



universität
wien

Diplomarbeit

Titel der Arbeit

Geschlechtsspezifische Leistungsunterschiede
in Piagets Water-Level Tasks:
Eine Meta-Analyse

Verfasserin

Stefanie Pavlovic

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, im April 2009

Studienkennzahl lt. Studienblatt:

A 298

Studienrichtung lt. Studienblatt:

Diplomstudium Psychologie

Betreuer:

Univ.-Ass. DDDr. MMag. Martin Voracek

DANKSAGUNG

Mein Dank ergeht an erster Stelle an meinen Betreuer Univ.-Ass. DDDr. MMag. Martin Voracek für seine große fachliche Expertise, die er für die Entstehung dieser Arbeit zur Verfügung gestellt hat. Für die Bereitstellung eines großen Teils und die Finanzierung zusätzlicher Anteile des äußerst umfangreichen Literaturkorpus gebührt mein Dank dem Arbeitsbereich Methodenlehre am Institut für Psychologische Grundlagenforschung und im Besonderen Univ.-Prof. Dr. Anton K. Formann MSc. Ihm danke ich weiterhin für die Idee zur Durchführung der vorliegenden Untersuchung. An der Zusammenstellung der Arbeitsgrundlagen für diese Analyse war Herr Matthias Mascherbauer in seiner Funktion als Studienassistent am Institut für Psychologische Grundlagenforschung maßgeblich beteiligt: Für seinen Einsatz bei der Beschaffung eines umfangreichen Teils des genannten Literaturkorpus sei ihm explizit gedankt.

Weiters gilt mein Dank Mag. Klug, Mag. Pietschnig und Herrn Walla BSc für das gewissenhafte Lektorat dieser Arbeit.

Dank gebührt nicht zuletzt meinem Freund Jakob und meinen Eltern Brigitte und Ernst für deren vielfältige Unterstützung während der Erstellung dieser Arbeit.

INHALT

EINLEITUNG	1
THEORETISCHER TEIL	5
1 Water-Level Tasks	6
1.1 Entwicklungshintergrund und Beschreibung der Testaufgaben	6
1.1.1 Stufentheorie des Lösungsverhaltens bei Water-Level Tasks	7
1.2 Geschichte des Forschungsgegenstandes	9
1.3 Theorien zur Entstehung geschlechtsspezifischer Leistungsunterschiede bei Water-Level Tasks	10
1.3.1 Biologische Korrelate	11
1.3.1.1 Körpergröße	11
1.3.1.2 Endokrine Faktoren	12
1.3.1.3 Biogenetische Faktoren	12
1.3.2 Erfahrung und Wissen	12
1.3.2.1 Frühe Erfahrung in räumlichen Aktivitäten	13
1.3.2.2 Allgemeines physikalisches Wissen	13
1.3.2.3 Spezielles physikalisches Wissen über das Invarianzprinzip	14
1.3.3 Räumliche Fähigkeiten	14
1.3.4 Soziokognitive Korrelate	17
1.3.4.1 Erfolgserwartung und Attributionsstil	17
1.3.4.2 Geschlechtsrolle	18
1.3.5 Perzeptive Korrelate	18
1.3.5.1 Kognitiver Stil: Feldabhängigkeit/Feldunabhängigkeit	19
1.3.6 Motorische Fähigkeiten	21
1.3.7 Zusammenfassung	22

1.4 Bisherige themenverwandte Analysen	23
2 Das Verfahren der Meta-Analyse	25
2.1 Verbreitung und Zweck des Verfahrens	25
2.2 Anwendung und Funktionsweise	26
2.2.1 Ablauf einer meta-analytischen Untersuchung	26
2.2.2 Ermittlung eines mittleren Gesamteffekts	27
2.2.2.1 Modell des festen Effekts vs. Modell zufallsvariabler Effekte	28
2.2.3 Ermittlung intervenierender Variablen	29
2.2.4 Die Publikationsverzerrung	30
2.3 Vorteile und Kritikpunkte	31
2.3.1 Vorteile gegenüber narrativen Reviews	31
2.3.2 Kritikpunkte	32
2.3.2.1 Garbage-In-Garbage-Out-Problem	32
2.3.2.2 Äpfel-und-Birnen-Problem	32
2.3.2.3 Abhängige Messungen	33
2.3.2.4 Schubladenproblem	33
2.3.2.5 Unzureichende primäre Ergebnisberichte	33
DURCHFÜHRUNG DER META-ANALYSE	35
3 Fragestellungen	36
4 Methoden	37
4.1 Systematik der Literatursuche	37
4.2 Datenextraktion und Kodierung	38
4.3 Ein- und Ausschlusskriterien für Primärstudien	40
4.4 Instrumente der statistischen Analyse	42
5 Ergebnisse	43

5.1 Beschreibung inkludierter Stichproben	43
5.2 Einzeleffekte und resultierender Gesamteffekt	44
5.2.1 Schätzung der Publikationsverzerrung	51
5.3 Intervenierende Variablen	56
5.3.1 Subgruppenanalysen	56
5.3.2 Effektmoderatorenregression	60
6 Diskussion	63
6.1 Ausmaß und Generalisierbarkeit des Geschlechtseffekts	63
6.2 Veränderung des Geschlechtseffekts mit dem Lebensalter	63
6.3 Veränderung des Geschlechtseffekts über die Zeit	65
6.4 Quellen des Geschlechtseffekts	65
6.5 Fazit und Ausblick	67
7 Zusammenfassung	70
APPENDICES	73
Appendix A: Seit Abschluss der Analyse erschienene relevante Quellen	74
Appendix B: Resultierender Suchstring für Literaturrecherche	75
Appendix C: Kodierschema für Primärstudien	77
Appendix D: Liste inspizierter, nicht inkludierter Studien	79
Appendix E: Beschreibung inkludierter Studien	98
Appendix F: Programmausgaben der Subgruppenanalysen	104
Appendix G: Programmausgaben der Effektmoderatorenregression	117

Literaturverzeichnis	120
Eidesstattliche Erklärung	134
Curriculum vitae	135

EINLEITUNG

Vor sechzig Jahren traten Jean Piaget und Bärbel Inhelder (1948/1956) unwissentlich eine Lawine los, die erst sechzehn Jahre später ins Rollen kommen sollte: Sie konzipierten die sogenannten Wasserstandsaufgaben (in der Fachliteratur auch bekannt als Water-Level Tasks und in der vorliegenden Arbeit in weiterer Folge als solche bezeichnet), um die Entwicklung räumlicher Konzepte bei Kindern zu testen. Beim Testmaterial dieses Instruments handelt es sich um mehrere Abbildungen von Gefäßen, die in unterschiedlichen Neigungswinkeln relativ zu einer horizontalen Referenzebene dargestellt sind. Die Testanweisung verlangt, in jedem Gefäß den Verlauf jenes Wasserspiegels einzuzeichnen, der sich ergibt, wenn das Gefäß bis zur Hälfte mit Flüssigkeit gefüllt wird. Piaget und Inhelders (1948/1956) Theorie zufolge müssten bereits Kinder ab einem Alter von neun Jahren über das Wissen verfügen, dass der Wasserspiegel unabhängig von der Neigung (und der Form) des Gefäßes in jedem Fall horizontal verläuft. Sechzehn Jahre später allerdings konnte Rebelsky (1964) erstmals zeigen, dass selbst Erwachsene erstaunliche Schwierigkeiten haben, diese simpel erscheinende Aufgabe zu lösen. In derselben Studie wurde außerdem ein Gefälle zwischen den Geschlechtern hinsichtlich der durchschnittlichen Testleistung offensichtlich – zu Ungunsten

weiblicher Testpersonen. Frauen zeigten signifikant größere Schwierigkeiten als Männer, die Testaufgaben korrekt zu lösen.

Eine Vielzahl von Folgeuntersuchungen beschäftigte sich nunmehr mit diesem Phänomen. Wenn auch ein großer Teil der replizierenden Befunde auf die tatsächliche Existenz einer geschlechtsspezifischen Leistungsdifferenz hinweist, erscheinen die Ergebnisse hinsichtlich des Ausmaßes dieser Differenz jedoch sehr variabel. Eine adäquate Zusammenführung der mittlerweile immensen Befundanzahl auf einen einzigen repräsentativen Gesamtwert ist mittels der Methode des narrativen Review ebensowenig zu bewerkstelligen wie eine eindeutige quantitative Klärung des Einflusses möglicher Moderatorvariablen auf beobachtete Leistungsdifferenzen. Die verhältnismäßig jüngere Methode der Meta-Analyse hingegen findet in den letzten Jahren vermehrt überall dort Anwendung, wo eindeutige, objektive, repräsentative und statistisch abgesicherte Befundintegrationen gefragt sind:

Die Anzahl wissenschaftlicher Publikationen steigt immer rasanter, woraus für Forscher die zunehmend größere Herausforderung resultiert, sich einigermaßen zeitökonomisch einen Überblick über den gegenwärtigen Stand jedweden wissenschaftlichen Themas zu verschaffen. Dies wird nicht zuletzt maßgeblich erschwert durch den Umstand, dass mit steigender Publikationsanzahl auch die Inkonsistenz der Ergebnisse einhergeht, unter anderem resultierend aus einem mehr oder minder großen Stichprobenfehler, dem jede Stichprobenuntersuchung von Natur aus zum Opfer fällt. Diese zunehmend heterogene Befundlage, zusammengesetzt aus einer sich stetig vermehrenden Anzahl von Primärstudien, deren methodische Güte ebenso stark variiert wie die berücksichtigten Variablen hinsichtlich eines bestimmten Forschungsthemas, verlangt nach einer systematischen, validen und reliablen Methode mit hoher statistischer Macht, die möglichst alle thematisch relevanten Befunde aggregieren und deren erwähnte methodische und inhaltliche Heterogenität angemessen berücksichtigen kann: der Meta-Analyse.

Nach einer theoretischen Einführung hinsichtlich des Forschungshintergrundes von Water-Level Tasks einerseits – mit besonderer Berücksichtigung von Theorien zur Entstehung geschlechtsspezifischer Leistungsunterschiede – und der Methode der Meta-Analyse andererseits, hat die vorliegende Arbeit zum Ziel, die Wissenslücke hinsichtlich des mittleren Ausmaßes

der geschlechtsspezifischen Leistungsdifferenz bei Piagets Water-Level Tasks, sowie des Einflusses mannigfaltiger Moderatorvariablen auf diese Differenz zu schließen. Durch ausführliche Literaturrecherchen wurde ein Literaturkorpus erstellt, das den Anspruch hat, alle zugänglichen relevanten Primärstudien, welche analysierbare Daten hinsichtlich des Geschlechtseffekts bei Water-Level Tasks berichten, zu beinhalten.

Die zentralen Phasen der durchgeführten Untersuchung werden dargestellt durch eine umfassende Literaturrecherche in mehreren Durchgängen sowie die Literaturbeschaffung, das Exzerpieren der Primärquellen hinsichtlich relevanter Information, die Beurteilung der Quellen anhand entwickelter In- und Exklusionskriterien für die Analyse, weiters durch die Kodierung inhaltlicher und methodischer Studienmerkmale der relevanten Quellen und schließlich durch die Integration aller extrahierten Daten mit Hilfe der statistischen Software Comprehensive Meta-Analysis (CMA) (Borenstein, Hedges, Higgins & Rothstein, 2005), zu einem globalen mittleren Geschlechtseffekt bezüglich der Leistung bei Water-Level Tasks. Es wurden weiters Auswertungen durchgeführt, um die Wirkung potenzieller Moderatorvariablen auf den globalen Geschlechtseffekt zu untersuchen und damit zur weiteren Aufklärung sowohl möglicher Ursachen des geschlechtsspezifischen Leistungsunterschiedes, seines Verlaufs über die Lebenszeit, als auch seiner langfristigen Veränderung in den letzten fünf Jahrzehnten beizutragen.

Als Erweiterung wurden die Ergebnisse außerdem auf eine mögliche Überschätzung durch die sogenannte Publikationsverzerrung untersucht: Im Allgemeinen wird davon ausgegangen, dass wissenschaftliche Untersuchungen, die starke und/oder statistisch signifikante Effekte berichten, mit größerer Wahrscheinlichkeit sowohl von Forschern zur Publikation eingereicht als auch von Herausgebern der Fachjournale zur Veröffentlichung angenommen werden. Thematisch identische Untersuchungen hingegen, deren Effekte schwach und/oder insignifikant ausfallen, bleiben eher unveröffentlicht, sind daher kaum oder nur erschwert zugänglich und können folglich in das Literaturkorpus einer Meta-Analyse meist nicht einbezogen werden. Dies wiederum hat verzerrende Wirkung auf den in einer Meta-Analyse berechneten mittleren Gesamteffekt. Mehrere Verfahren wurden durchgeführt, anhand welcher sich die Plausibilität einer derartigen Verzerrung im

vorliegenden Fall beurteilen lässt, und mit Hilfe derer gegebenenfalls eine Korrektur des Gesamteffekts vorgenommen wird.

Resultierende Schlussfolgerungen und Einordnung der Resultate in die bestehende Literatur sowie offene Wege für künftige Untersuchungen schließen die vorliegende Arbeit ab.

THEORETISCHER TEIL

1 Water-Level Tasks

1.1 Entwicklungshintergrund und Beschreibung der Testaufgaben

Mit der Absicht, Stufen der Entwicklung räumlicher Konzepte bei Kindern zu testen, entwickelten Piaget und Inhelder (1948/1956) sogenannte Wasserstandsaufgaben (in der Fachliteratur auch als Water-Level Tasks bekannt und in dieser Arbeit auch in weiterer Folge als solche bezeichnet). Die Standardform dieses Tests ist als Papier-Bleistift-Version konzipiert. Der Testperson werden typischerweise sechs oder acht gezeichnete Längsschnitte von handelsüblichen zylinderförmigen Flaschen vorgelegt, die relativ zu einer horizontalen Referenzebene in unterschiedlichen Neigungswinkeln abgebildet sind (Abbildung 1, oben). Aufgabe der Testperson ist es, innerhalb jedes Gefäßes den Verlauf jenes Wasserspiegels einzuzeichnen, der sich bei teilweiser Füllung der Gefäße ergeben würde (Abbildung 1, unten). Das allgemein gültige Lösungsgesetz sollte unmittelbar einsichtig sein: Unabhängig vom Neigungswinkel des Gefäßes verläuft der Wasserspiegel invariant horizontal.

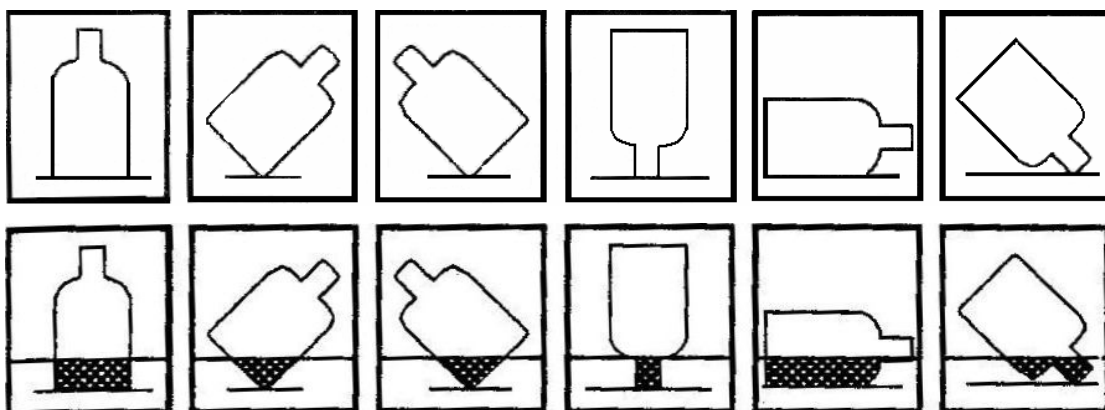


Abbildung 1: Beispiel für Testangabe (oben) und jeweils korrekte Lösung (unten) von Water-Level Tasks. Quelle: Ackermann, 1991.

In nachfolgenden Studien wurde die Standardtestversion auf viele Arten variiert. Von diesen Variationen betroffen sind zum Beispiel die Gefäßform, die Neigungswinkel der präsentierten Gefäße, die Itemanzahl oder auch die Verfügbarkeit einer Referenzebene. Es kommen sowohl Papier-Bleistift-Versionen zur Anwendung, in welchen die jeweilige Lösung händisch einzuzeichnen ist, als auch Multiple-Choice-Tests, in denen es gilt, pro Item das Gefäß mit dem korrekten Verlauf der Wasserlinie aus mehreren Alternativen auszuwählen, oder Testversionen, in denen mit Hilfe vorgefertigter Schablonen der Verlauf des Wasserspiegels richtig anzupassen ist. Auch reale Gefäße oder Fotografien von Gefäßen sind angewandte Abänderungen des Vorgabematerials. Es wird an dieser Stelle auf Kapitel 3 verwiesen, in dessen Rahmen ausführlich alle Variablen und deren Stufen behandelt werden, die in dieser Meta-Analyse berücksichtigt wurden. Kapitel 4.1.1 gibt darüber Aufschluss, nach welchen Kriterien die in der Analyse inkludierten Studien und deren Variablen ausgewählt wurden.

1.1.1 Stufentheorie des Lösungsverhaltens bei Water-Level Tasks

Wie eingangs bereits erwähnt, wurden Water-Level Tasks von Piaget und Inhelder (1948/1956) dazu entwickelt, um die kindliche Entwicklung räumlicher Konzepte zu testen. Wie in Kapitel 1.1 skizziert, ist für die korrekte Lösung dieser Testaufgaben die Kenntnis der invarianten Horizontalität von Flüssigkeiten unabdingbar. Diese Kenntnis setzt der Theorie folgend jedoch die Etablierung eines stabilen externen Referenzsystems voraus (Liben, 1991b), das der Testperson erlaubt, den Wasserspiegel innerhalb eines Gefäßes als parallel zur externen schwerkraftwirkenden Ebene, die horizontal verläuft, zu erkennen. Um den Entwicklungsstand eines derartigen Referenzsystems testen zu können, wandten Piaget und Inhelder (1948/1956) Water-Level Tasks an (mit dem Aufgabenzusatz, nicht nur die Wasserlinie einzuzeichnen, sondern auch die Flüssigkeit selbst). Genannter Entwicklungsstand lässt sich in mehrere Stadien einteilen, Abbildung 2 illustriert Piaget und Inhelders (1948/1956) Stufentheorie des kindlichen Lösungsverhaltens bei Water-Level Tasks.

Eine inhaltliche Komprimierung dieser Theorie, die den entwicklungspsychologischen Hintergrund zu Abbildung 2 zusammenfasst, geschah durch Liben (1991b) und wird nachfolgend im Überblick dargestellt:

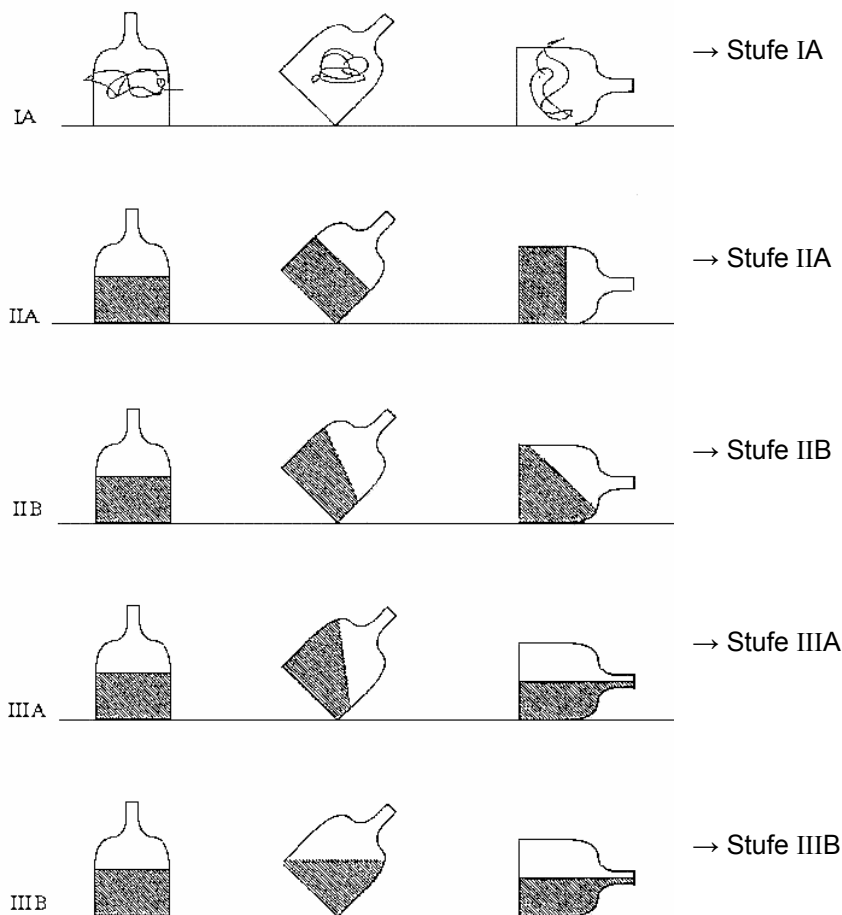


Abbildung 2: Stufentheorie des Lösungsverhaltens bei Water-Level Tasks:
Stufen IA bis IIIB (Piaget und Inhelder, 1948/1956).
Quelle: Liben, 1991b.

Stufe IA: Typischerweise zutreffend für ein Lebensalter zwischen vier und fünf Jahren. Jegliche Art von räumlicher Wahrnehmung, auch von planen Flächen, ist noch nicht möglich. Im Lösungsverhalten zeigt sich dies durch ungerichtete Krakel als Platzhalter für die Flüssigkeit im Gefäß, wobei diese Krakel auch über den Gefäßrand hinausreichen können.

Stufe IIA: Ein räumliches Referenzsystem (interner oder externer Art) fehlt. Folglich wird vom Kind auch keine Bewegung der Flüssigkeit wahrgenommen – weder relativ zum Gefäß (interne Referenz), noch relativ zur Umgebung außerhalb des Gefäßes (externe Referenz). Im Lösungsverhalten findet dies insofern seine Manifestierung, als der Wasserspiegel unabhängig vom Neigungswinkel des Gefäßes in jedem Fall parallel zum Gefäßboden eingezeichnet wird (selbst bei auf dem Kopf stehenden Gefäßen).

Stufe IIB: Ein räumliches Referenzsystem ist bereits etabliert, wenn auch noch sehr instabil. Relative Bewegungen des Wasserspiegels zum Gefäß sind bereits wahrnehmbar (aufgrund der Verwendung einer internen Referenz). Die Aufgabe ist jedoch nur korrekt lösbar, wenn das Gefäß vertikal stehend oder vertikal kopfüber präsentiert wird, alle anderen, horizontal liegenden und schrägen Neigungen zeigen zwar die Wahrnehmung einer Verschiebung des Wasserspiegels relativ zur Gefäßwand, werden jedoch nach wie vor falsch gelöst.

Stufe IIIA: Typischer Beginn im Alter zwischen sieben und acht Jahren. Diese Stufe wird markiert durch die Wahrnehmung einer horizontalen und vertikalen Achse. Externe Referenzpunkte werden jedoch meist noch ignoriert. Im Lösungsverhalten wird dies nun bei der korrekten Lösung für vertikal stehende und horizontal liegende Gefäße sichtbar. Im Falle von nicht-orthogonalen Neigungen des Gefäßes (weder vertikal noch horizontal) wird die Wasserlinie nur zufällig richtig eingezeichnet.

Stufe IIIB: Der Theorie folgend typischerweise im Alter von ca. neun Jahren. Ein stabiles externes Referenzsystem ist etabliert. Im Lösungsverhalten zeigt sich: Die invariante Horizontalität des Flüssigkeitsspiegels wird bewusst und bei allen Gefäßneigungen wahrgenommen, richtig vorausgesagt und reproduziert.

Piaget und Inhelders (1948/1956) Theorie zufolge müsste also Entwicklungsstufe IIIB im Alter von neun Jahren erreicht sein, womit eine korrekte Lösung aller Gefäßneigungen auch bei Erwachsenen einhergehen müsste. Das folgende Kapitel erläutert, inwiefern diese Theorie und die erst später geprüfte Praxis diesbezüglich auseinanderdriften.

1.2 Geschichte des Forschungsgegenstandes

In den 1960er Jahren nahm eine Untersuchungstradition zur Replizierung vieler von Piaget berichteten Erkenntnisse und Theorien seinen Ausgang. In diesem Rahmen widmete sich die Forschungsgemeinschaft auch den Water-Level Tasks (Liben, 1991b), allerdings zunächst nur an Stichproben, die ebenfalls zur Gänze aus Kindern bestanden. Diese Replikationsstudien bestätigten zwar die Einteilung des Lösungsverhaltens in die von Piaget und Inhelder (1948/1956) berichteten Stufen, jedoch konnte keine Bestätigung der Zuteilung zu den postulierten Altersangaben

gefunden werden. Vielmehr wurde an individuellen Testpersonen oftmals variierendes Lösungsverhalten beobachtet, das für mehrere unterschiedliche Entwicklungsstufen typisch ist (Dodwell, 1963). Demnach kann eine Testperson zu einem Testzeitpunkt mehrere Arten von Lösungsverhalten zeigen, die jeweils verschiedenen Entwicklungsstufen zuzuordnen sind.

Erst sechzehn Jahre nach der ersten Veröffentlichung der Water-Level Tasks durch Piaget und Inhelder (1948/1956) und den erwähnten folgenden Replikationsstudien an Kindern wurde dieser Test erstmals auch Erwachsenen vorgegeben. Eine große Überraschung war die Folge. 1964 führte Rebelsky an der Boston University den ersten Test an einer Stichprobe aus Erwachsenen durch, bestehend aus 128 PsychologiestudentInnen (59 männlich, 69 weiblich) im Alter zwischen 20 und 26 Jahren. Es zeigte sich, dass insgesamt mehr als 50 Prozent dieser Stichprobe beim Einzeichnen der Wasserlinie einen Abweichungsfehler produzierten, der größer als fünf Grad war. Ganz besonders fiel jenes Testitem auf, in dem das Gefäß um 60 Grad (von der Vertikalen aus) nach links gekippt präsentiert wurde: Der mittlere Abweichungsfehler betrug hierbei bei männlichen Testpersonen 11 Grad, während er bei weiblichen Testpersonen 20.7 Grad betrug. Mit Rebelsky (1964) wurde somit ein Gefälle zwischen Männern und Frauen hinsichtlich der Leistung bei Water-Level Tasks erstmals publiziert und seitdem von unzähligen Forschern weiteren Replikationen, Variationen und anderen Untersuchungen unterzogen (eine Liste der in der vorliegenden Meta-Analyse berücksichtigten Studien dieses Themas zeigt Tabelle 1).

Neben der Ermittlung des konkreten Ausmaßes der mittleren geschlechtsspezifischen Leistungsdifferenz sieht die Forschungsgemeinschaft heute die Ursachenklärung dieser Leistungsdifferenz als besondere Herausforderung. Das folgende Kapitel soll einen Überblick über einige Theorien geben, die diesbezüglich aufgestellt und zum Großteil bereits mehrfach geprüft wurden.

1.3 Theorien zur Entstehung geschlechtsspezifischer Leistungsunterschiede bei Water-Level Tasks

Welche Ursachen könnten hinter dem Leistungsunterschied zwischen den Geschlechtern stehen? Liben (1991b) weist darauf hin, dass das biologische Geschlecht allein nicht als kausal wirksame Variable ausreichen kann, um Leistungsunterschiede bei Water-Level Tasks grundsätzlich zu erklären. Vielmehr ist

das biologische Geschlecht als Indikator für andere Eigenschaften zu verstehen, die ihrerseits für unterschiedliche Leistung im Rahmen von Water-Level Tasks verantwortlich sind. Die Frage muss also lauten: Durch welche Variablen, die in engem Zusammenhang mit dem biologischen Geschlecht stehen, könnte die Leistung bei Water-Level Tasks beeinflusst werden?

Vor einem Überblick und thematischen Einteilung von Theorien, welche die Ursache geschlechtsspezifischer Leistungsunterschiede klären sollen, muss darauf hingewiesen werden, dass die einzelnen behandelten Variablen nicht grundsätzlich auf derselben Ursachenebene hinsichtlich des Geschlechtsunterschiedes zu sehen sind, sondern dass ursächliche Vernetzungen zwischen manchen von ihnen möglich und auch nicht zu verhindern sind. Auf eingehende Beleuchtung derartiger Vernetzungen wird jedoch zu Gunsten des Überblicks hier bewusst verzichtet.

1.3.1 Biologische Korrelate

Die hier vorgestellten potentiellen Ursachentheorien zum Geschlechtseffekt, die biologischen Korrelaten zuzuordnen sind, betreffen die folgenden drei Variablen: Körpergröße, endokrine Faktoren sowie biogenetische Faktoren. Sie werden im Folgenden näher erörtert.

1.3.1.1 Körpergröße

Ein Geschlechtsunterschied hinsichtlich eines biologischen Korrelats betrifft die Körpergröße: Frauen sind im Durchschnitt kleiner gewachsen als Männer (Percival & Quinkert, 1987; zitiert nach Robert & Harel, 1996). Dies könnte sich insofern auf die Testleistung bei Water-Level Tasks auswirken, als dass größer gewachsene Testpersonen, meist also Männer, aus evolutionärer Sicht gezwungen gewesen sein könnten, ein sensibleres System zur Wahrnehmung, Überwachung und Kontrolle der Balance zu entwickeln, um trotz ihres hochgewachsenen Körpers nicht das Gleichgewicht zu verlieren und umzufallen. Dieses (hypothetisch) feiner entwickelte Balance-System würde eine bessere Orientierung im Koordinatensystem, bestehend aus horizontalen und vertikalen Achsen mit sich bringen und könnte sich daher unter anderem in einer besseren Testleistung bei Water-Level Tasks zeigen, die ja vor allem darin besteht, die Horizontalität von Wasserspiegeln korrekt wiederzugeben. Körpergröße und Erfolg bei Water-Level Tasks müssten demnach positiv korreliert sein. Es wurden jedoch keine Hinweise auf eine Bestätigung dieser Theorie gefunden (Robert & Harel, 1996; Robert & Ohlmann, 1994).

1.3.1.2 Endokrine Faktoren

Mit dem Wissen, dass die bereits genannte mögliche biologische Ursachenvariable Körpergröße ebenfalls auf die Wirkung von Hormonen zurückgeführt werden kann, wird nun ein weiterer potentieller endokriner Faktor genannt: Sanders und Ross-Field (1986) führen Androgene und deren Wirkung auf den Fetus in einer kritischen pränatalen Phase ins Treffen. Konsistent mit der Theorie von Geschwind und Galaburda (1987; zitiert nach Pascual-Leone & Morra, 1991) ziehen sie mit der durch die virilisierenden Sexualhormone beeinflusste Lateralisation des Gehirns die Ausbildung auch jener Fähigkeiten in Betracht, die für die korrekte Beantwortung von Water-Level Tasks unabdingbar sind. Diese Theorie steht weiteren Untersuchungen in diese Richtung für eine definitive Klärung noch offen.

1.3.1.3 Biogenetische Faktoren

Ein X-chromosomales rezessives Gen ist Kern der von Thomas und Jamison (1981a) postulierten Kausalhypothese bezüglich des Geschlechtsunterschieds in der Testleistung bei Water-Level Tasks. Einige Studien folgten sofort, mit der Absicht, diese Theorie weiter zu untersuchen (Allen, Wittig & Butler, 1981; McGee, 1981; Thomas & Jamison, 1981b; Walker, Krasnoff & Peaco, 1981). Ergebnisse der Untersuchungen konnten die Theorie jedoch nicht eindeutig bestätigen (Pascual-Leone und Morra, 1991). Thomas und Jamison (1981b) schlossen daraus, dass ihre Hypothese korrekt war und die Unschärfe der Ergebnisse lediglich auf die Wirksamkeit zusätzlicher – unbekannter – Variablen zurückzuführen sei. Pascual-Leone und Morra (1991) schließen die Diskussion mit dem Argument, dass ohne die Kenntnis dieser möglichen weiteren Variablen eine Untersuchung der Hypothese mit eindeutigem Resultat nicht möglich ist.

1.3.2 Erfahrung und Wissen

Als zweiter möglicher Hauptaspekt hinsichtlich Ursachentheorien zum Geschlechtsunterschied in der Testleistung bei Water-Level Tasks wird neben biologischen Korrelaten Erfahrung und Wissen bezüglich jener Inhalte genannt, die zum Erfolg bei Water-Level Tasks beitragen können. Drei Variablen, die diesem Themenkreis zuzuordnen sind, werden hinsichtlich ihrer Berücksichtigung in der

Fachliteratur behandelt: Frühe Erfahrung in räumlichen Aktivitäten, allgemeines physikalisches Wissen sowie spezielles Wissen über das Invarianzprinzip.

1.3.2.1 Frühe Erfahrung in räumlichen Aktivitäten

Bereits im Kindesalter könnte sich anhand der gehäuft gewählten spielerischen Tätigkeiten zeigen, ob ein Erfolg in Sachen Water-Level Tasks wahrscheinlich ist oder nicht. Ob das bei Knaben häufiger vermutete Spiel mit Schwerpunkt auf räumlichen Aktivitäten tatsächlich in Zusammenhang mit besserem Erfolg in Water-Level Tasks steht, wurde vielfach untersucht, mit Resultaten, die sich entweder widersprechen (Beatty & Duncan, 1990; vs. Héroux & Robert, 1993) oder auf einen Zusammenhang zwischen räumlicher Aktivität und späterer Fertigkeit nur bei Frauen schließen lassen (Robert & Harel, 1996; Signorella, Jamison & Krupa, 1989; Signorella, Krupa, Jamison & Lyons, 1986; Voyer, Nolan & Voyer, 2000).

1.3.2.2 Allgemeines physikalisches Wissen

Bezüglich des in Ausbildung oder Beruf erworbenen physikalischen Wissens als Ursache des Geschlechtsunterschiedes zeigt sich in der Literatur ein ähnlich unstetes Bild wie schon bei der Variablen frühe Erfahrung in räumlichen Aktivitäten: Physikalisches Wissen trägt laut Kalichman (1991) zwar zum Erfolg bei Water-Level Tasks bei, dies jedoch abermals nur bei Frauen, nicht jedoch bei Männern. Da vor allem zwischen einzelnen Berufssparten, deren Wahl wiederum häufig geschlechtsspezifisch ausfällt, Unterschiede im (vor und während der Ausübung des Berufs erworbenen) physikalischen Wissen erwartet werden, wurde auch diese Variable hinsichtlich ihres Einflusses auf den Testerfolg untersucht: Die gewählte Berufssparte hängt laut Robert und Harel (1996) sowie laut Vasta, Rosenberg, Knott und Gaze (1997) zwar mit dem Testerfolg bei Water-Level Tasks zusammen, allerdings bestehen in den untersuchten Berufssparten weiterhin signifikante Geschlechtsunterschiede zu Gunsten von männlichen Testpersonen. Bezüglich der Leistungsunterschiede zwischen Personengruppen mit typischerweise unterschiedlichem Ausmaß an Erfahrung mit Wasserspiegeln ergab sich sogar ein der Erfahrungshypothese entgegengesetztes Resultat: Laut Hecht und Proffitt (1995) korrelieren Ausmaß der alltäglichen Erfahrung mit Wasserspiegeln und Erfolg bei Water-Level Tasks negativ (wobei der postulierte Geschlechtseffekt im Testerfolg innerhalb jeder Personengruppe weiterhin bestehen bleibt).

1.3.2.3 Spezielles physikalisches Wissen über das Invarianzprinzip

Sind Männer über die invariante Horizontalität des Wasserspiegels (im Fachjargon kurz als Invarianzprinzip bezeichnet) von vornherein besser informiert, und erleichtert ihnen dieses spezielle Wissen die richtige Beantwortung von Water-Level Tasks? Mehreren Untersuchungen zufolge trägt dieses spezielle Wissen zwar zur korrekten Beantwortung von Water-Level Tasks bei, es kann jedoch die Leistungsdifferenz zwischen den Geschlechtern nicht zur Gänze erklären (Liben & Golbeck, 1984; Lohaus & Kessler, 1996; McAfee & Proffitt, 1991; Myer & Higgins-Hensley, 1984; Pascual-Leone & Morra, 1991; Pulos, 1997; Vasta, Lightfoot & Cox, 1993).

1.3.3 Räumliche Fähigkeiten

Naheliegender ist die Annahme eines Zusammenhangs zwischen räumlichen Fähigkeiten und dem Testerfolg bei Water-Level Tasks. Ein Geschlechtseffekt zu Gunsten von Männern hinsichtlich räumlicher Fähigkeiten ist mehrfach nachgewiesen (z.B. Halpern, 2000; Hyde, 1981; Linn & Petersen, 1985; Voyer, Voyer & Bryden, 1995), wobei die stärksten geschlechtsspezifischen Leistungsunterschiede im Fähigkeitsbereich der mentalen Rotation auftreten, wie meta-analytische Untersuchungen zeigen konnten (Linn & Petersen, 1985; Voyer, Voyer & Bryden, 1995). Diese Geschlechtsdifferenz in Leistungen der mentalen Rotation kann bereits im frühen Kindesalter sichtbar gemacht werden (Koscik, O'Leary, Moser, Andreasen & Nopoulos, in press; Moore & Johnson, 2008; Quinn & Liben, 2008). Auch bei der Lösung von Water-Level Tasks kann die Beteiligung einer Komponente des räumlichen Fähigkeitsbereichs angenommen werden. In Untersuchungen, die Water-Level Tasks nicht als eigentlichen Forschungsgegenstand behandeln, sondern diese lediglich als Testinstrument für einen anderen Zweck einsetzen, wird häufig davon ausgegangen, dass Water-Level Tasks selbstverständlich zur Messung von allgemeinen räumlichen Fähigkeiten eingesetzt werden können (z.B. Mowrer-Popiel & DeLisi, 1984). In Untersuchungen, die sich speziell mit dem Phänomen des Geschlechtseffekts hinsichtlich der Leistung in Water-Level Tasks beschäftigen, findet sich bis dato allerdings nur selten ein Design, das gezielt die Theorie von räumlichen Fähigkeiten als UrsachenvARIABLE prüfen soll. Dies zeigt sich bei Sichtung der verfügbaren Literatur bezüglich Water-Level Tasks und wird weiterhin auch von Merriwether und Liben (1997) zusammenfassend festgestellt. Lediglich eine

Untersuchung fällt auf, die das Ziel, Geschlechtsunterschiede in den Leistungen bei Water-Level Tasks allein auf Geschlechtsunterschiede in den Leistungen bei einem Test zu räumlichen Fähigkeiten zurückzuführen, erreicht hat: Geiringer und Hyde (1976) setzten als Maß für räumliche Fähigkeiten den Spatial Relations Test, einen Untertest aus dem Primary Mental Abilities Test (Thurstone & Thurstone, 1962) ein. Die Beträge der berichteten Korrelationen zwischen Spatial Relations Test und Water-Level Tasks sind hoch, sie reichen von .42 bis .97. Im Zuge ihrer Untersuchung gelang es Geiringer und Hyde (1976) außerdem, durch Kontrolle der Leistungsunterschiede im Spatial Relations Test signifikante Geschlechtsunterschiede in der Leistung bei Water-Level Tasks vollständig zu beseitigen. Aufgrund dieser Ergebnisse gehen die Autoren davon aus, dass Water-Level Tasks im Grunde räumliche Fähigkeiten messen. Einschränkend geben sie allerdings zu bedenken, dass aufgrund fehlender Normalverteilung ihrer Daten bei der Ergebnisinterpretation Vorsicht geboten ist: Eine Verletzung der Normalverteilungs-Voraussetzung kann zu potentiell verzerrten Prüfgrößen und damit zu Fehlinterpretationen führen. Möglicherweise ist dieser Umstand die Ursache dafür, weshalb dieser Befund bei damals (im Jahr 1976) zeitgenössischen Forschern nicht als entscheidender Beitrag zur Lösung des Rätsels um den Geschlechtsunterschied bei Water-Level Tasks gesehen wurde und auch nicht zu inhaltlicher Diskussion oder Replikationsstudien anregen konnte. Generell fand diese Studie verhältnismäßig wenig Resonanz: Von sechszwanzig Zitierungen dieser Studie in nachfolgenden Untersuchungen wird lediglich in zwei Quellen (Krekling & Nordvik, 1992; Liben, 1991b) auf das Ergebnis hinsichtlich eines möglichen Zusammenhangs zwischen räumlichen Fähigkeiten und Water-Level Tasks Bezug genommen.

Der Zusammenhang zwischen einem weiteren Test aus dem Bereich räumlicher Fähigkeiten und Water-Level Tasks hingegen wurde mehrfach repliziert und dokumentiert, und dieser zeigt sich erheblich schwächer, als dies aufgrund der Schlussfolgerungen von Geiringer und Hyde (1976) zu erwarten wäre: Berichtete Korrelationen zwischen dem Mental Rotations Test und Water-Level Tasks bewegen sich mehreren Untersuchungen zufolge zwischen .39 und .57 (Casey, Nuttall & Pezaris, 2001; Lawton, 1994; Levine & Stern, 2002; Quaiser-Pohl, Lehmann & Eid, 2004; Voyer & Sullivan, 2003) und sind damit immerhin substantiell. Tabelle 1 zeigt diesbezüglich einen Überblick.

Tabelle 1: In Literatur berichtete Korrelationen zwischen Water-Level Tasks und MRT

Studie	Korrelation zwischen WLT und MRT (Stichprobe innerhalb der Studie)	<i>p</i>
Casey, Nuttall & Pezaris (2001)	.52	n.a.
Lawton (1994)	.39	<.01
Levine & Stern (2002)	.44	<.05
Quaiser-Pohl, Lehmann & Eid (2004)	.37	<.05
Voyer & Sullivan (2003)	.43 (timed testing)	<.05
	.57 (untimed testing)	<.05
	.52 (group testing)	<.05
	.39 (individual testing)	<.05

Für eine differenzierte Aussage über die Beteiligung räumlicher Fähigkeiten an der Lösung von Water-Level Tasks erscheinen daher genauere Untersuchungen aussichtsreich, die unterschiedliche Bereiche räumlicher Fähigkeiten in Zusammenhang mit Water-Level Tasks berücksichtigen: Meta-analytische Untersuchungen zu Geschlechtsunterschieden in räumlichen Fähigkeiten differenzieren räumliche Wahrnehmung, mentale Rotation und räumliche Visualisierung als voneinander abgegrenzte Kategorien innerhalb des gesamten räumlichen Fähigkeitsspektrums (Linn & Petersen, 1985; Voyer, Voyer & Bryden, 1995), die sich auch in der Größe der geschlechtsspezifischen Leistungsdifferenz voneinander unterscheiden. Es ist wahrscheinlich, dass die genannten Kategorien an der Lösung von Water-Level Tasks in unterschiedlichem Ausmaß beteiligt sind. Analysen in diese Richtung könnten zur weiteren Ursachenklärung des Geschlechtseffekts bei Water-Level Tasks maßgeblich beitragen.

Testtheoretische Beleuchtungen der beobachteten Resultate können außerdem weiteren Aufschluss über die tatsächliche Stärke der Beteiligung von Fähigkeiten der mentalen Rotation an der Lösung von Water-Level Tasks liefern (siehe Kapitel 6.4).

1.3.4 Soziokognitive Korrelate

Mögliche Ursachentheorien mit soziokognitiver Komponente betreffen Erfolgserwartung und Attributionsstil sowie die Geschlechtsrolle, worauf im Folgenden eingegangen wird.

1.3.4.1 Erfolgserwartung und Attributionsstil

Hinsichtlich mathematischer Fähigkeiten wurde Unterstützung für die Hypothese gefunden, dass individuelle Erfolgserwartungen mit den Testleistungen korrelieren, sodass die in Verbindung mit mathematischen Aufgaben typischerweise herabgesetzten Erfolgserwartungen von Frauen eine schlechtere Testleistung implizieren (Wigfield & Eccles, 1992; zitiert nach Casey, Nuttall & Pezaris, 2001). Bei Aufgaben im naturwissenschaftlichen und mathematischen Bereich allgemein sind für Mädchen variable Faktoren wie Glück oder Arbeitsanstrengung typische Ursachenzuschreibungen für ihren Testerfolg, während sie für ihr Scheitern stabile Faktoren wie mangelnde Fähigkeiten oder Schwierigkeit der Aufgabe als verantwortlich sehen (Ryckman & Peckham, 1987; zitiert nach Robert & Harel, 1996). Laut Whitley, McHugh und Frieze (1986) ist dieses geschlechtstypische Attributionsmuster in einer großen Auswahl weiterer Bereiche zu beobachten (unter anderem sehr ausgeprägt im Bereich räumlicher Fähigkeiten). Speziell betreffend das Aufgabenmaterial von Water-Level Tasks konnten Meehan und Overton (1986) zeigen, dass Männer sowohl eher mit ihrem eigenen Erfolg rechnen als auch ihren Testerfolg eher auf stabile Faktoren wie ihre Fähigkeit oder die Leichtigkeit der Aufgabe zurückführen, als dies Frauen tun. Dies konnte von Robert und Harel (1996) repliziert werden. Niedrigere Erfolgserwartung und ungünstiger Attributionsstil bei Frauen sind somit mit schlechteren Testleistungen korreliert. Bei der Lösung von Water-Level Tasks erscheinen Männer generell von der Richtigkeit jeder ihrer Aufgabenbeantwortungen stärker überzeugt als Frauen (Neumann, Fitzgerald, Furedy & Boyle, 2007). Um zu überprüfen, ob Erfolgserwartung und Testleistung bei Water-Level Tasks kausal in Zusammenhang gebracht werden können, variierten Meehan und Overton (1986) in einem weiteren Schritt die Erfolgserwartung von männlichen und weiblichen Testpersonen experimentell (nicht ohne die Wirksamkeit dieser Maßnahme zu überprüfen). Die Ergebnisse zeigen, dass eine herabgesetzte Erfolgserwartung bei Männern mit schlechteren Testleistungen bei Water-Level Tasks einhergeht, während sich hingegen trotz erhöhter Erfolgserwartung bei Frauen

deren Testleistungen kaum verändern. Der logische Schluss aus diesem Befund legt nahe, dass trotz eines beobachtbaren Zusammenhangs zwischen Erfolgserwartung und Testleistung bei Water-Level Tasks zumindest bei Frauen kein kausaler Einfluss der Erfolgserwartung auf die Testleistung – und damit auf den Geschlechtseffekt – zu sehen ist (siehe auch Liben, 1991b; Robert & Harel, 1996).

1.3.4.2 Geschlechtsrolle

Die Hypothese bezüglich Geschlechtsrolle als ausschlaggebendes Kriterium für den Geschlechtsunterschied in der Leistung bei Water-Level Tasks besagt, dass eine maskuline Rolle in der Gesellschaft einhergeht mit höherem Testerfolg, während niedrigerer Testerfolg bei Water-Level Tasks korreliert mit einer femininen Rolle. Maskulinität bzw. Femininität ist nicht als synonym mit dem biologischen Geschlecht zu verstehen (wenn auch stark abhängig davon), sondern als ein kulturell definiertes Rollenverständnis (Bem, 1981). Prinzipiell ist für Männer eine feminine Geschlechtsrolle ebenso möglich wie für Frauen eine maskuline. Typischerweise werden mit Maskulinität bzw. Femininität bestimmte Ausprägungen bezüglich sexueller Orientierung, Rolle in der Familie, Beziehungsstil, nonverbalem Verhalten und Auftreten sowie bezüglich bestimmter Interessen und Hobbies assoziiert (Lippa, 2001). Überprüft wurde die Theorie der Abhängigkeit besserer Leistungen von einer maskulinen Rolle für räumliche und mathematische Aufgaben im Rahmen einer Meta-Analyse von Signorella und Jamison (1986), deren Resultat hypothesenkonform ausfiel.

Bezüglich Water-Level Tasks im Speziellen zeigt sich hingegen abermals ein gemischtes Bild: Hypothesenunterstützend für beide Geschlechter fielen die Ergebnisse ebenso aus (Jamison & Signorella, 1980), wie sich in anderen Studien der postulierte Zusammenhang nur bei Frauen zeigte (Kalichman, 1989a; Signorella & Jamison, 1978; Signorella, Jamison & Krupa, 1989) oder aber nur bei Männern (Krasnoff, Walker & Howard, 1989), oder auch geschlechtsunabhängig überhaupt kein Hinweis auf den Einfluss der Geschlechtsrolle auf die Testleistung bei Water-Level Tasks zu finden war (Mowrer-Popiel & DeLisi, 1984; Walker & Krasnoff, 1978).

1.3.5 Perzeptive Korrelate

Betreffend Ursachentheorien zum Geschlechtseffekt bei Water-Level Tasks, die auf unterschiedliche Wahrnehmungsmuster Bezug nehmen, wurde vor allem der

kognitive Stil mit dem Konstrukt der Feldabhängigkeit bzw. Feldunabhängigkeit untersucht und wie folgt beschrieben.

1.3.5.1 Kognitiver Stil: Feldabhängigkeit/Feldunabhängigkeit

Die Begriffe Feldabhängigkeit und Feldunabhängigkeit als Ausdruck des kognitiven Stils sollen (neben anderen inhaltlichen Postulaten) ausdrücken, wie stark eine Person von Reizen in ihrem visuellen Blickfeld in ihrer Wahrnehmung beeinflusst und von anderen Reizen abgelenkt wird (Halpern, 2000). Ein Test, der das Vorherrschen des einen oder anderen kognitiven Stils bei Testpersonen zu messen vermag, ist der Rod-and-Frame Test (kurz RFT). Seine Testanforderung besteht darin, einen geraden Stab (*engl.* „Rod“) innerhalb eines in verschiedenen Neigungswinkeln präsentierten rechtwinkligen Rahmens (*engl.* „Frame“) exakt vertikal zu platzieren. Testpersonen mit feldabhängigem kognitiven Stil fällt diese Aufgabe sehr viel schwerer als jenen mit feldunabhängigem Stil – der Theorie nach aus folgendem Grund: Die Verwendung eines externen Referenzsystems (das ist der schwerkraftwirkende Boden, der Vertikalität und Horizontalität absolut gesehen definiert), das losgelöst vom unmittelbaren Aufgabenkontext existiert (dem rechtwinkligen Rahmen, der als vermeintlich externes Referenzsystem missverstanden werden kann), ist für Personen mit feldabhängigem kognitiven Stil ungleich schwieriger als für Personen mit feldunabhängigem kognitiven Stil. Eine inhaltlich ganz ähnliche Problematik ergibt sich bei der Beantwortung von Water-Level Tasks. Eine große Anzahl von in der Literatur berichteten signifikanten bis hochsignifikanten Korrelationen (von $r = .41$ bis $r = .75$) zwischen Rod-and-Frame Test und Water-Level Tasks, auch übersichtlich zusammengefasst in einem Review von Pascual-Leone und Morra (1991), stellt eindrucksvoll dar, dass zur Lösung von Water-Level Tasks und dem Rod-and-Frame Test ähnliche Fähigkeiten eingesetzt werden müssen (siehe Tabelle 2).

Tabelle 2: In Literatur berichtete Korrelationen zwischen Water-Level Tasks und Rod-and-Frame Test

Studie	Korrelation zwischen WLT und RFT (mittleres Alter bzw. Bildungsstufe der Stichprobe)	<i>p</i>
Abravanel & Gingold (1977)	.43 (9;11)	<.01
	.51 (10;2)	<.01
DeLisi, R. (1983)	.75 (6;9)	<.001
	.41 (8;7)	>.05
	.53 (10;7)	<.05
Linn & Kyllonen (1981)	.32 (Senior High School)	<.05
Quaiser-Pohl, Lehmann & Eid (2004)	.55 (9;9)	<.05
Sholl (1989)	.41 (Undergraduates)	<.01

Der kognitive Stil Feldabhängigkeit/Feldunabhängigkeit zeigt sich demzufolge in der individuellen Testleistung auch bei Water-Level Tasks: Personen mit feldabhängigem kognitiven Stil zeigen auch hier schlechtere Testleistungen als Personen mit feldunabhängigem kognitiven Stil (z.B. Pascual-Leone & Morra, 1991). Bei feldabhängigen Personen wird dies verursacht durch Nicht-Anwendung des externen Referenzsystems zu Gunsten gefäßinterner Referenzsysteme, welche die Produktion absolut horizontaler Wasserlinien verhindern. Betreffend die westliche Gesellschaft ist dieser Leistungsunterschied konform mit einem größeren Anteil von Frauen unter den feldabhängigen Personen, (Witkin & Goodenough, 1981; zitiert nach Robert & Harel, 1996). Somit ist ein Zusammenhang zwischen kognitivem Stil und Testerfolg bei Water-Level Tasks sichtbar und belegt. Hinsichtlich der Frage, ob der kognitive Stil auch als kausale Ursache für den Geschlechtsunterschied in Testleistungen bei Water-Level Tasks verstanden werden kann, wurden allerdings auch Zweifel geäußert: Es wird postuliert, dass Geschlechtsunterschiede im kognitiven Stil eine simple Widerspiegelung der (bekannten und mehrfach belegten) Geschlechtsunterschiede im visuell-räumlichen Bereich sein könnten, da angenommen wird, dass die zur Messung des kognitiven Stils verwendeten Tests auch eine starke räumliche Komponente aufweisen (Sherman, 1967; Halpern, 2000). Außerdem wird argumentiert, dass ein Teil der Varianz der Testergebnisse bei Water-Level Tasks auch bei Berücksichtigung des kognitiven Stils unerklärt verbleibt (Pascual-Leone & Morra, 1991). Ungeachtet dessen deuten einige Forschungsergebnisse darauf hin, dass das Konstrukt

Feldabhängigkeit/Feldunabhängigkeit nicht zu einer isolierten Dimension vereinfacht werden kann, sondern viel eher als komplexer Prozess zu sehen ist (Pascual-Leone 1969; Pascual-Leone & Goodman, 1979; zitiert nach Pascual-Leone & Morra, 1991): Feldabhängigkeit/Feldunabhängigkeit kann verstanden werden als eine Neigung, sich für eine von zwei einander widersprechenden Verarbeitungsstrategien zu entscheiden, die in bestimmten Situationen relevant erscheinen. Feldabhängigkeit wird beschrieben als die Neigung zur Wahl jener Strategie, die unmittelbar eingeleitet wird durch besonders auffällige Reize der Situation und in weiterer Folge zu einem Irrtum führt. Feldunabhängigkeit hingegen wird beschrieben als die Neigung zur Wahl jener Strategie, die dafür sorgt, dass aufgabenrelevante Schemata aktiviert und angewandt werden. Mentale aufmerksamkeitsbezogene Energie sowie ein aufmerksamkeitsbezogener Unterbrechungsmechanismus sollen weiters für die Wirksamkeit der passenden Schemata und damit zur Wahl der richtigen Strategie sorgen (Pascual-Leone, 1992).

1.3.6 Motorische Fähigkeiten

Als letzte der hier behandelten Theorien zum Geschlechtseffekt bei Water-Level Tasks soll auf motorische Fähigkeiten eingegangen werden: Können die schlechteren Testleistungen von Frauen darauf zurückgeführt werden, dass weibliche Testpersonen motorische Schwierigkeiten haben, eine horizontale Linie zu zeichnen? Liben und Golbeck (1980, 1986) prüfen diese Hypothese, indem sie explizit das Zeichnen einer horizontalen Linie (ohne ablenkenden Hintergrund eines geneigten Gefäßes) verlangen. In den Ergebnissen zeigt sich kein geschlechtsspezifischer Leistungsunterschied. Das Zeichnen einer horizontalen Linie beherrschen Frauen demnach ebenso gut wie Männer.

Weiters wurde von Thomas, Jamison und Hummel (1973) ein Apparat konstruiert, der sogenannte Thomas-Apparatus, der es erlaubt, Water-Level Tasks ohne zeichnerische Leistung zu lösen: In einer Tafel ist die ausgestanzte Silhouette einer handelsüblichen Bierflasche zu sehen. Hinter dieser Stanzung ist auf der Tafelrückseite eine kleine kreisförmige, um ihren Mittelpunkt drehbare Platte angebracht, die genau bis zur Hälfte eingefärbt ist. Diese Schablone dient als Platzhalter für die Wasserfüllung und deren Pegelverlauf innerhalb der Flasche. Die Testperson hat durch Adjustierung der drehbaren Platte den Verlauf des Pegels innerhalb der Flasche, die in unterschiedlichen Neigungen präsentiert wird,

anzugeben. In Untersuchungen, in denen der Wasserspiegel auf diese Weise zu adjustieren anstatt zu zeichnen ist, ergaben sich ebenfalls Geschlechtsunterschiede zu Gunsten von männlichen Testpersonen (z.B. Geiringer, 1976; Geiringer & Hyde, 1976; Thomas & Jamison, 1975; Thomas & Lohaus, 1993). Dieses Ergebnis kann auch in Hinblick auf einen Fund von Collaer, Reimer und Manning (2007) nicht überraschen, der darauf hinweist, dass Frauen in der Beurteilung von Linien hinsichtlich deren Winkel (und Position) größere Schwierigkeiten haben als Männer: Aus einer Vorlage bestehend aus mehreren (genau: fünfzehn), in einem Halbkreis aufgefächerten radial verlaufenden Linien in gleichen Abständen zueinander muss jene Linie identifiziert werden, die in Winkel und Position der jeweils einzeln vorgegebenen Stimuluslinie entspricht. Die korrekte Wahl der korrespondierenden Linie im Rahmen des sogenannten Judgement-of-Line-Angle-and-Position-15-Tests (JLAP-15) von Collaer, Reimer und Manning (2007) fällt weiblichen Testpersonen deutlich schwerer als männlichen: Ein d von 0.57 spricht für einen moderaten Geschlechtseffekt zu Gunsten von Männern hinsichtlich der beschriebenen Testleistung.

Natürlich erlaubt auch eine Testvorgabe in Form von Multiple-Choice-Aufgaben eine Beantwortung der Water-Level Tasks unter der Ausklammerung von motorischen Fähigkeiten. Ergebnisse von entsprechenden Untersuchungen zeigen ebenfalls Geschlechtsunterschiede in der bereits bekannten Richtung (Amponsah, 2000; Harris, Hanley & Best, 1977; Lawton, 1994; Vasta, Lightfoot & Cox, 1993).

Aufgrund der genannten Ergebnisse kann davon ausgegangen werden, dass motorische Fähigkeiten als Ursache für beobachtete Geschlechtsunterschiede hinsichtlich der Leistungen in Water-Level Tasks auszuschließen sind.

1.3.7 Zusammenfassung

Betreffend die oben genannten Theorien, welche die Leistungsunterschiede zwischen den Geschlechtern bei Water-Level Tasks erklären sollen, kann zusammenfassend gesagt werden: Körpergröße und motorische Fähigkeiten können als kausal wirksame Variablen mit großer Sicherheit ausgeschlossen werden. Anders verhält es sich mit den übrigen Variablen: endokrine und biogenetische Faktoren, frühe Erfahrung in räumlichen Aktivitäten, allgemeines physikalisches Wissen, spezielles physikalisches Wissen über das Invarianzprinzip, räumliche Fähigkeiten, Erfolgserwartung und Attributionsstil, Geschlechtsrolle und kognitiver Stil. Für jede

einzelne dieser Variablen gibt es unterstützende Hinweise (die unterschiedlich stark ausfallen) hinsichtlich ihrer Beteiligung am Geschlechtseffekt bei Water-Level Tasks. Allerdings vermag keine der bisher untersuchten Hypothesen den Geschlechtseffekt vollständig zu erklären. Diese Ansicht wird von Forschern vertreten, die sich in ihren Untersuchungen und Übersichtsarbeiten speziell diesem Phänomen widmen (Kalichman, 1988; Liben, 1991b; Merriwether & Liben, 1997; Robert & Harel, 1996; Vasta & Liben, 1996).

1.4 Bisherige themenverwandte Analysen

Mit dem viel zitierten Review von Maccoby und Jacklin (1974; zitiert nach Linn & Petersen, 1985) nahm die damals auch emanzipatorisch motivierte meta-analytische Forschung zu Geschlechtsunterschieden in psychologisch relevanten Fertigungsbereichen seinen Ausgang (Hyde, 1986). Hyde (1981) wandte sieben Jahre später das Verfahren der Meta-Analyse auf einen Teil des Literaturkorpus aus Maccoby und Jacklins (1974) Review an. In diesen beiden umfassenden Arbeiten befasste man sich allerdings noch nicht mit dem Thema der Water-Level Tasks und inkludierte daher auch keine dieser speziellen Aufgaben. Linn und Petersen (1985) widmeten sich zeitlich nachfolgend im Rahmen einer Meta-Analyse der Frage nach der Größe von Geschlechtsunterschieden speziell in räumlichen Fähigkeiten. In diese Analyse flossen zwar erstmals Daten zu Water-Level Tasks ein (die aus dreizehn publizierten Studien von 1975 bis 1982 stammen), allerdings werden für sie keine separaten Ergebnisse berichtet. Auch Signorella und Jamison (1986) inkludierten Water-Level Tasks in ihrer Meta-Analyse zu kognitiven Leistungen mit besonderer Berücksichtigung von Maskulinität, Femininität und Androgynie. Allerdings nahmen diese Autoren lediglich acht (vor 1986 publizierte) Studien, die Ergebnisse zu Water-Level Tasks berichten, in ihre Analyse auf und berichten überdies abermals keine Analyseergebnisse bezüglich Water-Level Tasks. Witters-Churchill und Witters' (1989) Meta-Analyse hat erstmals explizit die Leistungen in Water-Level Tasks zum Thema, allerdings befassen sich die Autoren ausschließlich mit besonderen Populationsgruppen wie Personen mit Hörschädigung, hochbegabten Personen mit Hörschädigung oder Personen mit Verhaltens- und/oder Lernstörung. Überdies wird der mittlere Geschlechtseffekt von lediglich fünf unabhängigen Stichproben berichtet (zwischen $d = 0.08$ und $d = 1.53$), wobei weiters keine Aussage über einen globalen Geschlechtseffekt gemacht wird.

Die erste und bisher einzige Untersuchung, die dezidierte Aussagen über die Größe des globalen Geschlechtsunterschiedes hinsichtlich der Leistung in Water-Level Tasks macht, wurde vor eineinhalb Jahrzehnten publiziert: Voyer, Voyer und Bryden (1995) sprechen in ihrer Meta-Analyse von einer Effektstärke $d = 0.42$ (zu Gunsten von Männern) hinsichtlich des geschlechtsspezifischen Leistungsunterschiedes bei Water-Level Tasks. Demnach lägen die Leistungen bei Water-Level Tasks von Männern und Frauen durchschnittlich 0.42 Standardabweichungen auseinander. Weiters liegt laut der genannten Autoren eine Verminderung der Geschlechtsdifferenz über die Zeit vor. Im Rahmen dieser Ergebnisse ist allerdings Folgendes zu berücksichtigen: Zunächst ist die Inaktualität des Befundes ins Treffen zu führen, sowohl aufgrund seines Publikationsdatums (1995) als auch der Tatsache, dass nur Studien bis zum Jahr 1993 in der Analyse berücksichtigt wurden. Weiters umfasst das Literaturkorpus zu Water-Level Tasks in dieser Untersuchung lediglich dreißig Stichproben (die überdies ausschließlich aus publizierten Studien stammen) – was in Relation zur vorliegenden Meta-Analyse als unvollständig bezeichnet werden kann. Drittens wurden drei der dreißig aufgenommen Studien fälschlicherweise mit Water-Level Tasks in Verbindung gebracht (nämlich: Olson & Eliot, 1986; Olson, Eliot & Hardy, 1988; Willemsen, 1974), und zusätzlich wurde eine Studie mit nicht korrekten Primärdaten inkludiert (nämlich: Willemsen & Reynolds, 1973), was das Literaturkorpus zu Water-Level Tasks in der Analyse von Voyer, Voyer und Bryden (1995) letztendlich auf sechszwanzig Studien reduziert.

Neben der Untersuchung von Voyer, Voyer und Bryden (1995), die als bisher einzige integrative Analyse einen globalen Geschlechtsunterschied in Water-Level Tasks in der Normpopulation thematisiert *und* quantifiziert (wobei die erwähnten Einschränkungen zu berücksichtigen sind), existieren außerdem drei narrative Reviewartikel zur Literatur der Water-Level Tasks (Kalichman, 1988; Liben, 1991b; Pascual-Leone & Morra, 1991), in deren Rahmen jeweils zwanzig bzw. 36 bzw. 32 Studien hinsichtlich des geschlechtsspezifischen Leistungsunterschiedes bei Water-Level Tasks berücksichtigt wurden. Übersichtsarbeiten dieser Art sind aufgrund der in Kapitel 2 erläuterten methodischen Gründe jedoch nicht mit den Resultaten einer Meta-Analyse gleichzusetzen.

2 Das Verfahren der Meta-Analyse

Das Verfahren der Meta-Analyse ist ein verhältnismäßig junges statistisches Instrument, gedacht für die Synthese mehrerer wissenschaftlicher Forschungsergebnisse. Im folgenden Kapitel wird ein kurzer Überblick über seinen Zweck, seine Anwendung und Funktionsweise sowie seine Abgrenzung zum narrativen Review gegeben.

2.1 Verbreitung und Zweck des Verfahrens

Der Begriff der Meta-Analyse wurde 1976 von Glass geprägt, allerdings reichen Untersuchungsmethoden, die mit diesem Verfahren zu vergleichen sind, bis um die Jahrhundertwende zurück (Cooper & Hedges, 1994).

Das Anwendungsfeld dieser Methode hat seinen Ursprung in der Psychologie und hat sich inzwischen auf eine große Anzahl inhaltlich ganz und gar unterschiedlicher Fachrichtungen ausgedehnt (Schulze, 2007). Sowohl die absolute Anzahl als auch der relative Anteil von Meta-Analysen an allen Publikationen ist in den letzten dreißig Jahren stetig gestiegen (Schulze, 2007). Doch nicht nur meta-analytische Publikationen mehren sich. Die Anzahl von wissenschaftlichen Veröffentlichungen generell steigt rasant an und wird auch in Zukunft weiterhin steigen. Damit wird auch die Herausforderung für Forscher immer größer, sich einigermaßen zeitökonomisch einen angemessenen Überblick über die Ergebnisse eines beliebigen Forschungsthemas zu verschaffen, da sich die Befunde durch ihre wachsende Anzahl auch immer inkonsistenter darstellen. Eine Meta-Analyse kann die gebündelte Information aller bis zu einem bestimmten Zeitpunkt erzielten Forschungsergebnisse bezüglich eines bestimmten Themas vermitteln. Sie lässt

dabei bekannte methodische Einschränkungen narrativer Reviews hinter sich. Näheres hierzu erläutert das folgende Kapitel.

2.2 Anwendung und Funktionsweise

Das Verfahren der Metaanalyse erlaubt einen Blick auf eine bestimmte wissenschaftliche Fragestellung von einer höheren Ebene aus: nicht etwa Testpersonen formen die in Frage stehende Population, deren Testwerte zusammengefasst werden, sondern bereits durchgeführte Studien stellen die zu analysierenden Untersuchungseinheiten dar. Anstelle der Befragung bzw. Beobachtung bzw. Testung der einzelnen Testpersonen steht hier die „Befragung“ der einzelnen thematisch relevanten Primärstudien: Interessierende Charakteristika und Ergebnisse dieser Studien werden exzerpiert, aggregiert und einer gemeinsamen Analyse zugänglich gemacht. Auf diese Weise kann der aktuelle Stand der Forschung auf einem bestimmten Gebiet integrativ dargestellt werden.

2.2.1 Ablauf einer meta-analytischen Untersuchung

Die wichtigsten Phasen einer solchen Analyse lassen sich wie folgt zusammenfassen (vgl. Rustenbach, 2003):

Eingrenzung des Forschungsgegenstandes, systematische und möglichst erschöpfende Literaturrecherche, Formulierung von In- und Exklusionskriterien für recherchierte Primärliteratur, Datenextraktion sowie Kodierung inhaltlich und methodisch relevanter Studienmerkmale, Durchführung der statistischen Analyse (d.h. Integration der Primärbefunde sowie Ermittlung eventuell intervenierender Variablen), Interpretation der Ergebnisse (unter Berücksichtigung einer eventuellen Publikationsverzerrung), Darstellung der Ergebnisse.

Rücklaufschleifen und Wiederholungen einiger Phasen, zum Beispiel bei notwendiger Überarbeitung von In- und Exklusionskriterien, sind nicht unüblich und können den gesamten Untersuchungsablauf maßgeblich verlängern.

Generell zählt eine derartige Untersuchung zu den aufwändigsten statistischen Analysen und erstreckt sich üblicherweise – je nach Zugänglichkeit und Umfang des verfügbaren Literaturkorpus zum in Frage stehenden Effekt – über eine längere Zeitspanne. Weiters ist zur Durchführung einer Meta-Analyse ausführliches Wissen hinsichtlich der in diesem Rahmen angewandten statistischen Methoden unbedingte Voraussetzung.

2.2.2 Ermittlung eines mittleren Gesamteffekts

Die durch eine möglichst erschöpfende Literaturerhebung recherchierten relevanten Ergebnisse und ermittelten Effekte aus jeder in die Analyse aufgenommenen Studie werden zu einem mittleren (und durch geeignete Gewichtung repräsentativen) Gesamtergebnis zusammengefasst – der gewichteten mittleren Effektgröße.

Es existieren verschiedene Arten von Effektgrößen, die sich alle ineinander überführen lassen. Ihr Einsatz hängt stark von der Fragestellung ab. Die Fragestellung der vorliegenden Meta-Analyse bezieht sich auf eine geschlechtsspezifische Leistungsdifferenz. Für Unterschiedshypothesen dieser Art wird die Effektstärke d angewandt. Sie ist auch bekannt als Cohen's d oder standardisierte Differenz der Mittelwerte. Ihre Berechnung zeigt Formel 1.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{gepoolt}}} \quad (1)$$

In der vorliegenden Meta-Analyse entspricht $\bar{x}_1$ bzw. $\bar{x}_2$ jeweils der durchschnittlichen Testleistung der männlichen bzw. weiblichen Testpersonen. Die Effektstärke d ergibt sich aus der Differenz der beiden Mittelwerte gebrochen durch die gepoolte Standardabweichung dieser beiden Untersuchungsgruppen.

Der Betrag von d ist ein standardisiertes Maß für die Größe der Mittelwertsdifferenz zwischen zwei Gruppen, oder anders formuliert, für die Stärke des Effekts der Gruppenzugehörigkeit (hier: des Geschlechts) auf die abhängige Variable (hier: die Testleistung). Bei einem d um einen Betrag von 0.20 (oder darunter) wird von einem kleinen Effekt gesprochen, ein d um 0.50 steht für einen mittleren Effekt, während man bei einem d um 0.80 (oder darüber) von einem großen Effekt spricht (Cohen, 1988; zitiert nach Rustenbach, 2003).

Die Richtung von d gibt an, welche Untersuchungsgruppe (hier: Männer oder Frauen) einen höheren Mittelwert erzielt hat (hier: eine bessere durchschnittliche Leistung erbracht hat). Die Interpretation hängt dabei maßgeblich davon ab, ob die Testleistung mittels Gut- oder Schlechtpunkten verrechnet wurde. In der vorliegenden Meta-Analyse bedeutet ein positives d eine bessere mittlere Testleistung von männlichen Testpersonen.

Wurde ein Effekt für jede aufgenommene Studie entweder auf diese Weise ermittelt, bereits von den Autoren berichtet oder aus einem anderen berichteten Resultat bzw. Effektmaß errechnet, wird nun ein mittlerer Gesamteffekt ermittelt. Die Repräsentativität des Gesamteffekts soll durch angemessene Gewichtung der einzelnen Studieneffekte sichergestellt werden. Abhängig vom zugrunde gelegten Modell bezüglich des Zustandekommens der gemessenen Studieneffekte wird eine von zwei möglichen Gewichtungsarten angewandt, was sich auf den resultierenden Gesamteffekt auswirkt. Nähere Erläuterungen hiezu gibt das folgende Kapitel.

2.2.2.1 Modell des festen Effekts vs. Modell zufallsvariabler Effekte

Beim Gesamteffekt, der im Zuge einer meta-analytischen Untersuchung errechnet werden soll, handelt es sich um einen gewichteten Mittelwert aller Einzeleffekte, die den berücksichtigten Studien entstammen. Bei einer ausreichend großen Anzahl von Studien ist davon auszugehen, dass sich deren Effekte voneinander unterscheiden – nicht alle Studien kommen zum selben Resultat. Dieser Umstand resultiert in einer gewissen Streuung der Einzeleffekte. (Genau diese Unterschiedlichkeit der Studienergebnisse ist der Grund für die Notwendigkeit einer Meta-Analyse.)

Im Modell des festen Effekts wird davon ausgegangen, dass alle Studien denselben „wahren“ Effekt in derselben Population zu messen beabsichtigten und ihre Resultate nur zufällig um diesen wahren Effekt variieren. Der Mittelwert all dieser zufällig variierenden Einzeleffekte kann den zugrunde liegenden wahren Effekt angemessen schätzen. Diese Annahme liegt dem Modell des festen Effekts zugrunde. In diesem Modell liefern alle Studien Information für ihren gemeinsamen wahren Effekt und können entsprechend der jeweiligen Güte ihrer Information gewichtet werden: Je präziser eine Studie ihr Resultat zu messen vermag, desto größer ist ihr Gewicht, mit dem sie ihre Information zur Schätzung des wahren Effekts einfließen lassen kann. Die Gewichtung richtet sich demnach in diesem Modell nach der Messgenauigkeit jeder Studie.

Das Modell zufallsvariabler Effekte hingegen geht davon aus, dass nicht alle Studien denselben wahren Effekt messen, sondern dass jeder Studie ein anderer wahrer Effekt aus einer eigenen Population zugrunde liegt. Der meta-analytisch berechnete Gesamteffekt stellt daher in diesem Modell nicht die Schätzung eines einzelnen wahren Effekts dar, sondern den Mittelwert mehrerer wahrer Effekte. Da in diesem Modell jede Studie wichtige Information für einen *eigenen* wahren Effekt

liefert, müssen die Gewichtungen, mit denen diese in den mittleren Gesamteffekt einfließen, sehr viel ausgewogener sein, als im Modell des festen Effekts.

Große Heterogenität zwischen den Einzeleffekten der Primärstudien wird oftmals als Hinweis dafür gesehen, dass zufällig variierte Effekte wirksam sind, deren Einflüsse nicht kontrollierbar sind, weswegen das Modell der zufallsvariablen Effekte als adäquater als jenes des festen Effekts angenommen wird (Sutton, Abrams, Jones, Sheldon & Song, 2000).

2.2.3 Ermittlung intervenierender Variablen

Vorliegende Heterogenität der Primärbefunde ist nicht immer unerwünscht: Sie kann explizit genutzt werden, um mögliche intervenierende Faktoren zu ermitteln, die Einfluss auf den untersuchten Effekt haben. Hierzu werden Befundgruppen nach methodischen oder inhaltlichen Studienkriterien gebildet, um herauszufinden, ob auf eine Abhängigkeit des untersuchten Effekts vom jeweiligen Studienkriterium geschlossen werden kann. Aufgrund signifikanter Unterschiede zwischen den gebildeten Gruppen wird auf einen Einfluss der jeweiligen Gruppierungsvariablen auf die abhängige Variable (den Effekt) geschlossen. Derartige Erkenntnisse können dazu beitragen, Wissen über den untersuchten Effekt zum Beispiel hinsichtlich begünstigender Faktoren zu erweitern. Als gruppierende Kriterien eignen sich sowohl Eigenschaften der Primärstudien (z.B. Publikationsmedium), der Testpersonen (z.B. Ausbildungsrichtung), der Intervention (z.B. Antwortmodus) als auch unterschiedliche Rahmenbedingungen (z.B. Gruppen- versus Einzelsetting). Die Anwendung der sogenannten Subgruppenanalyse ermöglicht durch diese Art der Gruppierung Rückschlüsse dahingehend, welche intervenierenden Variablen den untersuchten Effekt beeinflussen.

Eine weitere Möglichkeit, systematische Einflüsse auf den untersuchten Effekt aufzudecken, stellt die Effektmoderatorenregression dar. Gegenüber der Subgruppenanalyse hat sie den Vorteil, dass ihr Einsatz nicht auf kategoriale Moderatoren beschränkt ist, sondern dass auch metrische Variablen berücksichtigt werden können. Auch diese Methode hat zum Ziel, herauszufinden, welche inhaltlichen oder methodischen Variablen den untersuchten Effekt moderieren. Eine spezielle Form der Durchführung kann einflussnehmende Variablen nach ihrer Wichtigkeit für die Vorhersage der abhängigen Variable ordnen.

Beide beschriebenen Verfahren wurden in der vorliegenden Meta-Analyse kombiniert, um wirksame Moderatoren zu ermitteln.

2.2.4 Die Publikationsverzerrung

Ist auf meta-analytische Art und Weise ein Gesamteffekt errechnet, gilt es, bei dessen Interpretation eine mögliche Verzerrung durch einen gewissen Publikationstrend zu berücksichtigen: Rothstein, Sutton und Borenstein (2005) beschreiben die Publikationsverzerrung (auch bekannt als Publikationsbias) als „systematische Unrepräsentativität“ publizierter Studien hinsichtlich der Menge durchgeführter Studien überhaupt. Es ist bekannt und belegt, dass Untersuchungsergebnisse, die statistisch signifikante und/oder starke Effekte berichten, nicht nur mit größerer Wahrscheinlichkeit zur Publikation eingereicht werden, sondern auch größere Chancen auf tatsächliche, raschere und prominentere Veröffentlichung haben als Studienresultate, die statistisch nicht signifikante und/oder schwache Effekte berichten (Easterbrook, Berlin, Gopalan & Matthews, 1991; zitiert nach Sutton, Abrams, Jones, Sheldon & Song, 2000). Mit diesem Wissen wird die Gefahr offensichtlich, die auch bei der Interpretation eines meta-analytisch ermittelten Gesamteffekts besteht: Dass ein derartiges Resultat repräsentativ ist für die Population aller auf einem bestimmten Gebiet jemals getätigten Untersuchungen, könnte aus den folgendem Gründen in Frage gestellt werden. Wegen der oben erwähnten systematischen Zurückhaltung von nicht signifikant ausgefallenen Untersuchungen und Untersuchungen, die lediglich schwache Effekte nachweisen konnten, bleiben derartige Resultate in einer Meta-Analyse eher unberücksichtigt. In das Literaturkorpus finden mit größerer Wahrscheinlichkeit besser zugängliche, also vor allem publizierte Untersuchungen Eingang, die wie erwähnt tendenziell stärkere Effekte berichten als nicht publizierte Studien. Folgerichtig kann angenommen werden, dass der in einer Meta-Analyse ermittelte Gesamteffekt (der meist auf publizierten Einzeleffekten gründet), den wahren Gesamteffekt (der auch nicht publizierte und schwache Effekte einbeziehen würde) überschätzt.

Zur Abschätzung einer derartigen Verzerrung des Gesamteffekts stehen mehrere Verfahren zur Verfügung, von denen die folgenden in der vorliegenden Meta-Analyse zur Anwendung kamen: Funnel Plot, Trim-and-Fill-Methode, Rosenthals Failsafe N, Orwins Failsafe N, Eggers Regressionsmethode sowie Begg

und Mazumders Rang-Korrelations-Methode. Sie werden in Kapitel 5.2.1 anhand ihrer Resultate erläutert.

2.3 Vorteile und Kritikpunkte

2.3.1 Vorteile gegenüber narrativen Reviews

Narrative Reviews oder Übersichtsarbeiten haben ebenfalls zum Ziel, einen adäquaten Überblick über ein in Frage stehendes Forschungsgebiet zu geben und sollen, bei korrekter Durchführung, das Studium der darin vorwiegend qualitativ beschriebenen Einzelstudien vollkommen ersetzen (Glass, McGaw & Smith, 1981; zitiert nach Rustenbach, 2003). Es existiert jedoch ernst zu nehmende Kritik an der methodischen Vorgangsweise des traditionellen narrativen Review im Vergleich zum moderneren Verfahren der Meta-Analyse. Rustenbach (2003) gibt einen zusammenfassenden Überblick der Vorteile der Meta-Analyse gegenüber dem narrativen Review:

Meta-analytische Untersuchungen zeichnen sich zunächst und vor allem durch stärkere Objektivität aus als dies ein narrativer Review jemals bewerkstelligen könnte. Schon die subjektive Auswahl relevanter Studien im Review wird bei der Meta-Analyse durch eine möglichst erschöpfende Literaturrecherche ersetzt, womit die Repräsentativität der Analyseresultate für den gesamten Forschungsbereich gewährleistet werden kann. Die Integration kontradiktorischer und inkonsistenter Studienresultate stellt für die Meta-Analyse kein Problem dar – „problematisch“ zu integrierende Untersuchungen müssen nicht von der Analyse ausgeschlossen werden. Ein weiterer massiver Vorteil der Meta-Analyse besteht in der angemessenen Berücksichtigung des Stichprobenfehlers. Ebenso werden insignifikante Ergebnisse nicht unter- und signifikante Resultate nicht überschätzt – wie dies in narrativen Reviews häufig sehr wohl der Fall ist. Eine objektive, quantitative Analyse von Charakteristiken der Primärstudien im Zuge einer Meta-Analyse steht einer lediglich qualitativen Beschreibung im narrativen Review gegenüber. Methodische und inhaltliche Systematik des Vorgehens ist bei meta-analytischen Untersuchungen belegt und nachvollziehbar. Die quantitativen Schlüsse sind statistisch abgesichert, darin eingeschlossen sind eindeutige Aussagen über die Stärke von Effekten sowie eine größere Teststärke.

Hinsichtlich dieser Aspekte ist das Verfahren der Meta-Analyse vom methodischen Standpunkt aus dem des narrativen Review vorzuziehen.

2.3.2 Kritikpunkte

Häufig angeführte Kritikpunkte betreffend die Meta-Analyse, die es bei der Durchführung jeglicher meta-analytischer Vorhaben zu berücksichtigen gilt, sind im Folgenden erläutert.

2.3.2.1 Garbage-In-Garbage-Out-Problem

Als Garbage-In-Garbage-Out-Problem formuliert Hunt (1997; zitiert nach Rosenthal & DiMatteo, 2001) eine Widerspiegelung der Güte der inkludierten Primärstudien in der Güte des Resultats einer Meta-Analyse. Ein häufig angewandtes Vorgehen ist eine Gewichtung aller inkludierten Studien, wodurch die Stärke des Einflusses jeder Studie auf den Gesamteffekt kontrolliert werden kann. Laut Rosenthal und DiMatteo (2001) kann so für Studien von eingeschränkter methodischer Güte deren Einfluss zu Gunsten von methodisch einwandfreien Studien reduziert werden.

2.3.2.2 Äpfel-und-Birnen-Problem

Hinter dem Äpfel-und-Birnen-Problem (Hunt, 1997; zitiert nach Rosenthal & DiMatteo, 2001) verbirgt sich die Kritik an der variierenden Art der Operationalisierung des in Frage stehenden Konstrukts zwischen den inkludierten Studien: Die Sinnhaftigkeit einer Synthese von Studienergebnissen, die jeweils durch eine andere Art der Operationalisierung von zum Beispiel verbalen Fähigkeiten zustande gekommen sind, wird dabei angezweifelt. Anders formuliert, die Zusammenfassung von „Äpfeln und Birnen und der gelegentlichen Zitrone“ wird als nicht sinnvoll angesehen (Furberg & Morgan, 1987; zitiert nach Sutton, Abrams, Jones, Sheldon & Song, 2000). Gegenstimmen allerdings vertreten die Auffassung, dass heterogene Studien sehr wohl zu einem sinnvoll interpretierbaren Gesamteffekt integrierbar sind (Rushton, Brainerd & Pressley, 1983). Eine derartige Vorgangsweise könnte sogar der Validität des untersuchten Effekts zuträglich sein: Erweist sich ein Effekt, der unter unterschiedlichen und stark variierten Bedingungen mehrfach getestet wurde als konstant, so ist dieser als besonders robust anzusehen und gilt als generalisierbar (Campbell & Fiske, 1959; zitiert nach Hall, Tickle-Degnen, Rosenthal & Mosteller, 1994).

2.3.2.3 Abhängige Messungen

Die Synthese abhängiger Effekte zum mittleren Gesamteffekt hat eine Einschränkung der Validität zur Folge (Rustenbach, 2003). Abhängigkeit der Messungen liegt unter anderem vor bei Mehrfachtestung derselben Stichprobe: Berichtet eine Studie Ergebnisse wiederholter (relevanter) Messungen an derselben Stichprobe oder auch (bei nicht überliefertem Gesamtscore) Ergebnisse verschiedener Subskalen, die demselben Testverfahren angehören, so ist die Zusammenfassung dieser abhängigen Werte durch Mittelung eine adäquate Vorgangsweise (Rosenthal & DiMatteo, 2001).

Abhängigkeit liegt außerdem vor bei mehrfacher Publikation von Befunden derselben Stichprobe. Werden identische Ergebnisse derselben Stichprobe in unterschiedlichen Veröffentlichungen berichtet, sind derartige Duplikate zu exkludieren (Rustenbach, 2003).

2.3.2.4 Schubladenproblem

Das Schubladenproblem, das synonym mit dem Begriff der Publikationsverzerrung ist, wurde bereits in Kapitel 2.2.4 behandelt. Die Bezeichnung Schubladenproblem soll verbildlichen, dass eine systematische Auswahl durchgeführter Studien (nämlich jene mit insignifikanten Ergebnissen und schwachen Effekten) „in der Schublade“ verbleibt anstatt publiziert zu werden und so zur Verzerrung des errechneten Gesamteffekts führt, da letzterer in der Schublade zurückgehaltene Studien nicht berücksichtigen kann.

2.3.2.5 Unzureichende primäre Ergebnisberichte

Rosenthal und DiMatteo (2001) nennen eine mögliche zusätzliche Quelle der Verzerrung des Gesamteffekts: Das Problem unzureichend berichteter Angaben für die Berechnung einer Effektgröße. Primärstudien, die weder ihr Resultat in Form einer Effektgröße berichten, noch andere Angaben machen, aus denen eine Effektgröße nachträglich abgeleitet werden könnte, verhindern damit ihre Berücksichtigung in jeder meta-analytischen Untersuchung. Nur in einzelnen Fällen sind durch eine direkte Anfrage bei den Studienautoren die fehlenden Angaben in Erfahrung zu bringen, da Autoren älterer und besonders alter Studien (die am häufigsten von genanntem Manko betroffen sind) oft nicht mehr erreichbar sind. Inzwischen werden Forderungen nach dem Bericht von Studienergebnissen in einer Form, die deren Berücksichtigung in nachfolgenden meta-analytischen

Untersuchungen ermöglicht, immer lauter. In weiterer Zukunft darf daher eine Verringerung des prozentualen Anteils wissenschaftlicher Publikationen mit unzureichenden Ergebnisberichten erwartet werden.

DURCHFÜHRUNG DER META-ANALYSE

3 Fragestellungen

Alle Fragestellungen der vorliegenden Meta-Analyse beziehen sich auf den Geschlechtseffekt hinsichtlich der durchschnittlichen Leistung in Water-Level Tasks.

Die Analyse kann in zwei Abschnitte geteilt werden:

Einerseits wird die Frage nach dem konkreten Ausmaß des Geschlechtseffekts hinsichtlich der durchschnittlichen Leistung in Water-Level Tasks gestellt. Dies kann in Form eines mittleren Gesamteffekts beantwortet werden, der im Zuge einer integrativen Analyse (unter Einbeziehung aller verfügbaren relevanten Primärquellen) ermittelt wird. Anschließend sollen außerdem Existenz und Ausmaß einer etwaigen Publikationsverzerrung des ermittelten Gesamteffekts festgestellt werden (wofür mehrere Methoden zur Anwendung kamen, siehe Kapitel 5.2.1).

Die andere Frage betrifft Variablen, die einen Einfluss auf den Geschlechtseffekt haben könnten. Dies kann in Form einer Moderatoranalyse beantwortet werden, die auf zwei einander ergänzende Arten durchgeführt wurde (siehe Kapitel 5.3). Als potentiell einflussnehmende Variablen wurden sowohl Charakteristika der Studien selbst berücksichtigt als auch Eigenschaften der Testpersonen, der Testaufgaben, der Testauswertung sowie der Rahmenbedingungen der Testdurchführung. Insgesamt wurden achtzehn potentiell intervenierende Variablen hinsichtlich ihres Einflusses auf den Geschlechtseffekt untersucht, was in diesem Umfang bisher noch in keiner anderen Untersuchung zu diesem Thema durchgeführt wurde. Zwei darin enthaltene Fragestellungen sind an dieser Stelle besonders hervorzuheben: die Veränderung des Geschlechtseffekts über die Lebenszeit sowie seine langfristige Veränderung innerhalb der letzten fünf Jahrzehnte.

4 Methoden

4.1 Systematik der Literatursuche

Das umfangreiche Literaturkorpus der vorliegenden Meta-Analyse wurde schrittweise recherchiert: Der Großteil der Literatur wurde vom Arbeitsbereich Methodenlehre am Institut für Psychologische Grundlagenforschung für dieses Projekt zur Verfügung gestellt. Die beiden rezenten Recherchen in weiterer Folge fanden im Abstand von sechs Monaten im April und November 2007 statt. Eine zusätzliche relevante und aktuelle Studie aus 2008 wurde nachträglich in den Datensatz integriert und die gesamte Analyse wiederholt. Die seit Abschluss der Analyse erschienenen relevanten Quellen finden sich in Appendix A

Zur Identifizierung relevanter Primärstudien wurden elektronische Literaturdatenbanken herangezogen (ERIC, Medline, PSYINDEX, PsycInfo, PubMed, Sociofile, UMI, Web of Science), um Schlagwortsuchen durchzuführen. Der Katalog der Schlagworte konnte zu jedem Recherchezeitpunkt um zusätzliche Schlagworte erweitert werden, die in bereits identifizierten relevanten Primärquellen in engem Zusammenhang mit Water-Level Tasks verwendet wurden. Der zusammengestellte Katalog enthält die folgenden 38 Ausdrücke (in alphabetischer Reihenfolge):

Concept/s of horizontality, conception of horizontality, conservation of horizontality, draw water level/s, draw water line/s, draw water surface/s, horizontal task/s, horizontality item/s, horizontality of liquid/s, horizontality representation, horizontality task/s, invariance of liquid/s, liquid horizontality, perception of horizontality, Piagetian water task, Piagetian water test, representation of horizontality, Thomas apparatus, Wasserspiegelaufgabe/n, Wasserstandsaufgabe/n, water bottle task, water glass task, water glass test, water level assessment, water

level conceptualization, water level drawing task/s, water level problem/s, water level representation, water level task/s, water level test, water line representation, water line task/s, water line test, water line problem/s, water-level problem/s, water-level task/s, water-level test, WLT;

Der für diesen Katalog resultierende Suchstring findet sich in Appendix B.

Weitere relevante Quellen wurden außerdem mit Hilfe der Referenzlisten bereits identifizierter relevanter Studien ermittelt.

4.2 Datenextraktion und Kodierung

Zu berücksichtigende Daten wurden aus allen in Frage kommenden Primärquellen nach einem festgelegten Schema extrahiert und kodiert (Kodierschema siehe Appendix C). Natürlich eignen sich Variablen nur dann zur Berücksichtigung in einer Analyse, wenn deren Ausprägung in möglichst allen relevanten Studien berichtet wird oder aus anderen berichteten Angaben schätzbar ist. Da dies nicht bei allen zunächst berücksichtigten Variablen konsistent der Fall war, konnten letztendlich nicht alle extrahierten Daten in die Analyse aufgenommen werden.

Außer den bibliographischen Angaben der Untersuchungen, den jeweiligen Ergebnissen der Messungen zu Water-Level Tasks und den jeweiligen Stichprobengrößen (nach Geschlecht getrennt), wurden folgende Variablen im Weiteren erhoben (metrisch bzw. kategoriell in mindestens zwei Ausprägungen): Charakteristika der Studien selbst (Publikationsjahr, Publikationsstatus, Geschlecht des Erstautors), Eigenschaften der Testpersonen (Geburtsjahr, Nationalität, sozialer Status, Ausbildungsrichtung, Berufssparte), Eigenschaften der Testaufgaben (Itemanzahl, Anzahl von Musteritems ohne Feedback vor der Testung, Gefäßform, Vorgabemodus, Einbettung des Gefäßes in situativen Kontext, Verfügbarkeit eines realen Modells ohne sichtbaren Wasserspiegel, Verfügbarkeit einer horizontalen Referenz, Antwortmodus), Eigenschaften der Testauswertung (Auswertungskriterium zur Feststellung einer korrekten Aufgabenbeantwortung, Ergebnismodus) sowie Eigenschaften der Rahmenbedingungen der Testdurchführung (Gruppen-/Einzelsetting, Geschlecht des Testleiters, Instruktionsmodus, Speed-/Powertest).

In der Analyse nicht berücksichtigt werden konnten aufgrund der oben erwähnten zu großen Anzahl fehlender Werte folgende Eigenschaften der Testpersonen: Geburtsjahr, Nationalität, sozialer Status, Ausbildungsrichtung,

Berufssparte. Allerdings konnte anstatt des Geburtsjahres mit Hilfe anderer Angaben zur Stichprobe (z.B. Klassenstufe der Ausbildungseinrichtung) in allen Fällen die jeweilige Altersgruppe geschätzt werden, sodass dennoch ein Maß für das Alter der Testpersonen in die Analyse einfließen konnte. Die Eigenschaft Nationalität wurde zwar konsistent berichtet, allerdings ergaben sich zu häufig Nationalitätszuordnungen, die in allen inkludierten Studien nicht öfter als einmal genannt wurden, sodass aufgrund der zu schwachen Besetzung dieser Zellen die Berücksichtigung dieser Eigenschaft nicht sinnvoll erschien. Tabelle 3 gibt einen Überblick über Gruppierungen dieser Variable.

Tabelle 3: Nationalität und jeweilige Anzahl k der inkludierten Stichproben

Nationalität	k
Australien	1
Canada	23
China	3
Deutschland	21
Ghana	4
Großbritannien	1
Malaysia	1
nicht angegeben	1
Niederlande	4
Norwegen	4
Österreich	5
USA	84
USA mit chinesischen Wurzeln	2

4.3 Ein- und Ausschlusskriterien für Primärstudien

Für die Auswahl relevanter Studien aus der Menge in Frage kommender Quellen wurde ein Katalog von Ein- und Ausschlusskriterien erstellt. Dies geschah ebenfalls schrittweise, da sich bei fortschreitender Lektüre der Primärstudien immer weitere mögliche relevante Variablen zeigten, während andere ausgeschlossen werden mussten, um einerseits die Vergleichbarkeit der Messungen sicherzustellen, ohne jedoch andererseits ein Übermaß an Homogenität der aufgenommenen Quellen (und damit Informationsverlust hinsichtlich intervenierender Variablen) zu riskieren.

Für die Inklusion einer Untersuchung mussten folgende Kriterien zutreffen:

- Sprache der Studie ist Englisch oder Deutsch.
- Verfügbarkeit der Studie ist im Volltext gegeben.
- Water-Level Tasks fanden Anwendung und
- deren Ergebnisse werden in einer Form berichtet, die die Berechnung eines Geschlechtseffekts zulassen, falls dieser nicht bereits aufscheint.

Um die Vergleichbarkeit der untersuchten Stichproben zu sichern gilt für diese:

- Sie sind nicht-klinisch,
- nicht minder- oder hochbegabt,
- heterosexuell (soweit bekannt),
- hinsichtlich ihrer Leistung in Water-Level Tasks nicht vorselektiert (z.B. keine Testpersonen, die einen Vortest nicht bestanden haben und deshalb ausgewählt wurden).

Für die angewandte Testform der Water-Level Tasks gilt:

- Testvorgabe erfolgt entweder mittels 2D-Zeichnung auf Papier oder am PC, als Fotografie des Gefäßes oder als reales Gefäß.
- Die Testantwort (die Angabe des Verlaufs des Wasserspiegels) wird entweder händisch eingezeichnet (Papier-Bleistift-Test), aus einer Reihe möglicher Antwortalternativen ausgewählt (Multiple-Choice-Test) oder händisch mit Hilfe eines anderen Modells des Wasserspiegels (Lineal, Wollfaden o. ä.) angepasst. Alle genannten Arten der Testbeantwortung müssen mit Sichtkontakt des Vorgabematerials geschehen (geschieht dies mit verbundenen Augen, wird diese Messung ausgeschlossen – betrifft die Gruppe der „haptischen“ Testungen).
- Angewandte Testgefäße sind in Flaschenform, dieser ähnlich oder damit vergleichbar (keine amorphen Objekte oder Objekte bestehend aus mehreren

miteinander verbundenen Hohlräumen, keine geschwungenen Schläuche oder Rohre).

- Die vorgegebene Referenzebene ist horizontal (in Form einer Tischplatte oder horizontalen Referenzlinie unter dem Gefäß, jedoch kein keilförmiger Untergrund, steil abfallender Hang o.ä., auf dem das Gefäß platziert ist).
- Die Testdimension ist ausschließlich die Horizontale (innerhalb derselben Testaufgabe wird z.B. nicht gleichzeitig auch das Konzept der vertikalen Dimension getestet).
- Die Testinstruktionen sind vergleichbar (z.B. werden keine Messungen, bei denen laut Instruktion „Wasser in Bewegung“ ist, inkludiert).
- Die gemessene Testleistung betrifft ausschließlich die *Produktion* eines Wasserspiegels aus dem Gedächtnis, es wurden keine Messungen der *Wahrnehmungsleistung* von Wasserlinien berücksichtigt. Während der Testung darf somit keine direkte Vorlage mit sichtbarer korrekter Wasserlinie vorhanden sein (ein reales Gefäß ohne sichtbare Wasserlinie hingegen ist während der Testung als Modell erlaubt).
- Übungseffekte sollten weitgehend vermieden werden, indem nur Messungen vor einer etwaigen Trainingsintervention berücksichtigt wurden (Prätest-Messungen), auch Messungen mit zwischengeschaltetem Feedback nach jeder Testaufgabe wurden ausgeschlossen.

Weiters wurden Studien ausgeschlossen, bei denen offensichtlich eine

- wiederholte Verwendung derselben Stichprobe vorlag wie in zumindest einer anderen Untersuchung (siehe Kapitel 2.3.2.3).

Nachdem alle ursprünglich 263 in Frage kommenden Primärquellen in Hinblick auf die genannten In- und Exklusionskriterien einer eingehenden Lektüre (größtenteils inklusive einer Exzerption) unterzogen worden waren, resultierten 86 Studien, die nach Maßgabe dieses Kataloges in die gegenwärtige Analyse aufzunehmen waren. Innerhalb dieser 86 Studien fanden sich insgesamt 154 relevante unabhängige Stichproben, die als Untersuchungseinheiten Eingang in die vorliegende Meta-Analyse fanden. Appendix D listet all jene 177 Untersuchungen auf, die gefunden und inspiziert wurden, aber nach Maßgabe der oben genannten Kriterien nicht in der Analyse berücksichtigt werden konnten. Appendix E enthält eine Übersicht und Beschreibung aller 154 berücksichtigten Stichproben.

4.4 Instrumente der statistischen Analyse

Alle Analysen zur Ermittlung des mittleren Gesamteffekts, zur Abschätzung einer Publikationsverzerrung sowie zur Ermittlung intervenierender Variablen mittels Subgruppenanalysen wurden mit dem Programm Comprehensive Meta Analysis (CMA), Version 2.2.030 durchgeführt. Für deskriptivstatistische Auswertungen sowie für die Effektmoderatorenregression wurde das Programm Statistical Package for the Social Sciences (SPSS), Version 12.0 eingesetzt.

5 Ergebnisse

5.1 *Beschreibung inkludierter Stichproben*

Mit Hilfe der in Kapitel 4.3 erläuterten Ein- und Ausschlusskriterien konnten aus ursprünglich 263 in Frage stehenden Quellen 86 Studien als relevant klassifiziert und in die Analyse aufgenommen werden. Innerhalb der 86 Studien fanden sich 154 relevante unabhängige Stichproben. Diese stellen die Untersuchungseinheiten der Meta-Analyse dar. Die vorliegende Untersuchung schöpft somit aus Information, die von insgesamt 25864 Testpersonen erhoben wurde. Appendix E enthält eine Übersicht und Beschreibung der 154 berücksichtigten Stichproben.

Das zusammengestellte Literaturkorpus deckt 45 Jahre, und damit die Gesamtheit der speziellen Forschungsgeschichte der Water-Level Tasks seit Bekanntwerdung des Geschlechtseffektes (siehe Kapitel 1.2) bis heute, ab: Inkludierte Studien wurden in den Jahren 1964 bis 2008 publiziert bzw. erstellt. Teilnehmer der aufgenommenen Untersuchungen wurden zwischen den Jahren 1901 und 1995 geboren, womit eine Spannweite von 94 Jahren abgedeckt ist. Die jüngste Testperson war zum Zeitpunkt der Messung im Alter von 7 Jahren, die älteste im Alter von 89 Jahren. Die Anteile der drei bzw. vier Altersgruppen innerhalb der inkludierten Stichproben zeigt Abbildung 3. Insgesamt $k = 29$ Stichproben (entspricht 18.83%) von den aufgenommenen 154 Stichproben bestanden aus Kindern. Als Kinder klassifiziert wurden Teilnehmer bis zu einem Alter von 11 Jahren bzw. vor dem Ende der Grundschulausbildung. $K = 21$ Stichproben (entspricht 13.64%) bestanden aus Jugendlichen, die als solche im Alter zwischen 12 und 18 Jahren bzw. vor dem Ende der Schulausbildung klassifiziert wurden. $K = 100$

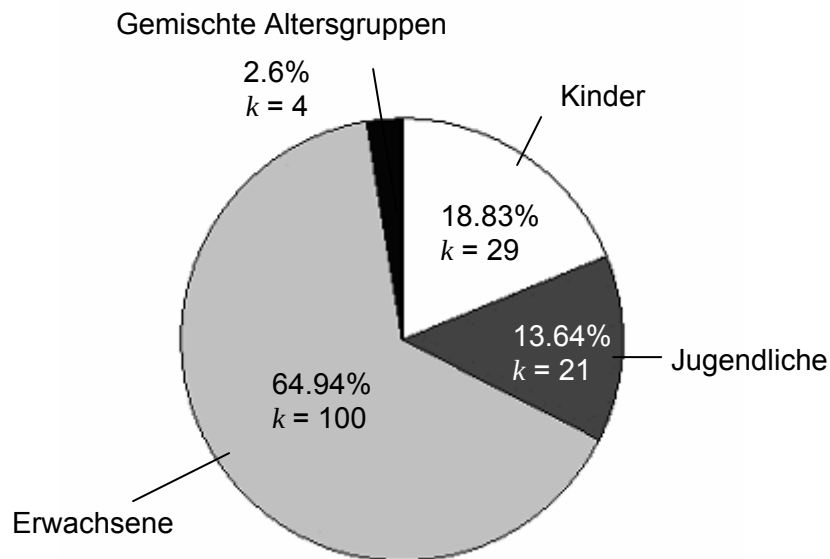


Abbildung 3: Anteile der Altersgruppen an der Menge aller inkludierten Stichproben.

Stichproben (oder 64.94%) bestanden aus Erwachsenen, das sind Teilnehmer ab einem Alter von 19 Jahren. Weiters wurde eine vierte Altersgruppe für jene Stichproben eröffnet, deren Messungen nicht getrennt nach den genannten drei Altersgruppen berichtet wurden, sondern sich mindestens über zwei verschiedene Altersgruppen erstreckten. Von derartigen Stichproben mit gemischten Altersgruppen gab es insgesamt $k = 4$ (entspricht 2.6%).

5.2 Einzeleffekte und resultierender Gesamteffekt

Die aus den Angaben der Primärstudien berechneten meta-analytischen Kennwerte für jede einzelne der inkludierten 154 Stichproben sind in Tabelle 4 aufgelistet (chronologisch nach Publikationsjahr geordnet): Neben Autor und Publikationsjahr der Studie sowie der Benennung der Stichprobe innerhalb der Studie findet sich der Geschlechtseffekt d jeder Stichprobe sowie dessen Standardabweichung s . Unter- und Obergrenze des Konfidenzintervalls, in dem sich der wahre Effekt der Population, aus der jede Stichprobe gezogen wurde, mit 95-prozentiger Wahrscheinlichkeit befindet, sind ebenfalls ablesbar. Die Wahrscheinlichkeit p gibt schließlich an, ob der beobachtete Einzeleffekt signifikant ausgefallen ist ($p < .05$) oder ob dieser auch die Ausprägung null annehmen könnte. Dies wird graphisch veranschaulicht durch den ebenfalls abgebildeten sogenannten Forest-Plot: Das Konfidenzintervall jedes Einzeleffektes ist hier durch eine horizontale Linie abgetragen, in dessen Mitte der beobachtete Einzeleffekt markiert

wurde. Je schmaler ein Intervall ausfällt, desto genauer wurde der Effekt gemessen. Schließt das Intervall die graue Nulllinie nicht mit ein, handelt es sich um einen signifikanten Effekt. Bei genauerer Durchsicht von Tabelle 4 finden sich unter allen 154 Einzeleffekten 94 signifikante Effekte und 60 nicht signifikante Effekte. Die Spannweite der beobachteten Effekte erstreckt sich von $d = -0.010$ bis $d = 1.837$, wobei ein negatives d in der vorliegenden Untersuchung für eine bessere Leistung weiblicher Testpersonen steht, während ein positives d eine bessere Leistung männlicher Testpersonen bedeutet. Insgesamt fanden sich in 154 Stichproben 152 positive Effekte, ein Effekt von $d = 0$, der für eine äquivalente Leistung von weiblichen und männlichen Teilnehmern steht, sowie lediglich ein negativer Effekt.

Tabelle 4: Meta-analytische Kennwerte mit Forest-Plot der inkludierten Stichproben (Ordnung chronologisch nach Publikationsjahr)

Studie	Stichprobe	d	s	Konfidenzintervall für d		p	Forest-Plot: d und Konfidenzintervall (95%)				
				Unter- grenze	Ober- grenze		-2,00	-1,00	0,00	1,00	2,00
Rebelsky (1964)	[whole sample]	0,148	0,202	-0,247	0,544	0,463			+		
Morris (1971)	group 1	0,442	0,253	-0,054	0,938	0,081			+		
Morris (1971)	group 2	1,101	0,268	0,575	1,626	0,000				+	
De Avila & Havassy (1974)	[whole sample]	0,026	0,058	-0,088	0,140	0,655			+		
Liben (1974)	group 1, "H"	0,510	0,288	-0,055	1,075	0,077			+		
Liben (1974)	group 2, "NH"	0,934	0,319	0,309	1,558	0,003				+	
Thomas & Jamison (1975)	experiment 3	0,736	0,179	0,385	1,087	0,000				+	
Geiringer (1976)	mature elderly	0,554	0,322	-0,077	1,186	0,085			+		
Geiringer (1976)	middle aged	0,624	0,264	0,105	1,142	0,018			+		
Geiringer (1976)	young elderly	0,465	0,320	-0,163	1,093	0,147			+		
Geiringer & Hyde (1976)	12th grade	0,424	0,261	-0,088	0,936	0,104			+		
Geiringer & Hyde (1976)	5th grade	0,512	0,262	-0,002	1,027	0,051			+		
Abravanel & Gingold (1977)	experiment 1	0,197	0,284	-0,359	0,754	0,487			+		
Abravanel & Gingold (1977)	experiment 2	0,416	0,286	-0,145	0,977	0,146			+		
Abravanel & Gingold (1977)	experiment 3	-0,010	0,366	-0,728	0,707	0,978					
Harris, Hanley & Best (1977)	experiment 1: adults	0,953	0,181	0,598	1,308	0,000				+	
Harris, Hanley & Best (1977)	experiment 1: children	0,812	0,267	0,288	1,335	0,002				+	
Harris, Hanley & Best (1977)	experiment 2	0,854	0,175	0,511	1,198	0,000				+	
Harris, Hanley & Best (1977)	experiment 3:	0,334	0,195	-0,049	0,717	0,087			+		
Harris, Hanley & Best (1977)	experiment 4	0,036	0,299	-0,551	0,622	0,905					
Harris, Hanley & Best (1977)	experiment 5	1,544	0,218	1,118	1,970	0,000				+	
Kelly J. & Kelly G. (1977)	[whole sample]	0,548	0,137	0,280	0,816	0,000				+	
McGillicuddy-DeLisi, DeLisi & Youniss (1978)	public college	0,979	0,473	0,051	1,907	0,039					
Walker & Krasnoff (1978)	[whole sample]	0,439	0,216	0,016	0,862	0,042			+		
McGiffert-Spradlin, Higgins-Hensley (1979)	[whole sample]	0,691	0,213	0,274	1,108	0,001				+	
Ray, Georgiou & Ravizza (1979)	[whole sample]	0,306	0,199	-0,085	0,697	0,125			+		
Treagust (1980)	[whole sample]	0,859	0,229	0,411	1,307	0,000				+	
Thompson, Mann & Harris (1981)	[whole sample]	0,522	0,194	0,142	0,901	0,007				+	
Walker, Krasnoff & Peaco (1981)	adults	0,152	0,176	-0,194	0,498	0,389			+		
Walker, Krasnoff & Peaco (1981)	children	0,668	0,181	0,314	1,023	0,000				+	
Krasnoff & Walker (1983)	[whole sample]	0,373	0,139	0,101	0,645	0,007				+	

Tabelle 4, Fortsetzung 1: Meta-analytische Kennwerte mit Forest-Plot der inkludierten Stichproben (Ordnung chronologisch nach Publikationsjahr)

Studie	Stichprobe	d	s	Konfidenzintervall für d		p	Forest-Plot: d und Konfidenzintervall (95%)				
				Unter- grenze	Ober- grenze		-2,00	-1,00	0,00	1,00	2,00
Linn & Pulos (1983)	[whole sample]	0,376	0,072	0,234	0,518	0,000			+		
Tegano & Faulkender (1983)	[whole sample]	0,808	0,202	0,412	1,204	0,000			+		
Kenyon (1984)	[whole sample]	1,383	0,406	0,586	2,180	0,001			+		
Liben & Golbeck (1984)	experiment 1: group "no rule, no example"	0,878	0,331	0,229	1,527	0,008			+		
Liben & Golbeck (1984)	experiment 1: group "no rule, with example"	0,631	0,324	-0,004	1,266	0,052			+		
Liben & Golbeck (1984)	experiment 2: "no rule" condition	0,854	0,234	0,396	1,312	0,000			+		
Mowrer-Popiel & DeLisi (1984)	[whole sample]	0,493	0,196	0,110	0,876	0,012			+		
Myer & Higgins-Hensley (1984)	[whole sample]	0,554	0,248	0,068	1,039	0,025			+		
Wittig & Allen (1984)	[whole sample]	0,573	0,150	0,278	0,868	0,000			+		
Golbeck (1986)	[whole sample]	0,721	0,258	0,215	1,227	0,005			+		
Kalichman (1986)	major: business	1,231	0,406	0,437	2,026	0,002			+		
Kalichman (1986)	major: engineering	1,670	0,823	0,058	3,282	0,042			+		
Kalichman (1986)	major: language arts	1,047	0,374	0,314	1,780	0,005			+		
Kalichman (1986)	major: natural sciences	1,123	0,465	0,212	2,034	0,016			+		
Kalichman (1986)	major: social sciences	1,271	0,365	0,555	1,987	0,000			+		
Liben & Golbeck (1986)	physical condition	0,697	0,326	0,058	1,335	0,032			+		
Meehan & Overton (1986)	experiment 1	0,531	0,222	0,096	0,966	0,017			+		
Meehan & Overton (1986)	experiment 2: controls	0,567	0,380	-0,177	1,311	0,135			+		
Murphy-Berman, Witters & Harding (1986)	vessel: curved bottle	1,191	1,447	-1,645	4,026	0,410			+		
Murphy-Berman, Witters & Harding (1986)	vessel: straight sided bottle	0,730	0,718	-0,677	2,136	0,309			+		
Sanders & Ross-Field (1986)	experiment 1: heterosexual group	1,714	0,585	0,568	2,860	0,003			+		
Otani & Leonard (1988)	experiment 1	0,440	0,196	0,056	0,824	0,025			+		
Krasnoff, Walker & Howard (1989)	[whole sample]	0,846	0,233	0,389	1,304	0,000			+		
Robert & Chaperon (1989)	"selectionphase"-sample	0,542	0,079	0,387	0,697	0,000			+		
Sholl (1989)	initial screening sample	0,446	0,153	0,145	0,747	0,004			+		
Signorella, Jamison & Krupa (1989)	[whole sample]	0,533	0,111	0,316	0,751	0,000			+		
Kalichman (1989a)	[whole sample]	0,788	0,169	0,456	1,120	0,000			+		
Kalichman (1989b)	[whole sample]	0,522	0,223	0,084	0,959	0,019			+		
Beatty & Duncan (1990)	[whole sample]	0,545	0,186	0,180	0,909	0,003			+		
Gladue, Beatty, Larson & Staton (1990)	heterosexual group	0,815	0,374	0,082	1,548	0,029			+		
Mowrer-Popiel (1990)	[whole sample]	0,305	0,170	-0,029	0,638	0,073			+		

Tabelle 4, Fortsetzung 2: Meta-analytische Kennwerte mit Forest-Plot der inkludierten Stichproben (Ordnung chronologisch nach Publikationsjahr)

Studie	Stichprobe	d	s	Konfidenzintervall für d		p	Forest-Plot: d und Konfidenzintervall (95%)				
				Unter- grenze	Ober- grenze		-2,00	-1,00	0,00	1,00	2,00
Robert (1990)	[whole sample]	0,733	0,205	0,331	1,135	0,000					
Robert & Tanguay (1990)	40 - 61 yrs	0,405	0,358	-0,296	1,107	0,258					
Robert & Tanguay (1990)	62 - 72 yrs	0,364	0,382	-0,385	1,112	0,341					
Robert & Tanguay (1990)	73 - 84 yrs	1,292	0,401	0,506	2,079	0,001					
Kalichman (1991)	[whole sample]	1,071	0,285	0,513	1,630	0,000					
Liben (1991)	experiment 1	0,490	0,199	0,100	0,881	0,014					
Liben (1991)	experiment 2	1,024	0,336	0,365	1,683	0,002					
McAfee & Proffitt (1991)	experiment 1	0,687	0,119	0,455	0,919	0,000					
McAfee & Proffitt (1991)	experiment 3	0,310	0,288	-0,254	0,874	0,282					
McCormick & Witelson (1991)	heterosexuals, consistent right-hand-preference	1,056	0,294	0,479	1,633	0,000					
Pennings (1991)	controls, pretest	0,941	0,527	-0,092	1,974	0,074					
Pennings (1991)	group "C", pretest	0,651	0,513	-0,355	1,656	0,205					
Pennings (1991)	group "R", pretest	0,741	0,517	-0,272	1,754	0,151					
Robert & Morin (1993)	"standard tasks"	0,516	0,211	0,102	0,929	0,014					
Thomas & Lohaus (1993)	experiment 1: 10 yrs	0,576	0,369	-0,147	1,298	0,119					
Thomas & Lohaus (1993)	experiment 1: 11 yrs	0,489	0,263	-0,027	1,005	0,063					
Thomas & Lohaus (1993)	experiment 1: 12 yrs	0,884	0,233	0,427	1,342	0,000					
Thomas & Lohaus (1993)	experiment 1: 13 yrs	0,422	0,216	-0,001	0,845	0,050					
Thomas & Lohaus (1993)	experiment 1: 14 yrs	0,258	0,264	-0,260	0,776	0,329					
Thomas & Lohaus (1993)	experiment 1: 15 yrs	0,279	0,249	-0,208	0,767	0,262					
Thomas & Lohaus (1993)	experiment 1: 16 & 17 yrs	0,055	0,290	-0,512	0,623	0,848					
Thomas & Lohaus (1993)	experiment 1: 6 & 7 yrs	0,610	0,332	-0,041	1,261	0,066					
Thomas & Lohaus (1993)	experiment 1: 8 yrs	0,531	0,275	-0,009	1,071	0,054					
Thomas & Lohaus (1993)	experiment 1: 9 yrs	0,203	0,346	-0,475	0,881	0,558					
Thomas & Lohaus (1993)	experiment 2: 11 & 12 yrs	0,425	0,339	-0,239	1,090	0,210					
Thomas & Lohaus (1993)	experiment 2: 13 & 14 yrs	0,539	0,284	-0,017	1,095	0,058					
Thomas & Lohaus (1993)	experiment 2: 15, 16 & 17 yrs	0,413	0,284	-0,143	0,969	0,145					
Thomas & Lohaus (1993)	experiment 2: 9 & 10 yrs	0,796	0,367	0,077	1,516	0,030					
Vasta, Lightfoot & Cox (1993)	"physikal task"	0,652	0,381	-0,095	1,398	0,087					
Lawton (1994)	[whole sample]	0,807	0,107	0,597	1,017	0,000					
Robert & Ohlmann (1994)	control condition	0,721	0,123	0,479	0,962	0,000					

Tabelle 4, Fortsetzung 3: Meta-analytische Kennwerte mit Forest-Plot der inkludierten Stichproben (Ordnung chronologisch nach Publikationsjahr)

Studie	Stichprobe	d	s	Konfidenzintervall für d		p	Forest-Plot: d und Konfidenzintervall (95%)				
				Unter- grenze	Ober- grenze		-2,00	-1,00	0,00	1,00	2,00
Robert, Pelletier, St-Onge & Berthiaume (1994)	wlt condition, "visual standard" group	1,174	0,239	0,704	1,643	0,000				+	
Hammer, Hoffer & King (1995)	major: architectural	0,000	0,305	-0,599	0,599	1,000					
Hammer, Hoffer & King (1995)	major: liberal arts	1,837	0,601	0,660	3,014	0,002					+
Hecht & Proffitt (1995)	graduate students	0,423	0,402	-0,365	1,211	0,293					
Hecht & Proffitt (1995)	group "experienced with liquids"	0,382	0,380	-0,363	1,127	0,315					
Hecht & Proffitt (1995)	group "non-experienced with liquids"	1,073	0,392	0,305	1,841	0,006					+
Parameswaran (1995)	pretest	0,413	0,240	-0,058	0,884	0,086					
Sholl & Liben (1995)	experiment 2: mass screening sample	0,693	0,190	0,320	1,066	0,000					
Sholl & Liben (1995)	experiment 3: mass screening sample	0,536	0,100	0,341	0,731	0,000					
Sholl & Liben (1995)	experiment 4: mass screening sample	0,729	0,079	0,574	0,883	0,000					
Wegesin (1995)	[whole sample]	0,731	0,243	0,255	1,207	0,003					
Lohaus, Thomas, Kessler & Gediga (1996)	grade 2	0,202	0,141	-0,074	0,479	0,152					
Lohaus, Thomas, Kessler & Gediga (1996)	grade 5	0,470	0,146	0,184	0,756	0,001					
Lohaus, Thomas, Kessler & Gediga (1996)	grade 8	0,794	0,147	0,507	1,082	0,000					
Parameswaran & DeLisi (1996)	[whole sample]	0,413	0,240	-0,058	0,884	0,086					
Robert & Harel (1996)	major: humanities & arts; vessel: bottles	0,657	0,100	0,461	0,852	0,000					
Robert & Harel (1996)	major: natural & applied sciences; vessel: bottles	0,671	0,091	0,494	0,849	0,000					
Robert & Harel (1996)	major: social sciences; vessel: bottles	0,502	0,095	0,316	0,688	0,000					
Vasta, Knott & Gaze (1996)	[whole sample]	0,016	0,149	-0,277	0,309	0,915					
Voyer (1996)	experiment 3	0,634	0,185	0,271	0,996	0,001					
Merriwether & Liben (1997)	[whole sample]	0,463	0,160	0,149	0,777	0,004					
Pulos (1997)	"high-science" group	0,256	0,230	-0,194	0,706	0,265					
Pulos (1997)	"low-science" group	0,324	0,231	-0,128	0,777	0,160					
Dessens et al. (1998)	controls	0,629	0,172	0,292	0,966	0,000					
Wegesin (1998)	heterosexual group	0,727	0,326	0,087	1,367	0,026					
Li, Nuttall & Zhao (1999a)	[whole sample]	0,281	0,091	0,102	0,460	0,002					
Li, Nuttall & Zhao (1999b)	group "BC" (Beijing, Chinese)	0,745	0,125	0,500	0,990	0,000					
Li, Nuttall & Zhao (1999b)	group "CCA" (Writing Chinese, US)	0,103	0,289	-0,463	0,669	0,721					
Li, Nuttall & Zhao (1999b)	group "NCA" (Non-writing Chinese, US)	0,024	0,177	-0,322	0,370	0,892					
Amponsah (2000)	Ghana, 12 yrs	0,463	0,125	0,218	0,708	0,000					
Amponsah (2000)	Ghana, 16 yrs	0,475	0,118	0,243	0,706	0,000					

Tabelle 4, Fortsetzung 4: Meta-analytische Kennwerte mit Forest-Plot der inkludierten Stichproben (Ordnung chronologisch nach Publikationsjahr)

Studie	Stichprobe	d	s	Konfidenzintervall für d		p	Forest-Plot: d und Konfidenzintervall (95%)				
				Unter- grenze	Ober- grenze		-2,00	-1,00	0,00	1,00	2,00
Amponsah (2000)	Ghana, 9 yrs	0,335	0,128	0,085	0,586	0,009			+		
Amponsah (2000)	Ghana, adults	0,631	0,104	0,427	0,834	0,000			+		
Amponsah (2000)	Norway, 12 yrs	0,690	0,120	0,456	0,924	0,000			+		
Amponsah (2000)	Norway, 16 yrs	0,790	0,119	0,558	1,022	0,000			+		
Amponsah (2000)	Norway, 9 yrs	0,327	0,145	0,042	0,611	0,024			+		
Amponsah (2000)	Norway, adults	0,594	0,103	0,393	0,796	0,000			+		
Golbeck & Sinagra (2000)	group "alone"	0,109	0,469	-0,810	1,027	0,816			+		
Voyer (2000)	[whole sample]	0,571	0,111	0,353	0,789	0,000			+		
Beisteiner (2001)	[whole sample]	0,277	0,072	0,136	0,417	0,000			+		
Casey, Nuttall & Pezaris (2001)	[whole sample]	0,670	0,150	0,375	0,964	0,000			+		
Quaiser-Pohl (2001)	[whole sample]	0,394	0,096	0,205	0,583	0,000			+		
Roberts & Bell (2001)	[whole sample]	0,606	0,256	0,105	1,107	0,018			+		
Levine & Stern (2002)	[whole sample]	0,616	0,226	0,174	1,058	0,006			+		
Seng & Tan (2002)	Chinese	0,350	0,268	-0,175	0,875	0,192			+		
Seng & Tan (2002)	Malays	0,347	0,334	-0,308	1,002	0,300			+		
Seng & Tan (2002)	nationality: "Others"	1,006	0,967	-0,889	2,900	0,298			+		
Formann (2003)	[whole sample]	0,176	0,108	-0,035	0,388	0,102			+		
Loehlin & McFadden (2003)	heterosexual group	0,322	0,186	-0,043	0,687	0,084			+		
Voyer & Sullivan (2003)	experiment 2: timed condition, group testing	0,782	0,179	0,431	1,133	0,000			+		
Voyer & Sullivan (2003)	experiment 2: timed condition, individual testing	0,453	0,188	0,085	0,820	0,016			+		
Voyer & Sullivan (2003)	experiment 2: untimed condition, group testing	0,582	0,147	0,294	0,869	0,000			+		
Voyer & Sullivan (2003)	experiment 2: untimed condition, individual	1,155	0,314	0,540	1,770	0,000			+		
Robert & Héroux (2004)	12 yrs	0,210	0,314	-0,406	0,825	0,504			+		
Robert & Héroux (2004)	15 yrs	0,221	0,334	-0,434	0,877	0,508			+		
Robert & Héroux (2004)	9 yrs	0,128	0,330	-0,519	0,775	0,698			+		
Robert & Longpré (2005)	control condition	0,949	0,236	0,486	1,411	0,000			+		
Neumann, Fitzgerald, Furedy & Boyle (2007)	[whole sample]	0,606	0,256	0,105	1,107	0,018			+		
Tran & Formann (2008)	adolescents	0,006	0,116	-0,222	0,235	0,956			+		
Tran & Formann (2008)	adults	0,592	0,059	0,476	0,707	0,000			+		
Tran & Formann (2008)	children	0,243	0,166	-0,082	0,567	0,143			+		
Modell zufallsvariabler Effekte: Gesamteffekt		0,547	0,024	0,501	0,594	0,000			+		

Unter Berücksichtigung aller 154 Einzeleffekte resultierte ein gewichteter mittlerer Gesamteffekt von $d = 0.547$ ($s = 0.024$; $KI : [0.501; 0.594]$), der hochsignifikant ausfällt ($p < .001$) (siehe Tabelle 4, Fortsetzung 4). Aufgrund starker Heterogenität zwischen den Stichproben ($Q = 361.836$; $df(Q) = 153$; $p < .001$) wurde das Modell zufallsvariabler Effekte interpretiert (siehe Tabelle 5).

Tabelle 5: Meta-analytische Kennwerte des Gesamteffekts im Modell des festen Effekts bzw. zufallsvariabler Effekte sowie Angaben zur Heterogenität der Einzeleffekte.

Modell	Stichproben- anzahl k	Gesamteffekt und Konfidenzintervall (95%)					Heterogenität		
		d	s	Unter- grenze	Ober- grenze	p	Q	$df(Q)$	p
Fester Effekt	154	0.506	0.014	0.479	0.532	<.001	361.836	153	<.001
Zufalls- variable Effekte	154	0.547	0.024	0.501	0.594	<.001			

Ein Effekt von $d = 0.547$ ist als mittelstark zu interpretieren (siehe Kapitel 2.2.2) und bedeutet im vorliegenden Fall, dass die durchschnittliche Leistung bei Water-Level Tasks von männlichen und weiblichen Testpersonen im Mittel um 0.547 Standardabweichungen differiert, wobei männliche Teilnehmer weibliche übertreffen. Der mittlere Geschlechtseffekt bei Water-Level Tasks fällt somit um ein Drittel größer aus als zuletzt von Voyer, Voyer und Bryden (1995) angenommen. Ihre Resultate stützen die genannten Autoren allerdings auf einen Datensatz, der nur ein Sechstel des Umfangs der vorliegenden Meta-Analyse aufweist (siehe Kapitel 1.4).

5.2.1 Schätzung der Publikationsverzerrung

Um den ermittelten mittleren gewichteten Gesamteffekt auf eine mögliche Überschätzung durch Publikationsverzerrung (siehe Kapitel 2.2.4 bzw. 2.3.2.4) zu prüfen und im gegebenen Fall deren Ausmaß zu schätzen, wurden mehrere Verfahren durchgeführt, die anhand ihrer Resultate im Folgenden erläutert werden. Vor Behandlung der einzelnen Verfahren soll festgehalten werden: Es existiert kein statistischer Test, durch dessen resultierenden Signifikanzwert eindeutig auf das Vorliegen bzw. die Abwesenheit einer Publikationsverzerrung geschlossen werden könnte. Lediglich die Plausibilität einer vorliegenden Verzerrung kann anhand einiger Verfahren auf mehrere Arten beurteilt werden, um letztendlich zu einer plausiblen

Annahme hinsichtlich Vorliegen und Ausmaß einer eventuellen Verzerrung zu gelangen. Einen zusammenfassenden Überblick über die ermittelten Kennwerte aller durchgeführten Verfahren zur Schätzung des Publikations-Bias gibt Tabelle 6.

Verfahren zur Ermittlung eines sogenannten Failsafe N geben die Anzahl von Effekten vom Ausmaß $d = 0$ (sogenannte Nulleffekte) an, die neben den bereits berücksichtigten Effekten zusätzlich in die Analyse inkludiert werden müssten (aber der Hypothese nach aufgrund eines vorliegenden Publikations-Bias nicht erhoben werden können), um den resultierenden mittleren Gesamteffekt auf ein bestimmtes Ausmaß zu reduzieren. Je größer die Anzahl zusätzlich notwendiger Nulleffekte ausfällt, desto unwahrscheinlicher ist es, dass der beobachtete Gesamteffekt in Wahrheit lediglich ein Artefakt der Publikationsverzerrung ist. In der klassischen Berechnung nach Rosenthal (1979) wird die Anzahl zusätzlicher Nulleffekte ausgegeben, die notwendig wäre, um den errechneten Gesamteffekt insignifikant ($p > .05$) erscheinen zu lassen. Im vorliegenden Fall beläuft sich diese Zahl auf $N = 48386$ (siehe Tabelle 6). Aus dieser hohen Anzahl ist zu schließen, dass das Vorliegen eines Effekts, der in Wahrheit insignifikant ist aber sich aufgrund des Publikations-Bias als signifikant darstellt, äußerst unwahrscheinlich ist.

Um den vorliegenden Geschlechtseffekt auf $d = 0.10$ und damit auf ein Ausmaß, das als vernachlässigbar anzusehen ist, zu reduzieren, wären laut einer alternativen Berechnung des Failsafe N nach Orwin (1983), die dem Untersuchenden mehrere Nebenbedingungen individuell zu variieren erlaubt, $N = 625$ zusätzliche Null-Effekte notwendig.

Die erläuterten Berechnungen zum Failsafe N lassen darauf schließen, dass es sich beim vorliegenden Gesamteffekt tatsächlich um einen robusten Effekt handelt. Das folgende Verfahren nach Begg und Mazumdar (1994) prüft etwas differenzierter, ob das Vorliegen einer Publikationsverzerrung plausibel ist; diese kann auch auf andere Weise wirksam sein als in Form der oben erwähnten fälschlichen Signifikanz des beobachteten Gesamteffekts. Es wird eine Rangkorrelation (Kendall's Tau) zwischen standardisierten Effekten und deren Standardfehlern berechnet und auf Signifikanz geprüft. Unter verzerrungsfreien Umständen werden Effekte sowohl aus Untersuchungen mit großem als auch mit kleinem Stichprobenumfang ungeachtet ihrer Stärke berücksichtigt. (Große Stichprobenumfänge ermöglichen genauere Messungen, dies bedeutet weiterhin kleinere Standardfehler der beobachteten Effekte.) Im Fall einer wirksamen

Publikationsverzerrung hingegen werden kleine Stichproben (mit größerem Standardfehler) nur berücksichtigt, wenn sie einen starken Effekt vorweisen können. Ein starker Zusammenhang zwischen Effekt und Standardfehler würde resultieren. Im vorliegenden Fall resultiert ein *Tau* von .07, das nicht signifikant ($p = .089$) und letztendlich vernachlässigbar ist.

Der Regression-Intercept-Test von Egger (1997) mit einem ähnlichen Hintergrund wie bereits bei Begg und Mazumdar (1994) arbeitet mit dem Verfahren der Linearen Regression: Es wird die Regression eines Maßes für die Präzision der Messung auf den standardisierten Effekt berechnet. Da im unverzerrten Fall große standardisierte Effekte gemeinsam mit hoher Präzision (kleinem Standardfehler, großem Stichprobenumfang) und kleine standardisierte Effekte gemeinsam mit geringer Präzision (großem Standardfehler, kleinem Stichprobenumfang) erwartet werden, müsste die Regressionsgerade durch den Ursprung verlaufen, womit der Intercept gleich null wäre. Bei Vorliegen eines Publikations-Bias hingegen hängen Effektgröße und Präzision indirekt proportional zusammen, womit die Regressionsgerade kippen und nicht mehr durch den Ursprung verlaufen würde. Im vorliegenden Fall beträgt der signifikante ($p < .001$) Intercept 0.839.

Wie oben erwähnt berechnet Begg und Mazumdar (1994) Methode – die im vorliegenden Fall mit einem *Tau* von .07 äußerst niedrig und nicht signifikant ausfiel – eine Rangkorrelation und verfügt somit laut Sterne und Egger (2005) als parameterfreies Verfahren im Allgemeinen über eine geringere Testmacht als der parametrische Regression-Intercept-Test von Egger (1997). Der signifikante Befund des Regression-Intercept-Test von Egger (1997) weist mit einem ebenfalls kleinen Betrag der Teststatistik (einem Intercept von 0.839) darauf hin, dass lediglich von einer äußerst schwachen Publikationsverzerrung ausgegangen werden kann (Sterne & Egger, 2005). Als gemeinsames Resultat der beiden genannten Tests darf angenommen werden: Eine vorliegende Verzerrung ist derartig gering ausgeprägt, dass deren Entdeckung Testinstrumenten mit entsprechender Macht vorbehalten ist.

Die Trim-and-Fill-Methode nach Duval und Tweedie (2000) identifiziert zunächst jene Effekte, die bei vorausgesetzter zufälliger Streuung aller beobachteten Effekte um den gemeinsamen mittleren Gesamteffekt ebenfalls vorhanden sein müssten, jedoch unter den beobachteten Effekten fehlen. Unter der Hypothese einer vorliegenden Publikationsverzerrung betrifft dies vornehmlich Effekte in der Gegenrichtung aus eher kleineren Stichproben. Nach Identifizierung dieser

systematisch fehlenden Effekte wird der mittlere Gesamteffekt anschließend unter Einbeziehung dieser zusätzlichen (schwächeren) Effekte adjustiert. Im vorliegenden Fall werden im angenommenen Modell zufallsvariabler Effekte 23 zusätzliche Effekte addiert, wodurch der beobachtete mittlere Geschlechtseffekt von $d = 0.547$ auf 0.496 adjustiert wird.

Diese Prozedur wird graphisch veranschaulicht durch den sogenannten Funnel-Plot (Abbildung 4), dessen Name von seiner trichterähnlichen Form rührt: Das Präzisionsmaß (der inverse Standardfehler) pro Studie wird auf der y-Achse gegen den jeweiligen beobachteten Einzeleffekt auf der x-Achse aufgetragen (dargestellt durch weiße Kreise im Funnel-Plot), wobei im unverzerrten Fall eine symmetrische Streuung aller Einzeleffekte um den mittleren Gesamteffekt (vertikale Linie in der Mitte) erwartet wird. Studien mit hoher Präzision (mit großem Stichprobenumfang), die den Effekt also genauer messen, befinden sich demnach näher am mittleren Gesamteffekt. Geringe Präzision (typisch für kleinere Stichproben) geht mit einer größeren, und im unverzerrten Fall symmetrischen Schwankungsbreite um den mittleren Gesamteffekt einher. Es resultiert die charakteristische verkehrt trichterförmige Form des Punkteschwarms. Bei Vorliegen einer Publikationsverzerrung allerdings fällt diese Schwankung von Studien mit geringer Präzision einseitig aus: Kleine Stichproben werden nur publiziert, wenn diese einen großen Effekt vorweisen können, womit sich am linken unteren Auslauf des Punkteschwarms (wo kleine Stichproben mit kleinen Effekten markiert sein müssten) eine auffällige Lücke ergibt. Diese Lücke wird entsprechend den Berechnungen nach Duval und Tweedie (2000) mit zusätzlichen Effekten gefüllt (dargestellt durch schwarze Punkte), um die Symmetrie um den mittleren Gesamteffekt herzustellen und die Verzerrung auszugleichen. Der unter Berücksichtigung der zusätzlichen Effekte erneut geschätzte Gesamteffekt ist im Vergleich zum weißen Rhombus unterhalb der x-Achse schwarz dargestellt. Man erkennt hier die kleine Linksverschiebung des adjustierten Effekts (schwarzer Rhombus) im Vergleich zum ursprünglich beobachteten (weißer Rhombus).

Alle oben genannten durchgeführten Verfahren zur Schätzung einer eventuellen Publikationsverzerrung legen gemeinsam nahe, dass beim beobachteten Gesamteffekt eine Überschätzung wie erwartet vorliegen könnte, diese Verzerrung jedoch insgesamt für den mittleren Geschlechtseffekt keine dramatische Wirkung zu haben scheint.

Tabelle 6: Resultate der durchgeführten Verfahren zur Schätzung der Publikationsverzerrung des Gesamteffekts

Failsafe N		Rangkorrelation (Begg & Mazumdar)		Regression Intercept (Egger)		Trim-and-Fill (Duval & Tweedie)			Modell
(Rosenthal)	(Orwin)	τ	p	Intercept	p	d beobachtet	Anzahl zusätzlicher Effekte	d adjustiert	
48386	625	0.073	.089	0.839	<.001	0.506	29	0.472	Fester Effekt
						0.547	23	0.496	Zufallsvariable Effekte

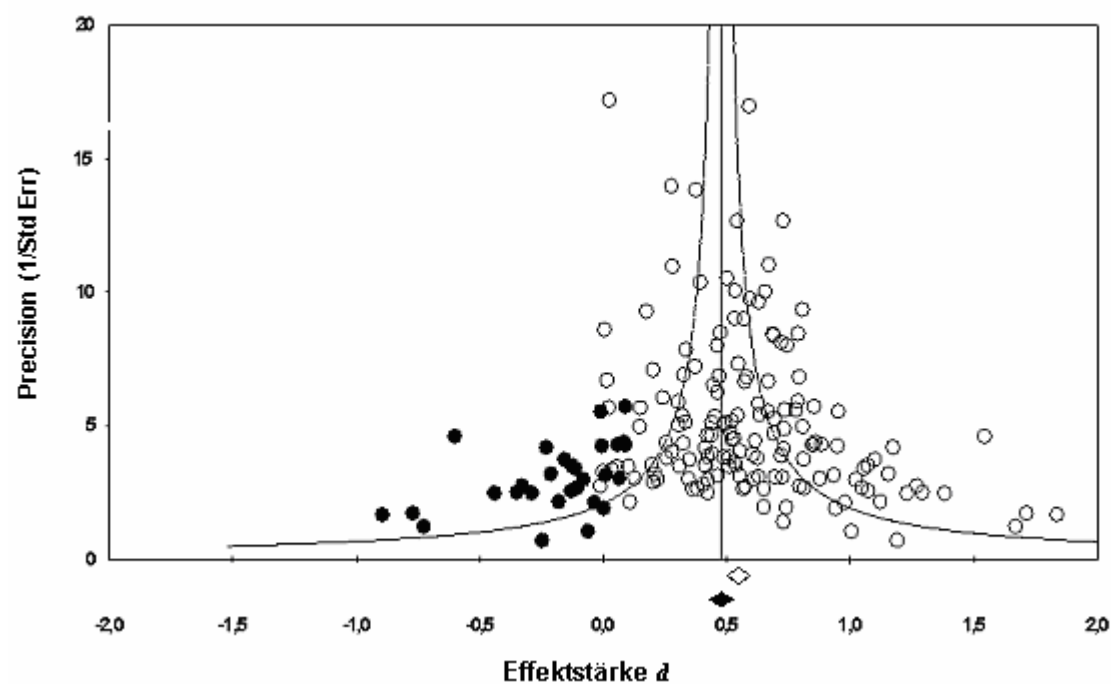


Abbildung 4: Funnel-Plot: Schätzung des Gesamteffekts (Rhombus weiß bzw. schwarz) ohne bzw. mit Berücksichtigung zusätzlicher Effekte (schwarze Punkte)

Aus welchen Gründen könnte der hier beobachtete Gesamteffekt nicht in stärkerem Ausmaß von einer Publikationsverzerrung beeinflusst werden? Einen triftigen Grund liefert vermutlich der Umstand, dass Water-Level Tasks im Rahmen der inkludierten Untersuchungen häufig nicht mit dem ausschließlichen Ziel der Erfassung des Geschlechtseffekts in der Leistung angewandt wurden, sondern diese Messung lediglich ein Nebenprodukt eines anderen inhaltlichen Zwecks ist. In der vorliegenden Analyse scheinen bei 24 von 86 inkludierten Studien in deren Titel Geschlechtsunterschiede hinsichtlich räumlicher Fähigkeiten oder speziell hinsichtlich der Testleistung in Water-Level Tasks als Untersuchungsfokus auf (Amponsah, 2000; Casey, Nuttall & Pezaris, 2001; Geiringer & Hyde, 1976; Golbeck, 1986; Harris, Hanley & Best, 1977; Kalichman, 1989a; Kenyon, 1984; Lawton, 1994; Levine & Stern, 2002; Liben & Golbeck, 1984; Linn & Pulos, 1983; McGiffert-Spradlin & Higgins-Hensley, 1979; Meehan & Overton, 1986; Parameswaran, 1995; Ray, Georgiou & Ravizza, 1979; Robert & Harel, 1996; Robert & Longpré, 2005; Robert & Morin, 1993; Seng & Tan, 2002; Tegano & Faulkender, 1983; Treagust, 1980; Vasta, Knott & Gaze, 1996; Vasta, Lightfoot & Cox, 1993; Voyer, 1996), wobei diese Studien 44 von 154 (entspricht 28.57%) relevanten Stichproben beinhalten. Die Stärke des gemessenen Geschlechtseffekts bei Water-Level Tasks steht somit bei der Mehrheit der inkludierten Untersuchungen (in über 70% der Fälle) nicht im Fokus der Betrachtung und unterliegt aus diesen Gründen auch nicht der systematischen Zurückhaltung aus publikationstechnischen Gründen im Falle einer schwächeren Ausprägung. Die Geltung dieser Überlegungen generell für Reviews von Geschlechtsdifferenzen wird auch von Hyde (1986) angenommen.

5.3 *Intervenierende Variablen*

5.3.1 *Subgruppenanalysen*

Im Zuge der durchgeführten Subgruppenanalysen konnten sechzehn Variablen, die kategoriale Merkmale der inkludierten Untersuchungen erfassen, als mögliche Moderatoren berücksichtigt und auf ihren Einfluss auf den Geschlechtseffekt hinsichtlich der durchschnittlichen Leistung in Water-Level Tasks überprüft werden. Die Resultate aller Subgruppenanalysen wurden jeweils aufgrund vorliegender starker Heterogenität in den Gruppen nach dem Modell zufallsvariabler

Effekte interpretiert. Die Kennwerte aller durchgeführten Prozeduren finden sich in Appendix F.

Hinsichtlich jener Variablen, die Charakteristiken der Studien selbst beschreiben, ergaben sich die folgenden Resultate:

Bezüglich Publikationsstatus der Untersuchungen wurde die Menge aller inkludierten Effekte gruppiert in publiziert – nicht publiziert. Gemäß dem Modell zufallsvariabler Effekte ergaben sich keine signifikanten Unterschiede zwischen den Gruppen ($Q = 0.005$; $df = 1$; $p = .943$). Dieser Befund ergänzt als direkter Test auf Vorliegen eines Publikations-Bias die Resultate der speziellen Testverfahren bezüglich Schätzung einer solchen Verzerrung aus Kapitel 5.2.1. Zusammenfassend kann nun festgehalten werden, dass eine sehr geringe Publikationsverzerrung vorliegen dürfte, die zwar von Verfahren diagnostiziert werden kann, welche sowohl differenziert hinsichtlich derartiger Verzerrungen messen als auch über adäquate Testmacht verfügen (siehe u. a. Regression-Intercept-Test in Kapitel 5.2.1), während jedoch andere Tests – so auch der hier besprochene direkte Gruppenvergleich publizierter versus nicht publizierter Untersuchungen – keine auffälligen Ergebnisse zeigen. Bezüglich Geschlecht des Erstautors (weiblich – männlich – keine Angabe) ergaben sich ebenfalls keine signifikanten Unterschiede ($Q = 2.985$; $df = 2$; $p = .225$). (*Anmerkung:* Bei Interpretation des Modells des festen Effekts ergibt sich ein signifikanter Unterschied hinsichtlich des beobachteten Geschlechtseffekts zwischen männlichen und weiblichen Erstautoren: Untersuchungen mit weiblichen Erstautoren finden mit $d = 0.548$ und $KI : [0.511; 0.585]$ einen signifikant größeren Geschlechtseffekt zu Gunsten von männlichen Testpersonen als Untersuchungen mit männlichen Erstautoren mit $d = 0.456$ und $KI : [0.418; 0.495]$.)

Hinsichtlich jener Variablen, die eine Eigenschaft der Testpersonen beschreibt, ergab sich das folgende Resultat:

Die Altersgruppe der Testpersonen (Kinder – Jugendliche – Erwachsene – gemischte Altersgruppen) hat gemäß dem Modell zufallsvariabler Effekte einen signifikanten Einfluss auf den beobachteten Geschlechtseffekt ($Q = 45.661$; $df = 3$; $p < .001$). Für Kinder ergab sich ein mittlerer Gruppeneffekt von $d = 0.363$, für Jugendliche $d = 0.498$, für Erwachsene $d = 0.606$ und für die gemischte Altersgruppe $d = 0.269$, wobei alle Gruppeneffekte hochsignifikant ausfielen (alle $p < .001$). Somit wurde ein Anstieg des Geschlechtseffekts mit dem Alter der Altersgruppe festgestellt.

Hinsichtlich jener Variablen, die Eigenschaften der Testaufgaben beschreiben, ergaben sich durch entsprechende Gruppierungen in den Subgruppenanalysen folgende Resultate:

Bezüglich der Anzahl von Musteritems ohne Feedback vor der Testung (kein – ein – mehr als ein Musteritem – keine Angabe) konnte kein Einfluss auf den Geschlechtseffekt festgestellt werden ($Q = 0.055$; $df = 3$; $p = .997$). Kein Einfluss auf den Geschlechtseffekt zeigte sich weiters bei den Variablen Gefäßform ($Q = 13.197$; $df = 8$; $p = .105$; gruppiert in: rechteckig – zylindrisch – kugelförmig – Weinflasche – Tasse – Kochflasche – Krug – gemischte Formen – keine Angabe), Vorgabemodus ($Q = 0.380$; $df = 5$; $p = .996$; gruppiert in: Zeichnung – Fotografie – reales Gefäß – computergestützt – gemischt – keine Angabe), Einbettung des Gefäßes in situativen Kontext ($Q = 0.553$; $df = 2$; $p = .758$; gruppiert in: mit Kontext – ohne Kontext – keine Angabe), Verfügbarkeit eines realen Modells ohne sichtbaren Wasserspiegel ($Q = 5.742$; $df = 4$; $p = .219$; gruppiert in: zu keiner Zeit – nur vor der Testung – vor und während der Testung – gemischt – keine Angabe), Verfügbarkeit einer horizontalen Referenz wie Tischplatte oder horizontale Linie unter dem Gefäß ($Q = 3.255$; $df = 2$; $p = .196$; gruppiert in: mit Referenz – ohne Referenz – keine Angabe) sowie Antwortmodus ($Q = 2.398$; $df = 4$; $p = .663$; gruppiert in: Zeichnen des Wasserspiegels – Anpassen des Wasserspiegels anhand eines Modells – Auswahl aus gegebenen Antwortalternativen – gemischt – keine Angabe).

Hinsichtlich jener Variablen, die die Testauswertung näher beschreiben, ergaben sich die folgenden Resultate:

Als Toleranzkriterium erhoben wurde die Größe des jeweils maximal erlaubten Abweichungsfehlers in Grad, der die gezeichnete Wasserlinie gerade noch als korrekte Antwort gelten lässt, wenn die Testantworten als richtig und falsch dichotomisiert wurden. Für die Subgruppenanalyse wurde gruppiert in kleine (3° bis 5°) und große (6° bis 15°) Toleranzkriterien. In anderen Fällen wurde die Testantwort entweder als der mittlere Abweichungsfehler in Grad oder dem Multiple-Choice-Format gemäß ausgewertet. Um zu prüfen, ob das Auswertungskriterium den Geschlechtseffekt beeinflusst, wurde gruppiert in: kleines Toleranzkriterium – großes Toleranzkriterium – mittlerer Abweichungsfehler – Multiple-Choice-Format – andere, nicht klassifizierbare Auswertungskriterien – keine Angabe. Aufgrund dieser Gruppierung ergaben sich keine signifikanten Unterschiede im Geschlechtseffekt

($Q = 5.481$; $df = 5$; $p = .360$). Ebenfalls keine signifikanten Unterschiede ergaben sich bezüglich dem Ergebnismodus ($Q = 6.198$; $df = 4$; $p = .185$; gruppiert in: Anzahl erfolgreicher Testpersonen – mittlere Anzahl korrekt gelöster Items – mittlerer Abweichungsfehler – anderer, nicht klassifizierbarer Ergebnismodus – keine Angabe).

Hinsichtlich jener Variablen, die die Rahmenbedingungen der Testdurchführung näher beschreiben, ergaben sich folgende Resultate:

Bezüglich Testsituation (gruppiert in: Gruppensetting – Einzelsetting – gemischt – keine Angabe) ergab sich kein Einfluss auf den Geschlechtseffekt ($Q = 4.627$; $df = 3$; $p = .201$). Weder Gruppierungen bezüglich Geschlecht des Testleiters ($Q = 1.675$; $df = 4$; $p = .795$; gruppiert in: weiblich – männlich – gleichgeschlechtlich mit Testpersonen – gemischt – keine Angabe), noch bezüglich Instruktionsmodus ($Q = 7.447$; $df = 5$; $p = .189$; gruppiert in: verbal – verbal und anschließend schriftlich – schriftlich – schriftlich und anschließend laut vorgelesen – andere, nicht klassifizierbare Instruktionsmodi – keine Angabe) oder Speed- versus Powertestung ($Q = 0.242$; $df = 1$; $p = .623$) konnten signifikante Gruppenunterschiede im Geschlechtseffekt aufdecken.

Zusammenfassend kann festgehalten werden: Im Zuge der Subgruppenanalysen, bei denen insgesamt sechzehn kategorielle potentielle Moderatorvariablen berücksichtigt wurden, zeigte lediglich der Moderator Altersgruppe einen signifikanten Einfluss auf den Geschlechtseffekt. Der mittlere Geschlechtseffekt von Kindern unterscheidet sich signifikant von jenem der Erwachsenen (siehe Konfidenzintervalle der Gruppen in Appendix F, Tabelle F.3). Ein Anstieg des Ausmaßes des Geschlechtseffekts mit dem Alter der Gruppen ist zu beobachten. Abbildung 5 veranschaulicht dies graphisch.

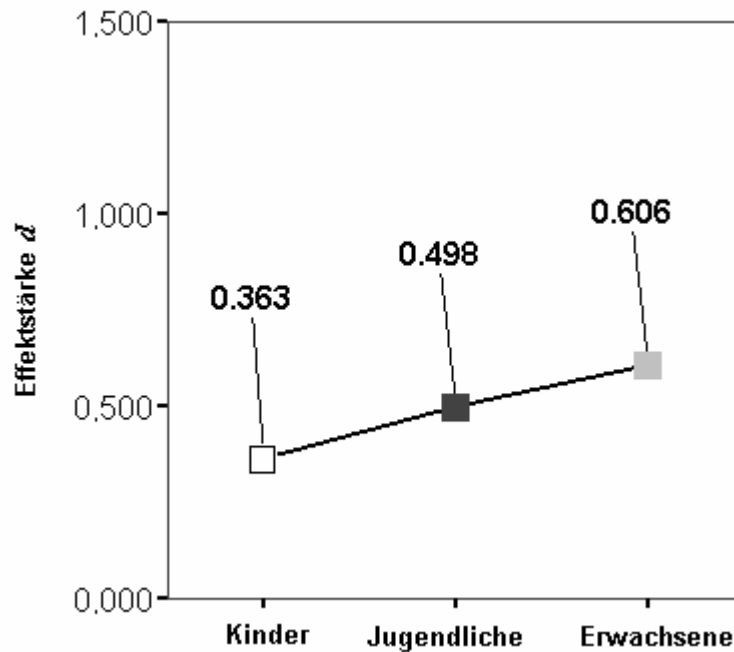


Abbildung 5: Anstieg des Geschlechtseffekts bei Water-Level Tasks mit dem Alter der Altersgruppe

Für die fünfzehn übrigen berücksichtigten kategoriellen Variablen konnte kein Einfluss auf den Geschlechtseffekt nachgewiesen werden.

5.3.2 Effektmoderatorenregression

Um bei der Prüfung möglicher wirksamer Moderatoren gemeinsam mit kategoriellen auch metrische Variablen berücksichtigen zu können, wurde eine Effektmoderatorenregression in Form einer schrittweisen gewichteten multiplen Regression durchgeführt, wobei die abhängige Variable der Geschlechtseffekt pro Studie war. Ziel dieser Prozedur ist es, durch schrittweise Ein- und Ausschlüsse aus der Menge aller berücksichtigten Prädiktoren den Prädiktor bzw. die Prädiktoren mit dem besten Vorhersagewert in Bezug auf die abhängige Variable zu ermitteln.

Da im Zuge dieser Analyse nur Fälle berücksichtigt werden, die keine fehlenden Werte in zumindest einer der aufgenommenen Variablen aufweisen, wurden diejenigen Variablen mit den höchsten Anzahlen fehlender Werte von dieser Analyse ausgeschlossen, um die Anzahl der berücksichtigten Fälle möglichst hoch zu halten. Elf der achtzehn erhobenen potentiell intervenierenden Variablen konnten schließlich aufgenommen werden, um die Anzahl der berücksichtigten Fälle von > 100 sicherzustellen. Diese Variablen sind (in Klammer bei kategorialer Ausprägung die Anzahl ihrer Kategorien, siehe auch Kapitel 5.3.1): Publikationsstatus (2), Publikationsjahr, Geschlecht des Erstautors (2), Altersgruppe der Stichprobe (3),

Gefäßform (8), Einbettung des Gefäßes in situativen Kontext (2), Anzahl der Testitems, Auswertungskriterium (5), Ergebnismodus (4), Testsituation (3), Speed- vs. Powertestung (2); Nicht berücksichtigt werden konnten aufgrund der Anzahl fehlender Werte die folgenden Variablen: Verfügbarkeit von Musteritems ohne Feedback, Vorgabemodus, Verfügbarkeit eines realen Modells, Verfügbarkeit einer horizontalen Referenz, Antwortmodus, Geschlecht des Testleiters, Instruktionsmodus;

Publikationsjahr sowie Anzahl der Testitems konnten als metrische Variablen nun in die Moderatoranalyse miteinbezogen werden. Die übrigen kategoriellen Variablen (mit jeweils zwei bis acht Kategorien, siehe oben) wurden dummy-kodiert, wodurch für jede kategorielle Variable mit c Kategorien jeweils $c-1$ Dummy-Variablen entstanden, sodass sich insgesamt schließlich 24 Prädiktoren ergaben. Ein Überblick über inkludierte Prädiktoren findet sich in Appendix G, Tabelle G.1. Bei Aufnahme dieser 24 Prädiktoren in die Regression resultierten 116 Fälle ohne fehlende Werte, die in der Regressionsanalyse berücksichtigt werden konnten. Die in der Regression inkludierten 116 Fälle dürfen hinsichtlich des Geschlechtseffekts als repräsentativ für die Menge aller 154 relevanten Einzeleffekte verstanden werden: Der zuvor ermittelte Gesamteffekt von $d = 0.547$ für 154 Einzeleffekte liegt innerhalb des Konfidenzintervalls ($KI : [0.510; 0.606]$) für den mittleren Gesamteffekt der inkludierten 116 Fälle, der sich auf $d = 0.558$ beläuft.

Die für die Effektmoderatorenregression angewandten relativen Gewichte beruhten auf den inversen Standardfehlern der berücksichtigten 116 Einzeleffekte und wurden derart konstruiert, dass ihre Summe die Anzahl der in der Regression berücksichtigten Fälle (116) ergibt.

Im Zuge dieser Prozedur in Form einer schrittweisen gewichteten multiplen Regression resultierten zwei Regressionsmodelle. Sowohl Modell 1 ($F = 10.295$; $df_1 = 1$, $df_2 = 114$; $p = .002$; $R^2 = .083$) als auch Modell 2 ($F = 8.696$; $df_1 = 2$, $df_2 = 113$; $p < .001$; $R^2 = .133$) fielen signifikant aus. Modell 1 zeigt einen signifikanten Einfluss der Altersgruppe ($\beta = .150$; $t = 3.209$; $p = .002$; $KI : [.058; .243]$). Modell 2 zeigt zusätzlich zur Altersgruppe ($\beta = .181$; $t = 3.826$; $p < .001$; $KI : [.087; .274]$) auch für die Variable Auswertungskriterium ($\beta = .134$; $t = 2.568$; $p = .012$; $KI : [.031; .238]$) signifikante Einflüsse auf den Geschlechtseffekt.

Aufgrund der Kollinearitätsstatistik $VIF < 10$ (in allen Fällen) darf angenommen werden, dass kein Kollinearitätsproblem vorliegt.

Programmausgaben dieser Prozedur finden sich in Appendix G, Tabellen G.2 bis G.5.

Der im Zuge der Subgruppenanalysen ermittelte signifikante Moderator Altersgruppe wurde somit im Rahmen der Effektmoderatorenregression als wichtigster Prädiktor, mit größtem Vorhersagewert für den Geschlechtseffekt, auch gegen potentielle metrische Moderatoren abgesichert, welche in dem in Kapitel 5.3.1 ausgeführten Analyseschritt der Subgruppenanalyse hinsichtlich ihres Einflusses auf den Geschlechtseffekt nicht gemeinsam mit potentiellen kategoriellen Moderatoren getestet werden konnten. Weiterhin wurde durch diese Analyse offensichtlich, dass das Publikationsjahr als (metrischer) Moderator keinen signifikanten Vorhersagewert für den Geschlechtseffekt besitzt. Demzufolge unterliegt der Geschlechtseffekt bei Water-Level Tasks in den letzten 45 Jahren (d. h. seit dem Beginn seiner Beforschung) keiner systematischen Veränderung. Insbesondere konnte keine kontinuierliche Verminderung festgestellt werden, wie sie von Voyer, Voyer und Bryden (1995) berichtet wird.

6 Diskussion

Die Einordnung der ermittelten Resultate in die bestehende Literatur sowie entwickelte Perspektiven für künftige Untersuchungen werden im Folgenden in den Abschnitten Ausmaß und Generalisierbarkeit des Geschlechtseffekts, Veränderung des Geschlechtseffekts mit dem Lebensalter, Veränderung des Geschlechtseffekts über die Zeit, Quellen des Geschlechtseffekts sowie Fazit und Ausblick dargestellt.

6.1 Ausmaß und Generalisierbarkeit des Geschlechtseffekts

Die aktuelle Meta-Analyse stützt sich auf ein Literaturkorpus mit sechsfachem Umfang jenes Korpus, das Voyer, Voyer und Bryden (1995) in der letzten (und vor Durchführung der vorliegenden Analyse einzigen) Untersuchung zu diesem Thema berücksichtigten. Der gewichtete mittlere Geschlechtseffekt betreffend die Leistung bei Water-Level Tasks wird durch die vorliegende Meta-Analyse mit einem hochsignifikanten $d = 0.547$ ($p < .001$) zu Gunsten männlicher Testpersonen um ein Drittel aufwärts korrigiert in Relation zum in Kapitel 1.4. erläuterten Fund von Voyer, Voyer und Bryden (1995). Die große Menge der in der vorliegenden Analyse zusätzlich berücksichtigten Variablen zeigt außerdem, dass von der Generalisierbarkeit des Geschlechtseffekts bei Water-Level Tasks ausgegangen werden darf.

6.2 Veränderung des Geschlechtseffekts mit dem Lebensalter

Eine Veränderung des Geschlechtseffekts bei Water-Level Tasks mit dem Lebensalter besteht laut den Befunden von Voyer, Voyer und Bryden (1995) nicht. Die vorliegende Untersuchung kommt zu einem anderen Ergebnis: Der Moderator

Altersgruppe fällt sowohl im Rahmen der durchgeführten Subgruppenanalysen als auch im Zuge der Effektmoderatorenregression signifikant aus und offenbart einen Anstieg des Geschlechtseffekts mit dem Lebensalter der Altersgruppen Kinder – Jugendliche – Erwachsene (siehe Kapitel 5.3). Dieses Resultat wird auch jüngst durch eine Untersuchung von Tran und Formann (2008) unterstützt: Eine Stichprobe von $N = 1704$ Testpersonen zeigt in drei Altersgruppen (jünger als 21 Jahre, 21 bis 60 Jahre, älter als 60 Jahre) eine mittlere Geschlechtsdifferenz von jeweils $d = 0.06$, $d = 0.32$ bzw. $d = 0.63$ und weist somit ebenfalls auf einen Anstieg des Geschlechtseffekts mit dem Lebensalter hin. In keiner anderen Primärstudie wurde der Geschlechtseffekt in dieser Breite des Altersspektrums beleuchtet, üblicherweise werden lediglich einzelne abgegrenzte Altersabschnitte berücksichtigt.

Ein Erklärungsmodell für den Fund ansteigender Geschlechtsunterschiede mit dem Lebensalter findet sich in der „Bent Twig“-Hypothese (*engl.* gekrümmter Zweig) nach Sherman (1978; zitiert nach Baenninger & Newcombe, 1995). Diese Theorie bezieht sich allgemein auf die Entstehung geschlechtsspezifischer Leistungsdifferenzen (vgl. Halpern, 2000). Von manchen Autoren wird dieses Modell auch speziell bezüglich des Geschlechtseffekts bei Water-Level Tasks als gültiges integratives Erklärungsmodell (siehe Kapitel 6.5) in Erwägung gezogen (Robert & Harel, 1996; Vasta & Liben, 1996). Das hier gefundene Resultat des Anstiegs des Geschlechtseffekts mit dem Alter erhöht die Plausibilität dieses Modells in besonderem Maße: Ausgehend von einer geschlechtsspezifischen biologischen Prädisposition für bestimmte Kompetenzen und Präferenzen, bereits beeinflusst durch unterschiedliche prä- und postnatale endokrine Faktoren, fließen laut der Theorie in weiterer Folge unterschiedliche Kindheitserfahrungen sowie unterschiedliche kognitive und soziokognitive Faktoren ein, die dazu beitragen, dass ein Individuum seine Schwächen keiner Förderung zugänglich macht bzw. seine Stärken noch weiter trainiert. Fortschreitende Verkümmern bzw. Vermehrung bestimmter Fähigkeiten sind die Folge, wodurch sich die fortschreitende Vergrößerung des geschlechtsspezifischen Leistungsunterschiedes ergibt. Dieses Modell ist besonders dafür geeignet, den hier erstmals meta-analytisch belegten Anstieg des Geschlechtseffekts mit dem Alter zu erklären.

6.3 Veränderung des Geschlechtseffekts über die Zeit

Laut Feingold (1988; zitiert nach Voyer, Voyer & Bryden, 1995) könnte das Ausmaß von geschlechtsspezifischen Differenzen in bestimmten kognitiven Leistungen beeinflusst werden von der sich über die Zeit verändernden gesellschaftlichen Einstellung gegenüber Frauen und deren rollentypischen Leistungen. Aufgrund dessen würden sich derartige Geschlechtseffekte in den letzten Jahren verringern. Der Befund von Harris (1981; zitiert nach Zinser, Palmer & Miller, 2004) unterstützt diese Theorie ebenfalls, während die Resultate von Colom, Quiroga und Juan-Espinosa (1999; zitiert nach Zinser, Palmer & Miller, 2004) der Annahme eines derartigen Trends der Verminderung kognitiver Geschlechtsunterschiede widersprechen. Voyer, Voyer und Bryden (1995) untersuchten diese These speziell bezogen auf die Leistungen in Water-Level Tasks und fanden eine Verringerung des beobachteten Geschlechtseffekts mit Ansteigen des Geburtsjahres der Testpersonen, was die Postulate von Feingold (1988) und Harris (1981) auch für die geschlechtsspezifische Leistungsdifferenz bei Water-Level Tasks zu bestätigen scheint. In der vorliegenden Meta-Analyse hingegen zeigte sich kein Einfluss der chronologischen Variable Publikationsjahr der Studien (das mit dem Geburtsjahr der Testpersonen hochsignifikant korreliert: Spearman's $r = .749$; $p < .001$) auf den Geschlechtseffekt (siehe Kapitel 5.3.2), wofür die Befunde von Colom, Quiroga und Juan-Espinosa (1999) sprechen. Eine Veränderung des Geschlechtseffekts über die Zeit liegt bei Water-Level Tasks demnach nicht vor. Die hier berücksichtigten Studien aus den Jahren 1964 bis 2008 decken über 45 Jahre hinweg die gesamte Zeit der Beforschung des Phänomens seit seinem Bekanntwerden ab, während die Untersuchung von Voyer, Voyer und Bryden (1995) mit einem Drittel des Umfangs nur 30 Forschungsjahre abdecken kann.

6.4 Quellen des Geschlechtseffekts

Die Suche nach der Quelle des Geschlechtseffekts hinsichtlich der Leistung in Water-Level Tasks blickt auf eine 45 Jahre alte Forschungstradition zurück. Betreffend die in Kapitel 1.3 vorgestellten Variablen endokrine und biogenetische Faktoren, frühe Erfahrung in räumlichen Aktivitäten, allgemeines physikalisches Wissen, spezielles physikalisches Wissen über das Invarianzprinzip, räumliche Fähigkeiten, Erfolgswartung und Attributionsstil, Geschlechtsrolle sowie kognitiver

Stil existieren in der einschlägigen Fachliteratur für jeden dieser Faktoren unterstützende Hinweise hinsichtlich seiner ursächlichen Beteiligung am Geschlechtseffekt bei Water-Level Tasks. Bis dato jedoch ist eine Hypothese ausständig, die den Geschlechtseffekt vollständig zu erklären vermag – wobei ein Erreichen dieses Ziels, dessen Forschungsgeschichte inzwischen beinahe ein halbes Jahrhundert überspannt, wohl als beschränkt realistisch angesehen werden muss. Intuitiv ist anzunehmen, dass bei der erfolgreichen Lösung von Water-Level Tasks vor allem eine starke räumliche Komponente beteiligt ist. In einer Studie von Geiser, Lehmann und Eid (2008) findet sich bezüglich des Mental Rotations Test (MRT) ein ähnliches Muster der Geschlechtsdifferenzen, wie es in der vorliegenden Arbeit auch für Water-Level Tasks gefunden wurde: Der Geschlechtseffekt im MRT zu Gunsten von Männern verstärkt sich laut der genannten Studie ebenfalls mit dem Lebensalter und erweist sich in den letzten Jahren ebenfalls als stabil. Wie bereits in Kapitel 1.3.3 erwähnt, wurden zwischen MRT und Water-Level Tasks substantielle Korrelationen von .39 bis .57 beobachtet, die jedoch trotz ihrer Substantialität schwächer ausfallen, als dies aufgrund der Resultate von Geiringer und Hyde (1976) zu erwarten wäre (siehe Kapitel 1.3.3). Allerdings entspricht laut Schumann (2006, S.32) „die empirisch gemessene Korrelation nur dann der Korrelation der wahren Werte, wenn alle Messfehler gleich null und beide Instrumente damit vollständig reliabel sind.“ Da jedoch weder Water-Level Tasks noch der MRT vollständig reliabel sind, wird die Korrelation zwischen diesen beiden Tests proportional zu deren Messfehler unterschätzt. Eine Bereinigung dieser Unterschätzung ist bei Kenntnis der Reliabilitäten beider Tests durch die sogenannte Verdünnungsformel (Formel 2) möglich:

$$\rho_{T_X T_Y} = \frac{r_{XY}}{\sqrt{r_{XX} \cdot r_{YY}}} \quad (2)$$

Die Reliabilität des Mental Rotations Test (MRT) wird von den Testautoren Vandenberg und Kuse (1978) mit .88 angegeben, die Reliabilität der Water-Level Tasks mit .80 (Casey, Nuttall & Pezaris, 2001; Krekling & Nordvik, 1992) bzw. mit .968 (U. Tran, persönliche Mitteilung, 18. März, 2009). Somit steigt die Korrelation $\rho_{T_X T_Y}$ zwischen Water-Level Tasks und MRT nach Anwendung der Verdünnungsformel auf die beobachteten Korrelationen .39 bis .57 (siehe Kapitel

1.3.3) auf nunmehr .47 bis .68 bzw. auf .42 bis .62. Nach der durchgeführten Korrektur ist der Zusammenhang zwischen Water-Level Tasks und MRT nun stärker und bewegt sich insgesamt zwischen .42 und .68. Die Annahme, dass Fähigkeiten zur mentalen Rotation maßgeblich zur Lösung von Water-Level Tasks beitragen, erscheint plausibel. Untersuchungen zu Geschlechtsunterschieden in räumlichen Fähigkeiten differenzieren räumliche Wahrnehmung, mentale Rotation und räumliche Visualisierung als voneinander abgegrenzte Kategorien innerhalb des gesamten räumlichen Fähigkeitsspektrums (Linn & Petersen, 1985; Voyer, Voyer & Bryden, 1995), die sich auch in der Größe der geschlechtsspezifischen Leistungsdifferenz voneinander unterscheiden. Es ist zu erwarten, dass die genannten Kategorien an der Lösung von Water-Level Tasks in unterschiedlichem Ausmaß beteiligt sind. Analysen in diese Richtung könnten zur weiteren Aufklärung des Geschlechtseffekts maßgeblich beitragen.

Anhand der vorliegenden Untersuchung konnte aus insgesamt 18 berücksichtigten Moderatoren bereits der Einfluss der Altersgruppe auf das Ausmaß des Geschlechtseffekts bei Water-Level Tasks aufgedeckt und mit unterschiedlichen Methoden erstmals meta-analytisch bestätigt werden. Es liegt ein generalisierbarer Geschlechtseffekt vor.

6.5 Fazit und Ausblick

Mit Hilfe der vorliegenden, bisher umfassendsten Analyse zum Thema geschlechtsspezifische Leistungsdifferenz bei Water-Level Tasks konnte ein moderierender Einfluss der Altersgruppe, die Stabilität über die Zeit sowie die Generalisierbarkeit des Geschlechtseffekts bei Water-Level Tasks, der tatsächlich um ein Drittel größer ist als ursprünglich angenommen, nachgewiesen werden.

Zur Klärung der Rolle spezieller räumlicher Fähigkeiten außer der mentalen Rotation stehen noch räumliche Wahrnehmung und räumliche Visualisierung weiteren Untersuchungen in Bezug auf die Leistung bei Water-Level Tasks offen.

Ebenso ist eine Überprüfung der von Sanders und Ross-Field (1986) eingebrachten Theorie des Einflusses endokriner Faktoren auf die geschlechtsspezifische Leistung bei Water-Level Tasks ausständig (siehe Kapitel 1.3.1.2).

Neben den soeben genannten Variablen, die bis dato empirisch noch nicht auf ihre Wirksamkeit bezüglich der Leistung bei Water-Level Tasks überprüft wurden,

glauben mehrere Autoren an den Beitrag eines integrativen Ansatzes für künftige Untersuchungen zur vollständigen Aufklärung des Phänomens. Merriwether und Liben (1997) vertreten die Ansicht, dass die Annahme verschiedener singulärer Kompetenzen bzw. Inkompetenzen und deren wiederholte Untersuchung in eine experimentelle Endlosschleife münden, die den Geschlechtsunterschied nicht vollständig aufklären kann. Eine mögliche Erklärung bietet die „Bent Twig“-Hypothese von Sherman (1978; zitiert nach Baenninger & Newcombe, 1995), die die Wirkung mehrerer Faktoren integriert. Sie wurde bereits in Kapitel 6.3. ausführlich erläutert. Dieses Modell erscheint auch aus einem anderen Blickwinkel in Bezug auf Water-Level Tasks bedeutsam, da es sich vor allem zur Erklärung des in dieser Analyse festgestellten Anstiegs des Geschlechtseffekts mit dem Lebensalter eignet.

Vasta und Liben (1996) vertreten die Meinung, dass Modelle, die Wechselwirkungen zwischen mehreren Variablen berücksichtigen, am besten dazu geeignet sind, die geschlechtsspezifische Leistungsdifferenz bei Water-Level Tasks aufzuklären, da eine Erklärung des Geschlechtseffektes durch eine oder mehrere isolierte Fähigkeiten bzw. Faktoren im Lichte der vielzähligen Untersuchungen der Forschungsgeschichte als unrealistisch anzusehen ist. Derartige Modelle sind vor allem dann angebracht, wenn untersucht werden soll, ob sich bestimmte Kombinationen von Faktorstufen unterschiedlich auf die Testleistung männlicher und weiblicher Testpersonen auswirken.

Ein weiterer integrativer Ansatz stammt von Kalichman (1988). Von ihm wurde ein Modell bestehend aus mehreren Kompetenzkomponenten zur Erklärung der Leistung bei Water-Level Tasks entwickelt. Dieses Modell geht davon aus, dass an der korrekten Lösung einer Wasserstandsaufgabe mehrere Fähigkeiten zu unterschiedlichen Zeitpunkten des Lösungsprozesses beteiligt sind. Folgende Kompetenzkomponenten müssen für die Aufgabenbewältigung laut Modell miteinander verbunden werden: visuelle Wahrnehmung, mentale Herauslösung des Gefäßes aus seinem unmittelbaren Umfeld (um seine Position relativ zu einer externen Referenz wahrnehmen zu können; hier kommt der kognitive Stil zum Ausdruck), Generierung einer mentalen Repräsentation (des Gefäßes samt Flüssigkeit), mentale Rotation (des Gefäßes), räumliche Koordination (um den Verlauf des Flüssigkeitspegels relativ zur geeigneten externen Referenz angemessen beurteilen zu können), Informationsabruf aus der Erinnerung betreffend die Wirkung der Gravitation; Dieses Multikomponentenmodell integriert somit

visuoperzeptive Faktoren, kognitiven Stil, räumliche Fähigkeiten sowie spezielles physikalisches Wissen über das Invarianzprinzip zu einer gemeinsamen Leistung. Es eignet sich zur Berücksichtigung in künftigen Untersuchungen zum Geschlechtseffekt bei Water-Level Tasks, nachdem in der vorliegenden Arbeit gezeigt wurde, dass die Synthese der bisher größten Anzahl von Studien zu Water-Level Tasks gemeinsam mit einer umfassenden Kontrolle konsistent berichteter Variablen zur Klärung dieses Phänomens maßgebliche Beiträge leisten kann.

7 Zusammenfassung

Piagets Wasserstandsaufgaben (auch bekannt als Water-Level Tasks, WLT) wurden ursprünglich entworfen, um die kindliche Entwicklung des externen Referenzsystems anhand der Beurteilung der invarianten Horizontalität des Wasserspiegels in geneigten Gefäßen zu testen. In darauf folgenden Studien wurde ein Geschlechtsunterschied in der Testleistung (zu Gunsten von männlichen Testpersonen) ab dem Alter von ca. 9 Jahren festgestellt, der auch im Erwachsenenalter fortbesteht. Weitere Untersuchungen zum globalen Geschlechtseffekt in der Testleistung der WLT entsprechen entweder als Review nicht den formalen Kriterien einer Meta-Analyse, oder ihr Datensatz ist unvollständig und veraltet. Die vorliegende, bisher umfassendste Meta-Analyse zu diesem Thema basiert auf 154 Stichproben (bestehend aus insgesamt beinahe 26,000 Testpersonen), wobei der aktuelle Umfang des Literaturkorpus' WLT-relevanter Studien jenen der letzten Analyse zu diesem Thema (Voyer, Voyer & Bryden, 1995) um das Fünffache übersteigt, und lässt damit bekannte Grenzen vorangegangener Untersuchungen hinter sich. Der resultierende – hochsignifikante – Geschlechtseffekt ist mit $d = 0.6$ von mittlerem Ausmaß und kann bisherige Erkenntnisse somit um ein Drittel nach oben korrigieren. Von der Generalisierbarkeit des Effekts darf ausgegangen werden. Die gefundene geschlechtsspezifische Leistungsdifferenz wird moderiert durch die Altersgruppe (Kinder ab 7 Jahren, Jugendliche und Erwachsene), wobei Kinder den kleinsten ($d = 0.4$) und Erwachsene den größten Geschlechtseffekt ($d = 0.6$) aufweisen. Weiterhin kann aufgrund der vorliegenden Resultate auf die Stabilität des Geschlechtseffekts über die Zeit geschlossen werden. Einflüsse einer großen Anzahl weiterer potenzieller Moderatorvariablen (Charakteristika der Studien selbst, Eigenschaften der Testpersonen, der

Testaufgaben, der Testauswertung sowie der Rahmenbedingungen der Testdurchführung) wurden berücksichtigt und überprüft. Erklärungsmodelle für diese erstmaligen meta-analytischen Befunde sowie resultierende Implikationen für künftige Untersuchungen werden diskutiert.

APPENDICES

Appendix A: Seit Abschluss der Analyse erschienene relevante Quellen

- Cherney, I. D., Brabec, C. M., & Runco, D. V. (2008). Mapping out spatial ability: Sex differences in way-finding navigation. *Perceptual and Motor Skills*, 107, 747-760.
- Morra, S. (2008). A test of a neo-Piagetian model of the water-level task. *European Journal of Developmental Psychology*, 5, 369-400.
- Rilea, S. L. (2008). A lateralization of function approach to sex differences in spatial ability: A reexamination. *Brain and Cognition*, 67, 168-182.

Appendix B: **Resultierender Suchstring für Literaturrecherche**

Der aus dem Schlagwortkatalog resultierende Suchstring hat die folgende Form:

```
(  
  (  
    water level* OR  
    water-level* OR  
    water line OR  
    water-line OR  
    water surface OR  
    water glass OR  
    water bottle  
  )  
  NEAR  
  (  
    task  
    OR problem  
    OR test  
    OR representation  
    OR assessment  
    OR conceptualization  
    OR draw*  
  )  
)  
OR  
(  
  (  
    horizontal*  
    OR invariance  
  )  
  NEAR  
  (  
    
```

concept*
OR conservation
OR task*
OR item*
OR liquid*
OR representation
OR perception
)
)
OR
(
(
Piaget*
)
NEAR
(
water task
OR water test
)
)
OR Thomas Apparatus
OR Wasserspiegelaufgabe*
OR Wasserstandsaufgabe*
OR WLT
NOT liver transplantation
NOT water load test
NOT world literature today

Anmerkung: Die Ausschlussausdrücke “liver transplantation”, “water load test” und “world literature today” werden in der Literatur ebenfalls unter dem Kürzel “WLT” bzw. „LT“ geführt, sodass bei Verwendung des Suchstrings „WLT“ auch eine große Anzahl von Untersuchungen als Suchergebnisse ausgegeben werden, die inhaltlich nichts mit Water-Level Tasks zu tun haben. Die Verwendung der genannten Ausschlussausdrücke erlaubt eine diesbezügliche Filterung der Suche.

Appendix C: Kodierschema für Primärstudien

Um das Erkennen und Extrahieren relevanter Informationen zu erleichtern und zu beschleunigen, wurde die (englische) Sprache des Kodierschemas an die vorherrschende Sprache der gesichteten Literatur angepasst.

Expression used for Piaget task			
Author			
Year of publication			
Journal name			
Language and Publication type			
Subjects' Sex	Male (#)	Female (#)	Total (#)
Age			
Year of birth			
Ethnicity			
Level or type of education/milieu			
Item examples (#)			
Items (#)			
Type, shape of vessel			
Mode of presentation (real/PP)			

Angles of tilt	
Horizontal cue (y/n)	
Mode of response (MC/ draw/ adjust WL or vessel)	
Test situation (Group/Individual)	
Sex of experimenter	
Mode of instruction (verbal..)	
Speed-/ Powertest	
Criterion of tolerance	
Criterion of success	
Results	

Appendix D: Liste inspizierter, nicht inkludierter Studien

Studie	Ausschlussgrund
Abravanel, E., & Gingold, H. (1982). Evidence for perceptual influences in understanding liquid level invariance. <i>Journal of Genetic Psychology</i> , 140, 237-247.	unrepräsentatives Design
Ackermann, E. (1981). <i>Statut fonctionnel de la représentation dans les conduites finalisées chez l'enfant</i> . Unpublished doctoral dissertation, University of Geneva, Switzerland.	Sprache = Französisch
Ackermann, E. (1991). From decontextualized to situated knowledge: Revisiting Piaget's water-level experiment. In I. Harel & S. Papert (Eds.), <i>Constructionism</i> (pp. 269-294). Westport, CT: Ablex.	keine Daten
Ackermann-Valladao, E. (1987). Que deviennent les idées a propos d'un phénomène une fois retraduites à travers différents medias? Le rôle du dessin, du récit et du geste dans la construction d'une connaissance [What happens to an idea if translated in various media? The role of drawing, language, and gesture in the construction of knowledge]. <i>Archives de Psychologie</i> , 55, 195-218.	keine Daten
Acredolo, C., & Acredolo, L. P. (1979). Identity, compensation, and conservation. <i>Child Development</i> , 50, 524-535.	keine Water-Level Tasks
Akiyama, M., Akiyama, H., & Goodrich, C. (1985). Spatial development across the life span. <i>International Journal of Aging and Human Development</i> , 21, 175-185.	keine ausreichenden Daten
Al-Jubaili, A. Y. (2002). Science misconceptions and working memory capacity among Saudi adolescents: A neo-Piagetian investigation [Abstract]. <i>Dissertation Abstracts International</i> , 63, 2126.	kein Volltext verfügbar
Allen, M. J., Wittig, M. A., & Butler, K. (1981). Comments on Thomas and Jamison's 'A test of the X-linked genetic hypothesis for sex differences on Piaget's water-level task'. <i>Developmental Review</i> , 1, 284-288.	keine Daten
Amponsah, B., & Krekling, S. (1994). Directional biases in adults'	unrepräsentatives

performance on a multiple-choice water-level task. <i>Perceptual and Motor Skills</i> , 79, 899-903.	Design
Andersson, K. E., Richards, H. C., & Hallahan, D. P. (1980). Piagetian task performance of learning disabled children. <i>Journal of Learning Disabilities</i> , 13, 501-505.	keine Water-Level Tasks
Baenninger, M., & Newcombe, N. (1995). Environmental input to the development of sex-related differences in spatial and mathematical ability. <i>Learning and Individual Differences</i> , 7, 363-379.	keine Water-Level Tasks
Baldy, R., Devichi, C., & Chatillon, J. F. (2004). Developmental effects in 2D versus 3D versions in verticality and horizontal tasks. <i>Swiss Journal of Psychology</i> , 63, 75-83.	keine Daten
Barna, J. D., & O'Connell, D. C. (1967). Perception of horizontality as a function of age and stimulus setting. <i>Perceptual and Motor Skills</i> , 25, 70-72.	keine ausreichenden Daten
Barsky, R. D., & Lachman, M. E. (1986). Understanding of horizontal in college women: Effects of two training procedures. <i>International Journal of Behavioral Development</i> , 9, 31-43.	Stichprobe rein weiblich
Beilin, H., Kagan, J., & Rabinowitz, R. (1966). Effects of verbal and perceptual training on water level representation. <i>Child Development</i> , 37, 317-329.	keine ausreichenden Daten
Beilin, H., Kagan, J., & Rabinowitz, R. (1968). Effects of verbal and perceptual training on water level representation. In I. E. Sigel & F. H. Hooper (Ed.), <i>Logical thinking in children: Research based on Piaget's theory</i> (pp. 382-394). New York: Holt, Rinehart and Winston.	Daten aus Beilin, Kagan & Rabinowitz (1966)
Bertamini, M., Spooner, A., & Hecht, H. (2003). Naive optics: Predicting and perceiving reflections in mirrors. <i>Journal of Experimental Psychology-Human Perception and Performance</i> , 29, 982-1002.	keine Water-Level Tasks
Berthiaume, F., Robert, M., St-Onge, R., & Pelletier, J. (1993). Absence of a gender difference in a haptic version of the water-level task. <i>Bulletin of the Psychonomic Society</i> , 31,	unrepräsentatives Design (haptic version)

57-60.	
Chen, S., Zhen, M., & Lin, Z. (1985). A study of experiments on the vertical and horizontal visual perception of children aged 5 to 9. <i>Information on Psychological Sciences</i> , 5, 6-11.	Sprache = Chinesisch
Cherney, I. D., & Collaer, M. L. (2005). Sex differences in line judgment: Relation to mathematics preparation and strategy use. <i>Perceptual and Motor Skills</i> , 100, 615–627.	keine Water-Level Tasks
Choi, J. (2000). <i>Route-learning strategies and spatial processing: A developmental perspective</i> . Unpublished doctoral dissertation, York University, Toronto, Canada.	unrepräsentatives Design (geneigte Referenzebene)
Clark, S. K. (1996). <i>Effect of spatial perception level on achievement using two and three dimensional graphics</i> . Unpublished doctoral dissertation, Pennsylvania State University.	keine Daten
Cohen, H. G. (1978). The scaling of six topological Piagetian groupings, as well as the effect that certain selected variables have on the attainment of these groupings and some of their homologs in the logical domain. <i>Journal of Research in Science Teaching</i> , 15, 115-127.	keine Water-Level Tasks
Croucher, C. J., Bertamini, M., & Hecht, H. (2002). Naive optics: Understanding the geometry of mirror reflections. <i>Journal of Experimental Psychology-Human Perception and Performance</i> , 28, 546-562.	keine Water-Level Tasks
DeAvila, E. A., & Havassy, B. (1974). <i>I.Q. Tests and Minority Children</i> . Austin, TX: Dissemination and Assessment Center for Bilingual Education.	nicht verfügbar
DeAvila, E. A., Havassy, B., & Pascual-Leone, J. (1976). <i>Mexican-American school children: A neo-Piagetian analysis</i> . Washington, DC: Georgetown University Press.	nicht verfügbar
DeLisi, R. (1983). Developmental and individual differences in children's representation of the horizontal coordinate. <i>Merrill-Palmer Quarterly</i> , 29, 179-196.	unrepräsentatives Design (Training)
DeLisi, R., & McGillicuddy-De Lisi, A. V. (1988). Individual differences in adolescents' horizontality representation:	unrepräsentatives Design (Training)

Associations with vocational major and gender. <i>Merrill-Palmer Quarterly</i> , 34, 437-449.	
DeLisi, R., McGillicuddy-De Lisi, A., & Mowrer-Popiel, E. (1995). Measuring grade-school children's ability to represent the Euclidean horizontal coordinate. <i>Journal of Genetic Psychology</i> 156, 231-240.	unrepräsentatives Design (Training)
DeLisi, R., Parameswaran, G., & McGillicuddy-De Lisi, A. (1989). Age and sex differences in representation of horizontality among children in India. <i>Perceptual and Motor Skills</i> , 68, 739-746.	unrepräsentatives Design (Training)
Dockweiler, C. J. (1978). <i>Using systems of reference with spatial concepts: Their development from 5th grade to college age</i> . Unpublished manuscript.	nicht verfügbar
Dodwell, P. C. (1963). Children's understanding of spatial concepts. <i>Canadian Journal of Psychology</i> , 17, 141-161.	keine Daten
Fabian, V. (1982). <i>Language development after 5: A neo-Piagetian investigation of subordinate conjunctions</i> . Unpublished doctoral dissertation, University of California, Berkeley.	nicht verfügbar
Farace, E. (1997). Gender differences in relationships between degree of brain lateralization and cognitive ability [Abstract]. <i>Dissertation Abstracts International</i> , 57, 5324.	kein Volltext verfügbar
Fernandez, T. (2004). Concepciones del planeta Tierra. Capacidades espaciales implicadas [Spatial capacities involved in conceptions of the earth planet]. <i>Infancia y Aprendizaje</i> , 27, 189-210.	Sprache = Spanisch
Foorman, B. R. (1981). A neo-Piagetian analysis of four-year-olds' performance on Piaget's water-level task. <i>Perceptual and Motor Skills</i> , 52, 631-639.	keine ausreichenden Daten
Ford, L. H. (1970). Predictive versus perceptual responses to Piaget's water-line task and their relation to distance conservation. <i>Child Development</i> , 41, 193-204.	keine ausreichenden Daten
Formann, A. K. (1998). <i>Modeling data from water-level tasks</i> . Unpublished doctoral dissertation, University of Sheffield,	Daten aus Formann, 2003

GB.	
Formann, A. K. (2001). Misspecifying latent class models by mixture binomials. <i>British Journal of Mathematical and Statistical Psychology</i> , 54, 279-291.	keine ausreichenden Daten
Golbeck, S. L. (1998). Peer collaboration and children's representation of the horizontal surface of liquid. <i>Journal of Applied Developmental Psychology</i> , 19, 571-592.	keine Daten
Goodman, D. (1971). <i>Cognitive style factors in linguistic performance with ambiguous sentences</i> . Unpublished master's thesis, York University, Toronto, Canada.	nicht verfügbar
Goodman, D. (1979). <i>Stage transitions and the developmental trace of constructive operators: An investigation of a neo-Piagetian theory of cognitive growth</i> . Unpublished doctoral dissertation, York University, Toronto, Canada.	nicht verfügbar
Goodrich, G. A. (1993). Knowing which way is up: Sex differences in understanding horizontality and verticality [Abstract]. <i>Dissertation Abstracts International</i> , 53, 4975.	kein Volltext verfügbar
Goodrich, G. A., Damin, P. B., Ascione, F. R., Thompson, T. M. (1993). Gender differences in Piagetian visual-spatial representation of verticality and horizontality. <i>Journal of Genetic Psychology</i> , 154, 449-458.	unrepräsentatives Design (geneigte Referenzebene)
Green, K. (1975). Age references for Piaget's horizontal reference task. <i>Bulletin of the North Carolina Psychological Association, Spring Issue</i> , 14-15.	nicht verfügbar
Halpern, D. F. (2000). <i>Sex differences in cognitive abilities</i> (3rd ed.). Mahwah, NJ: Erlbaum.	keine Daten
Hecht, H. (2001). Regularities of the physical world and the absence of their internalization. <i>Behavioral and Brain Sciences</i> , 24, 608-617.	keine Daten
Heller, M. A., Brackett, D. D., Scroggs, E., Allen, A. C., & Green, S. (2001). Haptic perception of the horizontal by blind and low-vision individuals. <i>Perception</i> , 30, 601-610.	unrepräsentatives Design (haptic version)
Heller, M. A., Calcaterra, J. A., Green, S. L., Barnette, S. L. (1999). Perception of the horizontal and vertical in tangible	unrepräsentatives Design (haptic version)

displays: Minimal gender differences. <i>Perception</i> , 28, 387-394.	
Héroux, G. & Robert, M. (1993, June). <i>Type of play predicts different visuo-spatial abilities in school-aged girls and boys</i> . Paper presented at the Annual Meeting of the American Psychological Society, Chicago.	nicht verfügbar
Héroux, G. & Robert, M. (1993, September). <i>Corrélat des habiletés visuo-spatiales chez les filles et les garçons d'âge scolaire: regard rétrospectif sur les jeux de la période préscolaire [Visuo-spatial ability correlates among school-aged girls and boys: a retrospective look at play over the pre-school period]</i> . Paper presented at the Annual Meeting of the Association de Psychologie Scientifique de Langue Francaise, Aix-en, Provence, France.	nicht verfügbar
Hogrefe, B. (1985). <i>Zur Entwicklung des Raumkonzepts beim Kind: Horizontale und Vertikale [Development of spatial concepts in children: Horizontalitry and Verticality]</i> . Unpublished doctoral dissertation, University of Salzburg, Austria.	unrepräsentatives Design
Howard, I. P. (1978). Recognition and knowledge of the water-level principle. <i>Perception</i> , 7, 151-160.	unrepräsentatives Design
Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using ω and d . <i>American Psychologist</i> , 36, 892-901.	Meta-Analyse, keine neuen Daten
Hyde, J. S., & Linn, M. C. (1986). <i>The psychology of gender: Advances through meta-analysis</i> . Baltimore: Johns Hopkins University Press.	keine Daten
Jahoda, G. (1976). Reproduction of kohs-type figures by Ghanaian children: orientation errors revisited. <i>British Journal of Psychology</i> , 67, 203-11.	Stichprobe rein männlich
Jamison, W., & Signorella, M. L. (1980). Sex-typing and spatial ability: The association between masculinity and success on Piaget's water-level task. <i>Sex Roles</i> , 6, 345-353.	keine ausreichenden Daten
Johnson, J. (1982). <i>The development of metaphor</i>	kein Volltext verfügbar

<i>comprehension: Its mental-demand measurement and its process-analytical models.</i> Unpublished doctoral dissertation, York University, Toronto, Canada.	
Johnson, W. S., Kelly, G. N., & Kelly, J. T. (1980). Perception of verticality by boys and girls in grade 6. <i>Perceptual and Motor Skills</i> , 51, 355-358.	keine Daten
Kalichman, S. C. (1987, March). <i>Water running-up-hill: A closer look at sex differences in perceiving the invariant horizontality of liquids.</i> Paper presented at the Annual Meeting of the Southeastern Psychological Association, Atlanta, GA.	Daten (aus Studie 1) aus Kalichman (1986) bzw. (aus Studie 2) aus Kalichman (1989a)
Kalichman, S. C. (1988). Individual differences in water-level task performance: A component-skills analysis. <i>Developmental Review</i> , 8, 273-295.	Review, keine neuen Daten
Kelly, G. N. (1979). <i>Analysis of college females' perception of liquid horizontality.</i> Unpublished doctoral dissertation, Washington State University.	Stichprobe rein weiblich
Kelly, G. N., Kelly, J. T., & Johnson, W. S. (1988). Coordination of perspective by sixth graders and teacher trainees. <i>Contemporary Educational Psychology</i> , 13, 358-370.	keine Water-Level Tasks
Kelly, J. T., & Kelly, G. N. (1978). Science activity centers: basic concepts of the physical world. <i>Science and Children</i> , 16(3), 25-26.	keine ausreichenden Daten
Kelly, R. R., & Davis, S. S. (1982). <i>Sex differences in elementary gifted children for perceiving liquid horizontality.</i> Unpublished manuscript.	nicht verfügbar
Kelly, R. R., & Witters, L. A. (1980, October). <i>Developmental and sex differences for perceiving liquid horizontality: Implications for instruction.</i> Paper presented at the meeting of the National Association for Gifted Children, Minneapolis, MN.	nicht verfügbar
Kelly, R. R., & Witters, L. A. (1981). Developmental and sex differences of gifted children's perception of liquid horizontality. <i>Journal for the Education of the Gifted</i> , 4, 85-	unrepräsentatives Design (simultane Testung der Konzepte

96.	
Kershner, J. R. (1970). Children's spatial representation of directional movement and figure orientations along horizontal and vertical dimensions. <i>Perceptual and Motor Skills</i> , 31, 641-642.	der Vertikalität und der Horizontalität) keine Water-Level Tasks
Kershner, J. R. (1973). Conservation of vertical-horizontal space perception in trainable retarded children. <i>American Journal of Mental Deficiency</i> , 77, 710-716.	keine Water-Level Tasks
King, M. A., & Yuille, J. C. (1980). <i>Context effects on children's memory of Piagetian concepts</i> . Paper presented at the 60th Annual Meeting of the Western Psychological Association, Honolulu, HI.	keine Water-Level Tasks
Kohlmann, R. (1984). <i>Zur Entwicklung von Horizontale und Vertikale anhand der Kinderzeichnung [The development of the vertical and the horizontal dimension by the example of children's drawings]</i> . Unpublished doctoral dissertation, University of Salzburg, Austria.	keine ausreichenden Daten
Krekling, S., & Nordvik, H. (1992a). Development of gender differences in performance on a water-level task [Abstract]. <i>International Journal of Psychology</i> , 27, 233.	kein Volltext verfügbar
Krekling, S., & Nordvik, H. (1992b). Observational training improves adult women's performance on Piaget's water-level task. <i>Scandinavian Journal of Psychology</i> , 33, 117-124.	Stichprobe rein weiblich
Krewaz, J. B. (1987). <i>A study into the relationships between field independence levels of thinking, and science achievement</i> . Unpublished master's thesis, Monash University, Melbourne, Australia.	nicht verfügbar
Larsen, G., & Abravanel, E. (1972). An examination of the developmental relations between certain spatial tasks. <i>Merrill-Palmer Quarterly</i> , 18, 39-52.	keine ausreichenden Daten
Lee, K. O. (1995). The effects of computer program training on the cognitive development of children aged 7-sup-8 yrs: Can Piaget's cognitive development process be improved by	Sprache = Koreanisch

training? <i>Korean Journal of Child Studies</i> , 16, 79-88.	
Leino, V., & Willemsen, E. (1976). Use of a perceptually based apparatus to train adult women's performance on a Piagetian measure of the horizontality concept. <i>Perceptual and Motor Skills</i> , 42, 363-369.	Stichprobe rein weiblich
Li, C. (2000). Instruction effect and developmental levels: A study on water-level task with Chinese children ages 9-17. <i>Contemporary Educational Psychology</i> , 25, 488-498.	relevante Daten aus Li, Nuttall & Zhao (1999a)
Liben, L. S. (1978). Performance on Piagetian spatial tasks as a function of sex, field dependence, and training. <i>Merrill-Palmer Quarterly</i> , 24, 97-110.	vorselektierte Stichprobe (failed pretest)
Liben, L. S. (1991). The Piagetian water-level task: Looking beneath the surface. In R. Vasta (Ed.), <i>Annals of Child Development</i> (Vol. 8, pp. 81-143). London: Kingsley.	keine Daten
Liben, L. S., & Golbeck, S. L. (1980). Sex differences in performance on Piagetian spatial tasks: Difference in competence of performance? <i>Child Development</i> , 51, 594-597.	keine ausreichenden Daten
Linn, M. C. (1981, April). <i>Male-female differences in formal thought</i> . Paper presented at the Biennial Meeting of the Society for Research in Child Development, Boston, MA.	keine ausreichenden Daten
Linn, M. C., & Delucchi, K. (1983). <i>The water level influences of task features on gender differences in performance</i> . Berkeley: University of California, Lawrence Hall of Science, Adolescent Reasoning Project.	nicht verfügbar
Linn, M. C., & Kyllonen, P. (1981). The field dependence-independence construct: Some, one, or none. <i>Journal of Educational Psychology</i> , 73, 261-273.	keine ausreichenden Daten
Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial abilities: A meta-analysis. <i>Child Development</i> , 56, 1479-1498.	Meta-Analyse, keine neuen Daten
Linn, M. C., & Pulos, S. (1983a). Aptitude and experience influences on proportional reasoning during adolescence: Focuses on male-female differences. <i>Journal of Research in</i>	Daten aus Linn & Pulos (1983b). <i>Journal of Educational</i>

<i>Mathematics Education</i> , 14, 30-46.	<i>Psychology</i> , 75, 86-96.
Lohaus, A., & Kessler, T. (1996). Zum Zusammenhang zwischen Lösungsverhalten und verbalisiertem Verständnis des Lösungsprinzips bei der Wasserspiegelaufgabe [Performance and verbally expressed understanding of the solution principle in the water-level task]. <i>Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie</i> , 28, 316-335.	keine Daten
Lohaus, A., Kessler, T., Thomas, H., & Gediga, G. (1994). Individuelle Unterschiede bei räumlichen Fähigkeiten im Kindesalter [Individual differences in spatial abilities of children]. <i>Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie</i> , 26, 373-390.	Stichprobe aus Lohaus, Thomas, Kessler & Gediga (1996)
Lohaus, A., Schumann-Hengteler, R., & Kessler, T. (1999). <i>Räumliches Denken im Kindesalter [Childhood spatial thinking]</i> . Göttingen: Hogrefe.	Stichprobe aus Lohaus, Thomas, Kessler & Gediga (1996)
Lord, T., Ferrero, J., Warner, L. & Williams, C. (2000). College students' performance on associated, corresponding tasks for horizontality. <i>Perceptual and Motor Skills</i> , 90, 1071-1078.	unrepräsentatives Design
Lynch, B. E. (1986). <i>Effects of selected filmic coding elements of TV on the development of the Euclidean concepts of horizontality and verticality in adolescents</i> . Paper presented at the Annual Convention of the Association for Educational Communications and Technology, Las Vegas, NV.	nicht verfügbar
Maher, C., & O'Brien, J. J. (1981). <i>Piagetian operative level related to horizontality and its influence on short-term and long-term memory</i> . Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles.	nicht verfügbar
Mansfield, E. A. (1997). <i>Working memory development in adolescence: A neo-Piagetian investigation</i> . Unpublished doctoral dissertation, University of Northern Colorado.	keine ausreichenden Daten
Maxwell, J. W., Croake, J. W., & Biddle, A. P. (1975). Sex	unrepräsentatives

differences in the comprehension of spatial orientation. <i>Journal of Psychology</i> , 91, 127-131.	Design
McGee, M. G. (1981). Predictions of X-linked inheritance for human characteristics. <i>Developmental Review</i> , 1, 289-295.	keine Water-Level Tasks
McJunkin, M. A. (1991). <i>Elementary children's concepts of the water cycle (water cycle concepts)</i> . Unpublished doctoral dissertation, Oklahoma State University.	keine ausreichenden Daten
Miller, E. F. , & Graybiel, A. (1966). Role of otolith organs in perception of horizontality. <i>American Journal of Psychology</i> , 79, 24-37.	keine Water-Level Tasks
Miller, R., Pascual-Leone, J., Campbell, C., & Juckes, T. (1989). Cross cultural similarities and differences on two neo-Piagetian cognitive tasks. <i>International Journal of Psychology</i> , 24, 293-313.	keine Water-Level Tasks
Molloy, G. N., Boyd, C. P., Lowe, A. R., & Molloy, R. F. (1996). Sex invariance of 13-year-old children on Piagetian task performance. <i>Journal of Genetic Psychology</i> , 157, 239-240.	keine Water-Level Tasks
Munsinger, H. (1974). Most California college women already know that the surface of still water is always horizontal. <i>American Journal of Psychology</i> , 87, 717-718.	unrepräsentatives Design
Murphy-Berman, V., & Witters, L. (1987). <i>Performance of adolescents with social-emotional problems on the Piagetian water line task</i> . Unpublished manuscript.	nicht verfügbar
Murphy-Berman, V., Witters, L., & Harding, R. (1985a). Effect of giftedness, sex, and bottle shape on hearing impaired adolescents' performance on the water line task. <i>Journal for the Education of the Gifted</i> , 8, 273-283.	nicht verfügbar
Murphy-Berman, V., Witters, L., & Harding, R. (1985b). Hearing-impaired students' performance on the Piagetian liquid horizontality test. In D. S. Martin (Ed.), <i>Cognition, education, and deafness</i> (pp. 59-63). Washington, DC: Gallaudet College Press.	keine Daten
Murphy-Berman, V., Witters, L., & Schillerstrom, C. (1988). <i>Learning disabled students' perception of liquid horizontality:</i>	nicht verfügbar

<i>An examination of the effects of gender, development, and training.</i> Unpublished manuscript.	
Newman, J. L., Attig, M., & Kramer, D. A. (1983). Do sex-role appropriate materials influence the Piagetian task performance of older adults? <i>Experimental Aging Research</i> , 9, 197-202.	keine Water-Level Tasks
Odom, R. D., Coon, R. C., & Guzman, R. D. (1970). Detection of an invariant among irrelevant dimensions of different salience value. <i>Perceptual and Motor Skills</i> , 30, 203-206.	keine Water-Level Tasks
Ohuche, N. M. (1984). Performance on the coordinate reference system: Are gender differences universal? <i>Journal of Cross-Cultural Psychology</i> , 15, 285-296.	unrepräsentatives Design
Olson, D. M., & Eliot, J. (1986). Relationships between experiences, processing style, and sex-related differences in performance on spatial tests. <i>Perceptual and Motor Skills</i> , 62, 447-460.	keine Water-Level Tasks
Olson, D. M., Eliot, J., & Hardy, R. C. (1988). Relationships between activities and sex-related differences in performance on spatial tests. <i>Perceptual and Motor Skills</i> , 567, 223-232.	keine Water-Level Tasks
Omari, I. M. (1975). Developmental order of spatial concepts among schoolchildren in Tanzania. <i>Journal of Cross Cultural Psychology</i> , 6, 444-456.	keine ausreichenden Daten
Onyehalu, A. S. (1985). A test of Piaget's horizontal decalage in Nigeria. <i>Perceptual and Motor Skills</i> , 61, 355-361.	keine Water-Level Tasks
Parameswaran, G. (2003). Age, gender, and training in children's performance of Piaget's horizontality task. <i>Educational Studies</i> , 29, 307-319.	Stichprobe vorselektiert (failed pretest)
Pascual-Leone, J. (1969). <i>Cognitive development and cognitive style: A general psychological integration.</i> Unpublished doctoral dissertation, University of Geneva, Switzerland.	unrepräsentatives Design
Pascual-Leone, J. (1992). Field dependence/independence and the water level task: A commentary to Pennings' paper. <i>Perceptual and Motor Skills</i> , 74, 1055-1058.	keine Water-Level Tasks

Pascual-Leone, J., & Goodman, D. (1979). Intelligence and experience: A neo-Piagetian approach. <i>Instructional Science</i> , 8, 301-367.	unrepräsentatives Design
Pascual-Leone, J., & Morra, S. (1991). Horizontality of water level: A neo-Piagetian developmental review. In H. Reese (Ed.), <i>Advances in child development and behavior</i> (vol. 23, pp. 231-276). San Diego, CA: Academic Press.	Review, keine neuen Daten
Perner, J., Kohlmann, R., & Wimmer, H. (1984). Young children's recognition and use of the vertical and horizontal in drawings. <i>Child Development</i> , 55, 1637–1645.	keine ausreichenden Daten
Perper, T., & Chase, J. (1974). Conceptual deficits in women. <i>Science</i> , 184, 809-810.	keine Water-Level Tasks
Pesch, H. G. (1985). Sex variation of eighth-grade students on Piagetian task performance. <i>Journal of Genetic Psychology</i> , 146, 141-142.	keine Water-Level Tasks
Peters, M., Lehmann, W., Takahira, S., Takeuchi, Y., Jordan, K. (2006). Mental rotation test performance in four cross-cultural samples (N=3367): Overall sex differences and the role of academic program in performance. <i>Cortex</i> , 42, 1005-1014.	keine Water-Level Tasks
Piaget, J., & Inhelder, B. (1956). <i>The child's conception of space</i> . (F. J. Langdon & J. L. Lunzer, Trans.). London: Routledge & Kegan Paul. (Original work published 1948).	keine Daten
Piaget, J., & Inhelder, B. (1975). <i>Die Entwicklung des räumlichen Denkens beim Kinde</i> (Gesammelte Werke, Bd. 6, Studienausgabe). Stuttgart: Ernst Klett.	keine Daten
Polak, R. J. (1982). <i>Relationship of obsessive compulsive disorder and compulsive trait to field dependence/independence and personality factors of dependence and anxiety</i> . Unpublished doctoral dissertation, Hofstra University, NY.	Stichprobe klinisch (obsessive compulsive disorder)
Pulos, S. (1992). Adults' understanding of conservation of horizontality. <i>Journal of General Psychology</i> , 119, 99-100.	keine ausreichenden Daten
Pulos, S. (1997). Adolescents' implicit theories of physical	keine Daten

phenomena: A matter of gravity. <i>International Journal of Behavioral Development</i> , 20, 493-507.	
Quaiser-Pohl, C., & Lehmann, W. (2002). Girls' spatial abilities: Charting the contributions of experiences and attitudes in different academic groups. <i>British Journal of Educational Psychology</i> , 72, 245–260.	keine Water-Level Tasks
Quaiser-Pohl, C., Lehmann, W., & Eid, M. (2004). The relationship between spatial abilities and representations of large-scale space in children - a structural equation modeling analysis. <i>Personality and Individual Differences</i> , 36, 95-107.	keine Daten
Randall, T. M. (1977). <i>Training the horizontality concept in a group of non-transitional subjects</i> . Paper presented at the Biennial Meeting of the Society for Research in Child Development, March 17-20 1977, New Orleans, LA.	Stichprobe vorselektiert (failed pretest)
Reith, E., Steffen, C., & Gillieron, C. (1994). Children's drawing of water level: Operatory knowledge, attention to visual image, and depiction skills. <i>Schweizerische Zeitschrift für Psychologie</i> , 53, 86-97.	keine ausreichenden Daten
Rilea, S. L. (2002). <i>Sex differences in spatial ability: A lateralization of function approach</i> . Unpublished doctoral dissertation, University of Alabama.	unrepräsentatives Design
Rittenhouse, R. K. (1978). <i>Horizontal decalage: The development of conservation in deaf students and the effect of the task instructions on their performance</i> . Unpublished doctoral dissertation, University of Illinois.	keine Water-Level Tasks
Robert, M. & Berthiaume, F. (2002). Analyse chronométrique de la resolution de la tâche d'horizontalité es liquids chez des filles et des garçons connaissant ou non l'invariance physique en jeu [A chronometric analysis of water-level performance in boys and girls knowing about the involved physical invariance or not]. <i>Année Psychologique</i> , 102, 657-692	Sprache = Französisch
Robert, M., & Morin, P. (1992). Performing the water-level task in a realistic context is detrimental not only to women. <i>Journal</i>	keine ausreichenden Daten

<i>of Genetic Psychology, 153, 237-241.</i>	
Robert, M., & Ohlmann, T. (1991). <i>Effect of gravitational cues from stimuli and body orientations on water-level representation</i> . Paper presented at the Annual Meeting of the American Psychological Association, Washington, DC.	nicht verfügbar
Robert, M., & Tremblay, S. (1990). Gender, amount and state of liquid and water-level representation [Abstract]. <i>Bulletin of the Psychonomic Society, 28, 498.</i>	kein Volltext verfügbar
Robert, M., & Tremblay, S. (1992). Gender differences in water-level representation as a function of information of state of liquid. <i>Journal of Genetic Psychology, 153, 231-235.</i>	unrepräsentatives Design
Roncadin, C. (1996). <i>A neo-Piagetian analysis of cognitive development in intellectual giftedness</i> . Unpublished doctoral dissertation, York University, Toronto, Canada.	keine ausreichenden Daten
Russell, J., & Maxwell, M. (1980). Children's water-level drawing: A problem of ignoring internal frameworks? <i>Perceptual and Motor Skills, 51, 1001-1002.</i>	keine ausreichenden Daten
Salbod, S. (2001). Correlations between cognitive style and performance on the water-level task by female graduate students. <i>Psychological Reports, 88, 747-748.</i>	Stichprobe rein weiblich
Sanders, G., & Ross-Field, L. (1986). Sexual orientation, cognitive abilities and cerebral asymmetry: A review and a hypothesis tested. <i>Monitore Zoologico Italiano, 20, 459-470.</i>	keine Water-Level Tasks
Saucier, D. M., McCreary, D. R., & Saxberg, J. K. J. (2002). Does gender role socialization mediate sex differences in mental rotations? <i>Personality and Individual Differences, 32, 1101-1111.</i>	keine Water-Level Tasks
Scardamalia, M. (1977). Information processing capacity and the problem of horizontal décalage: A demonstration using combinatorial reasoning tasks. <i>Child Development, 48, 28-37.</i>	keine Water-Level Tasks
Schütz, N. (1990). <i>Die Raumdarstellung in der Kinderzeichnung: Eine empirische Studie an vier- bis sechsjährigen Kindern zur Erklärung der Entwicklungsphänomene [Space</i>	keine Daten

<i>representation in children's drawings: An empirical study in children aged four to six concerning explanation developmental phenomena].</i> Essen: Verlag Die Blaue Eule.	
Schwartz, D. L. (1999). Physical imagery: Kinematic versus dynamic models. <i>Cognitive Psychology</i> , 38, 433-464.	keine Water-Level Tasks
Severtson, E. A. (1994). <i>Development of multiplicative reasoning</i> . Unpublished doctoral dissertation, York University, Toronto, Canada.	Stichprobe vorselektiert
Sheppard, J. L. (1974). The child's concept of horizontality with water levels: A training study. <i>Australian Journal of Psychology</i> , 26, 191-198.	Stichprobe vorselektiert
Sheppard, J. L. (1975). Children's performance on a test of the concept of horizontality, using water levels. <i>Australian Journal of Education</i> , 19, 191-192.	keine ausreichenden Daten
Sherman, J. A. (1967). Problems of sex differences in space perception and aspects of intellectual functioning. <i>Psychological Review</i> , 74, 290-299.	keine Water-Level Tasks
Signorella, M. L., & Jamison, W. (1986). Masculinity, femininity, androgyny, and cognitive performance: A meta-analysis. <i>Psychological Bulletin</i> , 100, 207-228.	Meta-Analyse, keine neuen Daten
Signorella, M. L., & Jamison, W. (1978). Sex differences in the correlations among field dependence, spatial ability, sex role orientation, and performance on Piaget's water level task. <i>Developmental Psychology</i> , 14, 689-690	Stichprobe vorselektiert (failed pretest)
Signorella, M. L., Krupa, M. H., Jamison, W., & Lyons, N. (1986). A short version of the spatial activities questionnaire. <i>Sex Roles</i> , 14, 475-479.	keine ausreichenden Daten
Skakich, S. (1978). <i>M-capacity, strategy, and style in two neo-Piagetian tasks</i> . Unpublished doctoral dissertation, York University, Toronto, Canada.	keine Daten
Smedslund, J. (1963). The effect of observation on children's representation of the spatial orientation of a water surface. <i>Journal of Genetic Psychology</i> , 102, 195-201.	keine ausreichenden Daten
Sneider, C. & Pulos, S. (1983). Children's cosmographies:	keine Water-Level

Understanding the earth's shape and gravity. <i>Science Education</i> , 67, 205-21.	Tasks
Sommerville, D., & Cox, M. V. (1988). Sources of error in the Piagetian water-level task. <i>Perception</i> , 17, 249-254.	unrepräsentatives Design (geneigte Referenzebene bzw. schiefes Gefäß)
Spradlin, L. M., & Hensley, J. H. (1979). An extension of evidence on sex differences in the failure of a water level conservation task by adults. <i>Journal of Genetic Psychology</i> , 135, 305-306.	ident mit McGiffert-Spradlin & Hensley (1979)
Stearns, K., & Borkowski, J. G. (1969). The development of conservation and horizontal-vertical space perception in mental retardates. <i>American Journal of Mental Deficiency</i> , 73, 785-790.	Stichprobe klinisch (mental retardates)
Sukel, K. B. (1984). <i>Relationship of intelligence quotient and chronological age to conservation, visual-motor integration, and Dial+ in gifted and non-gifted 4-, 5-, and 6-year-olds</i> . Unpublished doctoral dissertation, Northern Illinois University.	keine Daten
Thomas, H., & Jamison, W. (1981a). A test of the X-linked genetic hypothesis for sex differences on Piaget's water-level task. <i>Developmental Review</i> , 1, 274-283.	keine Daten
Thomas, H., & Jamison, W. (1981b). Genetic mediation of sex differences in horizontality: A rejoinder. <i>Developmental Review</i> , 12, 596-599.	keine Daten
Thomas, H., Jamison, W., & Hummel, D. D. (1973). Observation is insufficient for discovering that the surface of still water is invariably horizontal. <i>Science</i> , 181, 173-174.	Stichprobe vorselektiert
Thomas, H., Lohaus, A., & Kessler, T. (1999). Stability and change in longitudinal water-level task performance. <i>Developmental Psychology</i> , 35, 1024-1037.	keine ausreichenden Daten
Thomas, H., & Turner, G. F. W. (1991). Individual differences and development in water-level task performance. <i>Journal of Experimental Child Psychology</i> , 51, 171-194.	keine Daten
Thorstad, G. (1986). <i>The construction of the plan of a house test</i>	nicht verfügbar

<i>and an investigation into some of its developmental implications in children.</i> Paper presented at the Annual Conference of the British Psychological Society.	
Tubb, L. G. (1990). <i>Gifted deaf students: Case studies describing profiles of domains of intelligence.</i> Unpublished doctoral dissertation, University of South Florida.	Stichprobe klinisch, vorselektiert (gifted deaf)
Van Esch, L. (1978). <i>The measurement of individual processing capacity.</i> Unpublished doctoral dissertation, York University, Toronto, Canada.	keine Daten
Vasta, R., Belongia, C., & Ribble, C. (1994). Investigating the orientation effect on the water-level task: Who? when? and why? <i>Developmental Psychology</i> , 30, 893-904.	keine ausreichenden Daten
Vasta, R., & Liben, L. S. (1996). The water-level task: An intriguing puzzle. <i>Current Directions in Psychological Science</i> , 5, 171-177.	keine Daten
Vasta, R., Rosenberg, D., Knott, A., & Gaze, C. E. (1997). Experience and the water-level task revisited: Does expertise exact a price? <i>Psychological Science</i> , 8, 336-339.	keine ausreichenden Daten
Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. <i>Psychological Bulletin</i> , 117, 250-270.	Meta-Analyse, keine neuen Daten
Wegesin, D. J. (1998b). Relation between language lateralisation and spatial ability in gay and straight men and women. <i>Laterality: Asymmetries of Body, Brain, and Cognition</i> , 3, 227-239.	Daten aus Wegesin, (1998a). <i>Archives of Sexual Behavior</i> , 27, 91-108.
Weinstein, M. S. (1970). <i>Effects of training on the concepts of water level and horizontality in the classroom</i> (Report No. TR-128). Madison, WI: Research and Development Center for Cognitive Learning.	keine ausreichenden Daten
Willemssen, E. W. (1974). Judgment of the vertical as a function of stimulus characteristics, age, and sex. <i>Perceptual and Motor Skills</i> , 38, 1344.	keine Water-Level Tasks
Willemssen, E., Buchholz, A., Budrow, M. S., & Geannacopulos, N.	Stichprobe rein

(1973). Relationship between Witkin's Rod-and-Frame task and Piaget's water-line task for college women. <i>Perceptual and Motor Skills</i> , 36, 958.	weiblich
Willemsen, E., & Reynolds, B. (1973). Sex differences in adults' judgments of the horizontal. <i>Developmental Psychology</i> , 8, 309.	keine ausreichenden Daten
Witters-Churchill, L. J., Kelly, R. R., & Witters, L. A. (1983). Hearing-impaired students' perception of liquid horizontality: An examination of the effects of gender, development, and training. <i>Volta Review</i> , 85, 211-225.	Stichprobe rein weiblich, klinisch
Witters-Churchill, L. J., & Witters, L. A. (1989). Hearing-impaired students' performance on the Piagetian liquid horizontality test: An analysis and synthesis. In D. S. Martin (Ed.), <i>International Symposium on Cognition, Education, and Deafness, July 5-8 1989, Washington, DC</i> (Working papers, Vol. 2, pp. 614-637).	Meta-Analyse, keine neuen Daten

Appendix E: Beschreibung inkludierter Studien

Tabelle E. 1: Beschreibung in die Meta-Analyse inkludierter Studien

Studie	Stichprobe ^a	Publ ^b	Altersgr ^c	Gefäß ^d	Antwort ^e	Kriterium ^f	n(m) ^g	n(f) ^h	d ⁱ
Abravanel & Gingold (1977)	experiment 1	p	Ki	zyl	Zei	groß	26	24	0,197
	experiment 2	p	Ki	recht	Zei	groß	24	26	0,416
	experiment 3	p	Ki	zyl	Zei	groß	14	16	-0,010
Amponsah (2000)	Ghana, 9 yrs	p	Ki	recht	MC	MC	126	123	0,335
	Norway, 9 yrs	p	Ki	recht	MC	MC	94	99	0,327
	Ghana, 12 yrs	p	Jug	recht	MC	MC	149	117	0,463
	Norway, 12 yrs	p	Jug	recht	MC	MC	143	154	0,690
	Ghana, 16 yrs	p	Jug	recht	MC	MC	176	127	0,475
	Norway, 16 yrs	p	Jug	recht	MC	MC	139	171	0,790
	Ghana, adults	p	Erw	recht	MC	MC	206	184	0,631
	Norway, adults	p	Erw	recht	MC	MC	154	276	0,594
Beatty & Duncan (1990)	whole sample	p	Erw	recht	Zei	klein	59	61	0,545
Beisteiner (2001)	whole sample	np	gem	recht	Zei	klein	562	460	0,277
Casey et al. (2001)	whole sample	p	Jug	recht	Zei	klein	91	96	0,670
DeAvila & Havassy (1974)	whole sample	p	Ki	recht	Zei	kA	587	593	0,026
Dessens et al. (1998)	controls	p	Erw	recht	Zei	and	70	72	0,629
Formann (2003)	whole sample	p	gem	zyl	Zei	klein	227	204	0,176
Geiringer (1976)	middle aged	np	Erw	Koch	Anp	Abw	30	30	0,624
	young elderly	np	Erw	Koch	Anp	Abw	20	20	0,465
	mature elderly	np	Erw	Koch	Anp	Abw	20	20	0,554
Geiringer & Hyde (1976)	5th grade	p	Ki	Koch	Anp	Abw	30	30	0,512
	12th grade	p	Jug	Koch	Anp	Abw	30	30	0,424
Gladue et al. (1990)	heterosexual group	p	Erw	recht	Zei	Abw	16	15	0,815
Golbeck (1986)	whole sample	p	Erw	recht	Zei	klein	32	32	0,721
Golbeck & Sinagra (2000)	group "alone"	p	Erw	recht	Zei	Abw	6	19	0,109
Hammer et al. (1995)	major: liberal arts	p	Erw	recht	Zei	groß	27	27	1,837
	major: architectural	p	Erw	recht	Zei	groß	27	27	0,000

Tabelle E.1, Fortsetzung 1: Beschreibung in die Meta-Analyse inkludierter Studien

Studie	Stichprobe ^a	Publ ^b	Altersgr ^c	Gefäß ^d	Antwort ^e	Kriterium ^f	n(m) ^g	n(f) ^h	d ⁱ
Harris et al. (1977)	experiment 1: adults	p	Erw	gem	MC	MC	48	110	0,953
	experiment 2: adults	p	Erw	gem	MC	MC	53	105	0,854
	experiment 3	p	Erw	gem	MC	MC	45	65	0,334
	experiment 4	p	Erw	gem	MC	MC	14	55	0,036
	experiment 5	p	Erw	gem	MC	MC	58	52	1,544
	experiment 1	p	Ki	gem	MC	MC	33	28	0,812
Hecht & Proffitt (1995)	graduate students	p	Erw	recht	Ze	klein	20	20	0,423
	group "experienced with liquids"	p	Erw	recht	Ze	klein	20	20	0,382
	group "non-experienced with liquids"	p	Erw	recht	Ze	klein	20	20	1,073
Kalichman (1986)	major: business	p	Erw	gem	Ze	groß	25	25	1,231
	major: language arts	p	Erw	gem	Ze	groß	25	25	1,047
	major: social sciences	p	Erw	gem	Ze	groß	25	25	1,271
	major: natural sciences	p	Erw	gem	Ze	groß	25	25	1,123
	major: engineering	p	Erw	gem	Ze	groß	25	25	1,670
Kalichman (1989a)	whole sample	p	Erw	recht	Ze	groß	97	97	0,788
Kalichman (1989b)	whole sample	p	Erw	gem	Ze	groß	52	52	0,522
Kalichman (1991)	whole sample	p	Erw	recht	Ze	groß	38	38	1,071
Kelly & Kelly (1977)	whole sample	p	Erw	recht	Ze	klein	123	191	0,548
Kenyon (1984)	whole sample	p	Erw	zyl	Ze	kA	15	15	1,383
Krasnoff et al. (1989)	whole sample	p	Erw	recht	Ze	Abw	40	40	0,846
Krasnoff & Walker (1983)	whole sample	p	Erw	recht	Ze	Abw	115	98	0,373
Lawton (1994)	whole sample	p	Erw	gem	MC	MC	138	288	0,807
Levine & Stern (2002)	whole sample	p	Erw	recht	Ze	Abw	32	57	0,616
Li et al. (1999a)	whole sample	p	gem	recht	Ze	klein	256	230	0,281
Li et al. (1999b)	group "BC" (Beijing, Chinese)	p	Erw	recht	Ze	klein	107	188	0,745
	group "CCA" (Writing Chinese, US)	p	Erw	recht	Ze	klein	21	28	0,103
	group "NCA" (Non-writing Chinese, US)	p	Erw	recht	Ze	klein	60	69	0,024

Tabelle E.1, Fortsetzung 2: Beschreibung in die Meta-Analyse inkludierter Studien

Studie	Stichprobe ^a	Publ ^b	Altersgr ^c	Gefäß ^d	Antwort ^e	Kriterium ^f	n(m) ^g	n(f) ^h	d ⁱ
Liben & Golbeck (1984)	experiment 1: group "no rule, no example"	p	Erw	recht	Ze	klein	20	20	0,878
	experiment 1: group "no rule, with example"	p	Erw	recht	Ze	klein	20	20	0,631
	experiment 2: "no rule" condition	p	Erw	recht	Ze	klein	40	40	0,854
Liben & Golbeck (1986)	physical condition	p	Erw	recht	Ze	klein	20	20	0,697
Liben (1974)	group 1, "H"	p	Ki	recht	Ze	groß	27	23	0,510
	group 2, "NH"	p	Ki	recht	Ze	groß	20	24	0,934
Liben (1991)	experiment 1	p	Erw	recht	Ze	klein	50	54	0,490
	experiment 2	p	Erw	recht	Ze	klein	20	20	1,024
Linn & Pulos (1983)	whole sample	p	Jug	recht	Ze	kA	403	375	0,376
Loehlin & McFadden (2003)	heterosexual group	p	Erw	recht	Ze	kA	56	61	0,322
Lohaus et al. (1996)	grade 2	p	Ki	recht	Ze	groß	96	106	0,202
	grade 5	p	Jug	recht	Ze	groß	104	90	0,470
	grade 8	p	Jug	recht	Ze	groß	110	92	0,794
McAfee & Proffitt (1991)	experiment 1	p	Erw	recht	Ze	klein	214	228	0,687
	experiment 3	p	Erw	gem	Ze	and	34	19	0,310
McCormick & Witelson (1991)	heterosexuals, consistent right-hand-preference	p	Erw	recht	Ze	klein	38	38	1,056
McGiffert-Spradlin & Higgins-Hensley (1979)	whole sample	p	Erw	recht	Ze	kA	60	60	0,691
McGillicuddy-DeLisi et al. (1978)	public college	p	Erw	recht	Ze	klein	10	10	0,979
Meehan & Overton (1986)	experiment 1	p	Erw	recht	Ze	groß	42	42	0,531
	experiment 2: controls	p	Erw	recht	Ze	klein	12	18	0,567
Merriwether & Liben (1997)	whole sample	p	Erw	recht	Ze	groß	80	80	0,463
Morris (1971)	group 1	p	Erw	recht	Ze	kA	32	32	0,442
	group 2	p	Erw	recht	Ze	Abw	32	32	1,101
Mowrer-Popiel (1990)	whole sample	np	Erw	kug	Ze	Abw	70	70	0,305
Mowrer-Popiel & DeLisi (1984)	whole sample	p	Erw	kA	Ze	klein	39	86	0,493
Murphy-Berman et al. (1986)	vessel: curved bottle	p	Ki	Wein	Ze	klein	7	8	1,191
	vessel: straight sided bottle	p	Ki	recht	Ze	klein	7	8	0,730

Tabelle E.1, Fortsetzung 3: Beschreibung in die Meta-Analyse inkludierter Studien

Studie	Stichprobe ^a	Publ ^b	Altersgr ^c	Gefäß ^d	Antwort ^e	Kriterium ^f	n(m) ^g	n(f) ^h	d ⁱ
Myer & Higgins-Hensley (1984)	whole sample	p	Erw	recht	Zei	klein	41	44	0,554
Neumann et al. (2007)	whole sample	p	Erw	Tas	Zei	Abw	32	32	0,606
Otani & Leonard (1988)	experiment 1	p	Erw	gem	Zei	groß	54	99	0,440
Parameswaran (1995)	pretest	p	Erw	recht	Zei	kA	35	87	0,413
Parameswaran & DeLisi (1996)	whole sample	p	Erw	recht	Zei	kA	35	87	0,413
Pennings (1991)	group "R", pretest	p	Ki	gem	Zei	Abw	8	8	0,741
	group "C", pretest	p	Ki	gem	Zei	Abw	8	8	0,651
	controls, pretest	p	Ki	gem	Zei	Abw	8	8	0,941
Pulos (1997)	"low-science" group	p	Erw	recht	Zei	klein	40	40	0,324
	"high-science" group	p	Erw	recht	Zei	klein	40	40	0,256
Quaiser-Pohl (2001)	whole sample	p	Ki	recht	Zei	and	218	220	0,394
Ray et al. (1979)	whole sample	p	Erw	recht	Zei	kA	50	85	0,306
Rebelsky (1964)	whole sample	p	Erw	recht	Zei	klein	59	69	0,148
Robert (1990)	"younger men & women"	p	Erw	gem	Zei	klein	59	225	0,733
Robert & Chaperon (1989)	"selectionphase"-sample	p	Erw	zyl	Zei	klein	488	1032	0,542
Robert et al. (1994)	wlt condition, "visual standard" group	p	Erw	recht	Zei	klein	60	72	1,174
Robert & Harel (1996)	major: natural & applied sciences; vessel: bottles	p	Erw	gem	Zei	klein	290	230	0,671
	major: social sciences; vessel: bottles	p	Erw	gem	Zei	klein	200	266	0,502
	major: humanities & arts; vessel: bottles	p	Erw	gem	Zei	klein	194	232	0,657
Robert & Héroux (2004)	9 yrs	p	Ki	recht	Zei	klein	20	17	0,128
	12 yrs	p	Jug	recht	Zei	klein	19	22	0,210
	15 yrs	p	Jug	recht	Zei	klein	18	18	0,221
Robert & Longpré (2005)	control condition, block 1	p	Erw	recht	Zei	klein	40	40	0,949
Robert & Morin (1993)	"standard tasks"	p	Erw	recht	kA	klein	37	62	0,516
Robert & Ohlmann (1994)	control condition	p	Erw	recht	Zei	klein	140	140	0,721
Robert & Tanguay (1990)	40 - 61 yrs	p	Erw	gem	Zei	klein	17	15	0,405
	62 - 72 yrs	p	Erw	gem	Zei	klein	13	15	0,364
	73 - 84 yrs	p	Erw	gem	Zei	klein	15	15	1,292
Roberts & Bell (2001)	whole sample	p	gem	recht	Zei	groß	32	32	0,606
Sanders & Ross-Field (1986)	experiment 1: heterosexual group	p	Erw	recht	Zei	Abw	8	8	1,714

Tabelle E.1, Fortsetzung 4: Beschreibung in die Meta-Analyse inkludierter Studien

Studie	Stichprobe ^a	Publ ^b	Altersgr ^c	Gefäß ^d	Antwort ^e	Kriterium ^f	n(m) ^g	n(f) ^h	d ⁱ
Seng & Tan (2002)	Chinese	np	Ki	recht	Ze	klein	31	26	0,350
	Malays	np	Ki	recht	Ze	klein	23	15	0,347
	nationality: „Others“	np	Ki	recht	Ze	klein	2	3	1,006
Sholl (1989)	initial screening sample	p	Erw	gem	Ze	klein	161	316	0,446
Sholl & Liben (1995)	experiment 2: mass screening sample	p	Erw	gem	Ze	klein	52	125	0,693
	experiment 3: mass screening sample	p	Erw	Wein	Ze	klein	203	466	0,536
	experiment 4: mass screening sample	p	Erw	Wein	Ze	klein	379	523	0,729
Signorella et al. (1989)	whole sample	p	Erw	recht	Ze	klein	198	146	0,533
Tegano & Faulkender (1983)	whole sample	p	Erw	kA	Ze	kA	42	71	0,808
Thomas & Jamison (1975)	experiment 3	p	Erw	recht	Anp	Abw	62	72	0,736
Thomas & Lohaus (1993)	experiment 1: 6-7 yrs	p	Ki	recht	Ze	and	18	20	0,610
	experiment 1: 8 yrs	p	Ki	recht	Ze	and	30	25	0,531
	experiment 1: 9 yrs	p	Ki	recht	Ze	and	14	21	0,203
	experiment 1: 10 yrs	p	Ki	recht	Ze	and	21	12	0,576
	experiment 1: 11 yrs	p	Ki	recht	Ze	and	24	39	0,489
	experiment 1: 12 yrs	p	Jug	recht	Ze	and	43	38	0,884
	experiment 1: 13 yrs	p	Jug	recht	Ze	and	46	42	0,422
	experiment 1: 14 yrs	p	Jug	recht	Ze	and	31	27	0,258
	experiment 1: 15 yrs	p	Jug	recht	Ze	and	28	39	0,279
	experiment 1: 16-17 yrs	p	Jug	recht	Ze	and	26	22	0,055
	experiment 2: 9-10 yrs	p	Ki	kA	Anp	and	16	16	0,796
	experiment 2: 11-12 yrs	p	Ki	kA	Anp	and	20	16	0,425
	experiment 2: 13-14 yrs	p	Jug	kA	Anp	and	31	22	0,539
	experiment 2: 15-17 yrs	p	Jug	kA	Anp	and	30	22	0,413
Thompson et al. (1981)	whole sample	p	Erw	Krug	Ze	klein	38	98	0,522
Tran & Formann (2008)	adults	p	Erw	Koch	Ze	klein	542	672	0,592
	adolescents	p	Jug	Koch	Ze	klein	156	140	0,006
	children	p	Ki	Koch	Ze	klein	76	71	0,243
Treagust (1980)	whole sample	p	Jug	recht	Ze	kA	54	54	0,859
Vasta et al. (1993)	"physikal task"	p	Erw	recht	MC	MC	48	56	0,652
Vasta et al. (1996)	whole sample	p	Erw	recht	Ze	Abw	84	96	0,016

Tabelle E.1, Fortsetzung 5: Beschreibung in die Meta-Analyse inkludierter Studien

Studie	Stichprobe ^a	Publ ^b	Altersgr ^c	Gefäß ^d	Antwort ^e	Kriterium ^f	$n(m)$ ^g	$n(f)$ ^h	d ⁱ
Voyer & Sullivan (2003)	experiment 2: timed condition, group testing	p	Erw	recht	Ze	klein	55	85	0,782
	experiment 2: timed condition, individual testing	p	Erw	recht	Ze	klein	51	68	0,453
	experiment 2: untimed condition, group testing	p	Erw	recht	Ze	klein	90	105	0,582
	experiment 2: untimed condition, individual testing	p	Erw	recht	Ze	klein	21	27	1,155
Voyer (1996)	experiment 3	p	Erw	recht	Ze	klein	71	54	0,634
Voyer (2000)	whole sample	p	Erw	gem	Ze	klein	146	198	0,571
Walker & Krasnoff (1978)	whole sample	p	Erw	gem	Ze	Abw	44	44	0,439
Walker et al. (1981)	adults	p	Erw	recht	Ze	Abw	62	67	0,152
	children	p	Jug	recht	Ze	Abw	63	66	0,668
Wegesin (1995)	whole sample	np	Erw	recht	Ze	klein	40	40	0,731
Wegesin (1998)	heterosexual group	p	Erw	recht	Ze	klein	20	20	0,727
Wittig & Allen (1984)	whole sample	p	Erw	gem	gem	and	74	121	0,573

- a) Stichprobe: Beschreibung der in die Analyse eingegangenen unabhängigen Stichprobe innerhalb der genannten Studie
- b) Publikationsstatus der Arbeit: p = publiziert; np = nicht publiziert
- c) Altersgruppe der Stichprobe: Ki = Kinder; Jug = Jugendliche; Erw = Erwachsene; gem = gemischte Altersgruppe
- d) Gefäßform im Test: recht = rechteckig; zyl = zylindrisch; kug = kugelförmig; Wein = Weinflasche; Tas = Tasse; Koch = Kochflasche; Krug = Krug; gem = gemischte Formen; kA = keine Angabe
- e) Antwortmodus: Zei = Zeichnen des Wasserspiegels; Anp = Anpassen des Wasserspiegels anhand eines Modells; MC = Auswahl aus gegebenen Antwortalternativen; gem = gemischte Antwortmodi; kA = keine Angabe
- f) Kriterium: Größe des maximal erlaubten Abweichungsfehlers in Grad; klein = Kriterium zwischen 3° und 5°; groß = Kriterium zwischen 6° und 15°; Abw = Auswertung des mittleren Abweichungsfehlers; MC = Auswertung im Multiple-Choice-Format; and = andere, nicht klassifizierbare Auswertungskriterien; kA = keine Angabe
- g) $n(m)$: Anzahl männlicher Testpersonen je Stichprobe
- h) $n(f)$: Anzahl weiblicher Testpersonen je Stichprobe
- i) d : resultierendes Cohen's d bezüglich Geschlechtseffekt in der Testleistung (positives d = bessere Leistung männlicher Testpersonen; negatives d = bessere Leistung weiblicher Testpersonen)

Appendix F: Programmausgaben der Subgruppenanalysen

Tabelle F.1: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Publikationsstatus

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
not published	8	0,399	0,057	0,003	0,286	0,511	6,953	0,000	16,577	7	0,020	57,773	0,055	0,058	0,003	0,235
published	146	0,512	0,014	0,000	0,485	0,540	36,557	0,000	341,565	145	0,000	57,548	0,039	0,010	0,000	0,198
Total within									358,142	152	0,000					
Total between									3,694	1	0,055					
Random effects analysis																
not published	8	0,556	0,117	0,014	0,328	0,785	4,767	0,000								
published	146	0,548	0,025	0,001	0,500	0,596	22,344	0,000								
Total between									0,005	1	0,943					

Tabelle F.2: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Geschlecht des Erstautors

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
unknown	5	0,758	0,157	0,025	0,451	1,066	4,836	0,000	3,355	4	0,500	0,000	0,000	0,098	0,010	0,000
female	75	0,548	0,019	0,000	0,511	0,585	28,838	0,000	147,727	74	0,000	49,908	0,028	0,011	0,000	0,166
male	74	0,456	0,020	0,000	0,418	0,495	23,237	0,000	196,853	73	0,000	62,916	0,050	0,017	0,000	0,223
Total within									347,935	151	0,000					
Total between									13,901	2	0,001					
Random effects analysis																
unknown	5	0,758	0,157	0,025	0,451	1,066	4,836	0,000								
female	75	0,572	0,031	0,001	0,511	0,632	18,509	0,000								
male	74	0,518	0,037	0,001	0,447	0,590	14,196	0,000								
Total between									2,985	2	0,225					

Tabelle F.3: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Altersgruppe

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
adolescents	21	0,485	0,034	0,001	0,419	0,552	14,316	0,000	46,620	20	0,001	57,100	0,034	0,021	0,000	0,184
adults	100	0,601	0,017	0,000	0,567	0,635	34,613	0,000	168,040	99	0,000	41,085	0,021	0,008	0,000	0,146
children	29	0,255	0,035	0,001	0,185	0,324	7,185	0,000	40,954	28	0,054	31,630	0,020	0,019	0,000	0,140
mixed	4	0,269	0,049	0,002	0,173	0,365	5,494	0,000	2,508	3	0,474	0,000	0,000	0,009	0,000	0,000
Total within									258,122	150	0,000					
Total between									103,714	3	0,000					
Random effects analysis																
adolescents	21	0,498	0,057	0,003	0,385	0,610	8,685	0,000								
adults	100	0,606	0,025	0,001	0,556	0,655	23,857	0,000								
children	29	0,363	0,054	0,003	0,256	0,469	6,687	0,000								
mixed	4	0,269	0,049	0,002	0,173	0,365	5,494	0,000								
Total between									45,661	3	0,000					

Tabelle F.4: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Verfügbarkeit von Musteritems

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
9999 = missing	33	0,422	0,025	0,001	0,373	0,472	16,790	0,000	137,507	32	0,000	76,728	0,072	0,032	0,001	0,269
more than 1	6	0,569	0,050	0,003	0,470	0,667	11,322	0,000	5,956	5	0,311	16,049	0,003	0,013	0,000	0,057
no practice	55	0,512	0,025	0,001	0,463	0,560	20,800	0,000	126,573	54	0,000	57,337	0,047	0,021	0,000	0,217
one practice	60	0,561	0,024	0,001	0,514	0,607	23,586	0,000	73,829	59	0,093	20,086	0,009	0,008	0,000	0,093
Total within									343,865	150	0,000					
Total between									17,971	3	0,000					
Random effects analysis																
9999 = missing	33	0,543	0,058	0,003	0,429	0,658	9,327	0,000								
more than 1	6	0,559	0,057	0,003	0,447	0,672	9,769	0,000								
no practice	55	0,558	0,045	0,002	0,471	0,646	12,472	0,000								
one practice	60	0,557	0,029	0,001	0,501	0,613	19,506	0,000								
Total between									0,055	3	0,997					

Tabelle F.5: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Gefäßform

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
9999 = missing	6	0,591	0,104	0,011	0,386	0,796	5,655	0,000	2,387	5	0,793	0,000	0,000	0,043	0,002	0,000
cup	1	0,606	0,256	0,065	0,105	1,107	2,369	0,018	0,000	0	1,000	0,000	0,000	0,000	0,000	0,000
curved/wine	3	0,655	0,062	0,004	0,534	0,776	10,594	0,000	2,435	2	0,296	17,878	0,003	0,020	0,000	0,059
cylindrical	5	0,414	0,061	0,004	0,295	0,533	6,829	0,000	15,090	4	0,005	73,493	0,074	0,094	0,009	0,272
florence flask	8	0,460	0,047	0,002	0,369	0,551	9,877	0,000	22,428	7	0,002	68,789	0,058	0,058	0,003	0,241
mixed shapes	30	0,657	0,033	0,001	0,592	0,722	19,859	0,000	52,004	29	0,005	44,235	0,027	0,018	0,000	0,165
pitcher	1	0,522	0,194	0,038	0,142	0,901	2,693	0,007	0,000	0	1,000	0,000				0,000
rectangular	99	0,465	0,017	0,000	0,431	0,499	26,832	0,000	229,777	98	0,000	57,350	0,041	0,012	0,000	0,202
spherical bottle	1	0,305	0,170	0,029	-0,029	0,638	1,792	0,073		0,000	1,000	0,000				0,000
Total within									324,122	145	0,000					
Total between									37,714	8	0,000					
Random effects analysis																
9999 = missing	6	0,591	0,104	0,011	0,386	0,796	5,655	0,000								
cup	1	0,606	0,256	0,065	0,105	1,107	2,369	0,018								
curved/wine	3	0,649	0,075	0,006	0,502	0,796	8,652	0,000								
cylindrical	5	0,400	0,159	0,025	0,088	0,711	2,513	0,012								
florence flask	8	0,400	0,113	0,013	0,177	0,622	3,522	0,000								
mixed shapes	30	0,677	0,051	0,003	0,577	0,777	13,313	0,000								
pitcher	1	0,522	0,194	0,038	0,142	0,901	2,693	0,007								
rectangular	99	0,529	0,030	0,001	0,471	0,588	17,695	0,000								
spherical bottle	1	0,305	0,170	0,029	-0,029	0,638	1,792	0,073								
Total between									13,197	8	0,105					

Tabelle F.6: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Vorgabemodus

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
9999 = missing	5	0,520	0,126	0,016	0,274	0,767	4,146	0,000	0,792	4	0,940	0,000	0,000	0,058	0,003	0,000
drawing	139	0,503	0,014	0,000	0,476	0,531	36,200	0,000	351,000	138	0,000	60,684	0,042	0,011	0,000	0,205
mixed	1	0,573	0,150	0,023	0,278	0,868	3,811	0,000	0,000	0	1,000	0,000	0,000	0,000	0,000	0,000
PC	1	0,606	0,256	0,065	0,105	1,107	2,369	0,018	0,000	0	1,000	0,000	0,000	0,000	0,000	0,000
photograph	2	0,377	0,272	0,074	-0,156	0,911	1,386	0,166	7,434	1	0,006	86,548	1,461	2,387	5,696	1,209
real vessel	6	0,589	0,103	0,011	0,387	0,791	5,714	0,000	1,339	5	0,931	0,000	0,000	0,042	0,002	0,000
Total within									360,565	148	0,000					
Total between									1,271	5	0,938					
Random effects analysis																
9999 = missing	5	0,520	0,126	0,016	0,274	0,767	4,146	0,000								
drawing	139	0,548	0,025	0,001	0,498	0,598	21,648	0,000								
mixed	1	0,573	0,150	0,023	0,278	0,868	3,811	0,000								
PC	1	0,606	0,256	0,065	0,105	1,107	2,369	0,018								
photograph	2	0,846	0,916	0,838	-0,949	2,640	0,924	0,356								
real vessel	6	0,589	0,103	0,011	0,387	0,791	5,714	0,000								
Total between									0,380	5	0,996					

Tabelle F.7: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Einbettung des Gefäßes in situativen Kontext

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
9999 = missing	5	0,555	0,091	0,008	0,377	0,733	6,125	0,000	0,854	4	0,931	0,000	0,000	0,044	0,002	0,000
no context	142	0,501	0,014	0,000	0,474	0,529	35,852	0,000	355,018	141	0,000	60,284	0,043	0,011	0,000	0,207
context	7	0,608	0,078	0,006	0,455	0,762	7,764	0,000	3,850	6	0,697	0,000	0,000	0,029	0,001	0,000
Total within									359,721	151	0,000					
Total between									2,115	2	0,347					
Random effects analysis																
9999 = missing	5	0,555	0,091	0,008	0,377	0,733	6,125	0,000								
no context	142	0,547	0,025	0,001	0,498	0,597	21,619	0,000								
context	7	0,608	0,078	0,006	0,455	0,762	7,764	0,000								
Total between									0,553	2	0,758					

Tabelle F.8: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Verfügbarkeit eines realen Modells

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
9999 = missing	10	0,406	0,057	0,003	0,294	0,519	7,080	0,000	9,689	9	0,376	7,112	0,003	0,018	0,000	0,053
mixed	1	0,573	0,150	0,023	0,278	0,868	3,811	0,000	0,000	0	1,000	0,000	0,000	0,000	0,000	0,000
no real model	120	0,506	0,015	0,000	0,476	0,536	32,879	0,000	323,148	119	0,000	63,175	0,050	0,014	0,000	0,223
real model tilted, covered	18	0,549	0,038	0,001	0,475	0,623	14,563	0,000	18,761	17	0,342	9,384	0,003	0,010	0,000	0,052
real model taken away	5	0,467	0,088	0,008	0,294	0,640	5,295	0,000	5,521	4	0,238	27,554	0,017	0,044	0,002	0,130
Total within									357,119	149	0,000					
Total between									4,717	4	0,318					
Random effects analysis																
9999 = missing	10	0,411	0,061	0,004	0,290	0,531	6,690	0,000								
mixed	1	0,573	0,150	0,023	0,278	0,868	3,811	0,000								
no real model	120	0,567	0,029	0,001	0,510	0,625	19,402	0,000								
real model tilted, covered	18	0,545	0,041	0,002	0,465	0,625	13,370	0,000								
real model taken away	5	0,472	0,112	0,012	0,253	0,691	4,228	0,000								
Total between									5,742	4	0,219					

Tabelle F.9: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Verfügbarkeit einer horizontalen Referenz

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
9999 = missing	48	0,456	0,026	0,001	0,404	0,507	17,385	0,000	135,742	47	0,000	65,375	0,066	0,029	0,001	0,257
no horizontal reference	31	0,579	0,031	0,001	0,518	0,640	18,624	0,000	45,862	30	0,032	34,587	0,016	0,013	0,000	0,127
horizontal reference	75	0,505	0,019	0,000	0,468	0,541	27,225	0,000	171,034	74	0,000	56,734	0,035	0,012	0,000	0,186
Total within									352,638	151	0,000					
Total between									9,198	2	0,010					
Random effects analysis																
9999 = missing	48	0,610	0,051	0,003	0,510	0,710	11,945	0,000								
no horizontal reference	31	0,573	0,042	0,002	0,491	0,654	13,755	0,000								
horizontal reference	75	0,509	0,033	0,001	0,445	0,573	15,523	0,000								
Total between									3,255	2	0,196					

Tabelle F.10: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Antwortmodus

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
9999 = missing	1	0,516	0,211	0,044	0,102	0,929	2,445	0,014	0,000	0	1,000	0,000	0,000	0,000	0,000	0,000
adjust w/L	10	0,569	0,086	0,007	0,400	0,738	6,613	0,000	2,254	9	0,987	0,000	0,000	0,036	0,001	0,000
draw w/L	126	0,481	0,015	0,000	0,451	0,510	31,874	0,000	296,828	125	0,000	57,888	0,040	0,011	0,000	0,200
MC	16	0,628	0,035	0,001	0,559	0,697	17,786	0,000	47,323	15	0,000	68,303	0,044	0,025	0,001	0,210
mixed	1	0,573	0,150	0,023	0,278	0,868	3,811	0,000	0,000	0	1,000	0,000	0,000	0,000	0,000	0,000
Total within									346,406	149	0,000					
Total between									15,430	4	0,004					
Random effects analysis																
9999 = missing	1	0,516	0,211	0,044	0,102	0,929	2,445	0,014								
adjust w/L	10	0,569	0,086	0,007	0,400	0,738	6,613	0,000								
draw w/L	126	0,531	0,027	0,001	0,478	0,584	19,750	0,000								
MC	16	0,640	0,066	0,004	0,509	0,770	9,619	0,000								
mixed	1	0,573	0,150	0,023	0,278	0,868	3,811	0,000								
Total between									2,398	4	0,663					

Tabelle F.11: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Auswertungskriterium

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
3°-5°	65	0,528	0,019	0,000	0,490	0,565	27,685	0,000	130,782	64	0,000	51,064	0,025	0,010	0,000	0,159
6°-15°	22	0,567	0,050	0,002	0,469	0,664	11,412	0,000	40,299	21	0,007	47,890	0,051	0,035	0,001	0,227
9999 = missing	11	0,268	0,039	0,002	0,192	0,344	6,898	0,000	46,119	10	0,000	78,317	0,077	0,065	0,004	0,277
deviation	22	0,479	0,050	0,002	0,382	0,577	9,617	0,000	33,771	21	0,038	37,816	0,034	0,029	0,001	0,185
MC	16	0,628	0,035	0,001	0,559	0,697	17,786	0,000	47,323	15	0,000	68,303	0,044	0,025	0,001	0,210
other	18	0,466	0,052	0,003	0,364	0,567	8,998	0,000	10,549	17	0,879	0,000	0,000	0,018	0,000	0,000
Total within									308,843	148	0,000					
Total between									52,993	5	0,000					
Random effects analysis																
3°-5°	65	0,541	0,032	0,001	0,479	0,603	17,152	0,000								
6°-15°	22	0,607	0,075	0,006	0,460	0,754	8,098	0,000								
9999 = missing	11	0,481	0,104	0,011	0,277	0,684	4,629	0,000								
deviation	22	0,522	0,068	0,005	0,390	0,655	7,720	0,000								
MC	16	0,640	0,066	0,004	0,509	0,770	9,619	0,000								
other	18	0,466	0,052	0,003	0,364	0,567	8,998	0,000								
Total between									5,481	5	0,360					

Tabelle F.12: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Ergebnismodus

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
# successful SS	36	0,541	0,028	0,001	0,485	0,596	19,144	0,000	80,611	35	0,000	56,582	0,039	0,021	0,000	0,199
9999 = missing	3	0,648	0,099	0,010	0,454	0,841	6,560	0,000	0,888	2	0,641	0,000	0,000	0,030	0,001	0,000
mean # correct items	79	0,555	0,019	0,000	0,519	0,592	29,846	0,000	137,444	78	0,000	43,250	0,021	0,009	0,000	0,146
mean deviation	23	0,478	0,049	0,002	0,382	0,574	9,773	0,000	33,792	22	0,052	34,895	0,030	0,027	0,001	0,174
other	13	0,248	0,037	0,001	0,176	0,320	6,726	0,000	49,109	12	0,000	75,565	0,065	0,048	0,002	0,254
Total within									301,844	149	0,000					
Total between									59,992	4	0,000					
Random effects analysis																
# successful SS	36	0,606	0,051	0,003	0,505	0,706	11,829	0,000								
9999 = missing	3	0,648	0,099	0,010	0,454	0,841	6,560	0,000								
mean # correct items	79	0,558	0,028	0,001	0,503	0,612	20,044	0,000								
mean deviation	23	0,516	0,065	0,004	0,389	0,642	7,997	0,000								
other	13	0,374	0,091	0,008	0,196	0,552	4,119	0,000								
Total between									6,198	4	0,185					

Tabelle F.13: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Testsituation

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
9999 = missing	21	0,498	0,048	0,002	0,404	0,592	10,389	0,000	31,081	20	0,054	35,653	0,028	0,025	0,001	0,166
group	79	0,514	0,017	0,000	0,481	0,548	30,289	0,000	224,743	78	0,000	65,294	0,044	0,013	0,000	0,209
individual	46	0,561	0,036	0,001	0,490	0,633	15,387	0,000	56,911	45	0,110	20,929	0,016	0,017	0,000	0,128
mixed	8	0,414	0,037	0,001	0,342	0,486	11,305	0,000	40,272	7	0,000	82,618	0,059	0,047	0,002	0,243
Total within									353,007	150	0,000					
Total between									8,829	3	0,032					
Random effects analysis																
9999 = missing	21	0,545	0,063	0,004	0,420	0,669	8,581	0,000								
group	79	0,566	0,033	0,001	0,502	0,630	17,360	0,000								
individual	46	0,566	0,043	0,002	0,481	0,650	13,080	0,000								
mixed	8	0,344	0,100	0,010	0,149	0,540	3,457	0,001								
Total between									4,627	3	0,201					

Tabelle F.14: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Geschlecht des Testleiters

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
9999 = missing	102	0,534	0,018	0,000	0,499	0,569	30,291	0,000	190,094	101	0,000	46,868	0,028	0,010	0,000	0,168
female	23	0,538	0,039	0,002	0,461	0,614	13,762	0,000	28,919	22	0,147	23,926	0,012	0,015	0,000	0,109
male	5	0,467	0,088	0,008	0,295	0,638	5,326	0,000	3,565	4	0,468	0,000	0,000	0,038	0,001	0,000
mixed	17	0,416	0,028	0,001	0,361	0,470	14,897	0,000	115,409	16	0,000	86,136	0,090	0,050	0,003	0,299
same-sex as testpn	7	0,578	0,093	0,009	0,395	0,760	6,189	0,000	9,420	6	0,151	36,305	0,046	0,077	0,006	0,216
Total within									347,406	149	0,000					
Total between									14,430	4	0,006					
Random effects analysis																
9999 = missing	102	0,567	0,027	0,001	0,513	0,621	20,745	0,000								
female	23	0,535	0,049	0,002	0,439	0,632	10,849	0,000								
male	5	0,467	0,088	0,008	0,295	0,638	5,326	0,000								
mixed	17	0,516	0,084	0,007	0,352	0,680	6,160	0,000								
same-sex as testpn	7	0,500	0,139	0,019	0,227	0,773	3,591	0,000								
Total between									1,675	4	0,795					

Tabelle F.15: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Instruktionsmodus

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
9999 = missing	44	0,470	0,023	0,001	0,425	0,514	20,647	0,000	150,115	43	0,000	71,355	0,059	0,024	0,001	0,243
other	2	0,821	0,643	0,413	-0,439	2,081	1,277	0,202	0,082	1	0,775	0,000	0,000	1,844	3,401	0,000
verbal	50	0,472	0,026	0,001	0,420	0,523	18,031	0,000	91,212	49	0,000	46,279	0,031	0,016	0,000	0,176
verbal and written	15	0,544	0,037	0,001	0,472	0,616	14,834	0,000	23,535	14	0,052	40,514	0,014	0,014	0,000	0,119
written	27	0,616	0,032	0,001	0,553	0,680	19,027	0,000	66,727	26	0,000	61,035	0,047	0,026	0,001	0,216
written and read aloud	16	0,456	0,057	0,003	0,344	0,568	7,996	0,000	12,202	15	0,664	0,000	0,000	0,020	0,000	0,000
Total within									343,873	148	0,000					
Total between									17,963	5	0,003					
Random effects analysis																
9999 = missing	44	0,589	0,048	0,002	0,495	0,684	12,235	0,000								
other	2	0,821	0,643	0,413	-0,439	2,081	1,277	0,202								
verbal	50	0,496	0,041	0,002	0,416	0,577	12,047	0,000								
verbal and written	15	0,538	0,050	0,003	0,439	0,637	10,655	0,000								
written	27	0,644	0,061	0,004	0,525	0,764	10,558	0,000								
written and read aloud	16	0,456	0,057	0,003	0,344	0,568	7,996	0,000								
Total between									7,447	5	0,189					

Tabelle F.16: Subgruppenanalyse in Gesamtstichprobe gruppiert nach Speed-/Powertest

Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity				Tau-squared			
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed effect analysis																
power	141	0,496	0,015	0,000	0,467	0,524	33,884	0,000	342,366	140	0,000	59,108	0,044	0,012	0,000	0,210
speed	13	0,571	0,037	0,001	0,498	0,644	15,380	0,000	15,879	12	0,197	24,430	0,006	0,010	0,000	0,077
Total within									358,245	152	0,000					
Total between									3,591	1	0,058					
Random effects analysis																
power	141	0,545	0,026	0,001	0,494	0,597	20,711	0,000								
speed	13	0,571	0,044	0,002	0,485	0,656	13,082	0,000								
Total between									0,242	1	0,623					

Appendix G: Programmausgaben der Effektmoderatorenregression

Tabelle G.1: Berücksichtigte Prädiktoren und jeweilige Referenzkategorie (angegeben für kategorielle Variablen mit mehr als zwei Kategorien).

Prädiktoren	Referenzkategorie (nur angegeben bei > 2 Kategorien)
dummy_publication status	
year of publication [metr.]	XXXXXXXXXXXXXXXXXXXXXXXXXXXX
dummy_AU_sex	
dummy_age_adolescents	age_children
dummy_age_adults	
dummy_vessel_cylindrical	vessel_rectangular
dummy_vessel_curved	
dummy_vessel_pitcher	
dummy_vessel_florence flask	
dummy_vessel_spherical	
dummy_vessel_cup	
dummy_vessel_mixed	
dummy_context	
number of items [metr.]	XXXXXXXXXXXXXXXXXXXXXXXXXXXX
dummy_crittol_6° to 15°	crittol_3° to 5°
dummy_crittol_mean deviation	
dummy_crittol_MC	
dummy_crittol_other	
dummy_resmode_mean deviation	resmode_mean # of correct items
dummy_resmode_# of succeeded SS	
dummy_resmode_other	
dummy_testsit_individual	testsit_group
dummy_testsit_mixed	
dummy_speedpower	

Tabelle G.2: Programmausgabe: Aufgenommene Variablen und Methode der Aufnahme in schrittweiser gewichteter multipler Regression

Aufgenommene/Entfernte Variablen ^{a,b}			
Modell	Aufgenommene Variablen	Entfernte Variablen	Methode
1	dummy_age_adults	.	Schrittweise Auswahl (Kriterien: Wahrscheinlichkeit von F-Wert für Aufnahme $\leq$,050, Wahrscheinlichkeit von F-Wert für Ausschluß $\geq$,100).
2	dummy_crittol_MC	.	Schrittweise Auswahl (Kriterien: Wahrscheinlichkeit von F-Wert für Aufnahme $\leq$,050, Wahrscheinlichkeit von F-Wert für Ausschluß $\geq$,100).

a. Abhängige Variable: d d(sex)

b. Regression der gewichteten kleinsten Quadrate, gewichtet durch w_k116

Tabelle G.3: Programmausgabe: Modellzusammenfassung für schrittweise gewichtete multiple Regression

Modellzusammenfassung				
Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,288 ^a	,083	,075	,22570
2	,365 ^b	,133	,118	,22036

a. Einflußvariablen : (Konstante), dummy_age_adults

b. Einflußvariablen : (Konstante), dummy_age_adults, dummy_crittol_MC

Tabelle G.4: Programmausgabe: Signifikanzprüfung für Vorhersagewert der ausgegebenen Modelle

ANOVA^{c,d}

Modell		Quadrat summe	df	Mittel der Quadrate	F	Signifikanz
1	Regression	,524	1	,524	10,295	,002 ^a
	Residuen	5,807	114	,051		
	Gesamt	6,332	115			
2	Regression	,845	2	,422	8,696	,000 ^b
	Residuen	5,487	113	,049		
	Gesamt	6,332	115			

a. Einflußvariablen : (Konstante), dummy_age_adults

b. Einflußvariablen : (Konstante), dummy_age_adults, dummy_crittol_MC

c. Abhängige Variable: d d(sex)

d. Regression der gewichteten kleinsten Quadrate, gewichtet durch w_k116

Tabelle G.5: Programmausgabe: Signifikanzprüfung für Prädiktoren in ausgegebenen Modellen sowie Prüfung auf Multikollinearität

Koeffizienten^{a,b}

Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	T	Signifikanz	95%- Konfidenzintervall für B		Korrelationen			Kollinearitätsstatistik	
		B	Standard fehler	Beta			Unter- grenze	Ober- grenze	Nullter Ordnung	Partiell	Teil	Toleranz	VIF
1	(Konstante)	,452	,040		11,343	,000	,373	,531					
	dummy_age_adults	,150	,047	,288	3,209	,002	,058	,243	,288	,288	,288	1,000	1,000
2	(Konstante)	,402	,043		9,263	,000	,316	,488					
	dummy_age_adults	,181	,047	,346	3,826	,000	,087	,274	,288	,339	,335	,937	1,067
	dummy_crittol_MC	,134	,052	,232	2,568	,012	,031	,238	,145	,235	,225	,937	1,067

a. Abhängige Variable: d d(sex)

b. Regression der gewichteten kleinsten Quadrate, gewichtet durch w_k116

Literaturverzeichnis

(Mit* gekennzeichnete Quellen wurden in die vorliegende Meta-Analyse miteinbezogen.)

- *Abravanel, E., & Gingold, H. (1977). Perceiving and representing orientation: Effects of the spatial framework. *Merrill-Palmer Quarterly*, 23, 265-278.
- Ackermann, E. (1991). From decontextualized to situated knowledge: Revisiting Piaget's water-level experiment. In I. Harel & S. Papert (Eds.), *Constructionism* (pp. 269-294). Westport, CT: Ablex.
- Allen, M. J., Wittig, M. A., & Butler, K. (1981). Comments on Thomas and Jamison's „A test of the X-linked genetic hypothesis for sex differences on Piaget's water-level task“. *Developmental Review*, 1, 284-288.
- *Amponsah, B. (2000). A comparison of sex differences in visual-spatial performance from preadolescence to adulthood in Ghana and Norway. *South African Journal of Psychology*, 30(4), 25-31.
- Baenninger, M., & Newcombe, N. (1995). Environmental input to the development of sex-related differences in spatial and mathematical ability. *Learning and Individual Differences*, 7, 363-379.
- *Beatty, W. W., & Duncan, D. (1990). Relationship between performance on the everyday spatial activities test and on objective measures of spatial behavior on men and women. *Bulletin of the Psychonomic Society*, 28, 228-230.
- Begg, C., & Mazumdar, M. (1994). Operating characteristics of rank correlation test for publication bias. *Biometrics*, 50, 1088-1101.
- *Beisteiner, A. (2001). *Piaget's Wasserstandsaufgaben und Latent-Trait-Modelle*. Unpublished master's thesis, University of Vienna.
- Bem, S. (1981). Gender Schema Theory: A Cognitive Account of Sex Typing. *Psychological Review*, 88, 354-364.
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive Meta-Analysis, Version 2* [Computer software]. Englewood, NJ: Biostat.
- Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- *Casey, M. B., Nuttall, R. L., & Pezaris, E. (2001). Spatial-mechanical reasoning skills versus mathematics self-confidence as mediators of gender differences on

- mathematics subtests using cross-national gender-based items. *Journal for Research in Mathematics Education*, 32, 28-57.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. New York: LEA.
- Collaer, M. L., Reimers, S., & Manning, J. T. (2007). Visuospatial performance on an internet line judgment task and potential hormonal markers: Sex, sexual orientation, and 2D:4D. *Archives of Sexual Behavior*, 36, 177-92.
- Colom, R., Quiroga, M. A., & Juan-Espinosa, M. (1999). Are cognitive sex differences disappearing? Evidence from Spanish populations. *Personality and Individual Differences*, 27, 1189-1195.
- Cooper, H., & Hedges, L. (1994). Research Synthesis as a Scientific Enterprise. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 3-14). New York: Russel Sage Foundation.
- *DeAvila, E. A., & Havassy, B. (1974). *Intelligence of Mexican American children: A field study comparing neo-Piagetian and traditional capacity and achievement measures*. Austin, TX: Dissemination and Assessment Center for Bilingual Education.
- DeLisi, R. (1983). Developmental and individual differences in children's representation of the horizontal coordinate. *Merrill-Palmer Quarterly*, 29, 179-196.
- *Dessens, A., Cohen-Kettenis, P., Mellenbergh, G., van de Poll, N., Kopp, J., & Boer, K. (1998). Prenatal exposure to anticonvulsant drugs and spatial ability in adulthood. *Acta Neurobiologiae Experimentalis*, 58, 221-225.
- Dodwell, P. C. (1963). Children's understanding of spatial concepts. *Canadian Journal of Psychology*, 17, 141-161.
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455-463.
- Easterbrook, P., Berlin, J., Gopalan, R., & Matthews, D. (1991). Publication Bias in Clinical Research. *Lancet*, 337, 867-872.
- Egger, M., Smith, G., Schneider, M. & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629-634.
- Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, 43, 95-103.

- *Formann, A. K. (2003). Modeling data from water-level tasks: A test theoretical analysis. *Perceptual and Motor Skills*, 96, 1153-1172.
- Furberg, C. T., & Morgan, T. M. (1987). Lessons from overviews of cardiovascular trials. *Stat. Med.*, 6, 295-303.
- *Geiringer, E. R. (1976). *Performance on Piaget's water-level task in middle age and senescence*. Unpublished doctoral dissertation, Bowling Green State University, OH.
- *Geiringer, E. R., & Hyde, J. S. (1976). Sex differences on Piaget's water-level task: Spatial ability incognito. *Perceptual and Motor Skills*, 42, 1323-1328.
- Geiser, C., Lehmann, W., & Eid, M. (2008). A note on sex differences in mental rotation in different age groups. *Intelligence*, 36, 556-563.
- Geschwind, N., & Galaburda, A. M. (1987). *Cerebral lateralization*. Cambridge, MA.: MIT Press.
- *Gladue, B. A., Beatty, W. W., Larson, J., & Staton, R. D. (1990). Sexual orientations and spatial ability in men and women. *Psychobiology*, 18, 101-108.
- Glass, G., McGaw, B., & Smith, M. (1981). *Meta-Analysis in Social Research*. Beverly Hills, CA: Russel Sage Foundation.
- Glass, V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher*, 5, 3-8.
- *Golbeck, S. L. (1986). The role of physical content in Piagetian spatial tasks: Sex differences in spatial knowledge? *Journal of Research in Science Teaching*, 23, 365-376.
- *Golbeck, S. L., & Sinagra, K. (2000). Effects of gender and collaboration on college students' performance on a Piagetian spatial task. *The Journal of Experimental Education*, 69, 22-35.
- Hall, J., Tickle-Degnen, L., Rosenthal, R., & Mosteller, F. (1994). Hypotheses and problems in research synthesis. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 17-28). New York: Russel Sage Foundation.
- Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: Erlbaum.
- *Hammer, R. E., Hoffer, N., & King, W. L. (1995). Relationships among gender, cognitive style, academic major, and performance on the Piaget water-level task. *Perceptual and Motor Skills*, 80, 771-778.

- Harris, L. (1981). Sex-related variations in spatial skill. In L. S. Liben, A. H. Patterson, & N. Newcombe (Eds.), *Spatial representation and behavior across the life span: Theory and application* (pp. 83–125). New York: Academic Press.
- *Harris, L. J., Hanley, C., & Best, C. T. (1977). Conservation of horizontality: Sex differences in sixth graders and college students. In R. C. Smart & M. S. Smart (Eds.), *Readings in child development and relationships* (pp. 375-387). London: Macmillan.
- *Hecht, H., & Proffitt, D. R. (1995). The price of expertise: Effects of experience on the water-level task. *Psychological Science*, 6, 90-95.
- Héroux, G., & Robert, M. (1993, June). *Type of play predicts different visuo-spatial abilities in school-aged girls and boys*. Paper presented at the Annual Meeting of the American Psychological Society, Chicago.
- Hunt, M. (1997). *How science takes stock*. New York: Russell Sage Foundation.
- Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using ω and d . *American Psychologist*, 36, 892-901.
- Hyde, J. S. (1986). Introduction: Meta-Analysis and the psychology of gender. In J. S. Hyde & M. C. Linn (Eds.), *The Psychology of gender: Advances through meta-analysis* (pp. 1-13). Baltimore: Johns Hopkins University Press.
- Jamison, W., & Signorella, M. L. (1980). Sex-typing and spatial ability: The association between masculinity and success on Piaget's water-level task. *Sex Roles*, 6, 345-353.
- *Kalichman, S. C. (1986). Horizontality as a function of sex and academic major. *Perceptual and Motor Skills*, 63, 903-906.
- Kalichman, S. C. (1988). Individual differences in water-level task performance: A component-skills analysis. *Developmental Review*, 8, 273-295.
- *Kalichman, S. C. (1989a). Sex roles and sex differences in adult spatial performance. *Journal of Genetic Psychology*, 150, 93-100.
- *Kalichman, S. C. (1989b). The effects of stimulus context on paper-and-pencil spatial task performance. *Journal of General Psychology*, 116, 133-139.
- *Kalichman, S. C. (1991). Water levels, falling objects, and spiral tubes: An investigation of the general naiveté hypothesis of physical task performance. *Journal of Genetic Psychology*, 152, 255-262.
- *Kelly, J. T., & Kelly, G. N. (1977). Perception of horizontality by male and female college students. *Perceptual and Motor Skills*, 44, 724-726.

- *Kenyon, J. (1984). Paper-and-pencil tests of Piaget's water-level test: Sex differences and test modality. *Perceptual and Motor Skills*, 59, 739-742.
- Koscik, T., O'Leary, D., Moser, D., Andreasen, N., & Nopoulos, P. (in press). Sex differences in parietal lobe morphology: Relationship to mental rotation performance. *Intelligence*.
- *Krasnoff, A. G., & Walker, J. T. (1983). Visual spatial perception and paternal deprivation. *Personality and Individual Differences*, 4, 555-557.
- *Krasnoff, A. G., Walker, J. T., & Howard, M. (1989). Early sex-linked activities and interests related to spatial abilities. *Personality and Individual Differences*, 10, 81-85.
- Krekling, S., & Nordvik, H. (1992). Observational training improves adult women's performance on Piaget's water-level task. *Scandinavian Journal of Psychology*, 33, 117-124.
- *Lawton, C. A. (1994). Gender differences in way-finding strategies: Relationship to spatial ability and spatial anxiety. *Sex Roles*, 30, 765-779.
- *Levine, M. E., & Stern, R. M. (2002). Spatial task performance, sex differences, and motion sickness susceptibility. *Perceptual and Motor Skills*, 95, 425-431.
- *Li, C., Nuttall, R. L., & Zhao, S. (1999a). A test on the Piagetian water-level task with Chinese students. *Journal of Genetic Psychology*, 160, 369-380.
- *Li, C., Nuttall, R. L., & Zhao, S. (1999b). The effect of writing Chinese characters on success on water-level task: A study with Chinese and Chinese-American undergraduates. *Journal of Cross-Cultural Psychology*, 30, 91-105.
- *Liben, L. S. (1974). Operative understanding of horizontality and its relation to long-term memory. *Child Development*, 45, 416-424.
- *Liben, L. S. (1991a). Adults' performance on horizontality tasks: Conflicting frames of reference. *Developmental Psychology*, 27, 285-294.
- Liben, L. S. (1991b). The Piagetian water-level task: Looking beneath the surface. In R. Vasta (Ed.), *Annals of Child Development* (Vol. 8, pp. 81-143). London: Kingsley.
- Liben, L. S., & Golbeck, S. L. (1980). Sex differences in performance on Piagetian spatial tasks: Difference in competence of performance? *Child Development*, 51, 594-597.
- *Liben, L. S., & Golbeck, S. L. (1984). Performance on Piagetian horizontality and verticality tasks: Sex related differences in knowledge of relevant physical

- phenomena. *Developmental Psychology*, 20, 595-606.
- *Liben, L. S., & Golbeck, S. L. (1986). Adults' demonstration of underlying Euclidean concepts in relation to task context. *Developmental Psychology*, 22, 487-490.
- Linn, M. C., & Kyllonen, P. (1981). The field dependence-independence construct: Some, one, or none. *Journal of Educational Psychology*, 73, 261-273.
- Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial abilities: A meta-analysis. *Child Development*, 56, 1479-1498.
- *Linn, M. C., & Pulos, S. (1983). Male-female differences in predicting displaced volume strategy usage, aptitude relationships, and experience influences. *Journal of Educational Psychology*, 75, 86-96.
- Lippa, R. (2001). On deconstructing and reconstructing masculinity–femininity. *Journal of Research in Personality*, 35, 168–207.
- *Loehlin, J. C., & McFadden, D. (2003). Otoacoustic emissions, auditory evoked potentials, and traits related to sex and sexual orientation. *Archives of Sexual Behavior*, 32, 115-127.
- Lohaus, A., & Kessler, T. (1996). Zum Zusammenhang zwischen Lösungsverhalten und verbalisiertem Verständnis des Lösungsprinzips bei der Wasserspiegelaufgabe [Performance and verbally expressed understanding of the solution principle in the water-level task]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 28, 316-335.
- *Lohaus, A., Thomas, H., Kessler, T., & Gediga, G. (1996). Decomposing water-level responses: Field effects as separate influences. *Journal of Experimental Child Psychology*, 63, 79-102.
- Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.
- *McAfee, E. A., & Proffitt, D. R. (1991). Understanding the surface orientation of liquids. *Cognitive Psychology*, 23, 483-514.
- *McCormick, C., & Witelson, S. (1991). A cognitive profile of homosexual men compared to heterosexual men and women. *Psychoneuroendocrinology*, 16, 459-473.
- McGee, M. G. (1981). Predictions of X-linked inheritance for human characteristics. *Developmental Review*, 1, 289-295.
- *McGiffert-Spradlin, L., & Higgins-Hensley, J. (1979). An extension of evidence on sex

- differences in the failure of a water level conservation task by adults. *Journal of Genetic Psychology*, 135, 305-306.
- *McGillicuddy-DeLisi, A. V., DeLisi, R., & Youniss, J. (1978). Representation of the horizontal coordinate with and without liquid. *Merrill-Palmer Quarterly*, 24, 199-208.
- *Meehan, A. M., & Overton, W. F. (1986). Gender differences in expectancies for success and performance on Piagetian spatial tasks. *Merrill-Palmer Quarterly*, 32, 427-441.
- *Merriwether, A. M., & Liben, L. S. (1997). Adults' failures on Euclidean and projective spatial tasks: Implications for characterizing spatial cognition. *Journal of Adult Development*, 4, 57-69.
- Moore, D., & Johnson, S. (2008). Mental rotation in human infants: A sex difference. *Psychological Science*, 19, 1063-1066.
- *Morris, B. B. (1971). Effects of angle, sex, and cue on adults' perception of the horizontal. *Perceptual and Motor Skills*, 32, 827-830.
- *Mowrer-Popiel, E. L. (1990). *Adolescent and adult comprehension of horizontality as a function of cognitive style, training, and task difficulty*. Unpublished doctoral dissertation, Rutgers University, NJ.
- *Mowrer-Popiel, E., & DeLisi, R. (1984). An examination of spatial ability in relation to factors from the Bem Sex-Role Inventory. *Perceptual and Motor Skills*, 59, 131-136.
- *Murphy-Berman, V., Witters, L., & Harding, R. (1986). The effect of bottle shape, bottle position, and subject gender on intermediate-aged hearing-impaired students' performance on the water line task. *Volta Review*, 88, 37-46.
- *Myer, K. A., & Higgins-Hensley, J. H. (1984). Cognitive style, gender, and self-report of principle as predictors of adult performance on Piaget's water level task. *Journal of Genetic Psychology*, 144, 179-183.
- *Neumann, D. L., Fitzgerald, Z. T., Furedy, J. J., & Boyle, G. J. (2007). Sexually dimorphic effects of acute nicotine administration on arousal and visual-spatial ability in non-smoking human volunteers. *Pharmacology Biochemistry and Behavior*, 86, 758-765.
- Olson, D. M., & Eliot, J. (1986). Relationships between experiences, processing style, and sex-related differences in performance on spatial tests. *Perceptual and Motor Skills*, 62, 447-460.

- Olson, D. M., Eliot, J., & Hardy, R. C. (1988). Relationships between activities and sex-related differences in performance on spatial tests. *Perceptual and Motor Skills*, 567, 223-232.
- Orwin, R. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational and Behavioral Statistics*, 8, 157-159.
- *Otani, H., & Leonard, S. D. (1988). Judgments of horizontality of still water: Effects of perceptual and cognitive factors. *Perceptual and Motor Skills*, 66, 235-245.
- *Parameswaran, G. (1995). Gender differences in horizontality performance before and after training. *Journal of Genetic Psychology*, 156, 105-113.
- *Parameswaran, G., & DeLisi, R. (1996). Improvement in horizontality performance as a function of type of training. *Perceptual and Motor Skills*, 82, 595-603.
- Pascual-Leone, J. (1969). *Cognitive development and cognitive style: A general psychological integration*. Unpublished doctoral dissertation, University of Geneva, Switzerland.
- Pascual-Leone, J. (1992). Field dependence/independence and the water level task: A commentary to Pennings' paper. *Perceptual and Motor Skills*, 74, 1055-1058.
- Pascual-Leone, J., & Goodman, D. (1979). Intelligence and experience: A neo-Piagetian approach. *Instructional Science*, 8, 301-367.
- Pascual-Leone, J., & Morra, S. (1991). Horizontality of water level: A neo-Piagetian developmental review. In H. Reese (Ed.), *Advances in child development and behavior* (Vol. 23, pp. 231-276). San Diego, CA: Academic Press.
- *Pennings, A. H. (1991). Altering the strategies in embedded-figure and water-level tasks via instruction: A Neo-Piagetian learning study. *Perceptual and Motor Skills*, 72, 639-660.
- Percival, L., & Quinkert, K. (1987). Anthropometric factors, In M. A. Baker (Ed.), *Sex differences in human performance* (pp. 121-139). New York: Wiley.
- Piaget, J., & Inhelder, B. (1956). *The child's conception of space*. (F. J. Langdon & J. L. Lunzer, Trans.). London: Routledge and Kegan Paul. (Original work published 1948).
- *Pulos, S. (1997). Explicit knowledge of gravity and the water-level task. *Learning and Individual Differences*, 9, 233-247.
- *Quaiser-Pohl, C. (2001). Zum Einfluß des Wohnviertels auf die Raumvorstellung und die kognitiven Landkarten von 7- bis 12-Jährigen [The effect of neighborhoods on the spatial abilities and the cognitive maps of 7- to 12-year-old children].

Psychologie in Erziehung und Unterricht, 48, 280-297.

- Quaiser-Pohl, C., Lehmann, W., & Eid, M. (2004). The relationship between spatial abilities and representations of large-scale space in children - a structural equation modeling analysis. *Personality and Individual Differences*, 36, 95-107.
- Quinn, P., & Liben, L. (2008). A sex difference in mental rotation in young infants. *Psychological Science*, 19, 1067-1070.
- *Ray, W. J., Georgiou, S., & Ravizza, R. (1979). Spatial abilities, sex differences, and lateral eye movements. *Developmental Psychology*, 15, 455-457.
- *Rebelsky, F. (1964). Adult perception of the horizontal. *Perceptual and Motor Skills*, 19, 371-374.
- *Robert, M. (1990). Sex typing the water-level task: There is more than meets the eye. *International Journal of Psychology*, 25, 475-490.
- *Robert, M., & Chaperon, H. (1989). Cognitive and exemplary modeling of horizontality representation of the Piagetian water-level task. *International Journal of Behavioral Development*, 12, 453-472.
- *Robert, M., & Harel, F. (1996). The gender difference in orienting liquid surfaces and plumb-lines: Its robustness, its correlates, and the associated knowledge of simple physics. *Canadian Journal of Experimental Psychology*, 50, 280-314.
- *Robert, M., & Héroux, G. (2004). Visuo-spatial play experience: Forerunner of visuo-spatial achievement in preadolescent and adolescent boys and girls? *Infant and Child Development* 13, 49-78.
- *Robert, M., & Longpré, S. (2005). Sensory and postural input in the occurrence of a gender difference in orienting liquid surfaces. *Psychological Record*, 55, 67-89.
- *Robert, M., & Morin, P. (1993). Gender differences in horizontality and verticality representation in relation to initial position of the stimuli. *Canadian Journal of Experimental Psychology*, 47, 507-522.
- *Robert, M., & Ohlmann, T. (1994). Water-level representation by men and women as a function of rod-and-frame test proficiency and visual and postural information. *Perception*, 23, 1321-1333.
- *Robert, M., Pelletier, J., St-Onge, R., & Berthiaume, F. (1994). Women's deficiency in water-level representation: Present in visual conditions yet absent in haptic contexts. *Acta Psychologica*, 87, 19-32.
- *Robert, M., & Tanguay, M. (1990). Perception and representation of the Euclidean coordinates in mature and elderly men and women. *Experimental Aging*

Research, 16, 123-131.

- *Roberts, J. E., & Bell, M. A. (2002). The effects of age and sex on mental rotation performance, verbal performance, and brain electrical activity. *Developmental Psychobiology*, 40, 391-407.
- Rosenthal, R. (1979). The 'file drawer problem' and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- Rosenthal, R., & DiMatteo, M. (2001). Meta-Analysis: Recent Developments in Quantitative Methods for Literature Reviews. *Annual Review of Psychology*, 52, 59-82.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustment* (pp. 1-7). London: Wiley.
- Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, 94, 18-38.
- Rustenbach, S. (2003). *Metaanalyse: Eine anwendungsorientierte Einführung* (K. Pawlik, Ed.). Bern: Verlag Hans Huber.
- Ryckman, D. B., & Peckham, P. (1987). Gender differences in attributions for success and failure situations across subject areas. *Journal of Educational Research*, 81, 120-125.
- *Sanders, G., & Ross-Field, L. (1986). Sexual orientation and visio-spatial ability. *Brain and Cognition*, 5, 280-290.
- Schulze, R. (2007). The State and the art of meta-analysis. *Zeitschrift für Psychologie*, 215, 87-89.
- Schumann, S. (2006). *Repräsentative Umfrage: Praxisorientierte Einführung in empirische Methoden und Statistische Analyseverfahren*. Oldenbourg: Wissenschaftsverlag.
- *Seng, A. S. H., & Tan, L. C. (2002, April). *Cultural and gender differences in spatial ability of young children*. Paper presented at the Annual Meeting of the Association for Childhood Education International, San Diego, CA.
- Sherman, J. A. (1967). Problems of sex differences in space perception and aspects of intellectual functioning. *Psychological Review*, 74, 290-299.

- Sherman, J. A. (1978). *Sex-related cognitive differences*. Springfield, IL: Charles C. Thomas.
- *Sholl, M. J. (1989). The relation between horizontality and rod-and-frame and vestibular navigational performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 110-125.
- *Sholl, M. J., & Liben, L. S. (1995). Illusory tilt and Euclidian schemes as factors in performance on the water-level task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1624-1638.
- Signorella, M. L., & Jamison, W. (1978). Sex differences in the correlations among field dependence, spatial ability, sex role orientation, and performance on Piaget's water level task. *Developmental Psychology*, 14, 689-690.
- Signorella, M. L., & Jamison, W. (1986). Masculinity, femininity, androgyny, and cognitive performance: A meta-analysis. *Psychological Bulletin*, 100, 207-228.
- *Signorella, M. L., Jamison, W., & Krupa, M. H. (1989). Predicting spatial performance from gender stereotyping in activity preferences and self-concept. *Developmental Psychology*, 25, 89-95.
- Signorella, M. L., Krupa, M. H., Jamison, W., & Lyons, N. (1986). A short version of the spatial activities questionnaire. *Sex Roles*, 14, 475-479.
- Sterne J. A., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 99-110). London: Wiley.
- Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research*. Chichester, GB: Wiley.
- *Tegano, R. Z., & Faulkender, P. F. (1983). Adult sex differences in the concept of horizontality as related to mechanical ability, ACT scores and state anxiety. *Psychological Reports*, 53, 858-858.
- *Thomas, H., & Jamison, W. (1975). On the acquisition of understanding that still water is horizontal. *Merrill-Palmer Quarterly*, 21, 31-44.
- Thomas, H., & Jamison, W. (1981a). A test of the X-linked genetic hypothesis for sex differences on Piaget's water-level task. *Developmental Review*, 1, 274-283.
- Thomas, H., & Jamison, W. (1981b). Genetic mediation of sex differences in horizontality: A rejoinder. *Developmental Review*, 12, 596-599.

- Thomas, H., Jamison, W., & Hummel, D. D. (1973). Observation is insufficient for discovering that the surface of still water is invariably horizontal. *Science*, 181, 173-174.
- *Thomas, H., & Lohaus, A. (1993). Modeling growth and individual differences in spatial tasks. *Monographs of the Society for Research in Child Development*, 58(9, Serial No. 273).
- *Thompson, E. G., Mann, I. T., & Harris, L. J. (1981). Relationships among cognitive complexity, sex, and spatial task performance in college students. *British Journal of Psychology*, 72, 249-256.
- Thurstone, L. L., & Thurstone, T. G. (1962). *The primary mental abilities test*. Chicago: Science Research Associates.
- *Tran, U. S., & Formann, A. K. (2008). Piaget's water-level tasks: Performance across the lifespan with emphasis on the elderly. *Personality and Individual Differences* 45, 232-237.
- *Treagust, D. F. (1980). Gender-related differences of adolescents in spatial representational thought. *Journal of Research in Science Teaching*, 17, 91-97.
- Vandenberg, S., & Kuse, A. (1978). Mental Rotations, a Group Test of Three-Dimensional Spatial Visualization. *Perceptual and Motor Skills*, 47, 599-604.
- *Vasta, R., Knott, J. A., & Gaze, C. E. (1996). Can spatial training erase the gender differences on the water-level task? *Psychology of Women Quarterly*, 20, 549-567.
- Vasta, R., & Liben, L. S. (1996). The water-level task: An intriguing puzzle. *Current Directions in Psychological Science*, 5, 171-177.
- *Vasta, R., Lightfoot, C., & Cox, B. D. (1993). Understanding gender differences on the water-level problem: The role of spatial perception. *Merrill-Palmer Quarterly*, 39, 391-414.
- Vasta, R., Rosenberg, D., Knott, A., & Gaze, C. E. (1997). Experience and the water-level task revisited: Does expertise exact a price? *Psychological Science*, 8, 336-339.
- *Voyer, D. (1996). The relation between mathematical achievement and gender differences in spatial abilities: A suppression effect. *Journal of Educational Psychology*, 88, 563-571.
- *Voyer, D., Nolan, C., & Voyer, S. (2000). The relation between experience and spatial performance in men and women. *Sex Roles*, 43, 891-915.

- *Voyer, D., & Sullivan, A. M. (2003). The relation between spatial and mathematical abilities: Potential factors underlying suppression. *International Journal of Psychology*, 38, 11-23.
- 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.
- *Walker, J. T., & Krasnoff, A. G. (1978). The horizontality principle in young men and women. *Perceptual and Motor Skills*, 46, 1055-1061.
- *Walker, J. T., Krasnoff, A. G., & Peaco, D. (1981). Visual spatial perception in adolescents and their parents: The X-linked recessive hypothesis. *Behavior Genetics*, 11, 403-413.
- *Wegesin, D. J. (1995). *Functional verbal asymmetries and spatial ability in heterosexual and homosexual men and women: Evidence from event-related potentials*. Unpublished doctoral dissertation, University of Minnesota.
- *Wegesin, D. J. (1998). A neuropsychological profile of homosexual and heterosexual men and women. *Archives of Sexual Behavior*, 27, 91-108.
- Whitley, B. E., McHugh, M. C., & Frieze, I. H. (1986). Assessing the theoretical models for sex differences in causal attributions of success and failure. In J. S. Hyde & M. S. Linn (Eds.), *The psychology of gender: Advances through meta-analysis* (pp. 102-135). Baltimore: Johns Hopkins University Press.
- Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, 12, 265-310.
- Willemsen, E. W. (1974). Judgment of the vertical as a function of stimulus characteristics, age, and sex. *Perceptual and Motor Skills*, 38, 1344.
- Willemsen, E., & Reynolds, B. (1973). Sex differences in adults' judgments of the horizontal. *Developmental Psychology*, 8, 309.
- Witkin, H. A., & Goodenough, D. R. (1981). *Cognitive styles: Essence and origins*. New York: International Universities Press.
- Witters-Churchill, L. J., & Witters, L. A. (1989). Hearing-impaired students' performance on the Piagetian liquid horizontality test: An analysis and synthesis. In D. S. Martin (Ed.), *International Symposium on Cognition, Education, and Deafness, July 5-8 1989, Washington, DC* (Working papers, Vol. 2, pp. 614-637).
- *Wittig, M. A., & Allen, M. J. (1984). Measurement of adult performance on Piaget's

water horizontality task. *Intelligence*, 8, 305-313.

Zinser, O., Palmer, D. L., & Miller, C. R. (2004). Site distance, gender and knowledge of geographic sites. *Sex Roles*, 51, 661-686.

Eidesstattliche Erklärung

Ich bestätige, dass ich die vorliegende Diplomarbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe, und dass die Arbeit in gleicher oder ähnlicher Form noch in keiner anderen Prüfungsbehörde vorgelegen hat. Alle Ausführungen der Arbeit, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Wien, im April 2009

Stefanie Pavlovic

Curriculum vitae

Angaben zur Person

Name Stefanie Pavlovic
Staatsangehörigkeit A
Geburtsdatum, -ort 24.08.1981, Wien
Geschlecht weiblich
Familienstand ledig

Schul- und Berufsbildung

Daten Seit 2000
Bezeichnung der zu erwerbenden Qualifikation Diplomstudium Psychologie, Mag. rer. nat.
(1. Diplomprüfung 2003 mit Auszeichnung bestanden)
Ausbildungseinrichtung Fakultät für Psychologie, Universität Wien.
Stufe der internationalen Klassifikation ISCED 5

Daten 1992-2000
Bezeichnung der zu erwerbenden Qualifikation Matura (mit Auszeichnung bestanden)
Ausbildungseinrichtung Bundesgymnasium Wien VI, neusprachlicher Zweig.
Stufe der internationalen Klassifikation ISCED 3

Berufserfahrung

Daten Oktober 2004 – Jänner 2008
Funktion Studienassistentz
Tätigkeiten und Zuständigkeiten Mitarbeit in psychologischer Forschung und Lehre.
Arbeitgeber Institut für psychologische Grundlagenforschung, Fakultät für Psychologie, Universität Wien.

Daten April – Juni 2004
Funktion Fachpraktikum
Tätigkeiten und Zuständigkeiten Trainee im Bereich psychologische Forschung.
Arbeitgeber Institut für psychologische Grundlagenforschung, Fakultät für Psychologie, Universität Wien.

Daten	August 2002, September 2003
Funktion	Praktikum
Tätigkeiten und Zuständigkeiten	Teamassistenz für IT-Solutions in der Telekommunikation.
Arbeitgeber	T-Mobile Austria GmbH
Daten	Juli 2002
Funktion	Praktikum
Tätigkeiten und Zuständigkeiten	· Psychosoziale Betreuung ambulant aufgenommener psychiatrischer Patienten. · Sprechstundenhilfe.
Arbeitgeber	Sozialpsychiatrisches Ambulatorium Favoriten, Psychosozialer Dienst.
Universitäre Angaben	AUSZEICHNUNGEN Leistungsstipendium der Universität Wien (finanziert durch bm:bwk), 2003 Leistungsstipendium der Universität Wien (finanziert durch bm:bwk), 2006 Leistungsstipendium der Universität Wien (finanziert durch bm:bwk), 2007 PUBLIKATIONEN Voracek, M., & Pavlovic, S. (2007). The tell-tale hand: The relationship of 2D:4D to perceived attractiveness, sex typicality, and other attributes of palms. <i>Journal of Individual Differences</i>, 28, 88-97. Pavlovic, S., Voracek, M., & Formann, A. K. (2008). Meta-analysis of sex differences in Piaget's water-level tasks: Summarizing 50 years of research [Abstract]. <i>International Journal of Psychology</i>, 43, 67. KONGRESSBEITRÄGE Pavlovic, S., Voracek, M., & Formann, A. K. (2008). <i>Der Geist in der Flasche oder Meta-Analyse über Piagets Wasserstandsaufgaben: 50 Jahre Forschung auf den Punkt gebracht</i>. Vortrag auf der 8. Tagung der Österreichischen Gesellschaft für Psychologie, 24.-26. April 2008, Linz. Pavlovic, S., Voracek, M., & Formann, A. K. (2008). <i>Meta-analysis of sex differences in Piaget's water-level tasks: summarizing 50 years of research</i>. Paper presented at the 29th International Congress of Psychology, July 20–25 2008, Berlin, Germany.