



universität
wien

Diplomarbeit

Titel der Arbeit

Trainierbarkeit von Raumvorstellung im
Rahmen des Schulunterrichts:
Vergleich probabilistischer Prüfmodelle

Verfasserin

Ramona Knapp

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, im Februar 2012

Studienkennzahl: 298

Studienrichtung: Psychologie

Betreuer: Ao. Univ.-Prof. Dr. Georg Gittler

Inhaltsverzeichnis

Einleitung	4
Theoretischer Teil	5
1. Raumvorstellung	6
1.1 Überblick über die historische Entwicklung des Konstrukts Raumvorstellung	6
1.1.1 Spearman's Zwei-Faktoren-Theorie	6
1.1.2 Thurstone's Primärfaktoren	7
1.1.3 Guilford's Intelligenzstrukturmodell	9
1.1.4 Gardner's Theorie der multiplen Intelligenzen	11
1.2 Definition von Raumvorstellung	12
1.3 Geschlechtsspezifische Unterschiede in der Raumvorstellung	15
1.3.1 Entwicklungspsychologischer Verlauf	15
1.3.2 Biologische Erklärungsansätze der Geschlechtsunterschiede	17
1.3.2.1 Lateralisierung des Gehirns	18
1.3.2.2 Hormonelle Einflüsse	19
1.3.3 Psychische Erklärungsansätze der Geschlechtsunterschiede	20
1.3.3.1 Selbstvertrauen	20
1.3.3.2 Erfahrung	21
1.3.3.3 Lösungsstrategien	23
1.3.4 Weitere Erklärungsansätze	25
1.4 Verbesserungsmöglichkeiten der Raumvorstellung durch Training	26
2. Die Item Response Theorie (IRT)	27
2.1 Klassische Testtheorie vs. Probabilistische Testtheorie	27
2.2 Das dichotome logistische Modell von Rasch	29
2.2.1 Das Lineare logistische Testmodell	32
2.2.2 Veränderungsmessung nach Gittler und Fischer	33

Empirischer Teil	35
3. Zielsetzung und Forschungsfragen	36
4. Planung und Durchführung	37
4.1 Versuchsplan	37
4.2 Erhebungsinstrument	38
4.2.1 Der Dreidimensionale Würfeltest (3DW) von Gittler (1990)	38
4.2.2 Konstruktion des 3DW	39
4.2.3 Vergleich 3DW und I-S-T 70 (Würfelaufgaben)	41
4.3 Datenerhebung	42
4.4 Stichprobenbeschreibung	43
5. Ergebnisdarstellung	44
5.1 Fähigkeits- und Itemparameter	44
5.2 Analysen der Itemparameterdifferenzen mittels klassisch-statistischer Verfahren	47
5.2.1 Leistungsunterschiede zwischen den Testzeitpunkten	47
5.2.2 Leistungsunterschiede zwischen den Testgruppen	48
5.2.3 Geschlechtsspezifische Unterschiede in der Differenz der Fähigkeitsparameter zwischen den Testzeitpunkten	49
5.2.4 Geschlechtsspezifische Unterschiede in den Fähigkeitsparametern des ersten und zweiten Testzeitpunkts	50
5.2.5 Unterschiede in der Bearbeitungszeit zwischen den Testzeitpunkten	52
5.2.6 Unterschiede in der Bearbeitungszeit zwischen den Testgruppen	53
5.2.7 Zusammenhang zwischen Schulnoten und Testleistung	54
5.3 Analysen mittels Linearem logistischen Testmodell (LLTM)	55
5.4 Analysen mittels Veränderungsmessung nach Gittler und Fischer (2011)	60
6. Diskussion	62
Literaturverzeichnis	67
Sekundärliteraturverzeichnis	75
Tabellenverzeichnis	76
Abbildungsverzeichnis	78
Anhang	79
Abstract auf Deutsch	83
Abstract in English	84
Eidesstattliche Erklärung	85
Curriculum Vitae	86

Einleitung

Die vorliegende Arbeit beschäftigt sich mit dem Thema Raumvorstellung und der Verbesserung dieser im Rahmen des schulischen Unterrichts.

Die Bedeutung von Raumvorstellung wird an der enormen Menge an Forschungsarbeiten deutlich, die jährlich zu diesem Thema publiziert werden (Gittler & Fischer, 2011; Glück & Fabrizii, 2010; Titze & Heil, 2010; Tzuriel & Egozi, 2010; Van der Ham & Borst, 2011; Voyer, 2011).

Doch auch in verschiedenen Bereichen des täglichen Lebens wird Raumvorstellung immer wieder genutzt. Dies betrifft unter anderem sportliche Bereiche wie Bogenschießen oder Basketball (Baenninger & Newcombe, 1989), sowie die räumliche Orientierung beim Autofahren (Gittler, 1994).

Da im Rahmen der Entwicklungspsychologie hinreichend bekannt ist, dass es notwendig ist Fertigkeiten des täglichen Lebens bereits in jungen Jahren zu erlernen (Oerter & Montada, 2002), werden Trainingsmaßnahmen zur Raumvorstellung als essentiell erachtet.

Die Arbeit gliedert sich in einen theoretischen und einen empirischen Teil. In Ersterem wird das Konstrukt Raumvorstellung und dessen historische Entwicklung als Teilgebiet der Intelligenz, sowie Ursachen für geschlechtsspezifische Unterschiede in diversen Raumvorstellungstests untersucht. Hierbei wird differenziert zwischen biologischen sowie psychischen Ursachen. Weiterhin werden Möglichkeiten der Verbesserung von Raumvorstellungsleistung durch Training aufgezeigt.

Zusätzlich wird ein Einblick in die Item Response Theorie im Gegensatz zur Klassischen Testtheorie gegeben, wobei im Besonderen das dichotome logistische Modell von Rasch (1960) im Fokus steht.

Der empirische Teil widmet sich der Beschreibung der durchgeführten Studie und der Validierung dreier statistischer Modelle, der Auswertung anhand von Fähigkeitsparametern, anhand des Linearen logistischen Testmodells und der Veränderungsmessung nach Gittler und Fischer (2011). Auswertung und Ergebnisse werden im Verlauf dargestellt und interpretiert.

Theoretischer Teil

1. Raumvorstellung

1.1 Überblick über die historische Entwicklung des Konstrukts Raumvorstellung

Das Konstrukt „Raumvorstellung“ wurde seit Anfang des 20. Jahrhunderts im Zuge diverser Studien zur Intelligenz immer wieder in der psychologischen Forschung untersucht.

Eliot und Smith (1983) teilen die wissenschaftliche Forschung zur Raumvorstellung in drei große Phasen auf. In der Zeit zwischen 1904 und 1938 beschäftigten sich Forscher hauptsächlich mit der Frage, ob das Konstrukt „Raumvorstellung“ existiert oder ob ein globaler Faktor der Intelligenz ausgemacht werden kann. Im Laufe der Forschung konnten mehrere separate Intelligenzfaktoren beobachtet werden, zu welchen Raumvorstellung hinzugezählt wurde. In der darauffolgenden Phase (1938-1961) ging man dazu über, verschiedene Subfaktoren der Raumvorstellung aufzufinden und zu beschreiben, in welcher Art und Weise sich die verschiedenen Raumvorstellungsfähigkeiten voneinander unterscheiden. In der letzten Phase zwischen 1961 und 1982 lag der Fokus vor allem auf den Methoden mit denen die Leistungen in Raumvorstellungstests verbessert werden können.

Im Folgenden sollen die wichtigsten Intelligenztheorien im Zusammenhang mit Raumvorstellung kurz erläutert werden.

1.1.1 Spearman's Zwei-Faktoren-Theorie

Charles E. Spearman untersuchte als einer der ersten Forscher die Struktur der Intelligenz aus faktorenanalytischer Sicht. Er entwickelte Anfang des 20. Jahrhunderts eine monistische Intelligenztheorie, die aufgrund ihrer Plausibilität und Einfachheit allgemein bekannt geworden ist (Stemmler, Hagemann, Amelang, & Bartussek, 2011). Spearman untersuchte die Korrelationsmuster zwischen bestehenden Fähigkeitstests und stellte fest, dass diese meist positiv miteinander korrelierten. Mit Hilfe einer Faktorenanalyse versuchte er die gemeinsamen Komponenten der Intelligenz zu ermitteln. Damit schuf Spearman die Idee eines Generalfaktors g sowie weiterer testspezifischer Faktoren s , die nur für ganz spezielle Bereiche Gültigkeit haben (Spearman, 1904). Nach Spearman's Zwei-Faktoren-Theorie (oder auch General

faktortheorie) „all branches of intellectual activity have in common one fundamental function (or group of functions), whereas the remaining or specific elements of the activity seem in every case to be wholly different from that in all the others” (Spearman, 1904, S. 284).

Da sich jedoch zeigte, dass gewisse spezifische Fähigkeiten untereinander höher korrelieren, wurde der Generalfaktor *g* von späteren Forschern in sogenannte „Gruppenfaktoren“ aufgeteilt, zu denen verbale Faktoren und Raumvorstellungsfaktoren zählen (Stemmler et al., 2011). So benennt El-Koussy (1935, zitiert nach Eliot & Smith, 1983) den Gruppenfaktor der Raumvorstellung *K-Faktor*, den er als Fähigkeit bezeichnet, visuelle räumliche Bilder zu erlangen und zu nutzen.

Noch heute sind die Auswirkungen von Spearman's Zwei-Faktoren-Theorie dahingehend spürbar, dass in den meisten Intelligenztests ein Punktwert für den Gesamttest errechnet wird, der über die durchschnittliche Höhe der Intelligenz Auskunft geben soll (Stemmler et al., 2011).

1.1.2 Thurstone's Primärfaktoren

Thurstone (1931) ging im Gegensatz zu Spearman von einer multifaktoriellen Intelligenztheorie aus. Er baute die Zwei-Faktoren-Theorie Spearman's insofern aus, als dass er sich nicht von vornherein auf eine Zahl von Generalfaktoren festlegte, sondern von mehreren unabhängigen Primärfaktoren ausging, wobei die Anzahl der Primärfaktoren geringer sein sollte als die Anzahl der vorgegebenen Tests (Thurstone, 1931). Zur Bestätigung seiner Theorie entwickelte er das Verfahren der Multiplen Faktorenanalyse.

In seinen Untersuchungen identifizierte Thurstone zunächst neun Faktoren, die er im späteren Verlauf auf sieben reduzierte (Thurstone & Thurstone, 1941, zitiert nach Stemmler et al., 2011). Diese Primärfaktoren nannte er:

- **verbal comprehension**

Die Kenntnis von Wörtern, ihrer Bedeutung und ihrer angemessenen Verwendung wird durch den Faktor *verbal comprehension* zusammengefasst.

- **word fluency**

Der Faktor *word fluency* beschreibt das schnelle produzieren von Wörtern, die bestimmten strukturellen oder symbolischen Erfordernissen entsprechen müssen.

- **number**

Dieser Faktor beinhaltet Geschwindigkeit und Präzision bei einfachen arithmetischen Aufgaben.

- **space**

Die Bewältigung von Aufgaben des räumlichen Vorstellens, Orientierens und Erkennens unter anderem Bezugswinkels wird durch den Faktor *space* Rechnung getragen.

- **memory**

Der Faktor *memory* umfasst das Behalten paarweise gelernter Assoziationen.

- **perceptual speed**

Dieser Faktor beinhaltet Geschwindigkeit beim Vergleich visueller Konfigurationen.

- **reasoning**

Schlussfolgerndes Denken im Sinne des Auffindens allgemeiner Regeln wird durch *reasoning* zusammengefasst.

Alle sieben Primärfaktoren sind in Thurstone's Modell gleichberechtigt und bauen nicht aufeinander auf. Desweiteren teilte Thurstone den Primärfaktor *space* in die drei Subfaktoren *spatial relations*, *visualization* und *spatial orientation* auf (Thurstone, 1950, zitiert nach Maier, 1999).

- **spatial relations**

Der Faktor *spatial relations* bezeichnet die Fähigkeit ein Objekt zu identifizieren wenn es aus einem anderen Blickwinkel betrachtet wird. Hierbei handelt es sich um Umformungen ganzer Figuren, wobei der Betrachter außenstehend bleibt. Es geht damit in erster Linie um die räumliche Beziehung zwischen unbewegten, starren Objekten.

- **visualization**

Unter dem Begriff *visualization* versteht man die Fähigkeit, sich Bewegungen zwischen den Teilen einer Anordnung vorzustellen. Im Detail umfasst dies Rotationen, Verschiebungen und Faltungen. Im Gegensatz zu *spatial relations* handelt es sich hierbei um die Umformung von Teilen einer Figur.

- **spatial orientation**

Spatial orientation repräsentiert die Fähigkeit über räumliche Verhältnisse nachzudenken, wobei die körperliche Ausrichtung des Betrachters eine essenzielle Rolle spielt. Hierbei muss die eigene Person in die räumliche Situation eingeordnet werden, wobei der Betrachter innerhalb der Aufgabensituation hineinversetzt werden muss.

Thurstone kann als einer der ersten Forscher bezeichnet werden der den Faktor „Raumvorstellung“ in mehrere Subfaktoren aufgliederte und somit den Grundstein legte für eine immense Anzahl weiterer Untersuchungen, denen eine vergleichbare Vorgehensweise zugrunde liegt (Hegarty & Waller, 2004; Kaufman, 2007; Linn & Petersen, 1985; Parsons, Larson, Kratz, Thiebaut, Bluestein, Buckwalter, & Rizzo, 2004; Voyer, Voyer, & Bryden, 1995).

1.1.3 Guilford's Intelligenzstrukturmodell

Im Gegensatz zu Spearman und Thurstone lag der Fokus von Guilford's Forschungen auf der Informationsverarbeitung kognitiver Prozesse. Er entwickelte mit Hilfe der Faktorenanalyse eine multifaktorielle Intelligenztheorie, die er *structure-of-intellect model* nannte. Innerhalb dieses Modells versuchte Guilford das intellektuelle Potential eines Menschen anhand dreier Dimensionen vollständig zu beschreiben (Guilford & Hoepfner, 1971):

- **contents**

Dieser Bereich bezeichnet grundlegende Arten von Informationen, welche sich in figurale, symbolische, semantische und behaviorale Inhalte aufteilen lassen. Im Rahmen der Raumvorstellung sind besonders figurale Inhalte von Interesse, welche eine Figur-Grund-Organisation der Wahrnehmung voraussetzen.

- **operations**

Hierbei handelt es sich um die Hauptarten intellektueller Prozesse, welche sich in die Faktoren Gedächtnis, Kognition, divergente Produktion, konvergente Produktion und Evaluation aufgliedern lassen.

- **products**

Innerhalb des Organismus können Informationen die Form von Einheiten, Klassen, Beziehungen, Systemen, Transformationen und Implikationen annehmen.

Guilford und Hoepfner (1971) nennen das Intelligenzstrukturmodell morphologisch, da alle Fähigkeiten aus Kreuzklassifikationen bestehen. Insgesamt besteht das Modell aus $4 \times 5 \times 6 = 120$ Zellen, wovon jede ihre eigene Kombination aus *contents*, *operations* und *products* besitzt und eine unabhängige Fähigkeit repräsentiert (Abbildung 1). Im Rahmen der Raumvorstellung ist besonders der figurale Inhalt (*figural contents*) als essentiell zu betrachten.

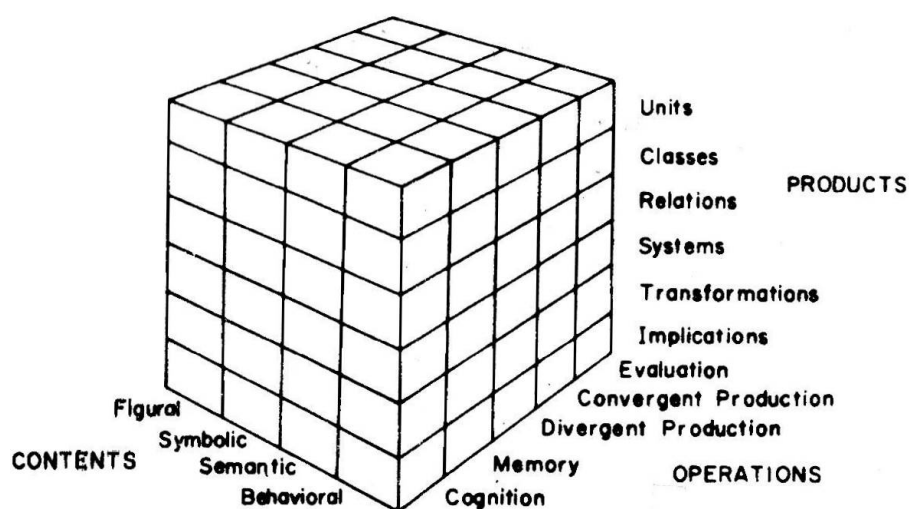


Abbildung 1: Intelligenzstrukturmodell (Guilford & Hoepfner, 1971, S. 19).

Guilford's Intelligenzstrukturmodell beeinflusste darauffolgende Forschungsansätze maßgeblich, da nie zuvor von einer solch großen Anzahl an unabhängigen Fähigkeiten ausgegangen worden war.

1.1.4 Gardner's Theorie der multiplen Intelligenzen

Im Gegensatz zu diversen anderen Intelligenzmodellen sticht im Besonderen Gardner's Theorie der multiplen Intelligenzen hervor. Sein Ziel lag darin, Intelligenzleistungen zu beschreiben die in allen Kulturen aufzufinden sind. In seiner Theorie betont Gardner vor allem die Eigenständigkeit verschiedener Intelligenzformen, wobei diese keine hierarchische Struktur im Sinne von Spearman aufweisen, sondern gleichberechtigt nebeneinander stehen. Dabei konnte Gardner (1985) sechs autonome Intelligenzen ausfindig machen.

- **linguistische Intelligenz**

Sie beinhaltet einen semantischen, phonologischen, syntaktischen und pragmatischen Umgang mit Sprache.

- **musikalische Intelligenz**

Sie dient der Unterscheidung von Melodien, Rhythmen und Timbres. Ebenfalls erfordert Musik eine horizontale und vertikale Erfassung von Tönen und Tonfolgen. Möglicherweise besteht ein Zusammenhang zwischen musikalischer und räumlicher Intelligenz.

- **logisch-mathematische Intelligenz**

Sie beinhaltet einen weiten Bereich von der sensomotorischen Beziehung zur Umwelt bis hin zu Abstraktionen.

- **räumliche Intelligenz**

Diese wird definiert als die Fähigkeit die visuelle Welt richtig wahrzunehmen, zu transformieren, modifizieren und visuelle Erfahrungen zu reproduzieren wenn die physische Stimulierung fehlt.

- **körperlich-kinästhetische Intelligenz**

Diese beinhaltet die Fähigkeit die Körperfunktionen optimal zu kontrollieren und Objekte geschickt zu manipulieren.

- **personale Intelligenz**

Diese Intelligenzform teilt Gardner wiederum in zwei verschiedene Faktoren auf.

- *intrapersonale Intelligenz*

Sie beinhaltet die Fähigkeit zur Introspektion. Damit ist die Fähigkeit gemeint, das eigene Gefühlsleben zu erfassen, zu verstehen, zu benennen und ausdrücken zu können.

- *interpersonale Intelligenz*

Sie beschreibt die Fähigkeit die Intention und Motivation anderer Menschen zu verstehen und darauf kompetent zu reagieren.

Gardner (1985) zufolge sind die verschiedenen Intelligenzen in bestimmten Gehirnregionen lokalisierbar und ermöglichen somit Spezialbegabungen. Raumvorstellungsfähigkeiten können im Allgemeinen unter Gardner's räumlicher Intelligenz subsummiert werden.

1.2 Definition von Raumvorstellung

Im Laufe der Intelligenzforschung wurde Raumvorstellung oftmals als essentieller Bestandteil des Konstrukts „Intelligenz“ erachtet. Im Allgemeinen werden räumliche Fähigkeiten als diejenigen Fertigkeiten bezeichnet, mit denen sich Menschen im Raum orientieren (Carroll, 1993). Eine Definition von Lohman (1979) beschreibt Raumvorstellung als „the ability to generate, retain, and manipulate abstract visual images“ (S. 126, zitiert nach Carroll, 1993).

Um eine exaktere Definition der Raumvorstellung möglich zu machen, gilt es zunächst herauszufinden, ob es sich bei Raumvorstellung tatsächlich um ein einheitliches Konstrukt handelt oder ob diese durch mehrere Faktoren beschrieben werden muss. Die meisten Autoren (Hegarty & Waller, 2004; Kaufman, 2007; Linn & Petersen, 1985; Parson et al., 2004; Voyer, Voyer, & Bryden, 1995) gehen heutzutage davon aus, dass

Raumvorstellung ein Zusammenspiel mehrerer Faktoren ist und die Theorie eines einheitlichen Konstrukts verworfen werden kann. Bei der Identifikation der relevanten Raumvorstellungsfaktoren herrscht in der Literatur jedoch keine Übereinstimmung.

In einer zusammenfassenden Übersicht von Studien zur Raumvorstellung geht McGee (1979) von einer starken Unterstützung für die Existenz von mindestens zwei unterschiedlichen Raumvorstellungsfaktoren aus:

- **spatial visualization**

Hierbei ist die Fähigkeit gemeint, Objekte mental zu rotieren, manipulieren oder zu drehen. In diesem Fall werden objektbezogene Transformationen durchgeführt.

- **spatial orientation**

Dabei handelt es sich um die Fähigkeit, die Anordnungen von Elementen in einem visuellen Muster zu verstehen, sowie die Fähigkeit, sich das Erscheinen von Objekten aus verschiedenen Perspektiven vorzustellen. In diesem Fall werden egozentrische Transformationen durchgeführt.

Lohman (1979, zitiert nach Carroll, 1993) konnte die Faktoren *spatial orientation* und *visualization* von McGee bestätigen, fügte jedoch noch einen weiteren Faktor, *spatial relation*, hinzu, den er als Fähigkeit beschreibt, Figuren und Objekte in der Vorstellung zu rotieren. Auffällig ist dabei, dass die drei Faktoren von Lohman den Subfaktoren von Thurstone (1950, zitiert nach Maier, 1999) stark ähneln.

Zimmerman (1954, zitiert nach Stumpf & Fay, 1983) stellte hingegen in seiner Kontinuumshypothese die Vermutung auf, dass es sich möglicherweise bei den beiden Subfaktoren *spatial visualization* und *spatial orientation* um ein Kontinuum handelt, bei dem die Faktoren sich nur in ihrer Komplexität und Schwierigkeit unterscheiden. So seien Aufgaben im Bereich *spatial orientation* leichter zu lösen als Aufgaben zu *spatial visualization*.

Eine weitere Einteilung der Raumvorstellungsfaktoren stammt von Linn und Petersen (1985), welche anhand einer Metaanalyse drei homogene Bereiche identifizieren, in die Raumvorstellungsfähigkeiten aufgliedert werden können:

- **spatial visualization**

Der erste Faktor beschreibt die Fähigkeit komplizierte, mehrteilige Manipulationen räumlicher Informationen vorzunehmen. Beispiele hierfür wären der Embedded Figure Test oder der Mental Paper Folding Test.

- **spatial perception**

Dieser Faktor beschreibt die Fähigkeit trotz ablenkender Informationen räumliche Verhältnisse in Bezug zur Orientierung des eigenen Körper zu setzen. Beispiele für derartige Aufgaben sind der Rod and Frame Test und Water Level Tests. Im Vergleich zu McGee (1979) kann *spatial perception* als Analogon zum Faktor *spatial orientation* gesehen werden.

- **mental rotation**

Hierbei handelt es sich um die Fähigkeit zwei- oder dreidimensionale Objekte zügig und richtig zu drehen. Ein Beispiel dafür wären die Items des Mental Rotation Test (MRT) von Vandenberg und Kuse (1978), die ausschließlich aus Aufgaben der *mental rotation* bestehen.

Hinsichtlich einer allgemeinen Definition verstehen Linn und Petersen (1985) Raumvorstellung als „skill in representing, transforming, generating, and recalling symbolic, nonlinguistic information“ (Linn & Petersen, 1985, S. 1482).

Eine weitere Definition stammt von Gittler (1994), welcher Raumvorstellung als diejenige Fähigkeit bezeichnet, „räumliche Gegebenheiten [...] in unserer Vorstellung statisch und dynamisch „abzubilden““ (S. 106), beziehungsweise als „nicht-sprachliche Fähigkeit der mentalen Repräsentation und Transformation räumlicher Gegebenheiten“ (S. 106).

Gemeinsam ist diesen Definitionen unter anderem die Tatsache, dass es sich um eine Fähigkeit handelt, die von Sprache unabhängig ist. Diese Theorie wird durch die Tatsache gestützt, dass im Rahmen biologischer Erklärungsmodelle (siehe Kapitel 1.3.2) generell die Ansicht vertreten ist, dass Raumvorstellungsfähigkeit in der rechten Hemisphäre, Sprache hingegen in der linken Hemisphäre des Gehirns lokalisiert ist (Vogel, Bowers, & Vogel, 2003).

Dennoch herrscht bis heute keine Einigkeit über Art und Anzahl der notwendigen Dimensionen von Raumvorstellung (Gittler, 1990). Dieses Problem ist ebenfalls ersichtlich an der großen Zahl an Raumvorstellungstests welche in psychometrischen

Studien verwendet werden (Voyer et al., 1995). So zählen Eliot und Smith bereits im Jahr 1983 in einem internationalen Verzeichnis 392 unterschiedliche Raumvorstellungstests auf.

1.3 Geschlechtsspezifische Unterschiede in der Raumvorstellung

Es zeigte sich in mehreren Studien (Arendasy, Sommer, & Gittler, 2010; Linn & Petersen, 1985; Voyer et al., 1995), dass Raumvorstellung eine der kognitiven Fähigkeiten zu sein scheint, die die größten und stabilsten Geschlechtsunterschiede aufzeigt.

Betrachtet man die geschlechtsspezifischen Differenzen der drei Raumvorstellungsbereiche *mental rotation*, *spatial perception* und *spatial visualization* nach Linn und Petersen (1985) genauer, so ist erkennbar, dass die größten Geschlechtsunterschiede bei Aufgaben bestehen, die eine zügige Manipulation symbolischer Informationen (*mental rotation*) erfordern. Die Effektgröße der Geschlechtsunterschiede liegt für Aufgaben zur dreidimensionalen Rotation bei $d = .90$, für zweidimensionale Rotation hingegen nur bei $d = .30 - .50$ (Linn & Petersen, 1985; Voyer et al., 1995).

Geringere geschlechtsspezifische Unterschiede können bei Aufgaben gefunden werden, die das Erfassen von Horizontalität und Vertikalität benötigen (*spatial perception*). Hierbei liegt die Effektgröße lediglich bei $d = .38 - .56$ (Linn & Petersen, 1985; Voyer et al. 1995). *Spatial visualization* hingegen zeigt mit einer Effektgröße von $d = .16$ die geringsten geschlechtsspezifischen Unterschiede auf.

Die Größe der Geschlechtsunterschiede variiert dabei stark über die verwendeten Tests (Kaufman, 2007). Dennoch scheinen die gefundenen Geschlechtsunterschiede sehr zuverlässig und umso stärker ausgeprägt zu sein, desto älter die untersuchten Testpersonen sind (Geiser, Lehmann, Corth, & Eid, 2008).

1.3.1 Entwicklungspsychologischer Verlauf

Für das Auffinden der Ursachen geschlechtsspezifischer Leistungsunterschiede erscheint es relevant herauszufinden, ab welchem Lebensalter erstmals Differenzen in der Raumvorstellung zu beobachten sind. Je früher im Entwicklungsverlauf der Kinder geschlechtsspezifische Unterschiede festgemacht werden können, desto eher ist

anzunehmen, dass genetische Ursachen eine mögliche Erklärung darstellen (Linn & Petersen, 1985).

Über das genaue Alter des Auftauchens von Geschlechtsunterschieden bezüglich der Raumvorstellungsleistung besteht jedoch keine Einigkeit in der wissenschaftlichen Forschung. Linn und Petersen (1985) verglichen in ihrer Metaanalyse die Leistung von Kindern hinsichtlich der drei von ihnen benannten Faktoren der Raumvorstellung (*mental rotation*, *spatial visualization* und *spatial perception*) und stellten fest, dass Geschlechtsunterschiede in *mental rotation* immer bereits dann festzustellen sind, wenn das entsprechende Testmaterial zur Prüfung dieses Bereiches auch valide eingesetzt werden kann. Diesbezügliche Studien geben ein Mindestalter von 13 Jahren an. Im Bereich *spatial perception* sind Geschlechtsunterschiede hingegen erst ab einem Alter von 18 Jahren erkennbar, für *spatial visualization* berichten Linn und Petersen (1985) keine nennenswerten Unterschiede.

Neuere Untersuchungen können signifikante Geschlechtsunterschiede in der Raumvorstellungsleistung vor allem im Bereich der *mental rotation* bereits ab einem Alter von 9 (Vederhus & Krekling, 1996), 10 (Titze, Jansen, & Heil, 2010), beziehungsweise 13 Jahren (Voyer et al., 1995) feststellen.

Besonders interessant erscheint die Untersuchung von Moore und Johnson (2008), die die Raumvorstellung bereits bei Säuglingen untersuchten. Sie maßen die Dauer der Aufmerksamkeit und Blickfixierung von Säuglingen zu Aufgaben der *mental rotation* und deuteten diese als Kennwerte der Wiedererkennung von Objekten. Hier konnten ebenfalls geschlechtsspezifische Unterschiede zugunsten der Buben aufgezeigt werden. Männliche Säuglinge können im Gegensatz zu weiblichen Säuglingen bereits ab einem Alter von fünf Monaten gedrehte räumliche Objekte wiedererkennen (Moore & Johnson, 2008).

Vor allem die zuletzt beschriebenen Ergebnisse lassen möglicherweise auf auslösende Faktoren schließen, die ihre Ursachen in sehr frühen Entwicklungsphasen oder genetischen Bedingungen haben (siehe Kapitel 1.3.2).

Aus methodischer Sicht ist jedoch anzumerken, dass die meisten Testverfahren zur Raumvorstellung für Erwachsene und Jugendliche entwickelt werden, so dass eine unveränderte Vorgabe für jüngere Kinder problematisch ist. Bodeneffekte können daher nicht ausgeschlossen werden (Vederhus & Krekling, 1996). Aufgrund dessen ist bei der Interpretation einschlägiger Studien darauf zu achten, welche Erhebungsinstrumente mit welcher Altersgruppe kombiniert werden.

Im Allgemeinen scheinen die Geschlechtsunterschiede im Bezug auf die besseren Leistungen der Männer in Raumvorstellungstests über die gesamte Lebensspanne sehr robust zu sein (Geiser, Lehmann, & Eid, 2008). Mit zunehmendem Alter scheinen männliche Probanden ihre Leistungen im Faktor *mental rotation* sogar stetig zu verbessern, wohingegen Frauen keine signifikanten Verbesserungen im zunehmenden Alter aufweisen (Titze et al., 2010).

Driscoll, Hamilton, Yeo, Brooks, und Sutherland (2005) stimmen mit diesen Ergebnissen prinzipiell überein, konnten jedoch bei einem Alter von 40-59 Jahren ein umgedrehtes Geschlechtsverhältnis ausmachen. In diesem Altersabschnitt besitzen Frauen durchschnittlich höhere Ergebnisse im MRT (Vandenberg & Kuse, 1978) als Männer. Dies mag Driscoll et al. (2005) zufolge mit hormonellen Veränderungen innerhalb der Wechseljahre in Zusammenhang stehen.

1.3.2 Biologische Erklärungsansätze der Geschlechtsunterschiede

Weitere Ursachen in Bezug auf unterschiedliche Leistungen im Bereich der Raumvorstellung können in biologischen Erklärungsansätzen gefunden werden. Eine der frühesten Hypothesen in diesem Zusammenhang besteht in der Annahme, dass unterschiedliche geschlechtsspezifische Raumvorstellungsleistungen mit rezessiven X-Chromosomen in Zusammenhang stehen (Stafford, 1961). Es wird angenommen, dass Verhaltensweisen, die durch rezessive X-Chromosome gesteuert werden, bei Männern (XY) eher zutage treten als bei Frauen (XX). Da Männer im Gegensatz zu Frauen nur ein X-Chromosom besitzen, besteht eine größere Wahrscheinlichkeit für das Auftreten eines rezessiven X-Chromosoms und der damit einhergehenden Verhaltensweisen, wie beispielsweise einer besseren Raumvorstellungsfähigkeit. Es konnten im weiteren Verlauf der Forschung jedoch keine konsistenten Nachweise für die Hypothese gefunden werden (Linn & Petersen, 1985).

Heutzutage konzentriert sich die biologische Forschung eher auf den Einfluss von Hormonen und der Lateralisierung des Gehirns als ausschlaggebende Gründe für die geschlechtsspezifischen Raumvorstellungsleistungen. Auf diese soll im Folgenden näher eingegangen werden.

1.3.2.1 Lateralisierung des Gehirns

In mehrere Studien (Hugdahl, Thomsen, & Ersland, 2006; Rilea, 2008; Vogel et al., 2003) konnte ein Zusammenhang zwischen der Lateralisierung des Gehirns und Raumvorstellungsfähigkeiten beschrieben werden, wobei eine starke Überlegenheit der rechten Hemisphäre bei Raumvorstellungsaufgaben sichtbar wird. Daher wird Raumvorstellungsfähigkeit generell in der rechten Hemisphäre vermutet.

Geschlechtsspezifische Differenzen können dahingehend gefunden werden, dass Männer eine starke Präferenz für die rechte Hemisphäre, Frauen hingegen keine Präferenz für eine der beiden Hemisphären bei Raumvorstellungsaufgaben aufzeigen (Vogel et al., 2003). Die Ergebnisse lassen Vogel et al. (2003) zu der Annahme kommen, dass möglicherweise Männer zur Lösung von Raumvorstellungsaufgaben die Figuren selbst nutzen, wohingegen Frauen zusätzlich auf die Sprache zurückgreifen, die in der linken Hemisphäre angesiedelt ist.

Zusätzlich betrachteten Vogel et al. (2003) unterschiedliche Faktoren der Raumvorstellung genauer. Bei Aufgaben die *spatial visualization* erforderten, zeigte sich keine Präferenz für eine der beiden Hemisphären, was dafür spricht, dass beide Hemisphären für diese Art von Aufgaben genutzt werden können. Im Gegensatz dazu zeigte sich bei Aufgaben zu *mental rotation* und zu *spatial perception* eine Präferenz für die rechte Hemisphäre. Diese Ergebnisse stehen im Einklang mit den Ergebnissen von Linn und Petersen (1985), in welchen Männer in *mental rotation* und *spatial perception* bessere Ergebnisse aufzeigen als Frauen, in *spatial visualization* jedoch nur sehr geringe Unterschiede zwischen den Geschlechtern zu beobachten sind.

Für den von Thurstone (1950, zitiert nach Maier, 1999) beschriebenen Bereich *spatial relation* hingegen, konnte ein linkshemisphärischer Vorteil für die Bearbeitungsstrategie der kategorialen Verarbeitung und ein rechtshemisphärischer Vorteil für Bearbeitungsstrategie der koordinierenden Verarbeitung gefunden werden (Van der Ham & Borst, 2011).

Was den Zusammenhang der Händigkeit zur Raumvorstellungsleistung angeht, kann eine Präferenz der Rechtshänder für die rechte Hemisphäre gefunden werden. Für Linkshänder kann hingegen keine Präferenz festgestellt werden (Vogel et al., 2003).

Eine weitere Hypothese, warum Frauen in Raumvorstellungsaufgaben schlechter abschneiden ist, dass Frauen zur Bearbeitung solcher Aufgaben Ressourcen der linken Hemisphäre hinzuziehen (Rilea, 2008). Dies führt zu einer bilateralen Verarbeitung von Raumvorstellungsaufgaben. Da Raumvorstellungsfähigkeit jedoch

nicht von sprachlichen Fähigkeiten abhängig ist, führt dies zu einer schlechteren Leistung bei Frauen (Rilea, 2008).

Anhand einer genaueren Betrachtung der Aktivierung des Gehirns mittels bildgebender Verfahren konnte beobachtet werden, dass bei mentaler Rotation besonders der rechte Parietallappen eine wichtige Rolle zu spielen scheint (Hugdahl et al., 2006).

Eine morphologische Betrachtung des Parietallappens zeigt einen Zusammenhang der Raumvorstellungsleistung mit der Menge an grauer Substanz im Parietallappen und dessen Oberflächenbereich auf (Koscik, O'Leary, Moser, Andreasen, & Nopoulos, 2009). So kann bei Frauen eine größere Menge an grauer Substanz im Parietallappen aufgefunden werden, was mit einer nachteiligen Wirkung in den Leistungen in mentaler Rotation in Zusammenhang gebracht wird. Männer hingegen besitzen einen größeren Oberflächenbereich des Parietallappens, was mit einer vorteilhaften Wirkung in mentaler Rotation in Verbindung steht. Koscik et al. (2009) gehen davon aus, dass der höhere proportionale Anteil an grauer Substanz bei Frauen eher mit einer höheren Fähigkeit in der räumlichen Merkfähigkeit als in der Raumvorstellungsleistung einher geht.

1.3.2.2 Hormonelle Einflüsse

Einige Autoren (Driscoll et al., 2005; Putz, Gaulin, Sporter, & McBurney, 2004; Sanders, Bereczkei, Csatho, & Manning, 2005) sehen hormonelle Veränderungen als auslösende Gründe für die Unterschiede in den Raumvorstellungsleistungen von Männern und Frauen. Dabei kann zwischen einem pränatalen, hormonellen Einfluss auf die Entwicklung und einem Einfluss fluktuierender Geschlechtshormone im Erwachsenenalter unterschieden werden (Sanders et al., 2005).

Unterstützt wird die erste Hypothese durch die Ergebnisse einer Studie von Sanders et al. (2005), die einen signifikanten negativen Zusammenhang zwischen der Raumvorstellungsleistung und dem Verhältnis der Länge des Zeigefingers zum Ringfinger bei Männern beobachten konnten. Ihre Ergebnisse konnten jeweils in drei unabhängigen Studien repliziert werden. Für Frauen wurde ein solcher Zusammenhang jedoch nicht aufgefunden. Die Unterschiede im Geschlecht sind dabei möglicherweise abhängig von weiteren unabhängigen Einflussfaktoren. Andere Autoren hingegen (Putz et al., 2004) konnten keinen verlässlichen Zusammenhang zwischen der Raumvorstellungsleistung und dem Längenverhältnis der Finger finden.

Auch Hines, Fane, Pasterski, Mathews, Conway, und Brook (2003) gingen der Hypothese nach, dass pränatale Androgene, unter anderem Testosteron, einen positiven Einfluss auf die Raumvorstellung haben. Hines et al. (2003) vermuteten, dass der Geschlechtsunterschied in der Raumvorstellungsleistung bei Mädchen mit einem pränatalen Überschuss an Androgenen, dem Adrenogenitalen Syndrom, verringert sein sollte. Diese Hypothese konnte jedoch nicht bestätigt werden. Es kann kein Zusammenhang zwischen einem erhöhten Androgenwert und guten Raumvorstellungsleistungen bei Mädchen festgestellt werden. Lediglich das Spielverhalten der Mädchen verändert sich hin zu eher burschenhaften Spielen. Burschen mit einem erhöhten Androgenwert zeigen sogar eine schlechtere Raumvorstellungsleistung auf als ihre Altersgenossen (Hines et al., 2008). Dem entspricht auch das Ergebnis von Hooven, Chabris, Ellison, und Kosslyn (2004), die ebenfalls von der Annahme ausgehen, dass ein hoher Testosteronwert keinen direkten Einfluss auf mentale Rotation hat. Hines et al. (2003) sehen die Ursachen für Geschlechtsunterschiede demnach eher in psychischen Ursachen.

1.3.3 Psychische Erklärungsansätze der Geschlechtsunterschiede

Ein weiterer Ansatzpunkt für das Auffinden der Ursachen von Geschlechtsunterschieden in der Raumvorstellung wird in psychischen Bereichen wie dem Selbstvertrauen, der Erfahrung und in unterschiedlichen Lösungsstrategien vermutet.

1.3.3.1 *Selbstvertrauen*

Selbstvertrauen scheint eine bedeutende Einflussvariable in Bezug auf Raumvorstellungsleistung zu sein. In einer Studie zum Antwortverhalten im Mental Rotation Test von Vandenberg und Kuse (1978) stellten Cooke-Simpson und Voyer (2007) fest, dass Männer nicht nur bessere Leistungen erzielen, sondern auch zuversichtlicher hinsichtlich ihrer Leistung sind. Frauen hingegen weisen niedrigere Leistungen und eine niedrigere Zuversicht auf. Daraus ergibt sich die Frage nach der Beeinflussungsrichtung, nämlich ob eine hohe Raumvorstellungsleistung das Selbstvertrauen steigert oder ob ein erhöhtes Selbstvertrauen mit einer höheren Raumvorstellungsleistung einher geht. Cooke-Simpson und Voyer (2007) gehen davon aus, dass eine Verbesserung der Raumvorstellungsleistung eine Zunahme des Selbstvertrauens fördert.

In engem Zusammenhang zu Selbstvertrauen steht der Einfluss von Vorurteilen. Verschiedene Autoren (Moè, 2009; Moè & Pazzaglia, 2006; Wraga, Helt, Jacobs, & Sullivan, 2006) konnten zeigen, welch großen Einfluss Vorurteile und Stereotype bezüglich der eigenen Leistung in Raumvorstellungstests haben. Dies gilt insofern, als dass Frauen ihre Leistungen verbessern, wenn ihnen mitgeteilt wird, dass sie dem anderen Geschlecht gegenüber im Test überlegen sind. Die Verbesserung der Frauen ergibt sich möglicherweise daraus, dass vorhandene Ängste reduziert und Zuversicht in die eigene Leistung gesteigert wird (Moè & Pazzaglia, 2006).

Ebenso tritt jedoch eine Verschlechterung in den Leistungen der Frauen auf, wenn ihnen mitgeteilt wird, dass Männer in Raumvorstellungstests tendenziell besser abschneiden. Da dies jedoch eine geläufige Annahme ist, besteht die Vermutung, dass Frauen in Raumvorstellungstests nicht ihre maximale Leistung ausschöpfen (Wraga et al., 2006).

Eine ähnliche Erklärung für die unterschiedlichen Leistungen sehen Moè, Meneghetti, und Cadinu (2009) in der Nutzung von inkrementellen Theorien, das heißt, ob die Überzeugung besteht, dass man sich in seinen Leistungen überhaupt verbessern kann. Der Glaube daran, dass die Möglichkeit besteht Fähigkeiten zu erwerben, die typisch männlich sind, führt dazu, dass Frauen effektivere Strategien verwenden und somit ihre Raumvorstellungsleistung tatsächlich auch verbessern (Moè et al., 2009).

Die Ergebnisse zeigen demnach, dass Faktoren wie Vorurteile und Stereotype dazu beitragen, die fast ausschließlich schlechteren Leistungen der Frauen in Raumvorstellungstests zu erklären (Wraga et al., 2006).

1.3.3.2 Erfahrung

Schon in den 80er Jahren konnte gezeigt werden, dass Erfahrung in Aktivitäten zu deren Durchführung Raumvorstellung erforderlich ist, beispielsweise Bogenschießen, Basketball wie auch Geometrie, mit einer guten Leistung in Raumvorstellungstests zusammenhängen (Baenninger & Newcombe, 1989). Dabei wird davon ausgegangen, dass die größere Erfahrung von Burschen in solchen Aktivitäten die geschlechtsspezifischen Differenzen in Raumvorstellungsaufgaben erklärt.

Alexander und Evardone (2008) beobachteten, dass Buben eher mit Sachgegenständen wie Bauklötzen spielen, wohingegen Mädchen sich eher mit menschenähnlichen Figuren beschäftigen. Aufgrund der häufigen Verwendung dieser Spielobjekte kommt es zu einer größeren Übung und Erfahrung was deren Form und Proportionen angeht.

Dies führt zu einem Vorteil der Burschen in mentalen Rotationsaufgaben, da diese meist aus unbelebten Gegenständen wie Würfeln bestehen. Um ihre Hypothese zu überprüfen, verwenden Alexander und Evardone (2008) zusätzlich zu typischen mentalen Rotationsaufgaben dreidimensionale menschliche Figuren. In einer Studie mit erwachsenen Probanden zeigt sich, dass die Leistung der Frauen in der Rotation menschlicher Figuren verbessert werden kann (Alexander & Evardone, 2008). Dies spricht für einen wesentlichen Effekt der Erfahrung in Aufgaben zur mentalen Rotation. Interessant erscheint auch, dass für Burschen eine regelmäßige Beschäftigung mit Computerspielen mit guten Leistungen in Raumvorstellungstests zusammenhängt (Quaiser-Pohl, Geiser, & Lehmann, 2006). Hierbei scheinen besonders Action-Computerspiele einen raumvorstellungsfördernden Effekt zu haben. Für Mädchen kann jedoch kein solcher Effekt anhand von Computerspielen ausgemacht werden.

Andere Studien hingegen (Feng, Spence, & Pratt, 2007) konnten wiederum eine größere Leistungssteigerung der Mädchen in Raumvorstellungstests nach einem Training von Computerspielen beobachten.

Weitere Forschung im Bereich der geschlechtsspezifischen Förderung von Computerspielen scheint daher vonnöten zu sein. Ein genereller positiver Effekt von Action-Computerspielen im Bereich der Raumvorstellungsleistung scheint jedoch belegt zu sein.

Desweiteren können Umgebungsfaktoren einen wesentlichen Einfluss auf die Erfahrung mit Raumvorstellung haben. Levine, Vasilyeva, Lourenco, Newcombe, und Huttenlocher (2005) konnten signifikante Geschlechtsunterschiede zugunsten der Burschen bei Kindern aus einer hohen sozioökonomischen Schicht feststellen, nicht jedoch bei Kindern aus einer niedrigeren sozioökonomischen Schicht. Die Gründe für die fehlenden Geschlechtsunterschiede bei Letzteren werden dabei in dem schlechteren Zugang zu Raumvorstellung förderndem Spielzeug vermutet, so dass Burschen und Mädchen Raumvorstellungsaktivitäten eher gleichartig nutzen. Bei Kindern aus einer höheren sozioökonomischen Schicht hingegen, beschäftigen sich Burschen öfter mit Aktivitäten der Raumvorstellung.

1.3.3.3 Lösungsstrategien

Eine Vielzahl an Autoren (Allen, 1974; Geiser, Lehmann, Corth, & Eid, 2008; Glück & Fabrizii, 2010; Glück, Machat, Jirasko, & Rollett, 2002; Linn & Petersen, 1985) geht davon aus, dass ein wesentlicher Faktor für die unterschiedlichen Leistungen in Raumvorstellungstest von Männern und Frauen in deren unterschiedlichen Lösungsstrategien liegt.

Generell kann von zwei verschiedenen Ansätzen ausgegangen werden. Einerseits besteht die Möglichkeit, dass Männer und Frauen die gleichen Strategien verwenden, Frauen diese jedoch ineffizienter anwenden. Andererseits könnten Männer und Frauen unterschiedliche Lösungsstrategien für Raumvorstellungsaufgaben verwenden (Allen, 1974). Bei Letzterem bestünde die Gefahr, dass unterschiedliche Fähigkeitsdimensionen bei Männern und Frauen gemessen werden, womit eine faire Interpretation der Ergebnisse unmöglich wäre (Gittler, 1994).

Einige Studien (Allen, 1974; Glück & Fabrizii, 2010; Linn & Petersen, 1985) konnten Hinweise für das Zutreffen der Theorie finden, dass Frauen und Männer die gleichen Lösungsstrategien in Raumvorstellungsaufgaben benutzen, Frauen jedoch dazu neigen diese weniger effektiv anzuwenden. Allen (1974) geht weiterhin davon aus, dass Frauen häufiger als Männer dazu neigen in Raumvorstellungstests zu raten oder aufzugeben wenn sie merken, dass sie nicht erfolgreich sind. Ebenfalls verwenden sie weniger abstrakte Methoden und testspezifische Strategien als Männer.

Bei einer genaueren Untersuchung des geschlechtsspezifischen Antwortverhaltens kann festgestellt werden, dass Männer bei mentalen Rotationsaufgaben wie dem MRT von Vandenberg und Kuse (1978), bei dem zwei korrekte Antworten aus vier Antwortalternativen ausgewählt werden müssen, dazu neigen die Antwortalternativen schnell und ungenau zu bearbeiten (*leaping response strategy*). Frauen hingegen neigen dazu genau zu untersuchen ob die gewählten Antwortalternativen wirklich richtig sind (*conservative response strategy*; Glück & Fabrizii, 2010; Hirnstein, Bayer, & Hausmann, 2009). Dies führt besonders in Testsituationen mit hohem Zeitdruck (speed-Testungen) zu Vorteilen hinsichtlich der Männer (Glück & Fabrizii, 2010; Voyer, 2011). Eine Metastudie von Voyer (2011) konnte zeigen, dass die Größe der Geschlechtsunterschiede indirekt proportional mit der zur Verfügung stehenden Zeit zusammenhängt, so dass jede Form der Zeitlimitierung Geschlechtsunterschiede in Raumvorstellungstests vergrößert.

Andere Autoren konnten wiederum Hinweise für die Theorie finden, dass Männer und Frauen unterschiedliche Antwortstrategien verwenden (Arendasy, Sommer, & Gittler, 2010; Geiser, Lehmann, Corth, & Eid, 2008; Heil & Jansen-Osmann, 2008). Bereits im Jahr 1953 versuchte Barratt (zitiert nach Quaiser-Pohl, 1998) verschiedene Lösungsstrategien für Raumvorstellungsaufgaben systematisch zu analysieren. Für die Lösung von *mental rotation* Aufgaben identifizierte er einen *part approach*, bei dem einzelne Details verglichen und beurteilt werden und einen *whole approach*, bei dem die Objekte als Ganzes verglichen werden.

Aktuelle Studien greifen dieses Thema auf, wobei sie zwischen einer holistischen und einer verbal-analytischen Vorgehensweise unterscheiden. Mit holistischen Strategien werden diejenigen Fähigkeiten bezeichnet, die Repräsentation einer Figur als Ganzes zu rotieren (Hirnstain et al., 2009). Die verbal-analytische Strategie wiederum bezeichnet die Fähigkeit, bestimmte Merkmale einer Figur mit der Zielfigur zu vergleichen. Der Begriff „verbal-analytische“ wird dabei als Abgrenzung zu räumlich-visuellen Strategien verwendet, da sich diese eher auf die Sprache und schlussfolgerndes Denken stützt (Grüßing, 2002).

Heutzutage scheint die Aufteilung in holistische und verbal-analytische Prozesse weite Verbreitung gefunden zu haben. So meinen Geiser, Lehmann, Corth, und Eid (2008), dass Frauen Rotationsaufgaben eher analytisch betrachten und mehr Ratestrategien verwenden als Männer. Da diese Strategien jedoch eher ineffizient in Bezug auf Raumvorstellungsaufgaben sind, könnte dies ein Grund für das schlechtere Abschneiden der Frauen sein. Ähnlich dazu konnten Arendasy et al. (2010) unterschiedliche Strategien zwischen Männern und Frauen bei der Lösung von Raumvorstellungsaufgaben beobachten. Während Männer die Aufgaben eher durch mentale Rotation lösen, also durch ein sukzessives Drehen der Figuren, bevorzugten Frauen die ineffizientere Strategie der Perspektivänderung, welche darin besteht sich die Figuren aus einer anderen Position vorzustellen.

Auch Heil und Jansen-Osmann (2008) legen Belege vor, dass Frauen analytische Prozesse verwenden, was dazu führt, dass die mentale Rotationsgeschwindigkeit bei komplexen Stimuli sinkt. Männer hingegen nutzen eher holistische Prozesse bei Raumvorstellungsaufgaben, wobei ihre mentale Rotationsgeschwindigkeit nicht durch die Komplexität der Stimuli beeinträchtigt wird.

Ein neuer Ansatz bei der Untersuchung von unterschiedlichen Lösungsstrategien stellt die Verwendung der Item Response Theorie dar. So verwendeten Glück et al. (2002) die IRT um zwei unterschiedliche Lösungsstrategien zur Lösung von Würfelaufgaben,

das Vergleichen von Mustern (*pattern strategy*) und eine tatsächlich räumliche Strategie (*spatial strategy*), zu untersuchen. Mit Hilfe des Linear Logistic Model with Relaxed Assumptions (LLRA) und des Mixed Rasch Models (MRM), konnte festgestellt werden, dass ein Training in Raumvorstellungsaufgaben nicht nur zu quantitativen, sondern auch zu qualitativen Strategieveränderungen führt. So wechselten rund zwei Drittel der Probanden im Posttest zur *spatial strategy*, obwohl sie im Pretest die ineffizientere *pattern strategy* verwendeten.

Zusammenfassend scheinen verschiedenste Untersuchungen dafür zu sprechen, dass Frauen und Männer bei Raumvorstellungsaufgaben unterschiedliche Lösungsstrategien verwenden. Diese können jedoch die geschlechtsspezifischen Unterschiede in Raumvorstellungstests nicht ausreichend erklären, so dass noch andere Faktoren eine ausschlaggebende Rolle spielen müssen (Hirnstein et al., 2009).

1.3.4 Weitere Erklärungsansätze

Eine weitere Erklärung geschlechtsspezifischer Unterschiede liegt in den testspezifischen Bedingungen. So konnte festgestellt werden, dass Frauen Aufgaben zur mentalen Rotation generell langsamer lösen, was dazu führt, dass sie unter zeitlimitierten Bedingungen nicht so gut abschneiden wie Männer (Peters, 2005).

Ein zusätzlicher Grund wird in der paper-pencil Vorgabe gesehen. So konnte gezeigt werden, dass geschlechtsspezifischen Leistungsunterschiede bei einer virtuellen Vorgabe von Aufgaben zur mentalen Rotation verschwinden (Parsons et al., 2004). Diese Ergebnisse liegen möglicherweise an den veränderten kognitiven Anforderungen, da in der virtuellen Testung ein Erstellen von dreidimensionalen Repräsentationen aus zwei-dimensionalen Objekten wegfällt (Parsons et al., 2004).

1.4 Verbesserungsmöglichkeiten der Raumvorstellung durch Training

Neben der Beschäftigung mit den Ursachen der unterschiedlichen Leistungen im Bereich der Raumvorstellung ist vor allem in letzter Zeit ein steigendes Interesse an den Möglichkeiten der Verbesserung von Raumvorstellungsfähigkeiten durch Training zu beobachten.

Baenninger und Newcombe (1989) untersuchten in einer der ersten Metaanalysen zum Thema Raumvorstellung Studien zur Trainierbarkeit von Raumvorstellungsfähigkeiten. Ihre Ergebnisse zeigen, dass ein generelles Training der Raumvorstellungsfähigkeit die Leistungen in Raumvorstellungstests bei Burschen wie auch Mädchen verbessert. Da jedoch nicht immer davon ausgegangen werden kann, dass das Erlernen einer allgemeinen Strategie auch einen Lerntransfer auf eine spezielle Strategie miteinschließt (Klauer, 2000), erscheint es sinnvoll ein testspezifisches Training anzuwenden.

So konnte gezeigt werden, dass ein intensives, langfristiges Training von circa 7 Stunden in der Woche über drei Wochen hinweg die Raumvorstellungsleistung verbessert (Wright, Thompson, Ganis, Newcombe, & Kosslyn, 2008). Dabei ist ein Transfer auf ähnliche Stimuli, sowie auf neue räumliche Testaufgaben zu beobachten. Einige Forscher (Feng et al., 2007; Tzuriel & Egozi, 2010; Neubauer, Bergner, & Schatz, 2010) beschäftigten sich weiterhin mit der Frage, wie Trainingsprogramme zur Raumvorstellung aussehen könnten. Aufgrund der erhöhten räumlichen Aufmerksamkeit innerhalb von Action Computerspielen, wurde ein Training anhand dieser entwickelt (Feng et al., 2007). Sowohl Männer als auch Frauen zeigen bei einem Training von mehreren Stunden Verbesserungen in Raumvorstellungstests wie dem MRT auf. Essentiell ist dabei, dass die geschlechtsspezifischen Leistungsunterschiede verringert werden können.

Ebenfalls wurden Verbesserungen nach einem Training festgestellt, das aus einer Mischung aus Würfelaufgaben, Bausteinen und computerisierten Tetris-Aufgaben bestand (Neubauer et al., 2010). Da in dieser Studie jedoch keine Kontrollgruppe verwendet wurde, ist ein reiner Wiederholungseffekt nicht auszuschließen.

Tzuriel und Egozi (2010) entwickelten ein Raumvorstellungstraining für Kinder im Alter von 6 Jahren. Ihr Trainingsprogramm hatte das Ziel, bereits im frühen Alter Strategien zur visuell-räumlichen Transformation zu erlernen. Die Strategien bestanden darin, das Arbeitsgedächtnis zu erweitern, räumliche Informationen global wahrzunehmen und

räumliche Stimuli aus unterschiedlichen Blickwinkeln zu erkennen. Es zeigte sich, dass das Training signifikant Verbesserungen der Kinder in der Raumvorstellungsleistung förderte und geschlechtsspezifische Unterschiede verringert wurden. Daraus ergibt sich, dass Raumvorstellungsfähigkeiten formbar zu sein scheinen und durch Training stark verbessert werden können (Neubauer et al., 2010). Aus diesen Ergebnissen ist erkennbar, dass Raumvorstellungsfähigkeiten schon in jungen Jahren gefördert werden sollten um Geschlechtsdifferenzen effizient zu überwinden (Tzuriel & Egozi, 2010).

Doch auch Erwachsene können ihre Raumvorstellungsleistungen durch einen effizienten Strategiewechsel verbessern. Gerade Frauen können anhand von Trainings die Anwendung holistischer Strategien erlernen und somit bessere Leistungen in Raumvorstellungstests erbringen (Alexander & Evardone, 2008).

2. Die Item Response Theorie (IRT)

Die Item Response Theorie eröffnet unter anderem die Möglichkeit die Wirksamkeit von Raumvorstellungstrainings zu ermitteln. Sie beinhaltet mehrere Vorteile im Vergleich zu klassischen Testtheorie, die im Folgenden im Detail erläutert werden sollen. Insbesondere das dichotome logistische Modell von Rasch und dessen Erweiterungen können zur Veränderungsmessung genutzt werden.

2.1 Klassische Testtheorie vs. Probabilistische Testtheorie

Die klassische Testtheorie ist die am weitesten verbreitete psychometrische Testtheorie. Sie geht von der Annahme aus, dass das Testergebnis eines Tests direkt dem wahren Ausprägungsgrad eines untersuchten Merkmals entspricht (Bortz & Döring, 2006). Zusätzlich beinhaltet jedes Testergebnis einen Messfehler, der getrennt von der wahren Merkmalsausprägung zu ermitteln ist. Die klassische Testtheorie ist demnach ein deterministisches Modell, da das Testergebnis ausschließlich auf der Merkmalsausprägung und dem Messfehler basiert (Bortz & Döring, 2006).

Gittler und Fischer (2011) nennen vier große Probleme der klassischen Veränderungsmessung:

- **Boden- und Deckeneffekte**

Überschreitungen des Messbereichs nach unten oder oben kommen dadurch zustande, dass die Veränderungen von den Ausgangswerten abhängig sind.

- **artifizielle Korrelationen**

Diese können zwischen Pretest-Scores und Differenz-Scores auftauchen wenn bereits a-priori Gruppenunterschiede bestehen und zu Fehlinterpretationen führen (Gittler, 2008).

- **geringe Reliabilität der Differenz-Scores**

Bei den *gain scores* handelt es sich um den Indikator für einen experimentellen Effekt. Er wird durch die Differenz des Rohscores zum zweiten Testzeitpunkt vom ersten Testzeitpunkt, $r_{v2} - r_{v1}$ errechnet. In der klassischen Testtheorie wird der *gain score* zur Veränderungsmessung verwendet, was jedoch zu mehreren Problemen führen kann. Im Zuge des Rasch-Modells (siehe Kapitel 2.2) kann der *gain score* zusammen mit dem Ausmaß der latenten Fähigkeit einer Person und den Itemparametern sinnvoll interpretiert werden.

- **fehlende Intervallskalierung der Rohwerte**

Statistische Analysen ordinal skalierte Differenz-Scores müssen zu inkonsistenten Ergebnissen führen, da ein nichtlinearer Zusammenhang zwischen Rohwerten und latenten Fähigkeiten besteht. Gleiche Rohwertdifferenzen entsprechen dabei unterschiedlichen Fähigkeitszuwächsen (Gittler, 2008). Erst eine Intervallskalierung der Fähigkeitsparameter lässt eine Veränderungsmessung zu.

Die probabilistische Testtheorie (auch unter Item Response Theorie bekannt) geht hingegen davon aus, dass die Wahrscheinlichkeit der Lösung eines Items größer ist bei einer Person mit einer hohen Merkmalsausprägung, als bei einer Person mit einer geringeren Merkmalsausprägung (Bortz & Döring, 2006). Die Itemcharakteristikkurve (ICC) gibt dabei einen Überblick über den Zusammenhang der Lösungswahrscheinlichkeit eines Items mit der Fähigkeit einer Person.

Im Gegensatz zur klassischen Testtheorie kann daher anhand der IRT geprüft werden, „ob die gegebenen Verrechnungsvorschriften eines Tests zu Testwerten führen, die verhaltensadäquate Relationen wiedergeben“ (Kubinger, 2006, S. 84). Anders

ausgedrückt, können anhand der IRT Modellprüfungen durchgeführt werden, um zu testen, ob ein psychologischer Test lediglich eine einzige Fähigkeitsdimension misst (Gittler & Glück, 1998). Die IRT „betrachtet die untersuchten Merkmale als latente Dimensionen und die einzelnen Testitems als Indikatoren dieser latenten Dimensionen“ (Bortz & Döring, 2006, S. 206). Latente Dimensionen werden dabei *latent traits* genannt und müssen über manifeste Merkmale geschätzt werden.

Eine weitere Anwendung der IRT liegt in der Veränderungsmessung (Gittler & Glück, 1998), worauf in dieser Arbeit ein besonderer Fokus gelegt werden soll. Im Vergleich zur klassischen Testtheorie kann die probabilistischen Testtheorie daher als leistungsfähiger bezeichnet werden (Gittler, 1990).

2.2 Das dichotome logistische Modell von Rasch

Das dichotome logistische Modell von Rasch (auch Einparameter-Logistisches-Modell oder Rasch-Modell) ist eines der am häufigsten verwendeten probabilistischen Testmodelle (Bortz & Döring, 2006). Es beschreibt die Wahrscheinlichkeit, dass eine Testperson ein Item löst, in Abhängigkeit der Fähigkeit der Person und der Schwierigkeit des Items (Kubinger, 2006). Die allgemeine Formel für das Rasch-Modell lautet demnach (Gittler & Fischer, 2011):

$$P(X_{vi} = 1 | S_v, I_i) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)} = \frac{\xi_v \epsilon_i}{1 + \xi_v \epsilon_i}$$

Sie beschreibt die Wahrscheinlichkeit einer richtigen Antwort der Person S_v auf das Item I_i .

$X_{vi} = 1$ bezeichnet eine richtige Beantwortung des Items I_i .

β_i ist der Schwierigkeitsparameter, ϵ_i der Einfachheitsparameter des Items I_i .

θ_v und ξ_v sind die Fähigkeitsparameter der Person S_v .

Innerhalb des Rasch-Modells kann jede Antwort in zwei Kategorien (richtig oder falsch) eingeordnet werden. Zusätzlich existiert ebenfalls das polytome logistische Modell, bei dem mehr als zwei Antwortkategorien existieren (Molenaar, 1995). Im weiteren Verlauf soll jedoch ausschließlich auf das dichotome logistische Modell eingegangen werden. Damit das Rasch-Modell gilt, müssen laut Gittler (1992; 1999) folgende Modellannahmen erfüllt sein:

- **Eindimensionalität der Items**

Alle Testaufgaben sollen für alle Testpersonen ein und dieselbe latente Fähigkeitsdimension messen (Gittler, 1990). Aus diesem Grund widerspricht eine Vorgabe von speed-Tests den Voraussetzungen des Modells, da mehrere Eigenschaftsdimensionen gemessen werden. Sind gewisse Items hingegen für Frauen schwieriger zu lösen als für Männer, so spricht dies nicht notwendigerweise gegen die Eindimensionalität der Items. So lange die Rangreihe und Relationen zwischen den Itemschwierigkeiten für beide Gruppen gleich sind, ist die Eindimensionalität gegeben (Gittler & Glück, 1998).

- **lokal stochastische Unabhängigkeit**

Die Wahrscheinlichkeit ein Item zu lösen, hängt nur von der Fähigkeit der Person und der Schwierigkeit des Items ab und ist unabhängig davon, welche Items davor gelöst wurden (Bortz & Döring, 2006). Dies ist besonders hilfreich für die Nutzung des LLTM, da aufgrund dieser Annahme Daten umorganisiert werden können (Gittler, 2008).

- **monotone Itemcharakteristik**

Die Itemcharakteristikkurve wächst stetig an und besitzt die asymptotischen Grenzen „0“ und „1“. Je fähiger die Person ist, desto wahrscheinlicher ist es für sie ein Item zu lösen.

- **Form der Itemcharakteristik**

Die Form der Itemcharakteristik entspricht einer logistischen Funktion der Differenz der Personenfähigkeit und der Itemschwierigkeit.

- **dichotome Testaufgaben**

Das Modell ist nur auf Items mit dichotomem Antwortformat anwendbar. Wird ein Item gelöst, wird es mit „1“ kodiert, wird es hingegen nicht gelöst, mit „0“.

Sind alle Modellannahmen erfüllt, besitzt das Rasch-Modell laut Gittler (1990; 1999) folgende Modelleigenschaften:

- **Spezifische Objektivität**

Der Test misst für alle Testpersonen die gleiche Fähigkeitsdimension unabhängig davon welche Personenstichprobe (Alter, Geschlecht) für die Schätzung der Itemparameter herangezogen wurde. Ebenfalls ist ein Vergleich zwischen zwei Personen davon unabhängig, auf welchen Items er basiert.

- **Erschöpfende Statistik**

Der Rohscore der sich für jede Person ergibt, ist ein faires Leistungsmaß. Er enthält alle notwendigen Informationen über den unbekannten Personenparameter. Somit haben alle Personen mit dem gleichen Rohscore den gleichen Fähigkeitsgrad, während Personen mit einem höheren Rohscore leistungstärker und mit einem niedrigeren Rohscore leistungsschwächer sind.

- **Intervallskala**

Die Item- und Personenparameter (Fähigkeitsgrade) liegen auf einer Intervallskala. Dies bedeutet, dass Lineartransformationen durchgeführt werden können. Damit können Personenparameterdifferenzen direkt interpretiert werden (Arendasy & Gittler, 2003).

Es gibt mehrere Modelltests zur Modellkontrolle des Rasch-Modells (für eine ausführliche Auflistung siehe Gittler (1986)). Zu diesen zählt unter anderem der Likelihood-Quotiententest (LQT) nach Andersen (1973), welcher die Modelleigenschaft der Stichprobenunabhängigkeit prüft. Hierbei betrachtet er die Gleichheit der Itemparameter in Subgruppen, wobei die Itemparameter auch gesamt geschätzt werden.

Unumgänglich für eine mögliche Modellgeltung Rasch-homogener Tests sind jedoch neben allen technischen Kriterien inhaltliche Überlegungen (Gittler, 1986). Nur durch a-priori-Hypothesen lässt sich die Frage beantworten, ob und warum Items zueinander passen und wie eine sinnvolle Itemselektion durchführbar ist. Wird dies hingegen vernachlässigt, kann es zu einer artifiziellen Modellanpassung kommen (Gittler, 1999). Dies führt dazu, dass die Rasch-Homogenität in anderen Untersuchungen mit dem gleichen Testmaterial nicht mehr gegeben sein kann.

2.2.1 Das Lineare logistische Testmodell

Bei dem Linearen logistischen Testmodell (LLTM) handelt es sich um eine Erweiterung des Rasch-Modells, wobei gewisse lineare Nebenbedingungen für die Itemparameter eingeführt werden (Fischer, 1995). Die allgemeine Formel für das LLTM lautet laut Gittler (2008): $\sigma_i = \sum_j q_{ij} \eta_j + c$

Sie beschreibt die Itemparameter σ_i als gewichtete Summe von m Basisparametern η_j und einer Normierungskonstante c . Das LLTM schätzt somit die Schwierigkeit η_j der kognitiven Operationen und prüft das Modell anhand eines LQT (Gittler & Glück, 1998). Das LLTM kann somit als Strukturmodell für Veränderungsmessung verwendet werden. Gleiche Testaufgaben, die unter verschiedenen Zeitpunkten vorgegeben werden, können dabei als strukturell unterschiedliche Aufgaben aufgefasst werden, da angenommen wird, dass sich die Schwierigkeit der Items unter veränderten Bedingungen mit verändert (Gittler & Glück, 1998). Voraussetzung dafür ist, dass es sich bei den Aufgaben um Rasch-homogene Items handelt.

Aus der Veränderung der Itemschwierigkeiten können durch eine lineare Zerlegung der Itemparameter anhand einer Gewichtungsmatrix (Q) Treatment- wie auch Trendeffekte geschätzt werden. Treatmenteffekte entstehen aus der Wirkung der Behandlung, wohingegen Trendeffekt durch eine Testwiederholung gegeben werden können (Gittler, 1994). Auch in Untergruppen, wie beispielsweise bei unterschiedlichen Geschlechtern, kann das Ausmaß der Veränderung geschätzt werden.

Ob die Annahmen der Gewichtungsmatrix stimmen, kann mittels eines LQT ermittelt werden. Dabei handelt es sich um einen χ^2 -Test, der prüft, „whether the model fits the data significantly better when the item parameters are estimated separately for subgroups of subjects (H1), than when they are estimated based on the whole sample (H0)“ (Gittler & Glück, 1998, S. 73).

Anhand der Auswertung mit dem LLTM können einige Probleme der klassischen Testtheorie vermieden werden (Gittler, 2008):

- **methodisch fundierte Messung**

Die Messung wird empirisch begründet und nicht einfach vorausgesetzt. Gilt das Modell, so entspricht das numerische Relativ dem empirischen Relativ.

- **Überwindung von Ausgangswert-Abhängigkeit**

Das Problem der Decken- und Bodeneffekte in der klassischen Veränderungsmessung wird durch die Modelleigenschaft der spezifischen Objektivität überwunden. Effektparameter und daraus resultierende Veränderungen können unabhängig von Personenparametern modelliert werden.

- **Separierbarkeit der Effekte**

Die Veränderung als solches kann in einzelne Teil-Effekte zerlegt werden wie Treatmenteffekte, geschlechtsspezifische Effekte und Wechselwirkungen.

Insgesamt gesehen, geht die Veränderungsmessung nach dem LLTM ähnlich vor wie die klassisch-statistischen Verrechnung, betrachtet diese jedoch etwas genauer und differenzierter.

2.2.2 Veränderungsmessung nach Gittler und Fischer

Ein Vorteil der Veränderungsmessung nach Gittler und Fischer (2011) begründet sich darin, dass eine Veränderungsmessung selbst bei sehr kleinen Gruppen exakt durchgeführt werden kann. Die Nullhypothese lautet dabei, dass sich die Personengruppen gleichartig voneinander unterscheiden (Homogenitätsannahme). Ebenfalls können sehr kurzfristige Effekte sichtbar gemacht werden (Gittler & Fischer, 2011). Zudem können auch individuelle Daten und Leistungen betrachtet werden.

Im Gegensatz zum LLTM können dadurch exaktere Effektparameter errechnet und Tendenzen erkennbar gemacht werden.

Empirischer Teil

3. Zielsetzung und Forschungsfragen

Die vorliegende Arbeit hat das Ziel herauszufinden, ob ein Treatment anhand von Konstruktionsbaukästen im Schulunterricht die Raumvorstellungsleistung von Schülerinnen und Schülern der siebten Schulstufe verbessern kann. Dabei sollen auch geschlechtsspezifische Unterschiede untersucht werden.

Zur Überprüfung sollen drei verschiedene Messverfahren angewandt werden, welche im Zuge der Untersuchung validiert werden sollen. Bei diesen Messverfahren handelt es sich um die klassisch-statistische Analyse anhand von Fähigkeitsparametern, die Veränderungsmessung anhand des logistischen linearen Testmodells (LLTM) und der Veränderungsmessung nach Gittler und Fischer (2011).

Die daraus folgenden Forschungsfragen lauten:

1. Unterscheiden sich die Versuchs- und die Kontrollgruppe in ihren Leistungen signifikant zum ersten und zweiten Testzeitpunkt?
2. Sind signifikante Unterschiede in den Leistungen der Versuchs- und Kontrollgruppe zwischen den beiden Testzeitpunkten beobachtbar?
3. Können signifikante Geschlechtsunterschiede in der Differenz der Fähigkeitsparameter zwischen den beiden Testzeitpunkten aufgefunden werden?
4. Unterscheiden sich die Fähigkeitsparameter der Burschen und Mädchen zum ersten und zweiten Testzeitpunkt signifikant voneinander?
5. Können signifikante Unterschiede in der Bearbeitungszeit zwischen den beiden Testzeitpunkten aufgefunden werden?
6. Sind signifikante Unterschiede der Versuchs- und Kontrollgruppe in der Bearbeitungszeit zu beobachten?
7. Gibt es einen signifikanten Zusammenhang zwischen den Schulnoten und der Testleistung zum ersten und zweiten Testzeitpunkt?

Die Forschungsfragen sollen im Folgenden anhand der drei oben genannten Messverfahren beantwortet werden.

4. Planung und Durchführung

4.1 Versuchsplan

In der vorliegenden Untersuchung wurden drei Klassen eines niederösterreichischen Gymnasiums auf ihre Raumvorstellungsfähigkeit getestet. Die Untersuchung wurde als Längsschnittstudie mit zwei Testzeitpunkten geplant. Die Klassen befanden sich zu beiden Testzeitpunkten in der siebten Schulstufe. Eine Schulklasse wurde vom zuständigen Fachlehrer ausgewählt ein Training zur Verbesserung der Raumvorstellung im Rahmen des Schulunterrichts des Technischen Werkens zu durchlaufen. Das Trainingsprogramm ersetzte in dieser Schulklasse den regulären Schulunterricht, weshalb die Gruppe im Folgenden als Versuchsgruppe (VG) bezeichnet wird. Die beiden anderen Klassen erhielten kein Trainingsprogramm und folgten dem vorgesehenen Lehrplan des Technischen Werkens. Diese beiden Klassen werden zusammengefasst als Kontrollgruppe (KG) bezeichnet.

Aufgrund der fehlenden Möglichkeit der Zuweisung der SchülerInnen in die Versuchs- und Kontrollgruppe, wurde die Studie als quasi-experimenteller 2 Gruppen-Versuchsplan mit Messwiederholung durchgeführt (Abbildung 2).

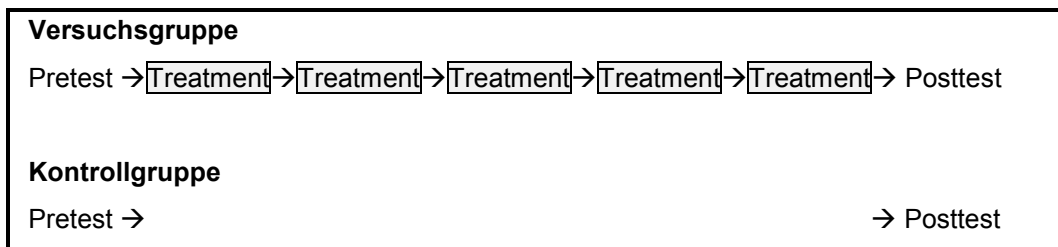


Abbildung 2: quasi-experimenteller 2 Gruppen-Versuchsplan mit Messwiederholung.

Die Treatments bestanden aus fünf separaten 90 minütigen Einheiten, welche durch den zuständigen Lehrer des Faches Technisches Werken entwickelt und durchgeführt wurden. Diese werden auszugsweise im Anhang 2 wiedergegeben. Die Treatments bestanden zum Großteil aus dem Nachbauen und Abzeichnen von Matador-Baukonstruktionen aus verschiedenen Perspektiven. Desweiteren mussten diverse Würfelaufgaben richtig gelöst werden. Der Fokus dieser Arbeit liegt in der Evaluation des Treatments, weswegen im Folgenden nicht näher auf den Inhalt der Treatments eingegangen werden soll.

4.2 Erhebungsinstrument

Um herauszufinden wie die Fähigkeiten einer Testpersonen in Bezug auf Raumvorstellung sind, wurde der Dreidimensionale Würfeltest (3DW) von Gittler (1990) vorgegeben. Aufgrund seiner testtheoretischen Fundierung im Rahmen der probabilistischen Testtheorie stellt er ein besonders wertvolles Erhebungsinstrument dar. Der Kategorisierung von Eliot und Smith (1983) zufolge, entspricht der 3DW der Kategorie *Block Rotation Task*.

4.2.1 Der Dreidimensionale Würfeltest (3DW) von Gittler (1990)

Bei dem 3DW handelt es sich um einen Powertest, der ab einem Alter von 13 Jahren empfohlen wird. Die intendierte Raumvorstellungsleistung liegt im „Manipulieren, Transformieren und/oder in Beziehung setzen von dreidimensionalen Objekten“ (Gittler, 1990, S. 8). Die Aufgabe der Testperson besteht darin, herauszufinden, welcher der Würfel A bis F den Würfel X in veränderter Lage darstellt, beziehungsweise ob keiner der Würfel dem Ausgangswürfel entspricht (Abbildung 3).

Beispiel II

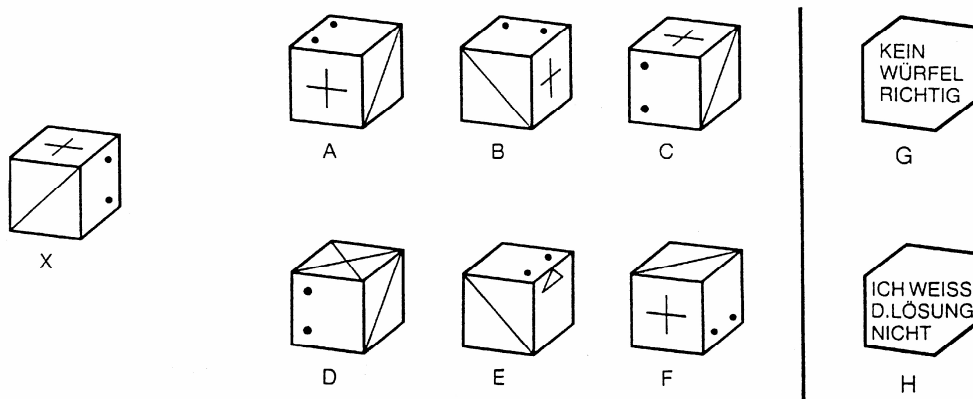


Abbildung 3: Übungsbeispiel II des 3DW.

Antwortwürfel G verhindert, dass die richtige Antwort nicht durch das Ausschließen falscher Antwortalternativen erworben wird (Falsifikationsstrategie), sondern durch Erkennen der richtigen Antwortwürfel (Verifikationsstrategie). Um die Ratewahrscheinlichkeit gering zu halten, wird noch eine weitere Antwortalternative H angeboten, die

dann gewählt werden soll, wenn die Testpersonen die Antwort nicht wissen, anstatt eine der anderen Lösungen zu wählen.

Der Test kann als Lang- oder Kurzform vorgegeben werden, wobei die Langform aus zwei Beispielaufgaben, einem Warmup-Item und 17 Testaufgaben besteht. Bei dem Warmup-Item handelt es sich um eine Flächenwürfelauflösungsaufgabe, die nicht in die Ermittlung des Rohwerts mit einbezogen wird (Gittler, 1990). Aufgrund des Vorliegens von generativen Regeln ist eine Entwicklung von Paralleltestversionen möglich.

Der 3DW ist für Individual- sowie Gruppentestungen geeignet. Da es sich um eine Powerleistung handelt und daher kein Zeitdruck zum Lösen der Aufgaben besteht, ist die Testdauer je nach Motivation, Leistungsfähigkeit und Arbeitstempo unterschiedlich. Es kann jedoch eine Bearbeitungszeit von 35-40 Minuten empfohlen werden.

Bei der Version des 3DW, die innerhalb der Studie Verwendung fand, wurde eine adaptierte Version des 3DW von Gittler (1990) vorgegeben (Anhang 1). Insgesamt wurden 15 Items vorgegeben, wobei es sich bei dem ersten Item um ein Warmup-Item handelte, so dass ein maximaler Score von 14 möglich war. Bei den Items handelte es sich um leichtere Aufgaben als die ursprünglichen Items des 3DW. Diese wurden aufgrund des jüngeren Alters der Testpersonen aus dem Itempool des 3DW ausgewählt und an das Niveau der Zielgruppe angepasst. Für die Bearbeitung der Aufgaben bestand keine Zeitlimitierung.

4.2.2 Konstruktion des 3DW

Der Dreidimensionale Würfeltest (3DW) wurde im Jahr 1990 von Gittler nach einer sechsjährigen Entwicklungszeit veröffentlicht. Hierbei basieren die Aufgaben auf dem Subtest 8 (Würfelauflösungsaufgaben) des Intelligenzstrukturtest (I-S-T 70) von Amthauer (1973). Bei der Generierung der Items anhand des Vorbilds des I-S-T 70 zeigte sich, dass drei verschiedene Aufgabenarten existierten. Bei der Drehung des Vorgabewürfels in den Lösungswürfel konnten insgesamt ein, zwei oder drei Lösungsschritte verwendet werden, wobei jeder Lösungsschritt eine 90°-Drehung erforderte. Dabei wurde vermutet, dass jeder zusätzliche Lösungsschritt die Aufgabe schwieriger zu lösen macht, so dass die Aufgaben mit nur einem Lösungsschritt am einfachsten und die Aufgabe mit drei Lösungsschritten am schwierigsten zu lösen sei. Dies war jedoch nicht der Fall. Da die Aufgaben mit zwei Lösungsschritten den Testpersonen leichter fielen als die Aufgaben mit einem Lösungsschritt, konnte das Kriterium der Eindimensionalität aller Testaufgaben nicht erzielt werden.

Zusätzlich zu den Lösungsschritten, konnten drei Bearbeitungsstrategien zu Lösung der Aufgaben eingesetzt werden (Gittler, 1990):

- **Raumstrategie**

Die Raumstrategie besteht darin, dass die Testpersonen die Bewegungen der Würfel in der Vorstellung nachvollziehen. Die Raumstrategie scheint dabei Ähnlichkeiten zum Faktor *spatial visualization* zu haben.

- **Relationsstrategie**

Die Relationsstrategie besteht darin, die Würfelflächen zueinander in Beziehung zu setzen und diese Relationen auf andere Objekte zu übertragen. Diese Strategie zeigt Ähnlichkeiten zum Faktor *spatial orientation* auf. Die Relationsstrategie scheint ebenfalls oft in Kombination mit der Raumstrategie angewendet zu werden.

- **Flächenstrategie**

Die Flächenstrategie besteht darin, den Würfel als zweidimensionales Objekt zu interpretieren und diesen lediglich um den Umkreismittelpunkt zu rotieren. Im Gegensatz zu der Raum- und Relationsstrategie kommen keine neuen Muster zu denen des Vorgabewürfels hinzu. Sie kann daher nur bei Aufgaben mit zwei Lösungsschritten angewendet werden.

Eine inhaltliche Analyse (Gittler, 1986) zeigte, dass bei Aufgaben mit ein- und dreiteiligen Lösungsschritten die Raum- und Relationsstrategie (Raumwürfelaufgaben) und bei zweiteiligen Lösungsschritten die Flächenstrategie (Flächenwürfelaufgaben) genutzt werden konnte.

Demnach schienen Raumwürfelaufgaben dazu geeignet zu sein räumliche Vorstellungsprozesse auszulösen, wohingegen Flächenwürfelaufgaben mit nicht-räumlichen Ausweichstrategien beantwortet werden konnten (Gittler, 1994). Damit waren die Flächenwürfelaufgaben kognitiv weniger aufwändig und leichter als die Raumwürfelaufgaben, was aber auch bedeutete, dass die Testaufgaben gemeinsam nicht Rasch-homogen sein können. Flächenwürfelaufgaben scheinen somit eine heterogene Subgruppe dazustellen, die nicht geeignet scheint, die beabsichtigte Raumvorstellungsdimension adäquat zu erfassen. Daher wurden die Flächenwürfelaufgaben aus

dem Testmaterial ausgeschlossen, so dass endgültig Rasch-homogene und konstruktvalide Raumwürfelaufgaben erstellt werden konnten (Gittler, 1999).

Essentiell ist demnach bei der Erstellung von Rasch-homogenen Aufgaben nicht nur ein technisches, sondern besonders auch ein theoriegeleitetes Vorgehen bei der Hypothesengenerierung. Eine technische Vorgehensweise bedeutet in dem Fall, dass auffällige Testaufgaben aufgrund statistischer Item-Kennzahlen (z-Werte Teststatistik) oder anhand von grafischen Modelltests ausgeschieden werden (Gittler, 1990). Dies heißt im Detail, dass es notwendig ist, vor der Durchführung der Modelltests (a-priori) Selektionshypothesen zu erstellen. Dazu gehört sich inhaltliche Gedanken zu den Items zu machen und sich zu fragen, ob sich diese homogen gegenüber dem Gesamtmaterial verhalten können oder ob einzelne Testaufgaben in Subgruppen aufgeteilt gehören. Bei einer rein technischen Vorgehensweise besteht demnach die Gefahr, dass es zu artifiziellen Modellanpassungen (Scheinhomogenität) kommen kann (Gittler, 1999), die bei einer Validierung nicht bestätigt werden kann.

4.2.3 Vergleich 3DW und I-S-T 70 (Würfelaufgaben)

Mittlerweile existieren vier Versionen des Intelligenzstrukturtests (IST). Der ursprüngliche IST wurde von Amthauer (1955) entwickelt. Weitere Versionen (IST 70, IST 2000, IST 2000 R) wurden im Laufe der Jahre hinzugefügt. Im Folgenden soll der 3DW mit dem IST 70 verglichen werden, auf dessen Aufbau der 3DW basiert.

Die Aufgabe der Testperson im I-S-T 70 in der Aufgabengruppe 8 (Würfelaufgaben) besteht darin, aus sechs verschiedenen Würfeln jenen Würfel auszuwählen, der dem vorgegebenen Würfel entspricht.

Eine genaue Betrachtung des Testmaterials des I-S-T zeigte jedoch einige Kritikpunkte auf. So konnten zwei unterschiedliche Strategien zur Lösung der Items angewendet werden, eine Verifikationsstrategie und eine Falsifikationsstrategie (Gittler, 1986; Gittler, 1990). Bei der Verifikationsstrategie, muss die Testperson sich aktiv für die richtige Antwort entscheiden, wohingegen bei der Falsifikationsstrategie die Testperson durch das systematisches Ausscheiden aller falschen Antwortkategorien die Aufgaben richtig lösen kann.

Letztere Antwortstrategie ist jedoch von dem her problematisch, als dass sie mit einem geringeren kognitiven Aufwand verbunden ist, da nur ein Widerspruch reicht um die Antwortalternative auszuschneiden (Gittler, 1990). Die Verifikationsstrategie verlangt

jedoch, dass eine völlige Widerspruchsfreiheit der Antwortalternative besteht, was mehrere Prüfschritte beinhaltet.

Einige Verbesserungsmaßnahmen die auch im 3DW durchgeführt worden sind, schlägt Gittler (1990) vor. So besteht eine Möglichkeit die Verifikationsstrategie bei Testpersonen zu erzwingen darin, zusätzliche eine Antwortalternative („keine Antwort richtig“) einzufügen. Weiterhin kann durch die Antwortalternative „ich weiß die Lösung nicht“, die Ratewahrscheinlichkeit gesenkt und damit eine Datenverfälschung verringert werden. Zusätzlich darf es nicht vorkommen, dass auf einem Würfel zweimal dasselbe Muster erscheint. Gleichzeitig darf im Vergleich zu den drei sichtbaren Mustern maximal ein neues Muster erscheinen. Schließlich sollten, um die Aufgabenschwierigkeit zu erhöhen, je Item sechs neue Antwortalternativen vorgegeben werden.

4.3 Datenerhebung

Die erste Testung fand im Februar 2011 (Testzeitpunkt 1: t1) und die zweite Testung im April 2011 (Testzeitpunkt 2: t2) statt, wobei zwischen den beiden Testzeitpunkten acht Wochen lagen. In dieser Zeit durchlief die Versuchsgruppe fünf Treatments á 100 Minuten.

Die Testung wurde anhand einer Papier-und-Bleistift-Version des 3DW im Gruppensetting durchgeführt, wobei jeweils eine kurze Einführung in das Testmaterial zu beiden Testzeitpunkten gegeben wurde.

Zusätzlich zur Vorgabe des 3DW wurde innerhalb der Testung das Alter, das Geschlecht, die Bearbeitungszeit, sowie die Schulnote im letzten Zeugnis in den Fächern Deutsch, Mathematik, Physik und Technisches Werken erhoben.

4.4 Stichprobenbeschreibung

Insgesamt nahmen 78 Kinder an der Untersuchung teil. 8 Personen mussten wegen Nicht-Erreichbarkeit zu einer der beiden Testzeitpunkte ausgeschieden werden und werden nicht weiter berücksichtigt. Für die weiteren Analysen standen somit $N = 70$ vollständige Daten zu Verfügung.

Die Stichprobe umfasst eine Versuchsgruppe von 25 Personen und eine Kontrollgruppe von 45 Personen. Die Verteilung des Geschlechts in der Gesamtstichprobe liegt bei 29 Burschen und 41 Mädchen, was einem Verhältnis von 41% zu 59% entspricht (Tabelle 1).

Tabelle 1: Verteilung des Geschlechts in der Versuchs- und Kontrollgruppe

	männlich	weiblich	Gesamt
VG	13 (52%)	12 (48%)	25 (35,7%)
KG	16 (35,6%)	29 (64,4%)	45 (64,3%)
Gesamt	29 (41,4%)	41 (58,6%)	70 (100%)

Anmerkung. Versuchsgruppe (VG), Kontrollgruppe (KG)

Die Verteilung der Geschlechter über die Gesamtstichprobe unterscheidet sich signifikant voneinander, $\chi^2(1) = 4.88$, $p < .05$. Der Altersmittelwert beträgt zum ersten Testzeitpunkt 12.89 Jahre und zum zweiten Testzeitpunkt 13.05 Jahre (Standardabweichung zu beiden Zeitpunkten $s = 0.36$). Tabelle 2 zeigt die Verteilung der Schulnoten in der Gesamtstichprobe.

Tabelle 2: Verteilung der Schulnoten in der Gesamtpopulation

		sehr gut	gut	befriedigend	ausreichend	nicht genügend	Gesamt
VG	Deutsch	7 (28%)	8 (32%)	3 (12%)	5 (20%)	2 (8%)	25 (100%)
	Mathematik	2 (8%)	8 (32%)	8 (32%)	6 (24%)	1 (4%)	25 (100%)
	Physik	6 (24%)	10 (40%)	6 (24%)	3 (12%)	-----	25 (100%)
	Technisches Werken	19 (76%)	4 (16%)	2 (8%)	-----	-----	25 (100%)
KG	Deutsch	9 (20%)	23 (51,1%)	11 (24,4%)	2 (4,4%)	-----	45 (100%)
	Mathematik	10 (22,2%)	18 (40%)	13 (28,9%)	4 (8,9%)	-----	45 (100%)
	Physik	43 (95,6%)	2 (4,4%)	-----	-----	-----	45 (100%)
	Technisches Werken	36 (80%)	6 (13,3%)	3 (6,7%)	-----	-----	45 (100%)

Anmerkung. Versuchsgruppe (VG), Kontrollgruppe (KG)

5. Ergebnisdarstellung

Im Folgenden werden zunächst die Fähigkeitsparameter anhand der Rohwerte ermittelt. Desweiteren wird eine statistische Prüfung der unterschiedlichen Fragestellungen anhand klassisch-statistischer Testverfahren vorgenommen.

Die Quantifizierung der Veränderung erfolgt zusätzlich anhand des linearen logistischen Testmodells (LLTM; Fischer, 1995) und der Veränderungsmessung nach Gittler und Fischer (2011).

5.1 Fähigkeits- und Itemparameter

Da für den 3DW das Rasch-Modell empirisch gilt (Gittler, 1990), können alle Modelleigenschaften und Modellannahmen angenommen und eine probabilistische Auswertung anhand des Rasch-Modells durchgeführt werden. Dieses besagt, dass die Wahrscheinlichkeit ein Item zu lösen ausschließlich von der Fähigkeit der Person (Fähigkeitsparameter) und der Schwierigkeit des Items (Itemparameter) abhängt (Bortz & Döring, 2006).

Zunächst werden die Itemparameter aus den Stichprobendaten geschätzt um nachträglich die Fähigkeitsparameter zurückrechnen zu können.

Tabelle 3: Schwierigkeits- und Leichtigkeitsparameter der Items

Item Nr.	σ	$-\sigma$
1	-2.0183	2.0183
2	-1.8546	1.8546
3	-1.0894	1.0894
4	-1.0387	1.0387
5	-.5176	0.5176
6	-.4900	0.4900
7	.0146	-0.0146
8	.0250	-0.025
9	.1622	-0.1622
10	.5660	-0.5660
11	.6328	-0.6328
12	.7047	-0.7047
13	1.1931	-1.1931
14	1.5658	-1.5658

Anmerkung. Schwierigkeitsparameter (σ), Leichtigkeitsparameter ($-\sigma$)

Zur Berechnung des LLTM und der Veränderungsmessung nach Gittler und Fischer (2011) werden die Schwierigkeits- beziehungsweise Leichtigkeitsparameter der Items verwendet (Tabelle 3). Negative Werte der Schwierigkeitsparameter σ lassen auf ein

leicht zu lösendes Item, positive Wert auf ein schwierig zu lösendes Items schließen. Für die Leichtigkeitsparameter gilt, dass negative Werte ein schwieriges Item, positive Werte ein leichtes Item kennzeichnen.

Die Fähigkeitsparameter liegen nach der Conditional Maximum Likelihood Schätzung (CML) im Bereich von $-\infty$ bis $+\infty$ (Warm, 1989). Da jedoch auch konkrete Schätzer für die Randwerte (alle beziehungsweise kein Item gelöst) erforderlich sind, werden zur weiteren Berechnung die Weighted Likelihood Schätzer (WLE) verwendet. Tabelle 4 zeigt die errechneten Schätzer für die $k = 14$ Items nach Warm (1989).

Tabelle 4: Conditional Maximum Likelihood Schätzer, sowie Weighted Likelihood Schätzer der Fähigkeitsparameter

Rohscores	CML	SE (CML)	WLE	SE (WLE)
0	999.0000	999.0000	-4.0902	1.8970
1	-3.1690	1.0788	-2.8269	1.0042
2	-2.3098	0.8176	-2.1463	0.7984
3	-1.7336	0.7116	-1.6365	0.7043
4	-1.2703	0.6547	-1.2086	0.6516
5	-0.8649	0.6215	-0.8266	0.6202
6	-0.4913	0.6029	-0.4713	0.6026
7	-0.1333	0.5956	-0.1297	0.5956
8	0.2222	0.5986	0.2089	0.5984
9	0.5879	0.6131	0.5552	0.6121
10	0.9804	0.6430	0.9223	0.6401
11	1.4264	0.6977	1.3301	0.6903
12	1.9811	0.8034	1.8151	0.7831
13	2.8164	1.0672	2.4683	0.9886
14	999.0000	999.0000	3.7042	1.8738
Gesamt	-1.9577		-2.3321	

Anmerkung. Conditional Maximum Likelihood Schätzer (CML), Standardabweichung der Conditional Maximum Likelihood Schätzer (SE(CML)), Weighted Likelihood Schätzer (WLE), Standardabweichung der Weighted Likelihood Schätzer (SE (WLE)),

Negative Werte der Fähigkeitsparameter lassen auf eine geringe Personenfähigkeit, positive Werte auf eine hohe Personenfähigkeit schließen. (Bortz & Döring, 2006). Erst die Fähigkeitsparameter ermöglichen es echte Differenzen zu ermitteln.

Für die Veränderung der Testleistung wurde die Differenz der Fähigkeitsparameter zum ersten und zweiten Testzeitpunkt errechnet.

Tabelle 5: Häufigkeiten der Rohscores zum ersten und zweiten Testzeitpunkt

	t1															
t2	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Gesamt
0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	3
1	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	3
2	1	0	1	0	2	1	0	0	0	0	0	0	0	0	0	5
3	0	3	1	1	2	0	0	0	0	0	0	0	0	0	0	7
4	1	2	2	1	0	0	0	0	0	0	0	0	0	0	0	6
5	0	0	2	1	6	2	2	1	0	0	0	0	0	0	0	14
6	0	0	0	0	3	2	1	1	1	1	0	0	0	0	0	9
7	0	0	0	0	1	2	2	1	0	0	0	0	0	0	0	6
8	0	0	0	1	2	0	1	0	0	0	0	0	0	0	0	4
9	0	0	0	0	0	2	0	0	0	1	0	0	1	0	0	4
10	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	3
11	0	0	0	0	0	1	0	1	0	0	0	2	0	0	0	4
12	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	2
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gesamt	3	8	6	4	17	11	7	4	1	4	0	3	1	1	0	70

Anmerkung. Testzeitpunkt 1(t1), Testzeitpunkt 2 (t2)

Tabelle 5 gibt die Häufigkeiten der Rohscores an, welche zum ersten und zweiten Testzeitpunkt erreicht wurden.

Vorab soll dem Leser die Veränderung der Kontroll- und Versuchsgruppe (Abbildung 4), sowie die Veränderung der Burschen und Mädchen (Abbildung 5) vom ersten zum zweiten Testzeitpunkt bildlich dargestellt werden.

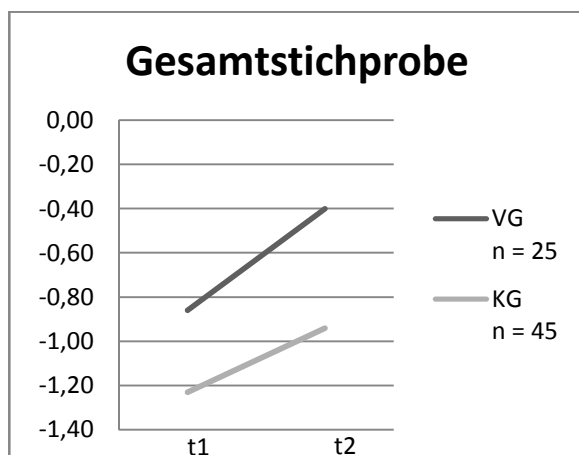


Abbildung 4: 3DW-Fähigkeitsparameter der Versuchs- und Kontrollgruppe.

Abbildung 4 verdeutlicht, dass die Versuchsgruppe zu beiden Testzeitpunkten höhere Werte bezüglich der Fähigkeitsparameter besitzt als die Kontrollgruppe.

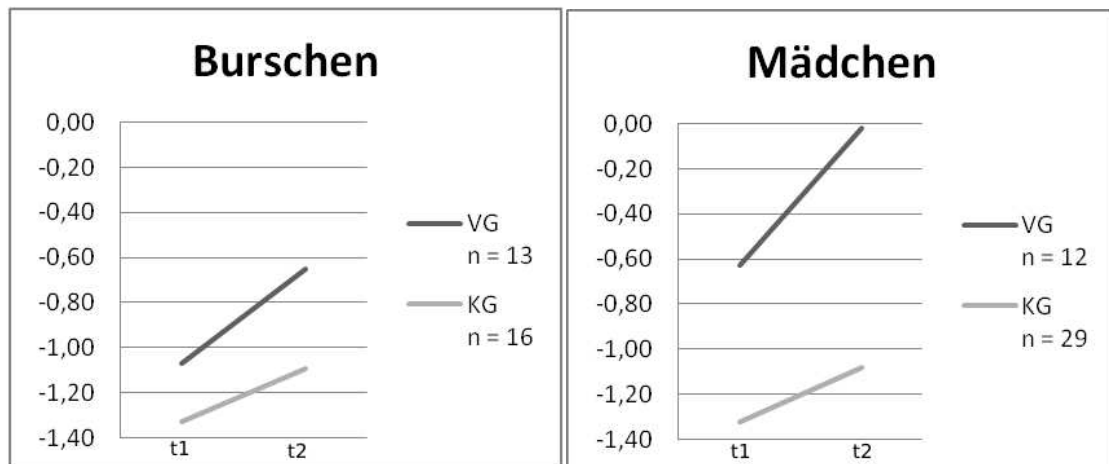


Abbildung 5: 3DW-Fähigkeitsparameter männlicher und weiblicher SchülerInnen zu beiden Testzeitpunkten.

Der graphische Verlauf der Fähigkeitsparameter der Burschen und Mädchen (Abbildung 5) lässt erkennen, dass die Versuchsgruppe der Burschen und Mädchen zu beiden Testzeitpunkten in ihren Leistungen über denen der Kontrollgruppe liegt.

5.2 Analysen der Itemparameterdifferenzen mittels klassisch-statistischer Verfahren

Im Folgenden sollen die unterschiedlichen Forschungsfragen anhand von klassisch-statistischen Verfahren mit Hilfe von Fähigkeitsparametern untersucht werden.

5.2.1 Leistungsunterschiede zwischen den Testzeitpunkten

Zunächst soll die Frage geklärt werden, ob es aufgrund der Testwiederholung zu einem Lerneffekt innerhalb der Versuchs- und Kontrollgruppe kommt. Zur Berechnung der Leistungsunterschiede zwischen den Testzeitpunkten, werden die Fähigkeitsparameter jeweils zu Testzeitpunkt 1 und Testzeitpunkt 2 verwendet. Aufgrund der fehlenden Normalverteilung der Fähigkeitsparameter in der Versuchsgruppe wird mit einem parameterfreien Verfahren gerechnet.

Tabelle 6: Mittelwertunterschiede der Fähigkeitsparameter der Versuchsgruppe zwischen t1 und t2 mittels Wilcoxon Test

Wilcoxon Test

		Fähigkeitsparameter			Z	p
		n	MW	s		
VG	t1	25	-.86	1.37	-2.10	.036*
	t2	25	-.40	1.39		

Anmerkung. Versuchsgruppe (VG), Testzeitpunkt 1 (t1), Testzeitpunkt 2 (t2), * signifikant bei $\alpha = 5\%$

Es zeigt sich ein signifikanter Unterschied zwischen dem ersten und zweiten Testzeitpunkt in der Versuchsgruppe (Tabelle 6). Demnach verbessert sich die Versuchsgruppe in ihren Leistungen im 3DW zum zweiten Testzeitpunkt.

Für die Kontrollgruppe wird aufgrund der fehlenden Normalverteilung ebenfalls mit einem parameterfreien Verfahren gerechnet.

Tabelle 7: Mittelwertunterschiede der Fähigkeitsparameter der Kontrollgruppe zwischen t1 und t2 mittels Wilcoxon Test

Wilcoxon Test

		Fähigkeitsparameter			Z	p
		n	MW	s		
KG	t1	45	-1.23	1.26	-2.22	.026*
	t2	45	-.94	1.21		

Anmerkung. Kontrollgruppe (KG), Testzeitpunkt 1 (t1), Testzeitpunkt 2 (t2), * signifikant bei $\alpha = 5\%$

Die Ergebnisse (Tabelle 7) zeigen auch in der Kontrollgruppe einen signifikanten Unterschied zwischen dem ersten und zweiten Testzeitpunkt.

Somit kann festgestellt werden, dass sowohl die Versuchs- als auch die Kontrollgruppe sich in ihren Leistungen vom ersten zum zweiten Testzeitpunkt signifikant verbessert haben.

5.2.2 Leistungsunterschiede zwischen den Testgruppen

Dieser Abschnitt beschäftigt sich mit der Frage, ob sich die Versuchs- und Kontrollgruppe signifikant voneinander in ihren Leistungen unterscheiden und somit ein Effekt des Treatments beobachtbar ist. Um die Leistungsveränderung feststellen zu können, werden die Fähigkeitsparameter des ersten Testzeitpunkts von dem des zweiten Testzeitpunkts subtrahiert, womit sich eine Differenz der Fähigkeitsparameter ergibt. Somit können inferenzstatistische Analysen angewendet werden (Gittler, 2008).

Tabelle 8: Mittelwertunterschiede der Differenz der Fähigkeitsparameter zwischen t1 und t2 bezüglich der VG und KG mittels U-Test nach Mann-Whitney

U-Test nach Mann-Whitney

	Differenz der Fähigkeitsparameter zwischen t1 und t2					
	n	MW	s	U	Z	p
VG	25	.46	0.92	535.50	-.33	.740
KG	45	.30	1.04			

Anmerkung. Kontrollgruppe (KG), Versuchsgruppe (VG), Testzeitpunkt 1 (t1), Testzeitpunkt 2 (t2)

Da die Normalverteilung für die Kontrollgruppe, überprüft mittels Kolmogorov-Smirnov-Test, nicht gegeben ist, muss mit einem parameterfreien Verfahren gerechnet werden. Das Ergebnis (Tabelle 8) zeigt, dass sich die beiden Testgruppen in der Differenz der Fähigkeitsparameter nicht signifikant voneinander unterscheiden.

5.2.3 Geschlechtsspezifische Unterschiede in der Differenz der Fähigkeitsparameter zwischen den Testzeitpunkten

Die Frage, ob es geschlechtsspezifische Unterschiede in der Versuchsgruppe bezüglich der Differenz der Fähigkeitsparameter zwischen dem ersten und zweiten Testzeitpunkt gibt, kann anhand eines parametrischen Verfahrens gerechnet werden, da die Normalverteilung und Homogenität der Varianzen gegeben ist.

Tabelle 9: Geschlechtsspezifische Mittelwertunterschiede bezüglich der Differenz der Fähigkeitsparameter zwischen t1 und t2 innerhalb der Versuchsgruppe mittels t-Test für unabhängige Stichproben

t-Test für unabhängige Stichproben

		Differenz der Fähigkeitsparameter zwischen t1 und t2				
		n	MW	s	t	p
VG	Burschen	13	.33	0.88	-.78	.446
	Mädchen	12	.61	0.97		

Anmerkung. Versuchsgruppe (VG), Testzeitpunkt 1 (t1), Testzeitpunkt 2 (t2)

In der Versuchsgruppe unterscheiden sich Burschen und Mädchen nicht signifikant voneinander in der Differenz der Fähigkeitsparameter zwischen dem ersten und zweiten Testzeitpunkt (Tabelle 9).

Für die Kontrollgruppe kann ebenfalls mit einem parametrischen Verfahren gerechnet werden, da die Normalverteilung und Homogenität der Varianzen gegeben ist.

Tabelle 10: Geschlechtsspezifische Mittelwertunterschiede bezüglich der Differenz der Fähigkeitsparameter zwischen t1 und t2 innerhalb der Kontrollgruppe mittels t-Test für unabhängige Stichproben

t-Test für unabhängige Stichproben

		Differenz der Fähigkeitsparameter zwischen t1 und t2				
		n	MW	s		
KG	Burschen	16	.41	1.27	.53	.599
	Mädchen	29	.24	0.91		

Anmerkung. Kontrollgruppe (KG), Testzeitpunkt 1 (t1), Testzeitpunkt 2 (t2)

Tabelle 10 zeigt, dass auch die Burschen und Mädchen der Kontrollgruppe sich nicht signifikant voneinander in der Differenz der Fähigkeitsparameter zwischen dem ersten und zweiten Testzeitpunkt unterscheiden.

Zusammenfassend kann festgestellt werden, dass sowohl in der Versuchs- als auch in der Kontrollgruppe keine signifikanten geschlechtsspezifischen Unterschiede in der Differenz der Fähigkeitsparameter aufgefunden werden konnten.

5.2.4 Geschlechtsspezifische Unterschiede in den Fähigkeitsparametern des ersten und zweiten Testzeitpunkts

Möglicherweise zeigt sich das Treatment eher für Mädchen wirksam als für Burschen. Ähnliche geschlechtsspezifische Effekte konnten bereits in einer Studie zum Unterricht von Geometrisch Zeichnen beobachtet werden (Gittler, Arendasy, & Hagen, 2006). Aus diesem Grund werden zusätzlich die geschlechtsspezifischen Leistungsunterschiede zu beiden Testzeitpunkten jeweils für Burschen und Mädchen separat errechnet. Da die Normalverteilung und Homogenität der Varianzen für die Mädchen in beiden Testgruppen gegeben ist, wird mit einem parametrischen Verfahren gerechnet.

Tabelle 11: Mittelwertunterschiede in den Fähigkeitsparametern der Mädchen der Versuchs- und Kontrollgruppe zu t1 mittels t-Test für unabhängige Stichproben

t-Test für unabhängige Stichproben

		Fähigkeitsparameter				
		n	MW	s		
t1	KG	29	-1.33	1.30	1.53	.133
	VG	12	-.63	1.38		

Anmerkung. Kontrollgruppe (KG), Versuchsgruppe (VG), Testzeitpunkt 1 (t1)

Für den ersten Testzeitpunkt zeigen sich keine Unterschiede in den Fähigkeitsparametern der Mädchen der Versuchs- und Kontrollgruppe (Tabelle 11).

Tabelle 12: Mittelwertunterschiede in den Fähigkeitsparametern der Mädchen der Versuchs- und Kontrollgruppe zu t2 mittels t-Test für unabhängige Stichproben

t-Test für unabhängige Stichproben

		Fähigkeitsparameter				
		n	MW	s	t	p
t2	KG	29	-1.09	1.15	2.66	.011**
	VG	12	-.02	1.15		

Anmerkung. Kontrollgruppe (KG), Versuchsgruppe (VG), Testzeitpunkt 2 (t2), ** signifikant bei $\alpha = 1\%$

Für den zweiten Testzeitpunkt (Tabelle 12) können jedoch signifikante Unterschiede für die Mädchen der Versuchs- und Kontrollgruppe aufgefunden werden.

Gleiches kann ebenfalls für die Burschen der Versuchs- und Kontrollgruppe berechnet werden. Da die Normalverteilung nicht gegeben ist, wird mit einem parameterfreien Verfahren gerechnet.

Tabelle 13: Mittelwertunterschiede in den Fähigkeitsparametern der Burschen der Versuchs- und Kontrollgruppe zu t1 mittels U-Test nach Mann-Whitney

U-Test nach Mann-Whitney

		Fähigkeitsparameter					
		n	MW	s	U	Z	p
t1	KG	16	-1.07	.75	75	-1.31	.19
	VG	13	-1.07	1.38			

Anmerkung. Kontrollgruppe (KG), Versuchsgruppe (VG), Testzeitpunkt 1 (t1)

Für die Burschen der zeigen sich hingegen keine Unterschiede in den Fähigkeitsparametern zum ersten Testzeitpunkt (Tabelle 13).

Tabelle 14: Mittelwertunterschiede in den Fähigkeitsparametern der Burschen der Versuchs- und Kontrollgruppe zu t2 mittels U-Test nach Mann-Whitney

U-Test nach Mann-Whitney

		Fähigkeitsparameter					
		n	MW	s	U	Z	p
t2	KG	16	-.65	1.31	102	-.11	.91
	VG	13	-.75	1.48			

Anmerkung. Kontrollgruppe (KG), Versuchsgruppe (VG), Testzeitpunkt 2 (t2)

Auch zum zweiten Testzeitpunkt sind keine signifikanten Unterschiede in den Fähigkeitsparametern der Burschen der Versuchs- und Kontrollgruppe zu beobachten (Tabelle 14).

Die Ergebnisse zeigen somit lediglich für die Mädchen zum zweiten Testzeitpunkt eine signifikante Veränderung auf. Diese sind dahingehend, dass die Mädchen der Versuchsgruppe höhere Fähigkeitsparameter zum zweiten Testzeitpunkt erzielen als Mädchen der Kontrollgruppe. Mögliche Ursachen könnten darin liegen, dass es sich

bei den Mädchen der Versuchsgruppe von vornherein um eine besonders leistungsstarke Gruppe handelt, da diese schon zum ersten Testzeitpunkt auffallen hohe Fähigkeitsparameter aufzeigen (siehe Abbildung 5).

5.2.5 Unterschiede in der Bearbeitungszeit zwischen den Testzeitpunkten

Es stellt sich weiterhin die Frage, ob sich die Versuchs- und Kontrollgruppe signifikant in ihrer Bearbeitungszeit vom ersten zum zweiten Testzeitpunkt unterscheiden. Tabelle 15 gibt einen Überblick über die Mediane der Bearbeitungszeit in Minuten.

Tabelle 15: Übersicht über die Mediane der Bearbeitungszeit

	Bearbeitungszeit		
	n	MD	
		t1	t2
VG	25	17.00	12.00
KG	45	15.00	12.00

Anmerkung: Versuchsgruppe (VG), Kontrollgruppe (KG), Testzeitpunkt 1 (t1), Testzeitpunkt 2 (t2)

Da bei Bearbeitungszeiten aufgrund der Unmöglichkeit negativer Zeiten generell eine schiefe Verteilung zu erwarten ist, soll in diesem Fall mit parameterfreien Verfahren gerechnet werden.

Tabelle 16: Mittelwertunterschiede in der Bearbeitungszeit der Versuchsgruppe zwischen den Testzeitpunkten mittels Wilcoxon Test

Wilcoxon Test

		Bearbeitungszeit				
		n	MW	s		
VG	t1	25	16.40	4.95	-3.44	.001**
	t2	25	12.12	5.58		

Anmerkung: Versuchsgruppe (VG), Testzeitpunkt 1 (t1), Testzeitpunkt 2 (t2), * signifikant bei $\alpha = 1\%$

Tabelle 16 zeigt, dass sich die Versuchsgruppe signifikant in ihrer Bearbeitungszeit vom ersten zum zweiten Testzeitpunkt unterscheidet. Die Mittelwerte zeigen, dass diese für beide Gruppen zum zweiten Testzeitpunkt deutlich reduziert ist.

Tabelle 17: Mittelwertunterschiede in der Bearbeitungszeit der Kontrollgruppe zwischen den Testzeitpunkten mittels Wilcoxon Test

Wilcoxon Test

		Bearbeitungszeit				
		n	MW	s	Z	p
KG	t1	45	15.20	5.96	-4.59	.001**
	t2	45	11.09	3.22		

Anmerkung. Kontrollgruppe (KG), Testzeitpunkt 1 (t1), Testzeitpunkt 2 (t2), * *signifikant bei $\alpha = 1\%$

Auch die Kontrollgruppe (Tabelle 17) unterscheidet sich signifikant in ihrer Bearbeitungszeit vom ersten zum zweiten Testzeitpunkt. Diese ist zum zweiten Testzeitpunkt im Vergleich zum ersten Testzeitpunkt reduziert.

Beide Testgruppen zeigen somit eine schnellere Bearbeitungszeit zum zweiten Testzeitpunkt auf.

5.2.6 Unterschiede in der Bearbeitungszeit zwischen den Testgruppen

Desweiteren stellt sich die Frage, ob sich die Versuchs- und die Kontrollgruppe in ihrer Bearbeitungszeit signifikant voneinander unterscheiden. Aufgrund der fehlenden Normalverteilung der Kontrollgruppe wird jeweils mit einem parameterfreien Verfahren gerechnet.

Tabelle 18: Mittelwertunterschiede in der Bearbeitungszeit der Kontroll- und Versuchsgruppe zu t1 mittels U-Test nach Mann-Whitney

U-Test nach Mann-Whitney

		Bearbeitungszeit					
		n	MW	s	U	Z	p
t1	VG	25	16.40	4.95	491	-.88	.379
	KG	45	15.20	5.96			

Anmerkung. Kontrollgruppe (KG), Versuchsgruppe (VG), Testzeitpunkt 1 (t1)

Tabelle 18 zeigt, dass sich die Versuchs- und Kontrollgruppe zum ersten Testzeitpunkt nicht signifikant voneinander in der Bearbeitungszeit unterscheiden.

Tabelle 19: Mittelwertunterschiede in der Bearbeitungszeit der Kontroll- und Versuchsgruppe zu t2 mittels U-Test nach Mann-Whitney

U-Test nach Mann-Whitney

		Bearbeitungszeit			U	Z	p
		n	MW	s			
t2	VG	25	12.12	5.58	525.5	-.46	.648
	KG	45	11.09	3.22			

Anmerkung. Kontrollgruppe (KG), Versuchsgruppe (VG), Testzeitpunkt 2 (t2)

Auch zum zweiten Testzeitpunkt unterscheiden sich die Versuchs- und Kontrollgruppe nicht signifikant in der Bearbeitungszeit der Testaufgaben (Tabelle 19).

Somit kann festgestellt werden, dass die Versuchs- und die Kontrollgruppe sich jeweils zum ersten und zum zweiten Testzeitpunkt nicht signifikant in ihrer Bearbeitungszeit voneinander unterscheiden.

5.2.7 Zusammenhang zwischen Schulnoten und Testleistung

Es stellt sich die Frage ob die Schulnoten in Zusammenhang stehen mit den Testleistungen im 3DW. Da Noten nicht intervallskaliert sind, kommt jeweils eine Spearman Korrelation zu Einsatz. Im Falle von Schulnoten ist ein negativer Zusammenhang zu erwarten, da mit besseren Schulnoten höhere Fähigkeitsparameter zu erwarten sind.

Tabelle 20: Korrelation der Schulnoten mit den Fähigkeitsparametern zu t1

Spearman Korrelation

	Fähigkeitsparameter zu t1				
	n	MW	s	r	p
Deutschnote	70	2.26	1.02	-.23	.059
Mathematiknote	70	2.46	.99	-.34	.004**
Physiknote	70	1.47	.83	-.08	.640
Note in Technischem Werken	70	1.30	.60	-.10	.417

Anmerkung. Testzeitpunkt 1 (t1), Testzeitpunkt 2(t2), ** signifikant bei $\alpha = 1\%$

Tabelle 21: Korrelation der Schulnoten mit den Fähigkeitsparametern zu t2

Spearman Korrelation

	Fähigkeitsparameter zu t2				
	n	MW	s	r	p
Deutschnote	70	2.26	1.02	-.22	.063
Mathematiknote	70	2.46	.99	-.34	.005**
Physiknote	70	1.47	.83	-.07	.577
Note in Technischem Werken	70	1.30	.60	-.14	.257

Anmerkung. Testzeitpunkt 1 (t1), Testzeitpunkt 2(t2), ** signifikant bei $\alpha = 1\%$

Einzig die Mathematiknote korreliert signifikant mit den Fähigkeitsparametern zum ersten (Tabelle 20) und zweiten Testpunkt (Tabelle 21). Je höher demnach der Fähigkeitsparameter, desto besser ist die Mathematiknote. Eine ähnliche Tendenz kann für die Deutschnote zu beiden Testzeitpunkten beobachtet werden. Für die Physiknote und die Note im Unterrichtsfach Technisches Werken konnte kein Zusammenhang festgestellt werden.

Dieses Ergebnis ist insofern auffällig, als dass in früheren Untersuchungen für Physik ein hoher Zusammenhang mit den Leistungen im 3DW festgestellt werden konnte (Gittler, 1990). Das Fehlen derartiger Ergebnisse in dieser Stichprobe mag damit zusammenhängen, dass 70% der SchülerInnen mit „sehr gut“ im Fach Physik benotet wurden, so dass eine Differenzierung kaum möglich ist.

5.3 Analysen mittels Linearem logistischen Testmodell (LLTM)

Eine weitere Analyse soll anhand des LLTM durchgeführt werden. Aufgrund der Annahme der lokalen stochastischen Unabhängigkeit des Rasch-Modells, können die Daten beider Testzeitpunkt so umorganisiert werden, dass die Datenmatrix des zweiten Testzeitpunkts an den ersten Testzeitpunkt angereiht wird (Gittler, 2008). Dabei geht man in diesem Fall von $2 \times 14 = 28$ strukturell unterschiedlichen (virtuellen) Items aus.

Zum ersten Testzeitpunkt gab es aufgrund der gleichen Itemschwierigkeiten für beide Testgruppen keine strukturellen Unterschiede, weswegen gleiche Itemnummern vergeben werden konnten. Zum zweiten Testzeitpunkt werden unterschiedliche Itemnummern vergeben, da es aufgrund der Wiederholungstestung zu unterschiedlichen Effekten kommen kann. Das quasi-gesättigte Rasch-Modell (Tabelle 22) geht davon aus, dass die Aufgaben zum zweiten Testzeitpunkt von jeweils einer Testgruppe bearbeitet wurden, weswegen die Daten unvollständig sind. Insgesamt geht man daher im quasigesättigten Rasch-Modell von $df = 42 - 1 = 41$ Freiheitsgraden aus. Die logarithmierte Likelihood der Daten beträgt $\ln L = -801.52$.

Tabelle 22: unvollständiges quasi-gesättigtes Rasch-Modell (qRM)

Gruppe	t1	t2	
	Item Nr.	Nr. der virtuellen Items	
KG	1 ... 14	15 ... 28	-----
VG	1 ... 14	-----	29 ... 42

Anmerkung. Testzeitpunkt 1 (t1), Testzeitpunkt 2(t2), Kontrollgruppe (KG), Versuchsgruppe (VG)

Das gesättigte Rasch-Modell (Tabelle 23) hingegen geht davon aus, dass die unterschiedlichen Testgruppen zum zweiten Testzeitpunkt gleiche Aufgaben bearbeiteten. Daher werden $df = 28 - 1 = 27$ Freiheitsgrade angenommen.

Tabelle 23: vollständiges gesättigtes Rasch-Modell (RM)

Gruppe	t1 Item Nr.	t2 Nr. der virtuellen Items
KG	1 ... 14	15 ... 28
VG	1 ... 14	15 ... 28

Anmerkung. Testzeitpunkt 1 (t1), Testzeitpunkt 2(t2), Kontrollgruppe (KG), Versuchsgruppe (VG)

Zunächst soll die Frage geklärt werden, ob das gesättigte Rasch-Modell (RM) die Daten mit weniger Parametern genauso gut beschreibt wie das quasi-gesättigte Rasch-Modell (qRM). Hierfür wird ein Likelihood-Quotiententest (LQT) von Andersen (1973) durchgeführt, durch den überprüft werden kann, ob die Daten durch ein restriktiveres Modell mit weniger unabhängigen Parametern genauso gut erklärt werden können (Tabelle 25). Fällt der Modelltest nicht signifikant aus, so sind keine Einbußen im Erklärungswert durch die geringe Anzahl an Parametern zu befürchten. Schließlich sollte jenes Modell bestimmt werden, welches mit möglichst wenigen Parametern auskommt.

Der LQT bezüglich des gesättigten Rasch-Modells (RM) und des quasi-gesättigten Rasch-Modells (qRM) fiel positiv aus, $p > .05$ (Tabelle 25), was bedeutet, dass die Daten durch das RM genauso gut beschrieben werden können, wie durch das qRM.

Um zu überprüfen, ob es notwendig ist 27 freie Parameter zur Beschreibung der Daten zu verwenden, oder ob diese auch als identisch zur Ersttestung mit einer zusätzlichen Zahl an Effektparametern zu beschreiben sind, lässt sich anhand einer Gewichtszahlenmatrix Q überprüfen (Tabelle 24).

Tabelle 24: LLTM Gewichtszahlenmatrix ($df = 16$) für Modell 1

	Item-Nr. im 3DW	virtuelle Item-Nr.	Itemparameter 1 2 3 4 5 6 7 8 9 10 11 12 13	Effektparameter <i>trend treatment sex</i>		
t1, KG männlich (n = 16)	1	1	1	0	0	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	13	13	.	.	.	.
t2, KG männlich (n = 16)	1	15	1	1	0	1
	.	.	.	.	.	.
	.	.	.	.	.	.
	13	27	.	.	.	.
t1, KG weibliche (n = 29)	1	1	1	0	0	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	13	13	.	.	.	.
t2, KG weiblich (n = 29)	1	29	1	1	0	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	13	41	.	.	.	.
t1, VG männlich (n = 13)	1	1	1	0	0	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	13	13	.	.	.	.
t2, VG männlich (n = 13)	1	43	1	1	1	1
	.	.	.	.	.	.
	.	.	.	.	.	.
	13	55	.	.	.	.
t1, VG weibliche (n = 12)	1	1	1	0	0	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	13	13	.	.	.	.
t2, VG weiblich (n = 12)	1	57	1	1	1	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	13	69	.	.	.	.
	14	70	0 0 0 0 0 0 0 0 0 0 0 0 0 0	1	1	0

Anmerkung. Testzeitpunkt 1 (t1), Testzeitpunkt 2 (t2), Versuchsgruppe (VG), Kontrollgruppe (KG)

Die Veränderungsmessung anhand des LLTM ergibt eine Gewichtungsmatrix Q mit 16 unabhängigen Basisparametern (13 Itemparameter und 3 Effektparameter), wobei der vierzehnte Itemparameter zur Normierung verwendet wurde. Die Eintragung einer „1“ in die Gewichtungsmatrix gibt an, welche Basisparameter für das technische Item (Item-Nr.) von Bedeutung sind. Die Eintragung einer „0“ drückt aus, dass keine Verbindung zwischen dem Basisparameter und dem technischen Item angenommen wird (Gittler, 1994). Die Itemparameter in der Gewichtungsmatrix außerhalb der Hauptdiagonalen die nicht explizit angegeben sind betragen Null.

Der Parameter *trend* beschreibt die Effekte durch die Testwiederholung, daher ist dieser jeweils zum zweiten Testzeitpunkt der Versuchs- und Kontrollgruppe mit „1“ beziffert. Der Parameter *treatment* erfasst hingegen die Effekte des Treatments, die innerhalb der Versuchsgruppe zu erwarten wären, weshalb die Versuchsgruppe (Burschen und Mädchen) zum zweiten Testzeitpunkt mit „1“ gewichtet wird. Schließlich beschreibt der Parameter *sex* die geschlechtsspezifischen Effekte des Treatments. Hierfür werden die Burschen der Versuchs- und Kontrollgruppe mit „1“ gewichtet, wohingegen die Mädchen ungewichtet bleiben.

Ob die Annahmen der Gewichtungsmatrix stimmen, kann jeweils anhand eines Likelihood-Quotiententests ermittelt werden. Der erste durchgeführte LQT misst, ob die Daten des quasi-gesättigten Rasch-Modells (qRM) durch das restriktivere gesättigte Rasch-Modell (RM) genauso gut erklärt werden können. Dafür werden die Daten in folgende Formel eingesetzt (Gittler & Glück, 1998):

$$\chi^2 = -2 \ln[L(\text{data} | \text{qRM}: \sigma_1, \dots, \sigma_{41}) / L(\text{data} | \text{RM}: \sigma_1, \dots, \sigma_{27})]$$

Die Freiheitsgrade entsprechen der Differenz der Parameter des qRM und des gesättigten RM, was in diesem Fall $df = 41 - 27 = 14$ entspricht.

Fällt der LQT insignifikant aus, erklärt das restriktivere Modell (RM) die Daten genauso gut wie das weniger restriktive Modell (qRM) und es können weitere, noch restriktivere Modelle getestet werden (Gittler & Glück, 1998). Das Ziel ist es hierbei jenes Modell zu finden, das die Daten am besten erklärt (Tabelle 25).

Tabelle 25: Ergebnisse der Hypothesenprüfung

Modell	k	lnL	Effektparameter			LQT		
			<i>trend</i>	<i>treatment</i>	<i>sex</i>	χ^2	df	p
quasi-gesättigtes Modell (qRM)	41	-801.52						
gesättigtes Modell (RM)	27	-809.65				zu qRM: 16.24	14	.299
Modell 1	16	-815.23	.36**	.14	-.04	zu RM: 11.16	11	.430
Modell 2a	15	-815.24	.35**	.13	-	zu Modell 1: .03	1	.875
Modell 2b	15	-815.39	.40**	-	-.01	zu Modell 1: .33	1	.563
Modell 3a	14	-815.40	.39**	-	-	zu Modell 2a: .31	1	.567
						zu qRM: 27.76	27	.423
Modell 3b	14	-818.26	-	.48**	-	zu Modell 2a: 6.04	1	.014*
Modell 3c	14	-818.89	-	-	.39*	zu Modell 2a: 7.30	1	.007**

Anmerkung. Parameteranzahl (k), Log-Likelihood (lnL), χ^2 -Wert (χ^2), Freiheitsgrade (df), Signifikanz (p), * signifikant bei $\alpha = 5\%$; ** signifikant bei $\alpha = 1\%$

Der insignifikante Modelltest des gesättigten Modells ($p = .299$) zeigt, dass die Daten durch das gesättigte Modell (RM) genauso gut erklärt werden können wie durch das quasi-gesättigte Modell (qRM). Der darauffolgende Modelltest 1 zeigt, dass auch das Modell mit $k = 16$ Basisparametern die Daten nicht schlechter erklärt als das gesättigte Modell mit $k = 27$ Parametern.

Im Modell 2a wird derjenige Effektparameter entfernt, der den geringsten Erklärungswert aufzeigt beziehungsweise am nächsten bei Null ist. Hierbei handelt es sich um den geschlechtsspezifischen Effekt *sex*. Es ist erkennbar, dass Modell 2a die Daten genauso gut erklärt wie Modell 1. Zur Kontrolle wird ebenfalls der Effekt des Treatments entfernt, was zu selbigem Ergebnis führt.

Entfernt man den geschlechtsspezifischen Effekt und den Effekt des Treatments (Modell 3a), ist zu beobachten, dass die Daten mit 13 Itemparametern und 1 Effektparameter *trend* (Parameterschätzung = .39) hinreichend beschrieben werden können.

Durch den Effektparameter *trend* wird eine globale Veränderung über beide Testgruppen erkennbar, die durch einen Übungseffekt der Testwiederholung zustande kommt. Biologische Reifungsprozesse werden aufgrund der Kürze des Untersuchungszeitraums ausgeschlossen. Die logarithmierte Likelihood der Daten beträgt: $\ln L = -815.40$. Damit ist nur der *trend* durch die Testwiederholung eine wirksame Einflussgröße bezüglich der Veränderung. Der insignifikante Modelltest bestätigt diese

Hypothese. Die Effekte des Treatments und geschlechtsspezifische Effekte waren nicht inhaltlich bedeutsam.

Zur Kontrolle wird ein Modelltest des Modells 3a mit dem quasi-gesättigte Modell (qRM) durchgeführt, der ebenfalls zu einem insignifikanten Ergebnis führt ($p = .423$). Das Modell 3a erklärt die Daten demnach genauso gut wie das qRM. Desweiteren werden die Daten anhand der Modelle 3b und 3c geprüft, die jedoch aufgrund signifikanter Modelltests verworfen werden müssen.

5.4 Analysen mittels Veränderungsmessung nach Gittler und Fischer (2011)

Die Ergebnisse der Veränderungsmessung anhand der Veränderungsmessung nach Gittler und Fischer (2011) ist in Tabelle 26 wiedergegeben.

Tabelle 26: Leistungsveränderung der Versuchs- und Kontrollgruppe

	Gruppe	Fisher-Test			Veränderungsmessung nach Gittler & Fischer		
		df	χ^2	p	η	KI	p
einseitig	VG	48	72.12	.014*	.475	(.189,-----)	.006**
	KG	86	160.16	< .01**	.348	(.125,-----)	.017*
zweiseitig	VG	48	55.64	.209	.475	(.097, .821)	.013*
	KG	86	77.89	.722	.348	(.061, .615)	.017*

Anmerkung. Versuchsgruppe (VG), Kontrollgruppe (KG), * signifikant bei $\alpha = 5\%$; ** signifikant bei $\alpha = 1\%$, Effektgröße (η).

Die Ergebnisse zeigen, dass sowohl in der einseitigen wie auch der zweiseitigen Testung signifikante Effekte der Veränderungsmessung nach Gittler und Fischer (2011) gefunden werden konnten. Die einseitige Hypothesenprüfung fällt genauer aus als bei der zweiseitigen Hypothesenprüfung. Im Gegensatz zur zweiten Hypothesenprüfung kann jedoch nur die Untergrenze des Konfidenzintervalls angegeben werden. Da von einer Verbesserung zum zweiten Testzeitpunkt ausgegangen werden kann, ist eine einseitige Hypothesenprüfung vorzuziehen.

Der Fisher-Test basiert auf der χ^2 -Verteilung und dient dem Vergleich der erwarteten Häufigkeiten mit den tatsächlich beobachteten (Fisher, 1995). Ebenso können anhand seiner, beobachtete Häufigkeiten auf ihre Signifikanz geprüft werden.

Der Fisher-Test ist jedoch für eine zweiseitige Testung nicht zu empfehlen, da eine Alternativhypothese in der zweiseitigen Testung nicht klar ersichtlich und eine Signifikanz damit nicht interpretierbar wäre (Gittler & Fischer, 2011). Diese Problematik

stellt sich auch in den vorliegenden Daten. In der einseitigen Testung fällt der Fisher-Test für die Versuchs-, wie für die Kontrollgruppe signifikant aus. In der zweiseitigen Testung zeigt dieser jedoch keine Signifikanz an.

Die Veränderungsmessung nach Gittler und Fischer (2011) prüft hingegen, ob sich die Personengruppen gleichartig verändern. Für die Versuchsgruppe ($\eta = .475$) und die Kontrollgruppe ($\eta = .348$) kann dabei eine Verbesserung der Testleistung vom ersten zum zweiten Testzeitpunkt beobachtet werden. Bei beiden Testgruppen kann daher von einem Lerneffekt durch die Testwiederholung ausgegangen werden.

Um den Effekt des Treatments allgemeiner zu interpretieren, sollen die Effektparameter der Versuchs- und Kontrollgruppe mittels Modellgleichung des Rasch-Modells in Odds Ratios umgewandelt werden. Anhand dieser kann das Quotenverhältnis beziehungsweise ein Chancenverhältnis errechnet werden, welches unabhängig von den gewählten Startwerten für die Personen- und Itemparameter immer dasselbe Verhältnis ergibt (Tabelle 27).

Tabelle 27: Interpretation der einseitigen Effektparameter anhand der odds ratios

	Fähigkeit ϑ	Itemschwierigkeit β	mittlerer Effekt η	p(+)	odds (p/(1-p))	odds ratio (KG/VG)	odds ratio (VG/KG)
KG	.00	.00	.348	.59	1.42	.88	
VG	.00	.00	.475	.62	1.61		1.14

Anmerkung. theta (ϑ), beta (β), eta (η), Wahrscheinlichkeit ein Items zu lösen (p(+))

Die Tabelle zeigt, dass die relative Chance ein Item und der Versuchsgruppen-Bedingung zu lösen das 1.14 fache beträgt im Vergleich zur Kontrollgruppen-Bedingung. Im direkten Vergleich der Bedingungen ist keine große Steigerung beobachtbar.

6. Diskussion

Raumvorstellung stellt seit dem Anfang des 20. Jahrhunderts ein zentrales Thema der wissenschaftlichen Forschung dar (Gardner, 1985; Gittler & Fischer, 2011; Linn & Petersen, 1985; Thurstone, 1931; Voyer et al., 1995). Unter Raumvorstellung versteht man ein Konstrukt, das aus mehreren Subfaktoren besteht und nicht anhand einer einzelnen globalen Fähigkeitsdimension beschrieben werden kann (Linn & Petersen, 1985). Oftmals replizierte Ergebnisse zeigen große geschlechtsspezifische Differenzen in den unterschiedlichen Dimensionen der Raumvorstellung. Vor allem werden diese im Faktor *mental rotation* beschrieben (Linn & Petersen, 1985; Voyer et al., 1995), wo ein deutlicher Vorteil der Männer im Vergleich zu Frauen sichtbar wird.

Von zunehmender Bedeutung scheint es, Raumvorstellungsfähigkeiten bereits in jungen Jahren anhand von einschlägigen Trainingsprogrammen zu fördern, um damit auch Geschlechtsunterschiede im weiteren Verlauf des Lebens möglichst gering zu halten (Neubauer et al., 2010; Tzuriel & Egozi, 2010).

In der vorliegenden Studie soll die Wirksamkeit eines Trainings zur Verbesserung von Raumvorstellung anhand von Konstruktionsbaukästen geprüft werden. Das Trainingsprogramm wurde hierbei von einem Schullehrer des Faches "Technisches Werken" vorgeschlagen, entwickelt und durchgeführt. An dem Training nahmen insgesamt 70 Schülerinnen und Schüler eines niederösterreichischen Gymnasiums im Alter zwischen 12 und 13 Jahren ($\text{mean} = 12.89$, $s = 0.36$) teil, welche in eine Versuchsgruppe ($n_1 = 25$) und Kontrollgruppe ($n_2 = 45$) aufgeteilt wurden. Zur Evaluierung wurden drei verschiedene testtheoretische Prüfverfahren verwendet, klassisch-statistische Verfahren, Analysen anhand des Linearen logistischen Testmodells sowie Analysen anhand der Veränderungsmessung nach Gittler und Fischer (2011).

Alle drei Prüfverfahren zeigen eine signifikante Verbesserung der Leistung bei einer zweiten Testvorgabe derselben Testaufgaben sowohl in der Versuchs- als auch in der Kontrollgruppe auf. Da dieser Effekt in beiden Gruppen zu beobachten ist, liegt die Vermutung nahe, dass es sich zum Teil um einen „trivialen Lerneffekt“ (Klauer, 2000, S.153) handelt. Dieser Effekt liegt dann vor, wenn Testaufgaben identisch sind und durch die erste Vorgabe ein Übungseffekt entsteht, der die Leistung bei der zweiten Vorgabe der Aufgaben beeinflusst (Klauer, 2000).

Bei einem Vergleich der beiden Testgruppen hinsichtlich einer Verbesserung der Raumvorstellungsleistung durch das Treatment, können jedoch keine trainings-spezifischen Effekte aufgefunden werden. Lediglich die Veränderungsmessung nach

Gittler und Fischer (2011) kann anhand der Odds Ratios zeigen, dass die relative Chance ein Item in der Versuchsgruppe zu lösen um das 1.14 fache größer ist als in der Kontrollgruppe. Da es sich jedoch hierbei nur um einen Trend und nicht um signifikante Unterschiede handelt, wäre es notwendig weitere Untersuchungen anhand größerer Stichproben durchzuführen.

Auch bezüglich geschlechtsspezifischer Effekte können im Großteil der statistischen Vergleiche keine signifikanten Unterschiede beobachtet werden. Dieses Ergebnis steht im Kontrast zu vielen bisherigen Studien (Geiser, Lehmann, Corth, & Eid, 2008; Voyer et al., 1995), die einen deutlichen Vorteil der Burschen gegenüber den Mädchen in Raumvorstellungstests feststellen konnten. Eine mögliche Begründung hierfür liegt darin, dass zumeist Geschlechtsunterschiede bezüglich der Raumvorstellung erst ab einem Alter von circa 13 Jahren - also nach Abschluss der Pubertät - aufzufinden sind (Linn & Petersen, 1985; Voyer et al., 1995). Möglicherweise sind die Testpersonen, insbesondere die Burschen, in der vorliegenden Stichprobe zu jung, um geschlechtsspezifische Unterschiede aufzuzeigen. Gerade bei Burschen dürften geschlechtsspezifische Reifungsprozesse notwendig sein, um eine hohe Raumvorstellungsfähigkeit zu entwickeln.

Hinsichtlich der Effekte der Bearbeitungszeit kann beobachtet werden, dass sowohl die Versuchs- als auch die Kontrollgruppe die Testaufgaben zum zweiten Testzeitpunkt deutlich schneller bearbeiteten. Dies liegt wahrscheinlich an dem zuvor schon beschriebenen Übungseffekt durch die wiederholte Vorgabe derselben Items. Schlussendlich kann auch eine indirekt proportionale Korrelation der Mathematiknote, sowie ein tendenzieller, indirekt proportionaler Zusammenhang der Deutschnote mit der Leistung im Raumvorstellungstest beobachtet werden. Für die Physiknote sowie für die Note im Fach Technischen Werken kann keine Korrelation beobachtet werden. Das Fehlen eines Zusammenhangs mit der Physiknote, mag mit einer undifferenzierten Notenvergabe zusammenhängen, da in anderen Studien eine negative Korrelation zwischen 3DW-Leistungen und Physiknote durchaus aufgefunden werden konnte (Gittler, 1990).

Es konnte gezeigt werden, dass alle drei statistischen Methoden (klassisch-statistische Auswertung anhand der Fähigkeitsparameter, Auswertung anhand des LLTM und Auswertung anhand der Veränderungsmessung nach Gittler und Fischer) essentiell zu denselben Ergebnissen führen. Während die Auswertung anhand der Fähigkeitsparameter jedoch nur Informationen über das Vorhandensein eines potentiellen Effekts liefert, erhält man in der Auswertung anhand des LLTM und der Veränderungs-

messung nach Gittler und Fischer (2011) detailliertere Ergebnisse über die Leistung der Stichprobe und mögliche Effekte.

Anhand des LLTM kann nach dem hierarchischen Aufbau der Modelle der Einfluss der Effektparameter ermittelt und diese bei fehlender Signifikanz sukzessive ausgeschlossen werden. Die Veränderungsmessung nach Gittler und Fischer (2011) lässt eine noch detaillierte Interpretation der Daten zu. Auch hier ist ein signifikanter Effekt durch die wiederholte Vorgabe in der Versuchs- und Kontrollgruppe erkennbar. Ein Effekt des Treatments in der Versuchsgruppe ist zwar zu vermuten, da die Leistungsveränderung vom ersten zum zweiten Testzeitpunkt auf dem 1% Niveau signifikant war, ist aber möglicherweise nicht stark genug, um sich signifikant von der Kontrollgruppe zu unterscheiden.

Eine Darstellung der Effekte anhand der Odds Ratios lässt erkennen, dass, obwohl keine signifikanten Effekte innerhalb des LLTM erkennbar waren, dennoch die Chance ein Item in der Versuchsgruppe zu lösen um das 1.14 fache größer war als in der Kontrollgruppe.

Aus inhaltlicher Sicht stellt sich die Frage, warum sich die Versuchs- und Kontrollgruppe dennoch nicht in ihren Leistungen signifikant voneinander unterscheiden. Dafür kommen mehrere mögliche Gründe in Frage: So liegt dies möglicherweise an einer ineffizienten Treatmentbedingung, sei es aufgrund der zu geringen Dauer oder Anzahl der Trainingseinheiten für die Versuchsgruppe. Ähnliche Studien zu Verbesserung des räumlichen Vorstellungsvermögens anhand eines Trainings mit Darstellender Geometrie (Gittler, 1992) oder Geometrisch Zeichnen (Gittler et al., 2006) verwendeten Längsschnittstudien über zwei Jahre um einen Effekt sichtbar zu machen. Auch eine Intensivierung der einzelnen Trainingseinheiten auf mehrere Stunden pro Woche kann eine Verbesserung der Raumvorstellungsleistung zu Folge haben (Wright et al., 2008). Ein intensives, langfristiges Treatment über mehrere Monate könnte somit möglicherweise die Raumvorstellungsleistung der Schüler effektiver beeinflussen und zu signifikanten Unterschieden im Vergleich zur Kontrollgruppe führen. Gerade bei Trainingsmaterial das nicht dem Testmaterial entspricht ist ein nichttrivialer Lerntransfer zu erwarten (Gittler, 1994) und erwünscht. Dieser lässt sich jedoch möglicherweise nur mit einem länger andauernden Training erzielen.

Ein weiterer möglicher Grund liegt in der Motivation der Schülerinnen und Schüler an dem Treatment teilzunehmen. Da das Treatment im Rahmen des Unterrichts zum Technischen Werken stattfand, mag die Konzentration der Schülerinnen und Schüler weniger hoch gewesen sein als in anderen Fächern des Schulunterrichts.

Auch ist der Inhalt der Trainingsmaßnahme kritisch zu beleuchten. Möglicherweise waren die einzelnen Aufgaben der Trainingseinheiten (siehe Anhang 2) zu wenig auf die Verbesserung von Raumvorstellungsfähigkeiten abgestimmt und bedürfen eingehender Überarbeitung.

Schlussendlich ist anzumerken, dass die Gesamtstichprobe generell sehr klein war. Obwohl ein Vorteil der Veränderungsmessung nach Gittler und Fischer (2011) darin besteht, Effekte auch in geringen Stichprobengrößen sichtbar zu machen, könnte die Effektgröße dennoch zu klein gewesen sein, um signifikante Unterschiede aufzuzeigen. Bortz und Döring (2006) schlagen für klassisch-statistische Verfahren einen optimalen Stichprobenumfang ab 310 Personen vor, um einen geringen Effekt über die Mittelwerte zweier unabhängiger Stichproben auf dem 5%igen Signifikanzniveau aufzufinden.

Als Ausblick für ähnliche Testungen wäre anzuraten, für den Posttest andere Items aus dem Itempool des 3DW auszuwählen als im Pretest Verwendung fanden. So können mögliche Übungseffekte besser ausgeschlossen und eventuelle Treatmenteffekte durch die Analysen anhand des LLTM und der Veränderungsmessung nach Gittler und Fischer (2011) sichtbar gemacht werden.

Literaturverzeichnis

- Alexander, G. M., & Evardone, M. (2008). Blocks and bodies: Sex differences in a novel version of the mental rotation test. *Hormones and Behavior*, 53, 177-184.
- Allen, M. J. (1974). Sex differences in spatial problem-solving styles. *Perceptual and Motor Skills*, 39, 843-846.
- Amthauer, R. (1955). *Intelligenz-Struktur-Test: I-S-T* (2., erw. u. verb. Aufl.). Göttingen: Hogrefe.
- Amthauer, R. (1973). *Intelligenz-Struktur-Test: I-S-T 70* (4., unveränd. Aufl.). Göttingen: Hogrefe.
- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123-140.
- Arendasy, M., & Gittler, G. (2003). Endlosschleifen: Psychometrische Grundlagen des Aufgabentyps E^P. *Diagnostica*, 49, 164-175.
- Arendasy, M., Sommer, M., & Gittler, G. (2010). Combining automatic item generation and experimental designs to investigate the contribution of cognitive components to the gender difference in mental rotation. *Intelligence*, 38, 506-512.
- Baenninger, M., & Newcombe, N. (1989). The role of experience in spatial test performance: A meta-analysis. *Sex Roles*, 20, 327-344.
- Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4., überarb. Aufl.). Heidelberg: Springer.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytical studies*. Cambridge, MA: Cambridge Univ. Press.

- Cooke-Simpson, A., & Voyer, D. (2007). Confidence and gender differences on the Mental Rotations Test. *Learning and Individual Differences*, 17, 181-186.
- Driscoll, I., Hamilton, D. A., Yeo, R. A., Brooks, W. M., & Sutherland, R. J. (2005). Virtual navigation in humans: the impact of age, sex, and hormones on place learning. *Hormones and Behavior*, 47, 326-335.
- Eliot, J., & Smith, I. M. (1983). *An international directory of spatial tests*. Windsor: Nfer-Nelson.
- Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, 18, 850-855.
- Fisher, R. A. (1995). *Statistical methods, experimental design, and scientific inference*. Oxford: Oxford Univ. Press.
- Fischer, G. H. (1995). The Linear Logistic Test Model. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch Models: Foundations, Recent Developments and Applications* (S. 131-155). New York: Springer.
- Gardner, H. (1985). *Frames of mind: the theory of multiple intelligences* (2., ed.). New York: Basic Books.
- Geiser, C., Lehmann, W., & Eid, M. (2008). A note on sex differences in mental rotation in different age groups. *Intelligence*, 36, 556-563.
- Geiser, C., Lehmann, W., Corth, M., & Eid, M. (2008). Quantitative and qualitative change in children's mental rotation performance. *Learning and Individual Differences*, 18, 419-429.
- Gittler, G. (1986). Inhaltliche Aspekte bei der Itemselektion nach dem Modell von Rasch. *Zeitschrift für Experimentelle und angewandte Psychologie*. 33, 386-412.

- Gittler, G. (1990). *Dreidimensionaler Würfeltest. Ein Rasch-skaliertes Test zur Messung des räumlichen Vorstellungsvermögens*. Weinheim: Beltz.
- Gittler, G. (1992). *Testpsychologische Aspekte der Raumforschung – Kritik, Lösungsansätze und empirische Befunde (Monographie)*. Unveröff. Habilitationsschrift, Universität Wien.
- Gittler, G. (1994). Intelligenzförderung durch Schulunterricht: Darstellende Geometrie und räumliches Vorstellungsvermögen. In G. Gittler, M. Jirasko, U. Kastner-Koller, C. Korunka, & Al-Roubaie (Hrsg.), *Die Seele ist ein weites Land* (S. 105-122). Wien: WUV-Univ.-Verl.
- Gittler, G. (1999). Sind Raumvorstellung und Reasoning separierbare Fähigkeitsdimensionen? Dimensionalitätsanalysen zweier Rasch-skalierten Tests: 3DW und WMT. *Diagnostica*, 45, 69-81.
- Gittler, G., Arendasy, M., & Hagen, F. (2006). *Fördert der Unterricht in „Geometrisch Zeichnen“ das räumliche Vorstellungsvermögen? – Eine IRT-basierte Längsschnittstudie*. Unveröff. Manuskript.
- Gittler, G. (2008). *Veränderungsmessung mit dem Linearen logistischen Testmodell (LLTM)*. Unveröff. Manuskript.
- Gittler, G., & Fischer, G. (2011). IRT-Based Measurement of Short-Term Changes of Ability, With an Application to Assessing the "Mozart Effect". *Journal of Educational and Behavioral Statistics*, 36, 33-75.
- Gittler, G., & Glück, J. (1998). Differential transfer of learning: Effects of instruction in Descriptive Geometry on spatial test performance. *Journal for Geometry and Graphics*, 2, 71-84.
- Glück, J., & Fabrizii, C. (2010). Gender differences in the Mental Rotations Test are partly explained by response format. *Journal of Individual Differences*, 31, 106-109.

- Glück, J., Machat, R., Jirasko, M., & Rollett, B. (2002). Training-related changes in solution strategy in a spatial test: An application of item response models. *Learning and Individual Differences*, 13, 1-12.
- Grüßing, M. (2002). Wieviel Raumvorstellung braucht man für Raumvorstellungsaufgaben? Strategien von Grundschulkindern bei der Bewältigung räumlich-geometrischer Anforderungen. *Zentralblatt für Didaktik der Mathematik*, 34, 37-45.
- Guilford, J. P., & Hoepfner, R. (1971). *The analysis of intelligence*. New York: McGraw-Hill.
- Heil, M., & Jansen-Osmann, P. (2008). 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, 683-689.
- Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32, 175-191.
- Hines, M., Fane, B. A., Pasterski, V. L., Mathews, G. A., Conway, G.S., & Brook, C. (2003). Spatial ability following prenatal androgen abnormality: targeting and mental rotations performance in individuals with congenital adrenal hyperplasia. *Psychoneuroendocrinology*, 28, 1010-1026.
- Hirnstein, M., Bayer, U., & Hausmann, M. (2009). Sex-specific response strategies in mental rotation. *Learning and Individual Differences*, 19, 225-228.
- Hooven, C. K., Chabris, C. F., Ellison, P. T., & Kosslyn, S. M. (2004). The relationship of male testosterone to components of mental rotation. *Neuropsychologia*, 42, 782-790.
- Hugdahl, K., Thomsen, T., & Ersland, L. (2006). Sex differences in visuo-spatial processing: An fMRI study of mental rotation. *Neuropsychologia*, 44, 1575-1583.

- Kaufman, S. B. (2007). Sex differences in mental rotation and spatial visualization ability: Can they be accounted for by differences in working memory capacity? *Intelligence*, 35, 211-223.
- Klauer, J. K. (2000). Das Huckepack-Theorem asymmetrischen Strategietransfers. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 32, 153-165.
- Koscik, T., O'Leary, D., Moser, D. J., Andreasen, N. C., & Nopoulos, P. (2009). Sex differences in parietal lobe morphology: Relationship to mental rotation performance. *Brain and Cognition*, 69, 451-459.
- Kubinger, K. D. (2006). *Psychologische Diagnostik. Theorie und Praxis psychologischen Diagnostizierens*. Göttingen: Hogrefe.
- Levine, S. C., Vasilyeva, M., Lourenco, S. F., Newcombe, N. S., & Huttenlocher, J. (2005). Socioeconomic status modifies the sex difference in spatial skill. *Psychological Science*, 16, 841-845.
- Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child development*, 56, 1479-1498.
- Maier, P. H. (1999). *Räumliches Vorstellungsvermögen. Ein theoretischer Abriß des Phänomens räumliches Vorstellungsvermögen. Mit didaktischen Hinweisen für den Unterricht*. Donauwörth: Auer.
- McGee, M. G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin*, 86, 889-918.
- Moè, A. (2009). Are males always better than females in mental rotation? Exploring a gender belief explanation. *Learning and Individual Differences*, 19, 21-27.

- Moè, A., & Pazzaglia, F. (2006). Following the instructions! Effects of gender beliefs in mental rotation. *Learning and Individual Differences*, 16, 369-377.
- Moè, A., Meneghetti, C., & Cadinu, M. (2009). Women and mental rotation: Incremental theory and spatial strategy use enhance performance. *Personality and Individual Differences*, 46, 187-191.
- Molenaar, I. W. (1995). Some Background for Items Response Theory and the Rasch Model. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 3-14). New York: Springer.
- Moore, D. S., & Johnson, S. P. (2008). Mental rotation in human infants: A sex difference. *Psychological Science*, 19, 1063-1066.
- Neubauer, A. C., Bergner, S., & Schatz, M. (2010). Two- vs. three-dimensional presentation of mental rotation tasks: Sex differences and effects of training on performance and brain activation. *Intelligence*, 38, 529-539.
- Oerter, R., & Montada, L. (2002). *Entwicklungspsychologie* (5., vollst. überarb. Aufl.). Weinheim: Beltz.
- Parsons, T. D., Larson, P., Kratz, K., Thiebaut, M., Bluestein, B., Buckwalter, J. G., & Rizzo, A. A. (2004). Sex differences in mental rotation and spatial rotation in a virtual environment. *Neuropsychologia*, 42, 555-562.
- Peters, M. (2005). Sex differences and the factor of time in solving Vandenberg and Kuse mental rotation problems. *Brain and Cognition*, 57, 176-184.
- Putz, D. A., Gaulin, S. J. C., Sporter, R. J., & McBurney, D. H. (2004). *Evolution and Human Behavior*, 25, 182-199.
- Quaiser-Pohl, C. (1998). *Die Fähigkeit zur räumlichen Vorstellung. Zur Bedeutung von kognitiven und motivationalen Faktoren für geschlechtsspezifische Unterschiede*. Münster: Waxmann.

- Quaiser-Pohl, C., Geiser, C., & Lehmann, W. (2006). The relationship between computer-game preference, gender, and mental-rotation ability. *Personality and Individual Differences*, 40, 609-619.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. København: Danmarks Paedagogiske Institut.
- Rilea, S. L. (2008). A lateralization of function approach to sex differences in spatial ability: A reexamination. *Brain and Cognition*, 67, 168-182.
- Sanders, G., Bereczkei, T., Csatho, A., & Manning, J. (2005). The ratio of the 2nd to 4th finger length predicts spatial ability in men but not women. *Cortex*, 41, 789-795.
- Spearman, C. (1904). "General intelligence" objectively determined and measured. *The American Journal of Psychology*, 15, 201-292.
- Stafford, R. E. (1961). Sex differences in spatial visualization as evidence of sex-linked inheritance. *Perceptual and Motor Skills*, 13, 428.
- Stemmler, G., Hagemann, D., Amelang, M., & Bartussek, D. (2011). *Differentielle Psychologie und Persönlichkeitsforschung* (7., vollst. überarb. Aufl.). Stuttgart: Kohlhammer.
- Stumpf, H., & Fay, E. (1983). *Schlauchfiguren. Ein Test zur Beurteilung des räumlichen Vorstellungsvermögens*. Göttingen: Hogrefe.
- Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, 38, 406-427.
- Titze, C., Jansen, P., & Heil, M. (2010). Mental rotation performance and the effect of gender in fourth graders and adults. *European Journal of Developmental Psychology*, 7, 432-444.
- Tzuriel, D., & Egozi, G. (2010). Gender differences in spatial ability of young children: The effects of training and processing strategies. *Child Development*, 81, 1417-1430.

- Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations: A group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, 47, 599-604.
- Van der Ham, I. J. M., & Borst, G. (2011). Individual differences in spatial relation processing: Effects of strategy, ability, and gender. *Brain and Cognition*, 76, 184-190.
- Vederhus, L., & Krekling, S. (1996). Sex differences in visual spatial ability in 9-year-old children. *Intelligence*, 23, 33-43.
- Vogel, J. J., Bowers, C. A., & Vogel, D. S. (2003). Cerebral lateralization of spatial abilities: A meta-analysis. *Brain and Cognition*, 52, 197-204.
- Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: a meta-analysis. *Psychonomic Bulletin & Review*, 18, 167-277.
- 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.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450.
- Wraga, M., Helt, M., Jacobs, E., & Sullivan, K. (2006). Neural basis of stereotype-induced shifts in women's mental rotation performance. *Social Cognitive and Affective Neuroscience*, 2, 12-19.
- Wright, R., Thompson, W. L., Ganis, G., Newcombe, N. S., & Kosslyn, S. M. (2008). Training generalized spatial skills. *Psychonomic Bulletin & Review*, 15, 763-771.

Sekundärliteraturverzeichnis

- Barratt, E. S. (1953). An analysis of verbal reports of solving spatial problems as an aid in defining spatial factors. *Journal of Psychology*, 36, 17-25.
- El-Koussy, A. A. H. (1935). The visual perception of space. *British Journal of Psychology. Monograph Supplement*, 20, 1-80.
- Lohman, D. F. (1979). *Spatial Ability; A Review and Re-analysis of Correlational Literature*. Aptitude Research Project, Report Nr.8, Stanford University.
- Thurstone, L. L. (1938). *Primary Mental Abilities*. Chicago: University of Chicago Press.
- Thurstone, L. L. (1950). *Some primary abilities in visual thinking* (Research Report 59). Chicago: Psychometric Laboratory.
- Thurstone, L.L., & Thurstone, T.G. (1941). *Factorial studies of intelligence*. Chicago: University of Chicago Press.
- Zimmerman, W. S. (1954). Hypotheses concerning the nature of the spatial factors. *Educational and Psychological Measurement*, 14, 196-400.

Tabellenverzeichnis

<i>Tabelle 1:</i> Verteilung des Geschlechts in der Versuchs- und Kontrollgruppe	43
<i>Tabelle 2:</i> Verteilung der Schulnoten in der Gesamtpopulation	43
<i>Tabelle 3:</i> Schwierigkeits- und Leichtigkeitsparameter der Items	44
<i>Tabelle 4:</i> Conditional Maximum Likelihood Schätzer, sowie Weighted Likelihood Schätzer der Fähigkeitsparameter	45
<i>Tabelle 5:</i> Häufigkeiten der Rohscores zum ersten und zweiten Testzeitpunkt	46
<i>Tabelle 6:</i> Mittelwertunterschiede der Fähigkeitsparameter der Versuchsgruppe zwischen t1 und t2 mittels Wilcoxon Test	48
<i>Tabelle 7:</i> Mittelwertunterschiede der Fähigkeitsparameter der Kontrollgruppe zwischen t1 und t2 mittels Wilcoxon Test	48
<i>Tabelle 8:</i> Mittelwertunterschiede der Differenz der Fähigkeitsparameter zwischen t1 und t2 bezüglich der VG und KG mittels U-Test nach Mann-Whitney	49
<i>Tabelle 9:</i> Geschlechtsspezifische Mittelwertunterschiede bezüglich der Differenz der Fähigkeitsparameter zwischen t1 und t2 innerhalb der Versuchsgruppe mittels t-Test für unabhängige Stichproben	49
<i>Tabelle 10:</i> Geschlechtsspezifische Mittelwertunterschiede bezüglich der Differenz der Fähigkeitsparameter zwischen t1 und t2 innerhalb der Kontrollgruppe mittels t-Test für unabhängige Stichproben	50
<i>Tabelle 11:</i> Mittelwertunterschiede in den Fähigkeitsparametern der Mädchen der Versuchs- und Kontrollgruppe zu t1 mittels t-Test für unabhängige Stichproben	50
<i>Tabelle 12:</i> Mittelwertunterschiede in den Fähigkeitsparametern der Mädchen der Versuchs- und Kontrollgruppe zu t2 mittels t-Test für unabhängige Stichproben	51
<i>Tabelle 13:</i> Mittelwertunterschiede in den Fähigkeitsparametern der Burschen der Versuchs- und Kontrollgruppe zu t1 mittels U-Test nach Mann-Whitney	51
<i>Tabelle 14:</i> Mittelwertunterschiede in den Fähigkeitsparametern der Burschen der Versuchs- und Kontrollgruppe zu t2 mittels U-Test nach Mann-Whitney	51
<i>Tabelle 15:</i> Übersicht über die Mediane der Bearbeitungszeit	52
<i>Tabelle 16:</i> Mittelwertunterschiede in der Bearbeitungszeit der Versuchsgruppe zwischen den Testzeitpunkten mittels Wilcoxon Test	52
<i>Tabelle 17:</i> Mittelwertunterschiede in der Bearbeitungszeit der Kontrollgruppe zwischen den Testzeitpunkten mittels Wilcoxon Test	53

<i>Tabelle 18:</i> Mittelwertunterschiede in der Bearbeitungszeit der Kontroll- und Versuchsgruppe zu t1 mittels U-Test nach Mann-Whitney	53
<i>Tabelle 19:</i> Mittelwertunterschiede in der Bearbeitungszeit der Kontroll- und Versuchsgruppe zu t2 mittels U-Test nach Mann-Whitney	54
<i>Tabelle 20:</i> Korrelation der Schulnoten mit den Fähigkeitsparametern zu t1	54
<i>Tabelle 21:</i> Korrelation der Schulnoten mit den Fähigkeitsparametern zu t2	54
<i>Tabelle 22:</i> unvollständiges quasi-gesättigtes Rasch-Modell (qRM)	55
<i>Tabelle 23:</i> vollständiges gesättigtes Rasch-Modell (RM)	56
<i>Tabelle 24:</i> LLTM Gewichtszahlenmatrix ($df = 16$) für Modell 1	57
<i>Tabelle 25:</i> Ergebnisse der Hypothesenprüfung	59
<i>Tabelle 26:</i> Leistungsveränderung der Versuchs- und Kontrollgruppe	60
<i>Tabelle 27:</i> Interpretation der einseitigen Effektparameter anhand der odds ratios	61

Abbildungsverzeichnis

<i>Abbildung 1:</i> Intelligenzstrukturmodell (Guilford & Hoepfner, 1971, S. 19).	10
<i>Abbildung 2:</i> quasi-experimenteller 2 Gruppen-Versuchsplan mit Messwiederholung.	37
<i>Abbildung 3:</i> Übungsbeispiel II des 3DW.	38
<i>Abbildung 4:</i> 3DW-Fähigkeitsparameter der Versuchs- und Kontrollgruppe.	46
<i>Abbildung 5:</i> 3DW-Fähigkeitsparameter männlicher und weiblicher SchülerInnen zu beiden Testzeitpunkten.	47

Anhang

Anhang 1: Antwortblatt des 3DW

Antwortblatt 3DW

Name: _____ Testform: _____ Datum: _____

Geburtstag: _____ Geschlecht: ☐ weiblich ☐ männlich (bitte ankreuzen)
TT / MM / JJJJ

Note im letzten Zeugnis in: Deutsch _____ Mathematik _____ Physik _____ Technisches Werken _____

Beispiel I	A	X	C	D	E	F	G	H
Beispiel II	A	B	C	X	E	F	G	H
Beispiel III	A	B	C	D	E	F	G	H

← im Aufgabenheft enthalten

-- Bitte auf weitere Anweisungen warten.--

Testanfang: _____
Stunde Minute

Aufgabe 1	A	B	C	D	E	F	G	H
Aufgabe 2	A	B	C	D	E	F	G	H
Aufgabe 3	A	B	C	D	E	F	G	H
Aufgabe 4	A	B	C	D	E	F	G	H
Aufgabe 5	A	B	C	D	E	F	G	H
Aufgabe 6	A	B	C	D	E	F	G	H
Aufgabe 7	A	B	C	D	E	F	G	H
Aufgabe 8	A	B	C	D	E	F	G	H
Aufgabe 9	A	B	C	D	E	F	G	H
Aufgabe 10	A	B	C	D	E	F	G	H
Aufgabe 11	A	B	C	D	E	F	G	H
Aufgabe 12	A	B	C	D	E	F	G	H
Aufgabe 13	A	B	C	D	E	F	G	H
Aufgabe 14	A	B	C	D	E	F	G	H
Aufgabe 15	A	B	C	D	E	F	G	H

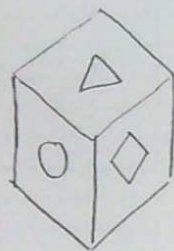
-- Bitte das Testende sofort eintragen! Keine Nachbearbeitung einzelner Aufgaben vornehmen.--

Testende: _____
Stunde Minute

-- Aufgabenheft und Antwortblatt sofort abgeben. Danke für die Mitarbeit! --

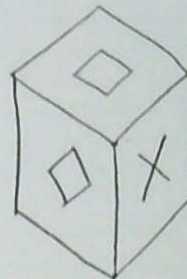
Anhang 2: Beispielaufgaben des Treatments

1.1. Versuche ausgehend von der Abb. 1.3. (Würfelnetz) das gefragte Symbol herauszufinden.



UNTERSEITE

A ☐
B ☐
C ☒
ZUTREFFENDES
EINRINGELN



UNTERSEITE

A ☐
B ☐
C ☒
ZUTREFFENDES
EINRINGELN

1.2. Kann der jeweils dargestellte Würfel aus dem Würfelnetz der Abb. 1.3. gebaut werden?
Zutreffendes einringeln!

MÖGLICH ODER UNMÖGLICH?

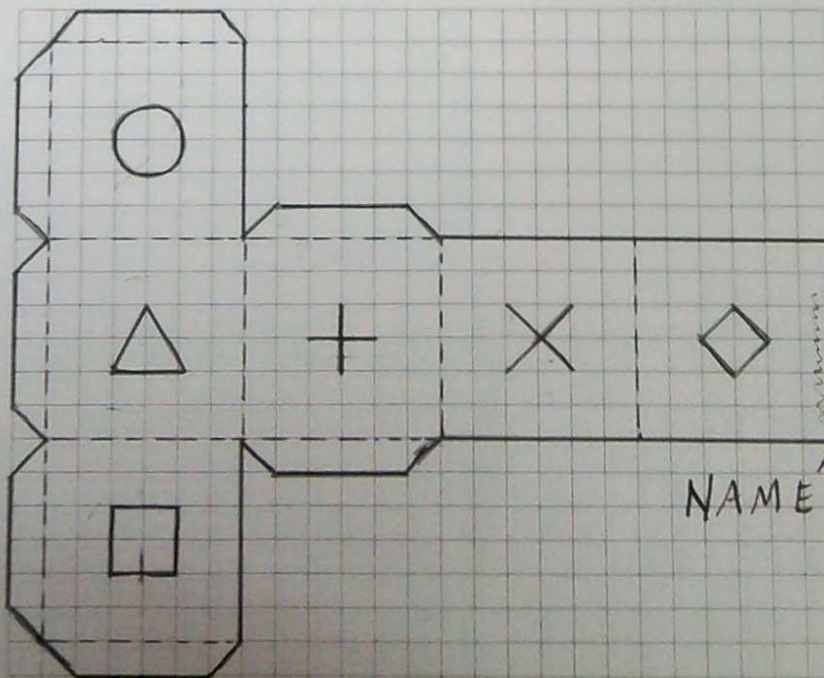


JA
NEIN ☒



JA
NEIN ☒

1.3. Würfelnetz zum Nachbauen. Übertrage die Vorlage laut Anleitung auf den Karton und baue ihn zusammen. Zuvor wie angezeichnet beschriften!



2.1. Versuche in einer Teamarbeit (2 Personen bilden ein Team) die Abbildungen mit Matadorbausteinen in der richtigen Anordnung und der richtigen Bausteinauswahl nachzubauen.

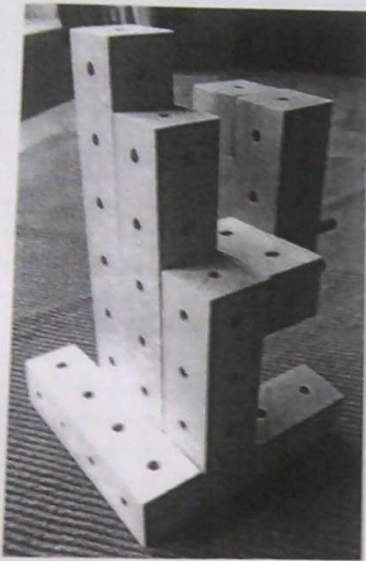


Abb. 1 Bauwerk 1

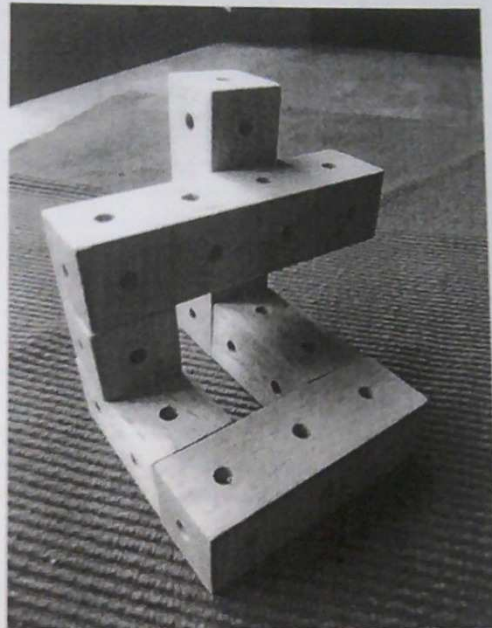


Abb. 2 Bauwerk 2

2.2. Zeichne nun mit Bleistift wie in der Abbildung 4 in dein Arbeitsheft jeweils 3 zusammenhängende Ansichten der Bauwerke 1 und 2

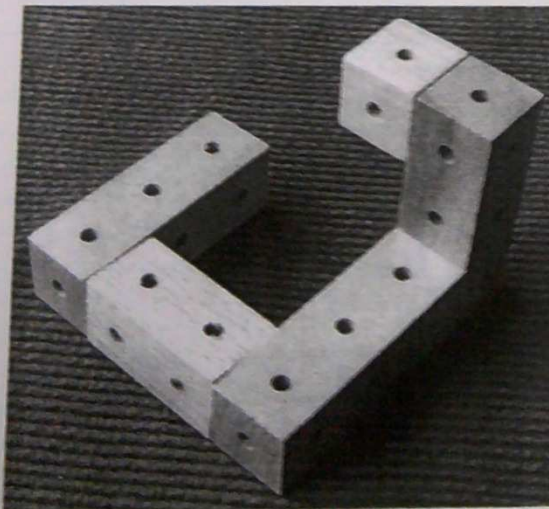


Abb.3 Bauwerk 3

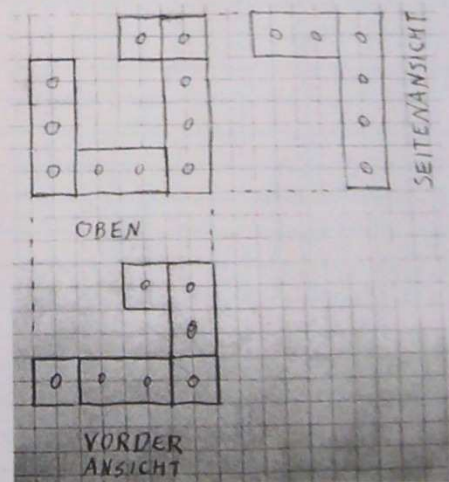
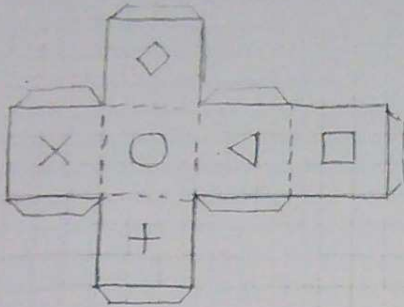


Abb. 4 Ansichten von Bauwerk 3

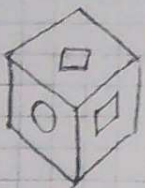
ARBEITSBLATT 3

3.1. WELCHER WÜRFEL PASST ZUM NETZ ?

RICHTIGE ANTWORT
EINRINGELN !



A



JA
NEIN

B



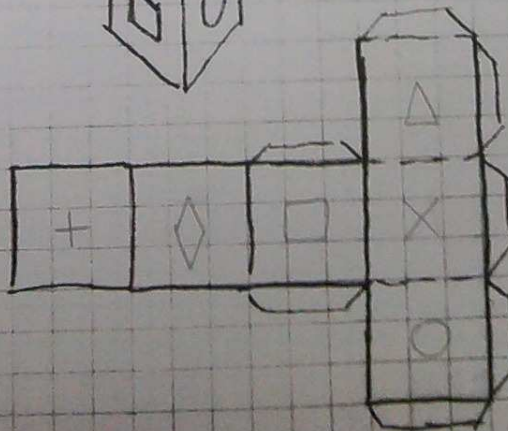
JA
NEIN

C



JA
NEIN

3.2. ZEICHNE DAS WÜRFELNETZ, WELCHES
ZU DEN 2 WÜRFELANSICHTEN PASST,
UND VERWENDE DIE SYMBOLE WIE BEI 3.1.



Abstract auf Deutsch

Die vorliegende Arbeit beschäftigt sich mit der Frage, ob ein Training anhand von Konstruktionsbaukästen einen Einfluss auf die Raumvorstellungsleistung von Schulkindern hat. Das Design entsprach einem quasi-experimentellen 2 Gruppen Versuchsplan mit Messwiederholung. Als psychologisches Erhebungsinstrument wurde der Dreidimensionale Würfeltest (Gittler, 1990) herangezogen. Insgesamt wurden 70 Kinder im Alter von 13 Jahren getestet, welche anhand dreier verschiedener statistischer Methoden ausgewertet wurden. Hierbei handelt es sich um eine klassisch-statistische Auswertung anhand von Fähigkeitsparametern, der Auswertung anhand des Linearen logistischen Testmodells (Fischer, 1995) und der Veränderungsmessung nach Gittler und Fischer (2011). Die Ergebnisse zeigen lediglich signifikante Übungseffekte auf. Es konnten keine signifikanten Effekte des Treatments oder geschlechtsspezifische Effekte aufgefunden werden. Alle drei Auswertungsmethoden stimmen in ihren Ergebnissen überein, was eine Validierung der Methoden ermöglicht.

Schlüsselwörter: Veränderungsmessung, Raumvorstellung, Rasch-Modell, 3DW

Abstract in English

The goal of the present study is to answer the question whether a training containing building blocks has an influence on the spatial ability of school children. For examination a quasi-experimental 2 group design with repetitive measuring was adopted. Altogether 70 children at the age of 13 were tested by means of the Three-Dimensional Cube Test (Gittler, 1990). Evaluation was done using three different statistical methods which are classical-statistical methods, measurement of change by means of the linear logistic test model (Fischer, 1995) and measurement of change according to Gittler and Fischer (2011). Results show significant effects by means of repetition. No significant effects could be shown by means of treatment or sex. All three statistical methods coincide in their results, thus enabling validation.

Keywords: measurement of change, spatial ability, Rasch Model, 3DC

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Wien, Februar 2012

(Ramona Knapp)

Curriculum Vitae

Persönliche Angaben

Ramona Knapp

geboren am 28. Juli 1986 in Marl, Deutschland

Staatsangehörigkeit: Deutschland

Familienstand: ledig

Ausbildung

2006 - 2012	Diplomstudium der Psychologie an der Universität Wien
2000 - 2006	Philipp-Matthäus-Hahn-Gymnasium, Echterdingen, Deutschland
	Abschluss: Abitur
1997 - 2000	Deutsche Botschaftsschule Peking, Gymnasium, China

Berufspraxis

Nov. 2009 bis Jan. 2010	Fachpraktikum im Österreichischen Institut für Sexualpädagogik
Mai bis Okt. 2009	Fachpraktikum in der Gesellschaft gegen Sekten- und Kultgefahren
Sept. bis Dez. 2008	Anstellung in der Psychologisch-Psychosomatischen Gemeinschaftsordination Frau Dr. Pfefferkorn