche Strukturierung ergiebig sein könnte. Andere Modelle, z. B. nach *Typen mathematischen Arbeitens*, nach Inhaltsbereichen o. ä., können für die *LSE 9* sinnvoll realisiert werden.

Bereichsspezifisches Fähigkeitsselbstkonzept Mathematik

Allgemeine und spezifische *Selbstkonzepte* sind aufgrund ihrer großen Bedeutung beim institutionellen Lernen, vor allem in der Schule, wichtige Forschungsgegenstände der pädagogischen Psychologie, aber auch wichtig für die Schulpädagogik und die Fachdidaktiken. Mit dem Konstrukt *Selbstkonzept* werden üblicherweise *Beschreibungen* einer Person über sich selbst, also selbstbezogene Kognitionen, verstanden, die sich auf Fähigkeiten und Eigenschaften beziehen (vgl. Moschner & Dickhäuser, 2006, S. 685 ff.). Die Bedeutung für den schulischen Bereich fassen D. H. Rost et al. (2004) wie folgt zusammen:

> „In Lern- und Leistungssituationen wirken sich Vorstellungen, die Personen über die Höhe ihrer eigenen Fähigkeiten haben (Selbstkonzepte), vielfältig auf Erleben und Verhalten aus […]. In Abhängigkeit von der Höhe ihres Selbstkonzepts erklären sich Lernende das Zustandekommen eigener Leistungen unterschiedlich […], sind unterschiedlich ausdauernd bei der Bearbeitung von Aufgaben […] und bilden unterschiedlich hohe Erfolgserwartungen aus […]. Weiterhin sind Effekte des Selbstkonzepts auf die Wahl von Fächern oder Aufgaben […] sowie auf spätere Leistungsmaße beobachtet worden“ (ebd., S. 44).

Unterhalb bzw. innerhalb eines allgemeinen *Selbstkonzepts*, das aus der Gesamtheit solcher Beschreibungen entsteht, lassen sich spezifische Bereiche in hierarchischen bzw. additiven Modellen ausdifferenzieren, z. B. das *akademische Selbstkonzept*, das sich vor allem auf fachliches Lernen bezieht. Dieses *akademische Selbstkonzept* kann weiter spezialisiert werden, sodass mit Blick auf das Fach Mathematik das „bereichsspezifische Fähigkeitsselbstkonzept Mathematik (FSK:M)“ entsteht. In dieser Spezialisierung werden die *bereichsspezifischen Fähigkeitsselbstkonzepte* als wichtige „Determinanten von Schulleistung“ Helmke & Schrader (2006) betrachtet. Dabei ist nahe liegend, dass nicht nur das *FSK:M* die *Mathematikleistung*, sondern auch das Erleben der eigenen Leistung das *FSK:M* beeinflussen kann. Diese Wechselwirkung ist mittlerweile auch empirisch stabil nachweisbar, wobei die inhaltlichen Wirkungsmechanismen noch nicht identifiziert sind:

> „In der Forschung zu Fähigkeitsselbstkonzepten herrscht weitestgehend Einigkeit über zwei Punkte: Erstens determiniert die vorangegangene Leistung einer Person in Teilen das Fähigkeitsselbstkonzept. Dieser Befund spricht für die Annahme des *skill development*-Ansatzes […], der zufolge das Fähigkeitsselbstkonzept zumindest teilweise aufgrund von kumulierten Erfahrungen entsteht. Zweitens kann aber ebenso wenig angezweifelt werden, dass das Fähigkeitsselbstkonzept nachfolgendes Erleben und Verhalten von Personen beeinflussen kann (der Effekt des Fähigkeitsselbstkonzepts auf Leistung wird als Prozess des *self enhancement* bezeichnet). Bei genauerer Inspektion der Literatur stellt sich allerdings heraus, dass wesentliche Mechanismen von *skill development*- und *self enhancement*-Prozessen unklar sind“ (Dickhäuser, 2006, S. 6).

Bereichsspezifische Fähigkeitsselbstkonzepte entstehen vor allem im sozialen und dimensionalen Vergleich. So wird das *FSK:M* dadurch beeinflusst, wie eine Person ihre *Mathematikleistung* und ihr Lernen von Mathematik einerseits im Vergleich zu anderen Personen und andererseits im Vergleich zu anderen Fächern wahrnimmt. Daraus resultiert, dass die gleiche objektiv festgestellte *Mathematikleistung* bei verschiedenen Personen zu ganz unterschiedlichen Ausprägungen des *FSK:M* führen kann:

> „Der Logik von Referenzrahmentheorien zufolge ist es nicht die (wahrgenommene) Leistung an sich, die eine Veränderung der Fähigkeitsselbstkonzepte bewirkt, sondern das Ergebnis des Leistungsvergleichs. So können bei gleicher Leistung ganz unterschiedliche Vergleichsergebnisse resultieren, je nach dem, mit welchen anderen Leistungen verglichen wird. Besonders deutlich wird dies an einem Phänomen, das man als «big-fish-little-pond-Effekt» bezeichnet […]. Zwei Schüler gleicher Leistungsstärke entwickeln unterschiedliche Fähigkeitsselbstkonzepte, je nachdem, wie leistungsstark ihre Vergleichsgruppe (z. B. ihre Klasse oder ihre Schule) ist: In einer leistungsschwachen Gruppe empfindet sich der Schüler ähnlich einem großen Fisch in einem kleinen Teich als vergleichsweise leistungsstark, während bei identischem Leistungsniveau in einer leistungsstarken Gruppe eine niedrigere Einschätzung eigener Leistungsfähigkeit resultiert“ (Dickhäuser, 2006, S. 6).

Es gibt weitere psychosoziale Konstrukte, die mit dem *Selbstkonzept* verwandt sind, in Wirkungsgefügen mit ihm stehen und zuweilen mit ihm verwechselt werden. An dieser Stelle werden daher das *Selbstwertgefühl* und die *Selbstwirksamkeit* vom *Selbstkonzept* abgegrenzt. Während das *Selbstkonzept* Beschreibungen von eigenen Fähigkeiten und Eigenschaften umfasst, entsteht das *Selbstwertgefühl* aus der Bewertung dieser Fähigkeiten und Eigenschaften (Moschner & Dickhäuser, 2006, S. 685). *Selbstwirksamkeit* unterscheidet sich vom *Selbstkonzept* vor allem dadurch, dass soziale und dimensionale Vergleiche keine Rolle spielen und die selbstbezogenen Kognitionen sich auf konkrete Anforderungssituationen beziehen, also kriteriale Bezugspunkte haben (vgl. Köller & Möller, 2006, S. 694).

Denkstile

Die „Theorie prädikativer versus funktionaler kognitiver Strukturen“ (Schwank, 1996) ist am Osnabrücker *Institut für Kognitive Mathematik* angeregt „durch langjährige Beobachtungen von Schülerinnen und Schülern beim Konstruieren und Analysieren von Algorithmen“ (Schwank, 2003b, S. 69) entstanden. Dabei werden zwei Arten des kognitiven Mo-

dellierens[95] unterschieden: *prädikatives Denken* und *funktionales Denken*. Innerhalb der mathematikdidaktischen Rezeption und Diskussion ist für dieses (inhaltlich noch auszuführende) Konzept auch die Bezeichnung „Denkstile" üblich (vgl. z. B. Hefendehl-Hebeker, 2003, S. 13). Aus mathematikdidaktischer Sicht ist zentral, „dass es bei vielen Menschen eine relativ stabile Vorliebe für eine der beiden kognitiven Strukturen gibt" (Schwank, 2003b, S. 69). Dies bedeutet, dass Lehrkräfte hinreichend für beide *Denkstile* sensibilisiert sein und beide Denkstile ansprechen können sollten, damit ihr Unterricht alle Schülerinnen und Schüler erreichen und fördern kann. Die beiden *Denkstile* lassen sich grob wie folgt charakterisieren (vgl. Schwank, 1996, 2003a):

- *Prädikatives Denken* kann auch als statisches Modellieren verstanden werden, wobei die Aufmerksamkeit auf gleiche bzw. ähnliche Merkmale (wie z. B. Formen oder Farben – allgemein: *Prädikate* im Sinne der klassischen Logik), auf Invariantes und auf strukturelle Zusammenhänge gerichtet ist. Personen, die so denken, erkennt man daran, dass sie mathematische Probleme „mental als Geflecht von *Prädikaten* modellieren" (Schwank, 2003b, S. 69) und dabei „das wiederholte Zutreffen von *Prädikaten* überprüft wird" (Schwank, 2003a, S. 70).
- *Funktionales Denken* kann hingegen als dynamisches Modellieren verstanden werden, wobei die Aufmerksamkeit auf Veränderungen und Wirkungsweisen, also auf Prozesse gerichtet ist. Personen, die so denken, erkennt man daran, dass sie mathematische Probleme durch Abfolgen von Handlungen bearbeiten, wobei „das wiederholte Funktionieren der Konstruktionsschritte getestet wird" (ebd.).

Welcher Denkstil bei der Lösung eines Problems Anwendung findet, lässt sich – je nach Aufgabenstellung – durch die Lösungen selbst, durch den Lösungsweg bzw. Begründungen für die Lösung und bei stärker visuellen Aufgaben durch Augenbewegungen und EEG-Messungen diagnostizieren (vgl. ebd., S. 73 ff.). Bei entsprechenden empirischen Untersuchungen hat Schwank (2003a) herausgefunden, „dass sich Mädchen häufig durch gute prädikative Leistungen auszeichnen, selten durch gute funktionale, und dass es bei Jungen, wenn auch nicht ganz so extrem, umgekehrt der Fall ist" (ebd., S. 75). Die Analyse von Lernmaterialien für den Mathematikunterricht zeigt, dass vor allem im Anfangsunterricht häufig prädikative Denkwege stärker durch das Material unterstützt werden, obwohl es auch alternative Materialien gibt (vgl. ebd., S. 75 ff.).

Analysen von Unterrichtsgeschehen an Schule und Hochschule, wie die von Büchter & Henn (2004), Gallin (2003), Hefendehl-Hebeker (2003) und Kaune (2003), zeigen, dass die Kategorien *funktionales Denken* und *prädikatives Denken* eine große Erklärungsmacht

[95] „Kognitives Modellieren" steht hier für bestimmte Arten mentaler Prozesse, die bei der Bearbeitung von Aufgaben ablaufen, und hat zunächst nichts gemein mit der mathematikdidaktischen Kategorie „Modellieren", die das Wechselspiel von „Mathematik" und dem „Rest der Welt" bezeichnet.

haben und zum Verstehen von individuellen Denkwege beitragen können. Büchter & Henn (2004) und Gallin (2003) zeigen dies anhand unterschiedlicher Aufgaben aus der Wahrscheinlichkeitsrechnung, die eher kombinatorisch-statisch mit der Betrachtung der Ergebnismenge oder eher dynamisch durch die Modellierung mit einem Baumdiagramm gelöst werden. Hefendehl-Hebeker (2003) und Kaune (2003) zeigen auch für Beispiele aus den Bereichen *Algebra und Funktionen* sowie *Elementargeometrie*, wie unterschiedliche mentale Modelle zu geeigneten Lösungen führen können. Die didaktische Herausforderung für Lehrkräfte besteht zunächst im Verstehen von Denkwegen, die nicht den eigenen entsprechen, und dann in der Auflösung von nur scheinbaren Gegensätzen bei unterschiedlichen Lösungsansätzen in einem diskursiven Unterrichtsgeschehen. Büchter & Henn (2004, S. 33 f.) konnten wiederholt feststellen, wie schwer es Lehramtsstudierenden fällt, sich auf einen Lösungsweg „der anderen Art“ gedanklich einzulassen.

Mit Blick auf das Konstrukt *Denkstile* können gerade innerhalb der Mathematikdidaktik an zwei Stellen Irritationen bezüglich der Bezeichnungen entstehen. So gibt es auch ein Konzept „mathematische Denkstile“ (vgl. z. B. Borromeo Ferri, 2004), das in anders geartetem Sinne verwendet wird. Noch größer ist die Verwechselungsgefahr beim *funktionalen Denken*, das in der Mathematikdidaktik – letztlich zurückgehend auf Felix Klein – auch mit „Vorstellungen zum mathematischen Funktionsbegriff“ bezeichnet werden könnte (vgl. Vollrath, 1989; Krüger, 2000).

Die Bedeutung der Denkstile für die vorliegende Arbeit ergibt sich nicht nur aus den oberflächlichen Parallelen mit *Raumvorstellung* wie figuralem Aufgabenmaterial und konsistenten Geschlechterunterschieden. Die Analyse von Bearbeitungsstrategien für Aufgaben zur *Raumvorstellung* (vgl. Kap. 3.3.3) legt vielmehr nahe, dass die dort als *holistisch* und *analytisch* bezeichneten Strategien raumvorstellungsbezogene Pendants von *funktionalem* bzw. *prädikativem Denken* sind. Schwank (2003a, S. 75) selbst stellt explizit die Analogie von *funktionalem Denken* und *mentaler Rotation* her. Wenn sich diese theoretisch plausiblen Betrachtungen auch empirisch nachweisen lassen, müssten männliche Versuchspersonen im Gegensatz zu weiblichen mental eher *holistisch* operieren, eher *funktional denken* und erfolgreicher beim *MRT* abschneiden. Vorliegende Einzelbefunde stützen diese Annahme bislang.

Eine für die vorliegende Arbeit zentrale und offene Frage im Zusammenhang mit *Denkstilen* ist, ob sie sich adäquat in einem *Paper and Pencil Test* erfassen lassen. Die Osnabrücker Diagnosematerialien (Schwank, 1999/2000) sind qualitativ orientiert, d. h. das beobachtete Testverhalten muss noch interpretiert werden. Dabei wird über rein verbale Informationen hinaus noch weiteres Testverhalten – seien es Gesten, Augenbewegungen oder EEG-Messungen – berücksichtigt. Schwank (2003a, S. 71 f.) verweist darauf, dass funktionale Lösungsansätze häufig nicht adäquat von den Versuchspersonen verbalisiert

werden können.[96] Bei der eigenen empirischen Arbeit muss also zunächst in der Voruntersuchung ein entsprechender *Paper and Pencil Test*, der üblichen Testgütekriterien entspricht, entwickelt werden.

4.1.3 Auswahl möglicher Instrumente

Nach der obigen Präzisierung der Fragestellung und der Festlegung der Konstrukte ist die Auswahl geeigneter Testinstrumente relativ einfach. Im Bereich *Raumvorstellung* müssen aus einem gewissen Überangebot von Tests diejenigen ausgesucht werden, die am besten zum festgelegten Konstrukt passen und die möglichst ergiebig für die Fragestellungen sind. Im Bereich *Denkstile* muss ein zusammengestelltes Instrument erst noch im Rahmen einer Voruntersuchung erprobt werden.

Raumvorstellung

In einer mathematikdidaktischen Arbeit könnte es naheliegend sein, Testmaterial zur *Raumvorstellung* auszuwählen oder zu entwickeln, das auf mathematischen oder mathematikdidaktischen Konzepten zur geometrischen Begriffsbildung aufbaut oder das nahe an den üblichen Lernmaterialien für den Geometrieunterricht ist. Die vorliegende Arbeit geht aus den beiden folgenden Gründen einen anderen Weg:

- Diese Arbeit basiert in ihren theoretischen Grundlagen zur *Raumvorstellung* vor allem auf einer langen psychometrisch und kognitionspsychologisch orientierten Tradition sowie den entsprechenden Konzepten, Tests und Befunden. Die Darstellungen in Kapitel 3 zeigen, dass für Fragestellungen in Kap. 4.1.1 an die Theoriebildung und an viele substanzielle Befunde aus dieser Forschungstradition angeknüpft werden kann. Die Befunde beziehen sich zum Teil – wie im Bereich *mentale Rotation* – direkt auf spezifische Referenztests. Diese Grundlagen können nicht ohne Weiteres genutzt werden, wenn gänzliches neues Testmaterial entwickelt wird.
- Je näher das Testmaterial an den gewohnten Aufgaben und Darstellungen des Mathematikunterrichts ist, desto größer werden nicht nur die inhaltlichen, sondern vermutlich auch die statistischen Zusammenhänge mit *Mathematikleistung*. Fraglich ist dann, welche Anteile dieser Zusammenhänge auf den Raumvorstellungsanteil des Testmaterials zurückgehen und welche auf den vertrauten Kontext.

Da sich das Konstrukt *Raumvorstellung* für den empirischen Teil dieser Arbeit am 3-Komponeten-Modell von Linn & Petersen (1985) orientiert, werden mögliche Tests zunächst nach diesen drei Komponenten ausgewählt:

[96] Dies lässt sich vermutlich mit den „Hirnhälftentheorien“ erklären, nach denen unterschiedliche Hirnhälften für figurales und verbales Denken „zuständig“ sind.

- Für die Komponente *räumliche Wahrnehmung* geben die Autorinnen den „Rod and Frame Test (RFT)“ (Witkin et al., 1962) und die „Water Level Tasks (WLT)“ (Piaget & Inhelder, 1971) als Referenztests an. Beim *RFT* soll in einen vorgegebenen schrägen Rahmen eine Stange eingezeichnet werden, die an der Mitte der oberen Rahmenseite frei beweglich eingehängt wird. Bei den *WLT* sollen die Versuchspersonen in schräg stehende Gefäße einen möglichen Wasserspiegel einzeichnen. Damit wird die Wahrnehmung der gravitativen Vertikalen bzw. Horizontalen durch diese beiden Tests potenziell erfasst. Beide Tests sollen zunächst Bestandteil der eigenen Untersuchung sein.
- Die Komponente *mentale Rotation* wird idealtypisch durch den „Mental Rotation Test (MRT)“ (Vandenberg & Kuse, 1978; Peters et al., 1995) erfasst. Der *MRT* ist sicherlich *der* Test für *dreidimensionale mentale Rotation* schlechthin. Linn & Petersen (1985) geben darüber hinaus noch Tests für die *zweidimensionale mentale Rotation* an. In Kap. 3.3.3 wurde bereits ausgeführt, dass entsprechende Tests nicht nur durch *mentale Rotation*, sondern auch sehr gut durch analytische Strategien bearbeitet werden können. Dadurch verwischt die Grenze zur Komponente *räumliche Visualisierung* in nicht gewünschtem Maße, sodass in der eigenen Untersuchung der *MRT* als einziger Referenztest für *mentale Rotation* verwendet wird.
- Für die *räumliche Visualisierung* wird schließlich der „Differential Aptitude Test – Subtest Spatial Relations (DAT:SR)“ (Bennett et al., 1973) verwendet, der von Linn & Petersen (1985) als Referenztest für diese Komponenten angeben wird und der seit jeher auch als Referenztest für S2 bzw. Vz genannt wurde. Beim *DAT:SR* sollen die Versuchspersonen entscheiden, welches von mehreren Objekten aus einer vorgegebenen Faltvorlage hergestellt werden kann. Dabei spielen nicht nur typische mathematische Figuren eine Rolle, sondern auch Gegenstände wie z. B. Bücher etc. Da der *DAT:SR* die *räumliche Visualisierung* ebenfalls idealtypisch und hinreichend klar abgegrenzt von *mentaler Rotation* erfasst, wird in der eigenen Untersuchung auf weitere Tests verzichtet.

Die drei Komponenten der *Raumvorstellung*, die der eigenen Untersuchung zugrunde liegen, werden mit den bisher benannten – und auch verfügbaren – Tests gut und trennscharf abgedeckt. Dennoch sollen zwei weitere Tests zunächst mit in die eigene Untersuchung aufgenommen werden, nämlich der „Dreidimensionale Würfeltest (3 DW)“ (Gittler, 1990) und ein „Zweidimensionaler Würfeltest (2 DW)“, der aus dem „Intelligenz-Struktur-Test – Untertest Würfelaufgaben (IST:WÜ)“ (Amthauer, 1973) entwickelt wurde. Dabei wurden ausschließlich „Flächenwürfel“ im Sinne von Putz-Osterloh (1977) berücksichtigt.

Mit dem *2 DW* und *3 DW* soll zunächst versucht werden, die Ergebnisse von Hosenfeld et al. (1997) und Köller et al. (1994) zu replizieren (siehe Kap. 3.3.3). Auf dieser Basis soll dann auch der *MRT* auf mögliche Strategieunterschiede bei der Bearbeitung untersucht werden. Sollten sich beim *2 DW* und *3 DW* wie in den beiden o. g. Studien „Analytiker“ und „Holistiker“ anhand ihrer Lösungsprofile in beiden Tests identifizieren lassen, so kann

anschließend untersucht werden, wie sich diese beide Gruppen beim *MRT* verhalten. Geiser et al. (2006) deuten mit ihrer empirisch gewonnenen Unterscheidung von „Rotators“ vs. „Nonrotators“ darauf hin, dass sich auch für den *MRT* eine Strategiefrage stellt. Die vorliegende Arbeit soll eine entsprechende Überprüfung der Konstruktvalidität des *MRT* leisten, die von den Ergebnissen des *2 DW* und *3 DW* ausgeht. Dies scheint bisher noch nicht in vergleichbarer Form erfolgt zu sein. Auf dem hier intendierten Weg kann der *MRT* hypothesengeleitet mit den beiden Fallgruppen untersucht werden. Würde man ausschließlich auf der Basis der *MRT*-Lösungsprofile arbeiten, wäre das Vorgehen eher explorativ und mögliche Ergebnisse wären vermutlich schlechter zu interpretieren. Die Konstruktvalidierung für den *MRT* scheint nicht zuletzt deshalb sinnvoll, da *mentale Rotation* in der Literatur die größten Geschlechterunterschiede und die stärksten Zusammenhänge mit *Mathematikleistung* aufweist.

Mit den hier vorläufig ausgewählten Tests für die eigene Untersuchung (*RFT*, *WLT*, *MRT*, *DAT:SR*, *3 DW*, *2 DW*) würde die Gesamterhebung in der Hauptuntersuchung aus forschungspragmatischer Sicht zu umfangreich ausfallen, sodass nach einer Voruntersuchung Tests – wie z. B. *2 DW* und *3 DW* nach der Konstruktvalidierung des *MRT* – entfallen oder gekürzt werden müssen.

Mathematikleistung

Die *Mathematikleistung* wird durch die *LSE 9* erfasst. Auf diesen Test kann in der Aufgabenstellung, Durchführung und Datenerhebung kein Einfluss genommen werden (vgl. Kap. 6.1.3). Allerdings kann die Auswertung der Daten über die im Rahmen von Vergleichsarbeiten übliche Auswertung hinausgehen, z. B. indem Aufgaben nach einem für die vorliegende Arbeit geeigneten Komponenten-Modell klassifiziert werden.

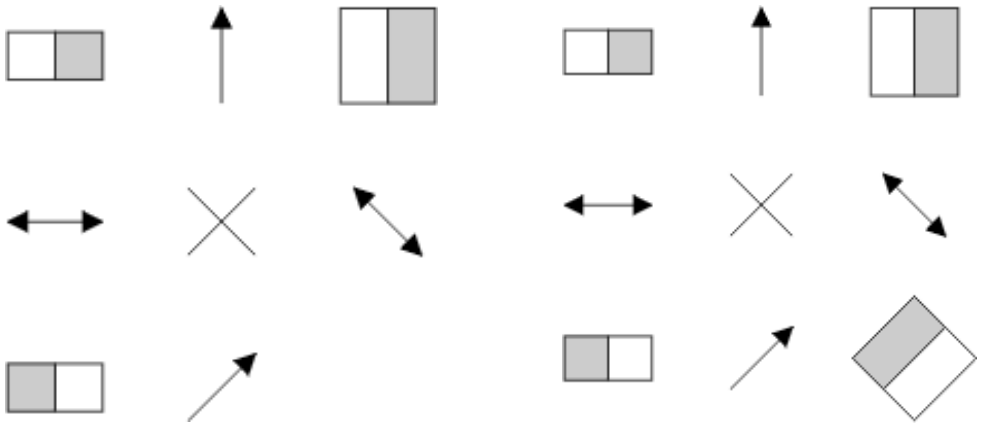
Bereichsspezifisches Fähigkeitsselbstkonzept Mathematik

Im Bereich der Selbstkonzeptforschung gibt es eine Vielzahl konkreter Testinstrumente. Dabei sollen die Versuchspersonen Aussagen wie „Mathematik ist eines meiner besten Fächer“ auf einer in der Regel vierstufigen Skala zwischen Polen wie „trifft überhaupt nicht zu“ und „trifft völlig zu“ einschätzen. Für das *FSK:M* können entsprechende Aussagen aus anderen Untersuchungen übernommen oder in Anlehnung daran formuliert werden.

Denkstile

Für den Bereich *Denkstile* muss ein *Paper and Pencil Test* entwickelt werden. Dabei kann auf das „Qualitatives Diagnoseinstrument für prädikatives versus funktionales Denken (QuaDiPF)“ (Schwank, 1999/2000) zurückgegriffen werden. Aufgaben aus dem Instrument *QuaDiPF* lassen sich z. B. im eBook „Kognitive Mathematik“ (Schwank, 1998) finden.

Abbildung 4.3: Aufgabe zur Diagnose von *Denkstilen* (links) mit einer möglichen Lösung (rechts) (Quelle: Schwank, 1998)



Da die Aufgabenlösungen allein nicht für eine Diagnose des Denkstils ausreichen, müssen die Versuchspersonen in einem *Paper and Pencil Test* dazu angeregt werden, ihre Denkwege zu beschreiben. Entsprechend muss neben den Aufgaben Platz für Beschreibungen sein und in die Aufgaben hinein gezeichnet werden dürfen.

4.2 Überlegungen zur Forschungsmethodik und Untersuchungsplanung

Im Folgenden werden zunächst einige forschungsmethodische Vorüberlegungen für die eigene empirische Untersuchung angestellt, da sie für die Gestaltung, Zusammenstellung, Durchführung und Auswertung der Datenerhebungen bedeutsam sein können. Dabei sollen mit Blick auf die präzisierte Fragestellung, auf die festgelegten Konstrukte und auf die möglichen Tests naheliegende statistische Modellierungen als potenzielles Methodeninventar bereitgestellt werden. In Kap. 2.1.2 wurde das *RM* mit einigen Varianten und Verallgemeinerungen vorgestellt und diskutiert, da seine Implikationen berücksichtigt werden müssen, wenn aktuelle Schulleistungsstudien auch methodenkritisch diskutiert werden sollen. Weitere statistische Verfahren werden im Rahmen dieser Arbeit nicht ausführlich diskutiert, sondern nur mit ihrer Kernidee skizziert und reflektiert angewendet.

4.2.1 Methodische Überlegungen zu geplanten Testbereichen

Die in Kap. 4.1.3 vorläufig ausgewählten Tests zur *Raumvorstellung* können alle mit *IRT*-Modellen skaliert werden. Im Gegensatz zu Fachleistungstests sind sie relativ homogen bezüglich der für die Aufgabenbearbeitung erforderlichen kognitiven Prozesse. Vor allem die Tests für *räumliche Wahrnehmung* und *mentale Rotation* sollten in diesem Sinne (weitgehend) homogen sein. Auf dieser Überlegung basierend ist die Testauswertung nach dem eindimensionalen zweikategoriellen *RM* nahe liegend, da die drei Tests (*RFT*, *WLT* und *MRT*) üblicherweise dichotom ausgewertet werden. Abweichend hiervon kann beim *RFT*

und beim *WLT* jeweils geprüft werden, ob die erhobenen Daten besser durch ein ordinales *RM* vorhergesagt werden können, in dem geringfügige Abweichungen von der Vertikalen bzw. Horizontalen beim *RFT* bzw. *WLT* als teilrichtige Lösung („partial credit“) gewertet werden. Ähnliche Überlegungen lassen sich beim *MRT* anstellen. Die *räumliche Visualisierung* hingegen ist durch größere Komplexität und größere Heterogenität der Denkwege gekennzeichnet. Dennoch sollten die Aufgaben des konkret ausgewählten Tests (*DAT:SR*) für eine Rasch-Modellierung hinreichend homogen sein – im Vergleich zu komplexen Fachleistungskonstrukten sind sie es in jedem Fall. Der *2 DW* und der *3 DW* sollten für sich auch jeweils mit dem eindimensionalen zweikategoriellen *RM* skalierbar sein; dies wurde durch eine strategiehomogene Aufteilung der ehemaligen *IST:WÜ*-Aufgaben in „Flächenwürfel“ (*2 DW*) und „Raumwürfel“ (*3 DW*) erreicht (vgl. Kap. 3.3.3).

Wenn sich alle Einzeltests wie oben angedeutet skalieren lassen, können die linearen Zusammenhänge zwischen den Tests in einem mehrdimensionalen *RM* als latente Korrelationen direkt messfehlerfrei geschätzt werden. Je nach Ausdifferenzierung eines solchen Modells muss nur berücksichtigt werden, dass die Zahl der zu schätzenden Parameter nicht zu groß wird, weil das Modell dann aus empirischer Sicht zu komplex würde.

Die Befunde aus Kapitel 3 deuten daraufhin, dass es nicht sinnvoll ist, einen Gesamttestwert für *Raumvorstellung* zu berechnen; deshalb werden Fragen wie die nach Geschlechterunterschieden komponentenweise beantwortet werden müssen. Ob der *LSE 9*- Mathematiktest sinnvoll in entsprechenden Komponenten aufgeteilt werden kann, ist noch offen; eine solche Differenzierung wird aber aus den in Kap. 2.3.3 ausgeführten Gründen angestrebt. Die Frage nach Geschlechterunterschieden kann für die einzelnen Leistungsbereiche in einem ersten Schritt z. B. mit Varianzanalysen untersucht werden (s. u.).

Bei den Raumvorstellungstests *2 DW* und *3 DW* soll in Analogie zu Hosenfeld et al. (1997) und Köller et al. (1994) nach Strategieunterschieden bei der Bearbeitung gesucht werden. Prinzipiell ist dies mit einer „Latent Class Analysis (LCA)“ möglich, die als klassifizierendes Verfahren auf die Lösungsprofile angewendet werden kann (s. u). Sollten sich auf diesem Wege unterschiedliche Klassen von Versuchspersonen identifizieren lassen, können die Testleistungen innerhalb dieser Klassen ggf. wiederum mit dem *RM* skalierbar sein. Dann hätte man einerseits eine qualitative Unterscheidung der Versuchspersonen, z. B. nach Bearbeitungsstrategien, und andererseits innerhalb der Klassen wieder eine Quantifizierung der Testleistung. Diese Kombination aus *LCA* und *RM* wird im „Mixed Rasch-Modell (MRM)“, dem gemeinsamen Obermodell von *LCA* und *RM*, realisiert (s. u.).

Bei der Auswertung des vorgegebenen *LSE 9*- Mathematiktests stellt sich im Besonderen die Frage, ob sich Aufgaben identifizieren lassen, bei denen es geschlechtsspezifisch unerwartet hohe oder unerwartet niedrige Lösungsquoten gibt. Dies lässt sich systematisch mit Analysen zum „Differential Item Functioning (DIF)“ (s. u.) untersuchen, die schon in Kap. 2.3.2 betrachtet wurden. Dabei müssen Gruppen von Versuchspersonen, z. B. bezüg-

lich des Geschlechts, *a priori* unterschieden werden. Für diese Gruppen werden dann entsprechende Berechnungen durchgeführt. Eine andere Möglichkeit, Komponenten in der *Mathematikleistung* zu identifizieren, ist auch hier potenziell eine *LCA* der Lösungsprofile.

Das aus der *Spatial Mediation Hypothesis* hervorgegangene erweiterte Modell zur Erklärung von Geschlechterunterschieden in der *Mathematikleistung* (vgl. Abb. 4.2, S. 122) soll schließlich, wenn alle anderen Annahmen zutreffen, Grundlage für die simultane Betrachtung verschiedener Mediatorvariablen zwischen *Geschlecht* und *Mathematikleistung* sein. Für eine empirische Analyse wird dies mithilfe von Strukturgleichungsmodellen zugänglich, die die Pfadanalysen als einen Spezialfall beinhalten (s. u.).

4.2.2 Ausgewählte Verfahren

Im Folgenden wird für die fünf zuvor genannten Verfahren bzw. Modelle (*Varianzanalyse*, *LCA*, *MRM*, *DIF*, *Strukturgleichungsmodelle*) die jeweilige Kernidee benannt und auf Literatur verwiesen, in der die zugehörigen Grundlagen der mathematischen Statistik dargestellt sind. Rechnerisch werden diese komplexen statistischen Verfahren mit aktuell üblichen Programmpaketen wie *SPSS*, *AMOS*, *ConQuest*, *Mplus* oder *WINMIRA* umgesetzt. Bei den eigenen Auswertungen wird in der Regel berichtet, mit welchem Programm die Ergebnisse berechet wurden.

Varianzanalyse

Möchte man Geschlechterunterschiede in einem Raumvorstellungs- oder Mathematiktest untersuchen, so sind Varianzanalysen eine mögliche Verfahrensgruppe (vgl. Fahrmeir, Hamerle & Nagl, 1996; Backhaus et al., 2008, Kap. 3). An dieser Stelle wird exemplarisch einer der einfachsten Fälle dargestellt, nämlich der Einfluss der unabhängigen Variable *Geschlecht* auf eine (als abhängig modellierte) Variable *Testleistung*.[97] Solche (univariaten) Varianzanalysen mit *einer* abhängigen Variablen werden nach dem englischen „Analysis of Varianz“ auch „ANOVA“ genannt. Die unabhängigen Variablen sind jeweils nominal- oder ordinal-skaliert (mit dem Spezialfall der dichotomen Variablen) und werden als *Faktoren* bezeichnet.

Im Rahmen einer *ANOVA* wird im fraglichen Fall ermittelt, welcher Anteil der gesamten Varianz der *Testleistung* auf den Faktor *Geschlecht* zurückgeführt werden kann. Dafür wird diese Gesamtvarianz in die Varianz zwischen den beiden *Geschlechtern* (= „Faktorstufen“) und die Varianz innerhalb der Faktorstufen zerlegt. Die Varianz zwischen den Faktorstufen erhält man nach der folgenden Idee: Wenn nur der Faktor *Geschlecht* die ge-

[97] Dieser einfachste Fall kann auch mit einem t-Test untersucht werden, da der Faktor Geschlecht nur zwei Ausprägungen hat und bei einem t-Test Mittelwerte für zwei Gruppen bezüglich signifikanter Unterschiede untersucht werden können.

samte Varianz der *Testleistung* V_g erklären würde, dann hätten alle weiblichen Versuchspersonen den gleichen Testwert θ_w und alle männlichen Versuchspersonen den Testwert θ_m.[98] Diese Konstellation wird nachgebildet, indem allen weiblichen Versuchspersonen der Mittelwert ihrer Gruppe μ_w zugewiesen wird und allen männlichen Versuchspersonen ebenfalls der Mittelwert ihrer Gruppe μ_m. Für den so erzeugten Gesamtdatensatz wird dann die Varianz V_f berechnet.

Setzt man diese Varianz V_f nun ins Verhältnis zur Gesamtvarianz V_g, so erhält man den Anteil der Varianz der *Testleistung*, der durch den Faktor *Geschlecht* erklärt wird. Die Berechnung kann entsprechend mit mehr als zwei Kategorien der unabhängigen Variablen, also mit mehr als zwei Gruppen, und auch mit mehr als einem Faktor durchgeführt werden. Ist einer dieser Faktoren intervallskaliert, so spricht man von einer Kovarianzanalyse, die z. B. zum Einsatz kommt, wenn die abhängige Variable *Mathematikleistung* mit den beiden Faktoren *Raumvorstellung* und *Geschlecht* untersucht werden soll. Jeweils wird mit geeigneten (F-verteilten) Prüfgrößen getestet, ob Gruppenunterschiede signifikant sind.

ANOVAs sind Standardverfahren, die in den gängigen Statistik-Programmpaketen implementiert sind. Bei Bedarf können *ANOVAs* aber auch auf Regressionsanalysen zurückgeführt werden. Dies wird für das Beispiel der Analyse von Geschlechterunterschieden skizziert: Man kodiert die unabhängige Variable *Geschlecht* mit den Werten „0“ und „1“ und verwendet sie wie einen intervallskalierten Prädiktor für die abhängige Variable *Testleistung*. Dann gibt das nicht-standardisierte Regressionsgewicht genau den Mittelwertunterschied zwischen den beiden Gruppen an. Die im Rahmen der Regressionsanalyse üblichen Signifikanzbetrachtungen führen zu den gleichen Resultaten wie die entsprechenden Betrachtungen bei einer analogen *ANOVA*.

Neben den hier beschriebenen univariaten Varianzanalysen gibt es auch multivariate Verfahren für Modelle mit mehreren abhängigen Variablen, die dann „MANOVA“ nach „Multivariate Analysis of Variance“ heißen.

Latent Class Analysis (LCA)

Skaliert man einen Test auf der Basis des *RM*, dann möchte man eine latente Personeneigenschaft wie z. B. *Mathematikleistung* quantifizieren. Es gibt aber auch latente Personeneigenschaften, die eher qualitativer Natur sind, z. B. können klassische Raumvorstellungsaufgaben mit Würfeln eher „analytisch“ oder eher „holistisch“ gelöst werden (vgl. Kap. 3.3.3). Die *LCA* ist ein *IRT*-Modell mit dem Versuchspersonen aufgrund ihres Testverhaltens (Lösungsprofile) entsprechend geclustert werden können (vgl. J. Rost, 2004, Kap. 3.1.2.2). Da der Schluss vom Testverhalten auf die zugrundeliegende qualitative la-

[98] Wäre dies der Fall, dann wäre der Test zwar ein absolut zuverlässiger Indikator für das Geschlecht der Versuchsperson, aber sicherlich kein Leistungstest für irgendeinen Bereich.

tente Personeneigenschaft immer mit Unsicherheit behaftet ist, werden die Personen den Klassen nicht deterministisch, sondern immer nur mit einer gewissen Wahrscheinlichkeit zugeordnet. Innerhalb jeder Klasse ist das (aufgrund der Klassenzugehörigkeit) prognostizierte Testverhalten für alle Versuchspersonen identisch.

Da die fragliche qualitative Personeneigenschaft *latent* ist, sind die folgenden drei zentralen Größen bei einer *LCA* prinzipiell unbekannt: *Anzahl* der latenten Klassen, *Größe* der latenten Klassen und *Wahrscheinlichkeit der Zugehörigkeit* der Personen zu den latenten Klassen. Vor der Durchführung einer *LCA* muss die *Anzahl* der latenten Klassen (nach Möglichkeit theoriegeleitet) als *Modellannahme* festgelegt werden. Anschließend können dann die *Größen* der latenten Klassen und die *Wahrscheinlichkeiten der Zugehörigkeit* der Personen zu den latenten Klassen als *Modellparameter* geschätzt werden.

Wenn vorab keine hinreichende Sicherheit bezüglich der *Anzahl* der Klassen besteht, kann die *LCA* nacheinander für verschiedene *Anzahlen* von Klassen durchgeführt werden, wobei anschließend die Güte der jeweiligen Modellanpassung an die beobachteten Daten überprüft wird (vgl. Kap. 4.2.3). So kann die *LCA* weitgehend explorativ eingesetzt werden. Liegen allerdings vielen plausible Annahmen vor, so kann neben der *Anzahl* der Klassen auch z. B. die *Größe* der Klassen oder das prognostizierte Testverhalten innerhalb der einzelnen Klassen vorab durch Parameterfestlegungen vorgegeben werden. Werden diese Annahmen dann über Modellanpassungstests abgesichert, verwendet man die LCA eher als strukturprüfendes Verfahren.

Mixed Rasch-Modell (MRM)

Wenn man im oben gewählten Beispiel mit den Würfelaufgaben davon ausgeht, dass es sowohl unterschiedlich leistungsstarke „Analytiker“ als auch unterschiedlich leistungsstarke „Holistiker“ gibt, sich die Versuchspersonen innerhalb der Klassen hinsichtlich ihrer Testleistung unterscheiden können, dann ist das *MRM* ein geeignetes Testmodell für die Auswertung der Daten. Das *MRM* ist ein gemeinsames Obermodell von *LCA* und *RM*. Mit dem *MRM* können die Versuchspersonen qualitativ (*LCA*) und innerhalb ihrer Klassen quantitativ (*RM*) unterschieden werden (vgl. Rost, 2004, Kap. 3.1.3.1).

Differential Item Functioning (DIF)

Eine im Bereich der Leistungstests wichtige Frage ist, ob einzelne Testaufgaben bestimmte Teilpopulationen systematisch bevorzugen oder benachteiligen und somit die Ergebnisse verzerren. Diese Frage der Verzerrung (engl. „bias“) wird dann besonders relevant, wenn Konsequenzen, wie Zertifizierung oder Zulassung, direkt an die Testleistung gekoppelt sind. Im Rahmen von internationalen Vergleichsstudien ist besonders wichtig, dass einzelnen Teilnehmerstaaten nicht aufgrund bestimmter Aufgabenstellungen besser oder schlechter abschneiden. Das Testinstrument soll schließlich die Staaten z. B. bezüglich der *Mathematikleistung* ihrer Schülerinnen und Schüler vergleichen, aber nicht bezüglich der Fra-

ge, wie vertraut die Schülerinnen und Schüler mit bestimmten Aufgabenkontexten oder -formulierungen sind (vgl. Baumert, Bos & Watermann, 2000, Kap. IV.3).

Aufbauend auf einer *IRT*-Skalierung der Testdaten besteht die Kernidee von *DIF*-Analysen darin, dass für Gruppen von Versuchspersonen (z. B. getrennt nach *Geschlecht*) das erwartete Testverhalten zu einzelnen Aufgaben oder Aufgabengruppen aufgrund der jeweiligen Leistungen im gesamten Test berechnet wird. Dieses erwartete Testverhalten wird dann in Relation zum tatsächlichen Testverhalten betrachtet. So lassen sich Aufgaben bzw. Aufgabengruppen identifizieren, bei denen eine Gruppe unerwartet gut oder unerwartet schlecht abschneidet. An diese statistische Analyse sollte dann eine theoretische Aufgabenanalyse anschließen, die ggf. klären kann, ob die Lösung der betroffenen Aufgabe(n) von Faktoren abhängt, die (a) nicht direkt auf die eigentlich zu messenden latente Personeneigenschaft zurückgeführt werden können und (b) ggf. bestimmte Gruppen systematisch benachteiligen.

Strukturgleichungsmodelle

Wenn man, z. B. im Bereich der Schulleistungsstudien, schon Annahmen über Wirkungsgefüge hat, also etwa theoretisch begründet Modelle zur Erklärung von *Mathematikleistung* wie in Abb. 2.7 (S. 38) oder Abb. 4.2 (S. 122) postuliert, so verfügt man über *Strukturmodelle*, deren statistische Erklärungskraft für real beobachtetes Testverhalten empirisch geprüft werden kann. Wenn in solchen Modellen neben manifesten (beobachtbaren) Variablen wie *Geschlecht* auch latente (nicht beobachtbare) Variablen wie *Mathematikleistung* eine Rolle spielen, dann muss auch spezifiziert werden, wie diese latenten Variablen gemessen werden können. Im Fall der *Mathematikleistung* kann dies z. B. über Testaufgaben (als Indikatoren für Leistung) geschehen, zu denen das Lösungsverhalten von Versuchspersonen beobachtet werden kann. Die Messung der *Mathematikleistung* geschieht dann durch statistische Rückführung (Regression) des beobachteten Testverhaltens auf diese latente Variable. Auf diesem Weg hat man ein *Messmodell* für *Mathematikleistung* spezifiziert (vgl. Brachinger & Ost, 1996; Backhaus et al., 2008, Kap. 11).

Neben der begrifflichen Unterscheidung von *Messmodell* und *Strukturmodell* unterscheidet man im *Strukturmodell* noch unabhängige Variablen, deren Verteilungen nicht innerhalb des Modells erklärt werden („exogene Variablen“), und abhängige Variablen, deren Verteilungen innerhalb des Modells erklärt werden sollen („endogene Variablen“). Auf der Basis von theoretisch angenommenen Wirkungsgefügen, die etwa wie in Abb. 4.2 (S. 122) veranschaulicht und noch um die Messmodelle für die latenten Variablen ergänzt werden, lassen sich Strukturgleichungsmodelle dann algebraisieren und algorithmisch analysieren.

Mit Strukturgleichungsmodellen lassen sich z. B. Regressionsanalysen, Pfadanalysen oder konfirmatorische Faktorenanalysen (CFA) realisieren. Während bei Regressionsanalysen die Verteilung einer endogenen auf eine oder mehrere exogene Variablen zurückgeführt werden soll, betrachten Pfadanalysen Wirkungsgefüge, in denen mehrere endogene Vari-

ablen enthalten sein können; Pfadanalysen im engeren Sinne enthalten dabei nur manifeste Variablen. Bei einer *CFA* werden die Korrelationen mehrerer vorab festgelegter latenter Faktoren untereinander untersucht. Dies geschieht z. B. im mehrdimensionalen *RM*. Die Messmodelle eines Strukturgleichungsmodells führen immer auf eine *CFA*, da die Messung eine Rückführung der Indikatoren auf die latenten Variablen darstellt, wobei die *Faktorladungen* (Trennschärfe des jeweiligen Indikators in Bezug auf den Faktor) geschätzt werden.[99]

4.2.3 Einschätzung der Modellgüte

Bei stringent geplanten empirischen Untersuchungen ergeben sich aus der inhaltlichen Fragestellung die relevanten Konstrukte und häufig auch schon die Instrumente, mit denen diese Konstrukte erfasst werden können, sowie die Auswertungsstrategien. Bei manchen Untersuchungen sind vorab Festlegungen der Auswertungsmethoden erforderlich, um ein aus testökonomischen Gesichtspunkten erforderliches Design realisieren zu können, wie z. B. bei *PISA 2000* ff. das „Multi-Matrix-Design“ auf der Basis des *RM* (vgl. Carstensen et al., 2007). Dabei ist vor der Testdurchführung und -auswertung in der Regel nicht klar, ob das präferierte Testmodell auch tatsächlich gut geeignet ist, um die beobachteten Daten zu erklären. Wenn die Parameter der jeweiligen Testmodelle mithilfe der beobachteten Daten geschätzt worden sind, kann die Modellanpassung an die beobachteten Daten empirisch mit geeigneten Kennwerten beschrieben werden. Auf dieser Grundlage kann einerseits die Güte der Anpassung eines Modells an die beobachteten Daten überprüft werden und andererseits können verschiedene Modelle bezüglich ihrer Anpassung verglichen werden.

Einschätzung der Anpassung eines Modells an die beobachteten Daten

Im Fall des *RM* werden unter anderem die Itemschwierigkeiten geschätzt, auf deren Basis anschließend die Passung der theoretisch – aufgrund der Modellgleichung – erwarteten *ICCs* zu den tatsächlich beobachteten Daten beurteilt werden kann. Im Kap. 5.4.1 wird diese graphische Betrachtung exemplarisch für ein *WLT*-Item durchgeführt (Abb. 5.13, S. 169). Etwaige Abweichungen des theoretischen Modells von den empirischen Daten können mit statistischen Maßen quantifiziert werden. Diese Idee liegt auch den sogenannten „Modellgeltungstests“ zugrunde, die aus erkenntnistheoretischer und aus mathematikdidaktischer Sicht besser „Test der Anpassungsgüte eines Modells“ heißen sollten.[100]

---

[99] Im *RM* sind diese Faktorladungen durch das Testmodell (mit parallel verschobenen *ICCs*) alle vorab festgelegt und gleichgesetzt. Daran erkennt man wiederum, dass das *RM* ein recht restriktives Modell ist, das seine vorteilhaften statistischen Eigenschaften nur durch erhebliche Reduktionen erreicht.

[100] Aus der Sicht des mathematischen Modellierens ist ein Modell nicht „richtig“ oder „falsch“, sondern es kann seine Zwecke mehr oder weniger gut erfüllen. Neben der möglichst guten (reduzierten) Beschreibung realer Gegebenheiten können z. B. auch – je nach Verwendungszweck – seine Einfachheit oder seine alge-

Dabei wird im Falle des *RM* nicht nur eine *ICC* betrachtet, sondern die gesamte Datenmatrix in Form der verschieden Antwortvektoren („Pattern“) und ihrer Häufigkeiten („Patternhäufigkeiten“). Auf der Basis des angenommenen Testmodells und der für dieses Modell geschätzten Parameter können die *erwarteten* Patternhäufigkeiten berechnet und mit den *beobachteten* Daten verglichen werden. Für die Pattern $\vec{x}$ mit den *beobachteten* Häufigkeiten $b_{\vec{x}}$ und den *erwarteten* Häufigkeiten $e_{\vec{x}}$ ist die Summe $\sum_{\vec{x}} \frac{\left(b_{\vec{x}} - e_{\vec{x}}\right)^2}{e_{\vec{x}}}$ eine asymptotisch χ^2-verteilte Prüfgröße (vgl. Büchter & Henn, 2007, Kap. 4.2.4), die im Falle des zweikategoriellen *RMs* für m Items und n_p unabhängige Modellparameter $m^2–n_p–1$ Freiheitsgrade hat.

Mit diesem χ^2-Test kann – unter bestimmten Voraussetzungen – geprüft werden, ob das durch die geschätzten Parameter konkretisierte Modell signifikant von den beobachteten Daten abweicht. Die Bedingung für eine asymptotische χ^2-Verteilung der Prüfgröße ist, dass die *erwarteten* Patternhäufigkeiten größer als Null sind. Wenn der Anteil der Pattern, die tatsächlich keinmal *beobachtet* werden konnten, zu groß ist, ist die Verteilung der Prüfgröße unbekannt. Dies ist schnell der Fall, da schon ein Test mit nur zehn dichotomen Items $2^{10} = 1\,024$ mögliche Pattern hat. In einem solchen Fall kann man die unbekannte Verteilung aber simulieren und auf der Basis der simulierten Verteilung den Test durchführen („Bootstrapping“, vgl. J. Rost, 2004, S. 336 ff.).

Vergleich mehrerer Modelle bezüglich der Güte der Anpassung

Wenn nicht nur ein Modell, sondern mehrere verschiedene Modelle theoretisch geeignet sind, die beobachteten Daten zu erklären, dann kann die Anpassungsgüte dieser Modelle miteinander verglichen werden. Erfüllen die Modelle, die miteinander verglichen werden sollen, bestimmte Voraussetzungen, so kann der Vergleich in Form eines einfachen Hypothesentests auf der Basis der χ^2-Verteilung durchgeführt werden. Alternativ stehen informationstheoretische Maße zur Verfügung, die an die Idee dieses χ^2-Tests anschließen.

Liegen mehrere konkurrierende Modelle mit theoretischer Plausibilität vor, dann kann die *Maximum-Likelihood-Idee* genutzt werden, die zuvor schon bei der Parameterschätzung zum Einsatz kommt: Aus Sicht der mathematischen Statistik steht bei der Parameterschätzung die Frage im Vordergrund, für welche Parameterwerte die gegebenen Daten die höchste Wahrscheinlichkeit des Auftretens haben (vgl. Büchter & Henn, 2007, Kap. 4.1.2). Dementsprechend kann man sich fragen, für welches konkretisierte Modell die beobachteten Daten (in Form der Datenmatrix oder der Varianz-Kovarianz-Matrix der manifesten Variablen) die höchste Wahrscheinlichkeit des Auftretens haben. Bei einem Vergleich der *Likelihoods* zweier konkretisierter Modelle muss allerdings beachtet werden, dass Modelle

braischen Eigenschaften wichtig sein. In der psychologischen Methodenlehre ist der Begriff „Modellgeltungstest“ aber üblich (vgl. J. Rost, 2004, Kap. 5).

mit vielen Parametern einerseits in der Regel flexibler sind und sich den Daten besser anpassen können, andererseits aber so komplex sind, dass sie inhaltlich kaum beschrieben werden können. Daher muss bei Modellvergleichen auch das *Kriterium der Sparsamkeit* berücksichtigt werden.

Wenn ein relativ „sparsames" Modell M_0 aus einem weniger „sparsamen" Modell M_1 durch Restriktionen hervorgeht,[101] dann kann man aus dem Quotienten der *Likelihoods* – unter bestimmten Voraussetzungen – eine asymptotisch χ^2-verteilte Prüfgröße gewinnen, die genauso viele Freiheitsgrade hat, wie das Modell M_1 mehr Parameter als das Modell M_0 hat. Dieser χ^2-Test heißt auch „Likelihood Ratio Test (LRT)", da er vom Quotienten $LR = \frac{L(M_0)}{L(M_1)}$ ausgehend die Prüfgröße

$$-2 \cdot \ln(LR) = -2 \cdot (\ln(L(M_0)) - \ln(L(M_1))) = -2 \cdot \ln(L(M_0)) - (-2 \cdot \ln(L(M_1)))$$

gewinnt, die asymptotisch χ^2-verteilt ist, wenn – wie beim obigen χ^2-Test für ein Modell – die *erwarteten* Patternhäufigkeiten größer als Null sind.

Da die Voraussetzungen für den *LRT* häufig nicht erfüllt sind, gibt es in der Literatur eine Vielzahl „informationstheoretischer Maße für den Modellvergleich" (J. Rost; 2004, Kap. 5.2), deren bekanntesten Vertreter (*AIC* und *BIC*) im Folgenden definiert werden. Diese Maße gehen zwar von den Größen $-2 \cdot \ln(L(M_i))$ aus, die auch beim *LRT* eine zentrale Rolle spielen, ihre Anwendung ist aber nicht an Voraussetzungen bezüglich Patternhäufigkeiten gebunden. Ihr Vorteil besteht darüber hinaus darin, dass die Modelle die verglichen werden sollen, nicht zwingend hierarchisch angeordnet sein müssen; ein Modell muss aus einem anderen also nicht zwingend durch Restriktionen hervorgehen. Neben der oben genannten Größe verwenden die beiden hier betrachteten Maße noch die Anzahl der Parameter n_p des jeweiligen Modells (und berücksichtigen dadurch das *Kriterium der Sparsamkeit*) sowie ggf. die Stichprobengröße N:

$$AIC(M_i) = -2 \cdot \ln(L(M_i)) + 2 \cdot n_p$$
$$BIC(M_i) = -2 \cdot \ln(L(M_i)) + \ln(N) \cdot n_p \,.$$

Beim Vergleich mehrerer konkurrierender Modelle gilt: Je kleiner die *AIC*- bzw. *BIC*-Werte sind, desto besser passt sich das Modell – im Sinne dieser informationstheoretischen Maße – an die beobachteten Daten an. Arithmetisch ist dies gut nachvollziehbar, da eine hohe Wahrscheinlichkeit des Auftretens der Daten bedeutet, dass die *Likelihood* $L(M_i)$ größer und die Größe $-2 \cdot \ln(L(M_i))$ somit kleiner wird. Die Parameterzahl geht „bestrafend" in die informationstheoretischen Maße ein.

[101] Dabei muss allerdings die stark einschränkende Bedingung erfüllt sein, dass die Restriktion nicht dadurch erfolgen darf, dass ein Parameter gleich Null gesetzt wird (vgl. J. Rost, 2004, S. 332).

Der *AIC* orientiert sich noch eng an der Idee der χ^2-verteilten Prüfgröße, bei der mit jedem Freiheitsgrad mehr, also jedem unabhängigen Parameter mehr, der kritische Wert ungefähr um 2 größer wird. Demgegenüber gewichtet der *BIC* die Parameterzahl für Stichprobengrößen ab $N = 8$ stärker und zwar in Abhängigkeit von der Stichprobengröße. Dadurch wird berücksichtigt, dass in größeren Stichproben auch mehr unterwartetes Antwortverhalten auftreten kann, was durch mehr Modellparameter besser erfasst werden kann, ohne dass dadurch die theoretische Plausibilität des Modells steigt. Das *Kriterium der Sparsamkeit* wird beim *BIC* also besonders berücksichtigt. Die Frage, welches der beiden Maße wann zum Einsatz kommen soll, beantwortet J. Rost (2004, S. 344): „Als grobes Auswahlkriterium kann gelten, dass der AIC bei kleinen Itemzahlen mit großen Patternhäufigkeiten, der BIC bei großen Itemzahlen mit kleinen Patternhäufigkeiten vorzuziehen ist."

Die Parameterzahl eines konkretisierten Modells hängt z. B. im Falle des *RM* nicht nur von der Anzahl der Itemschwierigkeiten, die geschätzt werden müssen, sondern auch von der Modellierung der Personenverteilung ab, da die gesamte Datenmatrix erklärt werden soll. Bei m dichotomen Items sind $m+1$ unterschiedliche Gesamtscores (Anzahl der richtigen Lösungen) möglich. Durch die Annahme einer Normalverteilung oder einer anderen geglätteten Verteilung lässt sich die Zahl der zu schätzenden Verteilungsparameter von $m+1$ auf 2 verringern. Durch eine Glättung der Verteilung entfernt sich das Modell zwar in der Regel etwas von den beobachteten Daten, wird dafür aber inhaltlich besser beschreibbar. Die sinkende *Likelihood* wird bei den informationstheoretischen Maßen durch die geringere Parameterzahl in der Regel gut kompensiert, solange die Glättung nicht zu erheblichen Abweichungen von den Daten führt.

Neben den hier vorgestellten Tests und Maßen für die Einschätzung der Modellgüte gibt es noch weitere wichtige Kriterien, vor allen den *Geltungsbereich einer Theorie* und die *Brauchbarkeit einer Theorie*. Da diese Kriterien bei den Modellen, die in der vorliegenden Arbeit verwendet werden, gleichermaßen erfüllt sind, wird hierauf nicht weiter eingegangen; entsprechende Ausführungen findet man bei J. Rost (2004, Kap. 5).

4.3 Grobplanung der Untersuchung

Da die vorliegende Arbeit ein Einzelvorhaben ohne Anschluss an ein größeres Forschungsprojekt ist, musste insbesondere bei der Untersuchungsplanung den begrenzten Ressourcen Rechnung getragen werden. Ausgehend von der präzisierten Fragestellung und den möglichen Testinstrumenten mussten pragmatische Entscheidungen getroffen werden, die sicherstellen, dass vor allem zum Kernbereich der Arbeit, also zur *Spatial Mediation Hypothesis*, belastbare Befunde erzielt werden können.

4.3.1 Anforderungen an die Stichproben

Der Zusammenhang von *Raumvorstellung* und *Mathematikleistung* soll unter besonderer Berücksichtigung von (möglichen) Geschlechterunterschieden und unter Einbeziehung weiterer Variablen (*FSK:M*, *Denkstile*) bei Schülerinnen und Schülern des 9. Jahrgangs untersucht werden. Dabei werden keine verallgemeinerbaren Aussagen zum Leistungsniveau dieser Population in Nordrhein-Westfalen oder Deutschland, sondern Zusammenhangsaussagen für die fraglichen Bereiche angestrebt. Für verallgemeinerbare Aussagen zum Leistungsniveau gibt es große und aussagekräftige Systemmonitoringstudien. Die Ziehung und Realisierung von Stichproben, die für solche Zwecke erforderlich sind, sprengen den Rahmen eines Einzelvorhabens.

Für die vorliegende Arbeit ist also lediglich wichtig, dass die Varianz der Testleistungen in den berücksichtigten Bereichen möglichst unverzerrt ist. Wäre die Varianz in einigen Testbereichen viel zu gering, dann würde eine Unterschätzung etwaiger Zusammenhänge damit einhergehen. Wäre die Varianz dagegen atypisch groß, dann würden Zusammenhänge tendenziell überschätzt. Eine möglichst typische Varianz in den fraglichen Testbereichen lässt sich z. B. sicherstellen, wenn man mehrere typische Schulen aus den Schulformen der Sekundarstufe I berücksichtigt und hier jeweils den gesamten 9. Jahrgang in die Untersuchung einbezieht. Damit hat man notwendig Klumpenstichproben und gewisse Effekte, die auf Schulformen (und Einzelschulen) als differenzielle Entwicklungsmilieus zurückgehen. Solche Effekte können aber zum einen statistisch kontrolliert werden und dürften zum anderen auf die Zusammenhangsanalysen keine (zu) große Wirkung haben.

4.3.2 Zeitplan für die Erhebungen

Die in Kap. 4.1.3 vorläufig ausgewählten Instrumente sind zusammen noch zu umfangreich, um in einer Untersuchung im Anschluss an die *LSE 9* eingesetzt zu werden. Außerdem steht nicht a priori fest, dass alle Instrumente für den Einsatz in der fragliche Zielgruppe geeignet sind. Daher wurde im Rahmen einer Voruntersuchung überprüft, wie sich die Testinstrumente bewähren und welche für die zentralen Fragen der vorliegenden Arbeit besonders ergiebig sind. Auf dieser empirischen Basis konnte eine zielgerechte Hauptuntersuchung geplant werden.

Da in der Hauptuntersuchung Zusammenhangsaussagen gewonnen werden sollen, ist es erforderlich, dass nicht zuviel Zeit zwischen der Erhebung der *Mathematikleistung* (*LSE 9*) und der Erhebung der *Raumvorstellung* sowie der weiteren berücksichtigten Bereiche liegt. Die *LSE 9*-Mathematik fand am 09.11.2004 statt. Die eigene Erhebung fand Ende November 2004 also binnen drei Wochen nach den *LSE 9*-Mathematik statt. Damit genügend Zeit zur Auswertung der Voruntersuchung blieb, ist diese direkt vor den Sommerferien 2003 im Juli durchgeführt worden.

4.3.3 Grobplanung der Voruntersuchung

Im Rahmen der Voruntersuchung sollte einerseits eine geeignete Auswahl aus dem Überangebot an Raumvorstellungstests getroffen werden und andererseits sollte überprüft werden, ob sich die *Denkstile* in einem schriftlichen Test hinreichend gut erfassen lassen. Die Anforderungen an die Stichprobe der Voruntersuchung sind also bezüglich der Varianz der Testleistungen weniger stark als in der Hauptuntersuchung. Für die Voruntersuchung konnten zwei mittelgroße Gymnasien und eine größere Gesamtschule gewonnen werden, in denen jeweils der komplette 9. Jahrgang untersucht wurde.

Der Zusammenhang von *Raumvorstellung* und *Mathematikleistung* wird vor allem in der Hauptuntersuchung analysiert. In der Voruntersuchung sollte *Mathematikleistung* aber zumindest über die Mathematiknoten der Schülerinnen und Schüler als externes Kriterium berücksichtigt werden, um Plausibilitätsbetrachtungen durchführen zu können. Dabei ist klar, dass Schulnoten nicht objektiv und kriterial vergeben werden, sondern vor allem mit Blick auf die Leistung eines Schülers bzw. einer Schülerin im Vergleich zum Rest der Lerngruppe.

4.3.4. Grobplanung der Hauptuntersuchung

Bei der Hauptuntersuchung mussten die oben genannten Anforderungen an die Stichprobe berücksichtigt werden. Aufgrund der Befunde von *PISA 2000* zur Verteilung der *Mathematikleistung* in den unterschiedlichen Schulformen (vgl. Klieme et al., 2001, S. 180) kann man davon ausgehen, dass die Varianz der *Mathematikleistung* hinreichend gut abgebildet wird, wenn alle Schulformen etwa in gleichem Umfang in der Stichprobe vertreten sind. Daher wurden für die Hauptuntersuchung jeweils eine Gesamt-, Haupt- und Realschule sowie ein Gymnasium für die Teilnahme an der Untersuchung gewonnen. Da es sich bei allen vier Schulen um typische Vertreterinnen ihrer Schulform handelt, ist von einer atypischen Verzerrung der Varianzen der Testleistungen nicht auszugehen.

Aufgrund der zeitlichen Belastung des Mathematikunterrichts durch die *LSE 9* wurde die eigene Erhebung so geplant, dass sie in einer üblichen Schulstunde, also in 45 Minuten, durchgeführt werden konnte. Zentral für die Hauptuntersuchung war, dass die Datensätze der eigenen Erhebung für jeden Schüler und jede Schülerin mit den Datensätzen der *LSE 9* verknüpft werden konnte.

5 Anlage und Befunde der Voruntersuchung

Die Voruntersuchung begann 2003 in der ersten Jahreshälfte mit der Vorbereitung der Datenerhebung, die im Juli stattfand. Anschließend wurden die Daten so ausgewertet, dass die eigenen Erhebungsinstrumente für die Hauptuntersuchung optimiert werden konnten. Inhaltlich können auf der Basis der Kapitel 2 und 3 sowie der Voruntersuchung Hypothesen für die Hauptuntersuchung formuliert werden.

Im Folgenden wird zunächst die Zielsetzung der Voruntersuchung differenzierter dargestellt, anschließend werden die Erhebungsinstrumente jeweils mit Beispielitems vorgestellt und die Durchführung und Auswertung der Daten dokumentiert. Schließlich werden die Befunde der Voruntersuchung in ihrer vorbereitenden Funktion für die Hauptuntersuchung zusammengefasst.

5.1 Zielsetzung der Voruntersuchung

Die in Kap. 4.1.3 vorläufig ausgewählten Erhebungsinstrumente würden vermutlich auf eine Durchführungsdauer von deutlich mehr als zwei Schulstunden (90 Minuten) hinauslaufen. Da aus den bereits genannten Gründen die Hauptuntersuchung aber nicht länger als eine Schulstunde (45 Minuten) sein soll, muss im Rahmen der Voruntersuchung mit Blick auf die zugrundeliegenden Fragen geklärt werden, welche Instrumente in welchem Umfang eingesetzt werden sollen. Außerdem soll die Hauptuntersuchung mit den Befunden der Voruntersuchung auch inhaltlich vorbereitet werden. Insgesamt lässt sich die Zielsetzung der Voruntersuchung wie folgt ausdifferenzieren:

- Die vorläufig ausgewählten Instrumente zur *Raumvorstellung* müssen sich zunächst prinzipiell in einer konkreten Untersuchung mit Schülerinnen und Schülern des 9. Jahrgangs bewähren. Hierbei stehen auch die Testinstruktionen, die Präsentation der Items im Testheft und ähnliche Fragen auf dem Prüfstand. Anhand der erhobenen Daten kann eingeschätzt werden, wie sich die Instrumente zur *Raumvorstellung* zueinander verhalten, z. B. ob die verschiedenen Komponenten der *Raumvorstellung* mit den ausgewählten Referenztests auch empirisch voneinander getrennt werden können. Anschließend kann auf empirischer Basis entschieden werden, welche Instrumente in der Hauptuntersuchung zum Einsatz kommen sollen und ob ggf. Kurzformen der Tests entwickelt werden können.
- In der Voruntersuchung sollen im Bereich *Raumvorstellung* auch die Tests *2 DW* und *3 DW* eingesetzt werden, da auf der Basis einer Replikation der Studien von Hosenfeld et al. (1997) und Köller et al. (1994) eine Konstruktvalidierung für den *MRT* durchgeführt werden soll (vgl. Kap. 4.1.3). Da sowohl der *2 DW* als auch der *3 DW* (in unterschiedlichen Anteilen) *mentale Rotation* und zugleich *räumliche Visualisierung* erfassen, werden diese beiden Tests in der Hauptuntersuchung vermutlich nicht zum Einsatz

kommen. Die anderen vorläufig ausgewählten Tests sind besser auf die berücksichtigten Komponenten der *Raumvorstellung* abgestimmt und eignen sich jeweils als nahezu idealtypische Referenztests.

- Im Bereich der *Denkstile* soll auf der Basis vorhandenen Diagnosematerials (Schwank, 1999/2000) ein schriftlicher Test entwickelt werden. Dieser Test muss im Rahmen der Voruntersuchung erprobt werden. Dabei ist zunächst offen, ob sich ein geeignetes Instrument für die Hauptuntersuchung entwickeln lässt.
- Für das *FSK:M* müssen noch konkrete Items ausgewählt und erprobt werden. Falls sich diese Items empirisch nicht bewähren, muss in der Hauptuntersuchung auf eine andere Skala zur Messung des *FSK:M* zurückgegriffen werden.
- Bei der Auswertung der Testdaten der Voruntersuchung kann jeweils überprüft werden, mit welchen Testmodellen und statistischen Verfahren die Daten angemessen skaliert und ausgewertet werden können. Entsprechende Analysen dienen u. a. der forschungsmethodischen Vorbereitung der Hauptuntersuchung. Dabei ist es möglich, dass z. B. die Datengrundlage – auch in der Hauptuntersuchung – nicht für alle Kap. 4.2.2 betrachteten Verfahren bzw. Modelle ausreicht.
- Auch wenn die Voruntersuchung primär der Zusammenstellung und Optimierung eines Erhebungsinstruments für die Hauptuntersuchung dient, liefert sie selbst schon erste inhaltliche Ergebnisse. So können die Daten der Voruntersuchung z. B. bezüglich etwaiger Geschlechterunterschiede in der *Raumvorstellung* untersucht werden. Darüber hinaus werden die Ergebnisse plausibilisiert, indem als externe Kriterien einige Schulnoten der Schülerinnen und Schüler hinzugezogen werden und auf dieser Basis die eigenen Ergebnisse ins Verhältnis zu den in der Literatur berichteten Befunden gesetzt werden können. Auf dieser Basis können die Hypothesen für die Hauptuntersuchung, die sich zunächst aus der Literatur ergeben, ggf. weiter präzisiert oder ausdifferenziert werden.

5.2 Instrumente der Voruntersuchung

Im Folgenden werden die einzelnen Erhebungsinstrumente der Voruntersuchung vorgestellt. Für die Tests zur *Raumvorstellung* und zu den *Denkstilen* werden jeweils einige Rahmendaten (zugrundeliegendes Konstrukt, Anzahl der Beispielitems, Anzahl der Testitems, Anzahl der Testseiten, geschätzte Bearbeitungsdauer), die schriftliche Testinstruktion und ein Beispiel für ein Testitem angegeben. Darüber hinaus wird ein Ausschnitt des Deckblatts mit der Skala für das *FSK:M* und weiteren Erhebungsfragen vorgestellt.

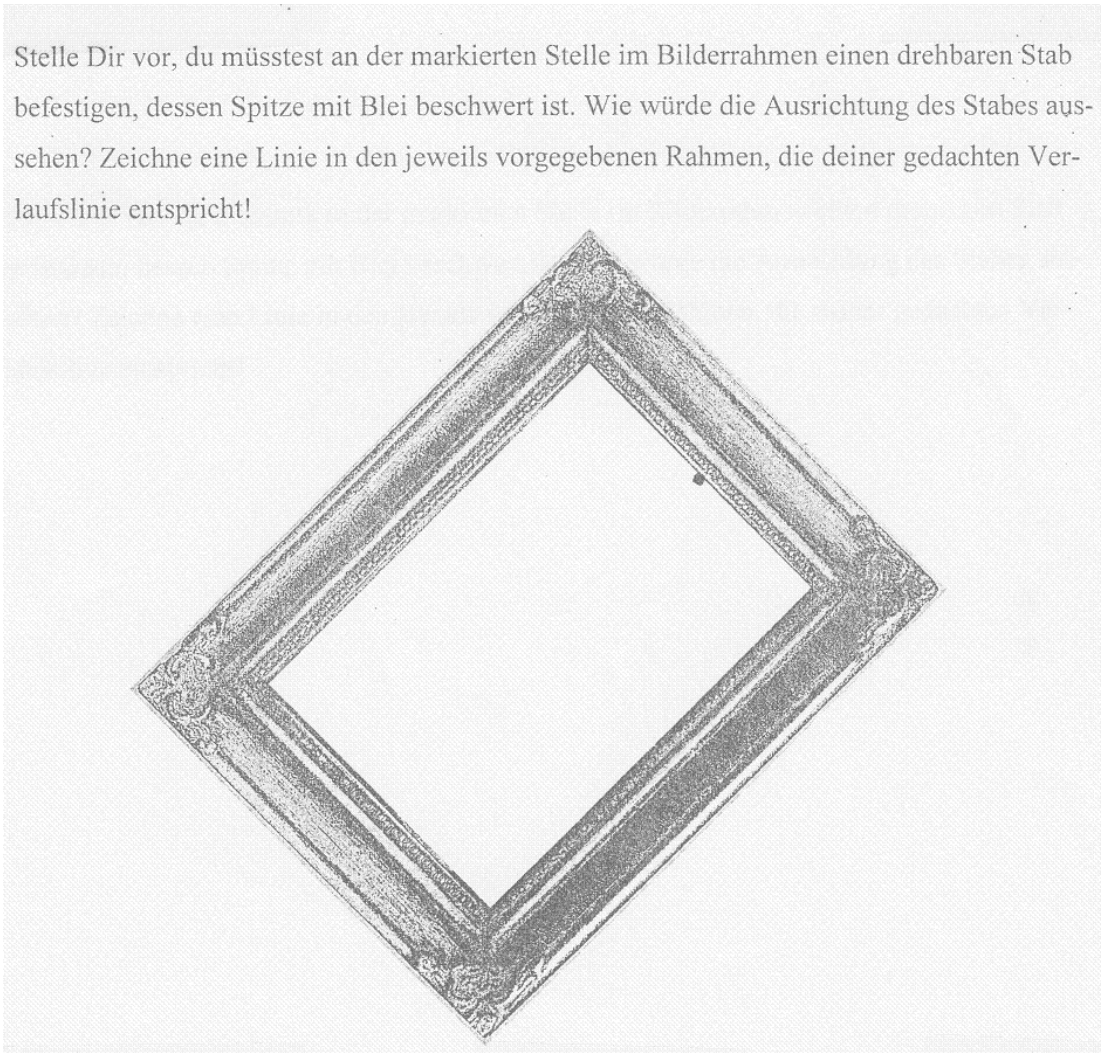
5.2.1 Instrumente zur Raumvorstellung

Die Auswahl der folgenden Tests basiert auf der inhaltlichen Begründung in Kap. 4.1.3.

Rod and Frame (RFT)

- Konstrukt: *Raumvorstellung – räumliche Wahrnehmung*
- Anzahl der Beispielitems: 0
- Anzahl der Items: 3
- Anzahl der Seiten: 1,5
- Geschätzte Bearbeitungsdauer: 2 Minuten

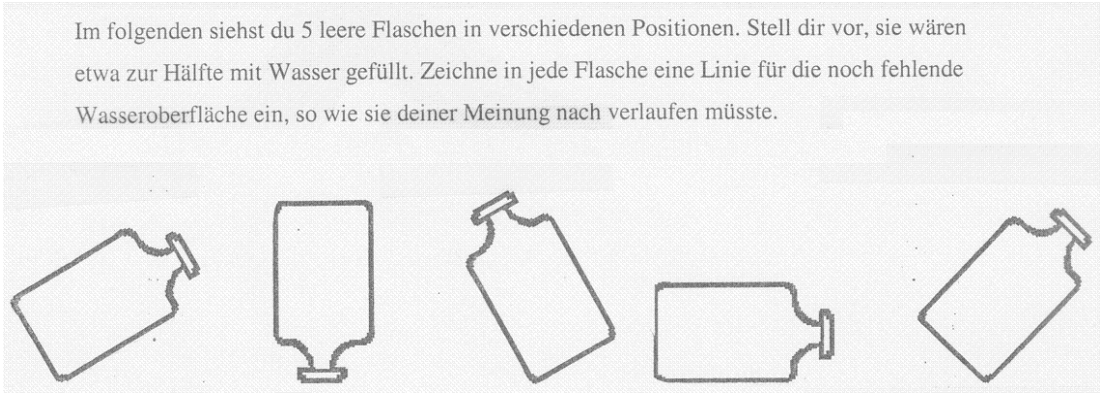
Abbildung 5.1: RFT – Testinstruktion und Beispiel für ein Testitem



Water Level Tasks (WLT)

- Konstrukt: *Raumvorstellung – räumliche Wahrnehmung*
- Anzahl der Beispielitems: 0
- Anzahl der Items: 5
- Anzahl der Seiten: 0,5
- Geschätzte Bearbeitungsdauer: 3 Minuten

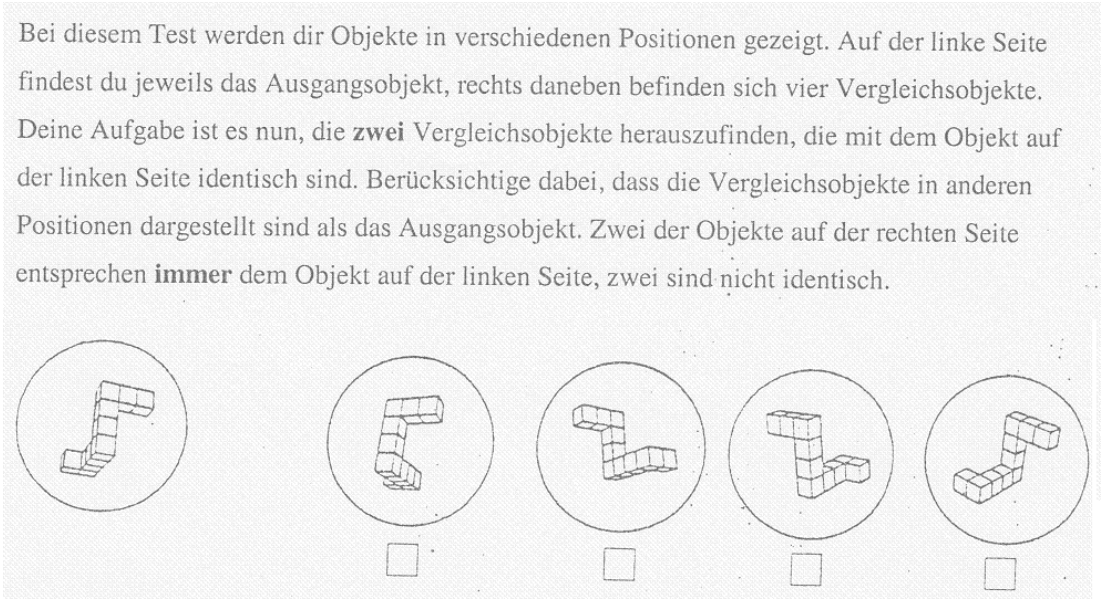
Abbildung 5.2: WLT – Testinstruktion und Testitems



Mental Rotation Test (MRT)

- Konstrukt: *Raumvorstellung – mentale Rotation*
- Anzahl der Beispielitems: 2
- Anzahl der Items: 10
- Anzahl der Seiten: 3
- Geschätzte Bearbeitungsdauer: 20 Minuten

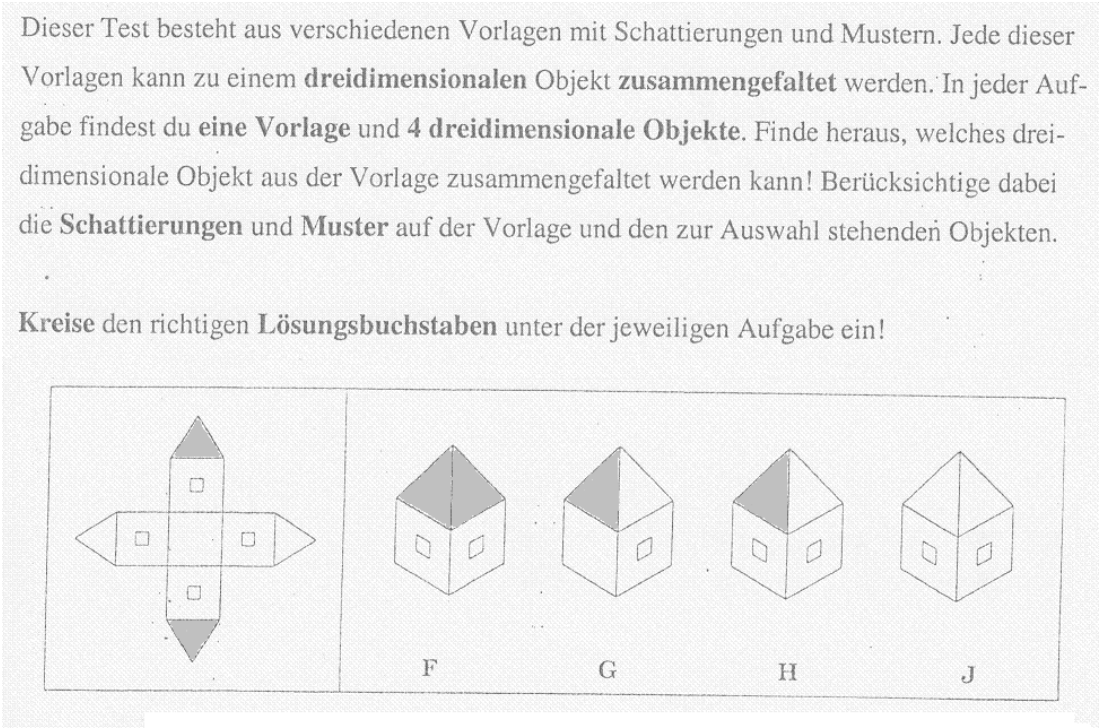
Abbildung 5.3: MRT – Testinstruktion und Beispiel für ein Testitem



Differential Aptitude Test – Subtest Spatial Relations (DAT:SR)

- Konstrukt: *Raumvorstellung – räumliche Visualisierung*
- Anzahl der Beispielitems: 2
- Anzahl der Items: 25
- Anzahl der Seiten: 6
- Geschätzte Bearbeitungsdauer: 20 Minuten

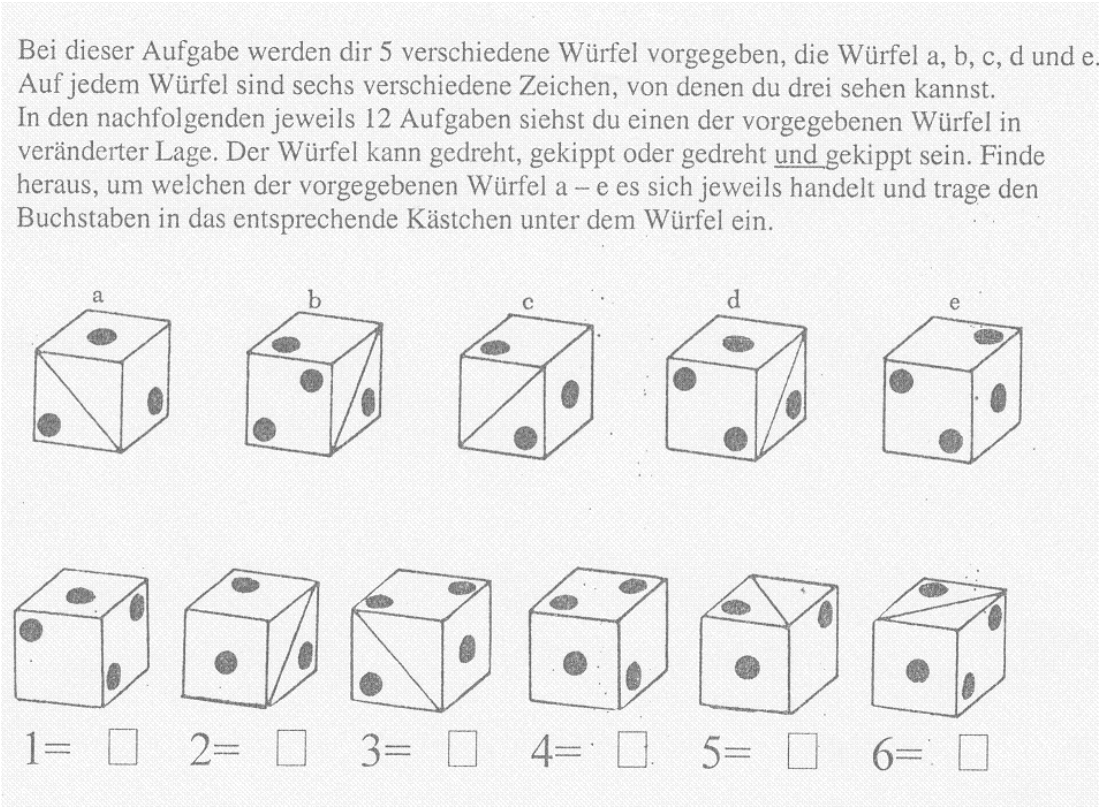
Abbildung 5.4: DAT:SR – Testinstruktion und Beispiel für ein Testitem



Zweidimensionaler Würfeltest (2 DW)

- Konstrukt: *Raumvorstellung – mentale Rotation* und *räumliche Visualisierung*
- Anzahl der Beispielitems: 0
- Anzahl der Items: 12
- Anzahl der Seiten: 1
- Geschätzte Bearbeitungsdauer: 15 Minuten

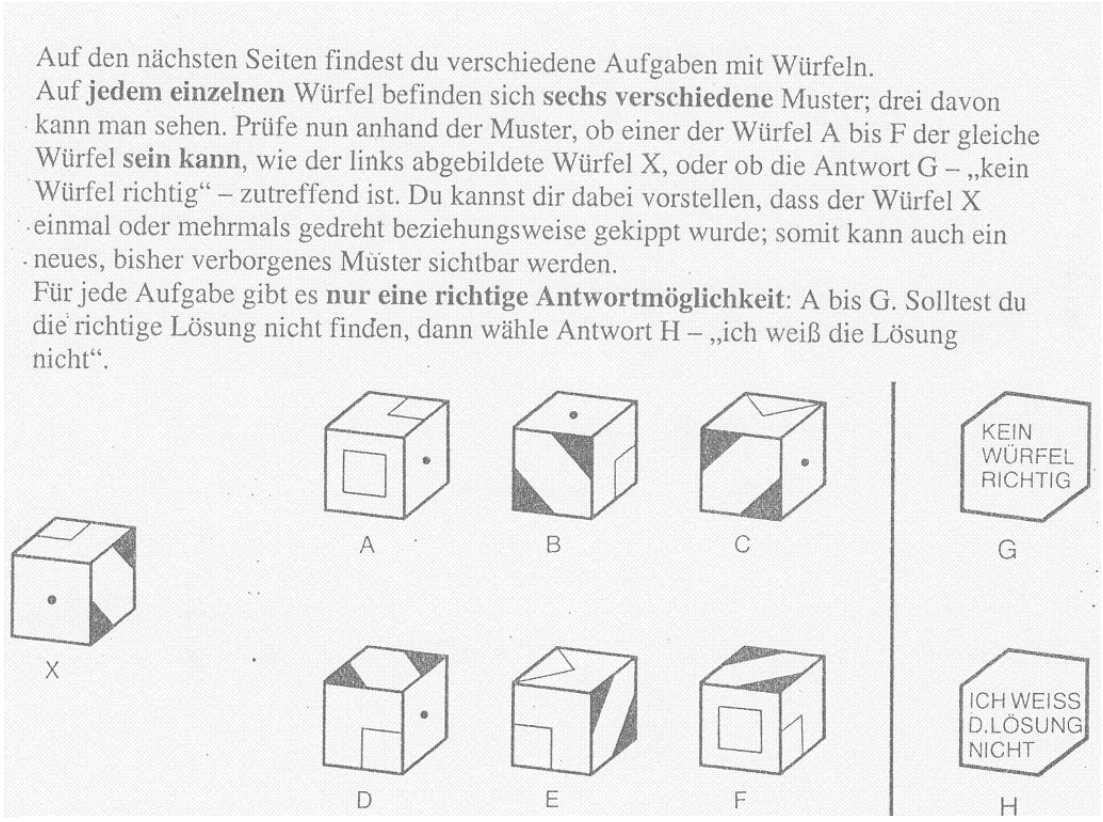
Abbildung 5.5: 2 DW – Testinstruktion und Beispiele für Testitems



Dreidimensionaler Würfeltest (3 DW)

- Konstrukt: *Raumvorstellung – mentale Rotation* und *räumliche Visualisierung*
- Anzahl der Beispielitems: 2
- Anzahl der Items: 10
- Anzahl der Seiten: 7
- Geschätzte Bearbeitungsdauer: 20 Minuten

Abbildung 5.6: 3 DW – Testinstruktion und Beispiel für ein Testitem



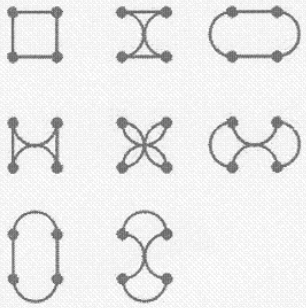
5.2.2 Instrument Denkstile

- Konstrukt: *Denkstile* (*prädikatives* vs. *funktionales Denken*)
- Anzahl der Beispielitems: 0
- Anzahl der Items: 4
- Anzahl der Seiten: 2
- Geschätzte Bearbeitungsdauer: 20 Minuten

Abbildung 5.7: Denkstile – Testinstruktion und Beispiel für ein Testitem

In den nachfolgend gezeigten Bildern fehlt jeweils unten rechts eine Figur. Suche eine möglichst sinnvolle Ergänzung und füge diese an der freien Stelle ein! Bitte notiere auf der rechten Seite kurz deine Begründung, warum du deine Lösung für sinnvoll hältst!

Aufgabe 1:

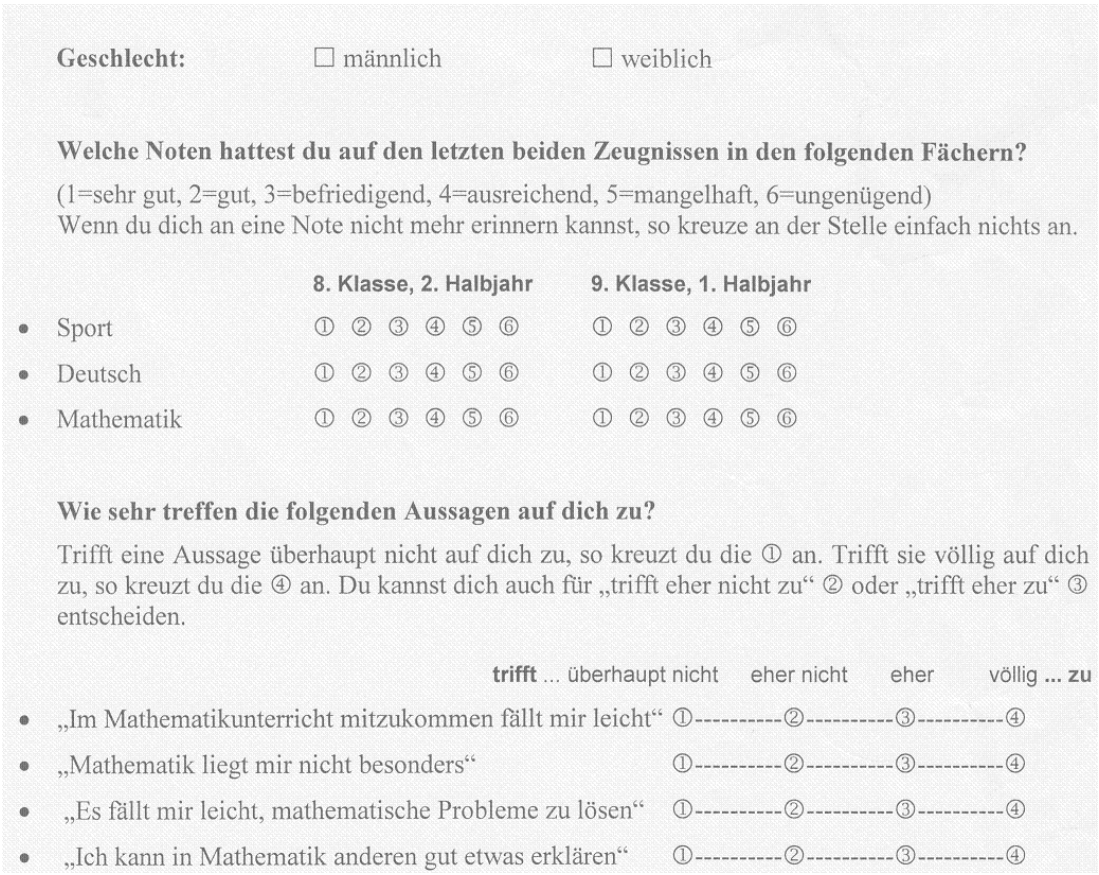


5.2.3 Weitere Instrumente

Auf dem Deckblatt der Testhefte steht zunächst eine Begrüßung mit einer kurzen Darstellung des Anliegens der Untersuchung und einigen allgemeinen Testinstruktionen. Anschließend sollen alle Schülerinnen und Schüler eine Kursnummer und eine individuelle Platznummer auf dem Deckblatt notieren. An diese Einleitung schließen die unten abgebildeten Fragen zum Geschlecht, zu ausgewählten Schulnoten und zum *FSK:M* an.

- Anzahl der Seiten: 1
- Geschätzte Bearbeitungsdauer: 10 Minuten

Abbildung 5.8: Ausschnitt vom Deckblatt der Testhefte



Geschlecht: ☐ männlich ☐ weiblich

Welche Noten hattest du auf den letzten beiden Zeugnissen in den folgenden Fächern?

(1=sehr gut, 2=gut, 3=befriedigend, 4=ausreichend, 5=mangelhaft, 6=ungenügend)
Wenn du dich an eine Note nicht mehr erinnern kannst, so kreuze an der Stelle einfach nichts an.

	8. Klasse, 2. Halbjahr	9. Klasse, 1. Halbjahr
• Sport	① ② ③ ④ ⑤ ⑥	① ② ③ ④ ⑤ ⑥
• Deutsch	① ② ③ ④ ⑤ ⑥	① ② ③ ④ ⑤ ⑥
• Mathematik	① ② ③ ④ ⑤ ⑥	① ② ③ ④ ⑤ ⑥

Wie sehr treffen die folgenden Aussagen auf dich zu?

Trifft eine Aussage überhaupt nicht auf dich zu, so kreuzt du die ① an. Trifft sie völlig auf dich zu, so kreuzt du die ④ an. Du kannst dich auch für „trifft eher nicht zu“ ② oder „trifft eher zu“ ③ entscheiden.

	trifft ... überhaupt nicht	eher nicht	eher	völlig **... zu**
• „Im Mathematikunterricht mitzukommen fällt mir leicht“	①	②	③	④
• „Mathematik liegt mir nicht besonders“	①	②	③	④
• „Es fällt mir leicht, mathematische Probleme zu lösen“	①	②	③	④
• „Ich kann in Mathematik anderen gut etwas erklären“	①	②	③	④

5.3 Durchführung und Auswertung der Voruntersuchung

Im Folgenden werden die Stichprobe, die Zusammenstellung der Erhebungsinstrumente in drei Varianten und die Verteilung dieser Varianten auf die Klassen der Stichprobe, die Durchführung der Datenerhebung, die Erfassung und Aufbreitung der Daten für die Auswertung mit entsprechenden Programmen sowie die Schritte der Auswertung der Daten dargestellt.

5.3.1 Beschreibung der Stichprobe

Die Voruntersuchung wurde an zwei Gymnasien und einer Gesamtschule durchgeführt. Ein Gymnasium und die Gesamtschule befinden sich in Großstädten im Ruhrgebiet, das zweite Gymnasium in einer Kleinstadt am Rande des Ruhrgebiets. Während die Gymnasien unter unauffälligen Rahmenbedingungen arbeiten können, wird die Gesamtschule überwiegend von Schülerinnen und Schülern mit ungünstigem sozialen Hintergrund besucht.

Das Großstadt-Gymnasium ist fünfzügig, aus organisatorischen Gründen konnten aber nur vier Klassen des 9. Jahrgangs an der Voruntersuchung teilnehmen. Vom Kleinstadt-Gymnasium konnten alle vier Klassen und von der Großstadt-Gesamtschule alle sechs Klassen des 9. Jahrgangs an der Voruntersuchung teilnehmen. Die Verteilung von Jungen und Mädchen innerhalb der drei Schulen ist typisch für die jeweilige Schulform. Die allgemein feststellbare höhere Schulbesuchsquote von Mädchen an Gymnasien schlägt sich aufgrund der pragmatischen Stichprobenfestlegung auch in den Gesamtzahlen nieder. So sind in der Voruntersuchung 43,6 % Jungen, 55,8 % Mädchen und 0,6 % Versuchspersonen ohne Angabe zum Geschlecht. Tabelle 5.1 gibt die Zusammensetzung der Stichprobe wieder.

Tabelle 5.1: Stichprobe der Voruntersuchung

Schule	männlich	weiblich	ohne Angabe	Summe
Großstadt-Gymnasium	42	64	0	106
Kleinstadt-Gymnasium	38	61	1	100
Großstadt-Gesamtschule	74	72	1	147
Summe	**154**	**197**	**2**	**353**

Da die Gesamtschule mit 6 Klassen in der Stichprobe vertreten ist (41,1 %) und ihre Schülerinnen und Schüler über eher ungünstige Lernvoraussetzungen verfügen, dürfte die Varianz der (nicht getesteten) Fachleistungen in dieser Stichprobe hinreichend groß sein, auch wenn die Leistungsverteilung vermutlich nicht repräsentativ für die „*PISA*-Population“ ist. Auch dürften die allgemeinen kognitiven Fähigkeiten insgesamt hinreichend heterogen für das Anliegen der Voruntersuchung sein.

5.3.2 Zusammenstellung der Erhebungsinstrumente und Sampling

Die in Kap. 5.2 vorgestellten Instrumente („Untertests“) der Voruntersuchung haben zusammen eine geschätzte Bearbeitungsdauer von 110 Minuten. Die Voruntersuchung sollte aber nicht länger als 90 Minuten dauern, da schulorganisatorische Bedingungen und die Konzentrationsfähigkeit der Schülerinnen und Schüler berücksichtigt werden mussten. Auf der Basis der geschätzten Bearbeitungsdauer wurden daher drei unterschiedliche Testhefte derart zusammengestellt, dass die Ziele der Voruntersuchung nicht unter dieser zeitlichen Beschränkung leiden mussten.

Im Testheft 1 wurde auf den *DAT:SR* verzichtet, im Testheft 2 auf den Untertest *Denkstile* und im Testheft 3 auf den *MRT*. Die drei genannten Tests haben alle eine geschätzte Bearbeitungsdauer von 20 Minuten, so dass für die verbleibenden Untertests in allen drei Testheften insgesamt 90 Minuten als geschätzte Bearbeitungsdauer übrig bleiben (vgl. Tab. 5.2). Diese pragmatische Auswahl orientierte sich daran, dass das Deckblatt natürlich obligatorischer Bestandteil jedes Tests ist, der Untertest *WLT*/*RFT* insgesamt sehr schnell zu bearbeiten ist und die Untertests *2 DW* und *3 DW* von allen Schülerinnen und Schülern bearbeitet werden sollten, um beide gemeinsam – wie in Kap. 4.2 dargestellt wurde – mit einer *LCA* bezüglich statistisch unterscheidbarer Bearbeitungsstrategien zu untersuchen.

Tabelle 5.2: Zusammenstellung der Testhefte für die Voruntersuchung

Untertest	Zeitbedarf	Testheft 1	Testheft 2	Testheft 3
UT 1: Deckblatt	10 min	X	X	X
UT 2: *2 DW*	15 min	X	X	X
UT 3: *WLT*/*RFT*	5 min	X	X	X
UT 4: *MRT*	20 min	X	X	–
UT 5: *Denkstile*	20 min	X	–	X
UT 6: *DAT:SR*	20 min	–	X	X
UT 7: *3 DW*	20 min	X	X	X

Die Testhefte 1-3 wurden klassenweise so eingesetzt, dass in jeder Schule alle drei Testhefte zum Einsatz kamen. Jedes Testheft wurde von zwei Gesamtschulklassen und von zwei oder drei Gymnasialklassen, insgesamt also von vier oder fünf Klassen bearbeitet. Testheft 1 wurde so von 126 Schülerinnen und Schülern bearbeitet, Testheft 2 von 130 Schülerinnen und Schülern und Testheft 3 von 97 Schülerinnen und Schülern. Jeder Untertest wurde von mindestens 223 Schülerinnen und Schülern bearbeitet.

5.3.3 Durchführung der Erhebung

Die Datenerhebung an den drei verschiedenen Schulen fand an fünf aufeinander folgenden Schultagen vor den Sommerferien 2003 statt. Da die Testhefte für eine Einsatzdauer von 90 Minuten konzipiert wurden, konnte die Erhebung jeweils in zwei aufeinander folgenden Schulstunden durchgeführt werden, ohne dass die Schülerinnen und Schüler eine große Pause durcharbeiten mussten.

In den beiden Gymnasien wurden die Klassen einzeln in ihren Klassenräumen getestet, sodass nicht erreicht werden konnte, dass die Schülerinnen und Schüler einzeln sitzen. Dennoch konnte ein Abschreiben von Testbearbeitungen des Nachbarn bzw. der Nachbarin nicht beobachtet werden. In der Gesamtschule wurden jeweils zwei Klassen parallel in einem sehr großen hörsaalartigen Raum getestet, sodass die Schülerinnen und Schüler einzeln sitzen konnten.

Die Tests wurden vom Autor der vorliegenden Arbeit selbst durchgeführt. In zwei Gymnasialklassen war keine Lehrkraft anwesend, ansonsten haben durchgehend Lehrkräfte mit Aufsicht geführt. In einer Doppelklasse an der Gesamtschule waren die Testbedingungen (trotz Aufsicht durch eine Lehrkraft) nicht optimal, da immer wieder größere Unruhe einkehrte. Dies ist in „Feldnotizen“ dokumentiert worden, um bei der Auswertung der Daten zu überprüfen, ob es hierdurch möglicherweise zu unbrauchbaren Daten gekommen ist.

Jedes Testheft bestand aus sechs Untertests (siehe Tab. 5.2), die durch „Stopp-Blätter“ getrennt waren. Schülerinnen und Schüler, die vor Ablauf der für den jeweiligen Untertest zur Verfügung stehenden Zeit die Bearbeitung beendet haben, durften nicht über diese „Stopp-Blätter“ hinaus im Testheft vorangehen. Diese Anweisung wurde überwiegend eingehalten, sodass in allen Klassen bzw. Doppelklassen allen Schülerinnen und Schülern gemeinsam die Instruktionen zu jedem einzelnen Test erläutert werden konnten.

Neben anderen jeweils klassenspezifischen Vorkommnissen konnten in allen Klassen bzw. Doppelklassen die folgenden Beobachtungen gemacht werden, die für die Auswertung der Voruntersuchung und für die Hauptuntersuchung relevant sind:

- Die Bearbeitungsdauer für das Deckblatt beträgt real – inklusive der Begrüßung und allgemeinen Einführung – deutlich weniger als 10 Minuten.

- Die Formulierung der Items zum *FSK:M* muss überdacht werden, da u. a. eine doppelte Verneinung auftaucht, die Schülerinnen und Schülern bei der Beantwortung Schwierigkeiten bereitet.
- Der Untertest *MRT* benötigt ebenfalls nicht 20 Minuten, sondern dürfte mit seinen 10 Items in der Regel in 15 Minuten bearbeitet werden können.
- Die Instruktion für die *RFT*-Aufgaben im Untertest *WLT*/*RFT* wurde von sehr vielen Schülerinnen und Schülern nicht verstanden. Schon während der Testbearbeitung hat sich angedeutet, dass dieser Test in der eingesetzten Form kaum verwertbare Ergebnisse liefern dürfte.
- Beim Untertest *Denkstile* sahen sich viele Schülerinnen und Schüler nicht dazu in der Lage, ihre Ergänzung der fehlenden Figur zu begründen. Viele äußerten sich dahingehend, dass sie ihre Überlegung schlichtweg nicht verbalisieren können. Zwar haben fast alle Schülerinnen und Schüler versucht, die Items zu bearbeiten, jedoch dürfte sich die Auswertung schwierig gestalten.

5.3.4 Erfassung und Aufbereitung der Daten

Die Daten der Voruntersuchung wurden für statistische Analysen in einer entsprechenden Datenmatrix erfasst. Als „Schlüssel“ für jeden Datensatz wurden die jeweilige Kursnummer und die individuelle Platznummern verwendet, die den Schülerinnen und Schüler vor der Testdurchführung mitgeteilt bzw. zugewiesen wurden. Die Erfassung der Daten erfordert bei Multiple-Choice-Items in der Regel nur eine Übertragung der angekreuzten Kategorie, wobei ggf. ein anderer Code für diese Kategorien verwendet wird. Anders stellt sich die Situation beim Untertest *WLT*/*RFT* und beim Untertest *Denkstile* dar. Beim Untertest *WLT*/*RFT* muss zunächst festgelegt werden, welche Bearbeitungen als richtig gelten, beim Untertest *Denkstile* ist ein Interpretationsschema für die Zuordnung von Bearbeitungen zu *prädikativem* bzw. *funktionalem Denken* erforderlich.

Beim Versuch, die Daten für den *RFT* und den Untertest *Denkstile* zu erfassen, zeigte sich, dass die in Kap. 5.3.3 dargestellten Beobachtungen bei der Testdurchführung zu Bearbeitungen geführt haben, die höchstens stark eingeschränkt ausgewertet werden können.

- Die Bearbeitungen des *RFT* zeigen deutlich, dass die Instruktionen nicht geeignet waren, um die Schülerinnen und Schüler zu Bearbeitungen im Sinne des Konstrukts *räumliche Wahrnehmung* anzuregen. Im Vergleich zum *WLT*, der die gleiche Komponente erfassen soll, fallen sowohl die hohe Anzahl nicht bearbeiteter Aufgaben als auch die extrem niedrige Anzahl der richtigen Bearbeitungen auf. Vermutlich ist die beschriebene Grundsituation beim *RFT* (beweglicher Stab im schief hängenden Bilderrahmen) für die Schülerinnen und Schüler nicht so zugänglich wie die Grundsituation beim *WLT* (Wasserspiegel in schräg gestellter Flasche). Da die Validität des verwendeten Testteils *RFT* im Gegensatz zum *WLT* insgesamt äußerst fraglich ist, wird der *RFT* nicht weiter

ausgewertet. In der Hauptuntersuchung kann die Komponente räumliche Wahrnehmung ggf. ausschließlich mithilfe des *WLT* erfasst werden.

- Die Bearbeitungen des Untertests *Denkstile* konnten zum Teil relativ sicher als *prädikatives* bzw. *funktionales Denken* interpretiert werden. Die theoretische Grundlage hierfür war die Konstruktbeschreibung in Kap. 4.1.2 in Verbindung mit den differenzierten Darstellungen von Schwank (1996, 1998, 2003a). Einige Bearbeitungen konnten aufgrund der konstruierten Lösungsfigur, andere im Zusammenhang mit der verbalen und zum Teil auch zeichnerischen Begründung gut interpretiert werden. Leider war eine derart leichte Interpretation aber nur in weniger als der Hälfte der Fälle möglich. Für etwa ein Viertel der Fälle ließ sich – bei größerer inhaltlicher Unsicherheit – eine Tendenz vermuten, die restlichen Fälle waren nicht bearbeitetet oder nicht auswertbar. Da insbesondere Versuchspersonen, die *funktional denken*, größere Schwierigkeiten haben (können), ihr Vorgehen zu verbalisieren (vgl. Schwank, 2003a, S. 71 f.), muss bei dieser hohen Quote der nicht oder nicht eindeutig zuordenbaren Versuchpersonen von einer systematischen Verzerrung ausgegangen werden. Für weitere quantitativ-empirische Auswertungen im Rahmen der Voruntersuchung können die Daten des Untertests *Denkstile* also nicht herangezogen werden. Für die Hauptuntersuchung ist zu prüfen, ob veränderte Instruktionen zu einer besseren Ausschöpfungsquote führen.

Signierung und Kodierung der Tests

Nachdem der Testteil *RFT* und der Untertest *Denkstile* in der Voruntersuchung nicht weiter ausgewertet werden, verbleibt als letzter Testteil, der sich nicht eines Multiple-Choice-Formats bedient, der *WLT*. Bei der Erfassung der Testleistung muss hier zunächst festgelegt werden, welche Bearbeitungen als richtig gelten sollen.[102] Laut Instruktion sollen die Versuchspersonen sich vorstellen, die beim *WLT* vorgegebenen Flaschen seien etwa zur Hälfte mit Wasser gefüllt. Zu dieser Situation passend sollen sie eine Linie einzeichnen, die die Wasseroberfläche darstellt. Da diese Aufgabe der Identifikation der gravitativen Horizontalen dient, können bestimmte Abweichungen von der Ideallösung noch akzeptiert und als richtig kategorisiert werden. So wurden auch Bearbeitungen akzeptiert, bei denen der eingezeichnete Wasserspiegel (a) nicht zur Teilinstruktion „etwa zur Hälfte gefüllt“ passte oder (b) statt durch eine gerade Linie durch eine Schlangenlinie dargestellt wurde. Bei der Auswertung wurde eine etwaige Schlangenlinie durch eine mittelnde Gerade ersetzt. Schließlich wurde als Genauigkeitsmaß für die Bearbeitung „± 3°“ festlegt, d. h. Linien, deren Richtung um nicht mehr als 3° von der gravitativen Horizontalen abweichen, gelten als richtig. Praktisch wurde dies mithilfe von Auswertungsschablonen realisiert (vgl. Abb. 5.9).

[102] Dieser Arbeitsschritt, bei dem die Vielzahl möglicher Bearbeitungen den Kategorien „richtig“ und „falsch“ zugeordnet werden, wird allgemein auch als „Signierung“ bezeichnet. Im Gegensatz dazu werden bei der „Kodierung“ bereits feststehende Kategorien z. B. durch Zahlen in einer Datenmatrix erfasst.

Abbildung 5.9: Auswertungsschablonen zum WLT

Mit diesen Auswertungsschablonen wurden die Bearbeitungen zu den einzelnen *WLT*-Items signiert und mit „0“ (= „falsch“) bzw. „1“ (= „richtig“) kodiert.

Bei den anderen Untertests, die sich allesamt des Multiple-Choice-Formats bedienen, gibt es grundsätzlich zwei Möglichkeiten der Kodierung. Für vertiefende Analysen des Bearbeitungsverhaltens (z. B. im Sinne von Bearbeitungsstrategien, vgl. Kap. 3.3.3) ist eine informative Kodierung der Items möglich, die angibt, wie viele und welche Antwortalternativen angekreuzt wurden. Für eine Skalierung der Testleistung auf der Basis des (ein- oder mehrdimensionalen) zweikategoriellen *RMs* genügt hingegen eine richtig/falsch-Kodierung, die ggf. berücksichtigt, ob ein Item bearbeitet wurde.[103] Da vor der Auswertung der Daten der Voruntersuchung nicht feststand, wie differenziert die Untertests ausgewertet werden sollen, wurde zunächst durchgehend die möglichst informative Kodierung gewählt. Diese nominale, mehrkategorielle Kodierung wurde später für die Skalierung der Testleistung dichotomisiert, wobei jeweils „1“ für die richtige Bearbeitung steht und „0“ für alle anderen bzw. fehlende Bearbeitungen. Für diese Umkodierung wurden neue Variablen angelegt.

Überlegungen zur Qualität der Daten

Da die Erfassung der Daten manuell erfolgte, kann trotz größter Sorgfalt nicht ausgeschlossen werden, dass es bei der Eingabe der Codes in die Datenmatrix zu einzelnen Fehlern gekommen ist. Dies kann zum Teil durch die verwendete Software abgefangen werden, etwa wenn ein unmöglicher Codes eingegeben werden soll (z. B. eine „7“, wenn nur „0“, „1“ und „9“ vorab als mögliche Codes festgelegt wurden). Darüber hinaus kann die Dateneingabe auch (stichprobenbasiert) durch einen Zweitkodierer überprüft werden. Auf

[103] Der Umgang mit fehlenden Bearbeitungen ist bei Leistungstests ein wichtiges Thema. Wenn eine Versuchsperson ein Item z. B. aus Zeitgründen nicht bearbeiten konnte, so ist dies zunächst nicht gleichwertig mit einer falschen Bearbeitung. Wenn ein Test allerdings so konzipiert ist, dass nahezu alle Versuchspersonen genügend Zeit zur Bearbeitung aller Items haben, so führt die Interpretation einer fehlenden Bearbeitung als „falsch“ kaum zu Verzerrungen bzgl. der Messung der Personeneigenschaft. In der vorliegenden Arbeit wurden fehlende Bearbeitungen in einem ersten Schritt separat kodiert und später wie falsche Bearbeitungen weiterverarbeitet.

diesen Schritt wurde im Rahmen der Voruntersuchung aus pragmatischen Gründen verzichtet, bei der Hauptuntersuchung wird die Eingabequalität aber überprüft.

Obwohl alle Untertests so konzipiert waren, dass eigentlich genügend Zeit zur Bearbeitung zur Verfügung stand, gab es Schülerinnen und Schüler, deren Testhefte eine hohe Zahl nicht bearbeiteter Items aufwiesen. Nach der Dateneingabe wurde daher zunächst eine neue Variable angelegt, in der die Anzahl der fehlenden Eingaben kodiert wurde. Je nach Testheft sollten die Schülerinnen und Schüler zwischen 54 und 75 unterschiedliche Angaben machen bzw. Items bearbeiten. Die maximale Anzahl nicht gemachter Angaben bzw. nicht bearbeiteter Items betrug 45. Bei Betrachtung dieses Falls wurde auch am Antwortmuster deutlich, dass die fehlenden Angaben motivational begründet waren und nicht auf fehlendes Leistungsvermögen zurückgeführt werden können. Dementsprechend wurde dieser Fall bei den weiteren Auswertungen nicht berücksichtigt. Anschließend wurden die Antwortmuster der Fälle mit den nächstkleineren Anzahlen fehlender Angaben betrachtet. In weiteren vier Fällen, die mehr als 15 fehlende Werte aufweisen, spielen vermutlich ebenfalls motivationale Aspekt eine wichtige Rolle, sodass diese Fälle aus dem Datensatz entfernt wurden. Bei allen anderen Fällen, die 15 oder weniger fehlende Angaben aufwiesen, können fehlende Angaben plausibel auf Erinnerungslücken (Angaben zu Noten) bzw. fehlendes Leistungsvermögen (Nicht-Bearbeitung schwieriger Items) zurückgeführt werden. Insgesamt verbleiben also 348 von ursprünglich 353 Fällen im Datensatz.

Schließlich wurden vor dem Hintergrund der Beobachtungen bei der Testdurchführung (vgl. Kap. 5.3.3) die beiden Klassen, in denen die Durchführungsbedingungen problematisch waren, bezüglich ihres Antwortverhaltens im Test mit den anderen vier Klassen derselben Schule verglichen. Da keine besonders auffälligen Abweichungen vorgefunden wurden, basiert die Auswertung der Voruntersuchung auch auf den Daten aus diesen beiden Klassen, also auf 348 Fällen.

5.3.5 Auswertung der Daten

Die Voruntersuchung soll zunächst die Frage klären, ob sich die Raumvorstellungstests bewähren und ggf. wie aus den Untertests der Voruntersuchung eine kürzere Testversion für die Hauptuntersuchung erstellt werden kann (vgl. Kap. 5.1). Darüber hinaus sollen erste Zusammenhänge und Gruppenunterschiede betrachtet werden, um die Hauptuntersuchung vorzubereiten. Die Analyse der Raumvorstellungstests kann dabei sowohl mit Methoden der *KTT* (vgl. Lienert & Raatz, 1998) als auch im Rahmen von *IRT*-Modellen (vgl. J. Rost, 2004) erfolgen. Wird ausschließlich innerhalb der *KTT* gearbeitet, so wird häufig die Anzahl richtig gelöster Items pro Untertest als Messwert betrachtet. Die Frage, ob ein Untertest eindimensional ist, wird dabei entweder per Augenscheinvalidierung geprüft oder mit Faktorenanalysen untersucht. Im Rahmen von *IRT*-Skalierungen werden entsprechende Annahmen und auch die Güte des zugrundegelegten Testmodells empirisch überprüft (vgl. Kap. 2.1.2 und Kap. 4.2.3).

Bei der Auswertung der Voruntersuchung wurden beide Wege beschritten. Da die Daten mit dem Programmpaket *SPSS* erfasst wurden, in dem sehr umfassend die Verfahren der *KTT* implementiert sind, ist eine deskriptive Analyse der Testrohwerte (Anzahl richtig gelöster Items) einfach und komfortabel umsetzbar. Auf der Basis geeigneter Kodierungen bzw. Umkodierungen sind aber auch die *IRT*-Skalierungen ohne großen Aufwand umsetzbar. Da die Algorithmen zur Schätzung der Modellparameter nicht immer transparent sind, wurden hier mehrere Programmpakete (*ConQuest*, *Mplus* oder *WINMIRA*) parallel eingesetzt, um die Skalierungen wechselseitig zu überprüfen und abzusichern. Die *IRT*-Skalierungen bieten sich in der Voruntersuchung an, um auffällige Items zu identifizieren und Kurzformen von Tests zu entwickeln. Für alle Raumvorstellungstests wurden dementsprechend sowohl *KTT*-Auswertungen als auch *IRT*-Skalierungen durchgeführt. Ausgewählte Ergebnisse werden in Kap. 5.4 dokumentiert.

Erste Auswertungen der vier Items zum *FSK:M* haben dazu geführt, dass dieses Konstrukt im Rahmen der Voruntersuchung nicht vertieft ausgewertet wird. Problematisch war dabei das zweite Item („*Mathematik liegt mir nicht besonders*“), das invers gepolt ist, d. h. im Gegensatz zu den drei anderen Items deuten hohe Zustimmungswerte zur Aussage dieses Items auf ein gering ausgeprägtes *FSK:M* hin. Die während der Testdurchführung beobachteten Schwierigkeiten der Schülerinnen und Schüler im Umgang mit der doppelten Verneinung („trifft nicht zu“ bezogen auf die Aussage des Items), kann auch in den Daten beobachtet werden: Konstruktkonform wäre gewesen, wenn auch das Ankreuzverhalten auf dieses Item invers zum Antwortverhalten auf die anderen Items gewesen wäre. Tatsächlich war das Antwortverhalten aber tendenziell gleichgerichtet. Zusätzlich weist das „Problem-Item“ auch eine größere Streuung des Antwortverhaltens auf, was vermutlich darin begründet liegt, dass einige Schülerinnen und Schüler es konstruktkonform verstanden haben, die überwiegende Mehrzahl aber nicht. In der Hauptuntersuchung müssen daher andere, konsistent formulierte Items verwendet werden.

5.4 Befunde der Voruntersuchung

Die Befunde zur Voruntersuchung werden im Folgenden mit Blick auf ihre dienende Funktion für die Hauptuntersuchung dargestellt. Zunächst wird die Voruntersuchung für die einzelnen Raumvorstellungstests separat ausgewertet. Dabei werden Statistiken zu den Testrohwerten (Anzahl richtig gelöster Items), die einen ersten Eindruck von der Schwierigkeit des Tests geben, um Ergebnisse aus den (eindimensionalen) Rasch-Skalierungen der einzelnen Tests ergänzt. Im Rahmen der *IRT*-Analyse werden auch Items mit schlechten Kennwerten genauer betrachtet und mögliche Kurzformen von Tests vorgeschlagen. Anschließend werden (lineare) Zusammenhänge zwischen den Untertests mithilfe des mehrdimensionalen *RMs* untersucht und Bearbeitungsstrategien über mehrere Untertests hinweg empirisch betrachtet (Konstruktvalidierung des *MRT* mithilfe von *2 DW* und *3 DW*). Schließlich werden vertiefende Analysen zu Geschlechterunterschieden und zum

Zusammenhang der Testleistungen mit Schulnoten durchgeführt. Differenzierte Betrachtungen zur Modellgüte werden im Rahmen der Voruntersuchung nicht berichtet.

Auf der Basis dieser Auswertungen kann die Hauptuntersuchung dann effektiv sowie theorie- und hypothesengeleitet durchgeführt werden. Der Mehrwert dieses Vorgehens lässt sich am Beispiel der Geschlechterunterschiede verdeutlichen: Ausgehend von den vorliegenden Befunden, die in Kap. 3.3.2 zusammengefasst wurden, können die Untertests der Voruntersuchung mit ersten Vermutungen über Zusammenhänge und Unterschiede ausgewertet werden. Die Ergebnisse dienen dann der Formulierung von Hypothesen für die Hauptuntersuchung. Wenn in der Hauptuntersuchung wiederum signifikante Ergebnisse erzielt werden, sind diese durch vorliegende Literatur und die Voruntersuchung gut vorbereitet und abgesichert – auch hinsichtlich der „α-Fehler-Kumulierung“[104].

Bei den Auswertungen der Voruntersuchung muss berücksichtigt werden, dass die Stichprobe in Klumpen organisiert ist – insbesondere sind zwei Schulformen beteiligt. Aufgrund der bekannten Befunde muss davon ausgegangen werden, dass sowohl allgemeine kognitive Fähigkeiten als auch Fachleistungen an den Gymnasien im Durchschnitt deutlich besser ausgeprägt sind als an der Gesamtschule. Da der Mädchenanteil an den Gymnasien (62,0 %) erheblich höher ist als an der Gesamtschule (49,3 %), müssen vor allem bei Betrachtungen zu Geschlechterunterschieden immer wieder Auswertungen nach Schulformen getrennt durchgeführt werden (vgl. „*TIMSS/II*-Paradoxon“, Kap. 2.3.2). Aber auch Besonderheiten in der Verteilung aller Schülerinnen und Schüler nach ihrer Testleistung (bzw. den Testrohwerten) können möglicherweise durch die Stichprobenstruktur erklärt werden.

Einige vertiefende Analysen zu den involvierten Tests bleiben der Hauptuntersuchung vorbehalten. So lassen sich z. B. die *WLT*-Aufgaben auch mit Blick auf den Drehwinkel der Flaschen auswerten. Vermutlich beeinflusst der Drehwinkel maßgeblich die Aufgabenschwierigkeit. Diese und ähnliche Betrachtungen werden ggf. anhand der optimierten Tests und Stichprobe der Hauptuntersuchung durchgeführt.

[104] Die „α-Fehler-Kumulierung“ – auch „α-Fehler-Inflation“ genannt – bezeichnet das Phänomen, dass bei einer hinreichend großen Zahl von Signifikanztests schon zufallsbedingt einige signifikante Ergebnisse erwartet werden müssen, obwohl die Nullhypothese gilt. Beim „üblichen“ Signifikanzniveau $\alpha = 0{,}05$ tritt erwartungsgemäß ein signifikantes Ergebnis pro 20 Tests auf. Mit den Möglichkeiten der heutigen Statistik-Programmpakete lassen sich fast beliebig viele Tests rechnen (ohne dass der Sinn der einzelnen Tests in jedem Fall hinterfragt wird). Dieses Problem kann zwar durch die Korrektur des Signifikanzniveaus gelöst werden, indem das eigentlich angedachte Niveau für einen Einzeltest bei k Einzeltests durch Division gleichmäßig aufgeteilt wird („Bonferroni-Korrektur“). Dadurch sinkt jedoch die Teststärke für das Aufspüren signifikanter Ergebnisse. Wenn die einzelnen Hypothesen jedoch aus der Literatur abgeleitet und zu älteren Befunden passend sind, kann auf die Korrektur verzichtet werden, da nicht alle kombinatorisch möglichen Testkonstellationen durchgeführt, sondern gezielt aus Theorie und Empirie gewonnene Konstellationen überprüft werden. In solchen Fällen kann ggf. auch mit einem großzügigeren Niveau als 5 % agiert werden.

5.4.1 Erprobung und Skalierung der Raumvorstellungstests

Für jeden der fünf Raumvorstellungstests, die im Rahmen der Voruntersuchung weitergehend ausgewertet wurden, wird zunächst jeweils die Verteilung der Stichprobe nach Gesamtscore[105] in einem Säulendiagramm und mit charakteristischen Kennwerten dargestellt. Dabei werden als Kennwerte der Median, das arithmetische Mittel, die Schiefe („skewness“) und der Exzess („kurtosis“) für die jeweilige Verteilung berichtet. Schiefe und Exzess dienen dem Vergleich der jeweiligen Verteilung mit der Standardnormalverteilung (Schiefe: rechtsschief – normal – linksschief; Exzess: breit – normal – schmal; vgl. Kelava & Moosbrugger, 2008). Bei Besonderheiten der Verteilungen werden ggf. noch weitergehende Analysen durchgeführt.

Zusätzlich werden für jeden Untertest die Ergebnisse der Rasch-Skalierung dargestellt. Dabei wird die Verteilung der Stichprobe nach Testleistung zusammen mit den Itemparametern (Itemschwierigkeit) auf der gemeinsamen (latenten) Dimension abgebildet; potenziell vorhandene (empirisch) problematische Items werden genauer betrachtet, da schlechte Itemkennwerte häufig darauf hindeuten, dass Items „handwerkliche Mängel“ aufweisen. Bei den Itemschwierigkeiten ist davon auszugehen, dass diese – anders als bei Fachleistungstests – innerhalb der Untertests relativ homogen sind, da innerhalb eines Untertests jeweils dieselbe Raumvorstellungskomponente erfasst werden soll. Idealtypisch ist dabei nur ein kognitiver Prozess zur Aufgabenlösung erforderlich, sodass es kaum schwierigkeitsbestimmende Faktoren für die Items gibt, die variiert werden können.

Die Parameter der Rasch-Skalierungen (Itemschwierigkeit und Personenfähigkeit), die für die einzelnen Untertests berichtet werden, ergeben sich aus bestimmten Varianten von *Maximum-Likelihood-Schätzungen*, die mit dem Programmpaket *WINMIRA* realisiert wurden. Die Werte für die Itemschwierigkeit werden üblicherweise so normiert, dass ihre Summe Null ergibt. Da die durch das *RM* gewonnene Skala eine Differenzskala ist, sind die Itemschwierigkeiten mit dieser Normierung festgelegt – und mit ihnen auch die geschätzte Testleistung der Versuchspersonen. Die Testleistung wurde für jeden möglichen Gesamtscore nach der *WLE*-Methode (vgl. Warm, 1989) ermittelt. Mit dieser Methode sind Schätzungen der Personeneigenschaft auch dann möglich, wenn eine Versuchsperson kein Item oder alle Items richtig gelöst hat (vgl. J. Rost, 2004, Kap. 4.2). Bei einer nicht modifizierten *Maximum-Likelihood-Methode* müsste für solche Versuchspersonen die Personeneigenschaft mit $-\infty$ (kein Item richtig) bzw. ∞ (alle Items richtig) geschätzt werden, da das Testverhalten nahe legt, dass die Lösungswahrscheinlichkeit für alle Items Null bzw. Eins beträgt. Bei der Analyse der einzelnen Untertests werden ggf. weitere Betrachtungen angestellt bzw. weitere Merkmale der Rasch-Skalierung am jeweiligen Fall diskutiert.

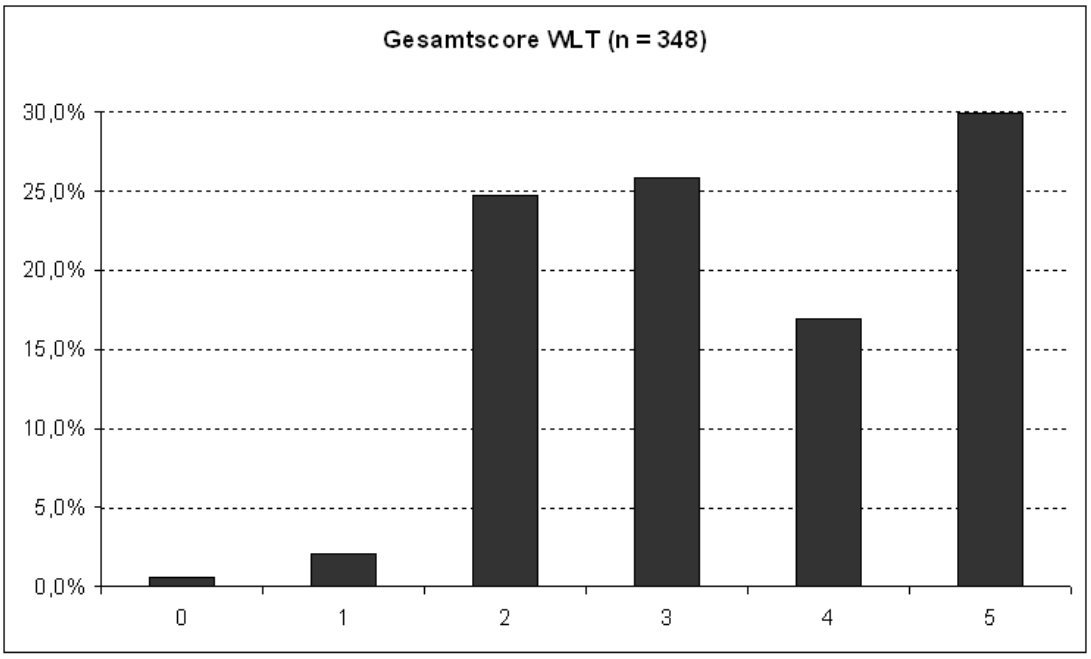
[105] Mit „Gesamtscore“ für einen Untertest wird für jede Versuchsperson die Anzahl richtig bearbeiteter Items bezeichnet.

Für den *DAT:SR* wird auf der Basis der Rasch-Skalierung und der darauf basierenden Itemanalyse eine mögliche Kurzform für die Hauptuntersuchung entwickelt. Der in der Vorsuchung verwendete Untertest *DAT:SR* ist mit 25 Items relativ lang, sodass eine Auswahl von Items gefunden werden soll, die möglichst informativ bei deutlich reduziertem Testumfang ist.

WLT

Betrachtet man die Verteilung der Versuchspersonen nach ihrem Gesamtscore im Testteil *WLT* zunächst graphisch im Säulendiagramm (Abb. 5.10), so fällt u. a. auf, dass (a) fast alle Versuchspersonen mindestens zwei Items richtig bearbeitet haben, dass (b) der *WLT* in seiner eingesetzten Form einen Deckeneffekt hat und bei Personen mit hoher Testleistung nicht weiter differenzieren kann sowie dass (c) die Verteilung „bimodal“ (zweigipflig) ist.

Abbildung 5.10: Verteilung der Versuchspersonen nach WLT-Gesamtscore



Phänomen (a) lässt sich damit erklären, dass die Items 2 und 4 sehr leicht sind.[106] Phänomen (b) kann bei dem verwendeten Testkonzept des *WLT* vermutlich nicht durch weitere Items mit anders gedrehten Flaschen, sondern nur durch höhere Anforderungen an die Präzision der Bearbeitung gelöst werden (vgl. Kap. 5.3.5); dann wäre aber fraglich, ob ausschließlich räumliche Wahrnehmung oder z. B. auch feinmotorische Kompetenzen erfasst

[106] Bei Item 2 ist die Flasche, in die der Wasserspiegel eingezeichnet werden soll, um 180° gegen den Uhrzeigersinn gedreht („auf dem Kopf“), bei Item 1 um 270° gegen den Uhrzeigersinn gedreht („auf der Seite“).

werden. Schließlich kann Phänomen (c) grundsätzlich mindestens zwei unterschiedliche Ursachen haben. Einerseits kann es sein, dass die Items so aufgebaut sind, dass die beiden schwierigsten Items in ähnlicher Weise von den Versuchspersonen gelöst werden können, sodass es wahrscheinlicher ist, alle fünf Items zu lösen, als nur vier. Andererseits kann auch die Struktur der Stichprobe zu einer bimodalen Verteilung führen. Daher werden die beiden großen Klumpen „Gesamtschule“ und „Gymnasium“ weiter unten noch einmal separat betrachtet.

Die beobachteten Phänomene spiegeln sich zum Teil auch in den ausgewählten Kennwerten der Verteilung wider (Tab. 5.3).

Tabelle 5.3: Kennwerte der Verteilung nach WLT-Gesamtscores (n = 348)

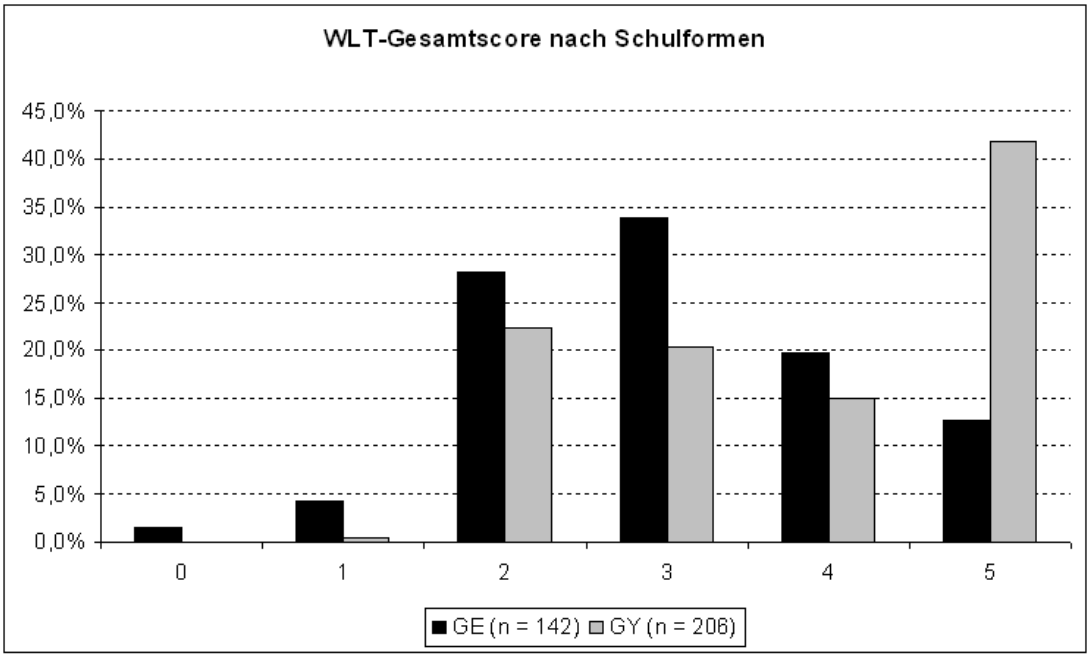
Test	Itemzahl	Median	arithm. Mittel	Schiefe	Exzess
WLT	5	3	3,46	–0,12	–1,12

Das arithmetische Mittel der Anzahl richtig bearbeiteter Items beträgt bei einer Gesamtzahl von fünf Items 3,46. Dies ist praktisch nur möglich, wenn die hohen Gesamtscores (4 und 5) sehr häufig auftreten und die niedrigen (0 und 1) sehr selten. Die Schiefe unterscheidet sich bei der vorliegenden Verteilung nicht signifikant von Null, obwohl deutlich sichtbar keine Normalverteilung vorliegt. Allerdings kann eine bimodale Verteilung mit zwei etwa gleich stark besetzten Gipfelbereichen genauso wenig schief sein, wie eine Normalverteilung. Die Bimodalität der Verteilung drückt sich aber deutlich im Exzess der Verteilung aus, da zwei Gipfel (zumindest bei einer diskreten Verteilung mit wenigen Ausprägungen) zu einer (im Vergleich zur Normalverteilung) relativ breiten Verteilung führen.

Auch wenn die Voruntersuchung primär der Vorbereitung der Hauptuntersuchung dient, soll an dieser Stelle exemplarisch das Phänomen (c) der Bimodalität untersucht werden.[107] Oben wurden zwei Vermutungen geäußert, von denen die zweite auf die Struktur der Stichprobe, die durch Schulformklumpen gekennzeichnet ist, hinweist. Abbildung 5.11 und Tabelle 5.4 stellen die Verteilung der Versuchspersonen nach *WLT*-Gesamtscore für beide Schulformen getrennt dar und unterstützen die genannte Vermutung.

[107] Die Verteilungen der Versuchspersonen auf die Gesamtscores der anderen Untertests werden dann jeweils kürzer dargestellt und nur bei großen Auffälligkeiten kommentiert.

Abbildung 5.11: Verteilung der Versuchspersonen nach WLT-Gesamtscore und nach Schulformen



Die getrennten Säulendiagramme zeigen, dass die beiden Verteilungen unterschiedliche Modalwerte haben, die so auch die gemeinsame Verteilung (Abb. 5.10) prägen. Die Verteilung für die Gesamtschule ist der Normalverteilung ähnlich, was durch die Werte für Schiefe und Exzess (Tab. 5.4) bestätigt wird. Der Deckeneffekt des *WLT* tritt vor allem im Gymnasium auf, wo über 40 % der Schülerinnen und Schüler alle fünf Items richtig bearbeiten. Bemerkenswert ist, dass der Gesamtscore 4 am Gymnasium seltener auftritt als der Gesamtscore 3 bzw. der Gesamtscore 2. Eine theoretische Aufgabenanalyse (Fokus: Drehwinkel der Flasche) deutet darauf hin, dass dies an zwei wesentlichen Typen von Items liegt: Horizontal bzw. vertikal ausgerichtete Flaschen können von schräg stehenden unterschieden werden. Bei tendenziell gut ausgeprägter *räumlicher Wahrnehmung* (wie am Gymnasium) können (häufig) alle oder (seltener) keine Items mit schräg stehenden Flaschen gelöst werden.

Tabelle 5.4: Kennwerte der Verteilungen nach WLT-Gesamtscores und nach Schulformen

	Itemzahl	**Median**	**arithm. Mittel**	**Schiefe**	**Exzess**
GE (n = 142)	5	3	3,04	0,03	–0,34
GY (n = 206)	5	4	3,75	–0,32	–1,46

Aufgrund der bisherigen Auswertung zum *WLT* kann man vermuten, dass die Bearbeitung der Items einerseits im Wesentlichen den gleichen kognitiven Prozess oder gleiche kognitive Prozesse erfordert und dass es zwei unterschiedliche Itemcluster gibt, die sich nach empirischer Schwierigkeit unterscheiden. Der erste Teil dieser Vermutung lässt sich nur schwer quantitativ-empirisch untersuchen, der zweite Teil wird aber durch die Ergebnisse der Rasch-Skalierung bestätigt. In Abbildung 5.12 werden die mit *WINMIRA* berechneten Werte in Anlehnung an den typischen Output des Programmpakets *ConQuest* dargestellt. In vertikaler Richtung wird die latente Dimension der fraglichen Personeneigenschaft und der Itemschwierigkeit („Logit-Skala“) dargestellt. Die sechs möglichen *WLT*-Gesamtscores sind durch die Skalierung (mit *WLE*-Schätzungen) so transformiert worden, dass nun ein Differenzskalenniveau vorliegt. An der Stelle der entsprechenden Testleistung (θ_j) ist pro 1 % der Stichprobe ein X gesetzt worden, sodass der linke Teil des Diagramms ein mit Blick auf das obige Säulendiagramm (Abb. 5.10) verzerrtes Balkendiagramm darstellt. Auf der rechten Seite stehen die Nummern der Testitems an den Positionen der zugehörigen Itemschwierigkeiten (σ_i).

Abbildung 5.12: Verteilungen der WLT-Testleistung und der Itemschwierigkeiten (n = 348)

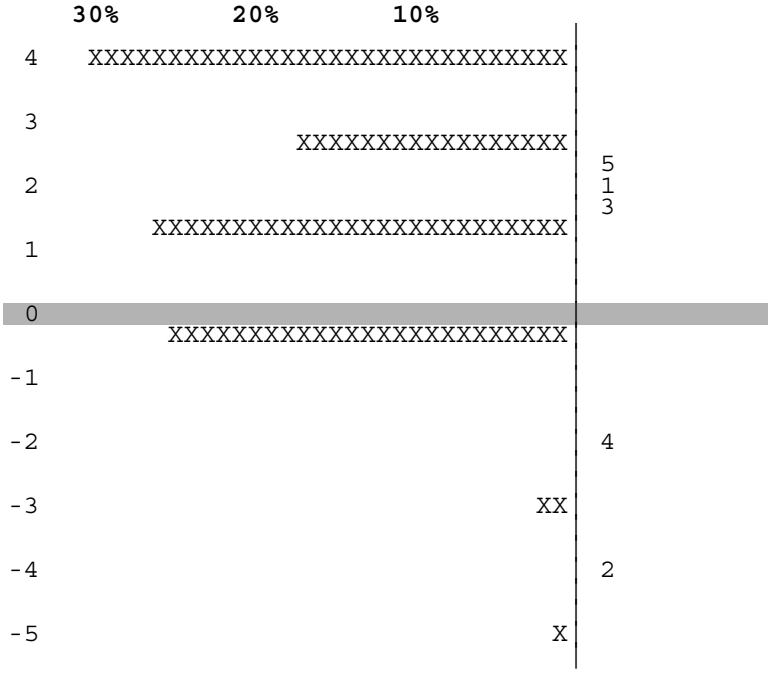
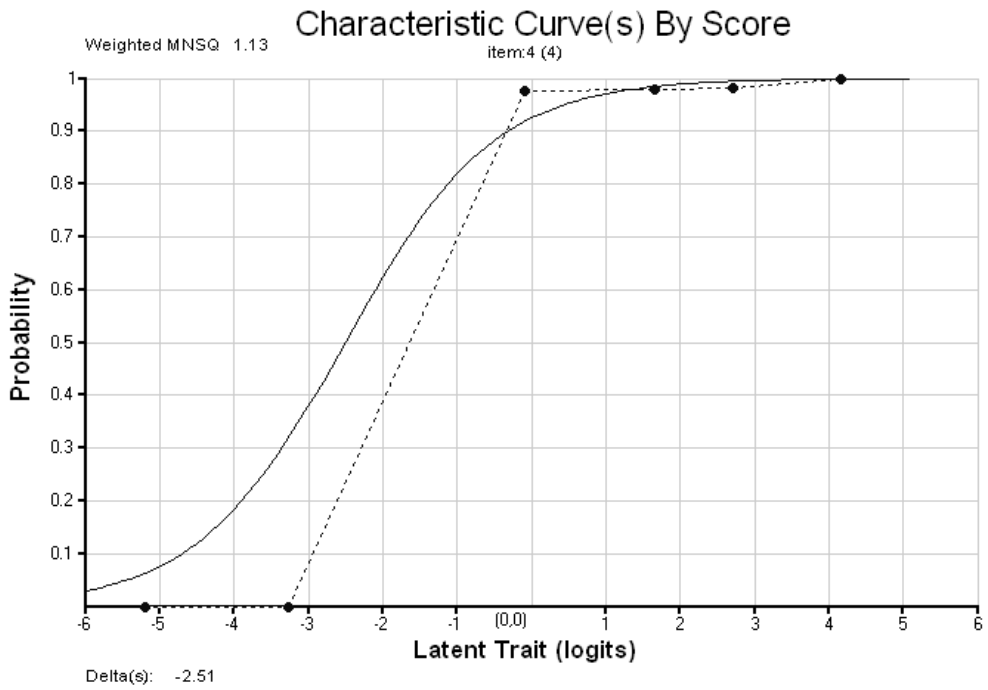


Abbildung 5.12 zeigt deutlich, dass es im oberen Bereich drei ähnlich schwierige Items (1, 3 und 5) gibt; hierbei handelt es sich um die Items mit schräg stehenden Flaschen. Im unteren Bereich sind die Items 2 und 4 relativ leicht, wobei Item 2 (Flasche „auf dem Kopf stehend“) noch einfacher ist als Item 4 (Flasche „auf der Seite liegend“). Die Vermutung zur Existenz dieser Itemcluster scheint tragfähig zu sein. Die durchschnittlich eher hohen Lösungsquoten beim *WLT*, mit dem arithmetischen Mittel 3,46, zeigen sich an der Verteilung auf der linken Seite. Die Itemschwierigkeiten sind – wie oben dargestellt wurde –

normiert mit Summe Null, die *WLE*-Schätzungen für die Testleistung sind hierdurch mit festgelegt und liegen in der Tendenz deutlich oberhalb von Null. Abbildung 5.12 zeigt auch, dass im oberen Leistungsbereich keine Items mehr liegen, die zur weiteren Differenzierung der Leistung beitragen könnten. Der Deckeneffekt des *WLT* lässt sich im Zusammenspiel der Verteilungen der Testleistung und der Itemschwierigkeit noch klarer feststellen.

Von den möglichen weitergehenden Test- und Itemanalysen, die auf der Basis der Rasch-Skalierung durchgeführt werden können, wird an dieser Stelle nur noch die Anpassung der Itemfunktionen (*ICCs*) an die empirischen Daten betrachtet (vgl. Kap. 4.2.3). Der mit dem *RM* theoretisch angenommene Verlauf dieser Funktionen wird – gemäß der Herleitung in Kap. 2.1.2 – durch die Funktion $p_{i,1} = \frac{\exp(\theta - \sigma_i)}{1 + \exp(\theta - \sigma_i)}$ beschrieben. Für jede beobachtete Ausprägung der Personeneigenschaft, die als Transformation aus den möglichen Gesamtscores entsteht, kann man aus den empirischen Daten relative Lösungshäufigkeiten für jedes Item bestimmen: Es wird der Anteil der Versuchspersonen mit dem jeweiligen Gesamtscore berechnet, der das fragliche Item richtig gelöst hat. Diese relative Lösungshäufigkeit kann mit der aufgrund der Modellgleichung erwarteten Wahrscheinlichkeit verglichen werden. Abbildung 5.13, die wie alle weiteren Visualisierungen dieses Typs mit dem Programmpaket *ConQuest* erstellt wurde, zeigt diesen Vergleich für das WLT-Item 4.

Abbildung 5.13: Theoretische und empirische ICC für das WLT-Item 4



Das *WLT*-Item 4 (Flasche „auf der Seite liegend“) ist gut zur Differenzierung zwischen Versuchspersonen mit Gesamtscore 1 und mit Gesamtscore 2 geeignet. Hat eine Versuchsperson (mindestens) zwei *WLT*-Items richtig bearbeitet, so ist fast sicher Item 4 darunter. Hingegen hat (praktisch) keine Versuchsperson mit Gesamtscore 1 dieses Item richtig bearbeitet. Zwischen Versuchspersonen mit Gesamtscore 0 mit Gesamtscore 1 oder zwischen Versuchspersonen mit hohen Gesamtscores kann Item 4 aber nicht mehr differenzieren.

Insgesamt passen theoretischer und empirischer Verlauf aber hinreichend gut zueinander. Dies drückt sich auch im Abweichungsmaß „Weighted MNSQ“ aus, das von *ConQuest* standardmäßig für alle Items berechnet wird, aus. Mit diesem gewichteten Abweichungsmaß können der theoretische und der empirische Verlauf auf signifikante Unterschiede getestet werden: Die Nullhypothese (theoretischer und empirischer Verlauf sind deckungsgleich) wird durch einen *Weighted MNSQ* von 1,00 repräsentiert. Bei *IRT*-Skalierungen ist es üblich, von einer hinreichend guten Passung eines Items auszugehen, wenn diese Nullhypothese nicht verworfen werden soll.[108] Der *Weighted MNSQ* von Item 4 beträgt 1,13 und liegt innerhalb des Intervalls [0,58; 1,24], in dem die Nullhypothese nicht verworfen werden soll. Diese Intervalle können von Item zu Item variieren und sind insbesondere bei sehr leichten Items relativ groß. Das Programm *WINMIRA* berechnet anstelle des *Weighted MNSQ* ein anderes Abweichungsmaß, nämlich den *Q-Index* bzw. die daraus gewonnenen standardisierten Werte Z_Q (vgl. J. Rost, 2004, Kap. 6.2), mit denen analoge Signifikanztests durchgeführt werden können und die im Wesentlichen zu denselben Resultaten führen.

Die analogen Berechnungen und Werte für die *WLT*-Items 1, 2, 3 und 5 deuten ebenfalls auf keine zu großen Abweichungen zwischen den jeweiligen theoretischen und empirischen Verläufen der *ICCs* hin. Der *WLT* hat sich für einen Einsatz in der Hauptuntersuchung also hinreichend bewährt, auch wenn er insgesamt im oberen Leistungsbereich kaum noch differenzieren kann. Das fehlende Differenzierungspotenzial dürfte sich in der Hauptuntersuchung vor allem wieder im Gymnasium zeigen; dort muss von einer Unterschätzung der Leistungsvarianz ausgegangen werden, die sich in einer Unterschätzung etwaiger linearer Zusammenhänge niederschlägt. In den anderen Schulformen dürfte dieses Phänomen weniger stark ausgeprägt sein oder gar nicht auftreten.

[108] Aus der Sicht der Logik des einfachen Hypothesentests ist dies nicht unproblematisch, da hier aus einem nicht signifikanten Testergebnis ein Schluss gezogen wird. Eigentlich ist es üblich, nur aus signifikanten Ergebnissen inhaltliche Schlüsse zu ziehen und nicht-signifikante Ergebnisse nicht weiter zu nutzen. Zumindest müsste aber der mögliche Fehler 2. Art betrachtet werden (vgl. Büchter & Henn, 2007, Kap. 4.2). Wie viele andere Konventionen bei der Testskalierung hat sich das übliche – aus puristischer Sicht problematische – Vorgehen aber anscheinend bewährt.

MRT

Für den *MRT* (und die weiteren Untertests der Voruntersuchung) wurden zunächst analoge Auswertungen wie für den *WLT* durchgeführt.

Abbildung 5.14: Verteilung der Versuchspersonen nach MRT-Gesamtscore

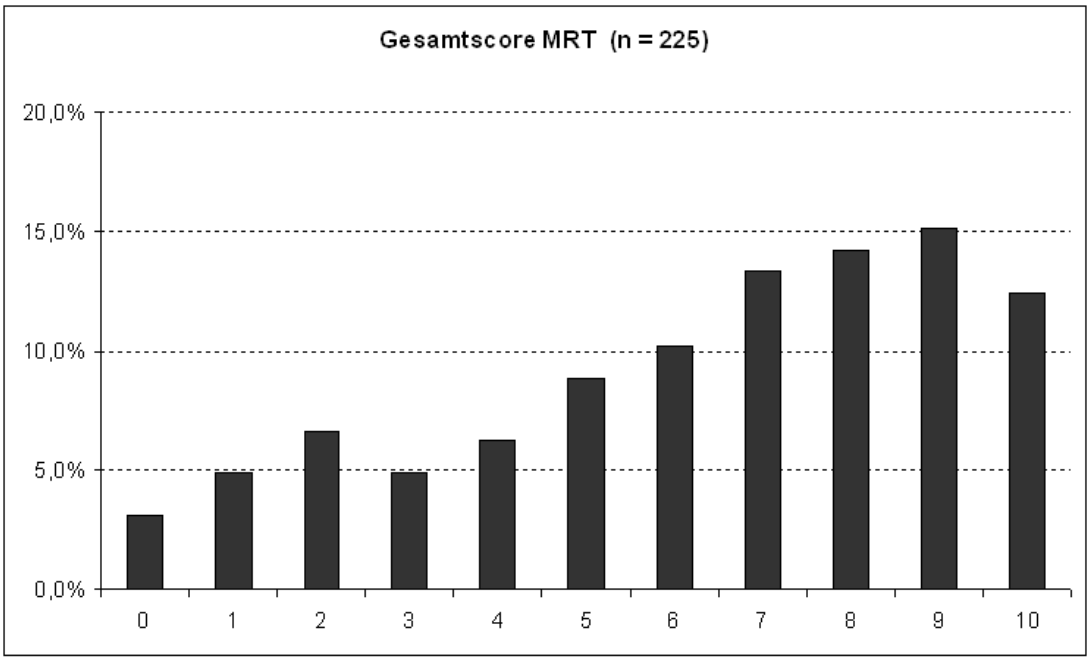


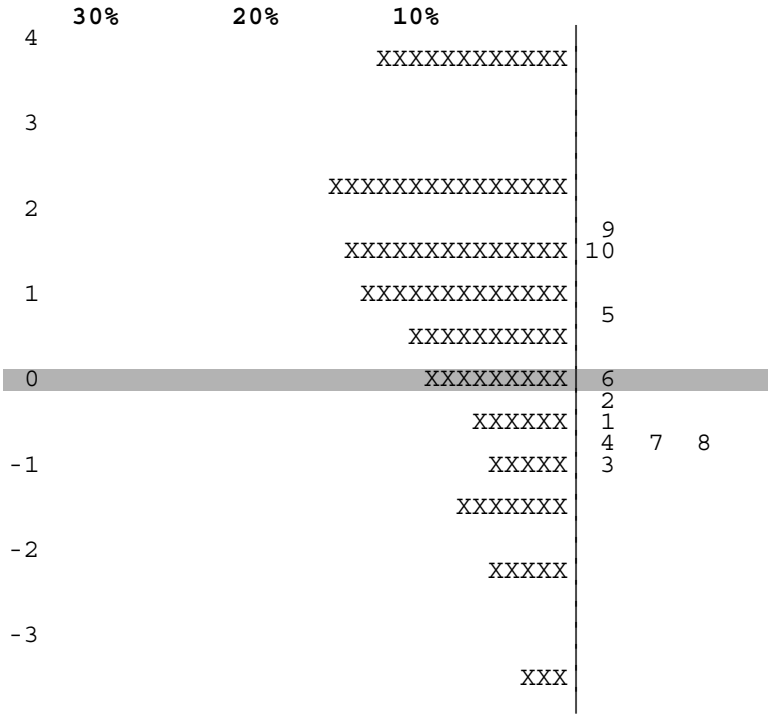
Abbildung 5.14 zeigt, dass es auch beim *MRT* einen moderaten Deckeneffekt gibt, der sich entsprechend in den Verteilungskennwerten niederschlägt (Tab. 5.5). In der *MRT*-Stichprobe konnten zwar einerseits 12,4 % der Versuchspersonen alle Items richtig bearbeiten. Auf der anderen Seite haben aber auch 3,1 % der Versuchspersonen alle Items falsch bearbeitet. Dies ist auch insofern bemerkenswert, als die theoretische Ratewahrscheinlichkeit für *MRT*-Items $^1/_6$ beträgt („genau zwei von vier sind richtig“), was für den gesamten *MRT* einem Erwartungswert von ca. 1,7 richtig bearbeiteten Items entspricht. Die obige Verteilung ist wiederum bimodal (allerdings weniger stark ausgeprägt als beim *WLT*). Dies lässt sich wieder auf die unterschiedlichen Verteilungen in den beiden Schulformen zurückführen.

Tabelle 5.5: Kennwerte der Verteilung nach MRT-Gesamtscores (n = 225)

	Itemzahl	**Median**	**arithm. Mittel**	**Schiefe**	**Exzess**
MRT	10	7	6,31	–0,58	–0,71

Die Ergebnisse der Rasch-Skalierung in Abb. 5.15 zeigen, dass die Items im mittleren Leistungsbereich gut differenzieren. Eine größere Spannweite der Itemschwierigkeit (hier 2,9 Logits zwischen Item 3 und Item 9) lässt sich mit Aufgaben des *MRT*-Typs vermutlich nicht erzielen. Da es im unteren Leistungsbereich keine adäquaten Items gibt, kann nicht ausgeschlossen werden, dass hier (ggf. geschicktes) Raten zur Unterscheidung zwischen den geringen Testleistungen führt.

Abbildung 5.15: Verteilungen der MRT-Testleistung und der Itemschwierigkeiten (n = 225)



Der *Q-Index* für Item 3 deutet auf eine etwas zu geringe Trennschärfe des Items hin, die allerdings nur knapp unterhalb des Signifikanzniveaus (0,045 < 0,05) liegt. Inhaltlich lässt sich dieses abweichende Verhalten nicht erklären. Zwar kann Item 3 nicht ausschließlich durch *mentale Rotation*, sondern auch über analytische Betrachtungen gelöst werden. Darin unterscheidet Item 3 sich aber nicht von allen anderen. Bei den schwierigsten Items 9 und 10 können Positionseffekte nicht ausgeschlossen werden; möglicherweise ließ bei einigen Versuchspersonen zum Ende dieses Testheftes die Konzentration nach. Eine genaue Betrachtung der Distraktoren offenbart aber auch, dass bei diesen Items aufgrund der Darstellung vermutlich die höchsten Anforderungen realisiert werden. Eine genauere Betrachtung der leichtesten Items (3, 4, 7 und 8) erfolgt im Rahmen der Analyse von Bearbeitungsstrategien in Kap. 5.4.2.

Insgesamt scheint auch der *MRT* hinreichend gut für die Hauptuntersuchung geeignet zu sein. Die empirischen Werte zu den einzelnen Items und zum gesamten Test sind inhaltlich hinreichend plausibel und die Deckeneffekte im Gymnasium weniger stark ausgeprägt als

beim *WLT*. Auch Item 3 ist nicht so problematisch, dass es entfernt werden müsste – ggf. fallen seine Kennwerte in der Stichprobe der Hauptuntersuchungen auch etwas besser aus. Die Frage möglicher Strategieunterschiede bei der Bearbeitung des *MRT* (und damit die Frage der Konstruktvalidität) wird in Kap. 5.4.2 vertieft.

DAT:SR

Der *DAT:SR* ist mit 25 Items der umfangreichste Untertest der Voruntersuchung. Da einige Items ungünstig dargestellt sind, wurde der *DAT:SR* bewusst zunächst in diesem Umfang eingesetzt. Im Rahmen der Auswertung der Voruntersuchung sollte dann eine Kurzform mit zehn Items für den Einsatz in der Hauptuntersuchung generiert werden.

Abbildung 5.16 und Tabelle 5.6 offenbaren zwar ebenfalls einen leichten „Deckeneffekt“ des *DAT:SR*, dieser beschränkt sich aber auf das Gymnasium. In der Gesamtstichprobe konnten 3,7 % der Versuchspersonen alle 25 Items richtig bearbeiten. Der Unterschied zwischen Gesamtscore 24 und Gesamtscore 25 kann dabei nicht durch ein besonders schwieriges Item erklärt werden, sondern wird vermutlich mit der großen Itemzahl zusammenhängen, ohne dass es Positionseffekte gibt.

Abbildung 5.16: Verteilung der Versuchspersonen nach DAT:SR-Gesamtscore

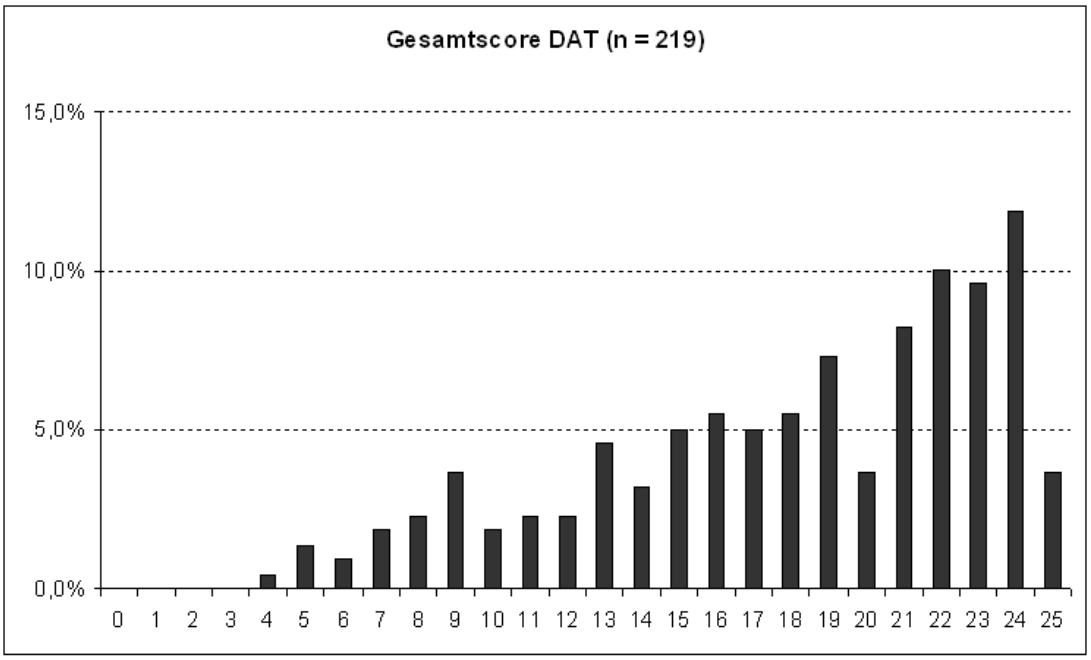
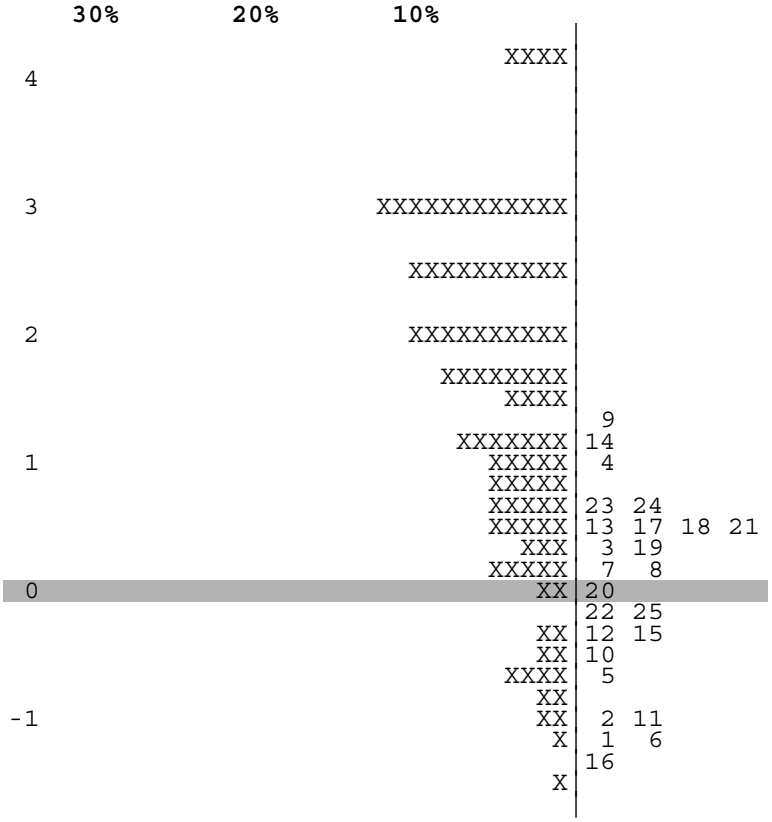


Tabelle 5.6: Kennwerte der Verteilung nach DAT:SR-Gesamtscores (n = 219)

	Itemzahl	**Median**	**arithm. Mittel**	**Schiefe**	**Exzess**
DAT:SR	25	19	18,02	–0,71	–0,47

Die Ergebnisse der Rasch-Skalierung in Abbildung 5.17 zeigen deutlich die Homogenität der Itemschwierigkeiten. Anders als beim *MRT* gibt es beim *DAT:SR* auch Items, die im unteren Leistungsbereich differenzieren. Dies hängt vor allem damit zusammen, dass der Test insgesamt relativ leicht ist. Die Verteilung der Testleistung hat ihren Schwerpunkt deutlich im positiven Bereich. Im oberen Leistungsbereich gibt es wiederum keine adäquaten Items; auch im Rahmen des *DAT:SR* könnten solche Items vermutlich nur mit konstruktfremden Anteilen entwickelt werden.

Abbildung 5.17: Verteilung der DAT:SR-Testleistung und der Itemschwierigkeiten (n = 219)



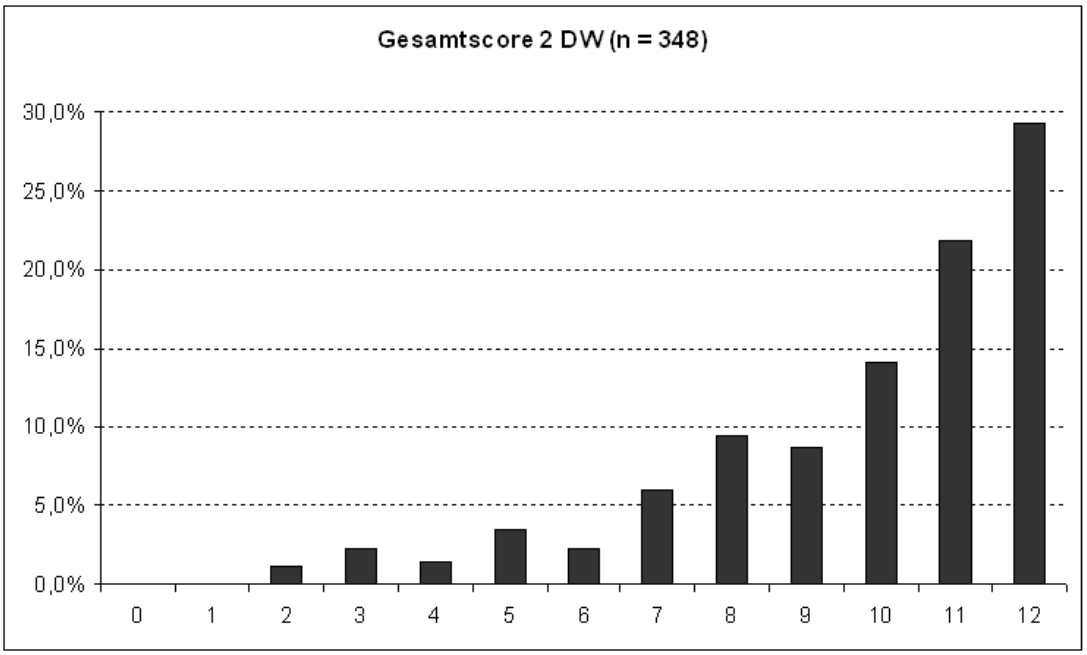
An der Verteilung der Testleistung lässt sich gut ein Effekt der Skalierung beobachten: Während im mittleren Bereich der Zusammenhang zwischen dem Gesamtscore und den transformierten Werten fast perfekt linear ist, verzerrt sich dieser Zusammenhang für extreme Gesamtscores deutlich. Da der niedrigste in Abb. 5.17 dargestellte Gesamtscore 5 ist, fällt dies nur im oberen Leistungsbereich auf. Dort werden die Abstände zwischen den zu zwei benachbarten Gesamtscores gehörigen Leistungswerten immer größer.

Da die Items 3, 7 und 24 problematische Kennwerte bezüglich ihrer Passung ins Modell aufweisen, werden diese drei Items sowie von den übrigen 22 die leichtesten 12 Items aus dem Test entfernt. So entsteht eine Kurzform des Tests, die ausschließlich modellkonforme Items enthält und das mögliche Differenzierungspotenzial nach oben ausschöpft.

2 DW

Der Untertest *2 DW* war in der Voruntersuchung unerwartet leicht. Zwar ist der *2 DW* aus dem *IST:WÜ* gerade durch Selektion der im Vergleich zu den „Raumwürfeln“ leichteren „Flächenwürfel“ entstanden (vgl. Kap. 5.2.1), dennoch sind die Lösungshäufigkeiten extrem hoch. Abbildung 5.18 zeigt einen starken Deckeneffekt.

Abbildung 5.18: Verteilung der Versuchspersonen nach 2 DW-Gesamtscore



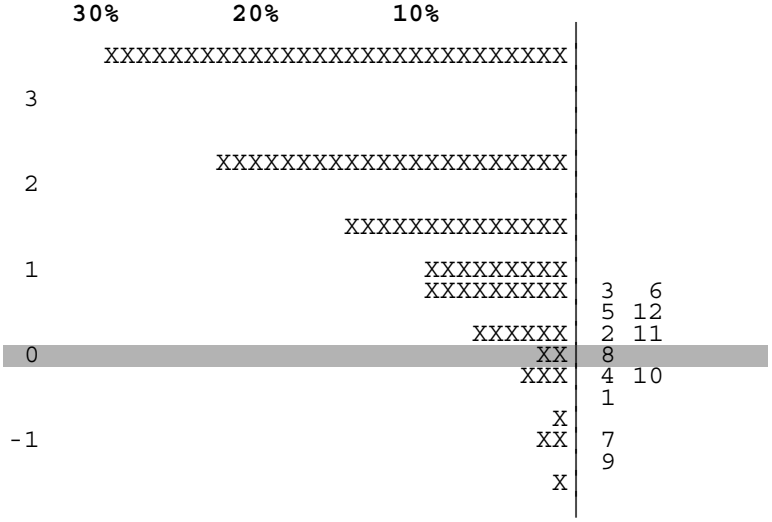
Obwohl der Test insgesamt 12 Items umfasst, erreichen 29,3 % der Versuchspersonen den maximalen Gesamtscore. Nimmt man den nächstniedrigeren Gesamtscore dazu so hat man bereits 51,1 % der Stichprobe erfasst. Dies drückt sich auch im Median der Verteilung sowie in der Schiefe aus (Tab. 5.7). Zwar gibt es auch beim *2 DW* Unterschiede bezüglich der beteiligten Schulformen, ein Deckeneffekt mit einhergehender eingeschränkter Varianz ist aber auch in der Gesamtschule zu beobachten.

Tabelle 5.7: Kennwerte der Verteilung nach 2 DW-Gesamtscores (n = 348)

	Itemzahl	**Median**	**arithm. Mittel**	**Schiefe**	**Exzess**
2 DW	12	11	9,74	–1,25	0,966

Die Ergebnisse der Rasch-Skalierung (Abb. 5.19) bestätigen dieses Bild und zeigen darüber hinaus die große Homogenität der Itemschwierigkeiten. Richtig schwierige Würfelaufgaben sind offensichtlich solche mit „Raumwürfeln“ (s. u., *3 DW*).

Abbildung 5.19: Verteilung der 2 DW-Testleistung und der Itemschwierigkeiten (n = 348)



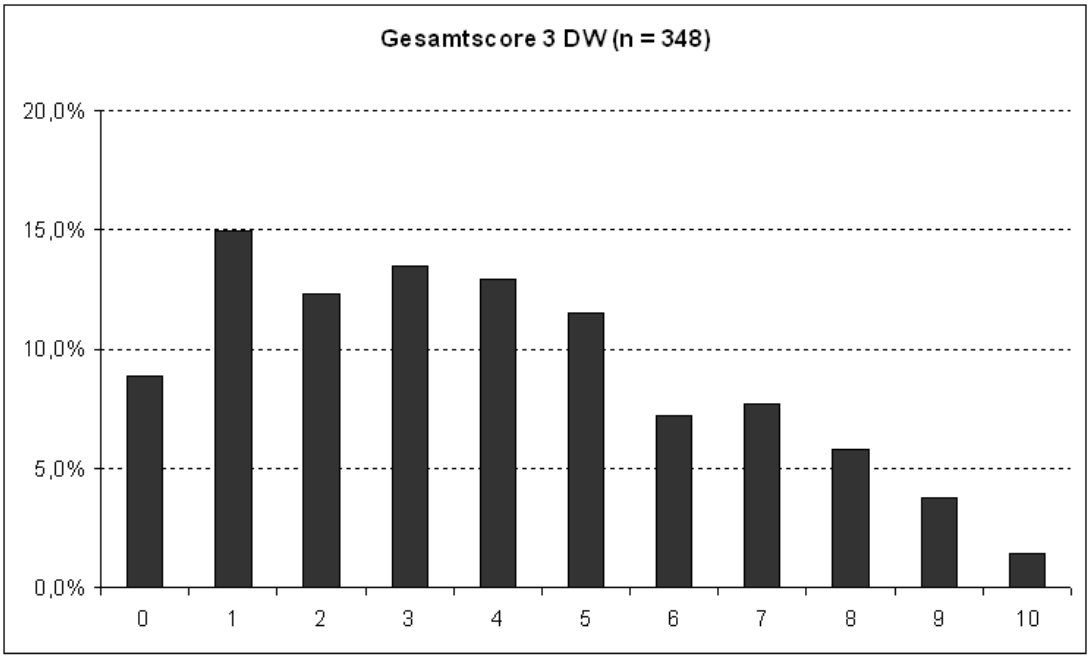
Auffällige Itemkennwerte zeigt lediglich Item 6, das zwar in der Verteilung als relativ schwieriges Item ausgewiesen wird, das aber vor allem von Versuchspersonen mit geringer Testleistung gut gelöst werden kann. Dementsprechend ist die reale Trennschärfe nicht modellkonform. Inhaltlich ist dies durchaus plausibel, da Item 6 schon durch sehr oberflächliche Betrachtungen richtig gelöst werden kann.

Insgesamt stellt die Tatsache, dass der *2 DW* ein extrem einfacher Untertest mit starkem Deckeneffekt ist, kein Problem dar. Der *2 DW* wurde vor allem deswegen in die Voruntersuchung aufgenommen, um in Kombination mit dem *3 DW* Ergebnisse zu Strategieunterschieden zu erzielen. Dies ist unter Umständen auch bei einer derart schiefen Verteilung möglich.

3 DW

Obwohl der *3 DW* ebenfalls aus dem *IST:WÜ* entstanden ist, sehen die Ergebnisse gänzlich anders als beim *2 DW* aus. Hier zeigt sich, dass „Raumwürfel“ empirisch deutlich schwieriger sind als „Flächenwürfel“ (Abb. 5.20).

Abbildung 5.20: Verteilung der Versuchspersonen nach 3 DW-Gesamtscore



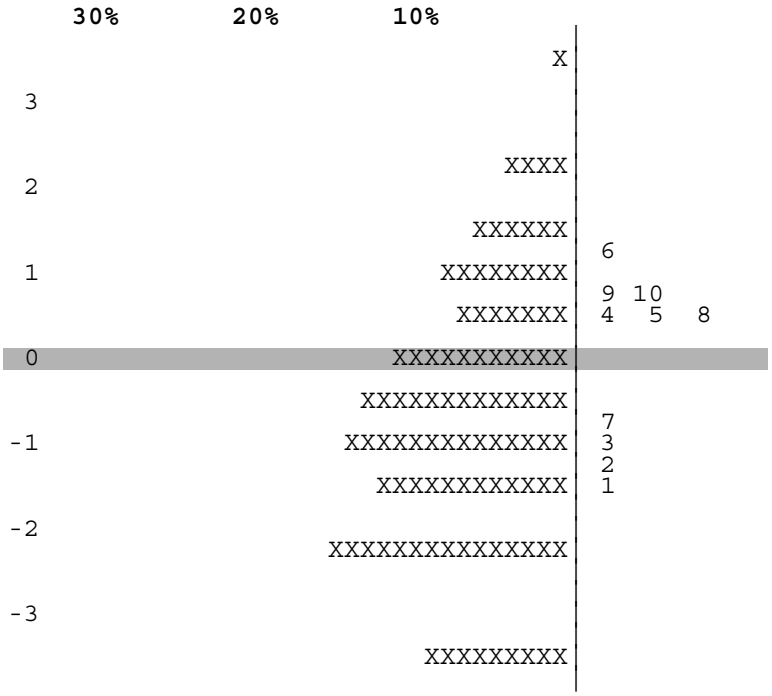
Obwohl die Verteilung nur ganz leicht bimodal ist, stecken dahinter deutlich sichtbar die typischen Effekte der Klumpenstichprobe. Während der Modalwert der Verteilung in der Gesamtschule bei 1 liegt, ergibt sich im Gymnasium nahezu eine Normalverteilung mit Modalwert 5. Der *3 DW* ist der einzige Untertest, der auf Ebene der gesamten Stichprobe eine positive Schiefe hat, die sich zudem signifikant von Null unterscheidet. Dies drückt sich auch im leichten Bodeneffekt aus, den der Test hat. Mit Blick auf die oben schulformspezifisch berichteten Mediane ist es nicht überraschend, dass dieser Bodeneffekt nahezu ausschließlich in der Gesamtschule auftritt, wo 19,0 % der Schülerinnen und Schüler kein Item richtig lösen. Inhaltlich ist diese hohe Zahl trotz Ratemöglichkeiten in einem Multiple-Choice-Test plausibel, da der *3 DW* hohe kognitive Anforderungen stellt und vermutlich schnell motivationale Effekte provoziert, die praktisch zu einem Testabbruch führen.

Tabelle 5.8: Kennwerte der Verteilung nach 3 DW-Gesamtscores (n = 348)

	Itemzahl	**Median**	**arithm. Mittel**	**Schiefe**	**Exzess**
3 DW	10	4	3,81	0,41	–0,74

Die Skalierungsergebnisse in Abb. 5.21 unterstreichen diesen Eindruck. Da die Verteilung der Testleistung insgesamt tendenziell im negativen Bereich liegt, können die Items vor allem im unteren Leistungsbereich nicht mehr differenzieren.

Abbildung 5.21: Verteilung der 3 DW-Testleistung und der Itemschwierigkeiten (n = 348)



Die Itemkennwerte der Voruntersuchung sprechen durchgängig für modellkonforme Items. Dies war auch insofern zu erwarten, als der *3 DW* als Rasch-homogener Test entwickelt wurde (vgl. Gittler, 1990). Insgesamt lassen auch die Ergebnisse zum *3 DW* die Möglichkeit zu, dass Unterschiede in den Bearbeitungsstrategien erfolgreich im Zusammenhang mit dem *2 DW* untersucht werden können.

5.4.2 Zusammenhänge zwischen den Raumvorstellungstests

Zusammenhänge zwischen Komponenten der *Raumvorstellung* oder zwischen konkreten Raumvorstellungstests können aus ganz unterschiedlichen Perspektiven betrachtet werden. Für die vorliegende Arbeit sind wiederum die psychometrische und die kognitionspsychologische Perspektive von besonderem Interesse. Aus psychometrischer Sicht wird unter „Zusammenhang“ vor allem die Kovarianz bzw. Korrelation als Maß für den linearen Zusammenhang verstanden. Aus kognitionspsychologischer Sicht ist eher interessant, ob für die erfolgreiche Bearbeitung unterschiedlicher Raumvorstellungstests gleiche oder vergleichbare mentale Prozesse erforderlich sind.

Die beiden zuvor genannten Perspektiven können sich in der Forschungspraxis wechselseitig sinnvoll ergänzen. Wenn etwa die Leistungen in zwei unterschiedlichen Raumvorstellungstests hoch miteinander korrelieren, so stellt sich die Frage, wie dies erklärt werden kann – möglicherweise spielen bei beiden Tests ähnliche kognitive Prozesse eine Rolle. Wird umgekehrt festgestellt, dass bei zwei unterschiedlichen Tests ähnliche kognitive Prozesse eine Rolle spielen, so sollte sich dies auch in einer höheren Korrelation ausdrücken. Zumindest moderate Korrelationen sind ohnehin zu erwarten, da – in einer hinreichend

heterogenen Stichprobe – kognitive Leistungen generell miteinander korrelieren. Diese generelle Korrelation zwischen den verschiedenen Komponenten bzw. Dimensionen kognitiver Leistungen stellen die empirische Grundlage für aktuelle Intelligenztheorien dar, die einen *g-Faktor* vorsehen (vgl. Kap. 3.1.1.).

Auch die bereits berichteten Ergebnisse der Voruntersuchung zeigen, vor allem wenn sie schulformspezifisch betrachtet werden, dass höhere Leistungen in einem Untertest tendenziell mit höheren Leistungen in den anderen Untertests einhergehen. In allen Untertests liegen die Leistungen am Gymnasium signifikant und mit großen Effektstärken über denen an der Gesamtschule. Dies drückt sich zwangsläufig auch in entsprechenden positiven Korrelationen aus. Im Folgenden werden zunächst solche statistischen Zusammenhänge zwischen den Untertests der Voruntersuchung mit dem mehrdimensionalen *RM* untersucht. Anschließend wird mit Blick auf die unterschiedlichen Bearbeitungsstrategien bei Würfelaufgaben (vgl. Kap. 3.3.3) untersucht, ob sich ähnliche Unterschiede auch beim *MRT* finden lassen. Auch dies geschieht auf der Basis eines *IRT*-Modells, nämlich einer *LCA*.

Insgesamt soll so das in Kap. 4.1.2 festgelegte Konstrukt *Raumvorstellung* bereits in der Voruntersuchung empirisch untersucht werden. Dabei dürfen die drei zugrundeliegenden Komponenten zwar korrelieren, allerdings sollten diese Korrelationen nicht zu hoch ausfallen. Mit der Analyse von Bearbeitungsstrategien soll der *MRT* als wichtiger Test für die Beobachtung und ggf. Erklärung von Geschlechterunterschieden in der *Raumvorstellung* vertieft untersucht werden.

Latente Korrelationen

Für eine Untersuchung der linearen Zusammenhänge zwischen den Untertests bietet sich das mehrdimensionale *RM* (vgl. Kap. 2.1.2) an. Diese Modellierung kann auch als konfirmatorische Faktorenanalyse (CFA) betrachtet werden, da a priori festgelegt wird, welche Items zu welcher Dimension gehören (oder in der Sprechweise der *CFA* „welche Items auf welchem Faktor laden“). Die latenten Korrelationen wurden paarweise für jeweils zwei der fünf Untertests in einem zweidimensionalen *RM* geschätzt. Diese paarweise Modellierung ist bei vorliegender „Einfachladungsstruktur“ („between-Modell“; vgl. Kap. 2.3.3), bei der jedes Items genau auf einer Dimension lädt, angemessen (vgl. Winkelmann & Robitzsch, 2009, S. 177 f.). Für die Komponenten *räumliche Wahrnehmung*, *mentale Rotation* und *räumliche Visualisierung* mit den zugehörigen Untertests *WLT*, *MRT* und *DAT:SR* lässt sich dieses Vorgehen auf der Basis der Festlegung der Konstrukte (vgl. Kap. 4.1.2) und der konkreten Instrumente (vgl. Kap. 5.2.1) rechtfertigen. Tabelle 5.8 gibt die mit dem Programmpaket *ConQuest* entsprechend geschätzten Korrelationen zwischen den Untertests wieder. Dabei sind die für die Hauptuntersuchung weniger relevanten Zusammenhänge zwischen dem *2 DW* bzw. dem *3 DW* und den anderen Untertests grau unterlegt.

Tabelle 5.9: Paarweise geschätzte latente Korrelationen zwischen den Untertests der Voruntersuchung

	2 DW	*3 DW*	***DAT:SR***	***MRT***
WLT	0,39 (n = 348)	0,49 (n = 348)	**0,44 (n = 219)**	**0,55 (n = 225)**
MRT	0,53 (n = 225)	0,74 (n = 225)	**0,64 (n = 96)**	
DAT:SR	0,61 (n = 219)	0,79 (n = 219)		
3 DW	0,42 (n = 348)			

Die latenten Korrelationen zwischen *WLT*, *MRT* und *DAT:SR* zeigen deutlich, dass durch diese Tests unterschiedliche Dimensionen erfasst werden. Dabei ist die Korrelation zwischen dem *WLT* und dem *DAT:SR* am niedrigsten, was mit Blick auf die konkreten Anforderungen der jeweiligen Aufgaben durchaus plausibel ist.[109] Ebenso plausibel ist, dass der Zusammenhang zwischen *MRT* und *DAT:SR* etwas enger ist.[110] Bei diesen vorsichtigen Vermutungen muss aber berücksichtigt werden, dass aufgrund des realisierten Samplings insbesondere die Teilstichprobe, die sowohl den *MRT* als auch den *DAT:SR* bearbeitet hat, mit 96 Schülerinnen und Schülern relativ klein war. Zwar sind alle in der Tabelle aufgeführten Korrelationen signifikant von Null verschieden, die Unterschiede zwischen den Werten sollten sich vor einer weitergehenden Interpretation aber erst in der Hauptuntersuchung bestätigen.

Von den Zusammenhängen zwischen dem *2 DW* bzw. dem *3 DW* und den anderen Untertests sind vor allem die relativ engen Zusammenhänge zwischen *3 DW* und *MRT* sowie zwischen *3 DW* und *DAT:SR* bemerkenswert und auch plausibel. So enthalten die Aufgaben des *3 DW* jeweils sowohl Anforderungen im Bereich *mentaler Rotation* (Drehen des Würfels) als auch Anforderungen im Bereich *räumlicher Visualisierung* (Überprüfen der Seitenmuster und -beziehungen).

Mit Blick auf das in Kap. 4.1.2 festgelegte Konstrukt *Raumvorstellung* kann auf jeden Fall festgehalten werden, dass die drei Komponenten, die in der Hauptuntersuchung im Zu-

[109] Die Identifikation der gravitativen Horizontalen, die beim *WLT* im Mittelpunkt steht, dürfte bei der Bearbeitung des *DAT:SR* keine Rolle spielen.

[110] Gewisse Teilleistungen bei der Bearbeitung der *DAT:SR*-Aufgaben bestehen in der mentalen Veränderung der räumlichen Lage von Teilobjekten (beim Falten der Netze). Diese Teilleistungen dürften näher an *mentaler Rotation* als an *räumlicher Wahrnehmung* sein.

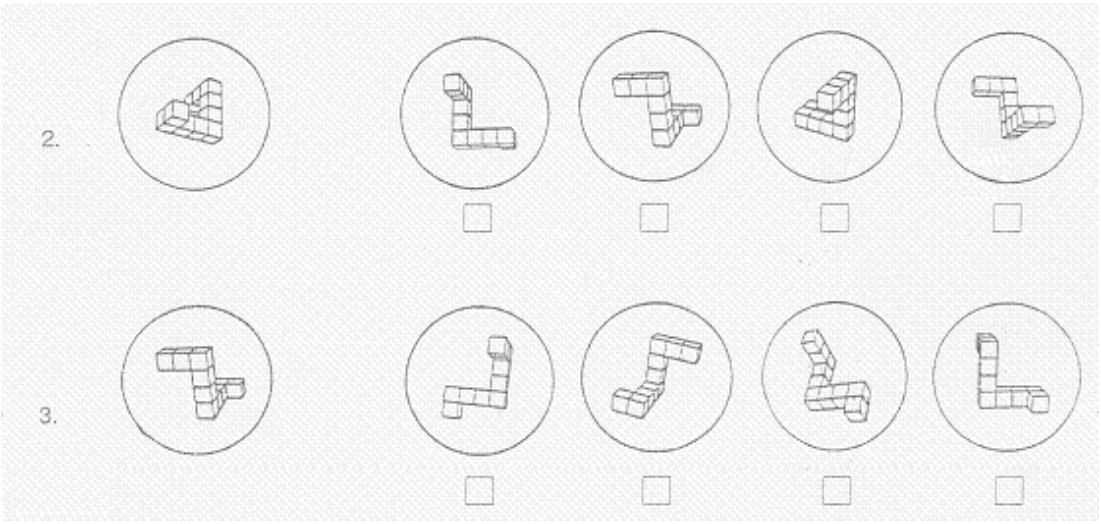
sammenhang mit *Mathematikleistung* betrachtet werden sollen, drei empirisch trennbare Dimensionen darstellen.

Bearbeitungsstrategien

In Kap. 3.3.3 wurde für Items aus dem *IST:WÜ* (vgl. Abb. 3.8, S. 100) dargestellt, dass es unterschiedliche Bearbeitungsstrategien geben kann, die in theoretischen Aufgabenanalysen identifiziert und bei geeigneten Designs bzw. Methoden auch empirisch vorgefunden werden können. Auf der Basis der Arbeiten von Barrat (1953) und Putz-Osterloh (1977) zeigen Köller et al. (1994) und Hosenfeld et al. (1997), dass sich bei den genannten Würfelaufgaben „Analytiker“ (auch: „Flächenstrategen“) von „Holistikern“ (auch: „Raumstrategen“) anhand ihrer Lösungsprofile im Rahmen einer *LCA* unterscheiden lassen. Putz-Osterlohs Kritik an der Strategieheterogenität des *IST:WÜ* folgend hat Gittler (1990) einen Würfeltest (*3 DW*) entwickelt, der ausschließlich „Raumwürfel“ (vgl. Kap. 3.3.3, S. 100 ff.) enthält. Eine Auswahl geeigneter „Flächenwürfel“ wurde im Rahmen der Voruntersuchung im Untertest *2 DW* zusammengefasst.

Für den *MRT* können Geiser et al. (2006) ebenfalls anhand theoretischer Aufgabenanalysen und empirischer Untersuchungen mit einer *LCA* zeigen, dass sich „Rotators“ von „Nonrotators“ unterscheiden lassen und dass sich dies wiederum in entsprechenden Items niederschlägt. Durch diesen Befund wird die Konstruktvalidität des *MRT* eingeschränkt, da dieser Test *der* Referenztest für („dreidimensionale“) *mentale Rotation* ist. Mögliche Bearbeitungsstrategien und entsprechende Item-Merkmale werden für die *MRT*-Items in Abb. 5.22 erläutert.

Abbildung 5.22: MRT-Items mit spiegelsymmetrischen (2) bzw. anders geformten (3) Falschlösungen



Bei Item 2 sind die beiden Falschlösungen spiegelsymmetrisch zu den richtigen Lösungen, während die Falschlösungen bei Item 3 grundsätzlich anders geformt sind als die richtigen Lösungen. Daher können „Nonrotators" das Item 3 auch durch andere mentale Prozesse als durch *mentale Rotation* richtig lösen. Die Vermutung liegt nahe, dass sich „Nonrotators" analytischer Strategien bedienen. Eine solche Strategiemöglichkeit könnte auch einen Beitrag zur relativ hohen latenten Korrelation zwischen *MRT* und *DAT:SR* leisten (s. o., Tab. 5.8), da dann nicht nur *mentale Rotation* bei *DAT:SR*-Items hilfreich ist, sondern auch die analytischen Anteile von *räumlicher Visualisierung* bei einigen *MRT*-Items.

Die Befunde von Geiser et al. (2006) zu „Rotators" und „Nonrotators" bei *MRT* scheinen also in einem anderen Test dieselben Strategieunterschiede zu identifizieren, die Putz-Osterloh (1977), Köller et al. (1994) und Hosenfeld et al. (1997) beim *IST:WÜ* gefunden haben. Die Voruntersuchung der vorliegenden Arbeit soll diese Frage weiter vertiefen (vgl. Kap. 4.1.3). Dazu sind drei Teilfragen zu beantworten:

- Können im Rahmen der Voruntersuchung bei einer gemeinsamen Auswertung der Untertests *2 DW* und *3 DW* „Analytiker" empirisch von „Holistikern" getrennt werden?
- Lassen sich bei der Auswertung des *MRT* analog „Nonrotators" von „Rotators" trennen?
- Handelt es sich in der Stichprobe bei „Analytikern" und „Nonrotators" tendenziell um dieselben Versuchspersonen?

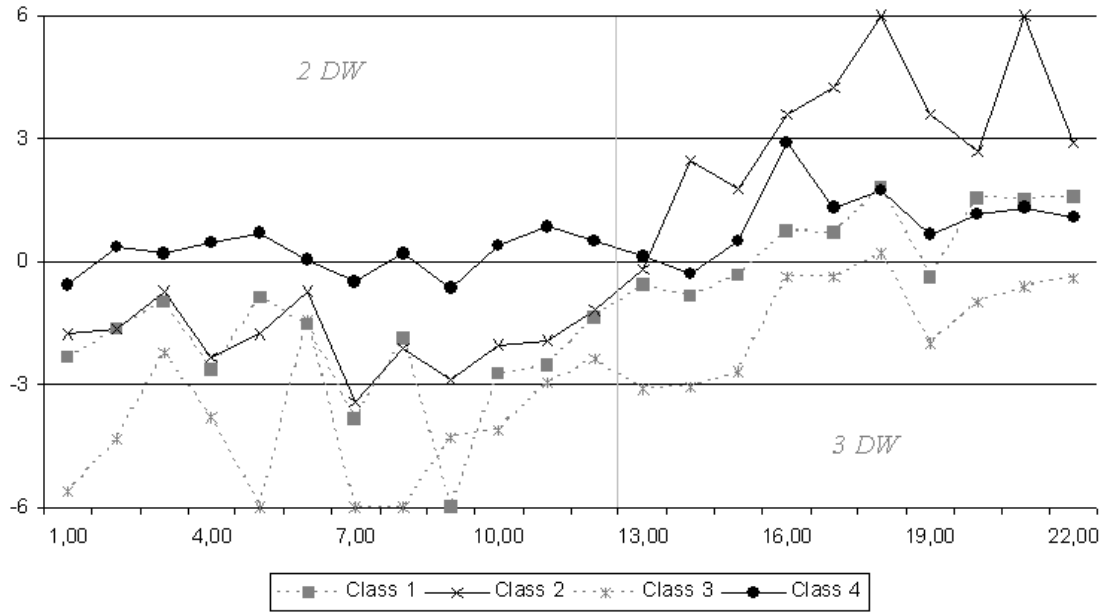
Der Reihenfolge der drei Teilfragen folgend werden zunächst die Ergebnisse einer gemeinsamen *LCA* für die Untertests *2 DW* und *3 DW* dargestellt. Im ersten Schritt wurden dabei die Lösungsprofile aller 348 im Datensatz verbliebenen Versuchspersonen berücksichtigt. Bei der angestrebten Analyse ist nicht a priori klar, wie viele Klassen für die Modellierung von unterschiedlichem Antwortverhalten (insbesondere im Sinne von Bearbeitungsstrategien) angemessen sind. Die obigen Ausführungen zu Bearbeitungsstrategien legen nahe, dass es mindestens zwei prinzipiell unterschiedliche Antwortmuster geben sollte. In der Untersuchung von Köller et al. (1994) hat eine 6-Klassen-Lösung die besten Modellanpassungswerte ergeben. Dabei muss berücksichtigt werden, dass die genannte Untersuchung mit einer erheblich größeren Stichprobe (und somit auch ggf. mit mehr qualitativ unterschiedlichem Antwortverhalten) arbeiten konnte.

Aufgrund dieser Ausgangslage scheint es angemessen zu sein, *LCA*-Lösungen mit verschiedenen Klassenanzahlen zu generieren und anschließend zu vergleichen. Dementsprechend wurden sieben *LCAs* mit 2 bis 8 Klassen (mit dem Programmpaket *WINMIRA*) gerechnet. Anschließend wurden die im *Bootstrapping* gewonnen Kennwerte für die Modellanpassung sowie das informationstheoretische Maß *BIC* für die verschiedenen Modelle miteinander verglichen (vgl. Kap. 4.2.2), wobei sich eine 4-Klassen-Lösung tendenziell als am besten passend herausstellte, ohne bereits befriedigende Werte für die Modellanpassung aufzuweisen. Bei einer Analyse der Lösungsprofile von Versuchspersonen, die besonders schlecht in das Modell passen, zeigte sich eine große Abweichung vom Modell mit

inhaltlich nicht plausiblen Profilen bei 15 Versuchspersonen (ca. 4 %). Nach Ausschluss dieser 15 Versuchspersonen zeigten sich für die verbleibende Stichprobe (n = 333) sehr gute Anpassungswerte.

Die Itemschwierigkeiten für die vier Klassen werden in Abb. 5.23 graphisch dargestellt. Dabei muss berücksichtigt werden, dass hohe Werte auf der Logit-Skala (vertikale Achse) auf eine hohe Aufgabenschwierigkeit und somit auf eine geringe Lösungshäufigkeit in der jeweils dargestellten Klasse hinweisen. Ein hoher Kurvenverlauf deutet also auf geringe Lösungsquoten für die entsprechenden Items hin. „Holistiker" sollten dabei relativ bessere Lösungsquoten beim *3 DW* aufweisen, „Analytiker" hingegen beim *2 DW*. Die ersten zwölf dargestellten Items stammen aus dem Untertest *2 DW*, die weiteren zehn aus dem Untertest *3 DW*.

Abbildung 5.23: Itemschwierigkeiten für 4 latente Klassen (Ergebnisse einer gemeinsamen LCA der Untertests 2 DW und 3 DW; n = 333)

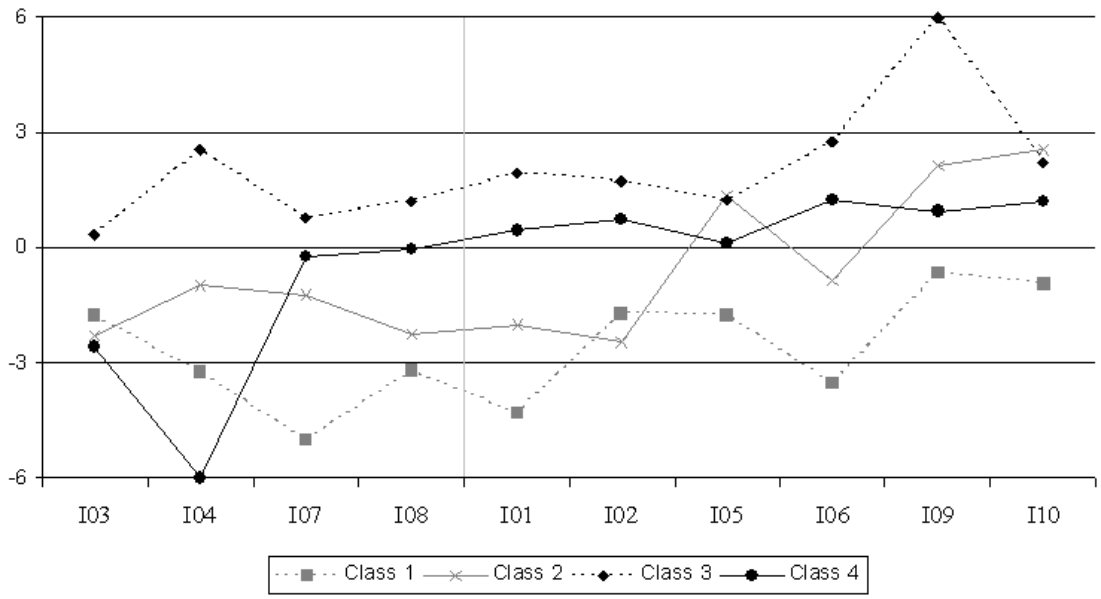


Die Klassen 1 und 3 unterscheiden sich vor allem durch unterschiedlich hohe Lösungsquoten. In Klasse 1 wurden durchschnittlich 14,30 von 22 Items (*2 DW* und *3 DW* zusammen) richtig bearbeitet, in Klasse 3 durchschnittlich 18,87. Die Kurven verlaufen überwiegend parallel. Bei den Klassen 2 und 4 zeigt sich hingegen deutlich der fragliche Strategieunterschied. Obwohl in Klasse 4 durchschnittlich nur 8,42 vom 22 Items richtig bearbeitet wurden, sind die eigentlich relativ schwierigen *3 DW*-Items (2,88 von 10 Items richtig) in dieser Klasse kaum schwieriger als die eigentlich relativ leichten *2 DW*-Items. Umgekehrt ist in Klasse 2 zwar die durchschnittliche Lösungsquote mit 11,11 von 22 Items etwas höher, aber die *3 DW*-Items sind hier besonders schwierig (0,96 von 10 Items richtig). Die Daten

deuten also darauf hin, dass in Klasse 2 vor allem „Analytiker“ sind, während in Klasse 4 vor allem „Holistiker“ sind. In den beiden anderen Klassen sind ggf. „Strategieflexible“ mit unterschiedlich guter Performanz.

Die zweite zu beantwortende Teilfrage bezieht sich auf den Untertest *MRT*. Von den 10 verwendeten Items sind die Falschlösungen bei den Items 3, 4, 7 und 8 grundsätzlich formverschieden, während es sich bei den Items 1, 2, 5, 6, 9 und 10 um Spiegelbilder handelt. Für „Nonrotators“ müssten die Items 3, 4, 7 und 8 einfacher sein als die übrigen Items. Auch für den *MRT* wurden mehrere *LCAs* mit 2 bis 6 Klassen gerechnet, wobei die Kennwerte wiederum für eine 4-Klassen-Lösung sprechen. Gute Modellanpassungswerte ergaben sich allerdings erst nach Entfernen von 4 der 225 *MRT*-Versuchspersonen aus der Stichprobe (ca. 2 %). Die Profile der Itemschwierigkeiten in Abb. 5.24 sind wie in Abb. 5.23 zu lesen. Die Items wurden dabei nach den beiden interessierenden Itemtypen sortiert.

Abbildung 5.24: Itemschwierigkeiten für 4 latente Klassen (n = 221)



Klasse 4 ist die einzige Klasse, in der jedes der Items 3, 4, 7 und 8 leichter ist als alle Items des Typs „Spiegelbilder“. In den anderen Klassen spielen die vier Items des Typs „Formverschieden“ hingegen keine Sonderrolle. Die Klassen 1, 2 und 3 stellen möglicherweise nur verschiedene Leistungsklassen mit Lösungsquoten 8,81 von 10 Items (Klasse 1), 6,20 von 10 Items (Klasse 2) und 1,70 von 10 Items (Klasse 3) dar. Dabei liegt Klasse 3 insgesamt auf dem Niveau der Ratewahrscheinlichkeit (eine von sechs möglichen Zweikombinationen ist richtig). Die Lösungsquote bei den „Nonrotators“ (Klasse 4) liegt hingegen bei durchschnittlich 4,97 von 10 Items, wobei durchschnittlich 3,01 der 4 Items vom Typ „Formverschieden“ und nur 1,96 von 6 anderen Items richtig bearbeitet werden.

Mit den verschiedenen *LCAs* ist es vermutlich[111] gelungen, „Nonrotators“ beim *MRT* und „Analytiker“ bei den Würfelaufgaben zu identifizieren. Gemäß den angestellten Vorüberlegungen müssten diese beiden Gruppen weitgehend aus denselben Versuchspersonen bestehen. Ein Blick auf die Gruppengrößen zeigt, dass die Gruppe der „Nonrotators“ ca. 13,6 % groß ist und die Gruppe der „Analytiker“ ca. 25,2 %. Da der *MRT* nur von einer Teilstichprobe bearbeitet wurde, soll nun ausgehend vom *MRT* untersucht werden, ob „Nonrotators“ tatsächlich überdurchschnittlich häufig zur Gruppe der „Analytiker“ zählen. Dabei ist zu berücksichtigen, dass die Fallzahlen relativ klein sind (Tab. 5.10).

Tabelle 5.10: Zusammenhang zwischen „Nonrotators“ und „Analytikern“

	andere	**Analytiker**	
andere	144	37	181
Nonrotators	14	15	29
	158	52	210

- Zwar sind die Fallzahlen mit insgesamt 29 „Nonrotators“ relativ klein, dennoch zeigt sich, dass in dieser Stichprobe[112] von den „Nonrotators“ ca. 51,7 % auch in der Klasse der „Analytiker“ sind, während von den anderen Versuchspersonen nur ca. 20,4 % „Analytiker“ sind.
- Liest man die Vierfeldertafel spaltenweise, so zeigt sich, dass ca. 28,8 % der „Analytiker“ auch „Nonrotators“ sind, während von den anderen Versuchspersonen nur ca. 8,9 % „Nonrotators“ sind.

Diese Ergebnisse können so interpretiert werden, dass sich die inhaltlich plausiblen Zusammenhänge zwischen „Nonrotators“ und „Analytikern“ auch in empirischen Tendenzen niederschlagen. Jedoch ist der Zusammenhang, der in der obigen Analyse identifiziert wurde, nicht besonders eng und die Aussagekraft der Analyse aufgrund der geringen Fallzahlen begrenzt. Möglicherweise können die theoretisch herausgearbeiteten Zusammenhänge in einer anders strukturierten und vor allem größeren Stichprobe durch die statistische Analyse von Lösungsprofilen auch empirisch besser nachgewiesen werden. Alternativ

[111] An dieser Stelle muss betont werden, dass alle hier geäußerten Vermutungen und Schlüsse zunächst nur statistisch plausibel sind. Möchte man eine größere Gewissheit über die tatsächlichen Bearbeitungsstrategien von Versuchspersonen haben, so müssen ergänzend z. B. qualitative Verfahren wie Interviews („Methode des lauten Denkens“) eingesetzt werden, bei denen die Versuchspersonen selbst Auskunft über ihre mentalen Prozesse geben.

[112] Die Stichprobe besteht in diesem Fall aus den 210 Versuchspersonen, die den Untertest *MRT* bearbeitet haben und die bei keiner der *LCAs* ausgesondert wurden.

dazu – oder auch ergänzend – können die vermuteten Zusammenhänge auch in einer kleinen Stichprobe mit qualitativen Methoden untersucht werden.

Damit ist die Konstruktvalidität des *MRT* theoretisch und empirisch leicht eingeschränkt. Geiser et al. (2006, S. 282 ff.) schlagen vor, dass ggf. nur noch Items des Typs „Spiegelbild“ verwendet werden sollten. In der Hauptuntersuchung der vorliegenden Arbeit wird trotzdem der unveränderte Untertest aus der Voruntersuchung eingesetzt. Dies hat im Wesentlichen drei Gründe:

- Zwar lassen sich im Rahmen von *LCAs* Klassifizierungen der Versuchspersonen erzielen, die auf unterschiedliche Bearbeitungsstrategien hindeuten, dennoch weist das eindimensionale *RM* sehr gute Modellanpassungswerte auf, sodass Leistungswerte mit guter empirischer Qualität gewonnen werden können. Zwar würde auch ein *MRM*, das die *LCA* mit dem *RM* verbindet, näherungsweise passen, aber mit schlechteren Anpassungswerten als das *RM*.
- Bei etwaigen Veränderungen des Untertests müssten zunächst weitere Voruntersuchungen durchgeführt werden, um auszuschließen, dass die empirische Güte des Untertests nicht leidet. Aus pragmatischer Sicht ist dieser zusätzliche Aufwand für die Ziele der vorliegenden Arbeit nicht erforderlich.
- Die Befunde aus anderen Studien, die inhaltlich in der vorliegenden Arbeit berücksichtigt werden, beruhen ebenfalls auf der heterogenen *MRT*-Version. Da hier eine Vergleichbarkeit bestehen bleiben soll, wird der klassische *MRT* verwendet.

5.4.3 Vertiefende Analysen zur Vorbereitung der Hauptuntersuchung

Die Auswertungen der Voruntersuchungen werden mit zwei inhaltlich vertieften Analysen abgeschlossen. Zunächst soll vor dem Hintergrund der in Kap. 3.3.2 zusammengefassten Befunde analysiert werden, ob bzw. wie stark ausgeprägte Geschlechterunterschiede in der *Raumvorstellung* vorliegen. Aus den in Kap. 2.3.2 dargestellten Gründen („*TIMSS/II*-Paradoxon“) muss dabei die Klumpung der Stichprobe in Schulformen berücksichtigt werden. Die in der Literatur berichteten Zusammenhänge zwischen *Raumvorstellung* und *Mathematikleistung* sollen schließlich über vorsichtige Zusammenhangsbetrachtungen mit Fachnoten untersucht werden.

Geschlechterunterschiede bei den Raumvorstellungstests

Die Fragen nach Geschlechterunterschieden in bestimmten Testleistungen kann deskriptiv über Mittelwerte, Standardabweichungen und Standardfehler betrachtet und z. B. mit einer Varianzanalyse zufallskritisch abgesichert werden (vgl. Kap. 4.2.2). Eine zweifaktorielle *ANOVA* bietet die Möglichkeit den Einfluss der Faktoren „Geschlecht“ und „Schulform“ gleichzeitig zu analysieren und dabei mögliche Interaktionseffekte zu berücksichtigen (z. B. „Stellen sich Geschlechterunterschiede in einer Schulform grundlegend anders dar als in der anderen?“).

Da die Testleistungen mit *IRT*-Verfahren skaliert wurden, bietet es sich an, die *WLE*-Schätzungen der jeweiligen Leistung als Zielvariable zu verwenden und nicht etwa die Gesamtscores für die jeweiligen Untertests. Mittelwertunterschiede und Effektstärken können messfehlerfrei bestimmt werden, indem sie als Parameter mitgeschätzt werden. Je nach verwendetem Programmpaket müssen dabei Varianzanalysen auf geeignete Regressionsanalysen zurückgeführt werden. Im Folgenden wird für die fünf Untertests zur *Raumvorstellung* jeweils separat dargestellt, welche Effekte signifikant sind und welche Effektstärken dabei vorliegen. Entsprechende Analysen sind mit *SPSS* (mit den *WLE*-Schätzungen der Testleistung) und *ConQuest* durchgeführt worden.

Auf der Basis der Befunde, die in Kap. 3.3 dargestellt wurden und der ersten Auswertungen der Voruntersuchungen, sind folgende Effekte zu erwarten:

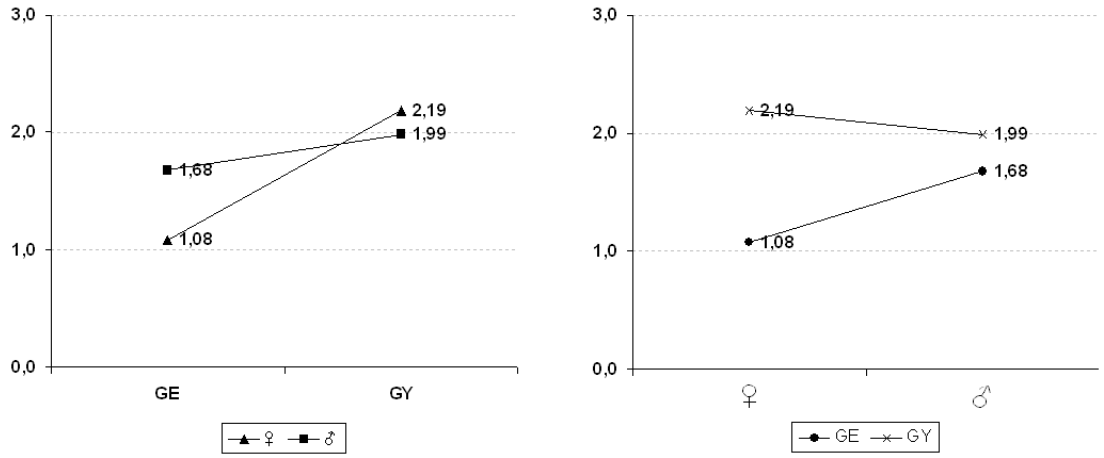
- signifikante Geschlechterunterschiede mit mittleren bzw. großen Effektstärken zugunsten der Jungen in den *WLT*- und *MRT*-Leistungen;
- keine signifikanten Geschlechterunterschiede in der *DAT:SR*-Leistung;
- keine signifikanten Geschlechterunterschiede bei der *2 DW-Leistung* (nicht zuletzt aufgrund des starken Deckeneffekts dieses Tests);
- ggf. signifikante Geschlechterunterschiede zugunsten der Jungen in der *3 DW*-Leistung (aufgrund des Anteils an *mentaler Rotation*) ;
- über alle Untertests hinweg signifikante Leistungsunterschiede zwischen den Schulformen (zugunsten der Gymnasien);
- keine signifikanten Interaktionseffekte zwischen „Geschlecht“ und „Schulform“.

An dieser Stelle muss betont werden, dass es ausdrücklich kein Ziel der vorliegenden Arbeit ist, verallgemeinerbare Aussagen über Schulformunterschiede zu generieren. In der Voruntersuchung sind nur eine Gesamtschule und zwei Gymnasien vertreten; für verallgemeinerbare Aussagen müsste eine erheblich größere und aufwändig gezogene Stichprobe realisiert werden. In dieser Arbeit steht nur der Zusammenhang zwischen kognitiven Leistungen, selbstbezogenen Kognitionen und dem Geschlecht der Versuchspersonen im Vordergrund. Die Berücksichtigung verschiedener Schulformen geschieht in der Absicht, eine angemessene Varianz der kognitiven Leistungen sicherzustellen. Bei der angestrebten Varianzanalyse müssen die Klumpen der Stichprobe aber berücksichtigt werden, da diese Struktur der Stichprobe möglicherweise Geschlechterunterschiede überlagert oder erst hervorbringt.

Die oben aufgelisteten, vermuteten Ergebnisse werden nun in umgekehrter Reihefolge betrachtet, da signifikante Interaktionseffekte die Interpretierbarkeit der Haupteffekte durch

die Faktoren beeinflussen.[113] Erwartungswidrig ergibt sich für den Untertest *2 DW* ein signifikanter Interaktionseffekt, der in Abb. 5.25 veranschaulicht wird:

Abbildung 5.25: Interaktionseffekt zwischen „Schulform“ und „Geschlecht“ beim 2 DW



Die graphischen Darstellungen der Mittelwerte nach den jeweiligen Faktorstufen („Interaktionsdiagramme“) zeigen deutlich, dass kein interpretierbarer Haupteffekt für den Faktor „Geschlecht“ vorliegt. Während Jungen an der Gesamtschule bessere Testleistungen zeigen als Mädchen, zeigen Mädchen an den Gymnasien bessere Testleistungen als Jungen. Welches Geschlecht „besser“ ist, hängt also vom betrachteten Klumpen der Stichprobe ab. Für den Faktor „Schulform“ zeigt sich hingegen ein (statisch signifikanter) Haupteffekt. So erzielen sowohl Mädchen als auch Jungen an den Gymnasien bessere Testleistungen als an der Gesamtschule, wobei der Unterschied zwischen den Schulformen für Mädchen deutlich stärker ausgeprägt ist. Da der Untertest *2 DW* nicht eindeutig eine der in dieser Arbeit fokussierten Komponenten der *Raumvorstellung* operationalisiert und daher auch in der Hauptuntersuchung keine Rolle mehr spielt, wird die Auswertung hier nicht weiter vertieft.

Bei den anderen vier Untertests zeigen sich erwartungskonform keine signifikanten Interaktionseffekte und – ebenfalls erwartungskonform – signifikante Schulformeffekte zugunsten des Gymnasiums. Die hier besonders interessierenden Geschlechtereffekte werden für die verbleibenden vier Untertests in Tab. 5.11 (aus Sicht der Jungen) dargestellt:

[113] Wenn z. B. für einen Untertest in einer Schulform die Mädchen signifikant besser wären als die Jungen und in der anderen Schulform die Jungen signifikant besser als die Mädchen, dann liegt offensichtlich kein signifikanter Haupteffekt des Faktors *Geschlecht* vor. Die Effekte dieses Faktors sind vielmehr nur in Abhängigkeit vom anderen Faktor inhaltlich interpretierbar.

Tabelle 5.11: Geschlechterunterschiede in der Raumvorstellung bei den Untertests WLT, MRT, DAT:SR und 3 DW (aus Sicht der Jungen)

	Stichprobe	**Effektstärke**	**Signifikanz**
WLT	n = 348	**0,62**	ja (p = 0,000)
MRT	n = 225	**0,52**	ja (p = 0,001)
DAT:SR	n = 219	**–0,35**	ja (p = 0,046)
3 DW	n = 348	–0,04	nein (p = 0,697)

Die Ergebnisse der Voruntersuchung zu Geschlechterunterschieden sind also nur zum Teil erwartungskonform:

- Eingetreten sind die signifikanten Unterschiede (mit „mittlerem“ Effekt) zugunsten der Jungen beim *WLT* und beim *MRT*, allerdings unerwartet (vgl. Kap. 3.3.2) mit der größeren Effektstärke beim *WLT*.
- Nicht erwartungskonform ist der signifikante Unterschied zugunsten der Mädchen beim *DAT:SR* (mit „kleinem Effekt“). Bei der Meta-Analyse von Linn & Petersen (1985) ergab sich für *räumliche Visualisierung* kein signifikanter Effekt, wobei einzelne in die Meta-Analyse eingegangene Studien ein ähnliches Bild wie die eigene Voruntersuchung ergaben.
- Ebenfalls zunächst nicht erwartungskonform, aber nach den Ergebnissen zum *DAT:SR* durchaus plausibel, ist das Fehlen signifikanter Effekte beim *3 DW*. Hier scheinen sich die Effekte *mentaler Rotation* und *räumlicher Visualisierung* im Sinne der zuvor genannten Ergebnisse gegenseitig zu kompensieren.

Die Ergebnisse der Voruntersuchung zu Geschlechterunterschieden in den verschiedenen Untertests zur *Raumvorstellung* zeigen zum einen, dass solche Unterschiede durchaus mit bedeutenden Effektstärken auftreten, und zum anderen, dass die Unterschiede vor allem in den Tests *WLT*, *MRT* und *DAT:SR* auftreten. Daher dürften diese Tests inhaltlich am interessantesten für die Hauptuntersuchung sein; dies gilt umso mehr, als mit dem *DAT:SR* in der Voruntersuchung auch ein Test vorliegt, bei dem Effekte zugunsten der Mädchen beobachtet werden konnten. Insgesamt deutet sich an, dass mit den drei für die Hauptuntersuchung ausgewählten Tests der Zusammenhang von *Raumvorstellung* und *Mathematikleistung* differenziert untersucht werden kann.

Zusammenhänge zwischen Raumvorstellung und Schulnoten

In der Hauptuntersuchung wird die Leistung in den dort verwendeten Raumvorstellungstests im Zusammenhang mit der Testleistung in Mathematik ausgewertet. In Kap. 3.3.5 wurde dargestellt, dass von einem entsprechenden substanziellen Zusammenhang ausge-

gangen werden kann, auch wenn dieser differenziert betrachtet werden muss. Ob ein solcher Zusammenhang potenziell für die Untertests der Voruntersuchung besteht, kann in erster Näherung durch den Zusammenhang mit Fachnoten untersucht werden. Dabei muss natürlich berücksichtigt werden, dass Fachnoten am ehesten eine Aussage über die leistungsbezogene Position eines Schülers bzw. einer Schülerin innerhalb der jeweiligen Lerngruppen machen. Wie Dicker (1977) speziell für den Mathematikunterricht gezeigt hat, ist die Note als Lehrerurteil über die Anordnung der Schülerinnen und Schüler nach ihrer Leistung innerhalb einer Lerngruppe recht zuverlässig. Allerdings sind Noten schon innerhalb einer Schule zwischen den Lerngruppen kaum vergleichbar.

Wäre man ausschließlich auf Fachnoten als Indikator für Fachleistung angewiesen, dann würde sich eine Modellierung von Zusammenhängen anbieten, die berücksichtigt, dass Schülerinnen und Schüler in Lerngruppen zusammengefasst sind, die in Schulen zusammengefasst sind, die wiederum in Schulformen zusammengefasst sind. Solche „Mehrebenenmodelle“ oder „Hierarchisch lineare Modelle“ sind heute in der empirischen Bildungsforschung bei entsprechenden Fragestellungen üblich. Für die Auswertung der Voruntersuchung kann methodisch allerdings auch bescheidener agiert werden, da im Rahmen der Hauptuntersuchung die *Mathematikleistung* durch einen externen Test lerngruppenübergreifend vergleichbar ist. An dieser Stelle soll nur überprüft werden, inwieweit die Mathematiknote mit den Raumvorstellungsleistungen auf Ebene der jeweiligen Lerngruppe zusammenhängen – bzw., anders gefragt, ob die Mathematiknote und die Raumvorstellungsleistung eine ähnliche Anordnung der Schülerinnen und Schüler einer Lerngruppe ergeben.

Eine solche heuristische Plausiblitätskontrolle wird im Folgenden für die fünf Untertests zur *Raumvorstellung* und die Noten in den Fächern Deutsch, Mathematik und Sport durchgeführt. Dabei sollen die Noten in den Fächern Deutsch und Sport als Kontrolle für die Ergebnisse mit den Mathematiknoten dienen. Der Bezugsrahmen „Lerngruppe“ wird statistisch realisiert, indem innerhalb jeder Lerngruppe sowohl die Noten als auch die Leistungen (*WLE*-Schätzungen) in den Raumvorstellungstests z-standardisiert werden (vgl. Büchter & Henn, 2007, S. 104). Streng genommen ist hierfür jeweils ein Intervallskalenniveau erforderlich, was eigentlich nur für die skalierte Testleistung zutrifft. Die Erfahrungen in der empirischen Bildungsforschung zeigen aber, dass keine gravierenden Verzerrungen entstehen, wenn auch für Fachnoten ein Intervallskalenniveau angenommen wird.

Durch die z-Standardisierung innerhalb jeder Lerngruppe liegt nun für jeden Schüler ein relativer Leistungswert in den fünf Untertests zur *Raumvorstellung* und ein relativer Notenwert für die drei Fächer vor. Dabei wurden die Noten vom Ende der Klasse 8 und dem Ende des ersten Halbjahres Klasse 9 gemittelt (was eigentlich wiederum Intervallskalenniveau voraussetzt). Durch diese Modellierung wurde die spezielle Struktur der Stichprobe hinreichend berücksichtigt, sodass die linearen Zusammenhänge nun über alle Versuchspersonen hinweg berechnet werden können. Die entsprechenden Korrelationskoeffizienten werden in Tab. 5.11 berichtet.

Tabelle 5.12: Korrelationskoeffizienten für innerhalb der Lerngruppen z-standardisierte Fachnoten und Testleistungen

	WLT	***MRT***	***DAT:SR***	*2 DW*	*3 DW*
Deutsch	–0,04 (n = 340)	–0,03 (n = 220)	–0,12 (n = 211)	–0,05 (n = 340)	**–0,17** (n = 340)
Mathematik	**–0,19** (n = 336)	**–0,24** (n = 217)	**–0,22** (n = 210)	**–0,19** (n = 336)	**–0,20** (n = 336)
Sport	–0,05 (n = 338)	–0,09 (n = 215)	**0,17** (n = 212)	–0,02 (n = 338)	–0,05 (n = 338)

In Tab. 5.12 wurden die rechnerischen Maße für Zusammenhänge, die nicht von primärem Interesse sind, grau unterlegt und signifikant von Null unterschiedliche Korrelationskoeffizienten fett gesetzt. Bei den Korrelationskoeffizienten muss beachtetet werden, dass die Notenskala negativ gepolt ist (kleine Werte entsprechen guten Leistungen), sodass negative Werte positive Leistungszusammenhänge anzeigen. Insgesamt zeigen sich schwach ausgeprägte Zusammenhänge, die überwiegend erwartungskonform sind. Die Zusammenhänge mit der Mathematiknote sind für alle Untertests am engsten und alle signifikant von Null verschieden. Ob die beiden anderen signifikanten Werte (Deutsch x *3 DW* und Sport x *DAT:SR*) zufallsbedingt aufgetreten sind oder auch inhaltlich substanziell sind, kann hier nicht aufgeklärt werden, ist für die Fragestellung der vorliegenden Arbeit aber auch ohne Bedeutung. Insgesamt deutet sich an, dass Zusammenhänge zwischen *Raumvorstellung* und *Mathematikleistung* mit den verwendeten Tests erfasst werden können und dass Leistungen im sprachlichen Bereich, hier repräsentiert durch die Deutschnoten, nicht oder nur schwach mit *Raumvorstellung* zusammenhängen. Dieses Ergebnis passt u. a. zu den Befunden von Lehmann et al. (2002).

Zusammenfassend lässt sich feststellen, dass die Hauptuntersuchung im Bereich der *Raumvorstellung* hinreichend durch die Voruntersuchung vorbereitet wurde. Dies gilt sowohl für die Instrumentenentwicklung als auch für die inhaltliche Vorbereitung. Für die Hauptuntersuchung bleibt zunächst der Bereich *Denkstile* kritisch, da weiterhin unklar ist, ob er durch einen *Paper and Pencil Test* angemessen operationalisiert werden kann. Die erforderliche neue Skala für das *FSK:M* kann hingegen aus den großen Schulleistungsstudien übernommen werden.

6 Anlage und Befunde der Hauptuntersuchung

Die Hauptuntersuchung basiert inhaltlich vor allem auf den Befunden, die in den Kapiteln 2 und 3 systematisch zusammengestellt worden sind. Im Rahmen der Voruntersuchung sind die Befunde aus Kapitel 3 u. a. mithilfe der Instrumente vertieft und ergänzt worden, die auch in der Hauptuntersuchung zur Erfassung der *Raumvorstellung* eingesetzt werden sollen. Insgesamt ergibt die vorliegende Literatur zusammen mit den Befunden der Voruntersuchung ein relativ konsistentes Bild, sodass – wie geplant – die Fragestellung der vorliegenden Arbeit mit der Hauptuntersuchung weiter verfolgt werden kann. Forschungsmethodisch knüpft die Hauptuntersuchung an die statistische Auswertung der Voruntersuchung an und nutzt die Modelle und Verfahren, die sich dort bewährt haben. Mit den Raumvorstellungstests *WLT*, *MRT* und *DAT:SR* liegen Instrumente vor, die inhaltlich klar und empirisch hinreichend trennscharf die drei Komponenten des Konstrukts *Raumvorstellung* erfassen, das in Kap. 4.1.2 festgelegt wurde.

Im Kern der Hauptuntersuchung werden die oben genannten Raumvorstellungstests zusammen mit einem fachdidaktisch und psychometrisch abgesicherten Mathematiktest, den nordrhein-westfälischen Lernstandserhebungen in der Jahrgangsstufe 9 (*LSE 9*), eingesetzt und ausgewertet. Insgesamt sollten auf diesem Weg für die „*PISA*-Population“ belastbare Ergebnisse zum Zusammenhang zwischen *Raumvorstellung* und *Mathematikleistung* unter Berücksichtigung möglicher Geschlechterunterschiede (und ggf. weiterer Faktoren) erzielt werden können. Im Folgenden werden zunächst die Hypothesen, die der Hauptuntersuchung zugrunde liegen, und Fragen, die im Rahmen der Hauptuntersuchung exploriert werden sollen, präzisiert. Anschließend werden die verwendeten Instrumente vorgestellt, wobei nur die *LSE 9* differenzierter dargestellt werden müssen. Nach der Dokumentation der Durchführung und Auswertung der Hauptuntersuchung werden die Grundauswertungen, die vertiefenden Analysen und die zugehörigen Befunde systematisch dargestellt.

6.1 Zugrundeliegende Hypothesen

Die Befunde aus Kapitel 3 und der Voruntersuchung deuten darauf hin, dass innerhalb der *Small-Scale Fähigkeiten* der *Raumvorstellung* vor allem bei der *mentalen Rotation* substanzielle Geschlechterunterschiede *und* substanzielle Zusammenhänge mit *Mathematikleistung* festgestellt werden können. Folglich hat die *mentale Rotation* das Potenzial, Geschlechterunterschiede in der *Mathematikleistung* statistisch zu erklären. Weniger stark ausgeprägt, aber noch klar feststellbar scheinen Geschlechterunterschiede bei der Komponente *räumliche Wahrnehmung* zu sein. Hingegen sind im Bereich *räumliche Visualisierung* zwar allgemein keine signifikanten Geschlechterunterschiede (vgl. Kap. 3.3.2) zu erwarten, möglicherweise aber doch mit geringen Effektstärken zugunsten der weiblichen Versuchspersonen beim konkret verwendeten *DAT:SR* (vgl. Voruntersuchung, Kap. 5.4.3). Dementsprechend dürfte *räumliche Visualisierung* kaum als Mediator für Geschlechterun-

terschiede in der *Mathematikleistung* infrage kommen. Dies bedeutet allerdings nicht, dass *räumliche Visualisierung* und *Mathematikleistung* nicht miteinander zusammenhängen. Der Zusammenhang sollte aufgrund der analytischen Bearbeitungsmöglichkeiten in beiden Bereichen sogar relativ eng sein – er erfüllt nur die oben angesprochene Mediatorfunktion nicht.

Inwieweit unterschiedliche *Denkstile* zu Unterschieden in der *Raumvorstellung* und der *Mathematikleistung* beitragen, kann nur dann untersucht werden, wenn sich das entsprechende Instrument mit modifizierten Instruktionen in der Hauptuntersuchung hinreichend bewährt. Ein verändertes Instrument des *FSK:M* sollte gewährleisten, dass dieses Konstrukt bei der empirischen Analyse der entsprechenden Modelle berücksichtigt werden kann. Die folgenden Hypothesen und das Explorationsanliegen werden nicht mehr im Einzelnen hergeleitet, da sie sich an die Kapitel 2 und 3 sowie die Auswertung der Voruntersuchung (Kap. 5.4) anschließen.

Die Hypothesen, die der Hauptuntersuchung zugrunde liegen, werden nach Untersuchungsbereichen getrennt zusammengestellt und im Sinne inhaltlich erwarteter Ergebnisse formuliert. Bei einer formalen Präzisierung für die empirische Überprüfung im Rahmen eines Hypothesentests müsste jeweils, soweit möglich, das logische Gegenteil der hier formulierten Hypothesen als Nullhypothese verwendet werden.

Wenn die inhaltlich formulierte Hypothese z. B. durch von Null verschiedene Korrelationskoeffizienten operationalisiert wird, besteht das logische Gegenteil einfach in der Annahme, der entsprechende Kennwert sei gleich Null. Diese Nullhypothese („$r = 0$“) kann dann auf dem jeweils festgelegten – d. h. in der Regel auf dem mit 5 % üblichen – Signifikanzniveau verworfen werden, wenn die erwarteten Ergebnisse eintreten und der Test genügend *Testpower* hat (vgl. Büchter & Henn, 2007, Kap. 4.2).

Geht die inhaltlich formulierte Hypothese dagegen davon aus, dass keine Unterschiede (im Sinne nicht von Null verschiedener Kennwerte) zu erwarten sind, dann kann das logische Gegenteil nicht einfach mithilfe einer konkreten Kennwertausprägung formuliert werden. In einer solchen Situationen könnte man sich – etwa im Falle eines Korrelationskoeffizienten – mit zwei Nullhypothesen der Art „$r < -0{,}1$“ oder „$r > 0{,}1$“ behelfen, was ausdrücken würde, dass man Korrelationskoeffizienten aus dem Intervall $[-0{,}1; 0{,}1]$ nicht als substanziell von Null verschieden betrachtet. Wenn beide genannten Nullhypothesen aufgrund der Daten und des festgelegten Signifikanzniveaus verworfen werden können, kann man im Sinne des Hypothesentests von einer vorläufigen Bestätigung seiner inhaltlich formulierten Hypothese ausgehen.

6.1.1 Hypothesen zur Raumvorstellung

H_01: Bei den Raumvorstellungstests *WLT* und *MRT* gibt es signifikante Leistungsunterschiede zugunsten der männlichen Versuchspersonen.

H_02: Beim Raumvorstellungstest *DAT:SR* gibt es signifikante Leistungsunterschiede zugunsten der weiblichen Versuchspersonen.

H_03: Der Zusammenhang zwischen je zwei der drei Raumvorstellungstests ist sowohl substanziell von Null verschieden als auch substanziell von Eins verschieden.

6.1.2 Hypothesen zur Mathematikleistung

H_04a: Beim *LSE 9*-Mathematiktest gibt es signifikante Leistungsunterschiede zugunsten der männlichen Versuchspersonen.

H_04b: Etwaige signifikante Leistungsunterschiede zugunsten der männlichen Versuchspersonen beruhen beim *LSE 9*-Mathematiktest auf entsprechenden Unterschieden im Bereich der (empirisch) schwierigeren Items.

H_04c: Etwaige signifikante Leistungsunterschiede zugunsten der männlichen Versuchspersonen beruhen beim *LSE 9*-Mathematiktest vor allem auf entsprechenden Unterschieden bei *rechnerischen Modellierungsaufgaben*.

6.1.3 Hypothese zum Fähigkeitsselbstkonzept Mathematik

H_05: Das *FSK:M* ist bei männlichen Versuchspersonen signifikant höher als bei weiblichen Versuchspersonen.

6.1.4 Hypothesen zum Zusammenhang der Konstrukte

H_06a: Die Leistung in allen Raumvorstellungstests hängt (positiv) mit der *Mathematikleistung* zusammen.

H_06b: Der Zusammenhang zwischen *Raumvorstellung* und *Mathematikleistung* ist für den *MRT* und den *DAT:SR* stärker als für den *WLT*.

H_06c: Der Zusammenhang zwischen der *MRT*-Leistung und der *Mathematikleistung* ist für empirisch schwierige *LSE 9*-Items enger als für empirisch leichte Items.

H_06d: Der Zusammenhang zwischen der *MRT*-Leistung und der *Mathematikleistung* ist für *rechnerische Modellierungsaufgaben* enger als für *technische Aufgaben*.

H_07a: Geschlechterunterschiede in der *Mathematikleistung* können nicht durch entsprechende Unterschiede in der *DAT*-Leistung erklärt werden.

H_07b: Geschlechterunterschiede in der *Mathematikleistung* können in geringerem Umfang durch entsprechende Unterschiede in der *WLT*-Leistung erklärt werden.

H_07c: Geschlechterunterschiede in der *Mathematikleistung* können in großen Anteilen durch entsprechende Unterschiede in der *MRT*-Leistung erklärt werden.

H_07d: Die Mediatorfunktion der *MRT*-Leistung für Geschlechterunterschiede in der *Mathematikleistung* ist für empirisch schwierige *LSE 9*-Items stärker als für empirisch leichte Items.

H_07e: Die Mediatorfunktion der *MRT*-Leistung für Geschlechterunterschiede in der *Mathematikleistung* ist für *rechnerische Modellierungsaufgaben* am stärksten.

H_08: In einem gemeinsamen Modell mit dem Konstrukt *Raumvorstellung* gibt es eine eigene Mediatorfunktion des *FSK:M* für Geschlechterunterschiede in der *Mathematikleistung*.

6.1.5 Explorationsanliegen zu Denkstilen

E_01: Für das erweiterte Modell zur Erklärung von Geschlechterunterschieden in der *Mathematikleistung* (vgl. Abb. 4.2, S. 122) ist von zentraler Bedeutung, ob sich *Denkstile* in einem *Paper and Pencil Test* erfassen lassen. Die Auswertung der Hauptuntersuchung nimmt zunächst diese Frage in den Blick.

6.2 Instrumente der Hauptuntersuchung

Mit Ausnahme der *LSE 9* sind die Instrumente der Hauptuntersuchung im Wesentlichen aus der Voruntersuchung bekannt, sodass nur noch etwaige Modifikationen gegenüber der Voruntersuchung dargestellt werden müssen. Der *LSE 9*-Mathematiktest wird hingegen ausführlicher dargestellt.

6.2.1 Instrumente zur Raumvorstellung

In der Voruntersuchung wurden insgesamt fünf Untertests zur *Raumvorstellung* eingesetzt, davon einer mit zwei Testteilen (*WLT*/*RFT*). Für die Ziele der Hauptuntersuchung werden die Untertests *2 DW* und *3 DW* nicht benötigt. Der Testteil *RFT* des Untertests *WLT*/*RFT* hat sich in der Voruntersuchung nicht bewährt, sodass *räumliche Wahrnehmung* in der Hauptuntersuchung nur noch durch den Testteil *WLT* erfasst wird, der sich in der Voruntersuchung hinreichend bewährt hat. Insgesamt werden also die Raumvorstellungstests *WLT*, *MRT* und *DAT:SR* eingesetzt. Dabei wird der *MRT* unverändert eingesetzt, der *WLT* mit einer leicht modifizierten Instruktion,[114] einem Item mehr und neu gezeichneten Items und der *DAT:SR* in einer Kurzform mit 10 Items, die in der Voruntersuchung ausgewählt wurden (vgl. Kap. 5.4.1).

[114] Die Modifikation bezieht sich lediglich darauf, dass explizit mitgeteilt wird, dass die Flaschen verschlossen sind und das Wasser somit nicht aus den Flaschen herauslaufen kann.

6.2.2 Instrument Denkstile

Das Instrument *Denkstile* hatte sich in der Voruntersuchung noch nicht hinreichend bewährt, da zu viele Bearbeitungen nicht auswertbar waren und dies vermutlich zu einer Verzerrung der Ergebnisse führt (vgl. Kap. 5.3.4). Ein weiteres Problem stellte der Bearbeitungsumfang des Untertests in der Voruntersuchung dar. Für die vier Items war eine Bearbeitungszeit von 20 Minuten vorgegeben, die von vielen Versuchspersonen auch annähernd ausgeschöpft wurde. Also muss auch in der Hauptuntersuchung, die maximal 45 Minuten dauern soll, mit ca. 5 Minuten je Item gerechnet werden. Aus Zeitgründen wurde daher für die Hauptuntersuchung ein Untertest *Denkstile* zusammengestellt, der aus nur zwei Items besteht, die beide potenziell sowohl *prädikatives Denken* als auch *funktionales Denken* sichtbar werden lassen können. Die Bearbeitungen der vier Items aus der Voruntersuchung zeigen, dass über alle vier Items hinweg relativ konsistent jeweils ein *Denkstil* diagnostiziert werden konnte, *wenn* eine solche Diagnose überhaupt möglich war. Daher dürfte die Beschränkung auf zwei Items das Diagnosepotenzial kaum verringern.

Die in der Voruntersuchung aufgetretenen Schwierigkeiten bei der Auswertung der Bearbeitungen, sollten in der Hauptuntersuchung durch ergänzende mündliche Instruktionen vermieden werden. So wurden die Schülerinnen und Schüler ausdrücklich dazu ermuntert, nicht nur Verbalisierungen ihrer Überlegungen zu notieren, sondern auch in die Aufgabenstellung hineinzuzeichnen, wenn ihnen dadurch die Beschreibung des eigenen Vorgehens leichter fällt. Auf diesem Weg sollen etwaige Schwierigkeiten beim Verbalisieren kompensiert und der Anteil auswertbarer Bearbeitungen gesteigert werden.

6.2.3 Instrument Lernstandserhebungen (LSE 9)

Die *Mathematikleistung* wird in der Hauptuntersuchung mit den nordrhein-westfälischen Lernstandserhebungen in Klasse 9 (*LSE 9*) erfasst. Die *LSE 9* wurden in Nordrhein-Westfalen im Jahr 2004 erstmalig durchgeführt und sollten mehrere Funktionen erfüllen: Neben ihrer primären Funktion, Leistungsdaten für innerschulische und landesweite Vergleiche von Lerngruppen bereitzustellen, sollten die *LSE 9* auch die im Jahr 2004 veröffentlichten Kernlehrpläne implementieren, indem sie deren Kompetenzerwartungen durch Leistungsaufgaben konkretisieren (vgl. Heymann & Pallack, 2007, S. 14 f.).

Die Kernlehrpläne für die Schulformen der Sekundarstufe I (MSJK, 2004a-d) setzen die KMK-Bildungsstandards (KMK, 2004, 2005a) auf Landesebene um. Sie enthalten für die Schulpraxis zwei wesentliche Neuerungen: Zum einen geben sie anstelle der zu unterrichtenden Inhalte („Inputsteuerung“) Kompetenzerwartungen vor, die zum Ende der Doppeljahrgangsstufen 5/6, 7/8 und 9/10 erreicht werden sollen („Outputsteuerung“). Zum anderen formulieren sie die Erwartungen im Bereich der „prozessbezogenen Kompetenzen“ (*Argumentieren/Kommunizieren*, *Modellieren*, *Problemlösen* und *Werkzeuge nutzen*) ge-

nauso differenziert und verbindlich wie im Bereich der „inhaltsbezogenen Kompetenzen“ (*Arithmetik/Algebra*, *Funktionen*, *Geometrie* und *Stochastik*).

Die *LSE 9* sind durch ihren engen Bezug auf die Kompetenzerwartungen am Ende der Doppeljahrgangsstufe 7/8 ein vollständig curricular valider Test, der die volle Breite der inhaltsbezogenen Kompetenzen berücksichtigt, im Bereich der prozessbezogenen Kompetenzen aber – bedingt durch die o. g. Implementationsfunktion der *LSE 9* – Schwerpunkte setzt. So soll in den Testjahren 2004 ff. jeweils einer der vier Prozessbereiche so differenziert durch Aufgaben erfasst werden, dass ein Kompetenzstufenmodell für diesen Bereich entwickelt werden kann. Der Schwerpunkt im Jahr 2004 lag im Bereich *Modellieren*; die zugehörige Modellieren-Skala wurde bereits in Kap. 2.1.3 (S. 34 f.) dargestellt.

Die *LSE 9* wurden im Jahr 2004 in Haupt-, Gesamt- und Realschulen sowie Gymnasien geschrieben, wobei zwei verschiedene Testhefte (Versionen A und B) mit einem großen Überlappungsbereich zum Einsatz kamen. Dies berücksichtigt, dass in den Kernlehrplänen eine breite schulformübergreifende Basis mathematischer Kompetenz identisch beschrieben wird, es aber auch bildungsgangspezifische Vertiefungen gibt. Beide Testhefte waren für eine Bearbeitungsdauer von 90 Minuten konzipiert und enthielten jeweils 39 Items, wobei in der Regel mehrere Items (im Sinne von Teilaufgaben) zu einem Itemstamm (Aufgabenkontext) gehörten. Von den jeweils 39 Items waren 20 Items in beiden Testversionen identisch („Ankeritems“). Sie wurden durch jeweils 19 Items ergänzt, die in Version A tendenziell leichter und in Version B tendenziell schwieriger als die Ankeritems waren. Version A wurde von Hauptschulen (Grund- und Erweiterungskurse) und Gesamtschul-Grundkursen bearbeitet, Version B von Gesamtschul-Erweiterungskursen, Realschulen und Gymnasien. Aufgrund der Ankeritems ist es rechnerisch möglich, die Testleistung aller Schülerinnen und Schüler auf einer gemeinsamen Skala abzubilden. Dabei muss aber geprüft werden, ob dies auch inhaltlich sinnvoll ist. Die folgende Tabelle fasst das *LSE 9*-Testdesign für das Jahr zusammen (Itemblock X = Ankeritems):

Tabelle 6.1: LSE 9-Testdesign (Testjahr 2004)

Itemblock	Itemzahl	Testheft A (HS, GE-GK)	Testheft B (GE-EK, RS, GY)
A	19	✓	--
X	20	✓	✓
B	19	--	✓

Die Entwicklung der *LSE 9* fand mit ähnlichem Aufwand statt wie die Entwicklung internationaler Schulleistungsstudien. Die Items wurden von Lehrkräften in einer Kommission

entwickelt, die wissenschaftlich (fachdidaktisch und psychometrisch) begleitet wurde, und in einer Präpilotierung und einer Pilotierung erprobt. Die Präpilotierung war eher qualitativ angelegt und fokussierte (a) auf die Verständlichkeit der Aufgabenstellung, (b) auf die auf Auswertbarkeit der Bearbeitungen und (c) ggf. auf die Generierung geeigneter Distraktoren.[115] In der Pilotierung wurden die verbliebenen, optimierten Aufgaben dann quantitativ im Rahmen von *IRT*-Skalierungen überprüft. Insgesamt wurden so zwei Testversionen entwickelt, die allen gängigen psychometrischen Anforderungen genügen (vgl. Fleischer et al., 2007).

Die *LSE 9* werden jeweils von den Fachlehrkräften ausgewertet, wobei die Pilotierung gezeigt hat, dass die Auswertungsobjektivität auf dem Niveau internationaler Schulleistungsstudien liegt. Die Daten werden über eine Web-Schnittstelle anonymisiert auf einen Server des Schulministeriums übertragen, sodass auf dieser Grundlage Rückmeldungen an die Schulen generiert werden können. Damit liegt für alle Schülerinnen und Schüler, die an diesem Test teilgenommen haben, nach Klassen und Schulen gebündelt, ein vollständiger Datensatz elektronisch vor. In der vorliegenden Arbeit werden die Datensätze der Schülerinnen und Schüler, die an der Zusatzerhebung „Raumvorstellung“ teilgenommen haben, mit den entsprechenden Datensätzen dieser Zusatzerhebung zusammengeführt.

Mit den *LSE 9* kann also ein Mathematiktest genutzt werden, der inhaltlich breit, nah an der Unterrichtspraxis und dabei im Sinne der „neuen Aufgabenkultur“ von SINUS, *PISA* etc. ist. Die starke empirische Absicherung des Tests gewährleistet eine zuverlässige und valide Messung der *Mathematikleistung*. Da die *LSE 9* im Jahr 2004 einen Schwerpunkt im Bereich „Modellieren“ hatte, sind in diesem Test *rechnerische* und *begriffliche Modellierungsaufgaben* neben *technischen Aufgaben* breit vertreten. Bei der Auswertung der *LSE 9* können also differenzierte Analysen mit entsprechenden Aufgabenklassifikationen durchgeführt werden.

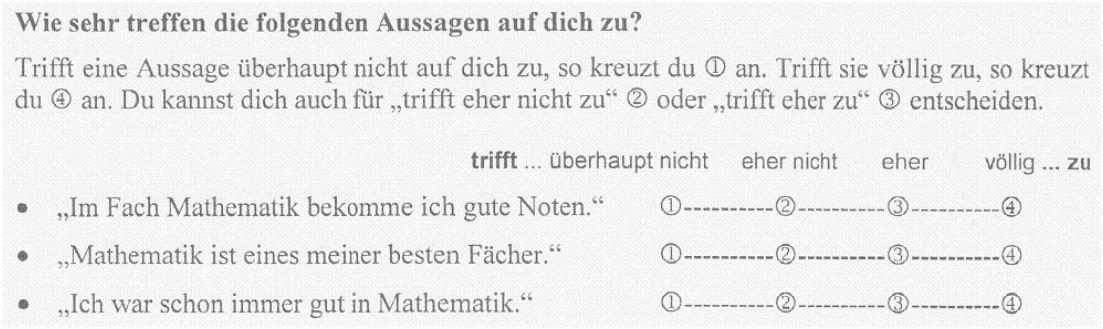
6.2.4 Weitere Instrumente

Auf dem Deckblatt des Testheftes wird – wie in der Voruntersuchung – nach einer Begrüßung das Anliegen der Untersuchung kurz dargestellt. Dabei wird darauf hingewiesen, dass die Daten im Zusammenhang mit den *LSE 9*-Ergebnissen ausgewertet werden und dass die Daten bei der Erfassung anonymisiert werden. Nach einigen allgemeinen Testinstruktionen und ersten Rahmendaten werden auf dem Deckblatt noch das *FSK:M* und die letzten Noten in den Fächern Deutsch, Mathematik und Sport erfasst.

[115] Im Gegensatz zu den großen Schulleistungsstudien ist der Anteil von Multiple-Choice-Items im LSE 9-Mathematiktest aber gering. Testtheoretisch haben die für den Mathematikunterricht typischen Kurzantwortformate („Ergebnis hinschreiben“) keine geringere Qualität und sogar gewisse Vorteile: Sie lassen die Vielzahl möglicher „Falschlösungen“ für die Lehrkraft sichtbar werden, da keine Beschränkung auf Distraktoren vorgegeben ist, und die Ratewahrscheinlichkeit beträgt dabei (im statistischen Sinne) Null.

Da sich die Items zum *FSK:M*, die in der Voruntersuchung eingesetzt wurden, nicht bewährt haben, mussten hier neue Items gefunden werden, die nach Möglichkeit ohne weitere Erprobung direkt eingesetzt werden können. Daher wurde auf die entsprechenden Items von *PISA 2000* zurückgegriffen, die dort eine Skala mit sehr guten empirischen Kennwerten gebildet haben (vgl. Kunter et al., 2002, S. 170).

Abbildung 6.1: FSK:M-Items aus PISA 2000 (Quelle: Kunter et al., 2002, S. 170)



6.3 Durchführung und Auswertung der Hauptuntersuchung

Analog zur Dokumentation der Voruntersuchung wird die Hauptuntersuchung mit der Beschreibung der Stichprobe, der Zusammenstellung des eigenen Testhefts, Anmerkungen zur Durchführung des eigenen Tests, Hinweise zur Erfassung der Daten und einer Skizze der Auswertung der Daten dargestellt.

6.3.1 Beschreibung der Stichprobe

Die Hauptuntersuchung wurde an vier Schulen durchgeführt, die typische Repräsentantinnen ihrer jeweiligen Schulform sind. Während die Hauptschule und die Realschule in ländlich geprägten Kleinstädten stehen, befinden sich das Gymnasium und die Gesamtschule in mittelgroßen Städten mit ländlichem Umfeld bzw. im Ruhrgebiet. Mathematik wird in der Jahrgangsstufe 9 am Gymnasium vierzügig, an der Haupt- und an der Realschule fünfzügig und an der Gesamtschule sechszügig unterrichtet. Von allen Schulen konnte jeweils der gesamte Jahrgang in die Hauptuntersuchung einbezogen werden.

Die Verteilung von Jungen und Mädchen ist innerhalb jeder Schule ungefähr typisch für die jeweilige Schulform, lediglich in der Realschule sind Mädchen aus Sicht der Schulform leicht überrepräsentiert. Zu allen Versuchspersonen liegen Angaben zum Geschlecht vor, da etwa fehlende Angaben aus dem Datensatz der *LSE 9* übernommen werden konnten. Tabelle 6.2 gibt für die Stichprobe der Hauptuntersuchung die Verteilung der Schülerinnen und Schüler auf die vier Schulen wieder.

Tabelle 6.2: Stichprobe der Hauptuntersuchung

Schule	männlich	weiblich	Summe
Hauptschule	55	48	103 (20,8 %)
Gesamtschule	67	70	137 (27,6 %)
Realschule	61	79	140 (28,2 %)
Gymnasium	44	72	116 (23,4 %)
Summe	**227** (45,8 %)	**269** (54,2 %)	**496**

Die Verteilung auf die Schulformen weicht zwar etwas von der Verteilung im Landesdurchschnitt ab (HS: 22,7 %; GE: 16,8 %; RS: 27,8 %, GY: 32,7 %; vgl. Köller et al., 2010, S. 3), diese Abweichung ist für die Ziele der Hauptuntersuchung aber unschädlich: Wie in Kap. 4.3.1 dargestellt wurde, kommt es vor allem darauf an, dass die Leistungsverteilung und insbesondere die Varianz in der Stichprobe möglichst repräsentativ ist, was der Fall sein dürfte und bei den Grundauswertungen der einzelnen Tests in Kap. 6.4 sichtbar werden müsste.

Bei der Beschreibung der Stichprobe muss noch berücksichtigt werden, dass zwar der eigene Test (Zusatzerhebung „Raumvorstellung“) von allen Schülerinnen und Schülern der Stichprobe bearbeitet wird, es bei den *LSE 9* aber zwei Testversionen gibt. Tabelle 6.3 stellt die Verteilung der Schülerinnen und Schüler auf die beiden Testheftversionen dar.

Tabelle 6.3: Stichprobe der Hauptuntersuchung nach LSE 9-Testheftversionen

Schule	männlich	weiblich	Summe
Heft A (HS-GK, HS-EK, GE-GK)	82	82	164 (33,1 %)
Heft B (GE-EK, RS, GY)	145	187	332 (66,9 %)
Summe	**227** (45,8 %)	**269** (54,2 %)	**496**

Die Verteilung der Stichprobe auf die beiden Testversionen entspricht damit im Wesentlichen dem Landesdurchschnitt, wobei auch Abweichungen unproblematisch wären, da die Leistungen auf einer gemeinsamen Skala abgebildet werden können.

6.3.2 Zusammenstellung des Testheftes

In Kap. 6.2 wurde dargestellt, wie die Instrumente für die Hauptuntersuchung aus den Instrumenten der Voruntersuchung entwickelt worden sind. Neben den leichten Modifikationen beim Untertest *WLT* wurden Kurzformen der Untertests *DAT:SR* und *Denkstile* entwickelt. Beim Deckblatt des Testhefts und beim (unveränderten) Untertest *MRT* werden in der Hauptuntersuchung die Bearbeitungszeiten aufgrund der Erfahrungen aus der Voruntersuchung angepasst. Für den Untertest *MRT* werden jetzt 15 Minuten (Voruntersuchung: 20 Minuten) vorgesehen. Damit liegt die Bearbeitungszeit für den *MRT* immer noch erheblich über der Standardbearbeitungszeit von 3 Minuten (jeweils für 10 Items). Der Untertest *MRT* soll in der eigenen Untersuchung bewusst als „Power-Test" und nicht als „Speed-Test" durchgeführt werden,[116] da die vorliegende Arbeit an der prinzipiellen Fähigkeit zur *mentalen Rotation* und nicht an der Konzentrationsfähigkeit und der Geschwindigkeit der ablaufenden mentalen Prozesse interessiert ist. Diese Überlegung ist auch vor dem Hintergrund der Analyse von Geschlechterunterschieden gerechtfertigt, da diese bei „Power-Varianten" des *MRT* nicht verschwinden (vgl. Titze et al., 2008, S. 130).

Die Tabelle 6.4 zeigt, dass das so zusammengestellte Testheft genau in einer Unterrichtsstunde bearbeitet werden kann. Dies war aus pragmatischen Gründen, nicht zuletzt auch wegen der Belastung des Mathematikunterrichts in der 9. Jahrgangsstufe durch die *LSE 9* und diese Zusatzerhebung, geboten.

Tabelle 6.4: Zusammenstellung des Testheftes für die Hauptuntersuchung

Untertest	Zeitbedarf
UT 1: Deckblatt	7 min
UT 2: *WLT*	3 min
UT 3: *MRT*	15 min
UT 4: *DAT:SR*	10 min
UT 5: *Denkstile*	10 min
Insgesamt	45 min

[116] Bei einem „Power-Test" soll das prinzipielle Leistungsvermögen der Versuchspersonen unabhängig von Performanzkomponenten wie Konzentrationsfähigkeit oder Verarbeitungsgeschwindigkeit erfasst werden. Bei einem „Speed-Test" werden hingegen bewusst genau solche Performanzkomponenten in den Blick genommen, z. B. durch Zeitrestriktionen.

6.3.3 Durchführung der Erhebung

Die Datenerhebung für die Hauptuntersuchung fand binnen drei Wochen nach Durchführung der *LSE 9*-Mathematik statt. Da der Autor der vorliegenden Arbeit dabei von einer studentischen Hilfskraft unterstützt wurde, konnte der 9. Jahrgang an jeder der vier teilnehmenden Schulen jeweils an einem Vormittag getestet werden. Dabei konnten durch Festlegung der Testinstruktionen und ggf. erlaubtes Antwortverhalten auf Klärungsfragen standardisierte Durchführungsbedingungen gewährleistet werden.

Die einzelnen Untertests waren wie in der Voruntersuchung durch „Stopp-Blätter" getrennt. So wurde erreicht, dass alle Schülerinnen und Schüler sich parallel mit den Testinstruktionen und ggf. Beispielitems auseinandersetzen konnten und dass bei Bedarf Rückfragen im Plenum geklärt werden konnten, *bevor* die eigentliche Testbearbeitung begann. Alle Beobachtungen während der Testdurchführung deuten darauf hin, dass die Bedingungen regulär waren und die Daten aus dieser Sicht uneingeschränkt nutzbar sind.

Während der Durchführung des Untertest *Denkstile* fiel aber wiederum auf, dass eine relevante Anzahl von Schülerinnen und Schülern ihre Überlegungen bei der Bearbeitung der Items nicht nachvollziehbar darstellen konnten. So gab es viele Rückfragen von Schülerinnen und Schülern, die zwar an den freien Stellen Figuren ergänzt haben, aber nicht wussten, was sie beschreiben sollen bzw. wie sie dies erledigen sollen. Da die Untersuchung als Gruppentest durchgeführt wurde, bestand auch keine Möglichkeit in Einzelgesprächen den jeweils bevorzugten *Denkstil* über die Methode des lauten Denkens zu diagnostizieren.

6.3.4 Erfassung und Aufbereitung der Daten

Für die angestrebten Zusammenhangsanalysen war die passende Zusammenführung der Daten aus der Zusatzerhebung und den *LSE 9* erforderlich. Da in den Schulen klassen- bzw. kursweise geführte Listen mit den Namen der Schülerinnen und Schüler und ihren *LSE 9*-Kennnummern vorlagen, sollten die Schülerinnen und Schüler auf dem Deckblatt des Testhefts ihre Namen und ihre Klasse bzw. ihren Kurs notieren. Über die *LSE 9*-Listen konnten die Datensätze dann mit den vorliegenden *LSE 9*-Datensätzen zusammengeführt und anschließend anonymisiert werden. Der gemeinsame Datensatz enthält dabei einige Schülerinnen und Schüler, die nur an den *LSE 9* oder nur an der Zusatzerhebung teilgenommen haben. Von den 496 Schülerinnen und Schülern der Stichprobe haben zwei nur an der Zusatzerhebung, 30 nur an den *LSE 9* und 464 an beiden Tests teilgenommen.

Bei der Erfassung der Daten sind zusätzlich zu den Betrachtungen, die bereits bei der Voruntersuchung angestellt wurden (vgl. Kap. 5.3.4), die Eingabequalität der Daten und die Frage der Auswertbarkeit des Untertests *Denkstile* besonders relevant. Die Dateneingabe wurde vom Autor der vorliegenden Arbeit selbst durchgeführt. Die Qualität dieser Dateneingabe wurde anschließend überprüft, indem eine studentische Hilfskraft die Daten mit entsprechenden Instruktionen für eine Teilstichprobe separat eingegeben und hinterher die

Überstimmung quantifiziert hat. Da der Untertest Denkstile erneut nicht in hinreichendem Umfang ausgewertet werden konnte (s. u.), können Unterschiede in der Dateneingabe nur aus Tippfehlern oder aus einer anderen Signierung beim Untertest *WLT* resultieren. Zwar wurden die *WLT*-Items wiederum mit einer Schablone ausgewertet, jedoch gab es einige Bearbeitungen, die eine „Grauzone“ bei der Signierung darstellten.

Pro Versuchsperson mussten 36 Daten aus den betrachteten Untertests und vom Deckblatt eingegeben werden. Für 100 zufällig ausgewählte Testhefte wurde die Zweitkodierung durchgeführt, sodass insgesamt 3600 Dateneingaben überprüft wurden. Dabei stellte sich heraus, dass ein Tippfehler und zwei abweichende Signierungen beim Untertest *WLT* (von insgesamt 600 überprüften Signierungen) vorlagen. Der Anteil unterschiedlich eingegebener Daten liegt in der ausgewählten Teilstichprobe also unterhalb von 0,1 %, was für eine sehr gute Auswertungsobjektivität spricht.

Der Untertest *Denkstile* wurde zunächst nur vom Autor der vorliegenden Arbeit signiert und kodiert, da eine Zweitauswertung durch eine studentische Hilfskraft einen erheblichen Schulungsaufwand vorausgesetzt hätte. Dies wäre im Falle eines hohen Anteils relativ eindeutig auswertbarer Bearbeitungen infrage gekommen. Zwar konnte der Anteil von Item-Bearbeitungen, die nicht oder nur mit großer Unsicherheit ausgewertet werden konnten, auf etwa ein Drittel gesenkt werden, dies reicht für eine angemessene statistische Weiterverarbeitung der Daten aber nicht aus. Dieses defensive Vorgehen ist insofern geboten, als unter den etwa zwei Dritteln auswertbarer Bearbeitungen ein sehr hoher Anteil von *prädikativen* Bearbeitungen war (ca. 75 %). Die *kann* ein Indiz dafür sein, dass unter den nicht oder nur unsicher auswertbaren Bearbeitungen sehr viele *funktionale* Bearbeitungen sind.

Die Konsequenz hieraus ist, dass das erweiterte Modell zur Erklärung von Geschlechterunterschieden in der *Mathematikleistung* (vgl. Abb. 4.2, S. 122) nicht in der vorgeschlagenen Form empirisch überprüft werden kann. Ohne das Konstrukt *Denkstile* stellt das verbleibende Modell die um das Konstrukt *FSK:M* ergänzte *Spatial Mediation Hypothesis* (vgl. Abb. 3.9, S. 107) dar. Das in Kap. 6.1.5 (S. 195) formulierte Explorationsanliegen E_01 ist somit bereits (mit negativem Ergebnis) umgesetzt worden.

6.3.5 Auswertung der Daten

Die Daten der eigenen Zusatzerhebung zu den *LSE 9* werden in der Hauptuntersuchung zunächst wieder einer Grundauswertung wie bei der Voruntersuchung unterzogen, die sich sowohl Methoden der *KTT* als auch Modellen und Verfahren der *IRT* bedient. In einem ersten Schritt werden die einzelnen Untertests separat betrachtet und darauf hin bewertet, ob sie aus empirischer Sicht eine belastbare Basis für vertiefende Analysen darstellen. Entsprechend werden auch die Daten der *LSE 9* für die Schülerinnen und Schüler der Hauptuntersuchung ausgewertet. Dabei muss inhaltlich und empirisch geprüft werden, ob beide Testheftversionen gemeinsam oder separat skaliert werden sollen.

Nach einer derartigen Grundauswertung der einzelnen Bestandteile der Hauptuntersuchung werden dann vertiefende Analysen durchgeführt, die direkt Bezug auf die in Kap. 6.1 formulierten Hypothesen nehmen. Daraus ergibt sich, dass wiederum Varianzanalysen (ggf. zurückgeführt auf geeignete Regressionsanalysen), mehrdimensionale *RM*s und ggf. *LCAs* zur Anwendung kommen. Für Zusammenhangsanalysen im Sinne der erweiterten *Spatial Mediation Hypothesis* kommen Kovarianzanalysen und, allgemeiner, Strukturgleichungsmodelle (mit dem Spezialfall der Pfadanalyse) in Betracht. Für die Exploration von Gruppenunterschieden bei *LSE 9*-Items können *DIF*-Analysen verwendet werden. Im Sinne der inhaltlichen Fragestellungen soll dabei die technische Berichterstattung auf das absolut notwendige Maß beschränkt bleiben.

6.4 Befunde der Hauptuntersuchung

Die Befunde der Hauptuntersuchung werden zunächst nach Testbereichen (*Raumvorstellung*, *Mathematikleistung*, *FSK:M*) getrennt und dann mit Blick auf den Zusammenhang zwischen den Konstrukten berichtet.

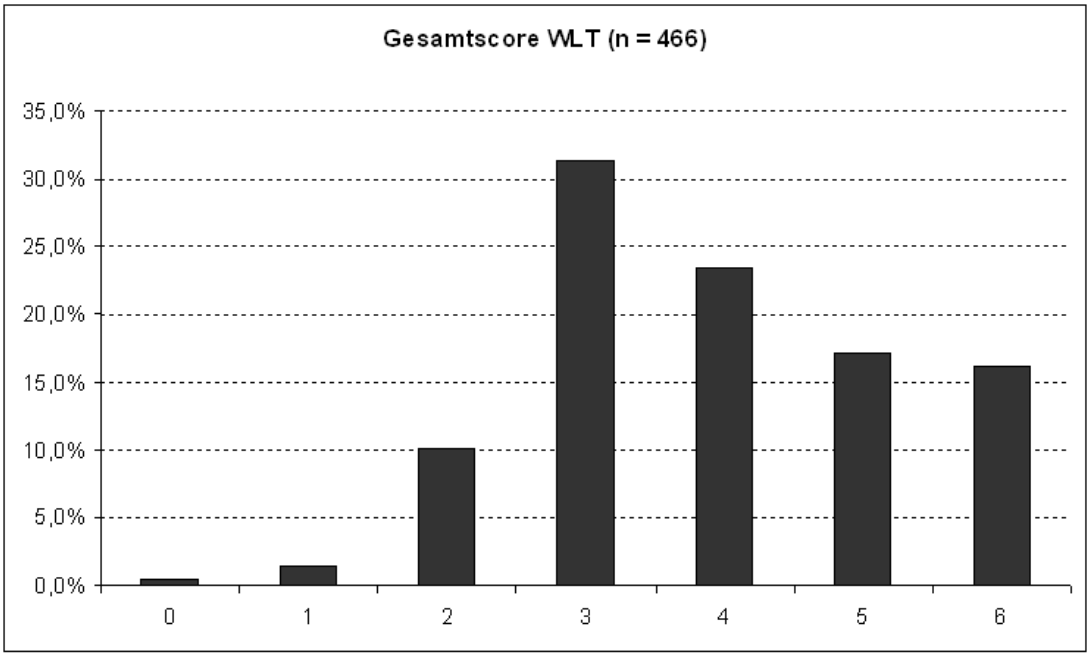
6.4.1 Raumvorstellung

Bei der Darstellung der Befunde zur *Raumvorstellung* werden, wie bei der Voruntersuchung, zunächst die Grundauswertungen der Untertests *WLT*, *MRT* und *DAT:SR* dokumentiert. Dabei wird die Verteilung der Stichprobe nach Gesamtscore wiederum graphisch (Säulendiagramm) und numerisch (Median, arithmetisches Mittel, Schiefe und Exzess) dargestellt. Anschließend werden die Ergebnisse der Rasch-Skalierung durch die Verteilungen der Items und der Versuchspersonen auf der gemeinsamen latenten Dimension betrachtet. An diese Grundauswertung schließen sich Analysen zu Geschlechterunterschieden und zum Zusammenhang der drei Komponenten der *Raumvorstellung* an, die jeweils direkt auf die zugrundeliegenden Hypothesen (vgl. Kap. 6.1) bezogen sind.

WLT

Der *WLT*, der als Referenztest für *räumliche Wahrnehmung* eingesetzt wurde, enthielt gegenüber der Voruntersuchung – neben einer leicht modifizierten Instruktion – neu gezeichnete Items und insgesamt ein Item mehr. Für die Stichprobe der Voruntersuchung ergab sich eine bimodale Verteilung mit starkem Deckeneffekt (vgl. Abb. 5.10, S. 165). Diese Verteilungseigenschaften konnten durch die Zusammensetzung der Stichprobe, zwei Gymnasien und eine Gesamtschule, erklärt werden. Der Deckeneffekt trat nur an den Gymnasien deutlich ausgeprägt mit dem Maximalscore als Modalwert auf. Die Stichprobe der Hauptuntersuchung ist anders zusammengesetzt und bildet die typische Verteilung von allgemeinen kognitiven Fähigkeiten und Fachleistungen innerhalb eines Altersjahrgangs besser ab. Das Säulendiagramm in Abb. 6.2 zeigt, dass die Verteilung in der Hauptuntersuchung eingipflig („unimodal“) ist, aber immer noch einen leichten Deckeneffekt aufweist.

Abbildung 6.2: Verteilung der Versuchspersonen nach WLT-Gesamtscore



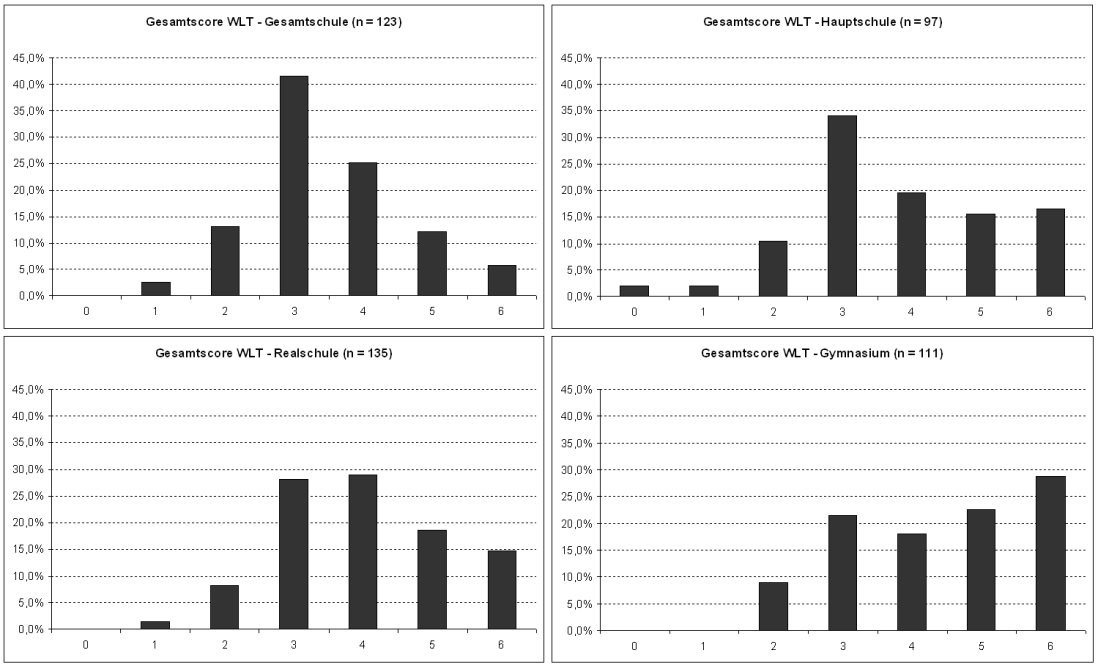
In der Hauptuntersuchung erreichen 16,1 % der Versuchspersonen den Maximalscore, während dies in der Voruntersuchung noch 29,9 % waren. Dabei dürfte das zusätzliche Item keine Rolle spielen, da es sich hierbei um das leichteste Item, eine aufrecht stehende Flasche, handelt. Dass die Verteilung der Stichprobe der Hauptuntersuchung „normaler“ als in der Voruntersuchung ist, bestätigen auch die Verteilungskennwerte in Tab. 6.5. Die Verteilung ist etwas breiter als die Standardnormalverteilung; bei einem Standardfehler von 0,23 unterscheidet sich der Exzess mit einem Wert von –0,67 signifikant von Null. Die Schiefe ist aber praktisch nicht von Null unterscheidbar (Standardfehler: 0,11).

Tabelle 6.5: Kennwerte der Verteilung nach WLT-Gesamtscores (n = 466)

Test	Itemzahl	Median	arithm. Mittel	Schiefe	Exzess
WLT	6	4	3,92	0,04	–0,67

Wie in der Voruntersuchung wird an dieser Stelle exemplarisch für den *WLT* gezeigt, wie sich die Gesamtverteilung der Stichprobe aus den Teilstichproben der vier Schulformen zusammensetzt. Abbildung 6.3 zeigt die entsprechenden Diagramme, deren Achsen jeweils gleich eingeteilt sind; die Diagramme sind von oben links nach unten rechts nach dem arithmetischen Mittel der Testergebnisse an der jeweiligen Schule angeordnet.

Abbildung 6.3: Verteilung der Versuchspersonen nach WLT-Gesamtscore und nach Schulformen



Offensichtlich liegt im Gymnasium auch in der Hauptuntersuchung ein stärkerer Deckeneffekt vor. Einen leichten Deckeneffekt gibt es auch in der Realschule und in der Hauptschule, allerdings kann der Test in diesen beiden Schulformen insgesamt gut differenzieren. Bemerkenswert ist, dass an der Hauptschule 16,5 % der Schülerinnen und Schüler den Maximalscore erzielen; dieser Anteil ist deutlich höher als an der Gesamtschule (5,7 %) und etwas höher als an der Realschule (14,8 %). Möglicherweise lässt sich das relativ gute Ergebnis an der Hauptschule durch den intensiven Technik- und Arbeitslehreunterricht sowie einen hohen berufspraktischen Anteil erklären. Es ist aber auch nicht auszuschließen, dass das Ergebnis auf die konkrete Zusammensetzung der Schülerschaft an *dieser* Hauptschule zurückzuführen ist. Da die Stichprobe nicht mit Blick auf repräsentative Leistungsdaten für Schulformen zusammengestellt wurde, können solche Ergebnisse nur mit größter Vorsicht und unter Betonung des hypothetischen Charakters gedeutet werden. Etwaige Auswirkungen von Technik- und Arbeitslehreunterricht sowie von berufspraktischen Anteilen müssten in einer spezifisch konzipierten Studie untersucht werden.

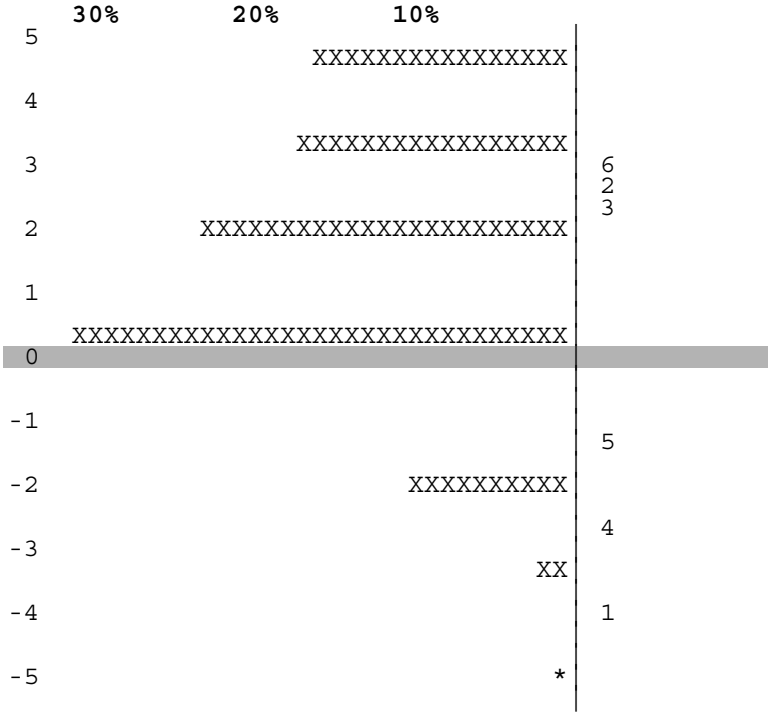
Die Kennwerte in Tab. 6.6 unterstreichen den optischen Eindruck von Abb. 6.3. Insbesondere die Unterschiede in den beiden berichteten Mittelwerten spiegeln die Unterschiede zwischen den Teilstichproben wider.

Tabelle 6.6: Kennwerte der Verteilungen nach WLT-Gesamtscores und nach Schulformen

	Itemzahl	Median	arithm. Mittel	Schiefe	Exzess
GE (n = 123)	6	3	3,49	0,37	–0,02
HS (n = 97)	6	4	3,79	–0,10	–0,32
RS (n = 135)	6	4	3,99	0,04	–0,66
GY (n = 111)	6	5	4,41	–0,28	–1,21

Eine Skalierung des *WLT* nach dem eindimensionalen zweikategoriellen *RM* ergibt gute Kennwerte für die Anpassung des Modells an die beobachteten Daten; dies gilt insbesondere für die verwendeten Items, deren empirische *ICCs* sehr gut durch die theoretischen *ICCs* modelliert werden. In Abb. 6.4 werden die mit *WINMIRA* geschätzten Itemschwierigkeiten und Personenfähigkeiten auf der gemeinsamen Logit-Skala dargestellt. Ein X repräsentiert ca. 1 % der Stichprobe. Beim Vergleich mit der analogen Darstellung bei der Auswertung der Voruntersuchung (Abb. 5.12, S. 168) muss berücksichtigt werden, dass die Itemnummern im Rahmen der Neuzeichnung des Tests verändert wurden.

Abbildung 6.4: Verteilung der WLT-Testleistung und der Itemschwierigkeiten (n = 466)



Arithm. Mittel (Testleistung): 1,64
Zugehörige Standardabweichung: 2,10

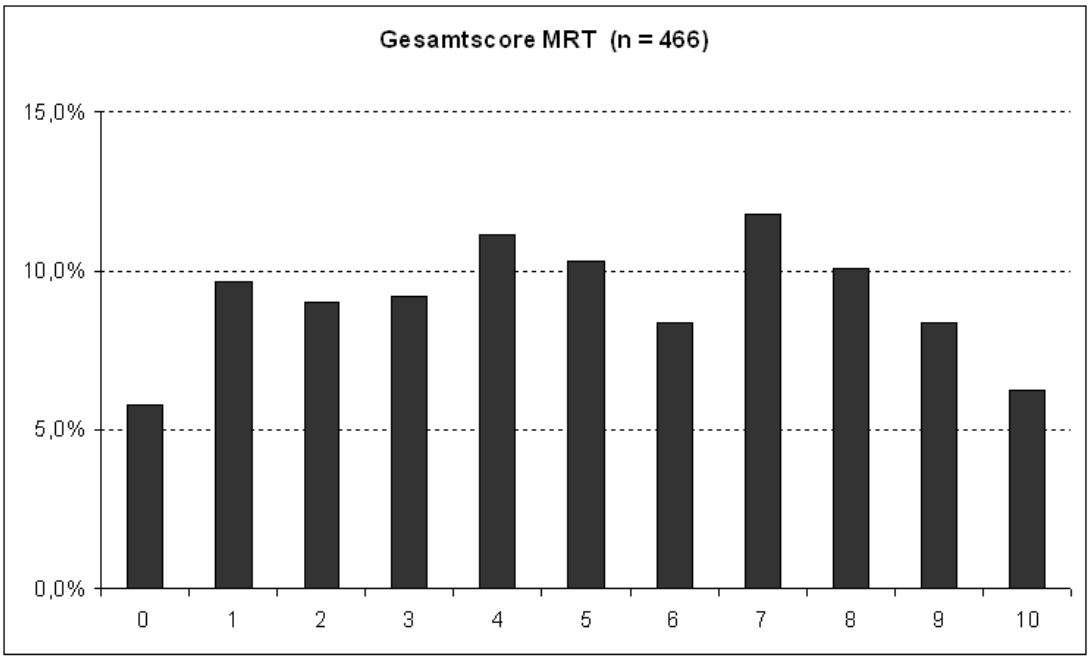
* 2 Versuchspersonen (ca. 0,4 %) erzielten eine Testleistung von ca. -5,11

Wie in der Voruntersuchung bilden sich hier zwei Itemcluster mit empirisch relativ einfachen und relativ schwierigen Items. Bei den Items 2, 3 und 6 handelt es sich um schräg stehende bzw. gekippte Flaschen, während die anderen drei Items Flaschen in einfachen Positionen zeigen (1: „aufrecht stehend“; 4: „auf dem Kopf“; 5: „auf der Seite“). Die Verteilung der Testleistung ermöglicht insgesamt die angestrebten Zusammenhangsanalysen.

MRT

Die Grundauswertung des *MRT*, der in der Hauptuntersuchung unverändert gegenüber der Voruntersuchung als Referenztest für *mentale Rotation* eingesetzt wurde, deutet ebenfalls darauf hin, dass die Stichprobe hinreichend gut für die Ziele der Untersuchung zusammengesetzt ist. Die Verteilung nach *MRT*-Gesamtscores (Abb. 6.5) ist relativ breit und weitgehend symmetrisch. Zwar sind ein moderater Deckeneffekt und ein moderater Bodeneffekt vorhanden, sie schränken das Differenzierungspotenzial des *MRT* in der betrachteten Stichprobe aber kaum ein. Den Maximalscore erzielen 6,2 % der Versuchspersonen, während 5,8 % gar kein Item richtig bearbeiten – und das, obwohl die Ratewahrscheinlichkeit $^1/_6$ beträgt.

Abbildung 6.5: Verteilung der Versuchspersonen nach MRT-Gesamtscore



Die näherungsweise Symmetrie der Verteilung schlägt sich auch in den Kennwerten nieder (Tab. 6.7): Der Median ist genau der mittlere mögliche Score, das arithmetische Mittel weicht praktisch nicht vom Median ab und die Schiefe ist praktisch gleich Null.

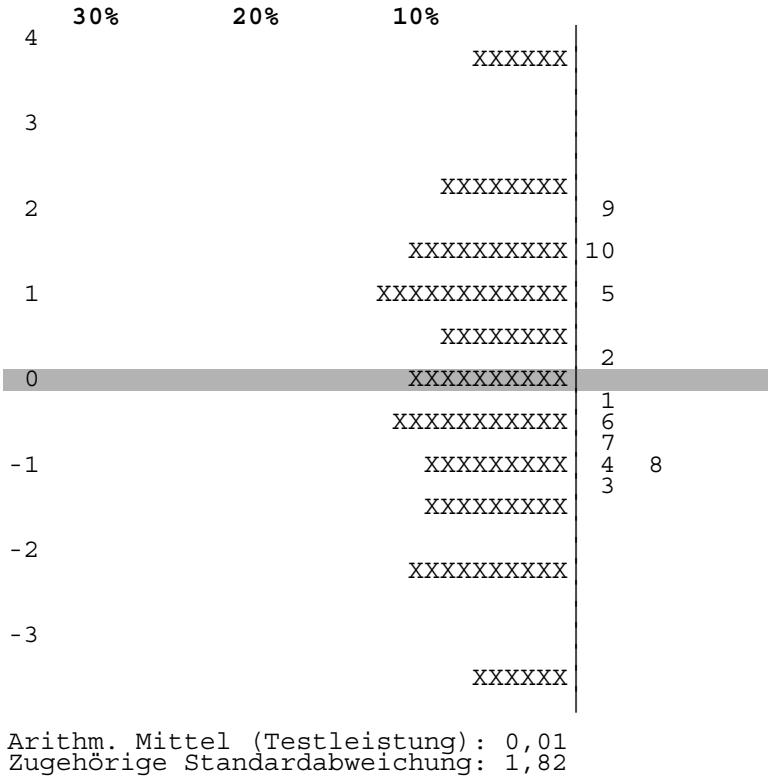
Tabelle 6.7: Kennwerte der Verteilung nach MRT-Gesamtscores (n = 466)

Test	Itemzahl	Median	arithm. Mittel	Schiefe	Exzess
MRT	10	5	5,03	–0,02	–1,12

Untersucht man die Verteilung wieder nach Schulformen getrennt, so zeigt sich im Gymnasium ein Deckeneffekt (17,2 % erreichen den Maximalscore). Dadurch dürfte die Varianz der Testleistung am Gymnasium leicht eingeschränkt sein. Dieser Befund gilt konstruktbedingt für alle drei Raumvorstellungstests. Insbesondere die *räumliche Wahrnehmung* und die *mentale Rotation* werden durch den *WLT* bzw. *MRT* mit Items erfasst, für die es kaum schwierigkeitsbestimmende Faktoren gibt. Eine Erhöhung der empirischen Schwierigkeit ließe sich daher nur konstruktfremd erzielen. Die Deckeneffekte sind aber auf das Gymnasium beschränkt und auch hier nicht so gravierend, dass das Untersuchungsziel gefährdet ist: Immerhin haben 82,8 % der Schülerinnen und Schüler nicht alle Items richtig gelöst und das arithmetische Mittel von 6,96 am Gymnasium zeigt, dass der Test auch an dieser Schule ein (leicht eingeschränktes) Differenzierungspotenzial hat.

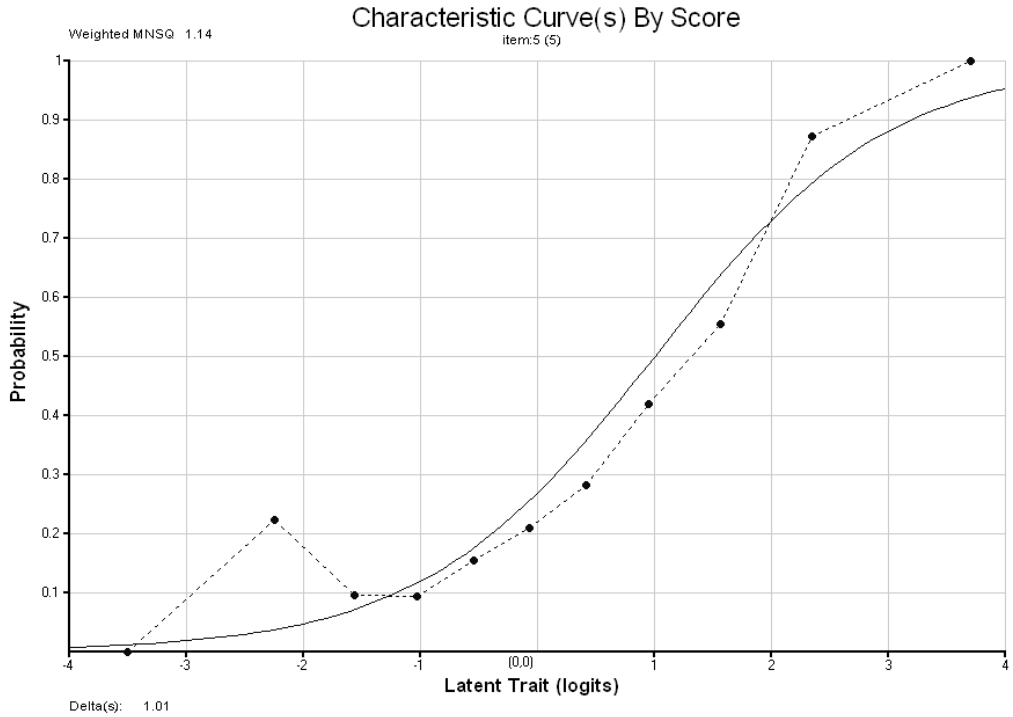
Die zuvor angesprochene relative Homogenität der Itemschwierigkeit zeigt sich beim *MRT* in der gemeinsamen Verteilung von Testleistung und Itemschwierigkeiten auf der Logit-Skala, die im Rahmen der Rasch-Skalierung gewonnen wurde.

Abbildung 6.6: Verteilung der MRT-Testleistung und der Itemschwierigkeiten (n = 466)



Die empirischen Werte weisen insgesamt auf eine gute Anpassung des Modells an die beobachteten Daten hin. Während Item 3, das in der Voruntersuchung auffällige Kennwerte hatte, in der Hauptuntersuchung unauffällig ist, weist der *Q-Index* darauf hin, dass Item 5 in der Hauptuntersuchung auffällig ist (0,019 < 0,050). Die *ConQuest*-Analyse weist einen *Weighted MNSQ* von 1,14 aus, der sich zwar signifikant von Eins unterscheidet, im Rahmen von internationalen Schulleistungsstudien aber als akzeptabel eingestuft werden würde.[117] Interessant bleibt aber die Frage, warum Item 5 empirisch auffällig ist. Da Item 5 bei einer theoretischen Anforderungsanalyse unauffällig ist, kann höchstens der statistische Zugang die relativ schlechten Kennwerte erklären. Hier bietet sich nach der Schätzung der Modellparameter wieder der Vergleich von theoretischer und empirischer *ICC* des Items an, der von *ConQuest* standardmäßig bereitgestellt wird (Abb. 6.7).

Abbildung 6.7: Theoretische und empirische ICC für das MRT-Item 5



Die elf eingezeichneten Punkte auf der empirischen *ICC* stellen die relativen Lösungshäufigkeiten für die (*WLE*-Schätzungen der) elf möglichen Gesamtscores dar. Bei Gesamtscore 1 mit der zugehörigen *WLE*-Schätzung von –2,241 (*ConQuest*) beträgt die relative Lösungshäufigkeit 22,2 %. Von den 45 Versuchspersonen, die nur ein Item richtig bearbeitet

[117] Bei *PISA 2006* etwa haben sich die beteiligten Wissenschaftlerinnen und Wissenschaftler darauf verständigt, dass der *Weighted MNSQ* für Items im Intervall von [0,8; 1,2] liegen soll (vgl. OECD, 2009, S.355).

haben, konnten zehn das fragliche Item 5 lösen. Hingegen konnten nur zwei dieser 45 Versuchspersonen (4,4 %) das empirisch leichtere Item 6 richtig lösen. Diese Betrachtung liefert zwar eine statistische Rekonstruktion des abweichenden Verhaltens von Item 5, eine inhaltliche Erklärung liegt damit aber weiterhin nicht vor. Da die empirische *ICC* für alle anderen Gesamtscores sehr gut zur theoretischen passt, muss als potenzielle Erklärung auch in Betracht gezogen werden, dass es sich um eine zufällige, wenn auch seltene, Abweichung handelt. Insgesamt kann der *MRT*, inklusive Item 5, für vertiefte Analysen verwendet werden.

DAT:SR

Der *DAT:SR* wurde in der Voruntersuchung mit insgesamt 25 Items als Referenztest für *räumliche Visualisierung* eingesetzt. Im Rahmen der Rasch-Skalierung wurde dort eine zehn Items umfassende Kurzform des Tests entwickelt, die in der Hauptuntersuchung eingesetzt wurde. Dabei wurden von den 22 Items, die unauffällige Kennwerte aufwiesen, die zehn schwierigsten ausgewählt, weil der *DAT:SR* einen Deckeneffekt aufwies. Abbildung 6.8 und Tabelle 6.8 zeigen, dass die Verteilung der Stichprobe nach *DAT:SR*-Gesamtscore in der Hauptuntersuchung relativ breit und nahezu symmetrisch ist sowie einen moderaten Deckeneffekt aufweist. Im Vergleich zum ebenfalls zehn Items umfassenden *MRT* haben mit 1,7 % deutlich weniger Versuchspersonen gar kein Item richtig bearbeitet, was auch an der höheren Ratewahrscheinlichkeit von $^1/_4$ liegen dürfte.

Abbildung 6.8: Verteilung der Versuchspersonen nach DAT:SR-Gesamtscore

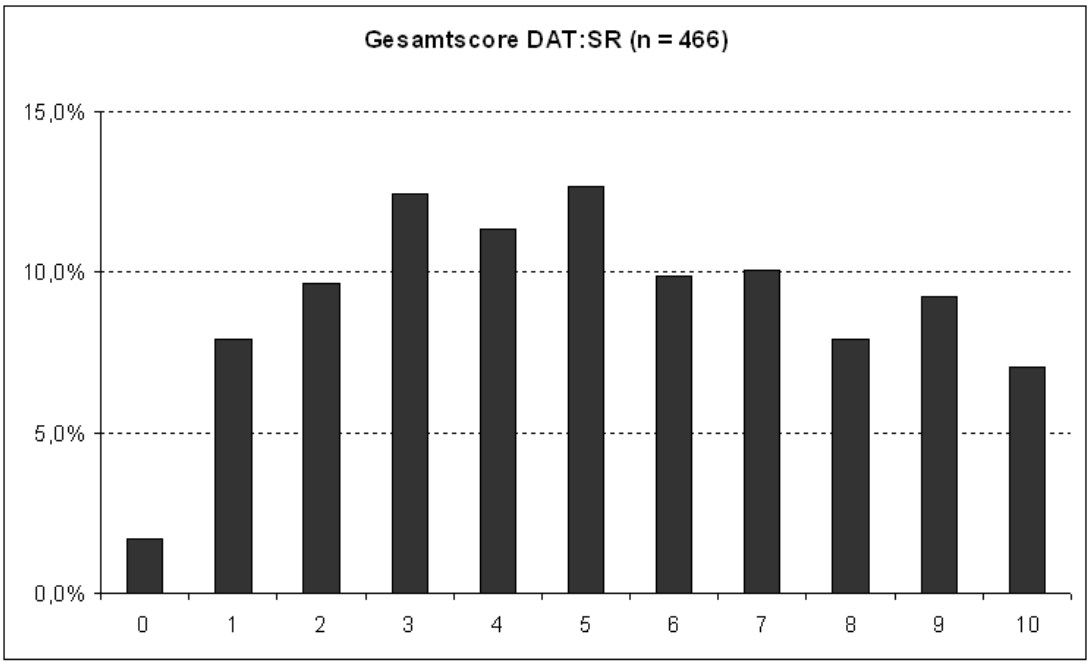
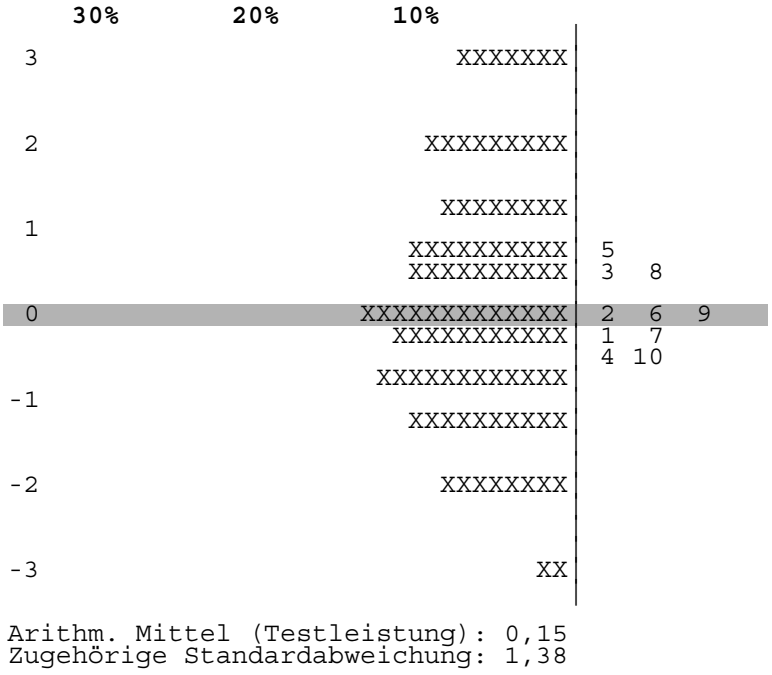


Tabelle 6.8: Kennwerte der Verteilung nach DAT:SR-Gesamtscores (n = 466)

Test	Itemzahl	Median	arithm. Mittel	Schiefe	Exzess
DAT:SR	10	5	5,21	0,10	–1,04

Da die Kurzform des *DAT:SR* auf der Rasch-Skalierung in der Voruntersuchung basiert, ist zu erwarten, dass die Items auch in der Hauptuntersuchung keine auffälligen Kennwerte aufweisen. Allerdings hat das obige Beispiel des *MRT*-Items 5 gezeigt, dass die Ergebnisse zwischen Voruntersuchung und Hauptuntersuchung nicht immer stabil sein müssen. Tatsächlich verhalten sich aber alle Items der Kurzform des *DAT:SR* im Rahmen der Rasch-Skalierung der Hauptuntersuchung hinreichend modellkonform. Bei der gemeinsamen Verteilung von Testleistung und Itemschwierigkeiten in Abbildung 6.9 konnte die Logit-Skala gröber als in der Voruntersuchung (Abb. 5.17; S. 174) eingeteilt werden, da bei der Kurzform des Tests weniger verschiedene Testleistungen auftreten können. Die Itemnummern geben die Position der Items im Untertest der Hauptuntersuchung wieder und sind daher nicht mit den Nummern in Abb. 5.17 vergleichbar.

Abbildung 6.9: Verteilung der DAT:SR-Testleistung und der Itemschwierigkeiten (n = 466)



Geschlechterunterschiede

Nachdem für die drei Referenztests für *räumliche Wahrnehmung*, *mentale Rotation* und *räumliche Visualisierung* die prinzipielle empirische Tauglichkeit festgestellt wurde, können nun vertiefende Analysen folgen. Zunächst werden die gezeigten Testleistungen auf

signifikante Geschlechterunterschiede untersucht. Dabei müssen aufgrund der Klumpenstruktur der Stichprobe *Geschlecht* und *Schulform* als Haupteffekte in einer zweifaktoriellen *ANOVA* modelliert werden. Die Haupteffekte können mit ConQuest latent geschätzt werden, indem die *ANOVA* (mit „Dummyvariablen“[118]) auf eine latente Regression zurückgeführt wird.

Signifikante Interaktionseffekte von *Geschlecht* und *Schulform* lagen für keinen der drei Untertests vor, sodass die Haupteffekte untersucht werden können. Erwartungsgemäß liegt bei allen Untertests ein signifikanter Haupteffekt des Faktors *Schulform* vor. Bei den paarweisen Vergleichen zwischen den Schulformen gibt es für keinen der drei Untertests signifikante Unterschiede zwischen der Gesamtschule und der Hauptschule. Bis auf den Vergleich der Hauptschule mit der Realschule im Untertest *WLT* sind alle anderen Unterschiede signifikant (vgl. Tab. 6.9). [119]

Tabelle 6.9: Ergebnisse der paarweisen Untersuchungen der Stufen des Faktors „Schulform“ auf signifikante Unterschiede

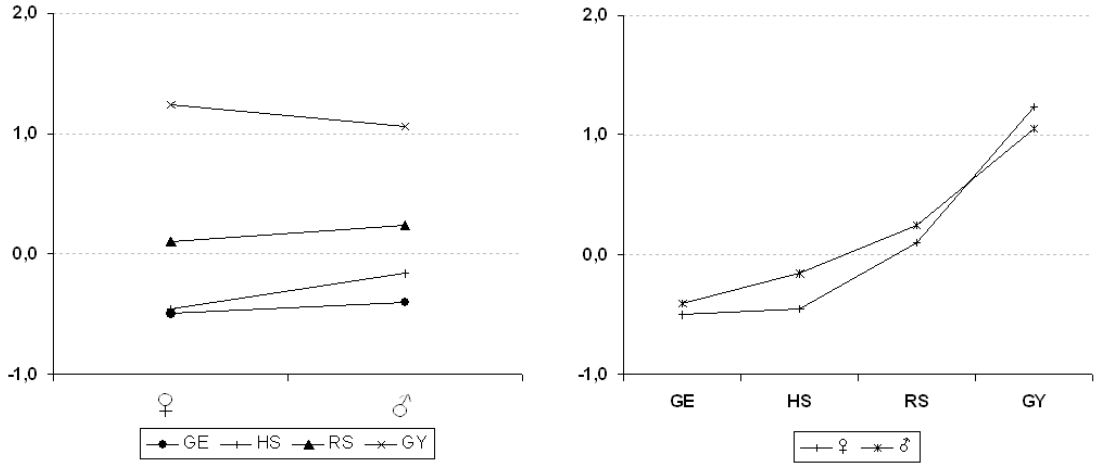
		GY	**RS**	**HS**
GE	*WLT*	ja	ja	nein
	MRT	ja	ja	nein
	DAT:SR	ja	ja	nein
HS	*WLT*	ja	nein	
	MRT	ja	ja	
	DAT:SR	ja	ja	
RS	*WLT*	ja		
	MRT	ja		
	DAT:SR	ja		

[118] Beispielsweise kann die kategorielle Variable *Schulform* mit ihren vier Ausprägungen durch drei dichotome Dummyvariablen für die latente Regression modelliert werden. Dabei kann eine Schulform als Referenzschulform verwendet werden und die drei dichotomen Variablen zeigen für die anderen Schulformen jeweils an, ob eine Versuchsperson zu dieser Schulform gehört oder nicht. Haben alle drei Dummies den Wert Null, so gehört eine Versuchsperson zur Referenzschulform.

[119] Dabei wurde das Signifikanzniveau nicht unter Berücksichtigung der Anzahl der Vergleiche korrigiert, da jeweils entsprechende Hypothesen zu den fraglichen Unterschieden aus der aktuellen Befundlage abgeleitet werden können.

Im Rahmen der vorliegenden Arbeit sind die Haupteffekte des Faktors *Geschlecht* von besonderer Bedeutung. Passend zur Hypothese H_01 (vgl. Kap. 6.1.1, S. 194) liegen signifikante Haupteffekte zugunsten der Jungen für die Untertests *WLT* und *MRT* vor. Entgegen der Hypothese H_02 liegt – anders als in der Voruntersuchung – kein signifikanter Haupteffekt (zugunsten der Mädchen) für den Untertest *DAT:SR* vor. Da das Ergebnis zum *DAT:SR* überraschend ist, lohnt sich ein Blick auf die zugehörigen Interaktionsdiagramme, obwohl keine signifikanten Interaktionseffekte vorliegen. Abbildung 6.10 zeigt, dass das Gymnasium die einzige Schule ist, an der die Mädchen bessere *DAT:SR*-Leistungen erzielen als die Jungen. In der Stichprobe der Voruntersuchung erzielten Mädchen auch an der Gesamtschule im Mittel (etwas) bessere Leistungen als Jungen.

Abbildung 6.10: Interaktionsdiagramme für den DAT:SR



Die zuvor angesprochenen Vergleiche zwischen den Schulformen finden sich optisch gut nachvollziehbar im linken Interaktionsdiagramm wieder. Das rechte Diagramm zeigt optisch zwar einen Interaktionseffekt (durch die gekreuzten Abschnitte zwischen „RS“ und „GY“) an, lässt aber auch erahnen, dass dieser nicht groß genug ist, um statistisch signifikant zu werden.

Da Signifikanztests noch nicht die Frage der Bedeutsamkeit von Unterschieden beantworten, ist ein Blick auf die zugehörigen Effektstärken wichtig. Tabelle 6.10 gibt die Effektstärken für die drei betrachteten Untertests an.

Tabelle 6.10: Geschlechterunterschiede in der Raumvorstellung bei den Untertests WLT, MRT und DAT:SR (aus Sicht der Jungen; n = 466)

	Effektstärke	**Signifikanz**
WLT	**0,40**	ja (p = 0,002)
MRT	**0,80**	ja (p = 0,000)
DAT:SR	0,08	nein (p = 0,462)

Die Effektstärken für Geschlechterunterschiede liegen damit in derselben Größenordnung der Effektstärken, die Linn & Petersen (1985, S. 1486) in ihrer Meta-Analyse ermittelt haben. Auf der Ebene der einzelnen Schulen liegen die Effektstärken für den besonders interessierenden Untertest *MRT* zwischen 0,61 (Gesamtschule) und 0,95 (Realschule).

Lineare Zusammenhänge zwischen den drei Komponenten

Für die Stichprobe der Voruntersuchung wurden die latenten Korrelationen zwischen je zwei der drei Untertests in einer mehrdimensionalen Modellierung mit Werten zwischen 0,44 und 0,64 geschätzt (vgl. Tab. 5.9, S. 180). Damit war zwar ein Zusammenhang zwischen den Untertests gegeben, dieser unterschied sich aber substanziell von einer Eindimensionalität des Konstrukts *Raumvorstellung*. Der engste Zusammenhang bestand zwischen *mentaler Rotation* (*MRT*) und *räumlicher Visualisierung* (*DAT:SR*), was inhaltlich auf die Anteile *mentaler Rotation* zurückgeführt werden kann, die bei *räumlicher Visualisierung* eine, wenn auch dem analytischen Vorgehen untergeordnete, Rolle spielen.

Im Rahmen der Hauptuntersuchung wurden die drei genannten und zum Teil optimierten Tests in einer Stichprobe eingesetzt, die bezüglich Leistungsvarianz und -verteilung repräsentativer ist; insbesondere dürften etwas größere Varianzen als in der Voruntersuchung und somit auch etwas größere Korrelationskoeffizienten auftreten. Auf dieser Basis sollen aktuelle Befunde zum Zusammenhang zwischen den drei Komponenten der *Raumvorstellung* bzw. zwischen den konkreten Referenztests generiert werden.

Da es zunächst prinzipiell auch sein könnte, dass die Bearbeitungen der drei Untertests hinreichend gut mit *einer* Globalskala *Raumvorstellung* modelliert werden können, wird ein eindimensionales Modell empirisch mit dem entsprechenden dreidimensionalen Modell verglichen. Die Skalierung nach dem ein- bzw. dreidimensionalen zweikategoriellen *RM* erfolgt dabei jeweils mit *ConQuest*. Tabelle 6.11 fasst die Kennwerte für die Güte der konkurrierenden Modelle zusammen (vgl. Kap. 4.2.3).

Tabelle 6.11: Kennwerte für die Güte der konkurrierenden Modelle

Modell	Parameter	–2·ln(L)	*AIC*	*BIC*
eindimensionales *RM* (mit geglätteter Verteilung)	27	12 899	12 953	13 065
dreidimensionales *RM* (mit geglätteter Verteilung)	32	12 607	12 671	12 804

Sowohl die Maße *AIC* und *BIC* als auch ein zu heuristischen Zwecken durchgeführter *LRT* sprechen eindeutig für das dreidimensionale Modell.[120] Für den zu heuristischen Zwecken durchgeführten *LRT* ergeben sich aus den Parameterzahlen 32 – 27 = 5 Freiheitsgrade für die χ^2-verteilte Prüfgröße, die den Wert 12 899 – 12 607 = 292 annimmt. Dies ergibt einen p-Wert von ca. 10^{-60}. Die *LRT*-Nullhypothese, das eindimensionale Modell passe besser zu den Daten, sollte also verworfen werden – so wie es die latenten Korrelationen in der Voruntersuchung schon nahe gelegt haben. Da die Voraussetzungen für die Durchführung eines *LRT* nicht hinreichend erfüllt sind, sollte dieser Test aber nur als zusätzliche heuristische Betrachtung einbezogenen werden und sich eine Entscheidung vornehmlich auf die informationstheoretischen Maße stützen.

Das Ergebnis des obigen „Modellgeltungstests“ wird auch anhand der latenten Korrelationen zwischen den Untertests in der Hauptuntersuchung verdeutlicht. Im dreidimensionalen Modell wurden die in Tab. 6.12 angegebenen Werte geschätzt.

Tabelle 6.12: Im dreidimensionalen RM geschätzte latente Korrelationen zwischen den Raumvorstellungstests WLT, MRT und DAT:SR (n = 466)

	DAT:SR	***MRT***
WLT	**0,59** (0,44*)	**0,55** (0,55*)
MRT	**0,74** (0,64*)	

* in Klammern: latente Korrelationen aus der Voruntersuchung

Die geschätzten Werte in Tab. 6.12 unterstützen die Hypothese H_03 (vgl. Kap. 6.1.1) deutlich: Sie sind sowohl substanziell von Null als auch substanziell von Eins verschieden. Im Vergleich zu den Korrelationen der Voruntersuchung (in Klammern) ergeben sich Än-

[120] Wären die Voraussetzungen für einen *LRT* erfüllt, könnte man auf die Maße *AIC* und *BIC* verzichten. Allerdings sind im vorliegenden Fall, die Bedingungen für eine asymptotische χ^2-Verteilung der Prüfgröße nicht hinreichend erfüllt (vgl. Kap. 4.2.3).

derungen bei den beiden Werten, an denen der *DAT:SR* beteiligt ist. Dies kann daran liegen, dass insbesondere die Ergebnisse zum *DAT:SR* in der Voruntersuchung nur eine eingeschränkte Varianz aufwiesen. Der Zusammenhang zwischen *MRT* und *DAT:SR* ist deutlich enger als die anderen beiden Zusammenhänge. Wie oben dargestellt wurde, ist dies aufgrund der Anteile *mentaler Rotation* bei der *räumlichen Visualisierung* inhaltlich plausibel und konstruktkonform.

6.4.2 Mathematikleistung

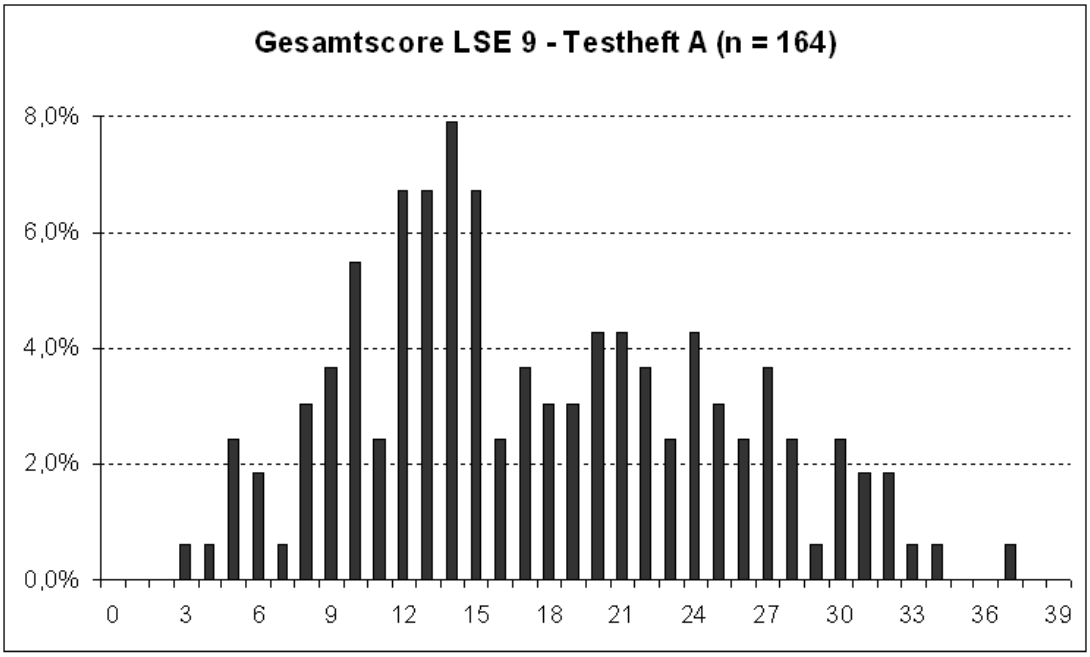
Die Auswertung des *LSE 9*-Mathematiktests wird im Folgenden zunächst nach beiden Testheften getrennt, wie bei den Raumvorstellungstests, durchgeführt. Anschließend werden die Ergebnisse der gemeinsamen Skalierung der gesamten Stichprobe für beide Testhefte zusammen berichtet.

LSE 9-Mathematik – Testheft A

Anhand der Mathematiktests aus den *LSE 9* kann gut verdeutlicht werden, was komplexe „Fähigkeitssyndrome“ (vgl. Kap. 2.3.3) wie *Mathematikleistung* von relativ eng gefassten kognitiven Leistungen, wie z. B. *mentaler Rotation*, aus der Sicht des Testens unterscheidet. Während die oben ausgewerteten Untertests zur *Raumvorstellung* relativ homogene Itemschwierigkeiten hatten, kann bei einem Mathematiktest über eine größere Anzahl schwierigkeitsbestimmender Faktoren (vgl. Kap. 2.3.1) ein Itemsatz konstruiert werden, der in allen Leistungsbereichen differenziert und somit Decken- und Bodeneffekte[121] vermeiden kann. Abbildung 6.11 zeigt eine hieraus resultierende Verteilung der Teilstichprobe für das Testheft A.

[121] Ein Bodeneffekt wäre bei einer Lernstandserhebung auch kein pädagogisch angemessenes Signal für die Schülerinnen und Schüler.

Abbildung 6.11: Verteilung der Versuchspersonen nach LSE 9-Gesamtscore – Testheft A



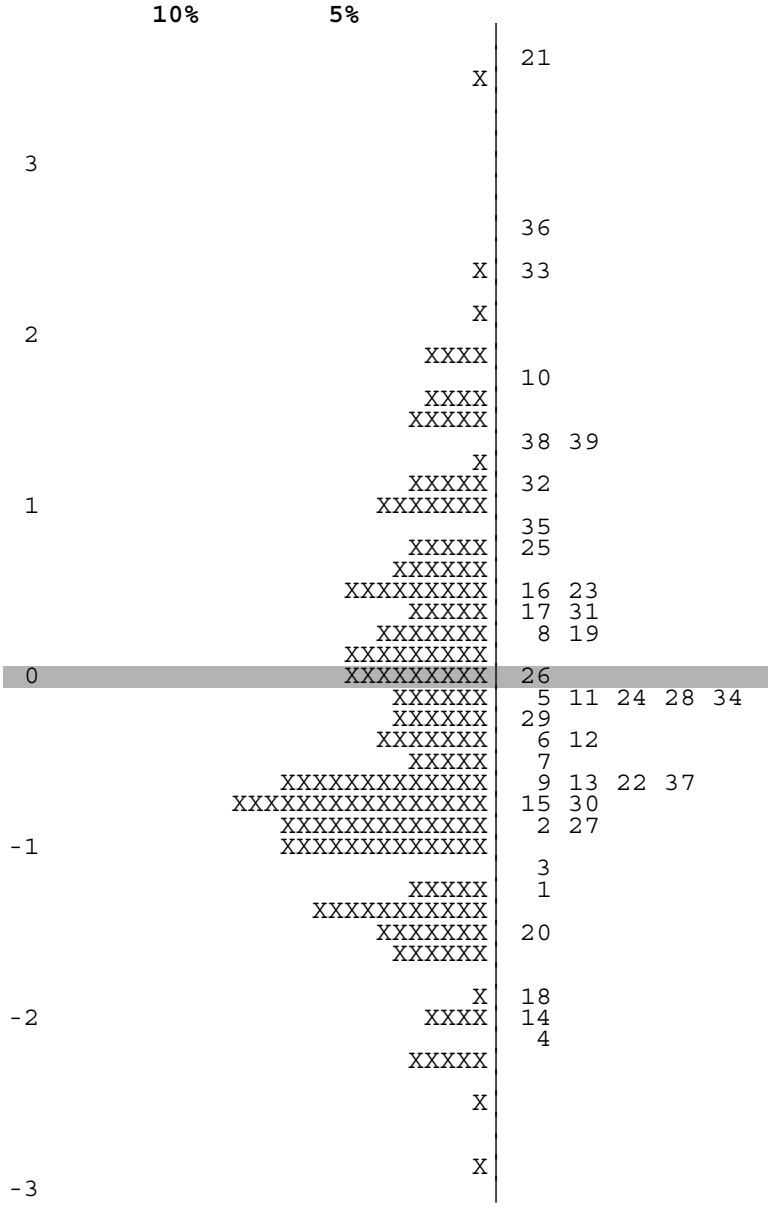
Das Testheft A war für die Teilstichprobe tendenziell schwierig, es gab aber keine Schülerinnen oder Schüler, die weniger als drei Items richtig bearbeitet haben. Die Tendenz „schwierig“ liegt vor allem in den 20 Ankeritems begründet, die sowohl in Testheft A als auch in Testheft B stehen. Der optische Eindruck der Verteilung wird wiederum durch die zugehörigen Kennwerte unterstützt (Tab. 6.13).

Tabelle 6.13: Kennwerte der Verteilung nach LSE 9-Gesamtscore – Testheft A (n = 164)

Test	Itemzahl	Median	arithm. Mittel	Schiefe	Exzess
***LSE 9* – TH A**	39	16	17,54	0,35	–0,65

Bei der gemeinsamen Verteilung von Testleistung und Itemschwierigkeiten auf der Logit-Skala in Abbildung 6.12 repräsentiert ein X ca. 0,5 % der Stichprobe. Die Verteilung zeigt gut, dass in praktisch allen Leistungsbereichen Items entsprechender Schwierigkeit zur Verfügung stehen. Lediglich im untersten Leistungsbereich hätten es noch ein oder zwei Items mehr sein können: Das empirisch leichteste Item hat eine Lösungsquote von 82,9 %; erfahrungsgemäß können aber auch Lösungsquoten von 90 % und darüber durch fachdidaktisch und psychometrisch angemessene Items realisiert werden.

Abbildung 6.12: Verteilung der LSE 9-Testleistung und der Itemschwierigkeiten für das Testheft A (n = 164)



Arithm. Mittel (Testleistung): -0,34
Zugehörige Standardabweichung: 1,15

Itemblock A: Items 1-19
Itemblock X: Items 20-39 (Ankeritems)

LSE 9-Mathematik – Testheft B

Für Testheft B führen die Betrachtungen, die für Testheft A angestellt wurden, überwiegend zu vergleichbaren Resultaten. Allerdings zeigen Abb. 6.13 und Tab. 6.14, dass das Testheft B für die Teilstichprobe tendenziell leicht war, was wiederum auf die 20 gemeinsamen Items mit Testheft A zurückzuführen ist.

Abbildung 6.13: Verteilung der Versuchspersonen nach LSE 9-Gesamtscore – Testheft B

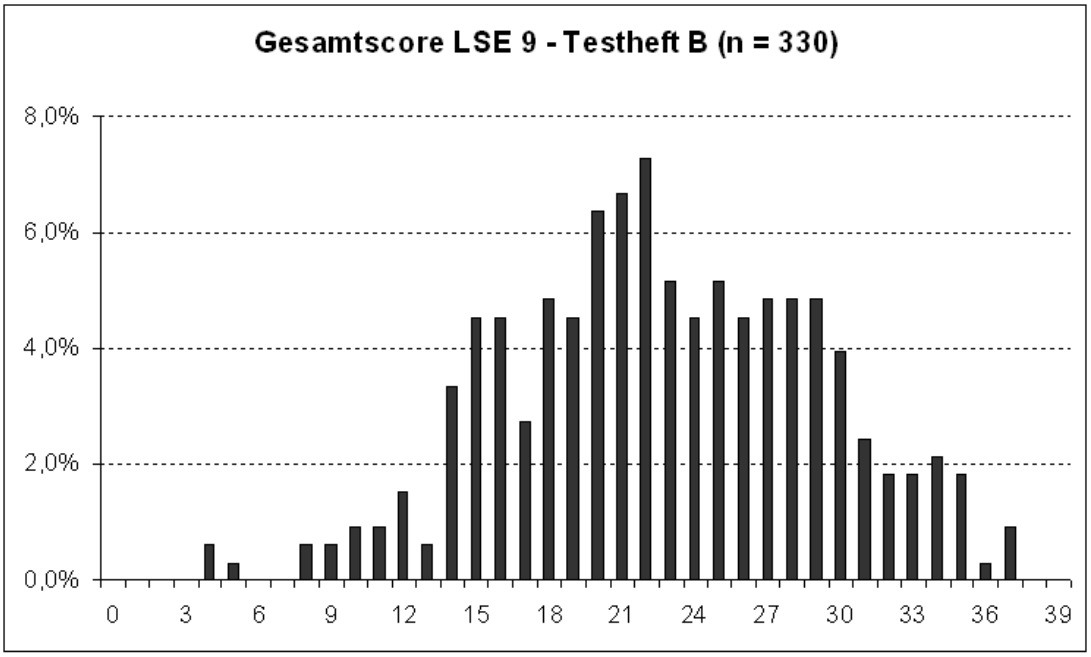


Abbildung 6.13 zeigt aber auch deutlich, dass weder Decken- noch Bodeneffekte aufgetreten sind. Wenn auch bei Testheft B und der zugehörigen Population alle Leistungsbereiche durch Items entsprechender Schwierigkeit abgedeckt sind, kann man wie bei Testheft A davon ausgehen, dass die in der Stichprobe der Hauptuntersuchung vorhandene Leistungsvarianz insgesamt angemessen erfasst wird.

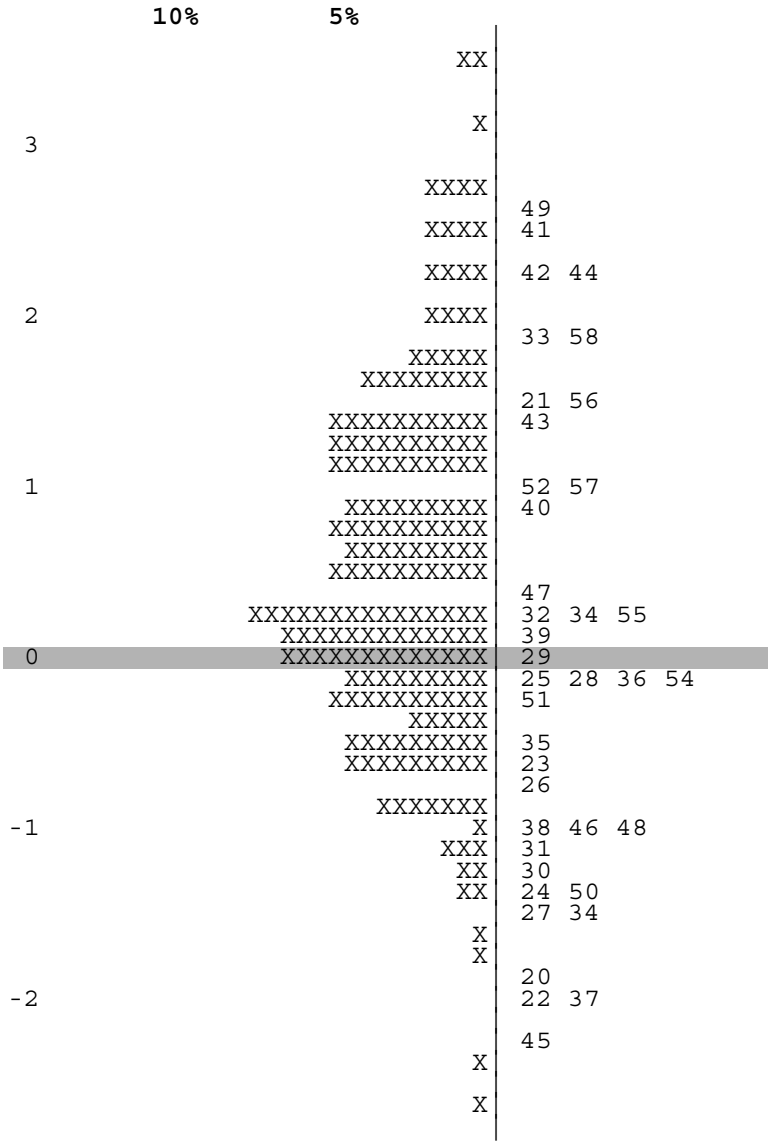
Tabelle 6.14: Kennwerte der Verteilung nach LSE 9-Gesamtscore – Testheft B (n = 330)

Test	Itemzahl	Median	arithm. Mittel	Schiefe	Exzess
***LSE 9* – TH B**	39	22	22,68	–0,11	–0,25

Die gemeinsame Verteilung von Testleistung und Itemschwierigkeiten auf die Logit-Skala in Abbildung 6.14, bei der ein X ca. 0,5 % der Stichprobe repräsentiert, zeigt, dass lediglich im obersten Leistungsbereich noch Items ergänzt werden könnten. Allerdings haben bereits die vorhandenen Items auch zwischen den besten Schülerinnen und Schülern der Stichprobe differenziert und die beiden theoretisch möglichen höchsten Gesamtscores wurden von keinem erzielt. Dies ist vor dem Hintergrund des verwendeten Testmodells plausibel: Die besten Testleistungen liegen auf der Logit-Skala ca. eine Einheit über den Schwierigkeiten der Items 49 und 41. Durch Einsetzen in die Modellgleichung (vgl. Kap. 2.1.2, S. 23) erhält man bei dieser Differenz eine Lösungswahrscheinlichkeit von 0,73

(für beide Items). Aufgrund der stochastischen Unabhängigkeit, die eine Modelleigenschaft ist, beträgt die Wahrscheinlichkeit, beide Items richtig zu lösen, nur noch 0,53.

Abbildung 6.14: Verteilung der LSE 9-Testleistung und der Itemschwierigkeiten für das Testheft B (n = 330)



Arithm. Mittel (Testleistung): 0,48
Zugehörige Standardabweichung: 1,07

Itemblock X: Items 20-39 (Ankeritems)
Itemblock B: Items 40-59

Aus testtheoretischer Sicht ist bei den Verteilungen zu Testheft A und Testheft B in den Abbildungen 6.12 und 6.14 unter anderem interessant, wie sich die 20 Ankeritems jeweils zueinander verhalten. Eine Eigenschaft des *RM* ist – bei perfekter Passung auf die Daten – die *spezifische Objektivität* (vgl. Kap. 2.1.2), die unter anderem bedeutet, dass die Relation zweier Items auf der Logit-Skala nicht von der Personenstichprobe abhängt, für die der Test skaliert wird und für die die entsprechenden Parameter geschätzt werden. D. h. die

Differenz der Itemschwierigkeiten ist invariant, nicht aber die Positionierung auf der Logit-Skala; dies entspricht gerade dem Differenzskalenniveau.

In der Praxis passt das *RM* aber immer nur näherungsweise und nicht perfekt auf beobachtete Daten. Daher können auch bei insgesamt akzeptablen Modellkennwerten Phänomene beobachtet werden wie bei den Items 21, 36 und 33. Während die Items 21 und 36 in der Testheft A-Teilstichprobe schwieriger sind als Item 33 (Differenzen: 1,27 für Item 21 und 0,27 für Item 36), ergibt sich in der Testheft B-Teilstichprobe ein umgekehrtes Bild (Differenzen: –0,39 für Item 21 und –1,94 für Item 36). Die Verschiebungen der relativen Itemschwierigkeiten können inhaltlich gut erklärt werden: Bei den Items 21 und 36 müssen funktionale Zusammenhänge algebraisch modelliert werden, was in den Bildungsgängen, die Testheft A bearbeitet haben, bis zum Testzeitpunkt (und in der Regel auch darüber hinaus) keine große Rolle spielt. Bei Item 33 muss hingegen ein Funktionsgraph qualitativ mit Bezug zu einer Realsituation interpretiert werden, was für alle Schülerinnen und Schülern beider Teilstichproben relativ schwierig ist.

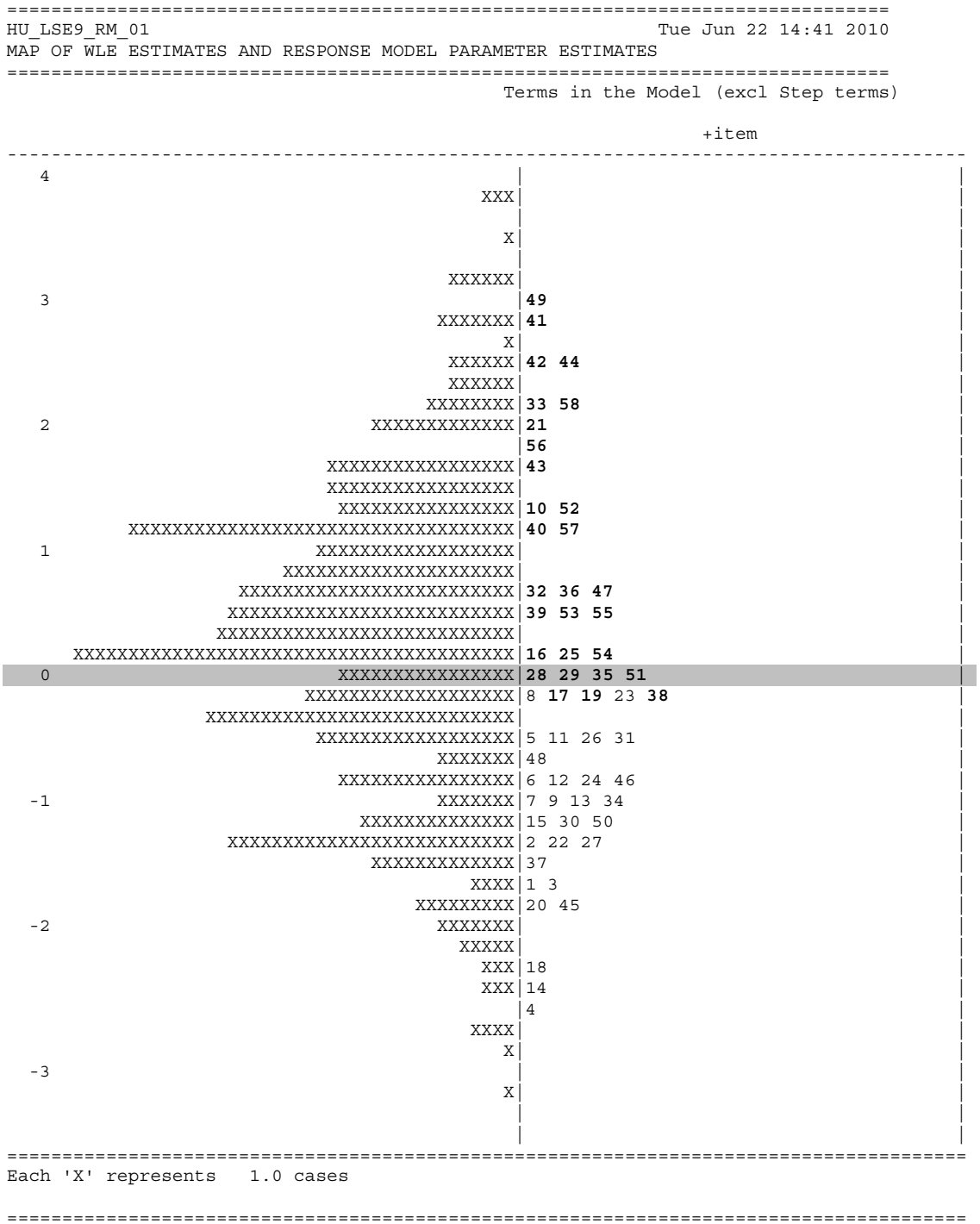
Solche Verschiebungen von relativen Itemschwierigkeiten (in Bezug auf ein anderes ausgewähltes Item) können bei enger gefassten Tests, wie z. B. dem *WLT*, der vermutlich genau einen mentalen Prozess anspricht, kaum beobachtet werden. Die größere Homogenität in dem Sinne, dass alle Items dieselbe Eigenschaft messen, ist bei eng gefassten Tests oft augenscheinlich gegeben. Fachleistungstests hingegen erfassen komplexe „Fähigkeitssyndrome“, sodass Verschiebungen der Itemschwierigkeiten z. B. zwischen Stichproben mit unterschiedlichen curricularen Voraussetzungen eher die Regel als die Ausnahme sind. Im Sinne eines Skalierungspragmatismus stellen einzelne Verschiebungen die Passung des verwendeten Modells nicht generell infrage, die Passung verschlechtert sich aber.

LSE 9-Mathematik – Gesamttest

Auch wenn die empirischen und theoretischen Betrachtungen der Ankeritems 21, 33 und 36 auf eine eingeschränkte Homogenität der Items und der Personen hindeuten, lässt sich der LSE 9-Mathematiktest mit hinreichend guten Kennwerten über beide Teilstichproben und beide Testhefte hinweg als *ein* Test skalieren. Mit *ConQuest* kann diese Skalierung des Gesamttests auf der Basis des *RM* ohne großen Aufwand umgesetzt werden. Hier wird die Möglichkeit des *RM* genutzt, dass verschiedene Testhefte auf einer Dimension skaliert werden können, wenn die Überlappungsbereiche (Ankeritems) zwischen je zwei der Testhefte hinreichend groß sind („Multi-Matrix-Design“).

Diese gemeinsame Skalierung entspricht auch der Auswertung der *LSE 9* und der Rückmeldung von Leistungsdaten an die Schulen (vgl. Fleischer et al., 2007). Die Modellieren-Skala für die *LSE 9*-Mathematik wurde nicht nach Bildungsgängen unterschieden, sondern passt dem Anspruch nach für alle Schülerinnen und Schüler im Land. Abbildung 6.15 gibt den *ConQuest*-Output wieder, wobei ein X eine Versuchsperson (ca. 0,2 %) repräsentiert.

Abbildung 6.15: Verteilung der LSE 9-Testleistung und der Itemschwierigkeiten für den Gesamttest (gemeinsame Skalierung der Testhefte A und B; n = 494)



Insgesamt zeigt Abb. 6.15, dass der Gesamttest für die Gesamtstichprobe sehr gut ausbalanciert ist: Das Leistungsspektrum wird im Wesentlichen durch Items adäquater Schwierigkeit abgedeckt (in den Bereichen sehr hoher bzw. niedriger Leistung kann durch die nächst gelegenen Items noch hinreichend gut differenziert werden) und im Bereich mit einer hohen Personendichte sind auch viele entsprechende Items angesiedelt.

Für die vertiefenden Analysen im Sinne der Hypothesen (vgl. Kap. 6.1) müssen die Items des *LSE 9*-Mathematiktests (a) nach empirischer Schwierigkeit und (b) nach Aufgabentypen (*Typen mathematischen Arbeitens*) klassifiziert werden. Die Klassifikation nach empirischer Schwierigkeit kann gut durch die „Split-Half-Idee" erfolgen, indem die 58 Items des Gesamttests in die empirisch schwierige und die empirisch leichte Hälfte getrennt werden. Dabei stellt –0,1 einen möglichen kritischen Wert dar, der zwischen den beiden Medianen der Itemreihe liegt: Das „leichteste der Schwierigen" ist Item 38 ($\sigma_{38} = -0{,}099$) und das „schwierigste der Leichten" ist Item 8 ($\sigma_8 = -0{,}138$).

Tabelle 6.15 zeigt, dass sich die beiden so gebildeten Testhälften komplementär aus den verschiedenen Itemblöcken zusammensetzen: Während in beiden Hälften jeweils zehn Ankeritems sind, finden sich vier Testheft A-Items in der empirisch schwierigen Hälfte und umgekehrt vier Testheft B-Items in der empirisch leichten Hälfte wieder. In der Tabelle sind auch die schwierigsten und leichtesten Items der jeweiligen Testhälfte und die mittlere Itemschwierigkeit angegeben. Aufgrund der Summennormierung der Itemschwierigkeiten haben die mittleren Itemschwierigkeiten beider Teiltests den gleichen Betrag. Die mittlere Schwierigkeitsdifferenz zwischen den beiden Testhälften beträgt 2,160 auf der Logit-Skala.

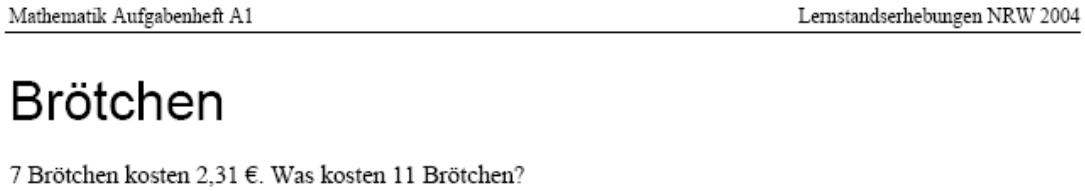
Tabelle 6.15: Aufteilung des LSE 9-Gesamttests nach empirischer Schwierigkeit

Testhälfte	Verteilung auf Itemblöcke		schwierigstes und leichtestes Items
empirisch schwierig	nur TH-B	15	Items 49 „BSKIE"[122], $\sigma_{49} = 3{,}021$ Item 38 „XAUTC2", $\sigma_{38} = -0{,}099$ (mittlere Schwierigkeit: 1,080)
	Ankeritems	10	
	nur TH-A	4	
empirisch leicht	nur TH-B	4	Item 8 „ARENC", $\sigma_8 = -0{,}138$ Item 4 „ABRO", $\sigma_4 = -2{,}558$ (mittlere Schwierigkeit: –1,080)
	Ankeritems	10	
	nur TH-A	15	

Zur inhaltlichen Konkretisierung des *LSE 9*-Tests und zur Veranschaulichung der Bandbreite der Anforderungen werden in den folgenden Abbildungen 6.16 und 6.17 das leichteste und das schwierigste Item des Gesamttests dargestellt.

[122] Die „Itemnamen" sind wie folgt aufgebaut: Der erste Buchstabe gibt den Itemblock an, zu dem das Item gehört (A, X oder B), die folgenden drei Buchstaben repräsentieren den Aufgabennamen aus dem Testheft und am Ende stehen ein oder zwei Buchstaben, die die jeweilige Teilaufgabe angeben.

Abbildung 6.16: Item 4 „ABRO“ (nur in Testheft A; $\sigma_4 = -2{,}558$)

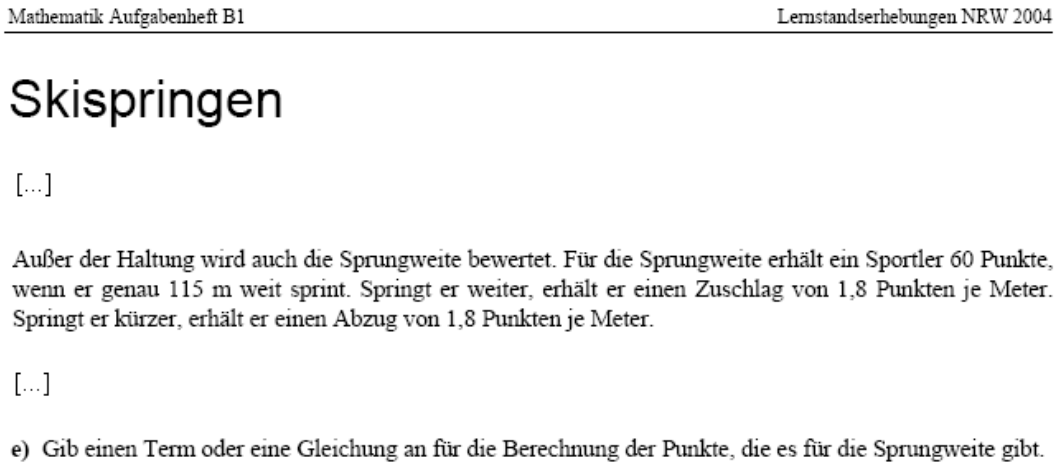


Mathematik Aufgabenheft A1 Lernstandserhebungen NRW 2004

Brötchen

7 Brötchen kosten 2,31 €. Was kosten 11 Brötchen?

Abbildung 6.17: Item 49 „BSKIE“ (nur in Testheft B; $\sigma_{49} = 3{,}021$)



Mathematik Aufgabenheft B1 Lernstandserhebungen NRW 2004

Skispringen

[...]

Außer der Haltung wird auch die Sprungweite bewertet. Für die Sprungweite erhält ein Sportler 60 Punkte, wenn er genau 115 m weit sprint. Springt er weiter, erhält er einen Zuschlag von 1,8 Punkten je Meter. Springt er kürzer, erhält er einen Abzug von 1,8 Punkten je Meter.

[...]

e) Gib einen Term oder eine Gleichung an für die Berechnung der Punkte, die es für die Sprungweite gibt.

Geschlechterunterschiede

Bei der Analyse von Geschlechterunterschieden in der *Mathematikleistung* kann genauso vorgegangen werden wie bei den Auswertungen zur *Raumvorstellung*. Dabei liegen für den *LSE 9*-Mathematiktest keine signifikanten Interaktionseffekte zwischen *Geschlecht* und *Schulform* vor, aber beide Haupteffekte sind signifikant. Vergleicht man die Schulformen paarweise, so liegt kein signifikanter Unterschied zwischen der Gesamtschule und der Hauptschule vor, während alle anderen Vergleiche signifikante Unterschiede aufweisen.

Der Haupteffekt des Faktors *Geschlecht* hat bei einer latenten Modellierung eine Effektstärke von 0,63 zugunsten der Jungen. Schulformspezifisch betrachtet liegt die Effektstärke zugunsten der Jungen zwischen 0,49 an der Gesamtschule und 0,94 an der Realschule. Diese Ergebnisse unterstützen die Hypothese H_04a, die von signifikanten Leistungsunterschieden zugunsten der Jungen ausgeht, ohne Einschränkung.

Da einige vorliegende Befunde (vgl. Kap. 2.3.2) darauf hindeuten, dass Mädchen bei empirisch leichten Items ähnlich gute Ergebnisse erzielen wie Jungen und die Geschlechterunterschiede bei schwierigen Items entstehen, werden die obigen Betrachtungen auch für die leichte und die schwierige Testhälfte durchgeführt. Dabei ergibt sich grundsätzlich das

gleiche Bild bezüglich vorhandener bzw. nicht vorhandener Interaktions- und Haupteffekte. Die Effektstärken für die signifikanten Leistungsunterschiede zugunsten der Jungen betragen in der Gesamtstichprobe 0,57 für die leichte und 0,67 für die schwierige Testhälfte. Diese Werte sind in ihrer Tendenz erwartungskonform, allerdings sind die signifikanten Geschlechterunterschiede in der empirisch leichten Testhälfte nicht verschwunden.

Da die Tendenz bei den geringen Unterschieden in den Werten auch zufallsbedingt auftreten kann, werden weitere Subtests nach Schwierigkeit gebildet. Dabei muss darauf geachtet werden, dass für die Skalierung in der Gesamtstichprobe in allen verwendeten Subtests hinreichend viele Items der Itemblöcke A, X und B sind. Unter diesen Rahmenbedingungen wurden die Subtests „sehr schwierig“ (2 x A, 6 x X, 14 x B), „sehr leicht“ (14 x A, 5 x X, 4 x B) und „sehr sehr leicht“ (7 x A, 5 x X, 2 x B) gebildet. Tabelle 6.16 stellt Itemzahlen, die mittleren Itemschwierigkeiten und die beobachteten Effektstärken dar.

Tabelle 6.16: Effektstärken für Geschlechterunterschiede in LSE 9-Subtests (empirische Schwierigkeit; n = 464)

Testversion	Anzahl der Items	Mittlere Itemschwierigkeit	Effektstärke
sehr sehr leicht	14	–1,578	0,44
sehr leicht	23	–1,271	0,52
leichte Hälfte	29	–1,080	0,57
Gesamttest	58	0,000	0,63
schwierige Hälfte	29	1,080	0,67
sehr schwierig	22	1,421	0,69

Die Tendenz bei der Effektstärke bestätigt sich also. Insofern wird die Hypothese H_04b, die im Einklang mit der Befundlage in Kap. 2.3.2 davon ausgeht, dass Geschlechterunterschiede in der *Mathematikleistung* auf empirisch schwierigen Items basieren, tendenziell unterstützt. Ein Verschwinden der Geschlechterunterschiede konnte aber auch bei der Testversion mit den 14 leichtesten *LSE 9*-Items nicht festgestellt werden. Dies kann möglicherweise an den aktuellen Testkonzepten liegen: Grundbildungsorientierte Mathematiktests wie die *LSE 9* enthalten nur in geringen Anteilen Aufgaben, die direkt algorithmisch bearbeitet werden können und die überwiegend aus dem Bereich *Arithmetik/Algebra* stammen; die Anzahl ist dabei so gering, dass sich aus diesen Items kein Subtest bilden lässt.

Da die verschiedenen nach empirischer Schwierigkeit zusammengestellten Subtests zu unterschiedlichen differenziellen Befunden führen, könnte man vermuten, dass eine mehrdimensionale Modellierung des *LSE 9*-Mathematiktests bessere empirische Kennwerte hat

als eine eindimensionale Modellierung. Dies wird für die Subtests „leichte Hälfte“ und „schwierige Hälfte“ untersucht, wobei die Items eines Subtests jeweils einer Dimension zugeordnet werden. Da die Zuordnung dann ausschließlich nach der empirischen Schwierigkeit aus der gemeinsamen Skalierung aller Items erfolgt, können die Dimensionen nicht inhaltlich beschrieben werden. Auch die empirischen Ergebnisse deuten nicht darauf hin, dass durch die beiden „Schwierigkeitshälften“ des Tests eigene Dimensionen beschrieben werden. Ein heuristisch durchgeführter *LRT* spricht – analog zur Dimensionalitätsanalyse der Raumvorstellungstests (vgl. S. 216) – ebenso wie die informationstheoretischen Maße eindeutig für die eindimensionale Modellierung. Auch die im weniger gut passenden zweidimensionalen Modell geschätzte latente Korrelation spricht mit einem Wert von 0,99 eindeutig gegen eine empirische Trennbarkeit der „Schwierigkeitsdimensionen“.

Eine zweite Suchrichtung für die inhaltliche Erklärung von Geschlechterunterschieden in der *Mathematikleistung* stellen die *Typen mathematischen Arbeitens* dar (vgl. Kap. 2.3.1). Stanat & Kunter (2001, S. 257; Zitat auf S. 55 der vorliegenden Arbeit) haben bei *DIF*-Analysen *technische Aufgaben* als relative Stärke von Mädchen und *rechnerische Modellierungsaufgaben* als relative Stärke von Jungen identifiziert. In Anlehnung an Neubrand et al. (2002) werden im Folgenden *technische Aufgaben*, *rechnerische Modellierungsaufgaben* und *begriffliche Modellierungsaufgaben* unterschieden. Eine Klassifikation der *LSE 9*-Aufgaben nach diesen Kategorien ergibt, dass alle drei Typen in allen drei Itemblöcken nahezu gleich vertreten sind (vgl. Tab. 6.17). Eine Konkretisierung der Kategorien anhand von Beispielitems leisten weiter unten die Abbildungen 6.18 bis 6.20.

Tabelle 6.17: Verteilung der LSE 9-Items nach Aufgabentypen

Itemblock	technisch	rechnerische Modellierung	begriffliche Modellierung
A	8	7	4
X	7	6	7
B	5	8	6
Gesamt	20	21	17

Da die Aufgabentypen ausgewogen über die Itemblöcke verteilt sind, können typspezifische Subtests gebildet und für die Gesamtstichprobe skaliert werden. Das arithmetische Mittel der Itemschwierigkeiten beträgt für die *technischen Aufgaben* –0,623, für die *rechnerischen Modellierungsaufgaben* 0,206 und für die *begrifflichen Modellierungsaufgaben* 0,563. Diese Tendenz der mittleren Itemschwierigkeiten für die drei Aufgabentypen lässt sich in den meisten Tests finden, allerdings gibt es z. B. auch empirisch schwierige *techni-*

sche Aufgaben (Item 21; $\sigma_{21} = 2{,}048$) oder empirisch leichte *begriffliche Modellierungsaufgaben* (Item 18; $\sigma_{18} = -2{,}280$).

Ein charakteristisches Beispiel für eine *technische Aufgabe* ist Item 20 (Abb. 6.18), bei dem sowohl das Verfahren, das angewendet werden soll, benannt wird als auch alle für die Rechnung erforderlichen Größen explizit gegeben sind. Im Unterschied zu *technischen Aufgaben* müssen die Schülerinnen und Schüler bei einer *rechnerischen Modellierung* wie bei Item 5 (Abb. 6.19) zunächst ein der Situation angemessenes Modell aufstellen und anschließend ein passendes Verfahren anwenden.

Abbildung 6.18: Beispiel „technische Aufgabe“ (Item 20 „XUMFA“; $\sigma_{20} = -1{,}705$)

Lernstandserhebungen NRW 2004 Mathematik Aufgabenheft B1

Umfangsterm

a) Die nebenstehende Abbildung zeigt ein Trapez. Berechne den Umfang.

Ergebnis: ☐ *cm*

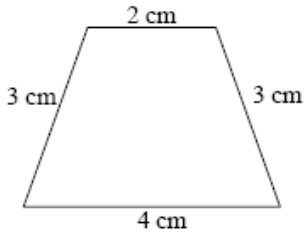
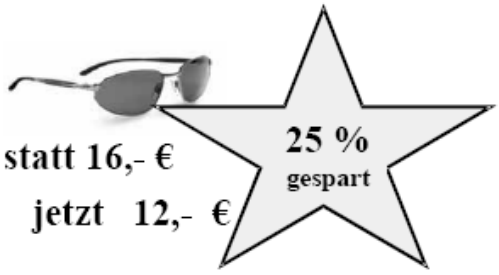


Abbildung 6.19: Beispiel „rechnerische Modellierung“ (Item 5 „ASCHL“; $\sigma_5 = -0{,}479$)

Lernstandserhebungen NRW 2004 Mathematik Aufgabenheft A1

Schlussverkauf

Im Schlussverkauf setzt ein Sportgeschäft die Preise vieler Artikel herunter. Die Werbung (siehe Abbildung rechts) zeigt den alten Preis, den neuen Preis und den Preisnachlass in Prozent.



Bei dem folgenden Artikel fehlt eine Angabe. Berechne diese Angabe.

statt 90,- €
jetzt ... €

Die Inliner kosten jetzt ☐ €.

Bei einer *begrifflichen Modellierung* steht ebenfalls zunächst das Aufstellen eines passenden Modells im Vordergrund, allerdings muss anschließend nicht algorithmisch, sondern vor allem konzeptuell agiert werden. Bei Item 44 (Abb. 6.20) ist der konzeptuelle Unterschied von relativer Häufigkeit und Wahrscheinlichkeit ebenso bedeutsam wie die Teilsymmetrie des Zylinders.

Abbildung 6.20: Beispiel „begriffliche Modellierung“ (Item 44 „BHOLB“; σ_{44} = 2,603)

Holzzylinder

Mit dem Holzzylinder in der Abbildung kann man „würfeln“.

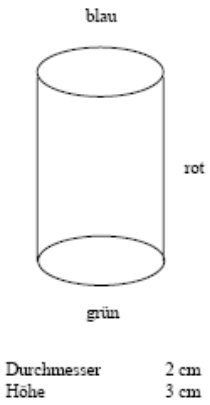
Du wirfst einen solchen Zylinder.

Die Farbe, die anschließend nach oben zeigt, ist das Ergebnis des Wurfes.

Der Zylinder wird 120mal geworfen.

a) Berechne aus den Häufigkeiten in der Tabelle die relativen Häufigkeiten.

b) Schätze dann sinnvolle Wahrscheinlichkeiten für die drei Farben.



Farbe	blau	rot	grün
Häufigkeit	27	63	30
a) relative Häufigkeit			
b) Wahrscheinlichkeit			

Für die drei so gebildeten fiktiven[123] Subtests *technische Aufgaben*, *rechnerische Modellierungsaufgaben* und *begriffliche Modellierungsaufgaben* können Analysen bezüglich möglicher Geschlechterunterschiede wieder nach dem bewährten Verfahren unter Berücksichtigung des zweiten Faktors (*Schulform*) durchgeführt werden.

Die Effektstärken wurden wiederum als Parameterschätzungen in latenten Modellen ermittelt und sind in Tab. 6.18 angegeben.

[123] Genauso wie die Subtests, die nach empirischer Schwierigkeit gebildet wurden, sind die Subtests nie in dieser Zusammenstellung eingesetzt, sondern nur statistisch konfektioniert worden.

Tabelle 6.18: Effektstärken für Geschlechterunterschiede in LSE 9-Subtests (Aufgabentypen; n = 464)

Testversion	Anzahl der Items	Mittlere Itemschwierigkeit	Effektstärke
technische	**20**	**–0,623**	**0,48**
rechnerische Modellierung	**21**	**0,206**	**0,72**
begriffliche Modellierung	17	0,563	0,65

Die Ergebnisse stehen im Einklang mit den Befunden in der Literatur, insbesondere mit denen von Stanat & Kunter (2001), und unterstützen die Hypothese H_04c. Der Unterschied zwischen *technischen Aufgaben* und *rechnerischen Modellierungsaufgaben* ist nicht nur signifikant, sondern auch bedeutsam. Da die Aufgabentypen und die empirischen Itemschwierigkeiten zwar theoretisch entkoppelt betrachtet werden können, in realen Tests aber die bereits betrachtete Tendenz aufweisen, können Interaktionseffekte zwischen diesen Aufgabenmerkmalen nicht ausgeschlossen werden. Die Itemmenge der *LSE 9* lässt keine vertiefende Analyse eines solchen Interaktionseffektes zu – hierzu bedürfte es größerer oder speziell für diesen Zweck konzipierter Itemmengen.

Eine heuristische Betrachtung deutet aber auf die substanzielle Bedeutung der Aufgabentypen hin: Bei den nach empirischer Schwierigkeit konfektionierten Subtests wies die mittlere Itemschwierigkeit eine Spannweite von ca. 2,7 und die Effektstärke einen Unterschied von 0,25 auf (Tab. 6.16). Der Unterschied in der mittleren Schwierigkeit von *technischen Aufgaben* und *rechnerischen Modellierungsaufgaben* beträgt hingegen nur ca. 0,83, der Effektstärkenunterschied aber 0,24. Hier scheinen die Aufgabentypen stärker zu wirken als die empirische Schwierigkeit. Zudem weist der Subtest *rechnerische Modellierungsaufgaben* von allen bisher betrachteten Subtests die größte Effektstärke auf.

Dieses Ergebnis wird auch durch eine *DIF*-Analyse (vgl. Kap. 4.2.2) für den *LSE 9*-Gesamttest unterstützt. Von den 58 Items des Gesamttests wiesen vier einen signifikanten *DIF*-Wert zugunsten der Mädchen und drei einen signifikanten *DIF*-Wert zugunsten der Jungen auf. Von den drei Items mit *DIF*-Werten zugunsten der Jungen waren zwei Items *rechnerische Modellierungsaufgaben* und ein Item eine *begriffliche Modellierungsaufgabe*. Von den vier Items mit *DIF*-Werten zugunsten der Mädchen waren drei *technische Aufgaben* und eins (erwartungswidrig) eine *rechnerische Modellierungsaufgabe*.

Anders als die Klassifikation der *LSE 9*-Items nach empirischer Schwierigkeit liefert die Klassifikation nach den *Typen mathematischen Arbeitens* nicht nur unterschiedliche differenzielle Befunde, sondern auch Subtests, die aus mathematikdidaktischer Sicht inhaltlich

beschreibbare Dimensionen darstellen. Zugleich liefert die entsprechende dreidimensionale Rasch-Skalierung des *LSE 9*-Mathematiktests bessere empirische Kennwerte für die Modellanpassung als das eindimensionale Modell. Der heuristisch durchgeführte *LRT* ergibt ein signifikantes Ergebnis (Prüfgröße: 55,3; Freiheitsgrade: 5; p-Wert: ca. 10^{-10}), das von den informationstheoretischen Maßen eindeutig unterstützt wird. Die mit *ConQuest* geschätzten latenten Korrelationen werden in Tab. 6.19 angegeben.

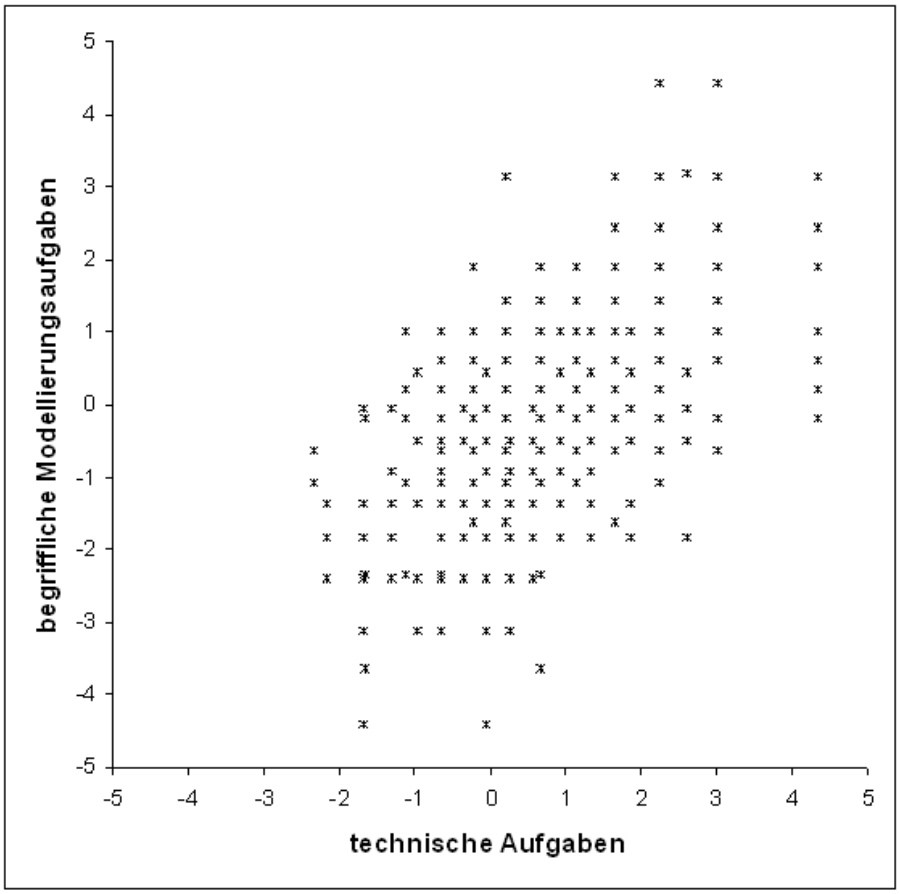
Tabelle 6.19: Im dreidimensionalen RM geschätzte latente Korrelationen zwischen den Typen mathematischen Arbeitens (n = 494)

	rechnerische Modellierungsaufgaben	***Begriffliche Modellierungsaufgaben***
technische Aufgaben	**0,93**	**0,84**
begriffliche Modellierungsaufgaben	**0,96**	

Auch wenn über die Interpretation der Werte für latente Korrelationen im Kontext von Dimensionalitätsanalysen keine Einigkeit besteht (vgl. Kap. 2.3.3), lassen die Werte in Tab. 6.19 bei Bedarf sicherlich zu, dass die Testleistungen im *LSE 9*-Mathematiktest im Sinne eines Skalierungspragmatismus eindimensional skaliert und berichtet werden können.[124] Auf der anderen Seite deutet die latente Korrelation zwischen *technischen Aufgaben* und *begrifflichen Modellierungsaufgaben* mit einem Wert von 0,84 darauf hin, dass der Test und die Daten (psychometrisch und mathematikdidaktisch) gut mit einer mehrdimensionalen Struktur beschrieben und interpretiert werden können. Dies verdeutlicht auch das Streudiagramm in Abb. 6.21, in dem für jede Versuchsperson die Kombination der Testleistung bei *technischen Aufgaben* mit der Testleistung bei *begrifflichen Modellierungsaufgaben* dargestellt wird. Die Punktwolke zeigt deutliche Unterschiede zwischen den beiden betrachteten Dimensionen: Für Personen mit einer mittleren Testleistung bei *technischen Aufgaben* ($\theta = 0$) streut die mögliche Testleistung bei *begrifflichen Modellierungsaufgaben* nahezu über das gesamte Leistungsspektrum.

[124] Zwar passt das dreidimensionale Modell besser zu den beobachteten Daten, die Verzerrungen, die durch eine eindimensionale Skalierung entstehen, dürften aber nicht erheblich sein.

Abbildung 6.21: Kontrastierung der Testleistungen im Subtest „technische Aufgaben“ mit den Testleistungen im Subtest „begriffliche Modellierungsaufgaben“ (n = 494)



6.4.3 Fähigkeitsselbstkonzept

Das *FSK:M* unterscheidet sich bezüglich der Testart und der Itemformate von den Leistungstests zur *Raumvorstellung* und zur *Mathematikleistung*. Die in Kap. 6.2.4. vorgestellten Items sollen selbstbezogene Kognitionen mithilfe eines Fragebogens erfassen. Es gibt also zunächst keine „richtigen“ oder „falschen“ Bearbeitungen, sondern nur fehlende oder unterschiedlich ausgeprägte Antworten. Des Weiteren werden die Items nicht dichotom (Statement trifft zu / trifft nicht zu), sondern abgestuft eingeschätzt. Dabei können die Antwortkategorien „1“ bis „4“ (für: Statement trifft überhaupt nicht / eher nicht / eher / völlig zu) genutzt werden. Wenn hieraus additiv ein Gesamtscore gebildet wird, so sind bei drei Items Werte zwischen 3 und 12 möglich.

Auch für das *FSK:M* wird zunächst eine Grundauswertung mit Verteilung der Stichprobe nach Gesamtscore sowie eine Rasch-Skalierung der drei Items vorgestellt, bevor im Sinne der Hypothese H_05 Geschlechterunterschiede betrachtet werden.

Verteilung der Stichprobe und Kennwerte der verwendeten Skala

Die Verteilung der Gesamtstichprobe auf die möglichen *FSK:M*-Gesamtscores ist zwar bimodal, jedoch nicht mit weit auseinander liegenden Gipfeln (vgl. Abb. 6.22). Insgesamt wird die Verteilung dadurch breit, aufgrund der nahezu symmetrischen Form aber nicht schief (vgl. Tab. 6.20). Bei der Auswertung wurden alle 440 Versuchspersonen berücksichtigt, die auf alle drei *FSK:M*-Items gültig geantwortet haben.

Abbildung 6.22: Verteilung der Versuchspersonen nach FSK:M-Gesamtscore

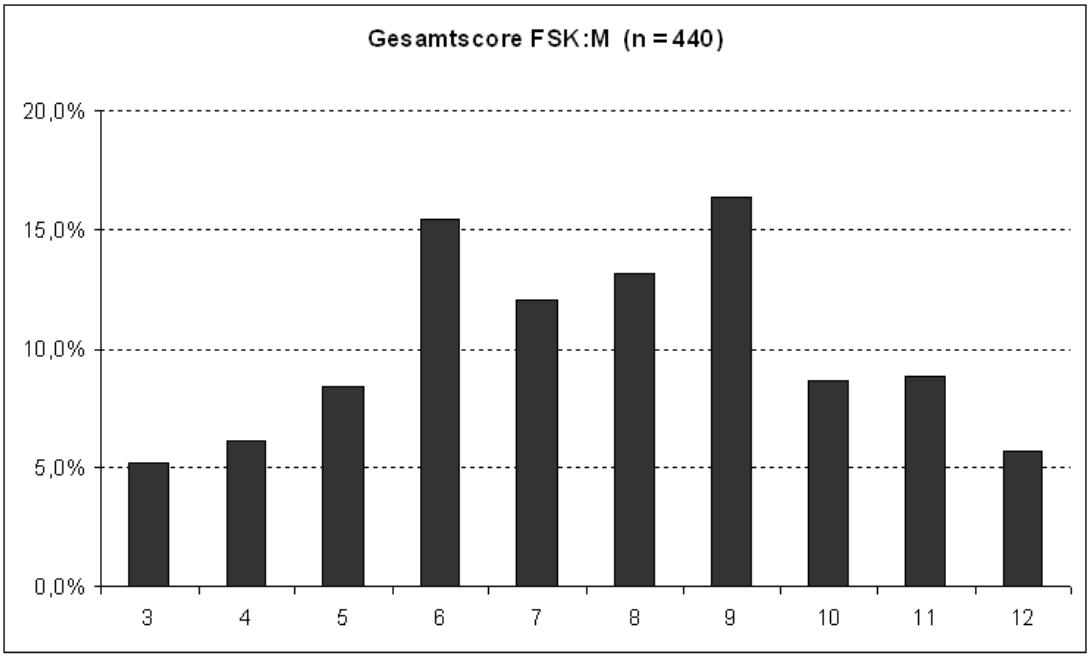


Tabelle 6.20: Kennwerte der Verteilung nach FSK:M-Gesamtscore (n = 440)

Skala	mögliche Scores	Median	arithm. Mittel	Schiefe	Exzess
FSK:M	3 bis 12	8	7,64	–0,06	–0,81

Die Gesamtscores und deren Mittelwerte in Tab. 6.20 können nicht direkt mit Bezug auf das Antwortformat interpretiert werden. Eine Division durch 3 (Anzahl der Items) bildet aus dem Gesamtscore das arithmetische Mittel der drei Antworten, das nun auf das Antwortformat bezogen werden kann. Forschungsmethodisch besteht aber das Problem, dass bei dieser Mittelwertbildung ein Intervallskalenniveau für die Antwortstufen vorausgesetzt wird, was inhaltlich kaum begründet werden kann. Sicher ist lediglich, dass die Antwortstufen geordnet sind. Somit können die drei Items in einem eindimensionalen ordinalen *RM* (mit vier Antwortkategorien) skaliert werden.

Für das *FSK:M* mit seinen gestuften Antwortkategorien, die für jedes Item die gleiche Bedeutung haben, gibt es neben dem ordinalen *RM* weitere plausible Testmodelle. Sie gehen durch Restriktionen aus dem ordinalen *RM* hervor und berücksichtigen dabei das konkrete Itemformat, vor allem die Bedeutung der Abstufungen: das *Dispersionsmodell*, das *Äquidistanzmodell* und das *Ratingskalen-Modell* (vgl. J. Rost, 2004, Kap. 3.3.2). Die Verteilung der Versuchspersonen auf die möglichen Gesamtscores kann bei den vorliegenden Daten gut durch eine geglättete Verteilung approximiert werden (*RMSEA* = 0,059)[125], sodass nur zwei Parameter für die Modellierung benötigt werden (vgl. Kap. 4.2.3).

Die Anzahl der zusätzlich benötigten Parameter hängt vom verwendeten Testmodell ab und wird in der folgenden Tabelle 6.21 zusammen mit den Kennwerten für die Güte der jeweiligen Modellanpassung angegeben:

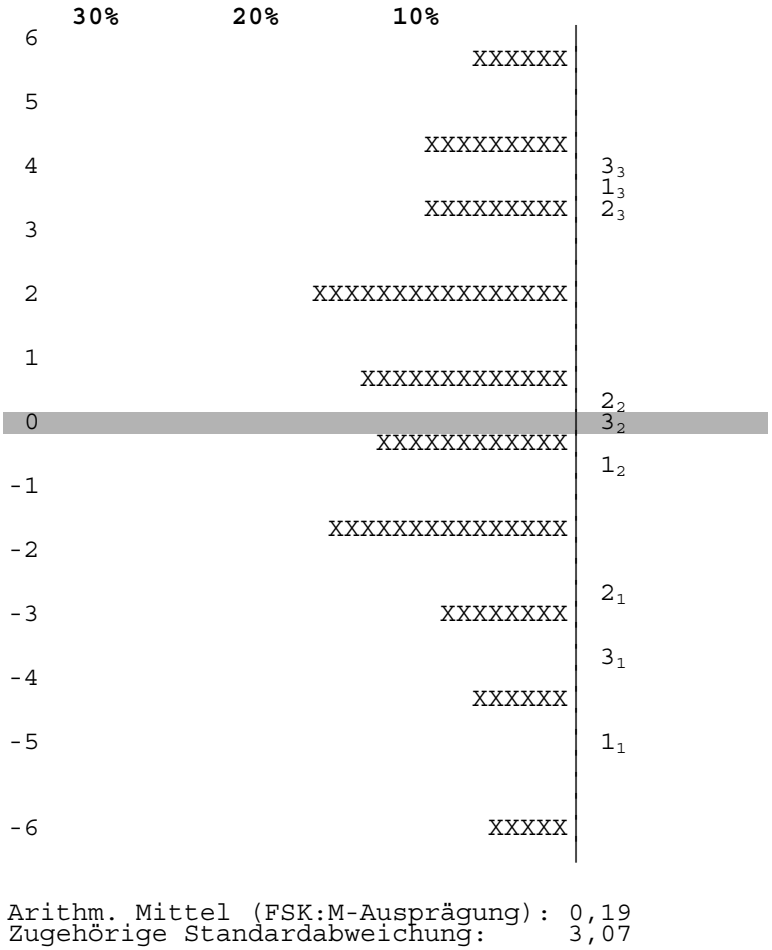
Tabelle 6.21: Kennwerte für die Güte der konkurrierenden ordinalen Modelle (n = 440)

Modell	Parameter	–2·ln(L)	*AIC*	*BIC*
Ordinales Rasch-Modell (mit geglätteter Verteilung)	2 + 8	2 695	2 715	2 756
Dispersionsmodell (mit geglätteter Verteilung)	2 + 6	2 706	2 722	2 755
***Äquidistanzmodell* (mit geglätteter Verteilung)**	**2 + 5**	**2 711**	**2 725**	**2 754**
Ratingskalen-Modell (mit geglätteter Verteilung)	2 + 4	2 750	2 762	2 787

Die *BIC*-Werte, die im vorliegenden Fall primär für den Modellvergleich verwendet werden sollten, sprechen knapp für das *Äquidistanzmodell*, wobei die Abstände zum *Dispersionsmodell* und zum ordinalen *RM* äußerst gering sind. Für das *Äquidistanzmodell* spricht darüber hinaus, dass es mit weniger Parametern auskommt und inhaltlich plausibel ist: Zusätzlich zu den Annahmen des ordinalen *RMs* geht das *Äquidistanzmodell* davon aus, dass bei jedem Item der Abstand zwischen den Antwortkategorien gleichgroß (auf der Logit-Skala) ist. Das *Äquidistanzmodell* bewährt sich zudem nicht nur im Vergleich zu den anderen Modellen, sondern auch bei einem Modelltest nach dem *Bootstrapping*-Verfahren (vgl. Kap. 4.2.3). Abbildung 6.23 zeigt die Ergebnisse der Skalierung als gemeinsame Verteilung von *FSK:M*-Ausprägungen und Itemschwellen (s. u.):

[125] Der „Root Mean Square Error of Approximation (RMSEA)“ ist ein Maß für den „Badness of Fit“, d. h. kleine Werte sprechen für eine gute Modellanpassung an die beobachteten Daten. Ein *RMSEA*-Wert von 0,059 wird allgemein als Beleg für einen guten Modell-Fit akzeptiert.

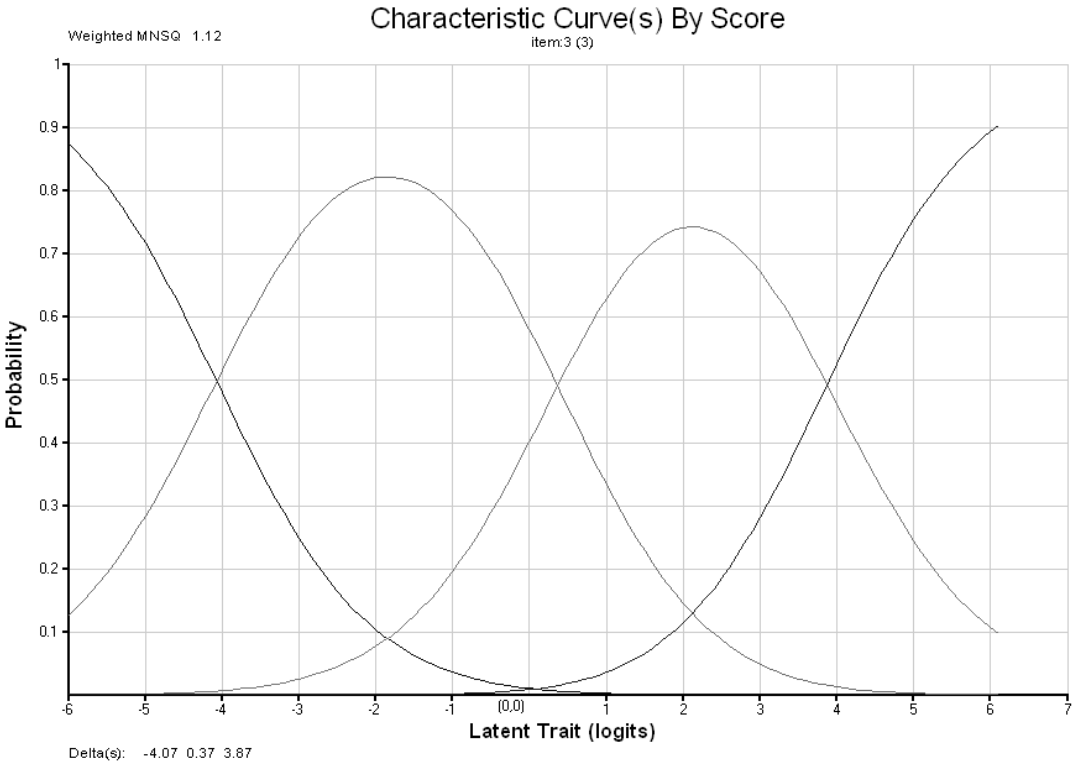
Abbildung 6.23: Verteilung der FSK:M-Ausprägung und der Itemschwellen (n = 440)



In der gemeinsamen Verteilung von *FSK:M*-Ausprägung und Itemschwellen repräsentiert ein X ca. 1 % der Stichprobe. „Itemschwellen“ lokalisieren die Übergänge von einer Antwortkategorie des vierstufigen Formats zur nächsten. Bei Item 1 ist es für Versuchspersonen mit *FSK:M*-Werten unterhalb des Schwellenwerts von 1_1 wahrscheinlicher, dass sie die Antwortkategorie „1“ wählen als eine andere. Liegt der *FSK:M*-Wert zwischen den Schwellenwerten von 1_1 und 1_2 so ist es am wahrscheinlichsten, dass die Antwortkategorie „2“ gewählt wird. Jenseits des Schwellenwerts von 1_3 ist es am wahrscheinlichsten, dass die Antwortkategorie „4“ gewählt wird.

Wenn man die theoretisch (aufgrund der geschätzten Parameter) erwarteten *ICCs* für die vier Antwortkategorien eines Items gemeinsam darstellt, so sind die Itemschwellen gerade die Schnittstellen der *ICCs* zweier benachbarter Antwortkategorien (vgl. Abb. 6.24). Bei guter Modellpassung gibt es zwischen zwei benachbarten Schwellen i_{m-1} und i_m jeweils einen hinreichend großen Bereich, in dem die Wahrscheinlichkeit für die Wahl der Antwortkategorie m am höchsten ist.

Abbildung 6.24: Theoretische ICC für das FSK:M-Item 3



Geschlechterunterschiede

Die aktuelle Forschungsliteratur deutet darauf hin, dass bei der Ausprägung des *FSK:M* zwar keine signifikanten Unterschiede zwischen den Schulformen, aber signifikante Unterschiede zugunsten der Jungen erwartet werden können. Geschlechterunterschiede können bei anderen *bereichspezifischen Fähigkeitsselbstkonzepten*, etwa für die sprachlichen Fächer, auch zugunsten der Mädchen beobachtet werden. Im Fach Mathematik können die Unterschiede zugunsten der Jungen sogar bei gleicher Fachleistung beobachtet werden (vgl. Moschner, 2001, S. 633).

Für das Konstrukt „bereichsspezifisches Fähigkeitsselbstkonzept“ sind, wie in Kap. 4.1.2 dargestellt wurde, soziale und dimensionale Vergleiche wichtig. Dabei führen die sozialen Vergleiche (innerhalb der Lerngruppe) zum „Big-Fish-Little-Pond-Effekt“, demzufolge gleiche objektiv feststellbare Fachleistungen zu ganz unterschiedlich ausgeprägten *bereichsspezifischen Fähigkeitsselbstkonzepten* führen können, je nach dem, wie leistungsstark die Gruppe ist, innerhalb derer sich ein Individuum vergleicht (vgl. Dickhäuser, 2006, S. 6; Lüdtke et al., 2002, S. 152). Bei gleicher objektiv gemessener *Mathematikleistung* sollte eine Schülerin in der Hauptschule in der Regel über ein erheblich besser ausgeprägtes *FSK:M* verfügen als eine Schülerin in einem Gymnasium – wie ein „großer Fisch in einem kleinen Teich“. Der doppelte Bezugsrahmen (innerhalb der Lerngruppe und zwi-

schen den Fächern) müsste dazu führen, dass das *FSK:M* in allen Schulformen im Mittel gleich ausgeprägt ist.

Eine zweifaktorielle *ANOVA* (*Geschlecht* x *Schulform*) zeigt, dass in der Stichprobe der Hauptuntersuchung kein signifikanter Interaktionseffekt zwischen *Geschlecht* und *Schulform* vorliegt. Durchaus überraschend ist, dass *beide* Haupteffekte signifikant sind, obwohl das Konstrukt und die Forschungsliteratur dies für den Faktor *Schulform* nicht erwarten ließen. Der signifikante Geschlechterunterschied zugunsten der Jungen wird im Rahmen einer latenten Modellierung mit einer Effektstärke von 0,63 geschätzt. Dieser Befund unterstützt die Hypothese H_05 ohne Einschränkung.

6.4.4 Zusammenhang der Konstrukte

Die bisher durchgeführten Auswertungen der Hauptuntersuchung zeigen, dass mit den verwendeten Instrumenten im Wesentlichen erwartungskonforme Ergebnisse im Sinne der Hypothesen H_01 bis H_05 erzielt werden konnten. Lediglich beim Raumvorstellungstest *DAT:SR* gab es in der Hauptuntersuchung – anders als in der Voruntersuchung – keine signifikanten Geschlechterunterschiede (zugunsten der weiblichen Versuchspersonen); dieses Ergebnis befindet sich aber im Einklang mit der Forschungsliteratur, die keine (bzw. keine eindeutigen) signifikanten Geschlechterunterschiede für die Raumvorstellungskomponente *räumliche Visualisierung* ausweist. Insgesamt scheinen die einzelnen Instrumente die jeweiligen Konstrukte in der Stichprobe der Hauptuntersuchung zuverlässig, valide und ohne gravierende Unter- oder Überschätzungen der Varianz zu erfassen, sodass auch die Zusammenhänge zwischen den Konstrukten mit diesen Instrumenten untersucht werden können.

Im Folgenden wird der Zusammenhang von *Raumvorstellung* und *Mathematikleistung* zunächst ohne Einbeziehung weitere Konstrukte, aber nach Komponenten differenziert untersucht (Hypothesen H_06a-d), wobei – nach den bisherigen Auswertungen – vor allem die *mentale Rotation* und die *rechnerischen Modellierungsaufgaben* von besonderem Interesse sind. Danach wird im Sinne der *Spatial Mediation Hypothesis* der Zusammenhang von *Raumvorstellung* und *Mathematikleistung* unter der Perspektive der Geschlechterunterschiede analysiert (Hypothesen H_07a-e). Schließlich wird mit Blick auf ein erweitertes Modell zur Erklärung von Geschlechterunterschieden in der *Mathematikleistung* das *FSK:M* in die Auswertungen einbezogen (Hypothesen H_08).

Raumvorstellung und Mathematikleistung

Die drei Komponenten der *Raumvorstellung* (*räumliche Wahrnehmung*, *mentale Rotation* und *räumliche Visualisierung*) konnten in Kap. 6.4.1 mit den verwendeten Referenztests *WLT*, *MRT* und *DAT:SR* empirisch voneinander getrennt werden. Dies war zum einen das Ergebnis des Vergleichs einer eindimensionalen mit einer dreidimensionalen Modellierung anhand der Kennwerte aus Tab. 6.11 (S. 216). Zum anderen folgt dies aus den geschätzten

Werten für die latenten Korrelationen zwischen den Komponenten (vgl. Tab. 6.12, S. 216). Der Zusammenhang zwischen *Raumvorstellung* und *Mathematikleistung* wird daher direkt nach den Raumvorstellungskomponenten differenziert betrachtet. Aufgrund der Skalierungsergebnisse für den *LSE 9*-Mathematiktest ist es sinnvoll, *Mathematikleistung* sowohl global (also eindimensional) als auch – aufgrund der differenziellen Befunde zu unterschiedlichen Subtests – nach Itemgruppen differenziert zu betrachten.

Der Zusammenhang zwischen den Raumvorstellungskomponenten und der *Mathematikleistung* wurde jeweils im Rahmen einer zweidimensionalen Skalierung nach dem *RM* als latente Korrelation (mit *ConQuest*) geschätzt. Die zweidimensionale Modellierung ist für die Fragestellung zwar naheliegend, zumindest aus theoretischer Perspektive aber nicht die einzig denkbare: Da *Raumvorstellung* ein klassischer (aktuell eher impliziter) Gegenstand der Schulmathematik ist,[126] wäre auch eine eindimensionale Modellierung denkbar, die *Raumvorstellung* als Teil eines Mathematiktests auffasst. Empirische Dimensionalitätsanalysen auf der Basis der Kennwerte für die jeweilige Modellanpassung zeigen aber eindeutig, dass für jede Komponente der *Raumvorstellung* ein zweidimensionales Modell erheblich besser zu den beobachteten Daten passt als ein eindimensionales. Die im jeweiligen zweidimensionalen Modell geschätzten latenten Korrelationen werden in Tab. 6.22 angegeben.

Tabelle 6.22: Latente Korrelationen zwischen dem LSE 9-Mathematiktest und den Raumvorstellungstests WLT, MRT und DAT:SR (n = 464)

	WLT	***MRT***	***DAT:SR***
LSE 9	**0,52**	**0,63**	**0,68**

Für alle drei Komponenten der *Raumvorstellung* besteht also ein substanzieller Zusammenhang mit *Mathematikleistung*, wobei die Zusammenhänge für die *MRT*-Leistung und die *DAT:SR*-Leistung enger sind als für die *WLT*-Leistung. Diese Ergebnisse unterstützen die Hypothesen H_06a und H_06b ohne Einschränkung. Das statistisch Gemeinsame der Raumvorstellungskomponenten mit *Mathematikleistung* ist damit ähnlich groß wie das statistisch Gemeinsame der Raumvorstellungskomponenten untereinander, das mit latenten Korrelationen zwischen 0,55 (*WLT* x *MRT*) und 0,74 (*MRT* x *DAT:SR*) geschätzt wurde.

Mit Blick auf die angestrebte statistische Erklärung von Geschlechterunterschieden in der *Mathematikleistung* durch die *Raumvorstellung* ist nach den vorliegenden Befunden aus

[126] Auf diesen Gegenstand wird allerdings aufgrund der aktuellen curricularen Entwicklungen im Bereich der Geometrie bzw. der Leitidee „Raum und Form“ immer weniger Unterrichtszeit verwendet. Tendenziell wird auch in der Geometrie immer mehr gerechnet.

den Kapiteln 2 und 3 sowie den eigenen Auswertungen die *mentale Rotation* die zentrale Raumvorstellungskomponente. Daher werden vor allem für den *MRT* vertiefende Analysen im Sinne der Hypothesen H_06c und H_06d (Zusammenhang von *MRT*-Leistung mit Itemschwierigkeit bzw. Aufgabentypen im *LSE 9*-Mathematiktest) durchgeführt. Soweit dies ohne zu großen Zusatzaufwand möglich ist, werden analoge Befunde für den *WLT* und den *DAT:SR* ebenfalls geniert und kontrastierend berichtet.

Die Ergebnisse in Tab. 6.23 zeigen, dass der Zusammenhang zwischen *mentaler Rotation* und *Mathematikleistung* beim *LSE 9*-Mathematiktest nicht von der Schwierigkeit der *LSE 9*-Items abhängt. So werden die latenten Korrelationen für die (rechnerisch konfektionierte) empirisch leichte und für die empirisch schwierige Testhälfte gleich geschätzt; die Hypothese H_06c ist vor diesem Hintergrund nicht haltbar. Für den *WLT* und den *DAT:SR* ergeben sich zwar geringfügige Unterschiede, diese sind aber nicht signifikant.

Tabelle 6.23: Latente Korrelationen zwischen LSE 9-Subtests (empirische Schwierigkeit) und den Raumvorstellungstests WLT, MRT und DAT:SR (n = 464)

	WLT	***MRT***	***DAT:SR***
***LSE 9* – leicht**	**0,50**	**0,62**	**0,67**
***LSE 9* – schwierig**	**0,55**	**0,62**	**0,71**

Bereits bei der Frage der Geschlechterunterschiede in der *Mathematikleistung* hat sich angedeutet, dass eine Unterscheidung von Subtests nach *Typen mathematischen Arbeitens* vermutlich ergiebiger ist als eine Unterscheidung nach empirischer Schwierigkeit – wobei mögliche Interaktionseffekte zwischen beiden Klassifikationen nicht mit dem verwendeten *LSE 9*-Mathematiktest untersucht werden können. Diese Vermutung bestätigt sich für den Zusammenhang zwischen der *MRT*-Leistung und der *LSE 9*-Leistung (vgl. Tab. 6.24).

Tabelle 6.24: Latente Korrelationen zwischen LSE 9-Subtests (Aufgabentypen) und den Raumvorstellungstests WLT, MRT und DAT:SR (n = 464)

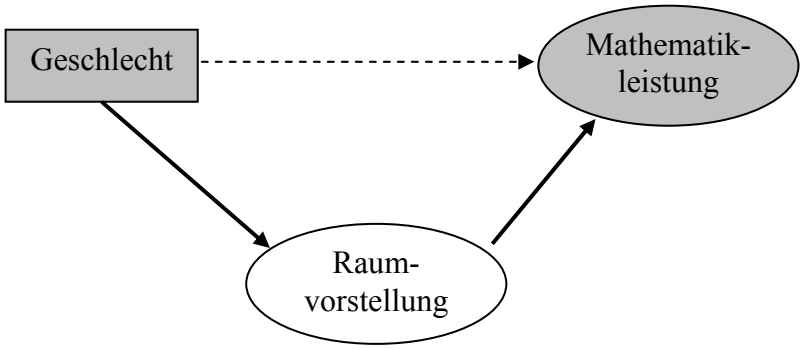
	WLT	***MRT***	***DAT:SR***
technische	**0,53**	**0,56**	**0,66**
rechnerische Modellierung	**0,55**	**0,66**	**0,68**
begriffliche Modellierung	0,46	0,59	0,66

Der Zusammenhang zwischen *MRT*-Leistung und *Mathematikleistung* ist für *rechnerische Modellierungsaufgaben* enger als für *technische Aufgaben*. Der Unterschied zwischen den beiden Werten ist bei der realisierten Stichprobengröße signifikant, wodurch die Hypothese H_06c unterstützt wird. Für den *WLT* und den *DAT:SR* ergeben sich marginale Unterschiede, die nicht signifikant sind. Die Ergebnisse zu Geschlechterunterschieden in der *MRT*-Leistung, zu Geschlechterunterschieden bei den verschiedenen *Typen mathematischen Arbeitens* und der obige Zusammenhang deuten darauf hin, dass für die *Spatial Mediation Hypothesis* vor allem der *MRT* und die *rechnerischen Modellierungsaufgaben* interessant sein dürften.

Spatial Mediation Hypothesis

In der *Spatial Mediation Hypothesis* fassten Burnett et al. (1979) die damals vorliegenden Befunde zusammen, die darauf hindeuteten, dass Geschlechterunterschiede in der *Mathematikleistung* zumindest in einem relevanten Anteil durch Geschlechterunterschiede in der *Raumvorstellung* erklärt werden können. Klieme (1986) hat zur Ausdifferenzierung dieser Vermutung beigetragen, indem er für verschiedene Komponenten der *Mathematikleistung* und der *Raumvorstellung* jeweils untersucht hat, welche Anteile der geschlechterbedingten Varianz der *Mathematikleistung* durch *Raumvorstellung* vermittelt werden. Schematisch wurde die *Spatial Mediation Hypothesis* in der vorliegenden Arbeit in Abb. 3.9 (S. 107) dargestellt, die hier noch einmal wiedergegeben wird (Abb. 6.25).

Abbildung 6.25: „Spatial Mediation Hypothesis“



Damit eine Variable in einem empirischen Zusammenhang tatsächlich eine *Mediatorvariable* ist, also einen *Mediatoreffekt* ausübt, müssen die folgenden Bedingungen erfüllt sein (vgl. Baron & Kenny, 1986, S. 1176):

- Der *Prädiktor* (in Abb. 6.25 *Geschlecht*) muss einen statistisch nachweisbaren Effekt auf das *Kriterium* (*Mathematikleistung*) haben, manifestiert in einem signifikant von Null verschiedenen Regressionskoeffizienten.

- Der *Mediator* (*Raumvorstellung*) muss ebenfalls einen im obigen Sinne statistisch nachweisbaren Effekt auf das *Kriterium* (*Mathematikleistung*) haben.
- Der Effekt des *Prädiktors* auf das *Kriterium* muss sich statistisch nachweisbar verringern, wenn der Effekt des *Mediators* berücksichtigt wird.

Die genannten Bedingungen implizieren, dass es einen statistisch nachweisbaren Effekt des *Prädiktors* auf den *Mediator* gibt. Wenn der Effekt des *Prädiktors* auf das *Kriterium* durch die Aufnahme des *Mediators* in ein gemeinsames Modell vollständig verschwindet, der entsprechende Regressionskoeffizient also praktisch nicht mehr von Null unterscheidbar ist, dann handelt es sich um eine *vollständige Mediation*. Wenn der Effekt sich statistisch nachweisbar verringert, aber nicht verschwindet, dann handelt es sich um eine *partielle Mediation*; in diesem Fall gibt es noch einen eigenständigen Effekt des *Prädiktors* auf das *Kriterium* oder weitere, noch nicht berücksichtigte, *partielle Mediationen*.

Bei einem *Mediationsmodell* wie der *Spatial Mediation Hypothesis* werden immer Wirkungsrichtungen in der oben beschriebenen Form angenommen. Dass dies für den Zusammenhang zwischen *Raumvorstellung* und *Mathematikleistung* eine durchaus starke, bislang inhaltlich nicht vollständig legitimierte Annahme ist, wurde bereits in Kap. 3.3.6 diskutiert. Dennoch soll im Anschluss an die vorliegende Forschungsliteratur zunächst mit diesen postulierten Wirkungsrichtungen gearbeitet werden. Methodisch werden *Mediatoreffekte* im Rahmen von Pfadanalysen bzw. Strukturgleichungsmodellen analysiert. Dabei stellt Abb. 6.25 bereits ein entsprechendes Pfaddiagramm bzw. Strukturmodell dar. Auf der Basis der Daten der Hauptuntersuchung kann ein Strukturgleichungsmodell realisiert werden, bei dem das obige Strukturmodell durch Messmodelle (vgl. 4.2.2) für (Komponenten der) *Raumvorstellung* und *Mathematikleistung* ergänzt wird.

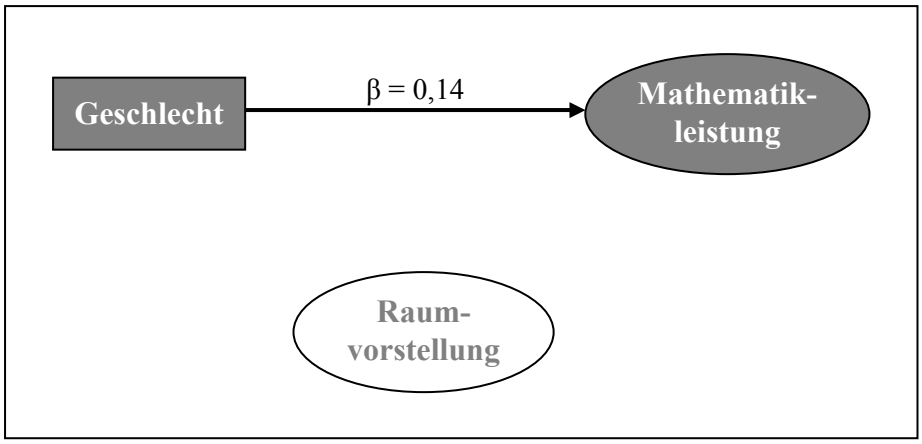
Die *Spatial Mediation Hypothesis* wird im Folgenden auf der Grundlage der Hypothesen H_07a-e differenziert analysiert, was insbesondere bedeutet, dass die Auswertungen nach den Komponenten der *Raumvorstellung* differenziert durchgeführt werden. *Mathematikleistung* wird zunächst wieder global (eindimensional) und anschließend differenziert nach *Typen mathematischen Arbeitens* bzw. nach empirischer Schwierigkeit der Items in die Untersuchung einbezogen. Auf diesem Weg kann im Rahmen der in der Hauptuntersuchung betrachteten Konstrukte ein theoretisch und empirisch fundiertes Modell zur Erklärung von Geschlechterunterschieden in der *Mathematikleistung* entwickelt werden, auf dessen Basis aktuelle Befunde generiert werden können. Im Sinne der obigen Definition von *Mediatorvariablen* werden zunächst die notwendigen Voraussetzungen für das Vorliegen von *Mediatoreffekten* überprüft.

Die folgenden Auswertungen wurden durchgängig auf der Basis entsprechender Strukturgleichungsmodelle mit dem Programmpaket *AMOS* realisiert. Aufgrund des durch das *LSE 9*-Testdesign bedingten höheren Anteils fehlender Werte bei den beobachteten Indikatoren für die *Mathematikleistung* (*LSE 9*-Items) hat *AMOS* keine „Fit-Statistiken“ (Kenn-

werte für die Güte der Anpassung des Modells an die beobachteten Daten) berechnet. Im Rahmen einer heuristischen Betrachtung wurden daher parallel zu den latenten Modellen die analogen Modelle mit manifesten Variablen[127] berechnet. Die Fit-Statistiken für diese manifesten Modelle sprechen alle für eine gute Anpassung der realisierten Modelle an die beobachteten Daten.

Notwendig für einen *Mediatoreffekt* der *Raumvorstellung* für Geschlechterunterschiede in der *Mathematikleistung* ist zunächst, dass solche Geschlechterunterschiede überhaupt statistisch nachweisbar sind und dass auch für die *Raumvorstellung* selbst ein statistisch nachweisbarer Zusammenhang mit *Mathematikleistung* besteht. Die Ergebnisse von Kap. 6.4.1 und Kap. 6.4.2 haben mit Varianzanalyse, latenten Regressionen und latenten Korrelationen diese Voraussetzungen bereits bestätigt. An dieser Stelle werden die Befunde aus der Sicht der Strukturgleichungsmodelle repliziert und entsprechend berichtet, d. h. es werden standardisierte Pfadkoeffizienten und der Anteil der jeweils erklärten Varianz in der Zielvariablen *Mathematikleistung* dargestellt. In Kap. 2.3.1 (S. 43) wurde bereits erwähnt, wie ein standardisierter Pfadkoeffizient β interpretiert werden kann: Wenn der Wert der unabhängigen Variablen um eine Standardabweichung (der unabhängigen Variablen) zunimmt, dann steigt der Wert der abhängigen Variablen um das β-fache einer Standardabweichung (der abhängigen Variablen). Abbildung 6.26 zeigt das Ergebnis der Regression der *LSE 9-Mathematikleistung* auf die Variable *Geschlecht*.

Abbildung 6.26: Statistischer Effekt des Prädiktors „Geschlecht“ auf das Kriterium „Mathematikleistung“ (n = 494)

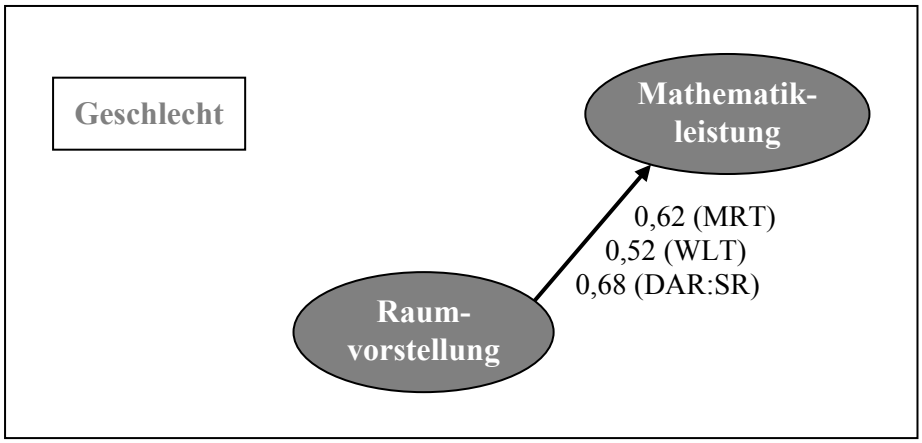


[127] Dabei wurden die latenten Variablen mit ihren zugehörigen Messmodellen durch manifeste Variablen ersetzt, deren Werte im Rahmen der Skalierungen der einzelnen Konstrukte gewonnen wurden: Beispielsweise wurde die latente Variable *Mathematikleistung* mit ihrem Messmodell, in dem das Antwortverhalten zu den *LSE 9*-Items auf sie zurückgeführt wird, durch die manifeste Variable „LSE_WLE“ ersetzt, die die mit *ConQuest* ermittelte *WLE*-Schätzung der Leistung im *LSE 9*-Mathematiktest enthält.

Der standardisierte Pfadkoeffizient, der den Effekt der Variablen *Geschlecht* auf die *Mathematikleistung* angibt, beträgt 0,14 und ist signifikant von Null verschieden. Die Varianz der *Mathematikleistung* wird durch die Variable *Geschlecht* zu ca. 2 % ($\approx 0{,}14^2$) erklärt. Generell gibt bei einer einfachen linearen Regression das Quadrat der standardisierten Pfadkoeffizienten den Anteil der durch die Regression erklärten Varianz an.

Die analoge Überprüfung der zweiten Voraussetzung für einen *Mediatoreffekt* der *Raumvorstellung* sollte aufgrund der bisherigen empirischen Befunde (u. a. in Kap. 6.4.1) nach den Komponenten der *Raumvorstellung* ausdifferenziert durchgeführt werden. Abbildung 6.27 zeigt das Ergebnis der entsprechenden Regressionen.

Abbildung 6.27: Statistischer Effekt der Raumvorstellungskomponenten auf die Mathematikleistung (n = 464)

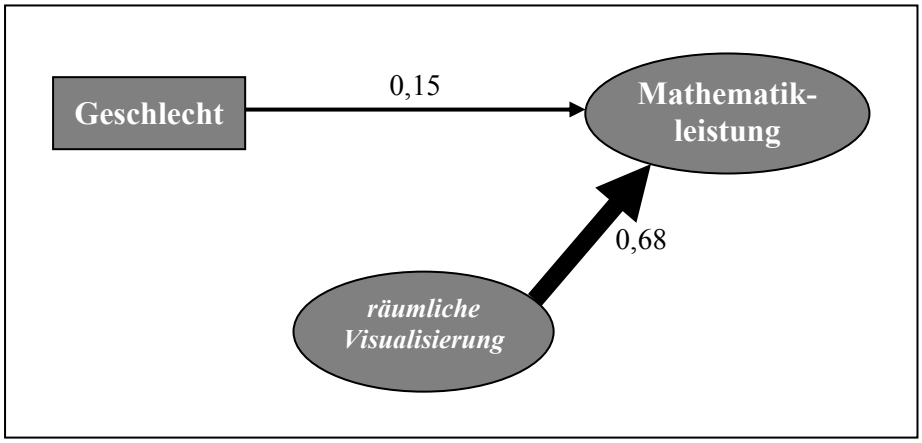


In Abb. 6.27 sind die drei standardisierten Pfadkoeffizienten für die drei in der Hauptuntersuchung verwendeten Untertests angegeben. Sie stimmen nahezu exakt mit den entsprechenden latenten Korrelationen zum Zusammenhang der Raumvorstellungskomponenten mit *Mathematikleistung* überein, die auf der Basis zweidimensionaler Rasch-Skalierungen mit *ConQuest* geschätzt und in Tab. 6.22 (S. 238) berichtet wurden. Die Quadrate der standardisierten Pfadkoeffizienten geben wieder den Anteil der durch den jeweiligen Raumvorstellungstest erklärten Varianz der *Mathematikleistung* an (*MRT* 39 %; *WLT* 27 %; *DAT:SR* 46 %). Zur Erinnerung: Der relativ hohe Anteil, der durch den *DAT:SR* statistisch erklärt wird, ist inhaltlich darauf zurückzuführen, dass *räumliche Visualisierung* in wesentlichen Anteilen auf analytische Denkprozesse zurückgreift – wie die *Mathematikleistung* auch. Nachdem beide notwendigen Voraussetzungen für *Mediatoreffekte* nun auch methodisch mit Strukturgleichungsmodellen realisiert wurden, kann die *Spatial Mediation Hypothesis* für die drei Komponenten der *Raumvorstellung* analog untersucht werden.

Für den Referenztest zur *räumlichen Visualisierung*, den *DAT:SR*, wurden in der Hauptuntersuchung keine signifikanten Geschlechterunterschiede festgestellt. Dementsprechend

wurde in Hypothese H_07a kein *Mediatoreffekt* des *DAT:SR* für Geschlechterunterschiede in der *Mathematikleistung* angenommen. Abbildung 6.28 zeigt, dass diese Hypothese durch die empirische Analyse bestätigt wird.

Abbildung 6.28: Mediationsdiagramm für räumliche Visualisierung (n = 464)



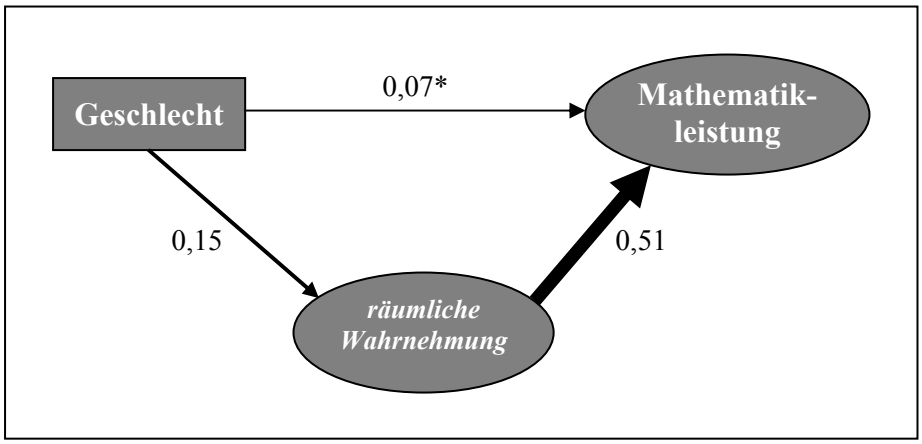
In Abb. 6.28 wird – wie in den folgenden Abbildungen auch – die Stärke der einzelnen Effekte durch die zugehörige Pfeile visualisiert. Die Linienstärke ist jeweils proportional zum standardisierten Pfadkoeffizienten. Dementsprechend werden Pfade, deren zugehörigen Koeffizienten praktisch nicht von Null verschieden sind, nicht gezeichnet. Im Vergleich zum Modell aus Abb. 6.26 hat sich der Pfadkoeffizient, der von der Variablen *Geschlecht* ausgeht, geringfügig verändert. Dieses Phänomen kann immer dann auftreten, wenn ein Modell durch weitere Variablen oder Pfade ergänzt wird. Insgesamt wird durch das Modell in Abb. 6.28, also durch die beiden Prädiktoren *Geschlecht* und *räumliche Visualisierung*, 49 % der Varianz der *Mathematikleistung* statistisch erklärt.[128] Bei mehr als einem Prädiktor entspricht dieser Anteil im Allgemeinen nicht der Summe der Quadrate der standardisierten Pfadkoeffizienten, sondern ist zumeist kleiner, da es häufig gemeinsame Varianzanteile beim Kriterium gibt, die durch mehrere Prädiktoren (jeweils einzeln) erklärt werden könnten. Dies wird später bei Modellen, die noch mehr Prädiktoren enthalten, deutlich sichtbar.

Bei der *räumlichen Wahrnehmung* wurden in Kap. 6.4.1 mittlere Geschlechterunterschiede und in Kap. 6.4.4 ein mittlerer Zusammenhang mit *Mathematikleistung* festgestellt, der kleiner war als die Zusammenhänge mit den anderen Komponenten der *Raumvorstellung*. Abbildung 6.29 zeigt passend zur Annahme von Hypothese H_07b, dass die Geschlechter-

[128] Dieses Bestimmtheitsmaß R^2 wird in der vorliegenden Arbeit durchgängig als korrigiertes R^2 angegeben, bei dem berücksichtigt wird, wie viele Regressoren die Varianz erklären (vgl. Backhaus et al., 2008, S. 71).

unterschiede in der *Mathematikleistung* dementsprechend in mittlerem Umfang durch *räumliche Wahrnehmung* (*WLT*) vermittelt werden.

Abbildung 6.29: Mediationsdiagramm für räumliche Wahrnehmung (n = 464)



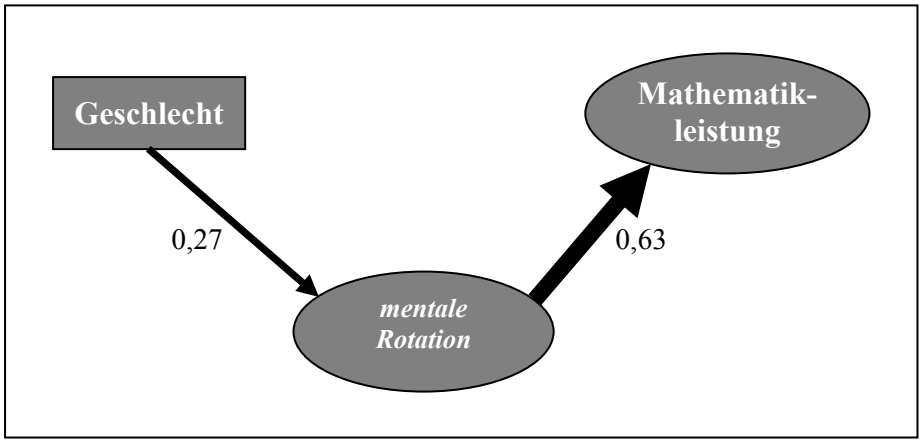
Der Effekt der Variablen *Geschlecht* auf die *Mathematikleistung* verringert sich von 0,14 (vgl. Abb. 6.26) auf 0,07 und ist auf dem 5 %-Signifikanzniveau nicht mehr von Null unterscheidbar (und daher mit einem „*“ gekennzeichnet). Im Diagramm ist er trotzdem aufgenommen worden, da er rechnerisch zur erklärten Varianz und zum „totalen Effekt“ der Variablen *Geschlecht* beiträgt.[129] Im Fall der *räumlichen Wahrnehmung* scheint also eine *partielle Mediation* vorzuliegen. Der totale Effekt berechnet sich als Summe von direkten und indirekten, also vermittelten („mediierten“) Effekten: $0{,}07 + 0{,}15 \cdot 0{,}51 \approx 0{,}14$.[130] Die erklärte Varianz der *Mathematikleistung* beträgt für das Modell in Abb. 6.29 ca. 27 %, ist also deutlich geringer als beim analogen Modell für *räumliche Visualisierung*. Dies ist auf den geringeren Zusammenhang von *räumlicher Wahrnehmung* und *Mathematikleistung* zurückzuführen.

Für die *mentale Rotation* war der Zusammenhang mit *Mathematikleistung* zwar ebenfalls geringer als für *räumliche Visualisierung*, dafür war der *MRT* aber der Raumvorstellungstest, der zu den größten Geschlechterunterschieden geführt hat. Dementsprechend ist mit Hypothese H_07c ein starker *Mediatoreffekt* angenommen worden, was durch die Ergebnisse, die in Abb. 6.30 dargestellt werden, bestätigt wird.

[129] Bei einem Signifikanztest mit der Nullhypothese „$\beta = 0$“ beträgt der entsprechende p-Wert 0,15. Unter Berücksichtigung eines möglichen Fehlers 2. Art (vgl. Bücher & Henn, 2007, Kap. 4.2) kann es sinnvoll sein, diesen Pfad inhaltlich zu berücksichtigen.

[130] Auf der Basis der angegebenen Werte müsste das gerundete Ergebnis 0,15 lauten. Beim gerundeten Wert 0,14 wurde berücksichtigt, dass die angegebenen Werte bereits Rundungsfehler enthalten.

Abbildung 6.30: Mediationsdiagramm für mentale Rotation (n = 464)



Die erklärte Varianz beträgt im obigen Modell ca. 39 %, wobei der Pfad von der Variablen *Geschlecht* auf die *Mathematikleistung* praktisch verschwindet ($\beta = -0{,}03$; $p = 0{,}50$), sodass der totale Effekt von der Variablen *Geschlecht* auf die *Mathematikleistung* (0,14) praktisch vollständig über die *mentale Rotation* vermittelt wird. Der *MRT* scheint also eine vollständige Mediation von Geschlechterunterschieden in der *Mathematikleistung* im Sinne der *Spatial Mediation Hypothesis* zu leisten – zumindest wenn keine weiteren Variablen im Strukturgleichungsmodell aufgenommen werden. Dies deutet darauf hin, dass die Komponente *mentale Rotation* im Bereich der *Raumvorstellung* das größte Potenzial hat, zur inhaltlichen Erklärung von Geschlechterunterschieden in der *Mathematikleistung* beizutragen. Daher werden für die *mentale Rotation* vertiefende Analysen im Sinne der Hypothesen H_07d-e durchgeführt, bei denen die *Mathematikleistung* differenziert betrachtet wird.

Bei der differenzierten Betrachtung von *Mathematikleistung* wurden bisher zwei alternative Klassifikationen verwendet. Die Unterscheidung der *LSE 9*-Items nach empirischer Schwierigkeit hat dabei zwar im Bereich der Geschlechterunterschiede in der *Mathematikleistung* eine gewisse Nützlichkeit (vgl. Tab. 6.16, S. 226), nicht aber bei der Betrachtung des Zusammenhangs von *mentaler Rotation* und *Mathematikleistung* (vgl. Tab. 6.23, S. 239). Bei beiden Fragestellungen war die Unterscheidung der Items nach *Typen mathematischen Arbeitens* ergiebiger, vor allem zwischen *technischen Aufgaben* und *rechnerischen Modellierungsaufgaben*. Dieses Potenzial bestätigt sich auch mit Blick auf unterschiedlich starke Zusammenhänge im Rahmen der *Spatial Mediation Hypothesis*.

Wird anstelle der globalen *LSE 9*-Testleistung im Strukturgleichungsmodell von Abb. 6.30 die Leistung in der empirisch leichten oder der empirisch schwierigen Testhälfte der *LSE 9* eingesetzt, so ändern sich die standardisierten Pfadkoeffizienten und die erklärte Varianz – entgegen einiger Befunde in der Forschungsliteratur – praktisch nicht; die Hypothese H_07d muss also verworfen werden.

Anders stellt sich die Situation bezüglich der Unterscheidung zwischen *technischen Aufgaben* und *rechnerischen Modellierungsaufgaben* dar. Die beiden zugehörigen Modelle sind mit den standardisierten Pfadkoeffizienten in den Abbildungen 6.31 und 6.32 dargestellt.

Abbildung 6.31: Mediationsdiagramm für mentale Rotation und technische Aufgaben (n = 464)

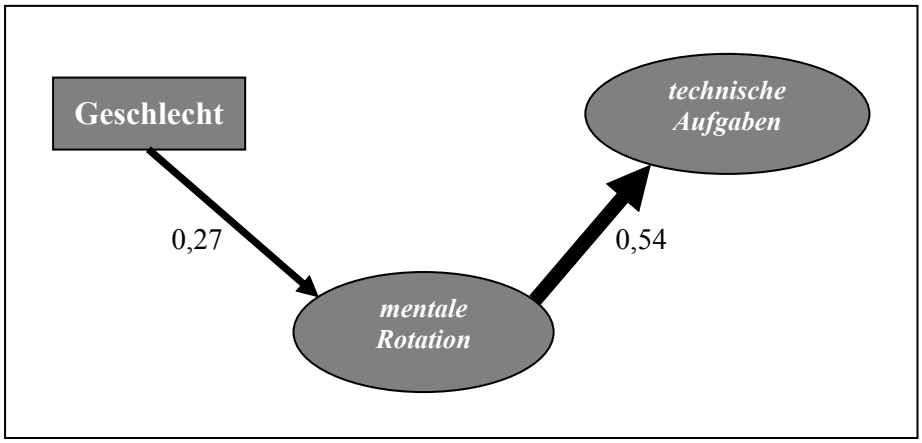
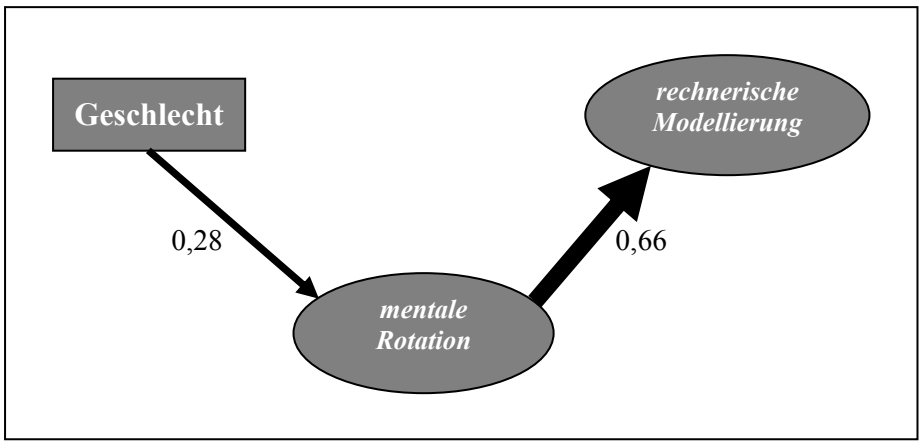


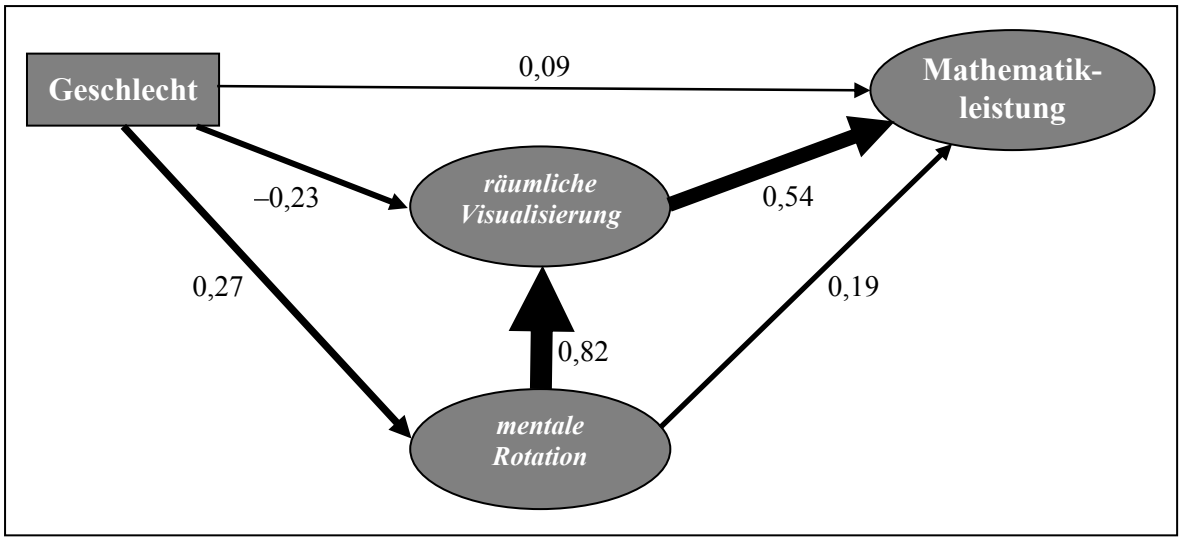
Abbildung 6.32: Mediationsdiagramm für mentale Rotation und rechnerische Modellierungsaufgaben (n = 464)



Die Ergebnisse unterstützen die Hypothese H_07e ohne Einschränkung: Der unterschiedlich starke Zusammenhang zwischen *mentaler Rotation* und den Items des jeweiligen Aufgabentyps führt dazu, dass für *technische Aufgaben* nur 28 % der Leistungsvarianz durch das angegebene Modell erklärt werden kann, während dies für *rechnerische Modellierungsaufgaben* immerhin 43 % sind. Als Zwischenfazit zur differenzierten Untersuchung der *Spatial Mediation Hypothesis* kann man also festhalten, dass die Kombination der Komponenten *mentale Rotation* und *rechnerische Modellierungsaufgaben* das größte Potenzial zur Erklärung von Geschlechterunterschieden in der *Mathematikleistung* hat.

Trotzdem scheint auch der Blick auf die anderen Modelle noch einmal lohnenswert: Die Abbildungen 6.28 bis 6.30 zeigen drei unterschiedliche Konstellationen für die drei Raumvorstellungskomponenten. Dabei hat die *mentale Rotation* zwar das größte Potenzial zur Erklärung von Geschlechterunterschieden in der *Mathematikleistung*, bedeutsam ist aber auch, dass die *räumliche Visualisierung* insgesamt am meisten Leistungsvarianz in der *Mathematikleistung* – unabhängig von der Frage nach Geschlechterunterschieden – erklären kann. Dies ist einerseits inhaltlich aufgrund der analytischen Denkprozesse, die jeweils erforderlich sind, plausibel und gibt andererseits Anlass dazu, auch räumliche Visualisierung in ein gemeinsames Mediationsmodell aufzunehmen. Damit wird auch berücksichtigt, dass *räumliche Visualisierung* in zwar untergeordneten, aber wahrnehmbaren Anteilen auf *mentale Rotation* zurückgreift (vgl. Delgado & Prieto, 1997). Abbildung 6.33 stellt das entsprechend erweiterte Strukturmodell und die damit gewonnenen Ergebnisse dar.

Abbildung 6.33: Erweitertes Mediationsdiagramm mit mentaler Rotation und räumlicher Visualisierung (n = 464)



An diesem erweiterten Modell sind einige Details bemerkenswert:

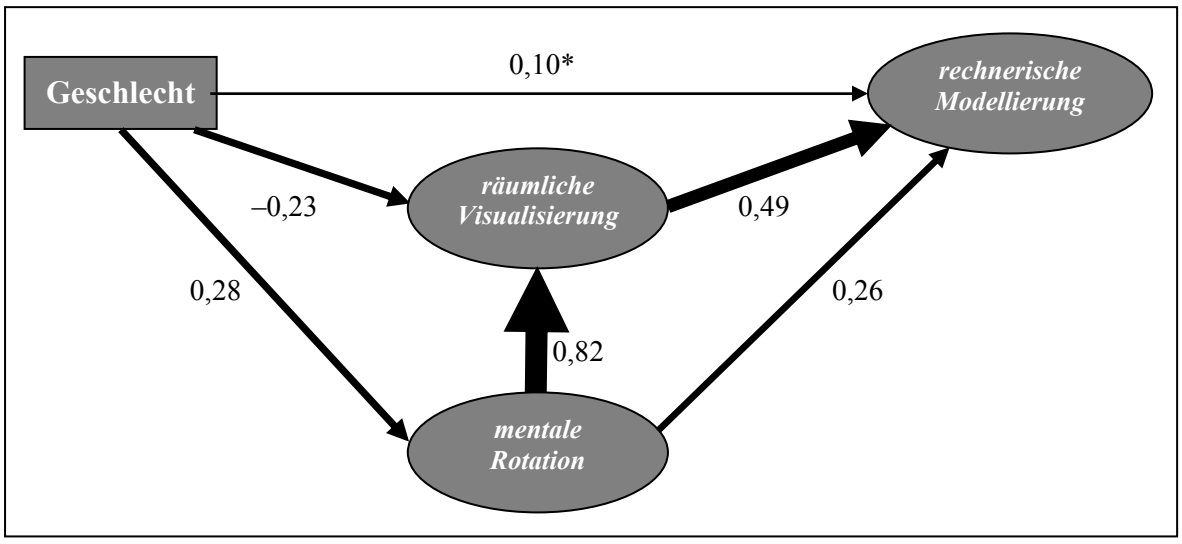
- Zunächst muss beachtet werden, wie negative Pfadkoeffizienten interpretiert werden müssen: Der Effekt von der Variablen *Geschlecht* auf die *räumliche Visualisierung* wirkt zugunsten der weiblichen, der entsprechende Effekt auf die *mentale Rotation* hingegen zugunsten von männlichen Versuchpersonen.
- Obwohl es praktisch keine Geschlechterunterschiede in der *räumlichen Visualisierung* gibt, liefert der standardisierte Pfadkoeffizient mit –0,23 einen relativ großen Betrag. Hierbei handelt es sich allerdings nur um den direkten Effekt. Der totale Effekt beträgt –0,01. Inhaltlich kann dies bedeuten, dass z. B. die analytischen Anteile im Konstrukt *räumliche Visualisierung* signifikante Unterschiede zugunsten weiblicher Versuchsper-

sonen erzeugen, die aber durch die Anteile *mentaler Rotation* im Konstrukt wieder ausgeglichen werden.

- Durch die Aufnahme von *räumlicher Visualisierung* in das Modell entsteht wieder ein nicht vermittelter Effekt von der Variablen *Geschlecht* auf die *Mathematikleistung*. Im reduzierten Modell sind anscheinend unter dem *Mediatoreffekt* der *mentalen Rotation* Komponenten statistisch zusammengefasst worden, die sich bei Berücksichtigung anderer Komponenten oder Konstrukte differenzieren lassen.
- Der totale Effekt der Variablen *Geschlecht* auf die *Mathematikleistung* setzt sich aus vier Wegen zusammen und ergibt insgesamt wieder den aus Abb. 6.26 bekannten Wert: $0{,}09 - 0{,}23 \cdot 0{,}54 + 0{,}27 \cdot 0{,}82 \cdot 0{,}54 + 0{,}27 \cdot 0{,}19 \approx 0{,}14$.
- Der Zusammenhang zwischen *mentaler Rotation* und *räumlicher Visualisierung* ist im obigen Modell stärker als die latente Korrelation, die keine weiteren Konstrukte berücksichtigt (vgl. Tab. 6.12, S. 216).
- Der direkte Effekt von *räumlicher Visualisierung* auf Mathematikleistung ist geringer als im reduzierten Modell in Abb. 6.28.
- Insgesamt werden durch dieses Modell 50 % der Varianz der *Mathematikleistung* statistisch erklärt, also kaum mehr als durch das reduzierte Modell in Abb. 6.28, dies aber differenzierter.

Da für die Frage der Geschlechterunterschiede besonders die Leistung bei *rechnerischen Modellierungsaufgaben* interessant zu sein scheint, wird das Modell aus Abb. 6.33 in Abb. 6.34 für diese Komponente der *Mathematikleistung* dargestellt.

Abbildung 6.34: Erweitertes Mediationsdiagramm mit mentaler Rotation, räumlicher Visualisierung und rechnerischen Modellierungsaufgaben (n = 464)



In diesem Modell ist der totale Effekt der *mentalen Rotation* auf die *rechnerischen Modellierungsaufgaben* mit ca. 0,66 deutlich höher als der totale (ausschließlich direkte) Effekt der räumlichen Visualisierung mit ca. 0,49. Der standardisierte Pfadkoeffizient von der Variablen *Geschlecht* auf die *Mathematikleistung* ist auf dem 5 %-Signifikanzniveau nicht von Null verschieden ($p = 0{,}06$), praktisch aber auch nicht gleich Null (s. o.), zumal er erhebliche Beiträge zum totalen Effekt der Variablen *Geschlecht* und zur Erklärung der Varianz der Zielvariablen leistet. Insgesamt werden durch das obige Modell ca. 52 % der Varianz der *Mathematikleistung* bei *rechnerischen Modellierungsaufgaben* erklärt.

Nach dieser Erweiterung des Mediationsmodells auf zwei Komponenten der *Raumvorstellung* ist die Frage naheliegend, ob nicht auch *räumliche Wahrnehmung* in das Modell aufgenommen werden soll. Hierfür fehlen allerdings die inhaltlichen Argumente und auch die explorative empirische Analyse ergibt nicht, dass dadurch ein weiterer Beitrag zur Erklärung von Geschlechterunterschieden oder Leistungsvarianz geleistet wird.

Erweitertes Modell mit FSK:M

Nachdem aus der Sicht der *Spatial Mediation Hypothesis* ein inhaltlich und empirisch leistungsfähiges Modell zur Erklärung von Geschlechterunterschieden in der *Mathematikleistung* vorliegt, ist es inhaltlich wünschenswert, auch Konstrukte jenseits der *Raumvorstellung* in ein solches Modell einzubeziehen. Nachdem das Konstrukt *Denkstile* in der Hauptuntersuchung nicht hinreichend gut erfasst werden konnte, steht im Rahmen der vorliegenden Arbeit noch das *FSK:M* exemplarisch als eines der wichtigsten psychosozialen Konstrukte im pädagogisch-psychologischen Bereich zur Verfügung. Grundlage für die Erweiterung der *Spatial Mediation Hypothesis* sind dabei die Modelle aus den Abbildungen 6.33 und 6.34.

Da es keine plausiblen Annahmen über Effekte zwischen Raumvorstellungskomponenten und dem *FSK:M* gibt, wird dieses Konstrukt als potenzieller weiterer Mediator zwischen der Variablen *Geschlecht* und der *Mathematikleistung* mitaufgenommen. Im Sinne der vorliegenden Fragestellung wird dabei als Wirkungsrichtung die vom *FSK:M* auf die Zielvariable *Mathematikleistung* angenommen, wohl wissend, dass es sich hierbei vermutlich um eine Wechselwirkung handelt (vgl. Kap. 4.1.2).

In den Abbildungen 6.35 und 6.36 wird einmal die globale (eindimensionale) *LSE 9*-Leistung (Abb. 6.35) und einmal die Leistung bei *rechnerischen Modellierungsaufgaben* als Zielvariable (Abb. 6.36) betrachtet.

Abbildung 6.35: Erweitertes Modell zur Erklärung von Geschlechterunterschieden in der Mathematikleistung (n = 440)

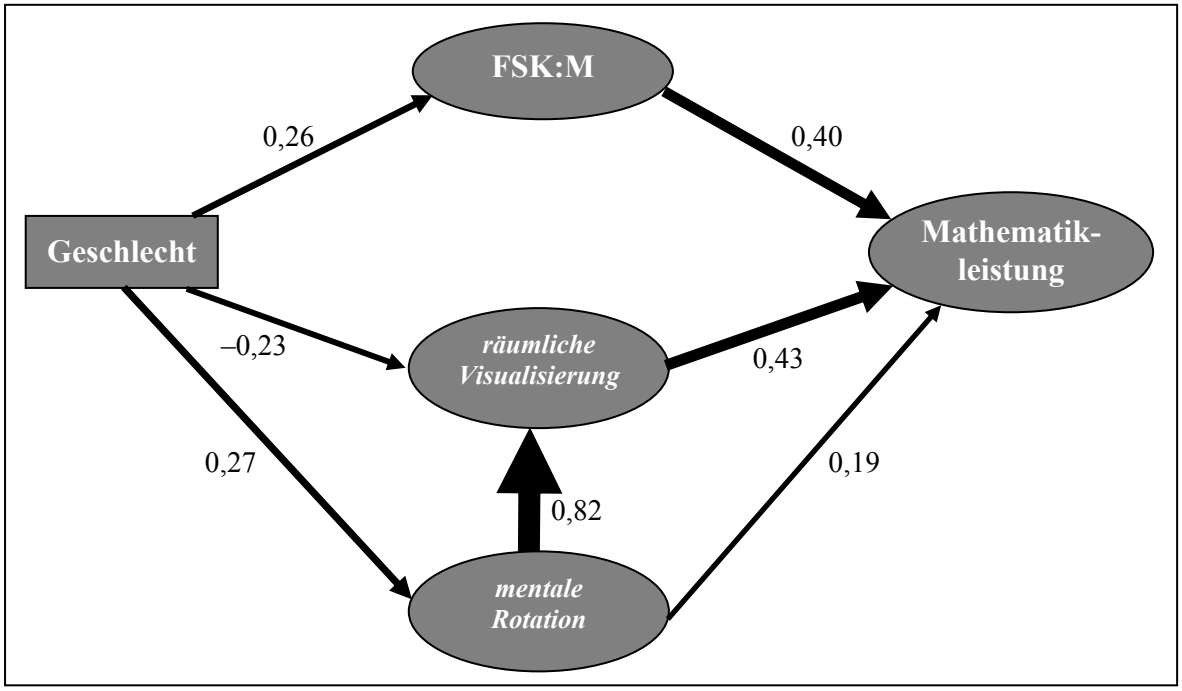
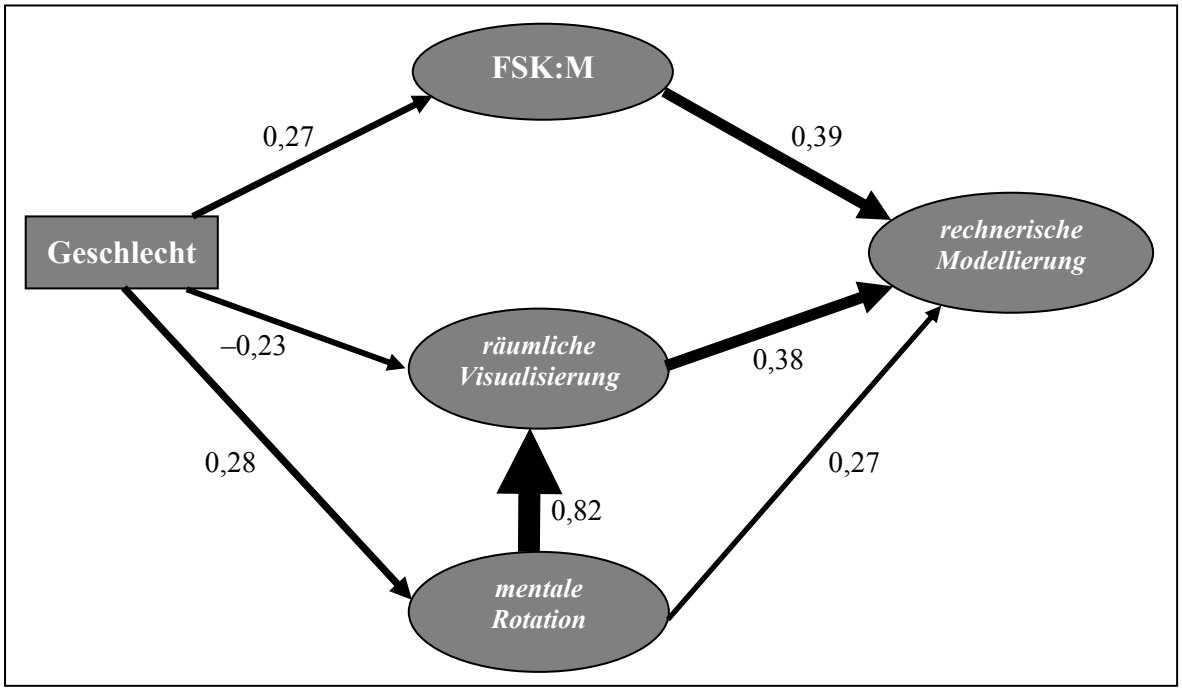


Abbildung 6.36: Erweitertes Modell zur Erklärung von Geschlechterunterschieden in der Mathematikleistung mit rechnerischen Modellierungsaufgaben (n = 440)



Nach der Aufnahme des *FSK:M* in das Modell zur Erklärung von Geschlechterunterschieden in der *Mathematikleistung* existiert bei keiner der beiden Zielvariablen ein direkter

Effekt der Variablen *Geschlecht* (jeweils $\beta = 0{,}00$). Zwar ist nicht auszuschließen, dass ein solcher Effekt durch die Aufnahme weiterer Variablen wieder auftreten könnte, aber die vorhandenen Mediatoren decken schon ein relativ breites Feld ab (figurale und analytische Denkprozesse sowie psychosoziale Aspekte). Der auffälligste Unterschied zwischen den beiden Modellen in den Abbildungen 6.35 und 6.36 ist sicherlich der deutlich stärkere direkte Effekt der *mentalen Rotation* auf *rechnerische Modellierungsaufgaben* als auf globale *Mathematikleistung*. Hier könnte also tatsächlich ein Schlüssel zum Verstehen von Geschlechterunterschieden in der *Mathematikleistung* liegen.

Bemerkenswert ist am Modell in Abb. 6.36 auch, dass der totale Effekt der Variablen *Geschlecht* in diesem Modell am größten ist. Dies ist vor allem durch die größeren Geschlechterunterschiede bei *rechnerischen Modellierungsaufgaben* bedingt. Der größte totale Effekt geht in diesem Modell von der *mentalen Rotation* mit 0,58 aus. Für das *FSK:M* und die *räumliche Visualisierung* betragen die totalen Effekte 0,39 bzw. 0,38; sie sind also etwa gleich groß. Insgesamt klärt das Modell 54 % der Varianz der *Mathematikleistung* auf und ist somit recht leistungsfähig. Dies gilt umso mehr, als dieses Modell nur zur Erklärung von Geschlechterunterschieden und nicht generell zur Erklärung von Leistungsvarianz entwickelt worden ist.

Wenn man im Modell noch weitere exogene Variablen berücksichtigt, die nicht innerhalb des Strukturmodells auf andere zurückgeführt werden, dann dürfte sich der Anteil erklärter Varianz noch steigern lassen. Beim Pfadmodell aus *PISA 2000* (vgl. Abb. 2.7, S. 38) etwa wurden als exogene Variablen noch die *kognitiven Fähigkeiten* und der *sozioökonomische Status* (sowie als endogene Variable die *Lesekompetenz*) berücksichtigt.[131] Der erklärte Varianzanteil bei diesem Modell lag bei 76 %; bei einer Aufnahme der fachnahen kognitiven Prozesse *mentale Rotation* und *räumliche Visualisierung* in das Modell aus Abb. 2.7 dürfte dieser ohnehin schon beachtliche Anteil noch weiter gesteigert werden können.

[131] Dass im Modell aus Abb. 6.36 der direkte Effekt des *FSK:M* erheblich höher ist als im Modell aus Abb. 2.7 ($\beta = 0{,}14$), kann vermutlich darauf zurückgeführt werden, dass die kognitiven Fähigkeiten im Modell aus Abb. 6.36 nicht explizit berücksichtig sind und entsprechend im (nicht mit modellierten) Hintergrund wirken.

7 Zusammenfassung der Befunde, Diskussion und Ausblick

Die vorliegende Arbeit geht im Kontext der Erforschung von *Mathematikleistung* von der Frage der potenziellen Erklärung von Geschlechterunterschieden in der *Mathematikleistung* durch entsprechende Unterschiede in der *Raumvorstellung* aus. Für diesen Zweck wurden in Kapitel 2 Grundlagen und Modelle sowie aktuelle Befunde und Entwicklungen der empirischen Bildungsforschung zusammengefasst und diskutiert. Kapitel 3 hat entsprechend Grundlagen und Befunde der Erforschung von *Raumvorstellung* dargestellt, wobei der Fokus auf die *Small-Scale Fähigkeiten* gerichtet wurde. Auf dieser Basis wurde ein eigenes Konstrukt von *Raumvorstellung* festgelegt, das Ausgangspunkt für die Auswahl und Weiterentwicklung von passenden Testinstrumenten war. Diese Instrumente wurden in einer Voruntersuchung erprobt, anschließend optimiert und in der Hauptuntersuchung in zeitlicher Nähe zu den *LSE 9* eingesetzt. Leider konnte das inhaltlich zur Fragestellung passende Konstrukt *Denkstile* nicht adäquat durch einen *Paper and Pencil Test* erfasst werden.

Zum Abschluss der vorliegenden Arbeit werden die empirischen Befunde, die zuvor differenziert dargestellt und lokal diskutiert worden sind (Kap. 5.4 und 6.4), zusammengefasst, in die theoretischen Grundlagen der Kapitel 2 und 3 integriert und mit Blick auf mögliche Konsequenzen für die empirische Bildungsforschung und die Mathematikdidaktik diskutiert. Schließlich werden in einem Ausblick solche Forschungsfragen und -vorhaben gestellt bzw. skizziert, die im Anschluss an die vorliegende Arbeit relevant und realistisch erscheinen.

7.1 Zusammenfassung der Befunde

Im Folgenden werden die Befunde stark verdichtet und nach den einzelnen Untersuchungsbereichen zusammengestellt berichtet. Vor den eigentlichen inhaltlichen Befunden wird dabei zunächst noch einmal der Einsatz der ausgewählten und weiterentwickelten Instrumente zur *Raumvorstellung* reflektiert.

7.1.1 Instrumente zur Erfassung der Raumvorstellung

Ausgehend von der Forschungsliteratur zur *Raumvorstellung* und der eigenen Fragestellung basiert die vorliegende Arbeit wesentlich auf dem 3-Komponenten-Modell der *Raumvorstellung* von Linn & Petersen (1985) mit den Komponenten *räumliche Wahrnehmung*, *mentale Rotation* und *räumliche Visualisierung*. Mit den Untertests *WLT*, *MRT* und *DAT:SR* wurden in der Hauptuntersuchung drei Tests eingesetzt, die über eine hohe Augenscheinvalidität verfügen und die sich beim Einsatz in der realisierten Stichprobe (Schülerinnen und Schüler der 9. Jahrgangsstufe aller Schulformen) voll bewährt haben:

- Die Verteilungen der jeweiligen Testleistungen zeigen, dass die zugrundeliegenden Konstrukte mit den Tests hinreichend differenziert erfasst werden können. Lediglich im Gymnasium gibt es einen moderaten Deckeneffekt, der bei Zusammenhangsanalysen zu einer leichten Unterschätzung der Zusammenhänge zwischen verschiedenen Konstrukten führen kann. Auf der Ebene des gesamten Altersjahrgangs sind die drei Tests aber ohne Einschränkung für Untersuchungsfragen wie die der vorliegenden Arbeit geeignet.
- Die drei eingesetzten Tests sind mit guten empirischen Kennwerten Rasch-skalierbar. Anstelle der oft üblichen klassischen Auswertung nur auf der Basis des Gesamtscores können also die Vorteile des *RM* genutzt werden. Solche Vorteile sind z. B. das höhere Skalenniveau der erhaltenen Messwerte oder eine gemeinsame (mehrdimensionale) latente Modellierung mit anderen Konstrukten. Die restriktiven Annahmen des *RM*, wie die parallel verlaufenden *ICCs*, sind bei solchen Tests, die relativ eng gefasste Bereiche kognitiver Fähigkeiten erfassen sollen, durchaus plausibel.
- Wenn man auf der Basis dieser oder vergleichbarer Testinstrumente Zusammenhangsuntersuchungen zwischen der *Raumvorstellung* und anderen Konstrukten durchführen möchte, stellt die Art der Stichprobenzusammenstellung aus der Hauptuntersuchung einen gut gangbaren Weg dar. Durch die Berücksichtigung der verschiedenen Schulformen des gegliederten Systems kann eine grobe Über- oder Unterschätzung der vorhandenen Varianz anscheinend vermieden werden.

7.1.2 Raumvorstellung

Die empirischen Befunde der Hauptuntersuchung bestätigen nachdrücklich, dass das 3-Komponenten-Modell von Linn & Petersen (1985) eine tragfähige, wenn auch sicherlich nicht einzig denkbare Grundlage für die Erforschung von Geschlechterunterschieden oder von Zusammenhängen der *Raumvorstellung* mit anderen kognitiven Leistungen darstellt:

- Für die von Linn & Petersen selbst angegebenen Referenztests *WLT* (*räumliche Wahrnehmung*), *MRT* (*mentale Rotation*) und *DAT:SR* (*räumliche Visualisierung*) wies die dreidimensionale Modellierung nach dem *RM* deutlich bessere empirische Kennwerte auf als die eindimensionale. Auch die latenten Korrelationen zwischen den Komponenten zeigen deutlich, dass es zwar ein statistisches Gemeinsames der Komponenten gibt, sie aber vor allem gut empirisch getrennt werden können. Hieraus folgt allerdings nicht, dass das verwendete Modell von Linn & Petersen grundsätzlich das beste für die empirische Erforschung von *Raumvorstellung* ist, da es nicht gegen andere Faktoren- oder Komponenten-Modelle getestet wurde. Für Fragestellungen, die eine inhaltliche Nähe zur vorliegenden Arbeit aufweisen, scheint es aber äußerst tragfähig und den anderen bekannten Modellen überlegen zu sein.
- Mit dem verwendeten Modell und den genannten Referenztests konnten in der Hauptuntersuchung insgesamt Ergebnisse gewonnen werden, die die vorhandene Forschungsliteratur konsistent ergänzen. Die signifikanten Geschlechterunterschiede zugunsten männ-

licher Versuchspersonen mit Effektstärken von 0,8 im *MRT* bzw. 0,4 im *WLT* bestätigen die Ergebnisse von Linn & Petersens Meta-Analyse ebenso wie das Fehlen signifikanter Unterschiede beim *DAT:SR*., das auf die komplexe Struktur der *räumlichen Visualisierung* zurückgeführt werden kann: Für den *DAT:SR* haben entsprechende Strukturgleichungsmodelle gezeigt, dass sich Geschlechterunterschiede zugunsten weiblicher Versuchspersonen im Bereich analytischer Prozesse mit Geschlechterunterschieden zugunsten männlicher Versuchpersonen im Bereich *mentaler Rotation* neutralisieren.

- Auch für die Zusammenhänge mit *Mathematikleistung* und die Frage der *Spatial Mediation Hypothesis* hat die Trennung der Komponenten *räumliche Wahrnehmung*, *mentale Rotation* und *räumliche Visualisierung* wertvolle Beiträge geleistet. Bei der Erklärung des Zusammenhangs von Geschlechterunterschieden in der *Mathematikleistung* scheint seitens der *Raumvorstellung* die *mentale Rotation* die zentrale Rolle zu spielen.

7.1.3 Mathematikleistung

Nach der Diskussion der Grundlagen der Erforschung von *Mathematikleistung* in Kapitel 2 war aus Sicht des Autors der vorliegenden Arbeit besonders interessant, wie sich der *LSE 9*-Mathematiktest in der Auswertung verhält. Die *LSE 9* wurden in der Entwicklung genau auf der Basis der dargestellten Grundlagen konzipiert und insbesondere im Rahmen der Pilotierung und für die Rückmeldung der Leistungsdaten in einem eindimensionalen *RM* skaliert, wobei etwaige Items mit schlechten Fit-Werten ausgesondert wurden. Dennoch konnte in der Hauptuntersuchung gezeigt werden, dass es inhaltlich und empirisch tragfähige mehrdimensionale Strukturen in diesem Test gibt. Gleichwohl passt auch eine eindimensionale Modellierung des Tests noch recht gut zu den beobachteten Daten.

- Die Geschlechterunterschiede zugunsten der Jungen weisen bei den *LSE 9*-Mathematik bei einer globalen (eindimensionalen) Auswertung eine Effektstärke von 0,63 auf.
- Bei einer statistischen Konfektionierung von Subtests nach empirischer Schwierigkeit zeigt sich, dass diese Effektstärke positiv mit der Testschwierigkeit korreliert ist. Beim Subtest mit den 14 empirisch leichtesten Items (von 58 Items) beträgt die Effektstärke 0,44; beim Subtest mit den 22 empirisch schwierigsten Items beträgt die Effektstärke 0,69. Ein vollständiges Verschwinden der Geschlechterunterschiede für empirisch leichte Items, so wie es in der Forschungsliteratur des Öfteren berichtet wird, kann allerdings nicht festgestellt werden. Dies kann vermutlich durch das Konzept eines grundbildungs- und kompetenzorientierten Tests erklärt werden.
- Bei einer Unterscheidung der *LSE 9*-Items nach *Typen mathematischen Arbeitens* zeigt sich – im Einklang mit der Forschungsliteratur –, dass Geschlechterunterschiede zugunsten der männlichen Versuchspersonen bei *rechnerischen Modellierungsaufgaben* mit einer Effektstärke von 0,72 deutlich stärker ausfallen als für *technische Aufgaben*.
- Die *Typen mathematischen Arbeitens* scheinen insgesamt eine ergiebige Klassifikation darzustellen, die inhaltlich und empirisch differenzierte Blicke ermöglicht. Bei einer

Einteilung der *LSE 9*-Items in *technische Aufgaben*, *rechnerische Modellierungsaufgaben* und *begriffliche Modellierungsaufgaben* weist eine entsprechende dreidimensionale Rasch-Skalierung bessere empirische Kennwerte auf als eine eindimensionale Skalierung, die aber im Sinne eines Skalierungspragmatismus insbesondere für Leistungsvergleiche genutzt werden kann. Die Ergiebigkeit der Differenzierung nach *Typen mathematischen Arbeitens* zeigt sich auch bei der Zusammenhangsuntersuchung mit den Komponenten der *Raumvorstellung* (s. u.).

7.1.4 Bereichsspezifisches Fähigkeitsselbstkonzept Mathematik

Das *FSK:M* wurde in der Hauptuntersuchung mit drei in *PISA 2000* bewährten Items erfasst, die eine Skala mit hoher innerer Konsistenz bilden. Die drei Items lassen sich mit sehr guten Kennwerten auf der Basis eines restringierten ordinalen *RMs* skalieren. Sie haben in der Hauptuntersuchung Ergebnisse geliefert, die nur zum Teil im Einklang mit der Forschungsliteratur stehen:

- Die Geschlechterunterschiede zugunsten männlicher Versuchspersonen waren vor diesem Hintergrund mit einer Effektstärke von 0,63 absolut erwartungskonform – und liegen dabei genau in der Größenordung der Unterschiede bei der globalen Auswertung des *LSE 9*-Mathematiktests. Dies ist aufgrund der Wechselwirkungen zwischen beiden Konstrukten äußerst plausibel. Allerdings bleibt auch bei statistischer Kontrolle der *Mathematikleistung*, z. B. durch eine Partialkorrelation, eine Kovarianzanalyse oder ein entsprechendes Strukturgleichungsmodell, ein eigenständiger Effekt der Variable *Geschlecht* auf die Ausprägung des *FSK:M* bestehen. Im Rahmen eines entsprechenden Strukturgleichungsmodells ist ein direkter Effekt mit einem standardisierten Pfadkoeffizienten $\beta = 0{,}18$ signifikant von Null verschieden. Der standardisierte totale Effekt beträgt 0,37. Inhaltlich bedeutet dies, dass das *FSK:M* bei gleicher *Mathematikleistung* bei männlichen Versuchspersonen tendenziell höher ausgeprägt ist als bei weiblichen.
- Vor dem Hintergrund der Forschungsliteratur ist *nicht* erwartungskonform, dass es signifikante Schulformunterschiede gibt, die zwischen dem beteiligten Gymnasium und der beteiligten Gesamtschule eine Effektstärke von ebenfalls 0,63 haben. Im Sinne der Referenzrahmentheorie wird das *FSK:M* durch soziale und dimensionale Vergleiche geprägt („Meine Mathematikleistung im Vergleich zu anderen Schülerinnen und Schülern und im Vergleich zur Leistung in anderen Fächern“). Diese Mechanismen müssten eigentlich bewirken, dass innerhalb jeder Schulform und jeder Schule das *FSK:M* im Wesentlichen gleich ausgeprägt ist. An dieser Stelle scheint vor einer Fortschreibung der alten Befunde weitere Forschung erforderlich zu sein.

7.1.5 Zusammenhang von Mathematikleistung und Raumvorstellung

Zum Zusammenhang von *Mathematikleistung* und *Raumvorstellung* kann die vorliegende Arbeit differenzierte Befunde beitragen, wobei das 3-Kompenten-Modell der *Raumvorstellung* und die *Typen mathematischen Arbeitens* eine hervorragende konzeptionelle Grundlage bilden. Dennoch können die Zusammenhänge mit den Methoden dieser Arbeit nur statistisch, nicht aber inhaltlich erklärt werden.

- Betrachtet man die *LSE 9-Mathematikleistung* global, so bestehen für alle drei Komponenten der *Raumvorstellung* substanzielle latente Korrelationen im jeweiligen zweidimensionalen *RM*, die höchste für den *DAT:SR* (*räumliche Visualisierung*) mit einem Wert von 0,68. Dies ist aufgrund der Anteile analytischer Denkprozesse an der Komponente *räumliche Visualisierung* inhaltlich plausibel.
- Die Zusammenhänge zwischen den drei Komponenten der *Raumvorstellung* und *Mathematikleistung* verändern sich nicht, wenn *LSE 9*-Subtests nach empirischer Schwierigkeit der Items gebildet werden.
- Betrachtet man die *Mathematikleistung* differenziert nach *Typen mathematischen Arbeitens*, so zeigt sich allerdings, dass *mentale Rotation* enger mit *rechnerischen Modellierungsaufgaben* zusammenhängt als mit *technischen Aufgaben*. Für die Komponenten *räumliche Wahrnehmung* und *räumliche Visualisierung* gibt es keine analogen Befunde.
- Die Befunde deuten insgesamt darauf hin, dass der Zusammenhang von *mentaler Rotation* und *rechnerischen Modellierungsaufgaben* mathematikdidaktisch besonders interessant ist. Dabei ist zunächst unklar, wie dieser Zusammenhang inhaltlich erklärt werden kann, zumal der *LSE 9*-Mathematiktest kein Item enthält, bei dem figurales Denken eine Rolle spielt. Solche Fragen der inhaltlichen Zusammenhänge müssen mit einem Untersuchungsdesign, das sich auch qualitativer Methoden bedient, bearbeitet werden.

7.1.6 Erklärung von Geschlechterunterschieden in der Mathematikleistung

Ausgehend von den differenzierten Befunden in den einzelnen Untersuchungsbereichen kann die Kernfrage der vorliegenden Arbeit untersucht werden. Geschlechterunterschiede in der *Mathematikleistung* sollen dabei möglichst umfassend durch Mediatorvariablen erklärt werden. Methodisch wurde dies mit entsprechenden Strukturgleichungsmodellen umgesetzt, die eine Schätzung der latenten Zusammenhänge ermöglichen.

- Für die *Spatial Mediation Hypothesis* zeigt sich, dass Befunde primär von der berücksichtigten Komponente der *Raumvorstellung* und sekundär von den *Typen mathematischen Arbeitens* abhängen. Statisch ergiebig – im Sinne einer Erklärung von Geschlechterunterschieden – ist das Mediationsmodell mit *mentaler Rotation* und *rechnerischen Modellierungsaufgaben*.

- Die *Spatial Mediation Hypothesis* wurde in einem ersten Schritt erweitert, indem neben der *mentalen Rotation* auch die *räumliche Visualisierung* (als zusätzliche Variable) und die Anteile *mentaler Rotation* an der *räumlichen Visualisierung* (als entsprechender Pfad) berücksichtigt wurden. So erhält man ein Strukturmodell, das unterschiedliche kognitive Prozesse berücksichtigt, die an der *Raumvorstellung* beteiligt sind: analytische Denkprozesse und *mentale Rotation*. Dies ist für die Erklärung von Geschlechterunterschieden inhaltlich von Interesse, da weibliche Versuchspersonen bei den analytischen Denkprozessen bessere Leistungen zeigen und männliche bei *mentaler Rotation*.
- Das so erhaltene Strukturmodell wurde in einem zweiten Schritt um das *FSK:M* als geschlechtersensitive Komponente aus dem Bereich der psychosozialen Konstrukte ergänzt. Damit kann die Mediation von Geschlechterunterschieden in der *Mathematikleistung* noch differenzierter erklärt werden. Insgesamt liegt so aus der Sicht der Erklärung von Geschlechterunterschieden ein leistungsfähiges Modell vor, das noch weiter ausgebaut werden kann.
- Es erscheint lohnenswert weitere Komponenten, insbesondere aus dem Bereich kognitiver Fähigkeiten und kognitiver Verarbeitung, in das Modell zu integrieren. Hier ist es inhaltlich besonders naheliegend, das Konstrukt *Denkstile* zu berücksichtigen, das aber ggf. nur in eher qualitativen Einzeltests erfasst werden kann. Bei anderen kognitiven Fähigkeiten im Sinne der typischen Intelligenztests muss vor einer Integration ins Modell genau geprüft werden, wo ggf. Dopplungen entstehen. So sind Aspekte figuraler, numerischer und analytischer Intelligenz bereits im Modell enthalten. Durch inhaltliche Dopplungen würden auch statistische Zusammenhänge in das Modell hineingetragen, die andere Effekte überlagern könnten.
- Bei der Arbeit mit Strukturgleichungsmodellen wie dem erweiterten Mediationsmodell aus der Hauptuntersuchung muss berücksichtigt werden, dass häufig Wirkungsrichtungen unterstellt werden, die inhaltlich nicht immer befriedigend geklärt sind. In der Regel werden Richtungen postuliert, die dem Anliegen entsprechen, möglichst viel Varianz der Zielvariablen zu erklären. Wenn dies bei der Diskussion der Ergebnisse berücksichtigt wird, entsteht daraus noch kein Schaden, häufig werden solche Pfeile aber auch als tatsächliche (kausale) Wirkungen interpretiert.[132]

[132] Ein prominentes Beispiel hiefür ist die Rezeption des „Pfadmodells zur Erklärung der Mathematikleistung“ aus dem Ergebnisbericht zu *PISA 2000*, das in Abb. 2.7 (S. 38) wiedergegeben ist. Der Pfeil von „Leistung Lesen“ auf „Leistung Mathematik“ wurde fortan so gelesen, dass aus einer Leseförderung automatisch eine Steigerung der *Mathematikleistung* folgt. So wichtig die Förderung der Unterrichtsprache ist, das Pfadmodell hatte nur den Zweck möglichst viel Varianz der *Mathematikleistung* zu erklären – und so wurde auch der fragliche Pfeil postuliert. Er könnte durchaus auch andersherum gezeichnet werden, wenn man bedenkt, dass *Lesekompetenz* bei *PISA* auch die Informationsentnahme aus Diagrammen beinhaltet.

7.2 Konsequenzen für die empirische Bildungsforschung

Die Befunde der vorliegenden Arbeit basieren wesentlich auf der Methodologie und auf ausgewählten Instrumenten und Verfahren der empirischen Bildungsforschung. Inhaltlich wurde die Arbeit durch entsprechende Befunde aus der empirischen Bildungsforschung angeregt, wie die Herleitung der Fragestellung in Kapitel 1 zeigt. Dementsprechend kann die Arbeit einen – als Einzelvorhaben eher bescheidenen – Beitrag zur Entwicklung der empirischen Bildungsforschung leisten. Die Auswertungen des *LSE 9*-Mathematiktests unterstützen einerseits einen Skalierungspragmatismus, der bei breit angelegten Mathematiktests und einer heterogenen Stichprobe eine eindimensionale Skalierung der Testleistung z. B. zum Zwecke von Leistungsvergleichen zulässt. Andererseits wurde mit den *Typen mathematischen Arbeitens* ein mehrdimensionales Modell gefunden, das bessere empirische Kennwerte als das eindimensionale hat und inhaltlich reichhaltiger ist.[133] Dementsprechend kann je nach Zweck der Leistungsmessung ein eindimensionales oder ein mehrdimensionales Modell angemessener sein, wobei es vermutlich auch tragfähige Alternativen zu den *Typen mathematischen Arbeitens* gibt.

Mögliche Konsequenzen aus der vorliegenden Arbeit werden mit Blick auf (a) die Ausdifferenzierung von Rahmenmodellen für die Erforschung von *Mathematikleistung* und (b) auf die inhaltliche Erklärung statistischer Zusammenhänge dargestellt.

7.2.1 Rahmenmodelle für die Erforschung von Mathematikleistung

Die vorliegende Arbeit hat neben der Erklärung von Geschlechterunterschieden in der *Mathematikleistung* durch ein geeignetes Mediationsmodell, das unter anderem *Raumvorstellung* berücksichtigt, als weiteres Ziel, durch die Entwicklung ebendieses Modells zur Ausdifferenzierung von Rahmenmodellen für die Erforschung von *Mathematikleistung* (vgl. Kap. 2.1.4) beizutragen. Innerhalb der empirischen Bildungsforschung ist es *eine* wichtige Aufgabe der beteiligten Fachdidaktiken, relevante kognitive Voraussetzungen und Verarbeitungsprozesse für das fachliche Lernen zu identifizieren und die verwendeten Rahmenmodelle an diesen Stellen auszudifferenzieren. Auf diesem Weg können Schulleistungsstudien über Leistungsvergleiche und die Erklärung von Leistungsunterschieden durch sozialstatistische Hintergrundmerkmale hinaus möglicherweise auch zur (zunächst statistischen und in der Folge ggf. inhaltlichen) Aufklärung von fachlichen Lernprozessen beitragen.

---

[133] Demgegenüber scheint die curriculare Strukturierung nach Inhaltsgebieten bzw. Leitideen, die bisher häufig (und in der Regel ohne Erfolg) als mehrdimensionale Alternative gegen das eindimensionale Modell getestet wurde, genauso wenig ergiebig zu sein wie eine Unterscheidung nach „prozessbezogenen“ bzw. „allgemeinen mathematischen Kompetenzen“. Dies ist kaum überraschend, da für eine dimensionale Struktur des Leistungskonstrukts eher unterschiedliche kognitive Prozesse identifiziert werden müssen. Curriculare Cluster sind aber fachsystematisch und nicht nach kognitiven Prozessen der Aufgabenbearbeitung gebildet. Die *Typen mathematischen Arbeitens* leisten hier eine plausible Strukturierung.

- Die Befunde der Hauptuntersuchung, vor allem das erweiterte Modell zur Erklärung von Geschlechterunterschieden in der *Mathematikleistung*, deuten darauf hin, dass der Bereich *Raumvorstellung* eine sinnvolle Ergänzung der Rahmenmodelle im Bereich der individuellen kognitiven Voraussetzungen bzw. der individuellen Verarbeitung darstellt. Dabei sollte vorrangig *mentale Rotation* und möglichst auch *räumliche Visualisierung* berücksichtigt werden. Wie in der vorliegenden Arbeit dürften dafür Tests mit ungefähr zehn Items völlig ausreichen. Für eine solche Ergänzung dürfte es eigentlich keine testökonomischen Hindernisse geben, da entsprechende Instrumente erprobt und bewährt vorliegen – und wenn man die Skalenhandbücher großer Schulleistungsstudien betrachtet, stellt man fest, dass es einige Konstrukte gibt, die für *Mathematikleistung* weniger relevant sind und trotzdem erfasst werden.
- Darüber hinaus scheint es geboten zu sein, *Mathematikleistung* nicht nur global, sondern auch nach Komponenten wie den *Typen mathematischen Arbeitens* differenziert zu betrachten. Solche differenzierten Analysen dürften – wie in der vorliegenden Arbeit – ein großes Potenzial zur Erklärung differentieller Befunde haben.
- Für das Konstrukt *Denkstile* sollte weiter versucht werden, einen geeigneten *Paper and Pencil Test* zu entwickeln, um die vermuteten Zusammenhänge zu untersuchen. Falls nur aufwändige Einzeltests möglich sind, sollte dies trotzdem für Teilstichproben überlegt werden. Der mögliche Mehrwert des Konstrukts *Denkstile* kann auf der Basis der empirischen Befunde der vorliegenden Arbeit allerdings nicht eingeschätzt werden.

7.2.2 Inhaltliche Erklärung von Zusammenhängen

Für eine inhaltlich ausgerichtete empirische Bildungsforschung, die nicht nur immer neue Leistungsdaten, sondern auch inhaltliche Beiträge für das Verstehen von fachlichem Lernen und Leisten generieren möchte, ist es wichtig, statistische Zusammenhänge auch inhaltlich zu erklären. Dazu gehört einerseits die differenzierte Betrachtung von Zusammenhängen und andererseits die inhaltliche Betrachtung von statistischen Modellen. Die statistische Erklärung von Leistungsvarianz erfolgt häufig über Regressionsanalysen, bei denen stets eine Wirkungsrichtung postuliert werden muss; auch in allgemeineren Strukturgleichungsmodellen wird häufig mit einseitig angenommenen Wirkungen gearbeitet. Wichtig ist es, die Zusammenhänge und Wirkungen inhaltlich zu verstehen und abzusichern.

- Wie wichtig die differenzierte Betrachtung von Konstrukten in entsprechenden Modellen ist, zeigt das Beispiel der *mentalen Rotation* und der *räumlichen Visualisierung*. Je nach dem, welche Komponente man in einem Modell berücksichtigt, kommt man zu ganz unterschiedlichen Ergebnissen bezüglich der Zusammenhänge mit *Mathematikleistung* oder bezüglich der Erklärung von Geschlechterunterschieden. Viele vorliegende Studien verwenden nur eine Komponente, generalisieren die Befunde dann aber für *Raumvorstellung*, sodass eine (vermutlich nur scheinbar) heterogene Befundlage nicht verwundert.

- Die vorliegende Arbeit hat auch einige Wirkungsrichtungen mit Blick auf die Zielvariable *Mathematikleistung* angenommen, die kritisch hinterfragt und inhaltlich geklärt werden müssen. Dies betrifft den Zusammenhang zwischen *Mathematikleistung* und *Raumvorstellung* ebenso wie den Zusammenhang zwischen *Mathematikleistung* und *bereichsspezifischem Fähigkeitsselbstkonzept*.[134]
- Die Ergebnisse der vorliegenden Arbeit deutet darauf hin, dass für die inhaltliche Erklärung von Zusammenhängen auch von Seiten der Fachdidaktik eine kognitive Sichtweise hinreichend stark berücksichtigt werden muss. Die *Typen mathematischen Arbeitens* zeigen das prinzipielle Potenzial solcher Betrachtungen. Warum *mentale Rotation* besonders eng mit den *LSE 9*-Items des Typ *rechnerische Modellierungsaufgaben* zusammenhängt, ist noch unklar, zumal diese Items kein figurales Denken erfordern. Hier muss offensichtlich die Aufgabenbearbeitung in beiden Tests intensiver aus kognitiver Sicht erforscht werden. Dabei können vermutlich auch die Strategieunterschiede, die vor allem bei den *IST:WÜ*-Aufgaben in der Voruntersuchung (*2 DW* und *3 DW*) betrachtet wurden, entsprechende Beiträge leisten. Mit den genannten Instrumenten ist eine statistische Klassifikation nach Bearbeitungsstrategien mithilfe von *LCAs* möglich, sodass man nicht ausschließlich auf (ressourcenintensive) Einzeltests angewiesen ist.

7.3 Konsequenzen für die mathematikdidaktische Forschung und Entwicklung

Die vorliegende Arbeit stellt aus mathematikdidaktischer Sicht Grundlagenforschung in großer Nähe zur Psychologie dar. Im Sinne der oben dargestellten potenziellen Bedeutung für Rahmenmodelle der Schulleistungsforschung kann sie auch als originär mathematikdidaktischer Beitrag zur empirischen Bildungsforschung verstanden werden. Die für jede Fachdidaktik zentrale konstruktive Entwicklungsforschung wird dabei zunächst nicht berührt. Allerdings stellen sich vor dem Hintergrund der in den Kapiteln 2 und 3 referierten und diskutierten Befunde sowie der eigenen Befunde Fragen an die Mathematikdidaktik, die auch als mögliche Konsequenzen formuliert werden können. Dabei spielen normative bildungstheoretische Aspekte ebenso eine Rolle wie die Entwicklungsforschung.

---

[134] Wozu die unreflektierte Verwendung von „Pfeilrichtungen“ in eigenen Modellen führen kann, zeigt die folgende Schlussfolgerung aus dem „Pfadmodell zur Erklärung von Mathematikleistung“ (vgl. Abb. 2.7, S. 38): „Festzuhalten bleibt: […] (c) Eine wichtige Funktion hat ferner das Selbstkonzept der mathematischen Begabung. Ein Teil der Geschlechterunterschiede und des Einflusses kognitiver Grundfähigkeiten ist über das Selbstkonzept vermittelt. Daraus ergibt sich die pädagogische Aufgabe, das mathematische Selbstkonzept gerade bei Mädchen sowie bei Schülerinnen und Schülern mit schwächeren kognitiven Grundfähigkeiten zu fördern“ (Klieme et al., 2001, S. 185). Aus mathematikdidaktischer Sicht – vielleicht aber auch vor dem Hintergrund inhaltlichen Denkens anstelle des Festhaltens an schnell gezeichneten Pfeilrichtungen – ist es sicherlich sinnvoller, das Mathematiklernen individuell zu fördern, wenn man *Mathematikleistung* verbessern möchte, statt „das mathematische Selbstkonzept … zu fördern“; dabei ist auch unklar, wie eine isolierte Förderung des Selbstkonzepts aussehen könnte.

7.3.1 Stellenwert der Raumvorstellung im Mathematikunterricht

Unter den 58 Items des *LSE 9*-Mathematiktests war kein einziges, das figurales Denken erfordert. Die vorhandenen geometrischen Aufgabenkontexte waren überwiegend Ausgangspunkte für arithmetische bzw. algebraische Tätigkeiten. Die abgebildeten Figuren waren direkt erkennbar, nicht in ungewohnten Lagen präsentiert und mussten auch nicht zunächst geometrisch strukturiert werden. Dieser Befund ist durchaus charakteristisch für die curricularen Veränderungen im Bereich der Geometrie der Sekundarstufe I und für den aktuellen Stellenwert von *Raumvorstellung* in der Sekundarstufe I. Geometrische Kontexte werden im intendierten wie im implementierten Curriculum immer häufiger als Anlässe zum Rechnen verstanden. Dem „curricularen Verlust“ expliziter Anteile von *Raumvorstellung*, Abbildungsgeometrie oder anderen „nicht-rechnenden“ Bereichen stehen allerdings „curriculare Gewinne“ in anderen Bereichen, wie der Stochastik oder dem qualitativen Umgehen mit funktionalen Zusammenhängen, gegenüber.

Da die curricularen Vorgaben auf der Seite der fachlichen Gegenstände nicht beliebig ausgeweitet werden können und der Ruf nach einer „Entschlackung der Lehrpläne“ immer aktuell ist, muss fortwährend ausgehandelt werden, welche fachlichen Gegenstände in der Schule verbindlich sein sollen. Hier ist vor allem die Mathematikdidaktik gefordert, über bildungstheoretische Klärungen und konsistente curriculare Konzepte dazu beizutragen, dass ausgewogene und durchdachte curriculare Vorgaben für einen *allgemeinbildenden* Mathematikunterricht erlassen werden. Die vorliegende Arbeit leistet keinen Beitrag zur Klärung dieser normativen Frage, auch wenn der Autor einen eindeutigen Standpunkt dazu hat. Die Bestandsaufnahme ist aus Sicht der *Raumvorstellung* aber eindeutig.

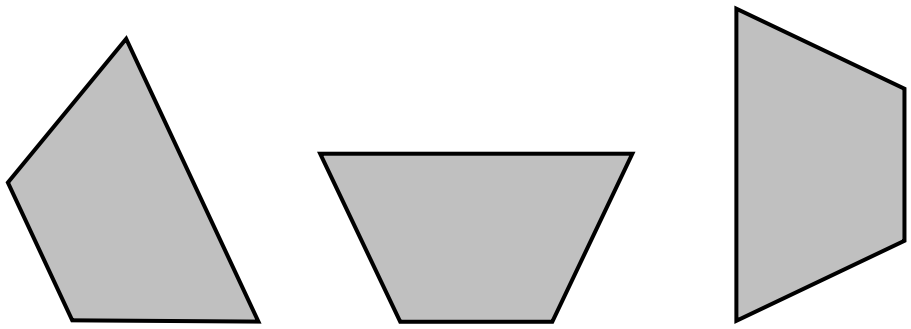
Sollten entsprechende Aushandlungsprozesse für den Mathematikunterricht der Sekundarstufe I – anders als aktuell für den Mathematikunterricht in der Primarstufe – zu dem Ergebnis führen, dass *Raumvorstellung* kaum noch zu den verbindlichen fachlichen Gegenständen zählen soll, dann kommen immer noch andere Fächer infrage, in denen *Raumvorstellung* explizit gefördert werden kann. Aufgrund der prinzipiellen Relevanz der *Raumvorstellung* als eine grundlegende kognitive Fähigkeit und Prädiktor für viele andere Leistungsbereiche ist es jedenfalls kaum vorstellbar, dass in der Sekundarstufe I entsprechende Fähigkeiten in keinem Fach explizit gefördert werden.

7.3.2 Konzeption und Evaluation von Fördermaßnahmen

Selbst wenn *Raumvorstellung* nicht mehr expliziter curricularer Bestandteil der Sekundarstufe I sein sollte, kann durch entsprechende Aufgabenkontexte immer wieder implizit eine Förderung entsprechender Fähigkeiten realisiert werden, ohne dass andere Ziele darunter leiden. Wenn z. B. Trapeze überwiegend als Anlässe zur Bestimmung geometrischer Maße im Unterricht auftauschen, dann kann und sollte über die räumliche (hier: zweidimensionale) Lage der Figuren auch *Raumvorstellung* gefördert werden. Erfahrungsgemäß haben

viele Schülerinnen und Schüler Schwierigkeiten, die Figuren in Abb. 7.1 als Trapeze zu identifizieren. Die implizite Förderung der „zweidimensionalen“ *mentalen Rotation* trägt darüber, dass solche Figuren überhaupt identifiziert werden können, auch dazu bei, dass die Maße der Figuren in unterschiedlichen Situation (und Lagen) bestimmt werden können.

Abbildung 7.1: Trapeze in „ungewohnter“ Lage



In der Primarstufe wird im Bereich der Geometrie eine Vielzahl gehaltvoller geometrischer Lernmaterialien und Spiele eingesetzt, die handlungsorientiert einen Vorstellungsaufbau fördern und dabei häufig auf die „prozessbezogenen“ bzw. „allgemeinen mathematischen Kompetenzen“ im Blick haben. Also bieten sich auch unabhängig von curricularen Veränderungen Möglichkeiten, *Raumvorstellung* zu fördern. Mit dem „Mathekoffer“ (Büchter & Henn, 2008) liegt ein Beispiel dafür vor, wie entsprechende Lernmaterialien aus der Primarstufe auch in der Sekundarstufe I produktiv genutzt werden können.

Sollte der Mathematikunterricht auch in der Sekundarstufe I *Raumvorstellung* explizit fördern wollen, so ist die mathematikdidaktische Entwicklungsforschung gefordert, konstruktive Vorschläge zu unterbreiten, die in der Regel nicht neu entwickelt, sondern nur aus Vorhandenem angepasst werden müssen. Aus allgemeiner lerntheoretischer Sicht und gemäß den mathematikdidaktischen Konzepten geometrischen Begriffslernens gibt es dabei einige Leitlinien, die die Entwicklung entsprechender Lernumgebungen unterstützen (vgl. Lompscher, 1988; van Hiele & van Hiele-Geldorf, 1978). Diese werden auch durch die empirischen Befunde zur Förderung der *Raumvorstellung* unterstützt.

Grundsätzlich hat sich bewährt, dass zunächst konkret handelnd mit realen Objekten Primärerfahrungen gesammelt werden. Auf diesen Primärerfahrungen können schrittweise Abstraktionen aufbauen, z. B. indem andere Darstellungen, insbesondere auch (zweidimensionale) Skizzen von dreidimensionalen Objekten, verwendet werden. Parallel dazu können Objekte nach ersten Eigenschaften klassifiziert und somit entsprechende Begriffe mit ihnen in Verbindung gebracht werden. Auch im Bereich figuralen Denkens spielen Verbalisierungen bei der mentalen Repräsentation von Objekten, Zusammenhängen und

Veränderungen eine wichtige Rolle. Schließlich sollen die Lernenden rein mental mit den (idealisierten) Modellen agieren können, wobei z. B. mental gewonnene Aufgabenlösungen ggf. wieder mit konkreten Objekten überprüft werden können.

Aus der Sicht des Lernens und Lehrens von Mathematik gibt es unter anderem die beiden folgenden offenen Fragen, zu deren Beantwortung die mathematikdidaktische Forschung und Entwicklung Beiträge leisten kann (und sollte):

- Wie wirkt sich Geometrieunterricht auf die allgemeine *Raumvorstellung* aus, wenn die implizite Förderung von *Raumvorstellung* zwar bewusst und intensiv gestaltet, auf zusätzliche explizite Maßnahmen aber verzichtet wird? Hier geht es um die Variation vorhandenen Aufgabenmaterials im Sinne der impliziten Förderung von *Raumvorstellung*, ohne dass andere fachliche Gegenstände weniger intensiv thematisiert werden.
- Wie wirkt sich die allgemeine Förderung von *Raumvorstellung* auf *Mathematikleistung* aus? Es gibt zwar – wie in der vorliegenden Arbeit – Befunde zum statistischen Zusammenhang, nicht aber darüber, ob eine (langfristig angelegte) gezielte Förderung der allgemeinen *Raumvorstellung* sich in der Breite positiv auf *Mathematikleistung* auswirkt.

Beide Fragen können seitens der Evaluation mit realistischem Aufwand mit Experimental- und Kontrollgruppen sowie Vor- und Nachtests praktisch umgesetzt werden. Die inhaltliche Vorbereitung der Interventionen ist auch gut zu leisten – wichtig, und nicht unproblematisch, ist eine gezielte Planung und Durchführung der Interventionen mit den jeweiligen Fachlehrkräften, aber auch das ist realisierbar.

7.4 Ausblick

Am Ende einer umfassenden wissenschaftlichen Arbeit, die neben neuen oder aktuellen Befunden natürlich viele offene, interessante und relevante Fragen entdeckt, kann man schnell ein breites Feld dessen, was künftig beforscht werden könnte oder sollte, darstellen. Die vorliegende Arbeit soll aber mit bescheideneren Perspektiven beendet werden, die hoffentlich hohe Realisierungschancen haben.

Die in Kap. 7.2 und Kap. 7.3 genannten möglichen – und aus Sicht der vorliegenden Arbeit sinnvollen – Konsequenzen für die empirische Bildungsforschung und die Mathematikdidaktik sind leistbar und sollten nach Möglichkeit umgesetzt werden. Dies betrifft insbesondere die Ausdifferenzierung von Rahmenmodellen für die Erforschung von *Mathematikleistung* und die differenzierten Auswertungsperspektiven. Die Interventionsstudien, die als mögliche Konsequenzen für die Mathematikdidaktik vorgeschlagen wurden, sind zwar gut realisierbar, bedürfen aber entsprechender (nicht unrealistischer) Ressourcen.

Die drängendste Frage, die sich aus den Befunden der vorliegenden Arbeit ergibt, und deren Beforschung gut realisierbar ist, lautet aber: Was ist das Gemeinsame von *mentaler Rotation* und *rechnerischen Modellierungsaufgaben*?

Literaturverzeichnis

Amthauer, R. (1953). *Intelligenz-Struktur-Test (I-S-T).* Göttingen: Hogrefe.

Amthauer, R. (1973). *Intelligenz-Struktur-Test (I-S-T 70).* Göttingen: Hogrefe.

Artelt, C., Demmrich, A. & Baumert, J. (2001). Selbstreguliertes Lernen. In Deutsches PISA-Konsortium (Hg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 271-299). Opladen: Leske + Budrich.

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In Deutsches PISA-Konsortium (Hg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69-140). Opladen: Leske + Budrich.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2008). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung.* 12. Auflage. Berlin u. a.: Springer.

Baron, R. M. & Kenny, D. A (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology, 51 (6)*, 1173-1182.

Barratt, E. S. (1953). An Analysis of Verbal Reports of Solving Spatial Problems as an Aid in Defining Spatial Factors. *The Journal of Psychology, 36 (2)*, 17-25.

Baumert, J. (2002). Deutschland im internationalen Bildungsvergleich. In N. Kilius, J. Kluge & L. Reisch (Hg.), *Die Zukunft der Bildung* (S. 100-150). Frankfurt am Main: Suhrkamp.

Baumert, J., Bos, W. & Lehmann, R. (Hg.) (2000a). *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit.* Opladen: Leske+Budrich.

Baumert, J., Bos, W. & Lehmann, R. (Hg.) (2000b). *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe.* Opladen: Leske+Budrich.

Baumert, J., Bos, W. & Watermann, R. (2000). Mathematische und naturwissenschaftliche Grundbildung im internationalen Vergleich. In J. Baumert, W. Bos & R. Lehmann (Hg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit* (S. 135-198). Opladen: Leske+Budrich.

Baumert, J., Köller, O., Lehrke, M. & Brockmann, J. (2000). Anlage und Durchführung der Dritten Internationalen Mathematik- und Naturwissenschaftsstudie zur Sekundarstufe II (TIMSS/III) – Technische Grundlagen. In J. Baumert, W. Bos & R. Lehmann (Hg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit* (S. 31-84). Opladen: Leske+Budrich.

Baumert, J., Kunter, M., Brunner, M., Krauss, S., Blum, W. & Neubrand, M. (2004). Mathematikunterricht aus Sicht der PISA-Schülerinnen und -Schüler und ihrer Lehrkräfte. In PISA-Konsortium Deutschland (Hg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (S. 314-354). Münster: Waxmann.

Baumert, J. & Lehmann, R. (Hg.) (1997). *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde.* Opladen: Leske + Budrich.

Baumert, J., Stanat, P. & Demmrich, A. (2001). PISA 2000: Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In Deutsches PISA-Konsortium (Hg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 15-68). Opladen: Leske+Budrich.

Baumert, J., Trautwein, U. & Artelt, C. (2003). Schulumwelten – Institutionelle Bedingungen des Lehrerns und Lernens. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann & M. Weiß (Hg.), *Pisa 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 261-331). Opladen: Leske + Budrich.

Bender, P. (2005). Die etwas andere Sicht auf PISA sowie TIMSS und IGLU. *Beiträge zum Mathematikunterricht, 2004*, 81-84.

Bennett, G. K., Seashore, H. G. & Wesman, A. G. (1973). *Differential Aptitude Test, Forms S and T*. New York: The psychological Corporation.

Besuden, H. (1973). Zur Raumgeometrie in der Schule. *Westermanns Pädagogische Beiträge, 25 (7)*, 390-393.

Besuden, H. (1979). Die Förderung der Raumvorstellung im Geometrieunterricht. *Beiträge zum Mathematikunterricht*, 64-67.

Birkel, P., Schein, S. A. & Schumann, H. (2002). *BST. Bausteine-Test. Ein Test zur Erfassung des räumlichen Vorstellungsvermögens*. Göttingen: Hogrefe.

Bishop, A. J. (1983). Space and Geometry. In R. Lesh & M. Landau (Hg.), *Acquisition of Mathematics Concepts and Processes* (S. 176-204). New York: Academic Press.

BLK (Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung) (Hg.) (1997). *Gutachten zur Vorbereitung des Programms „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts“*. Bonn: Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung.

Blum, W., Drüke-Noe, C., Hartung. R. & Köller, O. (Hg.) (2006). *Bildungsstandards Mathematik: Konkret – Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen*. Berlin: Cornelsen-Scriptor.

Blum, W., Drüke-Noe, C., Leiss, D., Wiegand, B. & Jordan, A. (2005). Zur Rolle von Bildungsstandards für die Qualitätsentwicklung im Mathematikunterricht. *Zentralblatt für Didaktik der Mathematik, 37 (4)*, 267-274.

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

Bonsen, M., Büchter, A. & Ophuysen, S. van (2004). Im Fokus: Leistung. Zentrale Aspekte der Schulleistungsforschung und ihre Bedeutung für die Schulentwicklung. In H.G. Holtappels, K. Klemm, H. Pfeiffer, H.-G. Rolff & R. Schulz-Zander (Hg.), *Jahrbuch der Schulentwicklung. Band 13. Daten, Beispiele und Perspektiven* (S. 187-223). Weinheim u. a.: Juventa.

Boring, E. G. (1923). Intelligence as the Tests Test it. *New Republic, 35,* 35-37.

Borneleit, P., Dankwerts, R., Henn, H.-W. & Weigand, H.-G. (2001). Expertise zum Mathematikunterricht in der gymnasialen Oberstufe. *Journal für Mathematikdidaktik, 22 , (1*), 73-90.

Borromeo Ferri, R. (2004). *Mathematische Denkstile. Ergebnisse einer empirischen Studie.* Hildesheim: Franzbecker.

Brachinger, W. & Ost, F. (1996). Modelle mit latenten Variablen: Faktorenanalyse, Latent-Structure-Analyse und LISREL-Analyse. In L. Fahrmeir, A. Hamerle & G. Tutz (Hg.), *Multivariate statistische Verfahren* (2., erweiterte Auflage, S. 639-766). Berlin: de Gruyter.

Bruder, R., Leuders, T. & Büchter, A. (2008). *Mathematikunterricht entwickeln. Bausteine für ein kompetenzorientiertes Unterrichten.* Berlin: Cornelsen Scriptor.

Brunner, M. (2006). *Mathematische Schülerleistung: Struktur, Schulformunterschiede und Validität.* Berlin: Max-Planck-Institut für Bildungsforschung. (Dissertation) (Quelle: http://edoc.hu-berlin.de/dissertationen/brunner-martin-2006-02-08/HTML/; 24.05.2010)

Büchter, A. (2004). Die Wissenschaft hat festgestellt …! Wie man sich vor Fehlschlüssen (nicht nur) in der Bildungsforschung wappnet. In G. Eikenbusch & T. Leuders (Hg.), *Lehrer-Kursbuch Statistik* (S. 103-107). Berlin: Cornelsen Scriptor.

Büchter, A. & Henn, H.-W. (2004). Stochastische Modellbildung aus unterschiedlichen Perspektiven. Von der Genueser Lotterie über Urnenaufgaben zur Keno Lotterie. *Stochastik in der Schule, 24 (3)*, 28-41.

Büchter, A. & Henn, H.-W. (2007). *Elementare Stochastik. Eine Einführung in die Mathematik der Daten und des Zufalls. 2., überarbeitete und erweiterte Auflage.* Berlin/Heidelberg: Springer.

Büchter, A. & Henn, H.-W. (Hg.) (2008). *Der Mathekoffer. Mathematik entdecken mit Materialien und Ideen für die Sekundarstufe I.* Seelze/Velber: Friedrich Verlag.

Büchter, A. & Leuders, T. (2005a). From students' achievement to the development of teaching: requirements for the feedback in comparative tests. *Zentralblatt für Didaktik der Mathematik, 37 (4)*, 324-334.

Büchter, A. & Leuders, T. (2005b). *Mathematikaufgaben selbst entwickeln. Lernen fördern – Leistung überprüfen.* Berlin: Cornelsen Scriptor.

Büchter, A., Leuders, T. & Bruder, R. (Hg.) (2005). Quality development in mathematics education by focussing on the outcome: new answers or new questions? *Zentralblatt für Didaktik der Mathematik, 37 (4)*.

Burnett, S. A., Lane, D. M. & Dratt, L. M. (1979). Spatial Visualiziation and Sex Differences in Quantitative Ability. *Intelligence, 3*, 345-354.

Carroll, J. B. (1993). *Human Cognitive Abilities. A Survey of Factor-Analytic Studies.* Cambridge, UK: Cambridge University Press.

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. Past, present, and future. In D. P. Flanagan & P. L. Harrison (Hg.), *Contemporary intellectual assessment. Theories, tests, and issues* (2. Auflage, S. 69–76). New York: Guilford.

Carstensen, C. H., Frey, A., Walter, O. & Knoll, S. (2007). Technische Grundlagen des dritten internationalen Vergleichs. In PISA-Konsortium Deutschland (Hg.), *PISA '06. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 367-390). Münster: Waxmann.

Cattell, R. B. (1963). Theory of Fluid and Crystallized Intelligence. A Critical Experiment. *Journal of Educational Psychology, 54 (1)*, 1-22.

Cattell, R. B. (1971). *Abilities. Their Structure, Growth, and Action.* Boston: Houghton Mifflin.

Collaer, M. L. & Nelson, J. D. (2002). Large Visuospatial Sex Difference in Line Judgment: Possible Role of Attentional Factors. *Brain and Cognition, 49*, 1-12.

Delgado, A. R. & Prieto, G. (1997). Mental rotation as a mediator for sex-related differences in visualization. *Intelligence, 24 (3)*, 393-479

Deutsches PISA-Konsortium (Hg.) (2001). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich.* Opladen: Leske+Budrich.

Dicker, H. (1977). Untersuchungen zur Beurteilung von Mathematikaufgaben. In K. Ingenkamp (Hg.), *Schüler- und Lehrerbeurteilung* (S. 171-193). Weinheim: Beltz.

Dickhäuser, O. (2006). Editorial zum Themenschwerpunkt. Fähigkeitsselbstkonzepte. Entstehung, Auswirkung, Förderung. *Zeitschrift für Pädagogische Psychologie, 20 (1/2)*, 5-8.

Dorst, J. (2006). *Identifizierung von eloquenten Kortexarealen der nicht-dominanten Hemisphäre mittels funktioneller transkranieller Dopplersonographie: Überprüfung einer mentalen Rotationsaufgabe und eines Memoryparadigmas.* Marburg: Philipps-Universität. (Dissertation)

Eccles, J., Adler, T. F., Futterman, F., Goff, S. B., Kaczala, C. M., Meece, J. L. & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Hg.), *Achievement and achievement motives. Psychological ans sociological approaches* (S. 78-146). San Francisco: W. H. Freeman & Co.

El Koussy, A. A. H. (1935). An Investigation into the Factors in Tests Involving the Visual Perception of Space. *British Journal of Psychology, Monograph Supplement No. 20*.

Ethington, C. A. (1992). Gender differences in a psychological model of mathematics achievement. *Journal for Research in Mathematics Education, 23 (2)*, 166-181.

Fahrmeir, L., Hamerle, A. & Nagl, N. (1996). Varianz- und Kovarianzanalyse. In L. Fahrmeir, A. Hamerle & G. Tutz (Hg.), *Multivariate statistische Verfahren* (2., erweiterte Auflage; S. 169-238). Berlin: de Gruyter.

Fahrmeir, L., Hamerle, A. & Tutz, G. (Hg.) (1996). *Multivariate statistische Verfahren.* 2., erweiterte Auflage. Berlin: de Gruyter.

Fay, E. (1992). Dreidimensionaler Würfeltest 3DW. Ein Rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens. *Diagnostica, 38 (2)*, 171-175.

Fend, H. (1980). *Theorie der Schule.* München: Urban & Schwarzenberg.

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

Fleischer, J., Wirth, J. & Leutner, D. (2007). Testmethodische Grundlagen der Lernstandserhebungen NRW. In MSW (Ministerium für Schule und Weiterbildung des Landes Nordhrein-Westfalen) (Hg.), *Lernstandserhebungen Mathematik in Nordrhein-Westfalen. Impulse zum Umgang mit zentralen Tests* (S. 91-113). Stuttgart: Klett Schulbuchverlage.

Frey, A., Asseburg, R., Carstensen, C. H., Ehmke, T. & Blum, W. (2007). Mathematische Kompetenz. In PISA-Konsortium Deutschland (Hg.),. *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (S. 249-276). Münster: Waxmann.

Gallin, P. (2003). Prädikatives und funktionales Denken in der Wahrscheinlichkeitsrechnung. *Zentralblatt für Didaktik der Mathematik, 35 (3)*, 110-119.

Galton, F. (1880). Statistics of mental imagery. *Mind, 5,* 300-318. (Quelle: http://psychclassics.asu.edu/Galton/imagery.htm; 21.04.2010)

Galton, F. (1883). *Inquiries into Human Faculty and its Development*. London: Macmillan. (Quelle: http://galton.org/books/human-faculty/index.html; 21.04.2010)

Gardner, H. (1983). *Frames of mind. The theory of multiple intelligences.* New York: Basic Books.

Gardner, H. (2006). *Multiple intelligences. New horizons.* New York: Basic Books.

Geiser, C., Lehmann, W. & Eid, M. (2006). Separating 'Rotators' from 'Nonrotators' in the Mental Rotations Test: A multigroup latent class analysis. *Multivariate Behavioral Research, 41*, 261-293.

Geiser, C., Lehmann, W., & Eid, M. (2008). A note on sex differences in mental rotation in different age groups. *Intelligence, 36*, 556-563.

Gittler, G. (1990). *3 DW. Dreidimensionaler Würfeltest. Ein rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens. Theoretische Grundlagen und Manual*. Weinheim: Beltz Test.

Goldstein, D., Haldane, D. &Mitchell, C. (1990). Sex differences in visual-spatial ability: The role of performance factors. *Memory and Cognition, 18*, 546–550.

Granzer, D, Köller, O., Bremerich-Vos, A., van den Heuvel-Panhuizen, M., Reiss, K. & Walther, G. (Hg.) (2009). *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule*. Weinheim u. a.: Beltz.

Grüßing, M. (2002). Wieviel Raumvorstellung braucht man für Raumvorstellungsaufgaben? Strategien von Grundschulkindern bei der Bewältigung räumlich-geometrischer Anforderungen. *Zentralblatt für Didaktik der Mathematik, 34 (2)*, 37-45.

Grüßing, M. (2005). Räumliche Kompetenzen und Mathematikleistung. *Sache-Wort-Zahl, 33 (71)*, 41-48.

Guay, R. B. & McDaniel, E. D. (1977). The relationship between mathematics achievement and spatial abilities among elementary school children. *Journal for Research in Mathematics Education, 8(3)*, 211-215.

Guilford, J. P. (1956). The Structure of Intellect. *Psychological Bulletin, 53*, 267-293.

Guilford, J. P. (1959). Three Faces of Intellect. *American Psychologist, 14*, 469-479.

Guilford, J. P. (1967). *The Nature of Human Intelligence.* New York, NY: McGraw-Hill.

Haertel, G. D., Walberg, H. J., Weinstein, T. (1983). Psychological models of educational performance. A theoretical sythesis of constructs. *Review of Educational Research, 53*, 75-92.

Halpern, D. F. (2000). *Sex differences in cognitive abilities.* 3. Auflage. Hillsdale, NJ: Erlbaum.

Hartmann, J. &Reiss, K. (2000). Auswirkungen der Bearbeitung räumlich-geometrischer Aufgaben auf das Raumvorstellungsvermögen. In D. Leutner & R. Brünken (Hg.), *Neue Medien in Unterricht, Aus- und Weiterbildung* (S. 85-93). Münster: Waxmann.

Hefendehl-Hebeker, L. (2003). Didaktik der Mathematik als Wissenschaft – Aufgaben, Chancen, Profile. *Jahresbericht der DMV, 105 (1),* 3-29.

Heil, M., & Jansen-Osmann, P (2008a). Gender differences in math and mental rotation accuracy but not in mental rotation speed in 8 years old children. *European Journal of Developmental Science, 2*, 195-201.

Heil, M. & Jansen-Osmann, P. (2008b). Sex differences in mental rotation with polygons of different complexity: Do men utilize holistic processes whereas women prefer piecemeal ones? *The Quarterly Journal of Experimental Psychology, 61 (5)*, 683-689.

Hellmich, F. & Hartmann, J. (2002). Aspekte einer Förderung räumlicher Kompetenzen im Geometrieunterricht. Ergebnisse einer Trainingsstudie mit Sonderschülerinnen und -schülern. *Zentralblatt für Didaktik der Mathematik, 34 (2)*, 56-61.

Helmke, A. & Schrader, F.-W. (2006). Determinanten der Schulleistung. In D. H. Rost (Hg.), *Handwörterbuch Pädagogische Psychologie* (3., überarbeitete und erweiterte Auflage, S. 83-94). Weinheim: Beltz PVU.

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Hg.), *Psychologie des Unterrichts und der Schule* (S. 71-176). Göttingen: Hogrefe.

Herzberg, F. & Lepkin, M. (1954). A Study of Sex Differences on the Primary Mental Abilities Test. *Educational and Psychological Measurement, 14*, 687-689.

Heymann, H.-W. (1996). *Allgemeinbildung und Mathematik.* Weinheim u. a.: Beltz.

Heymann, H.-W. & Pallack, A. (2007). Aufgabenkonstruktion für die Lernstandserhebung Mathematik. In MSW (Ministerium für Schule und Weiterbildung des Landes Nordhrein-Westfalen) (Hg.), *Lernstandserhebungen Mathematik in Nordrhein-Westfalen. Impulse zum Umgang mit zentralen Tests* (S. 14-46). Stuttgart: Klett Schulbuchverlage.

Hopmann, S., Brinek, G. & Retzl, M. (Hg.) (2007). *PISA zufolge PISA – PISA According to PISA. Hält PISA, was es verspricht? – Does PISA keep, what it promises?* Wien: LIT-Verlag.

Horn, W. (1962). *Leistungsprüfungssystem LPS*. Göttingen: Hogrefe.

Hosenfeld, I., Strauss, B. & Köller, O. (1997). Geschlechtsdifferenzen bei Raumvorstellungsaufgaben – eine Frage der Strategie? *Zeitschrift für Pädagogische Psychologie, 11 (2)*,. 85-94.

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

Jablonka, E. (2007). Mathematical Literacy. Die Verflüchtigung eines ambitionierten Testkonstrukts. In T. Jahnke & W. Meyerhöfer (Hg.), *Pisa & Co. Kritik eines Programms.* (2. Auflage, S. 247-280). Hildesheim: Franzbecker.

Jäger, A. O. (1967). *Dimensionen der Intelligenz.* Göttingen: Hogrefe.

Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau, 35*, 21–35.

Jäger, A. O., Süß, H.-M. & Beauducel, A. (1996). *Test für das Berliner Intelligenzstrukturmodell (BIS).* Göttingen: Hogrefe.

Jahnke, T. & Meyerhöfer, W. (Hg.) (2007). *PISA & Co. Kritik eines Programms.* 2. Auflage. Hildesheim: Franzbecker.

Jansen-Osmann, P., & Heil, M. (2007). Suitable stimuli to obtain (no) gender differences in the speed of cognitive processes involved in mental rotation. *Brain and Cognition, 64*, 217–227.

Jensen, A. R. (1998). *The g factor.* New York, NY: Praeger.

Kaune, C. (2003). Das Wissen um Unterschiede in den kognitiven Strukturen von Schülerinnen und Schülern als Erklärung von Unterrichtsbeiträgen. *Zentralblatt für Didaktik der Mathematik, 35 (3)*, 102-109.

Kelava, A. & Moosbrugger, H. (2008). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilung. In H. Moosbrugger & A. Kelava (Hg.), *Testtheorie und Testkonstruktion* (S. 73-98). Berlin: Springer.

Klauer, K. J. (2006a). Anlage und Umwelt. In D. H. Rost (Hg.), *Handwörterbuch Pädagogische Psychologie* (3., überarbeitete und erweiterte Auflage, S. 8-14). Weinheim: Beltz PVU.

Klauer, K. J. (2006b). Intelligenz und Begabung. In D. H. Rost (Hg.), *Handwörterbuch Pädagogische Psychologie* (3., überarbeitete und erweiterte Auflage, S. 275-280). Weinheim: Beltz PVU.

Klieme, E. (1985). Mathematisches Grundverständnis – Psychometrische Analysen und kognitionspsychologische Explorationsstudien. In G. Trost, F. Blum, E. Fay, A. Hengsen, E. Klieme, U. Maickle, H.-U. Nauels & H. Stumpf (Hg.), *Modellversuch „Tests für medizinische Studiengänge“ (TMS). Zehnter Arbeitsbericht: 1. April 1984 bis 30. September 1985.* Bonn: Institut für Test- und Begabungsforschung.

Klieme, E. (1986). Bildliches Denken als Mediator für Geschlechterunterschiede beim Lösen mathematischer Probleme. In H.-G. Steiner (Hg.), *Grundfragen der Entwicklung mathematischer Fähigkeiten. IDM-Reihe Band 13* (S. 133-151). Köln: Aulis Verlag Deubner.

Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physiktunterricht: Theoretische Grundlage, Kompetenzstufen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos & R. Lehmann (Hg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2:Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (S. 57-128). Opladen: Leske+Budrich.

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

Klieme, E., Neubrand, M. & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse. In Deutsches PISA-Konsortium (Hg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 139-190). Opladen: Leske+Budrich.

KMK (Kultusministerkonferenz) (2004). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss. Beschluss vom 4.12.2003.* München: Luchterhand.

KMK (Kultusministerkonferenz) (2005a): *Bildungsstandards im Fach Mathematik für den Hauptschulabschluss. Beschluss vom 15.10.2004.* München: Luchterhand.

KMK (Kultusministerkonferenz) (2005b): *Bildungsstandards im Fach Mathematik für den Primarbereich. Beschluss vom 15.10.2004.* München: Luchterhand.

Knoche, N., Lind, D., Blum, W., Cohors-Fresenborg, E., Flade, L., Loeding, W., Moeller, G., Neubrand, M. & Wynands, A. (2002). Die PISA-2000-Studie, einige Ergebnisse und Analysen. *Journal für Mathematik-Didaktik, 23 (3/4)*, 159-202.

Knoll, S. (1998) Anforderungsgestaltung im Mathematikunterricht. *mathematik lehren, 90*, 47-51.

Köller, O. (1998). *Zielorientierung und schulisches Lernen.* Münster: Waxmann.

Köller, O. & Baumert, J. (2002). Entwicklung schulischer Leistungen. In R. Oerter & L. Montada (Hg.), *Entwicklungspsychologie* (5., vollständig überarbeitete Auflage, S. 756-786). Weinheim: Beltz.

Köller, O. & Klieme, E. (2000). Geschlechterdifferenzen in den mathematisch-naturwissenschaftlichen Leistungen. In J. Baumert, W. Bos & R. Lehmann (Hg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (S. 373-404). Opladen: Leske+Budrich.

Köller, O., Knigge, M. & Tesch, B. (Hg.) (2010). *Sprachliche Kompetenzen im Ländervergleich Befunde des ersten Ländervergleichs zur Überprüfung der Bildungsstandards für den Mittleren Schulabschluss in den Fächern Deutsch, Englisch und Französisch. Zusammenfassung.* Berlin: Institut für Qualitätsentwicklung im Bildungswesen.

Köller, O. & Möller, J. (2006). Selbstwirksamkeit. In D. H. Rost (Hg.), *Handwörterbuch Pädagogische Psychologie* (3., überarbeitete und erweiterte Auflage, S. 693-698). Weinheim: Beltz PVU.

Köller, O., Rost, J. & Köller, M. (1994). Individuelle Unterschiede beim Lösen von Raumvorstellungsaufgaben aus dem IST- bzw. IST-70-Untertest „Würfelaufgaben“. *Zeitschrift für Psychologie, 202 (1)*, 65-85.

Koops, H. & Sorger, P. (1980). *Fallstudien zum mathematischen Fähigkeitsfaktor Räumliches Vorstellungsvermögen bei sechs- bis achtjährigen Schülern.* Opladen: Westdeutscher Verlag.

Koops, H., Mosel-Göbel, D. & Sorger, P. (1981). *Zur Interpretation und Konstruktion räumlicher Konfigurationen und ihrer ebenen Darstellungen. Entwicklungsstand und Entwicklungsverläufe bei Grundschülern.* Opladen: Westdeutscher Verlag.

Krauss, S., Neubrand, M., Blum, W., Baumert, J., Brunner, M., Kunter, M. & Jordan, A. (2008). Die Untersuchung des professionellen Wissens deutscher Mathematik-Lehrerinnen und -Lehrer im Rahmen der COACTIV-Studie. *Journal für Mathematik-Didaktik, 29 (3/4)*, 223-258.

Kunter, M., Dubberke, T., Baumert, J., Blum, W., Brunner, M., Jordan, A., Klusmann, U., Krauss, S., Löwen, K., Neubrand, M., & Tsai, Y.-M. (2006). Mathematikunterricht in den PISA-Klassen 2004: Rahmenbedingungen, Formen und Lehr-Lernprozesse. In PISA-Konsortium Deutschland (Hg.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (S. 161-196). Münster: Waxmann.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J. & Weiß, M. (2002). *PISA 2000. Dokumentation der Erhebungsinstrumente.* Berlin: Max-Planck-Institut für Bildungsforschung.

Krüger, K. (2000). *Erziehung zum funktionalen Denken. Zur Begriffsgeschichte eines didaktischen Prinzips.* Berlin: Logos.

Lange, H. (1999). Qualitätssicherung in Schulen. *Die Deutsche Schule, 91 (2),* 144-159.

Lehmann, W. & Jüling, I. (2002). Raumvorstellungsfähigkeit und mathematische Fähigkeiten – unabhängige Konstrukte oder zwei Seiten einer Medaille? *Psychologie in Erziehung und Unterricht, 49 (1)*, 31-43.

Lehmann, W., Jüling, I. & Knopf, H. (2002). Allgemeine und domänenspezifische kognitive Leistungen. Eine vergleichende Untersuchung in mathematisch und sprachlich orientierten Gymnasien. *Zeitschrift für Pädagogische Psychologie, 16*, 29-41.

Leopold, C. (2002). Untersuchungen zum Raumvorstellungsvermögen bei Studierenden der Ingenieurwissenschaften und der Mathematik unter geschlechtsvergleichender Perspektive. In L. Hermes, A. Hirschen & I. Meißner, I. (Hg.), *Gender und Interkulturalität. Ausgewählte Beiträge der 3. Fachtagung Frauen-Gender-Forschung in Rheinland-Pfalz*. Tübingen: Stauffenberg

Liben, L. S. (1978). Performance on Piagetian spatial tasks as a function of sex, field dependence, and training. *Merrill-Palmer quarterly, 24 (2)*, S. 97-110.

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

Lind, D. (2004). Welches Raten ist unerwünscht? Eine Erwiderung. *Journal für Mathematik-Didaktik, 25 (1)*, 70-74.

Linn, M. A. & Petersen, A. C. (1985). Emergence and Characterization of Sex Differences in Spatial Ability. A Meta-Analysis. *Child Development, 56 (6),* S. 1479-1498.

Linn, M. A. & Petersen, A. C. (1986). A Meta-Analysis of Gender Differences in Spatial Ability: Implications for Mathematics and Science Achievement. In S. Hyde & M. C. Linn (Hg.), *The Psychology of Gender. Advances Through Meta-Analysis* (S. 67-101). Baltimore: John Hopkins University.

Lohaus, A., Schumann-Hengsteler, R. & Kessler, T. (1999). *Räumliches Denken im Kindesalter.* Göttingen: Hogrefe.

Lompscher, J. (Hg.) (1988). *Persönlichkeitsentwicklung in der Lerntätigkeit*. Berlin: Volk und Wissen.

Lorenz, J. H. (1991). Rechenschwache Schüler in der Grundschule – Erklärungsversuche und Förderstrategie – Teil 1. *Journal für Mathematik-Didaktik, 12*, 3-34.

Lüdtke, O., Köller, O., Artelt, C., Stanat, P., & Baumert, J. (2002). Eine Überprüfung von Modellen zur Genese akademischer Selbstkonzepte: Ergebnisse aus der PISA-Studie. *Zeitschrift für Pädagogische Psychologie, 16 (3/4),* 151–164.

Lüthje, T. (2008). Räumliche Fähigkeiten von Kindern im Vorschulalter. Untersuchungsdesign und erste Ergebnisse. *Beiträge zum Mathematikunterricht 2008*, 581-584.

Lüthje, T. (2009). Geschlechtsspezifische Unterschiede im Vorschulalter bei der Bearbeitung von Raumvorstellungsaufgaben. *Beiträge zum Mathematikunterricht 2009*, 391-394.

Maier, P. H. (1994). *Räumliches Vorstellungsvermögen.* Frankfurt a. M: Peter Lang.

Maier, P. H. (1996). Die Relevanz der Raumvorstellung. *Pädagogische Rundschau, 50 (6)*, 745-751.

Maier, P. H. (1999a). Raumgeometrie und Raumvorstellung. Thesen zur Neustrukturierung des Geometrieunterrichts. *Der Mathematikunterricht, 45 (3)*, 4-18.

Maier, P. H. (1999b). *Räumliches Vorstellungsvermögen*. Donauwörth: Auer.

Manger, T. & Eikeland, O.-J. (1998). The effects of spatial visualization and students' sex on mathematical achievement. *British Journal of Psychology, 89 (1)*, S. 17-25.

McGee, M. G. (1979). *Human Spatial Abilities. Sources of Sex Differences.* New York: Praeger.

Meili, R. (1944). Grundlegende Eigenschaften der Intelligenz. *Schweizerische Zeitschrift für Psychologie und ihre Anwendungen, 2,* 166-175 u. 265-271.

Meili, R. (1964). Die faktorenanalytische Interpretation der Intelligenz. *Schweizerische Zeitschrift für Psychologie und ihre Anwendungen, 23,* 135-155.

Meißner, H. (2004). Messen von Kompetenzen – nur genau eine richtige Antwort? *Journal für Mathematik-Didaktik, 25 (3/4)*, 306.

Meißner, H. (2006). Projekt „DORF". Raumvorstellungen verbessern. *Journal für Mathematik-Didaktik, 27 (1),* 28-51.

Meyerhöfer, W. (2004). Zum Problem des Ratens bei PISA. *Journal für Mathematik-Didaktik, 25 (1)*, 62-69.

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

Michael, W. B., Guilford, J. P., Fruchter, B. & Zimmerman, W. S. (1957). The Description of Spatial-Visualization Abilities. *Educational and Psychological Measurement, 17 (2),* 185-199.

Moschner, B.(2001). Selbstkonzept. In D. H. Rost (Hg.), *Handwörterbuch Pädagogische Psychologie* (2., überarbeitete und erweiterte Auflage, S. 629-634). Weinheim: Beltz PVU.

Moschner, B. & Dickhäuser, O. (2006). Selbstkonzept. In D. H. Rost (Hg.), *Handwörterbuch Pädagogische Psychologie* (3., überarbeitete und erweiterte Auflage, S. 275-280). Weinheim: Beltz PVU.

MSJK (Ministerium für Schule, Jugend und Kinder des Landes Nordrhein-Westfalen) (Hg.) (2004a). *Kernlehrplan für das Gymnasium – Sekundarstufe I in Nordrhein-Westfalen – Mathematik.* Frechen: Ritterbach.

MSJK (Ministerium für Schule, Jugend und Kinder des Landes Nordrhein-Westfalen) (Hg.) (2004b). *Kernlehrplan für die Gesamtschule – Sekundarstufe I in Nordrhein-Westfalen – Mathematik.* Frechen: Ritterbach.

MSJK (Ministerium für Schule, Jugend und Kinder des Landes Nordrhein-Westfalen) (Hg.) (2004c). *Kernlehrplan für die Hauptschule– Sekundarstufe I in Nordrhein-Westfalen – Mathematik.* Frechen: Ritterbach.

MSJK (Ministerium für Schule, Jugend und Kinder des Landes Nordrhein-Westfalen) (Hg.) (2004d). *Kernlehrplan für die Realschule – Sekundarstufe I in Nordrhein-Westfalen – Mathematik.* Frechen: Ritterbach.

Neubrand, M. (2001). PISA – „Mathematische Grundbildung"/„mathematical literacy" als Kern einer internationalen und nationalen Leistungsstudie. In G. Kaiser, N. Knoche, D. Lind, & W. Zillmer (Hg.), *Leistungsvergleiche im Mathematikunterricht. Ein Überblick über aktuelle Studien* (S. 177-194). Hildesheim: Franzbecker.

Neubrand, M., Biehler, R., Blum, W., Cohors-Fresenborg, E., Flade, L., Knoche, N., Lind, D., Löding, W., Möller, G. & Wynands, A. (2001). Grundlagen der Ergänzung des internationalen PISA-Mathematik-Tests in der deutschen Zusatzerhebung. *Zentralblatt für Didaktik der Mathematik, 33(1)*, 45-59.

Neubrand, M., Klieme, E., Lüdtke, O., & Neubrand, J. (2002). Kompetenzstufen und Schwierigkeitsmodelle für den PISA-Test zur mathematischen Grundbildung. *Unterrichtswissenschaft, 30*, 100–119.

OECD (Organisation for Economic Co-operation and Development) (1999). *Measuring student knowledge and skills. A new framework for assessment.* Paris: OECD.

OECD (Organisation for Economic Co-operation and Development) (Hg.) (2009). *PISA 2006. Technical Report.* Paris: OECD.

Oehl, W. (1949). Erziehung zur Raumanschauung in der Grundschule. *Schola, 4*, 66-71.

Pawlik, K. (1968). *Dimensionen des Verhaltens. Eine Einführung in die Methodik und Ergebnisse faktorenanalytischer psychologischer Forschung.* Bern: Hans Huber.

Pekrun, R., vom Hofe, R., Blum, W., Götz, T., Wartha, S., & Jullien, S. (2006). Projekt zur Analyse der Leistungsentwicklung in Mathematik (PALMA) – Entwicklungsverläufe, Schülervoraussetzungen und Kontextbedingungen von Mathematikleistungen bei Schülerinnen und Schülern der Sekundarstufe I. In M. Pernzel & L. Allolio-Näcke (Hg.), *Untersuchungen von Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (S. 21-52). Münster: Waxmann.

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R. & Richardson, C. (1995). A redrawn Vandenberg and Kuse Mental Rotations Test: Different versions and factors that affect performance. *Brain and Cognition, 28*, 39-58.

Peters, M., Lehmann, W., Takahira, S., Takeuchi, Y., Jordan, K. (2006). Mental Rotation Test Performance in four cross-cultural samples (n=3367): Overall Sex Differences and the Role of Academic Program in Performance. *Cortex, 42(7)*, 1005-1014.

Piaget, J. & Inhelder, B. (1971). *Die Entwicklung des räumlichen Denkens beim Kinde*. Stuttgart, Klett.

Pinkernell, G. (2003). *Räumliches Vorstellungsvermögen im Geometrieunterricht. Eine didaktische Analyse mit Fallstudien.* Hildesheim/Berlin: Franzbecker.

PISA-Konsortium Deutschland (Hg.) (2004). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs*. Münster: Waxmann.

PISA-Konsortium Deutschland (Hg.) (2006). *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres*. Münster: Waxmann.

PISA-Konsortium Deutschland (Hg.) (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie.* Münster: Waxmann.

Putz-Osterloh, W. (1977). Über Problemlöseprozesse bei dem Test Würfelaufgaben aus dem Intelligenzstrukturtest IST und IST-70 von Amthauer. *Diagnostica, 23*, 252-265.

Quaiser-Pohl, C. (1998). *Die Fähigkeit zur räumlichen Vorstellung. Zur Bedeutung von kognitiven und motivationalen Faktoren für geschlechtsspezifische Unterschiede.* Münster: Waxmann.

Quaiser-Pohl, C., Lehmann, W. & Eid, M. (2004). The relationship between spatial abilities and representations of large-scale space – a structural equation modeling analysis. *Personality and Individual Differences, 36*, 95-107.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydicke.

Reed, S. (1974). Pattern recognition and categorization. *Memory and Cognition, 2 (2)*, 329-336.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 42-106). Weinheim u. a.: Beltz.

Rost, D. H. (1977). *Raumvorstellung. Psychologische und pädagogische Aspekte.* Weinheim u. a.: Beltz.

Rost, D. H. (2006). *Handwörterbuch Pädagogische Psychologie.* 3., überarbeitete und erweiterte Auflage. Weinheim: Beltz PVU.

Rost, D. H. (2008). Multiple Intelligenzen, multiple Irritationen. *Zeitschrift für Pädagogische Psychologie, 22 (2),* 97-112.

Rost, D. H. (2009). *Intelligenz. Fakten und Mythen.* Weinheim: Beltz PVU.

Rost, D. H., Dickhäuser, O., Sparfeldt, J. R., & Schilling, S. R. (2004). Fachspezifische Selbstkonzepte und Schulleistungen im dimensionalen Vergleich. Eine versuchsplanerische Überprüfung des I/E-Modells. *Zeitschrift für Pädagogische Psychologie, 18*, 43-52.

Schermelleh-Engel, K. & Werner, C. (2008). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hg.), *Testtheorie und Testkonstruktion* (S. 113-134). Berlin: Springer.

Scherer, P. (1999). Mathematiklernen bei Kindern mit Lernschwächen. Perspektiven für die Lehrerbildung. In C. Selter & G. Walther (Hg.), *Mathematikdidaktik als design science. Festschrift für Erich Christian Wittmann* (S. 170-179). Leipzig: Klett.

Schwank, I. (1996). Zur Konzeption prädikativer versus funktionaler kognitiver Strukturen und ihrer Anwendung. *Zentralblatt für Didaktik der Mathematik, 28 (6),* 168-183.

Schwank, I. (1998). *Kognitive Mathematik.* Osnabrück: Forschungsinstitut für Mathematikdidaktik. (eBook; Quelle: http://www.fmd.uni-osnabrueck.de/ebooks.html; 13.06.2010)

Schwank, I. (1999/2000). *QuaDiPF – Qualitatives Diagnoseinstrument für prädikatives versus funktionales Denken. Sets A/B/C/D*. Osnabrück: Forschungsinstitut für Mathematikdidaktik.

Schwank, I. (2003a). Einführung in funktionales und prädikatives Denken. *Zentralblatt für Didaktik der Mathematik, 35 (3*), 70-78.

Schwank, Inge (2003b). Vorwort. *Zentralblatt für Didaktik der Mathematik, 35 (3*), 69.

Shephard, R. N. & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge: Massachusetts Institute of Technology.

Shepard, R. N. & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science, Heft 171*, 701-703.

Sill, H.-D. (2010). Probleme und Erfahrungen mit „Mindeststandards“. *Mitteilungen der Gesellschaft für Didaktik der Mathematik, 88*, 5-11.

Spearman, C. (1904). „General Intelligence”, Objectively Determined and Measured. *American Journal of Psychology 15,* 201-293. (Quelle: http://psychclassics.yorku.ca/Spearman/index.htm; 21.04.2010) (Kap. 3)

Spearman, C. (1923). *The nature of intelligence and the principles of cognition.* London: Macmillan.

Spearman, C. (1927). *The abilities of man.* London: Macmillan.

Stanat, P. & Kunter, M. (2001). Geschlechterunterschiede in Basiskompetenzen. In Deutsches PISA-Konsortium (Hg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 249-269). Opladen: Leske+Budrich.

Stumpf, H. & Fay, E. (1983). *Schlauchfiguren – Ein Test zur Beurteilung des räumlichen Vorstellungsvermögens.* Göttingen: Hogrefe.

Stumpf. H. & Klieme, E. (1989). Sex related differences in spatial ability: More evidence for convergence. *Perceptual and Motor Skills, 69*, 915-921.

Sundermann, B. & Selter, C. (2006). *Beurteilen und Fördern im Mathematikunterricht. Gute Aufgaben. Differenzierte Arbeiten. Ermutigende Rückmeldungen.* Berlin: Cornelsen Scriptor.

Tartre, L. A. & Fennema, E. (1995). Mathematics Achievement and Gender: A Longitudinal Study of Selected Cognitvie ans Affective Variables [Grades 6-12]. *Educational Studies in Mathematics, 28*, 199-217.

Thurstone, L. L. (1934). Vectors of Mind. *Psychological Review, 41,* 1-32. (Quelle: http://psychclassics.asu.edu/Thurstone/; 21.04.2010)

Thurstone, L. L. (1936). The factorial isolation of primary abilities. *Psychometrika, 1,* 175–182.

Thurstone, L. L. (1938). *Primary Mental Abilities.* Chicago: University of Chicago Press.

Thurstone, L. L. (1949). *Mechanical Aptitude III. Analysis of Group Tests.* Psychometric Laboratory Report No. 55. Chicago: University of Chicago Press.

Thurstone, L. L. (1950). *Some primary abilities in visual thinking.* Psychometric Laboratory Report No. 59. Chicago: University of Chicago Press.

Thurstone, L. L. (1955). *The Differential Growth of Mental Abilities.* Psychometric Laboratory Report No. 14. Chapel Hill: University of North Carolina.

Titze, C., Heil, M. & Jansen, P. (2008). Gender Differences in the Mental Rotations Test (MRT) Are Not Due to Task Complexity. *Journal of Individual Differences, 29 (3),* 130-133.

Treumann, K. (1974). *Dimensionen der Schulleistung. Teil 2: Leistungsdimensionen im Mathematikunterricht.* Stuttgart: Klett.

van Hiele, P. M. & van Hiele-Geldorf, D. (1978). Die Bedeutung der Denkebenen im Unterrichtssystem nach der deduktiven Methode. In H. G. Steiner (Hg.), *Didaktik der Mathematik* (S. 127-139). Darmstadt: Wissenschaftliche Buchgesellschaft.

Vandenberg, S. G. & Kuse, A. R. (1978). Mental Rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 47,* 599-604.

Vernon, P. E. (1961). *The structure of human abilites.* 2. Auflage. London: Methuen.

Vollrath H.-J. (1989) Funktionales Denken. *Journal für Mathematikdidaktik, 10 (1)*, 3-37.

von Saldern, M. (1997). *Schulleistung in Deutschland – ein Beitrag zur Standortdiskussion.* Münster: Waxmann.

Voyer, D. & Doyle, R. A. (2010). Item type and gender differences on the Mental Rotations Test. *Learning and Individual Differences, 20 (5)*, 469-472.

Voyer, D. & Hou, J. (2006). Type of items and the magnitude of gender differences on the Mental Rotations Test. *Canadian Journal of Experimental Psychology, 60 (2)*, 91-100.

Voyer, D. & Saunders, K. (2004). *Gender differences on the mental rotations test. A factor analysis. Acta Psychologica, 117 (1)*, 79-94.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*, 250–270.

Wang, M.C., Haertel, G.D. & Walberg, H.J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63 (3)*, 249–294.

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika, 54*, 427-450.

Weinert, F. E. (2001). Schulleistungen – Leistungen der Schule oder der Schüler? In F. E. Weinert (Hg.), *Leistungsmessungen in Schulen* (S. 73-86). Weinheim u. a.: Beltz.

Wiedenbauer, G. (2006). *Manuelles Training mentaler Rotation*. Düsseldorf: Heinrich-Heine-Universität. (Dissertation)

Winkelmann, H. & Robitzsch, A. (2009). Modelle mathematischer Kompetenzen: Empirische Befunde zur Dimensionalität. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 169-196). Weinheim u. a.: Beltz.

Winkelmann, H. & van den Heuvel-Panhuizen, M. (2009). Geschlechtsspezifische mathematische Kompetenz. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 169-196). Weinheim u. a.: Beltz.

Winter, H. (1995). Mathematikunterricht und Allgemeinbildung. *Mitteilungen der Gesellschaft für Didaktik der Mathematik, 61*, 37-46.

Witkin, H. A., Dyk, R. B. & Fatersion, H. F. (1962). *Psychological Differentiation.* New York: Wiley.

Witkin, H. A., Oltman, P., Raskin, E. & Karp, S. (1971). *A manual for the Embedded Figure Tests*. Palo Alto, Calif.: Consulting Psychologists Press.

Wollring, B. (1999). Mathematikdidaktik zwischen Diagnostik und Design. In C. Selter & G. Walther (Hg.), *Mathematikdidaktik als design science. Festschrift für Erich Christian Wittmann* (S. 270-276). Leipzig: Klett.

Woschek, R. (2004). Ein Beitrag zur Diskussion des Rateproblems bei MC-Aufgaben. *Journal für Mathematik-Didaktik, 25 (2)*, 149-152.

Zimmer, K., Burba, D. & Rost, J. (2004). Kompetenzen von Mädchen und Jungen. In PISA-Konsortium Deutschland (Hg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs* (S. 211-224). Münster: Waxmann.

Zimmerman, W. S. (1954). Hypotheses concerning the nature of the spatial factors. *Educational and Psychological Measurement, 14,* 396-400.

Abkürzungsverzeichnis

1-PL	1-parametriges, logistisches Modell (= *RM*)
2 DW	Zweidimensionaler Würfeltest
2-PL	2-parametriges, logistisches Modell
3 DW	Dreidimensionaler Würfeltest (Gittler, 1990)
3-PL	3-parametriges, logistisches Modell
AIC	Akaike Information Criterion
ANOVA	Analysis of Variance – univariate Varianzanalyse
BIC	Bayesian Information Criterion
CAIC	Consistent Akaike Information Criterion
CFA	konfirmatorische Faktorenanalyse („confirmatory factor analysis“)
DAT:SR	Differential Aptitude Test – Subtest Spatial Relations (Bennett et al., 1973)
DIF	Differential Item Functioning
EFT	Embedded Figures Test
EK	Erweiterungskurs
FSK:M	Bereichspezifisches Fähigkeitsselbstkonzept Mathematik
GE	Gesamtschule
GK	Grundkurs
GY	Gymnasium
HS	Hauptschule
ICC	Item Characteristic Curve
IRT	Item Response Theory
I-S-T/I-S-T 70	Intelligenz-Struktur-Test (Amthauer, 1953, 1973)
IST:WÜ	Untertest Würfelaufgaben aus dem I-S-T bzw. I-S-T 70 (s. o.)
KTT	Klassische Testtheorie
LCA	Latent Class Analysis
LSE 9	nordrhein-westfälische Lernstandserhebungen in der Jahrgangsstufe 9
LLTM	Linear-logistisches Testmodell
LRT	Likelihood Ratio Test
MRM	Mixed Rasch-Modell (LCA x RM)
MRT	Mental Rotation Test (Vandenberg & Kuse, 1978; Peters et al., 1995)
QuaDiPF	Qualitatives Diagnoseinstrument für prädikatives versus funktionales Denken (Schwank, 1999/2000)
PISA	Programme for International Student Assessment (Erhebungsjahre 2000, 2003, 2006, 2009 ff.)
PTT	Probabilistische Testtheorie
RFT	Rod and Frame Test (Witkin et al., 1962)

RM	Rasch-Modell (in verschiedenen Varianten: eindimensional / mehrdimensional, zweikategoriell / mehrkategoriell, …)
RMSEA	Root Mean Square Error of Approximation
RS	Realschule
S1	Raumvorstellungsfaktor „Spatial Relations“ nach Thurstone (1950)
S2	Raumvorstellungsfaktor „Visualization“ nach Thurstone (1950
S3	Raumvorstellungsfaktor „Spatial Orientation“ nach Thurstone (1950
SIMS	Second International Mathematics Study
SR-O	Raumvorstellungsfaktor „Spatial Relations and Orientation“ nach Michael et al. (1957) – Zusammenführung von S1 und S3
TIMSS	früher: Third International Mathematics and Science Study (in diesem Sinne wird „TIMSS“ in der vorliegenden Arbeite verwendet); heute: Trends in International Mathematics and Science Study
TIMSS/I	Third International Mathematics and Science Study – Population I: Primarstufe
TIMSS/II	Third International Mathematics and Science Study – Population II: Sekundarstufe I
TIMSS/III	Third International Mathematics and Science Study – Population III: Sekundarstufe II
Vz	Raumvorstellungsfaktor „Visualization“ nach Michael et al. (1957) – weitgehend übereinstimmend mit S2
WLE	Weighted-Likelihood-Estimation / Weighted-Likelihood-Estimates (auch: Warm (Maximum) Likelihood Estimates of Rasch Measures)
WLT	WaterLevel Tasks (Piaget & Inhelder, 1971)

Abbildungsverzeichnis

Tabellenverzeichnis