Ursachen für Geschlechterdifferenzen in Tests des Allgemeinen Wissens

Dissertation zur Erlangung des Akademischen Grades des Dr. phil. durch

den Fachbereich Human- und Sozialwissenschaften

der Bergischen Universität Wuppertal

vorgelegt von

Philipp Meinolf Engelberg

aus Wuppertal

Februar 2015

Danksagung

Ich möchte mich an dieser Stelle bei den vielen Personen bedanken, die mich bei der Erstellung meiner Dissertation unterstützt haben: Herr Prof. Dr. Schulze hat mich in der gesamten Promotionszeit umfassend unterstützt und meine Fähigkeit zum selbstständigen wissenschaftlichen Arbeiten gefördert. Ebenso geht mein Dank an meine Kolleginnen und Kollegen, insbesondere an Susan Hellwig, mit der ich viele anregende Gespräche über die Themen meiner Arbeit führte.

Großer Dank gilt auch den Studierenden, die als Studentische Hilfskräfte oder im Rahmen von Abschlussarbeiten wertvolle Beiträge bei Datenerhebungen und bei der Entwicklung von Items leisteten, allen voran Tina Kleyboldt und Maike Pisters. Das gleiche gilt für die zahlreichen fachlichen Expertinnen und Experten verschiedener Themengebiete, die sich Zeit für die Durchsicht von Itementwürfen und ausführliche Gespräche über Verbesserungsmöglichkeiten nahmen.

Vielen Dank auch an die vielen Personen, die mir die Möglichkeit gaben, an Schulen in Bochum und in Xanten viele Probandinnen und Probanden für die Datenerhebungen gewinnen zu können.

Vor allem aber möchte ich mich bei meiner Frau Sandra, meinem Sohn und meinen Eltern bedanken, die mit ihrer emotionalen Unterstützung, mit ihrer enormen Geduld und ihrer Hilfsbereitschaft erst die Voraussetzungen geschaffen haben, diese Promotion abzuschließen.

Dir, Sandra, ist diese Arbeit gewidmet.

Zusammenfassung

In Tests des Allgemeinen Wissens wurden in der Vergangenheit häufig Geschlechterdifferenzen festgestellt, die einseitig zugunsten von Männern ausfallen. In dieser Dissertation wurden 2 potentielle Erklärungsmöglichkeiten hierfür überprüft, wobei es sich (a) um eine geringere Einschätzung des eigenen Wissens bei Frauen und (b) um eine stärkere Gewichtung von Interessengebieten von Männern in bestehenden Wissenstests handelt. In der ersten Untersuchung wurde experimentell überprüft, ob eine durch Vorgabe von leichten bzw. schwierigen Items manipulierte Einschätzung des eigenen Wissens zu höherer bzw. geringerer Leistung in einem Wissenstest führt. Im Ergebnis wiesen Selbsteinschätzung und Leistung generell einen positiven Zusammenhang auf. Die Selbsteinschätzung wurde zudem erfolgreich manipuliert. Dennoch zeigten Personen mit verringerter Selbsteinschätzung im Durchschnitt bessere Leistungen im Wissenstest als Personen mit erhöhter Selbsteinschätzung, weshalb die Hypothese, dass Geschlechterdifferenzen in Wissenstests durch Geschlechterdifferenzen in der Selbsteinschätzung verursacht werden, verworfen wird. Für die Überprüfung der zweiten Erklärungsmöglichkeit wurden die Interessen von Frauen und Männern mit einem hierfür entwickelten Interessenfragebogen erfasst und sämtliche Items des BOWIT und des Wissenstests des I-S-T 2000 R den Interessengebieten zugeordnet. Den Ergebnissen zufolge enthalten beide Tests mehrheitlich Items zu Themen, für die Männer sich durchschnittlich stärker interessieren. Im Anschluss wurde ein neuer Wissenstest entwickelt, der ausschließlich Items zu Interessengebieten von Frauen enthielt. Eine Testbatterie, bestehend aus diesem Wissenstest und dem Wissenstest des I-S-T 2000 R, weist eine 2-faktorielle Struktur auf. Frauen und Männern wird jeweils in einem Faktor der höhere Erwartungswert attestiert. Beide Faktoren korrelieren in vergleichbarer Höhe mit der Schulleistung. Die Parcels des I-S-T werden jedoch ausschließlich dem Faktor zugeordnet, in

dem Männer den höheren Erwartungswert aufweisen. Den Ergebnissen zufolge sind Interessenunterschiede der Geschlechter somit für Leistungsunterschiede in Wissenstests von Bedeutung. Aufgrund mangelhafter Passung eines 1-faktoriellen und guter Passung eines 2-faktoriellen Messmodells wird die Frage aufgeworfen, ob die Auffassung des Allgemeinen Wissens als einzelne Fähigkeit angemessen ist.

Inhaltsverzeichnis

Einleitung

Wissen ist heutzutage in vielen Bereichen des Lebens von zentraler Bedeutung. Nach Beauducel und Süß (2011) kann die gesellschaftliche Relevanz des Wissens "daran abgelesen werden, dass bei extrem vielen gesellschaftlichen Prüfungen Wissen abgefragt wird" (S. 265). Ackerman und Beier (2004) weisen darauf hin, dass Berufserfolg wesentlich durch das Wissen bestimmt wird, wobei es sich hier sowohl um das für den jeweiligen Beruf spezifische, aber auch um ein breitgefächertes, themenübergreifendes Wissen handeln kann, das im Folgenden als *Allgemeines Wissen* bezeichnet wird. Tests des Allgemeinen Wissens dienen in der psychologischen Diagnostik häufig auch als Markiervariablen der kristallinen Intelligenz. Die Bedeutung der kristallinen Intelligenz (bzw. des Wissens) für akademische und berufliche Leistungen wurde vielfach nachgewiesen (z.B. Beauducel & Kersting, 2002; Furnham, Monsen & Ahmetoglu, 2009). Für die Prüfung der Kriteriumsvalidität von Tests des Allgemeinen Wissens werden häufig Schulleistungen herangezogen. Nach Neidhardt-Wilberg (2005) stellt das Allgemeine Wissen eine Variable dar, "die für pädagogisch-psychologische Forschung und Praxis zumindest genauso relevant wie die allgemeine Intelligenz erscheint" (S. 146).

Frauen und Männer zeigen bei standardisierten psychologischen Tests des Allgemeinen Wissens häufig unterschiedliche Leistungen. In der Regel erreichen Männer hierbei die höheren Testscores (z.B. Ackerman, Bowen, Beier & Kanfer, 2001; Lynn & Irwing, 2002). Auch bei den Normen von Wissenstests findet man deutliche Geschlechterdifferenzen zugunsten von Männern (z.B. Feingold, 1993). In den Manualen zweier bekannter deutschsprachiger Wissenstests, dem Erweiterungsmodul des Intelligenz-Struktur-Tests

2000 R (I-S-T 2000 R) von Liepmann, Beauducel, Brocke und Amthauer (2007) und dem Bochumer Wissenstests (BOWIT) von Hossiep und Schulte (2008), wird über Mittelwertdifferenzen mit Effektstärken von 0.30 bzw. 1.02 Standardabweichungseinheiten zugunsten von Männern berichtet.

Geschlechterdifferenzen der Intelligenz und verschiedener Facetten davon, beispielsweise verbaler, numerischer oder figuraler Intelligenz, wurden in der psychologischen Forschung intensiv untersucht. Zusammenfassungen über Forschungsergebnisse bieten beispielsweise Halpern (2012), Hyde (2005) und Lippa (2005). Neidhardt-Wilberg (2005) merkt jedoch an, dass das Allgemeine Wissen "bislang selten explizite Berücksichtigung in der psychologischen Forschung" (S. 145) fand. Diese Feststellung ist nach wie vor aktuell, zumindest im Bereich der Differenziellen Psychologie. So finden sich auch nur sehr wenige Untersuchungen zu Geschlechterdifferenzen des Allgemeinen Wissens. Nichtsdestoweniger existieren zahlreiche Intelligenztests, die Untertests des Allgemeinen Wissens beinhalten, wie beispielsweise der Hamburg-Wechsler Intelligenztest für Erwachsene in der revidierten Version (zitiert nach Neidhardt-Wilberg, 2005), der Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe (BEFKI 8-10) von Wilhelm, Schroeders und Schipolowski (2014), oder der I-S-T 2000 R von Liepmann et al. (2007). Dem Umstand, dass Männer deutlich höhere Leistungen in Wissenstests zeigen, wird in der Testentwicklung teilweise durch die Erstellung geschlechtsspezifischer Normen begegnet, beispielsweise beim BOWIT (Hossiep & Schulte, 2008) oder beim BEFKI (Wilhelm et al., 2014). Eine durch die beschriebenen Geschlechterdifferenzen naheliegende Schlussfolgerung ist, dass das Allgemeine Wissen bei Frauen durchschnittlich geringer als bei Männern ausgeprägt ist. Es stellt sich jedoch die Frage, ob systematische Unterschiede von Frauen und Männern in Ergebnissen von Wissenstests den Schluss auf systematische Unterschiede in der Fähigkeit Wissen zulassen.

Verfügen Frauen im Durchschnitt tatsächlich über ein geringeres Wissen als Männer, wie es zahlreiche Befunde nahelegen, oder erfolgt in gebräuchlichen standardisierten Tests des Allgemeinen Wissens eine systematische Verzerrung der Testergebnisse zugunsten von Männern?

Eine mögliche Ursache für geschlechtsspezifische Verzerrungen in Testscores von Wissenstests sind Geschlechterdifferenzen in der Einschätzung des eigenen Wissens. In der bisherigen Forschung wurde wiederholt über eine geringere Einschätzung eigener kognitiver Fähigkeiten von Frauen im Vergleich zu Männern berichtet (z.B. Rammstedt & Rammsayer, 2000; von Stumm, Chamorro-Premuzic & Furnham, 2009). Valentine, DuBois und Cooper (2004) berichteten über positive korrelative Zusammenhänge zwischen der Einschätzung eigener Fähigkeiten und der tatsächlichen Leistung in Tests. Eine geringere Einschätzung des eigenen Wissens könnte Sorgen und Ängste auslösen, welche die kognitiven Ressourcen reduzieren, die der Person für die Bearbeitung der Aufgaben zur Verfügung stehen, was wiederum zu Leistungseinbußen führen könnte. Ashcraft und Kirk (2001) fanden Unterstützung für diesen Wirkmechanismus im Rahmen von Mathematikaufgaben. Für Wissenstests steht eine Überprüfung aus. In der vorliegenden Arbeit wurde im Rahmen eines Experiments die Einschätzung des eigenen Wissens der Teilnehmenden durch die Vorlage von schwierigen versus leichten Beispielitems manipuliert. Es wurde überprüft, ob Probandinnen und Probanden, deren Selbsteinschätzung durch die Manipulation verringert worden war, geringere Leistungen im Wissenstest zeigen als die Beteiligten, deren Selbsteinschätzung manipulativ erhöht worden war.

Eine Voraussetzung für die Vergleichbarkeit der Testscores von Personen unterschiedlicher Gruppen ist die Messinvarianz eines Tests. Beispielsweise verweisen Marsh (1994) und Lubke, Dolan, Keldermann und Mellenbergh (2003) auf die Wichtigkeit dieser psychometrischen Eigenschaft eines Messinstruments. Bei Wissen handelt es sich um ein

nicht direkt beobachtbares, latentes Merkmal. Die Beziehung zwischen den Ergebnissen eines Tests und dem latenten Merkmal, das durch den Test erfasst wird, kann in Form von Messmodellen dargestellt werden. Durch konfirmatorische Faktorenanalysen mit Mehrgruppen-Modellen lässt sich die Messinvarianz eines Tests überprüfen. Sofern sich der Test als messinvariant erweist, kann davon ausgegangen werden, dass für die verschiedenen Gruppen dasselbe Merkmal gemessen wird und Gruppenunterschiede in den Testscores als Gruppenunterschiede in dem Merkmal interpretiert werden können. Engelberg (2008) prüfte den Wissenstest des I-S-T 2000 R auf Messinvarianz für Frauen und Männer und kam zu einem uneingeschränkt zufriedenstellenden Ergebnis.

Das Vorliegen von Messinvarianz in einem Test lässt jedoch lediglich den Schluss zu, dass bei den verschiedenen Gruppen dasselbe psychologische Merkmal im gleichen Ausmaß für das Zustandekommen der Testscores relevant ist. Auch wenn Messinvarianz für einen Test als gegeben angenommen werden kann, bleibt die Frage offen, *welches* psychologische Merkmal relevant ist. So ist beispielsweise denkbar, dass in einem Wissenstest, der zur Erfassung des Allgemeinen Wissens dienen soll, tendenziell einseitig das Wissen zu einem bestimmten Themengebiet erfasst wird, während andere Themengebiete nicht abgedeckt werden. Ein solcher Test mag Messinvarianz aufweisen, könnte aber dennoch zu einer systematischen Verzerrung der Einschätzungen des Allgemeinen Wissens von Mitgliedern der beiden Gruppen führen. Das Problem würde hier nicht in der Vergleichbarkeit der Testergebnisse für die Gruppen liegen, sondern in der Interpretation der Testergebnisse als Allgemeines Wissen. Die inhaltliche Interpretation von Tests ist zentraler Gegenstand der Prüfung der Inhaltsvalidität. Wie Beauducel und Süß (2011) erläutern, wird die Sicherstellung der Inhaltsvalidität von Tests des Allgemeinen Wissens durch Probleme der Definition eines Itemuniversums erschwert. Es stellt sich daher die Frage, welche Wissensgebiete Bestandteil des Allgemeinen Wissens sind.

Cattell (1971/1987) formulierte in seiner Investmenttheorie die Annahme, dass die kristalline Intelligenz, welche in engem Zusammenhang mit Wissen steht, maßgeblich durch die zeitlichen und kognitiven Ressourcen bestimmt wird, die jemand in die Beschäftigung mit intellektuell anspruchsvollen Themen investiert. In den Themen, in die eine Person ihre Ressourcen investiert, erwirbt sie Wissen. Die Auswahl der Themen wiederum wird wesentlich durch die Interessen der Person bestimmt. Wie aus einer Meta-Analyse von Su, Rounds und Armstrong (2009) hervorgeht, liegen zahlreiche Belege dafür vor, dass Frauen und Männer sich hinsichtlich ihrer Interessen unterscheiden. Die Annahmen der Investmenttheorie und der unterschiedlichen Interessen von Frauen und Männern legen den Schluss nahe, dass auch das Wissen von Frauen und Männern sich auf unterschiedliche Bereiche konzentrieren könnte. Sofern keine Gründe dafür vorliegen, verschiedene Themengebiete bei der Erfassung des Allgemeinen Wissens unterschiedlich zu gewichten, sollte bei der Entwicklung von Wissenstests eine ausgewogene Berücksichtigung von Themen, für die Frauen bzw. Männer sich durchschnittlicher stärker interessieren, angestrebt werden. Der Wissenstest des I-S-T 2000 R und der BOWIT wurden in der vorliegenden Arbeit auf diese Ausgewogenheit überprüft. Aufgrund der aufschlussreichen Ergebnisse der Untersuchung wurde ein neuer Wissenstest entwickelt, der ausschließlich Items zu Interessengebieten von Frauen beinhaltet, und dessen psychometrische Qualitäten mit den Qualitäten der vorgestellten Wissenstests vergleichbar sind. Auch die Messinvarianz des Tests wurde geprüft und bestätigt. Im Anschluss wurde untersucht, wie die Leistungsunterschiede zwischen Frauen und Männern bei einer Testbatterie ausfallen, welche den Wissenstest des I-S-T 2000 R und den neu entwickelten Wissenstest beinhaltet. Die Ergebnisse warfen weitere Fragen bezüglich der Definition des Konstrukts des Allgemeinen Wissens auf.

In der vorliegenden Arbeit wurden somit zwei Merkmale, in denen sich Frauen und Männer unterscheiden, und welche beide in der bisherigen Forschung Relationen zum Wissen aufgewiesen haben, auf ihre ursächliche Bedeutung für Geschlechterdifferenzen in Wissenstests geprüft. Hierbei handelt es sich um die Einschätzung des eigenen Wissens und um Interessen. Es existieren zahlreiche weitere psychologische Merkmale, die ebenso wie Selbsteinschätzung und Interessen Zusammenhänge mit Wissen aufweisen, und bei deren Erfassung Mittelwertunterschiede zwischen Frauen und Männern gefunden wurden. Hierunter fallen beispielsweise Persönlichkeitsmerkmale, wie Extraversion und Typical Intellectual Engagement (siehe Kapitel 3.2.3). Es wird somit selbstverständlich kein Anspruch darauf erhoben, sämtliche potentiellen Ursachen für Verzerrungen von Geschlechterdifferenzen in bestehenden Tests des Allgemeinen Wissens zu überprüfen.

Die Arbeit ist folgendermaßen gegliedert: In Kapitel 2 wird zunächst der Begriff des Wissens näher erläutert, wobei insbesondere die Perspektive der Psychologie im Mittelpunkt steht. Es werden verschiedene Arten des Wissens beschrieben und die Wissensdiagnostik mit dabei auftretenden Problemen, beispielsweise bezüglich der Inhaltsvalidität, dargelegt. Außerdem wird Wissen zu verschiedenen anderen psychologischen Merkmalen in Beziehung gesetzt. In diesem Rahmen erfolgt auch eine detaillierte Beschreibung der Investmenttheorie von Cattell (1971/1987), welche für die Hypothese, dass Geschlechterdifferenzen in Wissenstests durch Geschlechterdifferenzen in Interessen erklärt werden können, von zentraler Bedeutung ist.

In Kapitel 3 wird auf das Phänomen der Geschlechterdifferenzen in Wissenstests näher eingegangen. Nach der Anführung zahlreicher Belege wird ein Überblick über eine Reihe von möglichen Erklärungen gegeben, die jedoch nicht Gegenstand der Untersuchungen dieser Arbeit sind. Die potentiellen Erklärungen der Leistungsunterschiede durch

Selbsteinschätzung und durch Interessen, die hier überprüft wurden, werden in den beiden folgenden Kapiteln jeweils gesondert behandelt.

Gegenstand von Kapitel 4 ist das Experiment, mit dem die Ursächlichkeit der Selbsteinschätzung für die Leistung in Wissenstests geprüft wurde. Zunächst erfolgt die Herleitung der Hypothese, dass die durchschnittlich geringere Einschätzung des eigenen Wissens von Frauen zu durchschnittlich geringeren Leistungen in Wissenstests führt. Im Anschluss werden das Experiment, die Ergebnisse der Untersuchung und die daraus gezogenen Schlussfolgerungen beschrieben.

Kapitel 5 umfasst drei Untersuchungen, in denen der Frage nachgegangen wurde, ob Interessenunterschiede zwischen Frauen und Männern und eine unausgewogene Berücksichtigung der entsprechenden Themen zu Geschlechterdifferenzen in Wissenstests führen könnten. Zunächst wird auf das Konstrukt Interesse eingegangen und Befunde zu Interessenunterschieden zwischen Frauen und Männern werden dargestellt. Anschließend erfolgen die Beschreibungen der drei Untersuchungen, jeweils in einem Methoden-, Ergebnis- und Diskussionsteil. Es sollte hierbei beachtet werden, dass die Planung der einzelnen Untersuchungen jeweils im Anschluss an die vorhergehende Untersuchung stattfand. Die Reihenfolge der Unterkapitel entspricht somit der praktischen Durchführung der drei Studien.

In Kapitel 6 werden abschließend die Ergebnisse und Schlussfolgerungen der Untersuchungen, die in den Kapiteln 4 und 5 beschrieben wurden, zusammengefasst. Außerdem werden Ideen für künftige Forschungsarbeiten zu dem Thema der Geschlechterdifferenzen in Wissenstests skizziert.

Der Begriff des Wissens

Zu Beginn der Arbeit wird zunächst der Begriff des Wissens näher beschrieben. Ausgangspunkt sind hier Definitionen, die sich im Duden finden. Nach einem kurzen Überblick über philosophische Aussagen zum Wissen erfolgt eine ausführliche Beleuchtung des Begriffs aus der Sicht der Psychologie. Hierbei wird auch auf die Diagnostik von Wissen eingegangen und das Konstrukt, wie es im Rahmen dieser Arbeit aufgefasst wird, näher spezifiziert. Außerdem wird der Begriff in Relation zur Intelligenz, zur Weisheit und zum Langzeitgedächtnis gesetzt.

2.1 Definition des Begriffs Wissen

Für das Substantiv Wissen findet sich im Duden von 1999 der Eintrag "*Gesamtheit der Kenntnisse, die jmd. [auf einem bestimmten Gebiet] hat*" (S. 4538). Als Beispiele hierfür werden unter anderem politisches und menschliches Wissen genannt. Für den Begriff Wissen als Verb findet man unter anderem einen Eintrag, in dem auf die Quelle des Wissens hingewiesen wird: "*durch eigene Erfahrung od. Mitteilung von außen Kenntnis von etw., jmdm. haben, sodass zuverlässige Aussagen gemacht werden können*" (S. 4537). Es finden sich auch weitere Begriffsbedeutungen, wie beispielsweise "*in der Lage sein, etw. zu tun*" (S. 4538) (Beispiel: Sie wussten, sich zu rechtfertigen.), die für die Begriffsbedeutung im Rahmen der vorliegenden Arbeit jedoch nicht von Bedeutung sind.

2.2 Wissen in der Philosophie

Der Hinweis im Duden auf die Möglichkeit, zuverlässige Aussagen zu treffen, findet sich auch in der Begriffsanalyse in der Philosophie wieder. In Platons Theaitetos wird eine Definition von Wissen als "wahre und gerechtfertigte Meinung" herausgearbeitet und

verworfen (siehe Baumann, 2006). Von Wissen lässt sich der Begriff des Glaubens abgrenzen. So stellt auch Glaube eine Überzeugung dar, die jedoch nicht wahr und auch nicht gerechtfertigt sein muss. Eine Überzeugung, die zufällig wahr ist, wie etwa die zufällig korrekte Vorhersage des Wetters, wird ebenfalls nicht als Wissen definiert. Ebenso fällt eine Überzeugung, die durchaus berechtigt und nachvollziehbar ist, jedoch nicht der Wahrheit entspricht, nicht unter den Begriff Wissen. Ein bekanntes Beispiel hierfür ist das geozentrische Weltbild. Die Definition des Begriffs Wissen durch die drei Elemente (a) Überzeugung, (b) wahr und (c) gerechtfertigt bildet in der Erkenntnistheorie die sogenannte "traditionelle Konzeption des Wissens" (Baumann, 2006, S. 39). Sie wurde "bis weit ins 20. Jh. hinein von vielen Philosophen fast als Selbstverständlichkeit akzeptiert" (Baumann, 2006, S. 40). Nichtsdestoweniger gab es auch Debatten über die Frage, ob die drei Elemente der Definition von Platon hinreichend sind (z.B. Gettier, 1963), worauf hier jedoch nicht näher eingegangen wird, da diese Diskussionen zu weit führen würden.

Ein Ansatz, der hier ausführlicher betrachtet wird, ist die Gruppe der sogenannten DIKW-Modelle. Die Abkürzung steht für die Begriffe *D*ata, *I*nformation, *K*nowledge und *W*isdom, welche sich in Form einer Pyramide hierarchisch anordnen lassen. Auf der untersten, breitesten Ebene steht die Wahrnehmung von Daten, womit sensorische Stimuli wie beispielsweise Symbole gemeint sind. Daten können zu inhaltlicher Information zusammengefasst werden. So ergeben Symbole Wörter, welche Sätze und damit Aussagen bilden können. Informationen können wiederum miteinander in Verbindung gebracht, organisiert und in einem Kontext gespeichert werden, was das Wissen darstellt. Wie bereits erwähnt, bilden die DIKW-Modelle eine Gruppe von Modellen. Unterschiede bestehen beispielsweise in der Anzahl der hierarchischen Ebenen. So stellt das Wissen bei einem Teil der Modelle bereits die höchste Hierarchiestufe dar, weshalb diese auch als DIK-Modelle bezeichnet werden. Bei anderen Modellen ist zusätzlich über dem Wissen noch die Ebene der

Weisheit angeordnet. Bellinger, Castro und Mills (2004) zufolge betrachtete Ackoff *understanding* als eine weitere Ebene zwischen Wissen und Weisheit. Understanding definierte er als "appreciation of why" und Weisheit als "evaluated understanding" (zitiert nach Bellinger et al., 2004). Zeleny (1987) fasste die vier Ebenen folgendermaßen zusammen: "To manage wisely implies knowing why to do something; to manage effectively implies knowing what to do; to manage efficiently implies knowing how to do it (and to muddle through implies nothing and having 'lots of data' around)" (Zeleny, 1987, S. 60). Bellinger et al. geben folgendes Beispiel für die Veranschaulichung der hierarchischen Ebenen: Die Wahrnehmung, dass es regnet, lässt sich der Ebene der *Daten* zuordnen. Die Tatsache, dass die Temperatur um 15 Grad fiel und es anschließend anfing zu regnen, findet sich auf Ebene der *Information*, welche nun wiederum in den Kontext von Kenntnissen über physikalische Gesetze gebracht werden kann. So mag eine Person *wissen*, dass wenn bei hoher Luftfeuchtigkeit die Temperatur stark fällt, die Feuchtigkeit nicht in der Atmosphäre gehalten wird und es zu Regen kommt. Das Verständnis der Gesetzmäßigkeiten der Zusammenhänge zwischen Regen, Verdunstung, Luftströmungen, Temperaturabfall etc. wird der Ebene der *Weisheit* zugerechnet. Es wurden weitere Definitionen für diese Begriffe vorgeschlagen, die hier nicht vollständig beschrieben werden können. Allen Auffassungen ist jedoch gemeinsam, dass Wissen als eine Gesamtheit organisierter, durch Daten erworbener Information betrachtet wird, die durch Vernunft geleitetes Handeln ermöglicht. Wissen befähigt demzufolge zu zielgerichtetem Verhalten. Kritik an den DIKW-Modellen bezieht sich häufig auf die unklare Differenzierung zwischen den verschiedenen Ebenen. So lässt sich beispielsweise die Frage stellen, ob eine mathematische Gleichung wie 2 + 2 = 4 unter Information oder unter Wissen fällt (Zins, 2007). Auch die Abgrenzung zwischen Wissen und Weisheit erscheint aus ähnlichen Gründen problematisch.

2.3 Wissen in der Psychologie

Wissen ist Gegenstand unterschiedlicher Teilbereiche der Psychologie. So verweisen Beauducel und Süß (2011) auf Bezüge zur Allgemeinen Psychologie, Gedächtnispsychologie, Pädagogischen Psychologie, Entwicklungspsychologie und Differenziellen Psychologie, wobei letztere auch im Zentrum dieser Arbeit steht. Spada und Mandl (1988) differenzieren zwischen vier verschiedenen Teilbereichen der Wissenspsychologie: der Wissensrepräsentation, dem Wissenserwerb, der Wissensanwendung und der Wissensveränderung. *Wissensrepräsentation* umfasst die Organisation des Wissens über Sachverhalte und über Handlungen im Gedächtnis. Spada und Mandl weisen darauf hin, "dass unter dem Begriff 'Wissen' keineswegs nur relativ statisches Faktenwissen im Sinne von Begriffs-Merkmalsstrukturen subsumiert wird, sondern auch Wissen um Ursachen und Wirkungen, algorithmische Fertigkeiten, heuristisches Wissen usw." (S. 2). Unter *Wissenserwerb* wird die Aneignung neuen Wissens verstanden. Neben der Informationssuche, -aufnahme und -verarbeitung werden als Beispiele für Forschungsinhalte hier unter anderem "die Rolle emotionaler und motivationaler Faktoren sowie Effekte der Aufgabenorientierung und des Kontextes" (S. 3) genannt. Der Bereich der *Wissensanwendung* umfasst die Aktivierung und den Abruf vorhandenen Wissens. Als Beispiele für Forschungsbereiche nennen Spada und Mandl hier unter anderem "die induktive Generierung von Hypothesen auf der Basis von Einzelerfahrungen" (S. 3) und "emotionale und motivationale Aspekte einer erfolgreichen (und erfolglosen) Wissensanwendung" (S. 3) Der vierte Teilbereich der *Wissensveränderung* beinhaltet beispielsweise "die Umwandlung von Faktenwissen in Handlungswissen, d.h. die Prozeduralisierung von deklarativem Wissen, wie sie beim Erwerb geistiger Fähigkeiten postuliert wird" (S. 3).

2.3.1 Arten des Wissens

Wie an den Beschreibungen dieser Teilgebiete der Wissenspsychologie bereits deutlich wird, gibt es je nach der Perspektive, aus welcher Wissen betrachtet wird, unterschiedliche Differenzierungen zwischen Wissensarten. Einige der Begriffe, die in der Literatur anzutreffen sind, werden hier näher erläutert. Eine häufig verwendete Differenzierung ist die zwischen deklarativem und prozeduralem Wissen. Die Unterscheidung wurde von Ryle (1949) getroffen, wobei er nicht die Begriffe des deklarativen bzw. prozeduralen Wissens verwendete, sondern zwischen *knowing that* und *knowing how* differenzierte: "We speak of learning how to play an instrument as well as of learning that something is the case" (S. 28). Während bewusstseinsfähige, beschreibbare Inhalte Gegenstand des deklarativen Wissens (knowing that) sind, umfasst das prozedurale Wissen (knowing how) die "Kompetenz oder Fertigkeit, eine kognitive und/oder motorische Operation bzw. Handlung auszuführen" (Süß, 1996, S. 63).

Eine ähnliche Unterscheidung geht auf Broadbent, FitzGerald und Broadbent (1986) zurück. Sie differenzierten zwischen explizitem (bewusstseinsfähigem und verbalisierbarem) Wissen und implizitem (nicht bewusstseinsfähigem und nicht verbalisierbarem) Wissen und veranschaulichten die Unterscheidung anhand von Entscheidungsprozessen. Explizites Wissen äußert sich in einer Analyse der Konsequenzen jeder möglichen Entscheidung. Diese Analyse basiert auf verbalisierbarem Wissen über die entscheidenden Merkmale (*key features*) der aktuellen Problemsituation und dem "knowledge of the structure of the world" (Broadbent et al., 1986, S. 49), welches hier zielgerichtet eingesetzt wird. Eine effektive Entscheidung kann jedoch auch ohne das explizite Wissen getroffen werden, sofern die Person in der Vergangenheit ähnlichen Situationen ausgesetzt war und damit implizites Wissen erworben hat: "Thus, action based on matching the current situation to one from the past will give better than chance performance, although it is incapable of supporting question

answering" (Broadbent et al., 1986, S. 48). Süß (1996) weist darauf hin, dass explizites Wissen deckungsgleich mit deklarativem Wissen ist, während implizites Wissen als eine Teilmenge des prozeduralen Wissens aufgefasst werden kann, welche lediglich das Wissen umfasst, welches "nicht durch Selbstbeobachtung bewusst gemacht und expliziert werden kann" (S. 64).

Tulving (1972) führte eine Unterscheidung zwischen semantischem und episodischem Gedächtnis ein, woraus auch eine Unterscheidung zwischen semantischem und episodischem Wissen folgte. "Episodic memory receives and stores information about temporally dated episodes or events, and temporal-spatial relations among these events" (S. 385). Im Gegensatz dazu bezeichnet Tulving das sematische Gedächtnis als das Gedächtnis, "necessary for the use of language. It is a mental thesaurus, organized knowledge a person possesses about words and other verbal symbols" (S. 386). Auch die Relationen der Begriffe und Symbole zueinander sind dem semantischen Gedächtnis zuzuordnen. Semantisches Wissen bezeichnet also sprachliches Wissen über die Welt, während episodisches Wissen individuelle Erfahrungen des Individuums umfasst.

Eine weitere Wissensform, die in der Literatur vielfältig ausgelegt wird, ist *stilles Wissen*, welches auch als *tacit knowledge* bezeichnet wird. Als zentrale Merkmale des Begriffs beschreiben Beauducel und Süß (2011) den impliziten Erwerb, die schwierige Verbalisierbarkeit und damit die Kategorisierung als nondeklaratives Wissen. Ambrosini und Bowman (2001) weisen in ihrem Artikel zur Operationalisierung stillen Wissens auf die große Anzahl an Synonymen hin, die in der Forschung hierfür vorgeschlagen wurden. Beispiele hierfür wären die Begriffe skill, know-how und 'recipe', was mit "Erfolgsrezept" übersetzt werden könnte. Der interessierte Leser sei auf diesen Artikel verwiesen, der detaillierte Information über weitere Auslegungen des Begriffs enthält.

2.3.2 Wissensdiagnostik

Bezüglich der Diagnostik von Wissen werden hier vier verschiedene Methoden der Wissenserfassung beschrieben, zwischen denen Kluwe (1988) differenziert. Hierbei handelt es sich um lautes Denken (thinking aloud), Befragen (probing), Kategorisieren (sorting) und freie Reproduktion (free recall). Das laute Denken umfasst die laute Aussprache eigener Gedankengänge während der Bearbeitung einer Aufgabe: "Die Methode wird eingesetzt, um im Verlauf eines Lern- oder Lösungsprozesses Daten über das von einer Person aktivierte Wissen und über dessen Veränderungen zu erhalten" (Kluwe, 1988, S. 362). Unter der Methode der Befragung können verschiedene Verfahren zusammengefasst werden, welche sich in zeitlicher Hinsicht (nach einem Lernprozess; während eines Lernprozesses; unabhängig von Lernprozessen) und im Hinblick auf die Spezifität der Instruktion, also dem Ausmaß, in dem "das von der Versuchsperson zu aktivierende Wissen eingegrenzt wird" (Kluwe, 1988, S. 370), unterscheiden können. Die Methode der Kategorisierung beinhaltet die Gruppierung von Reizmaterial. Bei der Methode der freien Reproduktion werden die Versuchspersonen zur Wiedergabe von vorab dargebotenen Informationen aufgefordert. Das Verfahren wird vor allem eingesetzt, um die Struktur und die Organisation vorhandenen Wissens zu untersuchen. Ein Beispiel hierfür ist die Analyse von chunks durch die Reproduktion der Positionen vorgegebener Objekte in einem Raum (McNamara, Hardy und Hirtle, 1989).

Die verschiedenen von Kluwe (1988) beschriebenen diagnostischen Verfahren eignen sich für unterschiedliche Forschungsinhalte und für die Erfassung unterschiedlicher Arten des Wissens. So dürfte beispielsweise die Methode des lauten Denkens für Veränderungen des Wissens während einer Tätigkeit und zur Erfassung (nicht-impliziten) prozeduralen Wissens, oder auch stillen Wissens geeignet sein. Hingegen dürfte die Befragung die Methode der Wahl für die Messung interindividueller Unterschiede des deklarativen Wissens sein.

Beauducel und Süß (2011) stellen die wichtigsten Wissensarten mit den verschiedenen Formen der Erfassung von Wissen in einer Grafik dar, die in Abbildung 1 wiedergegeben ist. Gegenstand der vorliegenden Arbeit ist lediglich die Form des Wissens, die auch Beauducel und Süß in ihrer Abbildung hervorgehoben haben: das semantische deklarative Wissen, das über die Methode des Befragens erfasst wird. Die Befragung findet in standardisierten Testverfahren zur Erfassung des Wissens am häufigsten in Form von Mehrfachwahlaufgaben statt. Auf eine Frage werden hier mehrere konkrete Antwortoptionen geboten, von denen die Person eine auswählen soll. Diese Form der Befragung ist durch hohe Spezifität gekennzeichnet und ist unabhängig von einem bestimmten Lernprozess. Es wird erfasst, ob die Person eine Frage beantworten kann, nicht aber, woher sie das hierfür benötigte Wissen hat bzw. welche kognitiven Prozesse dem Erwerb oder dem Abruf des Wissens zugrunde liegen.

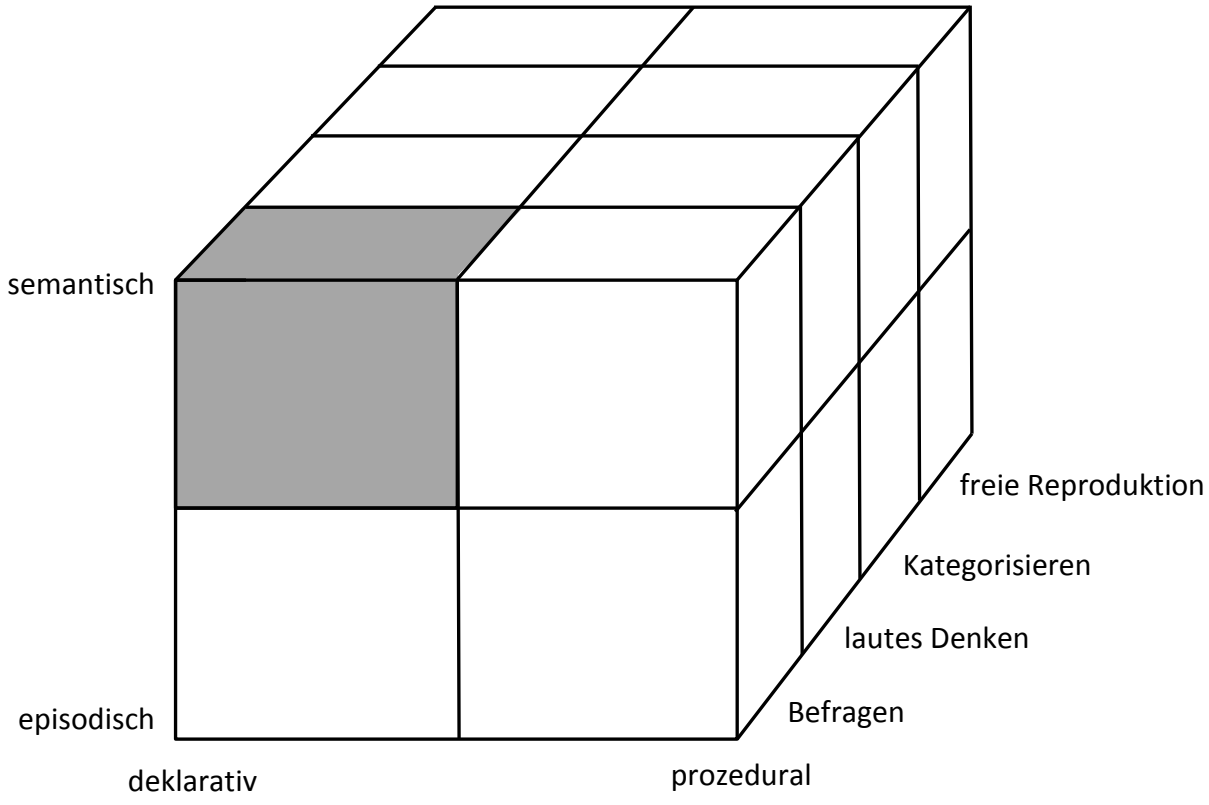

Abbildung 1. Wissensarten und Methoden der Wissensmessung.

Die Verwendung des Begriffs Wissen beinhaltet somit im Folgenden – soweit nicht anders vermerkt – das deklarative semantische Wissen. In den meisten standardisierten Tests, die zur Erfassung dieser Art des Wissens Wissens dienen, werden Mehrfachwahlaufgaben verwendet, "because of their high reliability, objectivity, and straightforward scoring rubrics" (Liu, Lee & Linn, 2011, S. 1083).

Es folgt an dieser Stelle eine Behandlung der Merkmale von Multiple-Choice Wissenstests im Hinblick auf die drei Hauptgütekriterien Objektivität, Reliabilität und Validität. Als Beispiele werden – soweit verfügbar – jeweils die numerischen Ausprägungen der einzelnen Kriterien für den Wissenstest des I-S-T 2000 R (Liepmann et al., 2007) und für den BOWIT (Hossiep & Schulte, 2008) angegeben, da diese Tests in der vorliegenden Arbeit weitere Verwendung fanden. Hierbei besteht kein Anspruch auf Vollständigkeit. In den Manualen der Tests finden sich detailliertere Informationen.

Die hohe Auswertungs- und Durchführungsobjektivität von Wissenstests, die aus Mehrfachwahlaufgaben bestehen, dürften selbsterklärend sein. Bei eindeutigen Instruktionen und klaren Anweisungen, wie beispielsweise bei nachträglichen Korrekturen der Versuchspersonen vorzugehen ist, erreicht die Auswertungsobjektivität ein Maximum. Im Hinblick auf die Durchführungsobjektivität kann eine Maximierung angestrebt werden, indem möglichst detaillierte Anweisungen zu mündlichen Instruktionen und zum Umgang mit speziellen Situationen, wie beispielsweise Verständnisfragen von Seiten der Versuchspersonen, umzugehen ist. Die Interpretationsobjektivität ist im Wesentlichen an die Inhaltsvalidität gebunden, die auf Seite 26 näher beschrieben wird.

Wissenstests weisen bei der Schätzung der Reliabilität spezielle Schwierigkeiten auf. Hier muss zwischen verschiedenen Reliabilitätsmaßen unterschieden werden: so sind die interne Konsistenz, die Retest-, Paralleltest- und Split-half-Reliabilität bei thematisch breit angelegten Wissenstests unterschiedlich gut geeignet. Eine niedrige interne Konsistenz lässt

speziell bei inhaltlich heterogenen Tests nicht zwangsläufig auf eine geringe Reliabilität schließen. Dennoch weisen Tests des Allgemeinen Wissens häufig Werte für Cronbachs α auf, die mit den Werten von Intelligenztests mit wesentlich homogeneren Aufgaben, wie beispielsweise Matrizentests, vergleichbar sind. So beträgt Cronbachs α für den Wissenstest des I-S-T 2000 R α = .93 und für den BOWIT α = .95. Die Prüfung der Paralleltest-Reliabilität ist bei der Testentwicklung und die Prüfung der Retest-Reliabilität bei der Datenerhebung mit einem hohen Aufwand verbunden. Für die Retest-Reliabilität ergibt sich bei Wissenstests zusätzlich das Problem, dass Wissen ein psychologisches Merkmal ist, das zeitlichen Veränderungen unterliegt. Die Schätzung der Reliabilität in Form von zeitlicher Stabilität ist bei der Erfassung eines zeitlich veränderlichen Merkmals problematisch. Dieses Problem lässt sich theoretisch mit einem möglichst kurzen zeitlichen Abstand zwischen den beiden Erhebungen minimieren. Das bringt jedoch wiederum die Schwierigkeit mit sich, dass Erinnerungen an die erste Testdurchführung bei der zweiten Durchführung relevant werden können und somit eine systematische Überschätzung der Reliabilität erfolgen kann. Für den Wissenstest des I-S-T 2000 R wurden die Parallel- und die Retest-Reliabilität nicht berechnet. Beim BOWIT beträgt die Paralleltestreliabilität r = .91 und die Retest-Reliabilität r = .96. Bei der Schätzung der Reliabilität durch die Split-half-Methode ist bei Wissenstests speziell zu beachten, dass beide Testhälften zu jedem Thema jeweils die gleiche Anzahl an Items enthalten. Sofern in den beiden Testhälften unterschiedliche Themen dominieren, kann dies zu einer systematischen Unterschätzung der Reliabilität führen. Für die Skalen des Wissenstests des I-S-T 2000 R liegen die Schätzungen der (nach der Spearman-Brown-Formel korrigierten) Split-half-Reliabilität zwischen r = .84 (Skala für figurales Wissen) und r = .93 (Gesamtskala). Beim BOWIT wurde eine Korrelation von r = .96 zwischen den beiden Testhälften berechnet, wobei hier auf eine Korrektur nach Spearman-Brown verzichtet wurde.

Bezüglich der Validität geben Beauducel und Süß (2011) ausführliche Informationen zu den für Multiple-Choice Wissenstests typischen Merkmalen. Insbesondere die Inhaltsvalidität wird hier ausführlicher behandelt, da sie im Rahmen der Hypothesenbildung der vorliegenden Arbeit von zentraler Bedeutung ist. Die Inhaltsvalidität lässt sich für Wissenstests bestimmen, sofern ein eindeutig definiertes Itemuniversum vorliegt. Als Beispiel geben Beauducel und Süß die vier Grundrechenarten an. Hier ist eindeutig feststellbar, ob ein Wissenstest, der die Beherrschung letzterer erfassen soll, inhaltsvalide ist. Je heterogener die Themen sind, zu denen das Wissen abgefragt werden soll, desto schwieriger lässt sich ein Itemuniversum und damit die Inhaltsvalidität bestimmen. Die Entwickler von Wissenstests begegnen dem Problem teilweise durch die Anwendung der sogenannten *curricularen Validität* (Beauducel & Süß, 2011). Das Itemuniversum, das als Grundlage für die Itemerstellung dient, ist dann durch Lehrpläne, die beispielsweise an Schulen Verwendung finden, definiert. Der Umstand, dass sich die Curricula unterschiedlicher Schulen unterscheiden, erschwert die Erfassung der Inhaltsvalidität auf diesem Weg. Eine Beschränkung auf die Inhalte, die die Curricula unterschiedlicher Schulen gemeinsam haben, wäre möglich, würde jedoch die thematische Heterogenität des Tests einschränken. Hossiep und Schulte (2008) verwendeten für die Beurteilung der Inhaltsvalidität des BOWIT die Korrelationen zwischen Skalen einzelner Domänen, zu denen der BOWIT Items enthält, und den Interessen der Versuchspersonen an den entsprechenden Domänen. Jener Erfassung der Inhaltsvalidität liegt die Annahme zugrunde, dass die Interessen einer Person von maßgeblicher Bedeutung für die Auswahl der Domänen sind, mit denen sich die Person beschäftigt und in denen sie somit Wissen erlangt. Eine hohe Korrelation deutet somit nach Meinung von Hossiep und Schulte auf eine hohe Inhaltsvalidität hin. Sämtliche im Zusammenhang mit der Inhaltsvalidität berechneten Korrelationen des BOWIT sind positiv und auf dem 1‰-Niveau signifikant. Das Minimum betrug $r = .19$ für die Skala Ernährung/Bewegung/Gesundheit (Form A) und das Maximum

r = .54 für die Skalen Mathematik/Physik (Form A) und Technik/EDV (Form B). Zur Inhaltsvalidität des Wissenstests des I-S-T 2000 R geben Liepmann et al. (2007) keine Informationen. Wie auf Seite 24 bereits angemerkt, hängt die Interpretationsobjektivität eng mit der Inhaltsvalidität zusammen. Aufgrund der für Tests des Allgemeinen Wissens charakteristischen Schwierigkeiten bezüglich der Inhaltsvalidität sind auch der Interpretationsobjektivität von derartigen Tests Grenzen gesetzt. So kann in Frage gestellt werden, ob die in einem Wissenstest erbrachten Leistungen (ausschließlich) auf das Allgemeine Wissen zurückzuführen sind. Dieses Thema wird auch in der vorliegenden Arbeit in Kapitel 5 aufgegriffen.

Für die Beurteilung der Kriteriumsvalidität von Wissenstests werden häufig Korrelationen mit Schulleistungen, beruflichen Leistungen und anderen Wissenstests verwendet. Liepmann et al. (2007) berichten über Korrelationen zwischen Schulnoten und der mit dem I-S-T 2000 R erfassten Kristallinen Intelligenz, für die der Wissenstest des I-S-T Markiervariablen enthält[1]. Die Korrelationen zwischen Kristalliner Intelligenz und Schulnoten liegen zwischen -.29 (Mathematik) und -.38 (Erdkunde). Außerdem korreliert gc zu .68 mit dem Wissenstest des revidierten Hamburg-Wechsler-Intelligenztests für Erwachsene (HAWIE-R). Bezüglich der Kriteriumsvalidität des BOWIT berichten Hossiep und Schulte (2008) beispielsweise über eine Korrelation von -.28 zwischen den Rohwerten im BOWIT und der Abiturnote. Die Rohwerte des BOWIT korrelieren außerdem zu .24 mit der Anzahl der abgeschlossenen Studiengänge und zu .17 mit dem Bruttoentgelt im Jahr.

[1]Liepmann et al. (2007) verweisen darauf, dass der Wissenstest des I-S-T 2000 R Markiervariablen der Kristallinen Intelligenz enthält, nicht aber mit ihr gleichzusetzen sei. Für die Erfassung der Kristallinen Intelligenz bedarf es Liepmann et al. zufolge des Einsatzes des vollständigen I-S-T 2000 R.

Die Prüfung der Konstruktvalidität beinhaltet die Prüfung von Hypothesen, die sich auf die Struktur des Tests beziehen. Hierzu gibt es vielfältige Ansätze. Im vorherigen Abschnitt zur Kriteriumsvalidität wurden unter anderem Wissenstests als Kriterium erwähnt. Im Rahmen der Konstruktvalidität können zusätzlich weitere Leistungstests herangezogen werden, bei welchen erwartet wird, dass sie im Vergleich zu anderen Wissenstests gering mit dem Wissenstest korrelieren. Dies ermöglicht die Prüfung der konvergenten und der diskriminanten Validität. Als Beleg der diskriminanten Validität wird im Manual des I-S-T 2000 R über eine Korrelation zwischen der Kristallinen Intelligenz und dem d2 von -.09 berichtet. Da die Kristalline Intelligenz zu .68 mit dem Wissenstest des HAWIE-R korreliert, wird auch die konvergente Validität als gegeben angesehen. Weitere Beispiele für Verfahren, die zur Prüfung der Konstruktvalidität verwendet werden, sind exploratorische und konfirmatorische Faktorenanalysen. Eine mit dem BOWIT durchgeführte exploratorische Hauptkomponentenanalyse mit Oblimin-Rotation ergab eine 2-faktorielle Struktur, wobei die Faktoren von Hossiep und Schulte (2008) als "Gesellschafts- und Geisteswissenschaftliches Wissen" (S. 28) und als "naturwissenschaftlich-technisches Wissen" (S. 29) bezeichnet wurden. Die Höhe der Korrelation zwischen den beiden Faktoren wird nicht angegeben. Im Manual des BOWIT wird nicht Stellung dazu genommen, wie eine 2-faktorielle Struktur mit der Absicht, einen Test des Allgemeinen Wissens zu erstellen, zu vereinbaren ist. Dementsprechend sollte auch erwähnt werden, dass die beschriebene Hauptkomponentenanalyse im Manual des BOWIT nicht im Rahmen der Überprüfung der Konstruktvalidität, sondern im Rahmen der Testkonstruktion behandelt wird. Liepmann et al. (2007) haben nicht die Konstruktvalidität des Wissenstests, wohl aber die Konstruktvalidität des gesamten I-S-T 2000 R in Form einer exploratorischen und einer konfirmatorischen Faktorenanalyse geprüft. Hierbei zeigte sich erwartungsgemäß eine 2-faktorielle Struktur, wobei der eine Faktor als Fluide Intelligenz und der andere Faktor als Kristalline Intelligenz

interpretiert werden kann. Der Wissenstest zeigt hierbei höhere Faktorladungen auf dem Faktor der Kristallinen Intelligenz. Im Folgenden wird die Beziehung zwischen Allgemeinem Wissen und dem Konstrukt der Kristallinen Intelligenz näher erläutert.

2.3.3 Wissen und Intelligenz

Tests des deklarativen semantischen Wissens werden häufig zur Erfassung der Kristallinen Intelligenz eingesetzt. Auch die Wissenstests, welche in den Untersuchungen dieser Arbeit zum Einsatz kamen – der Wissenstest des I-S-T 2000 R und der BOWIT – können ihren Autoren zufolge hierfür verwendet werden. Wie in Kapitel 2.3.2 erläutert, ist der Wissenstest des I-S-T 2000 R nach Liepmann et al. (2007) jedoch nicht mit Kristalliner Intelligenz gleichzusetzen, sondern enthält lediglich die Markiervariablen. Auch bezeichnen die Autoren die mit dem Wissenstest des I-S-T erfasste Fähigkeit nicht explizit als *Allgemeines* Wissen. Jedoch soll hiermit "die Fähigkeit zum Wissenserwerb in unserer Kultur" (Liepmann et al., 2007, S. 17) erfasst werden. Außerdem soll er "möglichst viele Themenbereiche umfassen, um den Anteil sehr spezifischer Interessen und Lerngelegenheiten an Wissenserfassung (-abbildung) zu reduzieren" (S. 49). Mit Ausnahme der Kulturspezifität wird somit offenbar keine thematische Eingrenzung angestrebt, weshalb die Fähigkeit, zu deren Erfassung der Test entwickelt wurde, wohl auch als *Allgemeines* Wissen bezeichnet werden kann.

Der Begriff der Kristallinen Intelligenz geht auf Cattell (1943) zurück. Cattell entwickelte, später in Zusammenarbeit mit Horn (z.B. Horn & Cattell, 1966, 1967, 1982), ein Modell der Intelligenz, welches als eine Synthese der 2-Faktoren-Theorie von Spearman (1904) und der Theorie mehrerer Faktoren von Thurstone (1938) aufgefasst werden kann (Amelang & Bartussek, 2001). Daher sollen im Folgenden auch kurz die Theorien von Spearman und Thurstone erläutert werden. Spearman veröffentlichte 1904 eine Arbeit über die korrelativen Zusammenhänge zwischen Schulleistung und sensorischer Diskrimination,

die später die Grundlage der 2-Faktoren-Theorie der Intelligenz bildete (siehe Carroll, 1993). Dieser Theorie zufolge sind für die Lösung einer Aufgabe aus dem kognitiven Leistungsbereich jeweils ein aufgabenspezifischer Faktor und der sogenannte g-Faktor von Bedeutung. Der g-Faktor wurde von Spearman als *general intelligence* bezeichnet und umfasst sowohl angeborene als auch auf Erfahrungen basierende intellektuelle Fähigkeiten, wie beispielsweise Schulleistungen. Demgegenüber stand ein Modell mehrerer Faktoren von Thurstone, das über viele Jahre als nicht vereinbar mit Spearmans 2-Faktoren-Theorie angesehen wurde (Carroll, 1993). Thurstone (1938) führte mit verschiedenen kognitiven Leistungsaufgaben Faktorenanalysen durch und verwendete hierbei orthogonale Rotationsmethoden. Hierbei wurden sieben Primärfaktoren (*primary mental abilities*) identifiziert, die Thurstone mit S (Space), P (Perceptual Speed), N (Number Facility), V (Verbal Relations), W (Word Fluency), M (Memory) und I (Induction) bezeichnete. Während Thurstone zu Beginn die Unabhängigkeit der Faktoren postulierte, räumte er ab 1947 ein, dass Korrelationen zwischen ihnen möglich wären und dass diese auf Spearmans g-Faktor gegründet sein könnten (siehe Carroll, 1993). Ein Beispiel für einen Intelligenztest, in dem sowohl der g-Faktor von Spearman als auch, auf spezifischerer Ebene, eine Reihe der Faktoren von Thurstone Berücksichtigung finden, ist der Berliner-Intelligenzstruktur-Test (BIS) von Jäger, Süß und Beauducel (1997). Das Wissen wird mit dem Test nicht erfasst.

Cattell (1943) stellte das Modell der Fluiden und der Kristallinen Intelligenz vor. Die Fluide Intelligenz (gf) beschreibt die Fähigkeit zur Lösung von neuen Problemen, für die in der Vergangenheit erworbene Fähigkeiten nicht von Bedeutung sind. Sie ist biologisch und physiologisch bedingt und wird durch das Erleben und Verhalten des Individuums in der Vergangenheit nicht beeinflusst. Hiervon ist die Kristalline Intelligenz (gc) abzugrenzen. Die Kristalline Intelligenz basiert auf Wissen und Erfahrungen, die der Mensch durch die Akkulturation im Laufe seines Lebens sammelt. Cattell (1963) berichtete über die Ergebnisse

einer exploratorischen Faktorenanalyse mit obliquer Rotation, in die zahlreiche kognitive Fähigkeitstests – darunter auch Tests zur Erfassung von fünf primary mental abilities nach Thurstone – sowie Tests zur Erfassung von Persönlichkeitsmerkmalen, die nicht Fähigkeiten zuzurechnen sind, Eingang fanden. Die Tests wurden Jugendlichen der siebten und achten Klasse vorgelegt. Das Ergebnis der Faktorenanalyse beinhaltete unter anderem zwei Faktoren, die miteinander zu .47 korrelierten und auf Basis der Faktorladungen eindeutig als Fähigkeiten interpretiert werden konnten. Die Faktorladungen der einzelnen Fähigkeitstests veranlassten Cattell (1963) außerdem zur Interpretation der beiden Faktoren als gf und gc. Während die Fluide Intelligenz vorwiegend für Tests relevant ist, die neue, unbekannte Problemstellungen beinhalteten, "crystallized ability loads more highly those cognitive performances in which skilled judgment habits have become crystallized (whence its name) as the result of earlier learning application of some prior, more fundamental general ability to these fields" (Cattell, 1963, S. 2 f.). Auch in Untersuchungen mit Personen anderer Altersgruppen wurden positive Korrelationen zwischen gf und gc gefunden, welche im Durchschnitt etwa .30 bis .40 betrugen (Cattell, 1971/1987). Den Zusammenhang zwischen beiden Fähigkeiten erklärte Cattell (1963, 1971/1987) mit der Investmenttheorie: Diese Theorie bezieht sich hauptsächlich auf die Entwicklung kognitiver Fähigkeiten bei Kindern und Jugendlichen. Kristalline Intelligenz entsteht durch die Investition von Fluider Intelligenz in intellektuelle Tätigkeiten, die vor allem in der Schule gefordert werden. Allerdings ist die Fluide Intelligenz nicht das einzige Merkmal, das für die Kristalline Intelligenz bedeutsam ist, "because years at school, interest in school work and other influences will also determine, perhaps substantially, the level of crystallized abilities" (Cattell, 1971/1987, S. 139). Auch Persönlichkeitsmerkmale wie Extraversion und Selbstvertrauen sind hier von Bedeutung (Cattell, 1963). Durch die Fluide Intelligenz und durch die Investitionen in komplexe Lernsituationen, deren Ausmaß durch Interessen und die genannten Persönlichkeitsmerkmale

bestimmt wird, erwirbt der Mensch Erfahrungen. Die hierauf basierende kognitive Leistungsfähigkeit wird als Kristalline Intelligenz bezeichnet. Je umfangreicher der Erfahrungsschatz und das Wissen einer Person sind, desto höher ist ihre Kristalline Intelligenz. Fluide Intelligenz erleichtert den Erwerb von Wissen. Je höher die Fluide Intelligenz einer Person ist, umso größer ist – unter sonst gleichen Bedingungen – die Menge an Informationen, die sie sammeln, verarbeiten und lernen kann. Daher korrelieren gf und gc positiv miteinander. In der Investmenttheorie differenzierte Cattell (1971/1987) explizit zwischen früheren und aktuellen Interessen der Person. Für die Kristalline Intelligenz sind die früheren und nicht die aktuellen Interessen relevant. Frühere Interessen entscheiden darüber, in welche Themen eine Person in der Vergangenheit investiert hat und in welchen Themen sie somit Wissen erlangt hat. Die Interessen, die zum Zeitpunkt der Messung von Kristalliner Intelligenz erst seit kurzer Zeit bestehen, haben in der Vergangenheit nicht zur Investition von Zeit und Fluider Intelligenz in die entsprechenden Themen beigetragen. Für die Kristalline Intelligenz zum Zeitpunkt der Messung sind sie daher irrelevant. Auf diese Annahme von Cattell (1971/1987) wird in Kapitel 5.1.2 noch näher eingegangen.

Sowohl unter gf als auch unter gc verstand Cattell Generalfaktoren. Wie im vorigen Absatz beschrieben, bezieht sich die Investmenttheorie hauptsächlich auf Investitionen während der Schulzeit. Ein zentrales Problem sah Cattell in der Messung Kristalliner Intelligenz nach Abschluss der Schule: "The crystallized intelligence factor then goes awry both conceptually and in regard to the practical predictions to be made from traditional intelligence tests" (Cattell, 1971/1987, S. 143). Kristalline Intelligenz lässt sich demzufolge somit auf beliebig viele Themenbereiche erweitern, weshalb kein einzelnes Themengebiet herangezogen werden kann, um gc in der Population zu messen. Für die Messung bei Kindern und Jugendlichen favorisierte Cattell (1963) die schulischen Leistungen. Was die Messung bei Erwachsenen betrifft, sah er 1971/1987 die folgenden Möglichkeiten: (a) die

Erweiterung der thematischen Breite der Tests "(an approach which, in practice, might amount to producing as many different tests as there are occupations, etc.)" (S. 143), (b) die Beschränkung auf die Erfassung der Fluiden Intelligenz und (c) das Festhalten an der Erfassung der Kristallinen Intelligenz in der Form, wie sie zum Ende der Schulzeit vorliegt, auch bei Erwachsenen. Wie in Kapitel 2.3.2 erläutert, wird die Inhaltsvalidität von Wissenstests teilweise über die curriculare Validität angestrebt. Hierin kann eine Wahl der Option (c) von Cattell gesehen werden. Horn (1988) beschreibt die Kristalline Intelligenz als "large number of behaviors that point to an individual’s breadth of knowledge, experience, sophistication, judgment, skills of communication, understanding of conventions and capacity for reasonable thinking" (S. 658). Horn betont hier somit explizit die Breite des Wissens. Als Fähigkeiten, die als Indikatoren für gc angesehen wurden, nennt Horn hier unter anderem "Verbal Knowledge" (S. 659) und "Information about the Humanities, Social and Physical Science, Business and Culture in General" (S. 659). Horn weist für die Messung von gc nicht auf eine Beschränkung auf Tests hin, die sich auf schulisches Material konzentrieren. Tests der Kristallinen Intelligenz sind in der Regel thematisch heterogen und nicht auf das Wissen beschränkt, das im Rahmen der Schule erworben werden kann. Thematische Heterogenität zeigt sich auch bei Betrachtung empirischer Ergebnisse bezüglich der Entwicklung der Kristallinen und der Fluiden Intelligenz über die Lebensspanne. So zeigt Horn (1988), dass die Kristalline Intelligenz im Laufe des Lebens über einen vergleichsweise langen Zeitraum hinweg gesteigert werden kann. Die Fluide Intelligenz nimmt im Gegensatz dazu etwa ab Mitte des dritten Lebensjahrzehnts ab, was mit der Investmenttheorie konform geht, da dieser zufolge die Kristalline Intelligenz im Gegensatz zur Fluiden Intelligenz erfahrungsbasiert ist. Würde sich die Messung der Kristallinen Intelligenz ausschließlich auf das zum Ende der Schulzeit vorhandene Wissen konzentrieren, wäre eine derartige Entwicklung abwegig. Auch der BOWIT und der Wissenstest des I-S-T 2000 R erfassen Wissen in einem breiten Kontext.

So enthält der Wissenstest des I-S-T beispielsweise Items zu den Themenbereichen Wirtschaft und Alltag. Der BOWIT erfasst unter anderem das Wissen zu den Themen Verkehr, Archäologie, Ernährung und Wirtschaft. Hierbei handelt es sich um Themen, die nicht, oder zumindest nicht länderübergreifend, als schulische Fächer existieren.

Horn und Noll (1997) verweisen darauf, dass das Modell der Fluiden und Kristallinen Intelligenz seit Beginn seiner Entwicklung modifiziert worden ist: "At first it was a theory of two intelligences. Today, it would be better labeled the theory of several intelligences" (S. 82). Hierbei handelt es sich, neben gf und gc, um "visual intelligence (Gv), auditory intelligence (Ga), the intelligence of short-term apprehension and retention (SAR), the intelligence of fluency and retrieval from long-term storage (TSR), and quantitative intelligence (Gq)" (S. 82) sowie um "process of intellectual speed (Gs), and speed of decision making (QDS)" (S. 82). Ein intensiv untersuchtes Merkmal der genannten Fähigkeiten ist ihre Entwicklung über die Lebensspanne. Bezüglich der Fluiden und der Kristallinen Intelligenz berichten Horn und Noll (1997) über ähnliche Ergebnisse wie Horn (1988). Für die Leistungen des Langzeitgedächtnisses (TSR) findet sich ein Entwicklungsverlauf, der eng an den der Kristallinen Intelligenz angelehnt ist. Diese Fähigkeiten, die in weiterentwickelten Versionen des Modells der Fluiden und Kristallinen Intelligenz aufgezeigt wurden, sind für die vorliegende Arbeit jedoch nicht von Bedeutung. Der Fokus liegt hier ausschließlich auf der Kristallinen Intelligenz.

Ackerman (1996) entwickelte die sogenannte PPIK-Theorie, in die Intelligenz, Persönlichkeitsmerkmale und Wissen Eingang fanden, und welche deutlich an die Investmenttheorie von Cattell angelehnt ist. Als zentrale Merkmale beschreiben Ackerman und Beier (2004) (a) die Auffassung von Intelligenz als gf und gc (Intelligence-as-*P*rocess), (b) die Berücksichtigung von "trait complexes" (Ackerman & Beier, 2004, S. 130), welche mit der Intelligenz und der Entwicklung von Intellektualität zusammenhängen (*P*ersonality

und *I*nterests) sowie (c) die Berücksichtigung von Wissen (Intelligence-as-*K*nowledge). Wie aus dem ersten Merkmal hervorgeht, werden Fluide Intelligenz und Kristalline Intelligenz als Intelligence-as-Process betrachtet. Ackerman (2000b) distanziert sich hier explizit von der Trennung zwischen gf und gc: "There is an inherent danger in segregating process from knowledge, just as has been demonstrated in the attempts by Cattell and his followers to separate gf from gc" (S. 443). Nach Ackerman (2000b) dürften manche Tests in Abhängigkeit von der Stichprobe unterschiedliche Ladungen auf den als gf und gc bezeichneten Faktoren aufweisen. Einen Beleg für diese Annahme über Cattells Theorie gab Ackerman (2000b) allerdings nicht. Ackerman (2000b) betrachtet Intelligence-as-Process als "speeded aspects of intelligence that decline during normal adults development" (S. 443). Hierunter fallen somit auch Schlussfolgerndes Denken, Gedächtnisleistungen, Wahrnehmungsgeschwindigkeit und räumliche Rotation. Sämtliche Fähigkeiten, die Ackerman unter Intelligence-as-Process zusammenfasst, erreichen ihren Höhepunkt im jungen Erwachsenenalter. Es bleibt anzumerken, dass damit ein wesentliches Merkmal der Kristallinen Intelligenz nicht gegeben ist. Als wichtige Persönlichkeitsfaktoren nennen Ackerman und Beier (2004) zum einen Neurotizismus und Testangst, die mit Intelligence-as-Process korrelieren und zum anderen Typical Intellectual Engagement (Goff & Ackerman, 1992) und Offenheit, die mit Intelligence-as-Knowledge korrelieren. Bezüglich der Interessen orientiert sich Ackerman an dem RIASEC-Modell von Holland (1959, 1997). Eine nähere Beschreibung hiervon findet sich in Kapitel 5.1.1 der vorliegenden Arbeit. Für die Erläuterung des PPIK-Modells reicht hier zunächst der Hinweis aus, dass Holland zwischen sechs Interessengebieten differenziert, welche beispielsweise Rolfhus und Ackerman (1996) zufolge in unterschiedlichen Ausmaßen mit Intelligence-as-Process und Intelligence-as-Knowledge korrelieren.

Die Konstrukte Intelligence-as-Knowledge und Kristalline Intelligenz unterscheiden sich Ackerman und Beier (2004) zufolge vor allem in der Breite: "That is, there are as many domains of knowledge that could potentially be assessed as there are occupations, avocations, and life skills" (S. 132). Um einen umfassenden Überblick über die Ausprägung der Kristallinen Intelligenz bei einer einzelnen Person, die einen bestimmten Beruf ausübt, zu bekommen "one would have to provide for two types of tests: a deep test of professional knowledge and a broad array of more shallow tests outside the profession" (Ackerman, 2000b, S. 445). Wie im Rahmen der Investmenttheorie beschrieben, sah Cattell (1971/1987) eine Möglichkeit der Messung der Kristallinen Intelligenz bei Erwachsenen in thematisch breit angelegten Tests. Die Messung von Intelligence-as-Knowledge ist mit dieser Alternative von Cattell vergleichbar, weshalb Intelligence-as-Knowledge auch als Synonym für Kristalline Intelligenz betrachtet werden kann. Lediglich die von Ackerman und Beier vorgeschlagene hohe Detailliertheit eines Wissenstests zu berufsrelevantem Wissen einerseits und die geringere Detailliertheit eines Wissenstests zu anderen Themengebieten andererseits bildet hier einen Unterschied. Eine solche Herangehensweise würde allerdings praktische Schwierigkeiten mit sich bringen: Die Interessen eines Menschen entscheiden auch über seine Berufswahl und damit über das Thema, zu dem ein besseres Wissen erwartet würde. Damit würde auch bestimmt, zu welchem Themengebiet ein Test mit höheren Anforderungen vorgelegt wird und welche Themengebiete folglich die übrigen sind, bei denen Tests mit geringer Detailliertheit zur Anwendung kommen würden, um Intelligence-as-Knowledge zu messen. Da unterschiedliche Berufe in unterschiedlichem Ausmaß Wissen erfordern, würden hiermit erhebliche diagnostische Schwierigkeiten im Rahmen der Normierung einhergehen. Das Problem spiegelt sich auch in den verschiedenen Arbeiten von Ackerman und Kollegen wider, da in sämtlichen empirischen Untersuchungen zur PPIK-Theorie für die Messung von Intelligence-as-Knowledge thematisch breit angelegte Wissenstests verwendet wurden. Die

Übereinstimmung von Intelligence-as-Knowledge und Kristalliner Intelligenz findet auch hierdurch Bestätigung. Aus den genannten Gründen werden in der vorliegenden Arbeit im Folgenden die PPIK-Theorie und das Konstrukt Intelligence-as-Knowledge nicht gesondert behandelt. Sämtliche Aussagen zur Kristallinen Intelligenz lassen sich auch auf das Konstrukt Intelligence-as-Knowledge übertragen. Ebenso werden Investmenttheorie und PPIK-Theorie als hinreichend parallel betrachtet, um nicht separat behandelt werden zu müssen.

2.3.4 Wissen und Weisheit

Ein weiteres Konstrukt, dessen Zusammenhang mit Wissen hier näher betrachtet wird, ist Weisheit. Zwei Auffassungen dieses Begriffs in der Psychologie stammen von Baltes und Smith (1990) und von Sternberg (1998) und werden im Folgenden nacheinander behandelt.

Die Perspektive von Baltes und Smith (1990) ist eng mit dem Begriff des Wissens verknüpft: Sie fassen Weisheit als "*expert knowledge involving good judgment and advice in the domain, fundamental pragmatics of life*" (S. 95) auf. Als Kriterien der Weisheit werden "rich factual knowledge, rich procedural knowledge, life span contextualism, relativism" (S. 95) und die Fähigkeit "to understand and manage uncertainty" (S. 96) genannt. Demnach lässt sich Weisheit eindeutig als eine Fähigkeit beschreiben. Rich factual knowledge bezeichnet stark ausgeprägtes Faktenwissen, das im Langzeitgedächstnis gespeichert ist und das mit einer "multiple cross-referenced encyclopedia" (Baltes & Smith, 1990, S. 100) verglichen werden kann. Hier findet sich also ein eindeutiger Bezug zum deklarativen semantischen Wissen, welches auch als ein Merkmal der Kristallinen Intelligenz begriffen werden kann und Gegenstand der hier betrachteten Tests des Allgemeinen Wissens ist. Rich procedural knowledge beinhaltet das prozedurale Wissen, welches ebenfalls bereits in Kapitel 2.3.1 beschrieben wurde. Baltes und Smith (1990) fassen es zusammen als ein Repertoire an mentalen Prozessen, die für die Auswahl und Strukturierung von Informationen verwendet

werden, und die damit zur Entscheidungsfindung dienen. Life span contextualism beinhaltet "understanding that life development and life events are embedded in multiple life span contexts . . . involving thematic and temporal relationships" (S. 102). Unter relativism verstehen Baltes und Smith die Kenntnis von individuellen und kulturellen Unterschieden, beispielsweise in Bezug auf Werte, Interessen und Prioritäten. Mit uncertainty ist das Wissen über Einschränkungen der Planbarkeit und Vorhersagbarkeit des Lebens gemeint sowie die Einsicht, dass das Wissen über ein Problem oder über ein Individuum niemals vollständig sein kann. Die beschriebenen fünf Kriterien bilden gemeinsam die Weisheit nach der Definition von Baltes und Smith (1990). Das deklarative semantische Wissen, das Gegenstand der vorliegenden Arbeit ist, bildet ein Kriterium. Eine hohe Ausprägung hierin ist eine Voraussetzung für Weisheit. Hinzu kommen hier jedoch weitere Formen des Wissens in Bezug auf unterschiedliche Lebenssituationen, auf interindividuelle Unterschiede und auf Einschränkungen der Vorhersagbarkeit. Hohes deklaratives semantisches Wissen alleine ist somit Baltes und Smith zufolge nicht hinreichend für Weisheit.

Sternberg (1998) fasst Weisheit als eine Form des stillen Wissens auf, welches auch in Kapitel 2.3.1 beschrieben wurde. Nach Sternberg besteht die Weisheit einer Person in der Anwendung ihres stillen Wissens zur Ausbalancierung (a) intrapersonaler, (b) interpersonaler und (c) extrapersonaler Interessen. Intrapersonale Interessen umfassen Ziele der Person selbst. Interpersonale Interessen bestehen in Zielen anderer Personen. Extrapersonale Interessen können nicht einzelnen Personen zugeordnet werden. Sie sind situations- und kontextbezogen. Beispiele hierfür wären die Stadt, das Land oder die Religion einer Person. Auf Weisheit basierendes Handeln ist nicht vollständig auf die Interessen einer der drei Parteien ausgerichtet, sondern hat die ausgewogene Berücksichtigung der Interessen aller Parteien zum Ziel. Die Weisheit besteht darin, einen Lösungsweg zu finden, der das

gemeinsame Wohl aller Betroffenen – Sternberg bezeichnet dies als *common good* – zum Ziel hat.

Weisheit nach der Auffassung von Sternberg (1998) ist somit ein Merkmal, das nicht eindeutig als Fähigkeit klassifiziert werden kann. Als Quellen interindividueller Unterschiede in Weisheit nennt Sternberg beispielsweise die Ziele (*goals*) einer Person. Während manche Menschen hauptsächlich nach der Umsetzung eigener Ziele, wie beispielsweise der beruflichen Karriere, streben, messen andere Menschen dem common good größere Bedeutung bei. Jene Eigenschaft ist von zentraler Bedeutung für Weisheit, ist jedoch kein Leistungs-, sondern vielmehr ein Persönlichkeitsmerkmal. Weitere Quellen für interindividuelle Unterschiede sind nach Sternberg (2000) beispielsweise Wissen, kreative und analytische Fähigkeiten, aber auch die Motivation für weises Handeln und andere Persönlichkeitsmerkmale. Es sind also sowohl Fähigkeitskomponenten als auch Persönlichkeitseigenschaften, durch welche die individuelle Ausprägung der Weisheit einer Person bestimmt wird. Sternberg (2000) verweist auf eine Studie von Staudinger, Smith und Baltes (1992), bei der keine Unterschiede zwischen jüngeren und älteren Erwachsenen in der Weisheit festgestellt wurden. Diese Ergebnisse legen Sternberg (2000) zufolge nahe, dass Weisheit stärker mit Kristalliner als mit Fluider Intelligenz zusammenhängen dürfte.

Worin besteht nun der Zusammenhang zwischen Wissen und Weisheit nach der Auffassung von Sternberg? Sternberg (2000) beschreibt sieben Prozesse, die der Weisheit, wenn sie in Aktion tritt, zugrunde liegen: Sie beinhalten (a) die Feststellung, dass ein Problem vorliegt, (b) die Präzisierung der Art des Problems, (c) das Abrufen von Informationen über das Problem, (d) das Entwerfen einer Lösungsstrategie, (e) das Ansammeln von Ressourcen für die Lösung des Problems, (f) die Überwachung der Problemlösung und (g) bewertendes Feedback in Bezug auf die Lösung. Ein Bestandteil der Weisheit ist somit die Verfügbarkeit von Informationen über das Problem. Hierunter fallen

beispielsweise Informationen über die Ursache des Problems und über die Folgen, die mögliche Handlungsweisen mit sich bringen würden. Sternberg (2000) merkt jedoch an, dass "acquisition of knowledge can be in the service of wisdom, but it is not in itself wise" (S. 639 f.). Weisheit ist eine Eigenschaft, die nicht situationsspezifisch ist. Eine Person, die sich in einem bestimmten Kontext weise verhalten kann, kann dies auch in anderen Situationen mit anderen Problemen, obwohl hierfür möglicherweise andere Informationen erforderlich sind. Das Wissen, das für weises Verhalten erforderlich ist, kann in unterschiedlichen Situationen variieren. Wie kann eine Person jedoch situationsübergreifend weise handeln, wenn die erforderliche Information variiert? In Situationen, in denen einer weisen Person das Wissen fehlt, welches für eine Reaktion mit einer ausbalancierten Berücksichtigung intra-, inter- und extrapersonaler Interessen erforderlich wäre, äußert sich Weisheit in der Fähigkeit, diese Grenze wahrzunehmen und entsprechend zurückhaltend über die Angemessenheit von verschiedenen Handlungsmöglichkeiten in der Situation zu urteilen. Eine weise Person "will therefore know not only when to give advice but when not to . . . because the individual will know the limitation of his or her own tacit knowledge" (Sternberg, 2000, S. 640). Hier findet sich somit ein gemeinsames Element mit der Auffassung von Weisheit nach Baltes und Smith (1990) wieder, wo in Form des Kriteriums der uncertainty den Kenntnissen der Grenzen des eigenen Wissens eine hohe Bedeutung beigemessen wird. Sternbergs Theorie zufolge ist explizites Wissen zwar für Weisheit nicht irrelevant, von wesentlich größerer Bedeutung ist jedoch das stille Wissen, über das eine Person verfügt. Deklaratives semantisches Wissen über den Kontext eines Problems ist hilfreich, aber weder erforderlich noch hinreichend für Weisheit (Sternberg, 2000).

Eine relativ allgemeine Beschreibung des Zusammenhangs zwischen Weisheit und Wissen, die für beide hier beschriebenen Ansätze der Weisheit gelten mag, gibt Sternberg (1990): "The wise person is someone who *probes inside knowledge* – who understands the

meaning of what is known. . . . They [Weise Personen] understand the significance of the knowledge they do and do not and can and cannot have as well as its potential meaning for life endeavors" (S. 152).

2.3.5 Wissen und Langzeitgedächtnis

Für den Wissensbestand einer Person ist das Langzeitgedächtnis bedeutsam. Ähnlich wie zwischen verschiedenen Wissensarten differenziert wird, lassen sich auch verschiedene Arten des Langzeitgedächtnisses unterscheiden. So wird in der vorliegenden Arbeit das deklarative semantische Gedächtnis betrachtet. Die hierin enthaltene Information bildet das deklarative semantische Wissen einer Person. Forschung in der Allgemeinen Psychologie zum Thema Gedächtnis konzentrierte sich großteils auf das semantische Gedächtnis und auf die Frage, in welcher Form Wissen im Gedächtnis repräsentiert ist. Wie auch in Kapitel 2.3.1 beschrieben, beschreibt Tulving (1972) das semantische Gedächtnis als das Wissen über Worte und andere verbale Symbole. Anderson (2000/2001) bezeichnet in seinem Lehrbuch zur kognitiven Psychologie die Enkodierung dieses Wissens als *bedeutungsbezogene Wissensrepräsentationen*. Klix (1988) weist darauf hin, dass bei Tieren die Bedeutung, die sie einem Objekt zuweisen, mehrheitlich angeboren ist. Bei höher entwickelten Lebewesen kann die Bedeutung hingegen korrigiert und vergessen werden, was als *Lernen* bezeichnet wird. Während Tiere für ein und dasselbe Objekt ausschließlich eine Klassifizierung vornehmen, können Menschen abhängig von ihren Lernerfahrungen zahlreiche Klassifizierungen vornehmen. Klix (1988) gibt hierfür folgendes Beispiel: Eine Blume ist für eine Biene immer eine Blume, während der Mensch eine Blume beispielsweise als Geschenk, als Zierpflanze, als Heilpflanze, als Symbol für Liebe und als Symbol für Totenehrung klassifizieren kann. So können für ein Objekt allgemeine Merkmale mit situationsspezifisch verschiedenen Merkmalen durch Wortmarken verknüpft werden, wodurch ermöglicht wird,

"jene Merkmale zu akzentuieren, auf deren situationsbezogene Bedeutung es neben der allgemeinen ankommt. Die Bindung begrifflicher Merkmalssätze durch Worte ermöglicht es also, die Vielfalt realer Objekteigenschaften bedarfsgerecht im Gedächtnis zu fixieren" (Klix, 1988, S. 21).

Ein häufig verwendeter Begriff für die so entstehenden Strukturen des Gedächtnisses sind sogenannte *semantische Netzwerke*. Ohne Anspruch auf Vollständigkeit werden hier einige zentrale Theorien zu semantischen Netzwerken erläutert, wie sie von Baddeley (1997) zusammengefasst werden. Frühe Theorien waren hierarchisch strukturiert und lassen sich in Klassische Theorien und Prototypen-Theorien einteilen. Die hierarchischen Strukturen beinhalten auf verschiedenen Ebenen sogenannte *Konzepte*, wie beispielsweise den Begriff "Vogel". Klassischen Theorien zufolge können Konzepten bestimmte Merkmale eindeutig zugeordnet werden (z.B. "kann fliegen"). Jene Merkmale werden als notwendig und hinreichend für die Zuordnung eines Objekts zu einem Konzept gesehen. Konzepte können wiederum durch weitere Konzepte näher spezifiziert werden (z.B. "Kanarienvogel", "Strauß"). Alle Merkmale, die für ein Konzept aus der höheren Ebene gelten, müssen auch bei sämtlichen näher spezifizierten Konzepten vorzufinden sein. Das genannte Beispiel, welches an Collins & Quillian (1969) angelehnt ist, verdeutlicht, dass diese Annahme nicht haltbar ist. Zum einen können Strauße nicht fliegen. Außerdem können nicht nur Vögel, sondern auch viele Insekten fliegen. Eine Alternative zu Klassischen Theorien bilden daher Prototypen-Theorien. Hier wird davon ausgegangen, dass verschiedene Beispiele eines Konzeptes in unterschiedlichem Ausmaß prototypisch für das Konzept sind. So wäre der Kanarienvogel beispielsweise ein prototypischeres Beispiel für das Konzept Vogel als der Strauß. Nichtsdestoweniger weisen die Merkmale verschiedener Beispiele für ein Konzept Überlappungen auf. So haben beispielsweise sowohl der Kanarienvogel als auch der Strauß einen Schnabel und ein Gefieder, was eine Zuordnung des Straußes zum Konzept Vogel

ermöglicht. Sämtlichen Theorien, welche die Strukturen des Gedächtnisses in Form von semantischen Netzwerken modellieren, ist gemeinsam, dass Begriffe durch andere Begriffe umschrieben werden, womit die Möglichkeiten des Transfers der Modelle auf das reale Leben beschränkt sind. Baddeley (1997) gibt hierzu folgendes Beispiel: Aus dem Wissen, dass A sich rechts von B und B sich rechts von C befinden, könnte geschlussfolgert werden, dass C sich ebenfalls rechts von A befindet. Sofern es sich bei A, B und C um Personen handelt, die alle an einer Seite eines Tisches sitzen, ist die Schlussfolgerung korrekt. Sofern sie aber um einen runden Tisch sitzen, ist sie nicht korrekt. Johnson-Laird, Herrmann und Chaffin (1984) bezeichneten dieses Problem als *symbolic fallacy* – "the assumption that the mere translation of sentences into symbols constitutes a useful account of their meaning" (S. 310).

Eine weitere Gruppe von Theorien zur Enkodierung des Wissens stützt sich auf sogenannte *Schemata*. Baddeley (1997) nennt einige Merkmale von Schemata, die unterschiedliche Theorien, welche auf dem Konzept beruhen, gemeinsam haben. Beispielhaft seien hier folgende genannt: (a) Schemata haben Variablen, die teilweise fix und teilweise veränderlich sind. Als Beispiel wird das Schema des Restaurantbesuchs genannt. Die Ausgabe von Geld ist hier eine fixe Variable, während der zu zahlende Betrag variabel ist. (b) Schemata können ineinander geschachtelt sein. So ist das Schema Auge Bestandteil des Schemas Gesicht, welches wiederum Bestandteil des Schemas Kopf ist. (c) Schemata können auf allen Abstraktionsebenen existieren. So kann es ein Schema zu direkt wahrnehmbaren Objekten, wie z.B. dem Gesicht, geben, aber auch zu abstrakteren Begriffen, wie beispielsweise Steuererklärung. Neben Begriffsschemata gibt es auch Ereignisschemata, die einem bestimmten Geschehen Variablen zuschreiben, und die auch als *scripts* bezeichnet werden. Sie beinhalten die Abfolge von Teilereignissen des Geschehens und sind bei fehlenden oder falschen Informationen hilfreich (Anderson, 2000/2001).

Je nach der kognitionspsychologischen Theorie, deren Perspektive eingenommen wird, lassen sich interindividuelle Differenzen im Wissen somit unterschiedlich darstellen. Aus Perspektive der Theorien semantischer Netzwerke würden Leistungsunterschiede in Wissenstests durch Unterschiede in der Anzahl der Verknüpfungen im semantischen Netzwerk abgebildet. Durch jedes Item eines Wissenstests würden demnach ein oder mehrere semantische Netzwerke angeregt, welche bei den Individuen unterschiedlich ausgeprägt sind, und deren Verknüpfungen sich interindividuell unterscheiden. Aus Perspektive der Schema-Theorien können Personen, die ein höheres Wissen haben, den Schemata, die Inhalt von Items sind, eine größere Anzahl an Variablen zuweisen. Welchem der beschriebenen Modelle der bedeutungsbezogenen Wissensrepräsentation der Vorzug zu geben ist, ist nicht Gegenstand dieser Arbeit. In der kognitiven Psychologie, die das Prinzip der Organisation von Wissen zum Thema hat, wird das Gedächtnis in der Regel auf relativ abstrakter Ebene untersucht. Interindividuelle Unterschiede im Wissen von Menschen stellen hier keinen Forschungsgegenstand dar. Vielmehr liegt das Ziel hier in der Modellierung der Organisation von Informationen. Inhalte, die im Rahmen der Experimente abgefragt werden, wie beispielsweise das Schema eines Restaurantbesuchs, weisen sehr geringe interindividuelle Differenzen auf. Klix (1988) weist darauf hin, dass bei dieser Forschung der Umstand, "daß die natürliche Wissensbasis eines Menschen eine innere Architektur hat, daß die Begriffe des bestehenden Vorwissens miteinander vernetzt sind, daß es verschiedene Beziehungen zwischen ihnen gibt, die Erkennung begünstigen, behindern oder gar verhindern" (S. 22), in der Regel nicht thematisiert wird. Es ist davon auszugehen, dass die "innere Architektur" des Wissens bei Tests des Allgemeinen Wissens zum Tragen kommt. Das bei jenen Tests abgefragte Wissen dürfte deutliche idiosynkratische Zusammenhänge aufweisen, sei es in Form von semantischen Netzwerken oder Schemata. Nach Kenntnis des Autors wurden Wissenstests bisher nicht zur Forschung in der Gedächtnispsychologie eingesetzt. Sie dienen

zur Diagnostik des Wissens und eng damit verknüpfter psychologischer Merkmale (beispielsweise Kristalliner Intelligenz) auf individueller Ebene oder auf der Ebene von Gruppen. Unabhängig davon, welche der vorgestellten Theorien der Realität entsprechen mag, weist das so erfasste Wissen starke interindividuelle Differenzen auf, denen zahlreiche Ursachen zugrunde liegen können.

Ein anderer Teilbereich der Gedächtnispsychologie konzentriert sich auf die *Prozesse*, die für Gedächtnisleistungen relevant sind (Buchner & Brandt, 2002). Forschungsgegenstand sind hier Faktoren, die beim Behalten, bei der Speicherung und beim Abruf von Informationen relevant sind. Beispiele für Themen sind die Effekte der Verarbeitungstiefe (z.B. Craik & Lockhart, 1972), der Kontextspezifität (z.B. Godden und Baddeley, 1975) und verschiedener Formen von Interferenzen (z.B. Loftus & Loftus, 1980) auf die Gedächtnisleistungen. Dieser Forschungsbereich ist für die Untersuchung der vorliegenden Arbeit von wesentlich größerer Bedeutung als die im vorigen Abschnitt beschriebenen Theorien zur Organisation des Wissens im Langzeitgedächtnis. Es wird hier insbesondere auf die Theorie der Verarbeitungstiefe näher eingegangen, da sie in engem Zusammenhang mit der Hypothesenbildung dieser Arbeit steht. Nach Craik und Lockhart (1972) ist die Tiefe der Verarbeitung von Information von zentraler Bedeutung für den Erfolg beim späteren Abruf. Eine tiefe Verarbeitung beinhaltet hier die Verarbeitung der Bedeutung von Information, während bei der oberflächlichen Verarbeitung lediglich perzeptuelle Aspekte berücksichtigt werden. Um die Unterscheidung zwischen tiefer und oberflächlicher Verarbeitung in Experimenten zu operationalisieren, ließen Hyde und Jenkins (1973) die Probandinnen und Probanden beim Lernprozess angeben, ob sich bestimmte Buchstaben in den gelernten Wörtern fanden (oberflächliche Verarbeitung), oder ob es sich bei dem Wort um ein Verb oder ein Substantiv handelte (semantische Verarbeitung). Beim Abruf der gelernten Worte zeigte sich, dass die Begriffe, die mit oberflächlicher Verarbeitung gelernt worden waren,

schlechter erinnert wurden als die Worte, die beim Lernen semantisch verarbeitet worden waren. Diese Befunde wurden häufig repliziert (siehe Buchner & Brandt, 2002). Morris, Bransford und Franks (1977) wiesen auf eine Einschränkung der Theorie der Verarbeitungstiefe hin. Sie machten in ihrer Untersuchung deutlich, dass die Kongruenz zwischen der Verarbeitung beim Lernen und beim Abruf ebenfalls ein wichtiger Prädiktor der Erinnerungsleistung ist. Sofern beim Lernen hauptsächlich perzeptuelle Merkmale des Materials beachtet wurden, ist der Abruf erfolgreicher, wenn er ebenfalls auf perzeptueller Ebene stattfindet. Die Operationalisierung des Abrufs auf perzeptueller Ebene bestand in der Frage, ob sich ein Wort aus der Lernphase des Experiments auf ein neu präsentiertes Wort reimte. Eine semantische Verarbeitung war somit hier nicht notwendig. Sofern das Material jedoch unter Beachtung der Bedeutung gelernt worden war, fiel die Erinnerungsleistung höher aus, wenn auch beim Abruf semantische Verarbeitung erforderlich war. Morris et al. zufolge ist somit der "Grad der Überlappung von kognitiven Prozessen bei der Enkodierung und dem Abruf von Information" (Buchner & Brandt, 2002, S. 511) von großer Relevanz für die Erinnerungsleistung. 1990 veröffentlichten Lockhart und Craik einen Artikel, in dem sie ihren Ansatz mit dem von Morris et al. kombinierten. Sie verwiesen hier darauf, dass beim Vergleich der Erinnerungsleistung unter kongruenten Bedingungen die semantische Verarbeitung der oberflächlichen Verarbeitung überlegen war.

Die prozessorientierte Theorie der Verarbeitungstiefe steht im Einklang mit der in Kapitel 2.3.3 vorgestellten Investmenttheorie. Wie bereits erläutert, nahm Cattell (1971/1987) an, dass stärkere Interessen an bestimmten Themen zur ausführlicheren Beschäftigung – zu einem stärkeren "Investment" in die entsprechenden Themen – führen. Es liegt nahe, dass stärkere Interessen auch zu einer tieferen Verarbeitung des Materials führen und damit bessere Erinnerungsleistungen zur Folge haben. Gerade bei den hier untersuchten Tests des Allgemeinen Wissens sollte eine Person von tieferer Verarbeitung profitieren, da hierbei

ebenfalls verschiedene Antwortoptionen inhaltlich beurteilt und miteinander verglichen werden müssen. Die Beschäftigung mit dem Material auf semantischer Ebene – die höhere Investition in das Thema – führt somit zu einer höheren Transferangemessenheit und damit zu einer besseren Erinnerungsleistung in der Testsituation. Des Weiteren liegt es nahe, dass stärkeres Investment auch zu einem höheren Ausmaß an Elaborationen, also zu einer stärkeren Anreicherung des zu behaltenden Materials um zusätzliche Information, führen dürfte. Dies mag man sich anhand eines semantischen Netzwerks vorstellen, das zusätzliche Verknüpfungen aufweist. Wie zahlreiche Experimente belegen, werden Inhalte, zu denen Elaborationen gebildet wurden, besser erinnert (siehe Anderson, 2000/2001).

Nachdem in Kapitel 2 der Begriff des deklarativen semantischen Wissens herausgearbeitet und in Relation zur Psychologischen Diagnostik, zur Allgemeinen und zur Differenziellen Psychologie gesetzt wurde, liegt der Fokus in Kapitel 3 auf der Differenziellen Psychologie: Frauen und Männer weisen in zahlreichen Untersuchungen Leistungsunterschiede in Tests des Allgemeinen Wissens auf, die im Folgenden näher betrachtet werden.

Geschlechterdifferenzen des Allgemeinen Wissens

Leistungsunterschiede zwischen Frauen und Männern in Tests des Allgemeinen Wissens sind der zentrale Gegenstand der vorliegenden Arbeit. Im Folgenden werden zunächst zahlreiche empirische Befunde für dieses Phänomen aufgezeigt. Im Anschluss werden mögliche Erklärungen und Forschungsergebnisse vorgestellt, welche die Erklärungen stützen oder in Frage stellen. Zuletzt erfolgt eine kurze Umschreibung der Erklärungsmöglichkeiten, die in der vorliegenden Arbeit überprüft wurden. Die ausführlichen Erläuterungen finden sich in den jeweiligen Kapiteln.

3.1 Belege für Geschlechterdifferenzen des Allgemeinen Wissens

Für die Leistungen von Frauen und Männern in Tests des Allgemeinen Wissens wurde in zahlreichen Untersuchungen und Normierungsstichproben über Mittelwertunterschiede, oder über Unterschiede auf latenter Ebene berichtet. Die Effektstärke der Unterschiede variierte dabei, die Richtung des Effekts wies jedoch fast durchgehend auf eine höhere Leistung von Männern hin. Im Folgenden wird ein Überblick über Untersuchungen gegeben, in denen Geschlechterdifferenzen in Tests des Allgemeinen Wissens geprüft wurden. Soweit möglich, wird auch das hierbei verwendete Testmaterial detailliert beschrieben. Als Effektstärkemaß wurde in der Regel der Koeffizient *d* nach Cohen (1988) verwendet. Einige Autoren subtrahierten bei dessen Berechnung den Mittelwert der Frauen von dem Mittelwert der Männer, während andere Autoren umgekehrt vorgingen. Um in dieser Arbeit einheitlich über die Effektstärken zu berichten, wurde, soweit notwendig, die von den Autoren berichtete Effektstärke mit -1 multipliziert. Effektstärken mit positivem Vorzeichen weisen auf eine höhere Leistung der Frauen, Effektstärken mit negativem Vorzeichen auf eine höhere Leistung der Männer hin.

3.1.1 General Knowledge Test (GKT)

Irwing, Cammock und Lynn (2001) entwickelten den General Knowledge Test (GKT), welcher insgesamt 216 Items im freien Antwortformat enthält. General knowledge wurde hier aufgefasst als "culturally valued knowledge, communicated by a range of non-specialist media. Ephemera or knowledge confined solely to one medium, such as 'television soaps', were explicitly excluded by this conceptualization, as was information so specialist [*sic*] as to require extensive training for it to be acquired" (S. 859). Irwing et al. bestimmten 18 Themen, die diesem Kriterium genügten. Hierbei handelte es sich um History of Science, Politics, Sport, History, Classical Music, Art, Literature, General Science, Geography, Cookery, Medicine, Games, Discovery and Exploration, Biology, Film, Fashion, Finance und Popular Music. Anhand einer Stichprobe von 718 Personen (70.9% weiblich) prüften die Autoren die Struktur des GKT. Im ersten Schritt wurde für die Items jedes einzelnen Themas jeweils eine Hauptkomponentenanalyse durchgeführt. Nach entsprechender Itemselektion konnten die Items jedes Themas jeweils einer einzelnen Hauptkomponente zugeordnet werden. Eine Ausnahme bildete das Thema Popular Music. Da die Hälfte der Items hier auf der ersten Hauptkomponente positive und die andere Hälfte negative Ladungen aufwies, entschieden sich die Autoren für die Differenzierung zwischen zwei Themen, die sie mit Popular Music und Jazz and Blues bezeichneten. Durch mehrere darauf folgende exploratorische und konfirmatorische Faktorenanalysen mit 19 manifesten Variablen erstellten Irwing et al. in mehreren Schritten ein 6-faktorielles Modell, dessen Faktoren für interpretierbar befunden wurden. Die sechs Faktoren wurden als *Current Affairs*, *Fashion, Family, Physical Health and Recreation*, *Arts* und *Science* bezeichnet und wiesen Korrelationen zwischen .37 (Fashion und Science) und .82 (Current Affairs und Arts) auf. Eine weitere Faktorenanalyse zweiter Ordnung ergab einen einzelnen Faktor, der als *General Knowledge* bezeichnet wurde.

Eine konfirmatorische Faktorenanalyse des Modells zeigte eine zufriedenstellende Passung (SRMR = 0.05, RMSEA = 0.07). In Anhang A sind die 19 Themen des GKT nach Irwing et al. aufgelistet, geordnet nach den sechs Faktoren, denen sie zugeordnet wurden. Sämtliche manifesten Variablen wiesen ausschließlich auf dem ihnen jeweils zugeordneten Faktor Ladungen ungleich 0 auf. Sämtliche anderen Ladungen waren auf 0 fixiert. Mit Ausnahme des Themas Film konnten alle manifesten Variablen eindeutig einem Faktor zugeordnet werden. Die Variable Film wurde den zwei Faktoren Fashion und Arts zugeordnet. Dieser Test bildete die Grundlage zahlreicher Untersuchungen von Geschlechterdifferenzen im Wissen:

Lynn, Irwing und Cammock (2002) setzten den GKT in einer Stichprobe von 636 Studierenden (73.7% weiblich) aus Nordirland in einer gekürzten Version ein, welche lediglich 182 Items beinhaltete. Über den Grund der Kürzung und für die Verteilung der 182 Items auf die 19 Faktoren geben Lynn et al. keine Information. Die oben beschriebene faktorielle Struktur wurde jedoch auch hierfür überprüft und bestätigt. Die Effektstärke der Summenscores sämtlicher Items von Frauen und Männern betrug d = -0.51. Die Effektstärken, welche für die Summen der Items der sechs Faktoren berechnet wurden, variierten deutlich. Sie betrugen -0.82 (Current Affairs), 0.01 (Fashion), 0.46 (Family), -0.75 (Physical Health and Recreation), -0.31 (Arts) und -0.58 (Science).

Der GKT wurde außerdem ins Deutsche übersetzt (GKT-G) und in einer weiteren Untersuchung von Lynn, Wilberg und Margraf-Stiksrud (2004) von 302 Schülerinnen und Schülern (50.7% weiblich) im Alter von durchschnittlich 17.6 Jahren bearbeitet. Vorab wurde eine Pilotstudie durchgeführt, deren Ergebnis eine auf 95 Items gekürzte Version war. Die Itemauswahl wurde hierbei auf Basis von Schwierigkeit und Trennschärfe der Items durchgeführt. Die 95 Items des GKT-G konnten 17 Themen zugeordnet werden. Die Themen History of Science und General Science wurden zu dem Thema Physics and Chemistry

zusammengefasst. Außerdem wurden die Themen Popular Music und Jazz and Blues zu dem Thema Popular Music zusammengefasst. Mit der deutschsprachigen Version des GKT wurden keine Hauptkomponenten- bzw. Faktorenanalysen durchgeführt. Die Effektstärke des gesamten Tests betrug in der Arbeit von Lynn et al. (2004) d = -0.60.

Tabelle 1

Effektstärken (Cohens d) der Geschlechterdifferenzen der 17 im GKT-G erfassten Themen in den drei Untersuchungen von Lynn et al. (2002), Lynn et al. (2004) und Lynn et al. (2005)

Thema	Lynn et al. (2002)	Lynn et al. (2004)	Lynn et al. (2005)
Politics	-0.69	-0.33	-0.45
History	-0.72	-0.62	-0.58
Geography	-0.41	-0.67	-0.58
Discovery and Exploration	-0.69	-0.56	-0.44
Finance	-0.69	-0.85	-0.47
Cookery	0.48	0.50	0.26
Medicine	0.32	0.20	-0.10
Sport	-0.84	-1.12	-0.57
Games	-0.54	-0.82	-0.32
Biology	-0.42	-0.49	-0.09
Popular Music	0.15	-0.04	-0.18
Fashion	0.05	-0.03	0.16
Film	-0.13	-0.06	0.01
Classical Music	-0.08	0.03	-0.25
Art	-0.07	0.16	-0.15
Literature	-0.49	0.09	-0.28
Physics/Chemistry	-0.63	-0.55	-0.44

In einer weiteren Untersuchung von Lynn, Wilberg-Neidhardt und Margraf-Stiksrud (2005) wurde der GKT-G an der Universität in Lüneburg eingesetzt, um die Leistungsunterschiede von Frauen und Männern zu überprüfen. Die Stichprobe umfasste 233

Studierende (62.2% weiblich). Hier wurde für den gesamten Test eine Effektstärke von d = -0.49 berechnet.

Lynn et al. (2004) und Lynn et al. (2005) berichteten auch über die Effektstärken der 17 einzelnen Themen, denen die Items des GKT-G zugeordnet wurden. In beiden Artikeln wurde auch über die Effektstärken der jeweiligen Themen berichtet, die sich unter Berücksichtigung der im GKT-G verwendeten Items aus den Daten der irischen Stichprobe von Lynn et al. (2002) berechnen ließen. Die Effektstärken aller drei Untersuchungen werden hier in Tabelle 1 zusammengefasst. Die Mehrzahl der Themen zeigte in allen drei Untersuchungen Effektstärken zugunsten von Männern. Das Thema Cookery war das einzige, für das in allen drei Untersuchungen durchschnittlich höhere Leistungen von Frauen festgestellt wurden. Bei den Themen Medicine, Popular Music, Fashion, Film, Classical Music, Art und Literature variierte die Gruppe, die höhere Leistungen zeigte.

Auf latenter Ebene prüften Lynn und Irwing (2002) den Effekt des Geschlechts auf die Leistungen im GKT, wobei hier eine auf 72 Items gekürzte Variante des GKT verwendet wurde. 1047 Personen (56.7% weiblich) nahmen an der Untersuchung teil. In einer konfirmatorischen Faktorenanalyse wurde der GKT durch sechs Faktoren erster Ordnung und einen Faktor zweiter Ordnung abgebildet. Die sechs Faktoren erster Ordnung entsprachen bei dieser gekürzten Version des GKT jedoch nicht den sechs Faktoren aus der Untersuchung von Irwing et al. (2001). Die gekürzte Version des GKT, die Lynn und Irwing (2002) verwendeten, "was devised by choosing one domain of general knowledge to represent each of the first-order factors" (S. 549). Die hier verwendeten Themen waren Finance (zur Repräsentation des Faktors Current Affairs), Fashion (zur Repräsentation des Faktors Fashion), Medicine (Family), Games (Physical Health and Recreation), Literature (Arts) und General Science (Science). Der Faktor zweiter Ordnung wurde auch hier mit General Knowledge betitelt. Eine weitere Variable, die in das Modell Eingang fand, war die Fluide

Intelligenz, die durch den 3-minütigen Grammatical Reasoning Tests von Baddeley (1968) operationalisiert wurde. Ein MIMIC-Modell, bei dem für das Geschlecht und die Fluide Intelligenz ausschließlich Ladungen auf dem Faktor zweiter Ordnung zugelassen wurden, zeigte keine zufriedenstellende Passung. Bei dem revidierten Modell, das eine gute Passung aufwies (SRMR=.03, RMSEA=.06), zeigte die Variable Geschlecht eine negative Ladung auf dem Faktor zweiter Ordnung, was als geringere Leistung der Frauen zu interpretieren ist. Zusätzlich wurden hier jedoch auch direkte Ladungen des Geschlechts auf drei der sechs Faktoren erster Ordnung zugelassen. Hierbei handelte es sich um die Faktoren Literature, Fashion und Medicine. Diese Ladungen waren positiv, woraus zu schlussfolgern ist, dass "females do better than predicted by the overall trend on these aspects of general knowledge" (S. 552). Auf den anderen drei Faktoren erster Ordnung, General Science, Games und Finance, wurden keine direkten Ladungen des Geschlechts zugelassen. Die Schlussfolgerung von Lynn und Irwing (2002), dass "the effect of sex on general knowledge ability confirms the results obtained by Lynn *et al.* (2002), that males have a large advantage on general knowledge" (S. 552) bedarf somit der Einschränkung, dass Teilgebiete des Wissens offensichtlich auch durch einen oder mehrere andere Faktoren als das im GKT erfasste Allgemeine Wissen beeinflusst werden, die bei Frauen stärker ausgeprägt sind. Lynn und Irwing (2002) weisen dementsprechend in der Diskussion darauf hin, dass es möglicherweise eine "different sort of general knowledge from that represented in our test" (S. 554) gibt. Es sei in diesem Zusammenhang an die in Kapitel 2.3.2 berichteten Ergebnisse einer exploratorischen Faktorenanalyse erinnert, die Hossiep und Schulte (2008) im Rahmen der Konstruktion des BOWIT durchführten. Hierbei zeigte sich kein einzelner Faktor, der als Allgemeines Wissen bezeichnet werden könnte. Vielmehr ergab sich eine 2-faktorielle Struktur. Die Faktoren wurden von Hossiep und Schulte als Gesellschafts- und Geisteswissenschaftliches Wissen und Naturwissenschaftlich-technisches Wissen bezeichnet

(siehe Seite 28). Eine konfirmatorische Faktorenanalyse mit einem 2-faktoriellen Modell wurde von Lynn und Irwing (2002) für den GKT nicht durchgeführt. Die Frage, ob möglicherweise also auch ein anderes Modell, beispielsweise ein 2-faktorielles Modell wie beim BOWIT, die Daten ebenso gut erklären könnte, bleibt hier somit offen.

Zusammenfassend kann für den GKT festgehalten werden, dass Männer insgesamt eine deutlich höhere Leistung im Gesamtscore zeigten. Auf latenter Ebene konnte jedoch ein Modell, in dem sämtliche Faktoren erster Ordnung auf einen einzelnen Faktor zweiter Ordnung zurückgeführt wurden, nicht zufriedenstellend bestätigt werden.

3.1.2 Testbatterie von Rolfhus und Ackerman

Ein weiterer Wissenstest, der zur Überprüfung von Leistungsunterschieden zwischen Frauen und Männern eingesetzt wurde, wurde von Rolfhus und Ackerman (1999) entwickelt. Im Rahmen seiner Dissertation entwickelte Rolfhus 20 themenspezifische Wissenstests. Der Schwerpunkt lag hierbei auf wissenschaftlichen Themen, "because one potential application was prediction and classification in academic situations" (Rolfhus & Ackerman, 1999, S. 513). Die 20 Wissenstests umfassten die Themen American Government, American History, American Literature, Art, Astronomy, Biology, Business/Management, Chemistry, Economics, Electronics, Geography, Law, Music, Physics, Psychology, Statistics, Technology, Tools/Shop, Western Civilization und World Literature. Nach einem umfangreichen Entwicklungsprozess lagen 20 Wissenstests vor, die jeweils aus 35 bis 100 Multiple-Choice-Items bestanden, welche der Schwierigkeit nach geordnet waren. Die Bearbeitung der einzelnen Tests startete mit dem einfachsten Item und wurde nach drei falschen Antworten abgebrochen. Die 20 Tests wurden einer Stichprobe von 143 Personen (65.7% weiblich) vorgelegt. Sämtliche Tests wiesen leichte bis moderate positive Korrelationen miteinander auf. Rolfhus und Ackerman (1999) zogen den Schluss, dass dieses

Ergebnis "suggests a general factor underlying the knowledge scales" (S. 515). Eine exploratorische Faktorenanalyse mit obliquer Rotation ergab vier Faktoren, die die Autoren als *Humanities*, *Science*, *Civics* und *Mechanical* bezeichneten, und mit denen eine weitere Faktorenanalyse durchgeführt wurde. Hier zeigte sich eine 1-faktorielle Lösung. Dieser Faktor wurde als *General Knowledge* bezeichnet. Detailliertere Beschreibungen der Testentwicklung, der Inhalte der einzelnen Themengebiete und der Vorgehensweise bei den Faktorenanalysen finden sich bei Rolfhus und Ackerman.

Das im vorigen Absatz beschriebene Testmaterial wurde in einer leicht abgewandelten Version von Ackerman, Bowen, Beier und Kanfer (2001) in einer Untersuchung verwendet, bei der unter anderem die Geschlechterdifferenzen im Allgemeinen Wissen überprüft wurden. Die Abwandlung bestand in der Elimination der Tests zu den Themen Statistics und Tools/Shop. Stattdessen wurde ein neuer Test zu dem Thema Current Events (1990s) hinzugefügt. Die Themen American Government, American History und American Literature finden sich in dieser Testbatterie ebenfalls nicht wieder, jedoch werden U.S. Government, U.S. History und U.S. Literature als Themen genannt. Daher ist anzunehmen, dass hier die gleichen Tests wie bei Rolfhus und Ackerman (1999) verwendet wurden. Insgesamt 320 Personen (52.5% Frauen) bearbeiteten den Test. Eine exploratorische Faktorenanalyse mit obliquer Rotation ergab auch hier eine 4-Faktoren-Lösung. Die Faktoren wurden mit den Begriffen *Physical Science/Technology*, *Biology/Psychology*, *Humanities* und *Civics* bezeichnet. Tabelle 2 zeigt die Zuordnung der Skalen zu den einzelnen Faktoren und die Effektstärken, die für die Skalen berechnet gefunden. Detaillierte Informationen zu der Vorgehensweise bei der Zuordnung der Skalen zu den Faktoren finden sich bei Ackerman et al. (2001). Da für die Mehrzahl der Themen die Effektstärken nicht in numerischer Form berichtet werden, sondern aus einem Balkendiagram abzulesen sind, werden sie hier in kategorisierter Form dargestellt.

Tabelle 2

Effektstärken (Cohens d) der Geschlechterdifferenzen der 19 von Ackerman et al. (2001) untersuchten Wissensgebiete

Faktor	Kategorien von Cohens *d*					
Thema	< -0.8	-0.8 – -0.6	-0.6 – -0.4	-0.4 – -0.2	-0.2 – 0	> 0
Phys. Sc./Tech.						
Technology	x					
Electronics	x					
Physics		x				
Astronomy		x				
Chemistry				x		
Humanities						
Geography		x				
West. Civ.			x			
Cur. Ev. (1990s)				x		
Music				x		
World Literature					x	
Art					x	
U.S. Literature					x	
Civics						
U.S. History			x			
Law				x		
Economics				x		
U.S. Government				x		
Business					x	
Bio./Psy.						
Biology				x		
Psychology						x

Anmerkungen. Phys. Sc./Tech. = Physical Science/Technology; West. Civ. = Western Civilization; Cur. Ev. (1990s) = Current Events (1990s); Bio./Psy. = Biology/Psychology.

Wie in Tabelle 2 zu erkennen ist, wiesen die Themen der vier Faktoren Effektstärken unterschiedlicher Höhe auf. Die Themen des Faktors Physical Science/Technology zeigten die höchsten Effektstärken. Dem Faktor Humanities wurden Themen zugeordnet, deren Effektstärken stark streuten. Die Themen des Faktors Civics zeigten eher geringe Effektstärken und der Faktor Biology/Psychology enthielt ein Thema mit geringer Effektstärke (Biology) und als einziger Faktor ein Thema, das eine (minimale) Effektstärke zugunsten der Frauen aufwies (Psychology). Die Faktoren korrelierten positiv miteinander, mit Ausnahme des Faktors Humanities, der mit allen anderen Faktoren negativ korrelierte. Die ausschließlich positiven Korrelationen der 19 einzelnen Skalen sahen Ackerman et al. (2001) jedoch als eindeutigen Hinweis auf einen einzelnen Faktor höherer Ordnung, der als general knowledge bezeichnet werden könnte, weshalb "a global knowledge composite, Gk, was created with a unit-weighted *z*-score sum of all 19 knowledge tests" (S. 805). Für die Leistungsunterschiede von Männern und Frauen über sämtliche Tests hinweg wurde eine Effektstärke von d = -0.68 berechnet. Im Vergleich zum GKT von Irwing et al. (2001) fiel sie somit bei Ackerman et al. (2001) etwas höher aus.

3.1.3 Normierungsstichproben des I-S-T 2000 R und des BOWIT

Der Wissenstest des I-S-T 2000 R von Liepmann et al. (2007) und der BOWIT von Hossiep und Schulte (2008) wurden auch im Rahmen der Untersuchungen für die vorliegende Arbeit verwendet. In Kapitel 2.3.2 wurden bereits die Hauptgütekriterien der beiden Tests beschrieben. In den Testmanualen finden sich außerdem Informationen über die jeweiligen Leistungsunterschiede zwischen Frauen und Männern.

Die Normierungsstichprobe des I-S-T, auf deren Basis die Geschlechterdifferenzen berechnet wurden, umfasste 661 Personen (55.4% weiblich). Liepmann et al. (2007) konstatieren für den Gesamtwert des Wissenstests, dass "der Durchschnitt der männlichen

Probanden etwa drei Standardwertpunkte über dem Durchschnitt der weiblichen Probanden" (S. 69) lag. Da die Standardabweichung in der Skala der Standardwerte σ = 10 beträgt, kann hieraus geschlossen werden, dass Cohens *d* hier bei etwa -0.30 lag, Ein *t*-Test lieferte mit "$t_{561.39}$ = -4.20, p < .001" (S. 69) ein hochsignifikantes Ergebnis. Liepmann et al. sahen von der Erstellung geschlechtsspezifischer Normen ab, da ihnen "die vorliegende Stichprobe für ein derartiges Vorhaben noch nicht umfassend genug erschien, das einerseits besonders komplex ist . . . und andererseits weitreichende gesellschaftspolitische Implikationen haben kann" (S. 69).

Für die Überprüfung der Geschlechterdifferenzen beim BOWIT stand Hossiep und Schulte (2008) eine Normierungsstichprobe von 2425 Personen (40.7% weiblich) zur Verfügung. Bei Männern ergab sich für die Rohwerte ein Mittelwert von 106.71 Punkten (*SD* = 20.00). Bei Frauen betrug der Mittelwert 84.43 Punkte (*SD* = 23.70). Hieraus lässt sich eine Effektstärke von *d* = -1.04 berechnen. Dieser Unterschied wurde auch über einen *t*-Test "(t = 18,04, p < 0,001)" (S. 52) inferenzstatistisch abgesichert. Für den BOWIT wurden geschlechtsspezifische Normen erstellt.

3.1.4 Sonstige Belege

Im Folgenden werden weitere Untersuchungen beschrieben, die Beispiele für die einseitigen Geschlechterdifferenzen in Tests des Allgemeinen Wissens darstellen. Die Untersuchungen wurden im Rahmen unterschiedlicher Forschungsfragen durchgeführt und stammen aus verschiedenen Teildisziplinen der Psychologie.

Steinmayr und Amelang (2006) prüften die Kriteriumsvalidität der ersten Auflage des I-S-T 2000 R von 2001. Hierbei lagen ihnen die Daten von 219 Personen (51.6% weiblich) vor. Für den Wissenstest zeigte sich hierbei eine Effektstärke von *d* = -0.78.

Feingold (1993) publizierte einen Artikel, in dem er einen Überblick über Geschlechterdifferenzen kognitiver Fähigkeiten in verschiedenen Generationen gab. Grundlage hierfür bildeten die "norms from past and present standardization of . . . the Wechsler scales to assess temporal trends in the magnitude of gender differences for each of three broad age groups (children, adolescents, adults)" (S. 100). Die früheren Normen, auf die Feingold zurückgriff, sind der Wechsler Intelligence Scale for Children (WISC) von 1949 und der Wechsler Adult Intelligence Scale (WAIS) von 1955 zuzuordnen. Die späteren Normen sind jeweils den revidierten Versionen, dem WISC-R von 1974 und dem WAIS-R von 1981, zuzuordnen. Das Allgemeine Wissen wurde hier durch den Subtest Verbal Information erfasst. In Tabelle 3 ist der Ausschnitt der von Feingold berichteten Ergebnisse wiedergegeben, der sich auf diesen Subtest bezieht.

Tabelle 3

Effektstärken (Cohens d) der Geschlechterdifferenzen bei Normierungsstichproben von vier Wechsler Intelligenztests für die Subskala Verbal Information

	WISC (1949)	WISC-R (1974)
Kinder	-.07	-.23
Jugendliche	-.38	-.51
	WAIS (1955)	WAIS-R (1981)
Erwachsene	-.20	-.28

In Tabelle 3 sind für alle drei Altersgruppen Geschlechterdifferenzen zu erkennen, die in ihrer Richtung übereinstimmen, in ihrer Höhe jedoch variieren. Am geringsten fielen die Unterschiede in der Gruppe der Kinder aus, deren Daten die Normen des WISC von 1949 bildeten. Der größte Unterschied wurde bei den Jugendlichen in den Normen des WISC-R

von 1974 gefunden. Insgesamt ist bei allen drei Altersgruppen ein Anstieg der Leistungsunterschiede in den beiden jeweiligen Normierungsstichproben festzustellen.

Oswald, Rupprecht und Hagen (1997) untersuchten in einer Studie mit 697 Teilnehmenden im Alter von 62 bis 64 Jahren (48.2% weiblich) Zusammenhänge zwischen verschiedenen kognitiven Fähigkeiten und soziodemographischen sowie Persönlichkeitsmerkmalen. Die Untersuchung wurde im Rahmen des Projekts ILSE (Interdisziplinäre Längsschnittstudie des Erwachsenenalters über Bedingungen des gesunden und zufriedenen Älterwerdens) durchgeführt. Wie aus dem Abschlussbericht des Projekts von Schmitt, Wahl und Kruse (2008) hervorgeht, wurde zur Erfassung des Allgemeinen Wissens der entsprechende Subtest des revidierten Hamburg-Wechsler Intelligenztests für Erwachsene (HAWIE) eingesetzt. Die Leistungsunterschiede im Allgemeinen Wissen wiesen nach den Ergebnissen von Oswald et al. eine Effektstärke von d = -0.79 auf und waren auf dem 1‰-Niveau signifikant.

Furnham, Christopher, Garwood und Martin (2007) setzten die in Kapitel 3.1.1 beschriebene Kurzversion des GKT von Irwing et al. (2001) ein, um Zusammenhänge zwischen Allgemeinem Wissen, dem Lernverhalten und Persönlichkeitsmerkmalen zu untersuchen. Die Stichprobe umfasste hierbei 430 Studierende (72.8% weiblich). Furnham et al. berichteten nicht über Mittelwerte oder Cohens d, jedoch über eine Korrelation von .24 ($p < .01$) zwischen dem Geschlecht und Allgemeinem Wissen. Das positive Vorzeichen weist hierbei auf eine höhere Leistung der Männer hin.

Evans, Schweingruber und Stevenson (2002) verglichen die Interessen und das Wissen von Frauen und Männern in den Vereinigten Staaten, in Taiwan und in Japan. Für die Erfassung des Wissens wurde der General Information Test konstruiert. Dieser bestand aus 12 Items im offenen Format und "was constructed by members of a research team that included bilingual male and female native speakers of Chinese, Japanese and English" (Evans

et al., 2002, S. 155). Bei der Testentwicklung wurde besonderer Wert auf die kulturelle Fairness gelegt. So wurde der Test simultan in allen drei Sprachen erstellt und enthielt ausschließlich Items mit Inhalten, die in allen drei Kulturen leicht verfügbar waren. Daher wurde kein Wissen zu kulturspezifischen Themen, wie beispielsweise Kunst, Musik oder Literatur abgefragt. Es wurden Themen aufgenommen, von denen man annahm, dass sie weltweit verfügbar sind und dass Mädchen und Jungen sich in gleichem Ausmaß dafür interessieren. So enthielt der Test beispielsweise Fragen zur Geschichte im sozialen/politischen Kontext (z.B. Gandhi), nicht aber zur Geschichte im militärischen Kontext. In Bezug auf Wissenschaften und das Weltgeschehen wurden "questions with lifestyle implication, such as on Chernobyl and everyday science (e.g., sweat, blankets)" (S. 156) verwendet. Auf Basis von Augenscheinvalidität konnten sechs Items des General Information Test nach Meinung von Evans et al. dem Themengebiet der Global Cultural Literacy und sechs Items der Science Literacy zugeordnet werden. Der Test wurde von 1469 Personen aus Taiwan (58.0% weiblich), 1119 aus Japan (42.9% weiblich) und 1003 aus den USA (52.5% weiblich) bearbeitet. Alle Personen waren Schülerinnen und Schüler im Alter von 16 bis 17 Jahren. Die Ergebnisse zeigten kulturübergreifend bessere durchschnittliche Leistung der Jungen bei allen 12 Items. Eine 2-faktorielle Varianzanalyse für die Gesamtscores ergab einen hochsignifikanten Haupteffekt des Geschlechts ($p < .001$). Der Haupteffekt der Kultur und der Interaktionseffekt waren auf dem 1%-Niveau signifikant. Am schwächsten fielen die Geschlechtsunterschiede in den USA aus mit $d = -0.46$. In Taiwan betrug die Effektstärke $d = -0.58$ und in Japan fiel sie mit $d = -0.88$ am stärksten aus.

Die in den vorherigen Abschnitten aufgeführten zahlreichen Belege geben ein eindeutiges Bild der Geschlechterdifferenzen in Tests des Allgemeinen Wissens. Die Leistungsunterschiede zwischen Frauen und Männern in Tests des Allgemeinen Wissens fallen insgesamt einseitig zugunsten von Männern aus und wurden sowohl zeitlich konsistent

als auch kulturübergreifend festgestellt. Es ist jedoch zu beachten, dass bei differenzierter Betrachtung einzelner Themen durchaus Unterschiede in den Geschlechterdifferenzen zu finden sind, was in den Abschnitten 3.1.1 und 3.1.2 für den General Knowledge Test von Irwing, Cammock und Lynn (2001) und für den Wissenstest von Rolfhus und Ackerman (1999) dargelegt wurde. Stumpf und Stanley (1998) berichten über ähnliche Ergebnisse im Rahmen der 1996 vom College Board durchgeführten Advanced Placement Tests für 29 verschiedene Fächer. Auch hier finden sich Unterschiede in der Richtung und Stärke der Geschlechterdifferenzen. Die höchste Effektstärke zugunsten von Männern findet sich mit d = -0.52 im Fach Physics C (Mechanics). Weitere Beispiele für Fächer, in denen die Leistungen von Männern mit d < -0.20 höher als die der Frauen ausfallen, sind Economics, Computer Science A, Chemistry, Government (U.S.), Mathematics und Biology. In einigen Fächern, beispielsweise Music Theory, Art: Studio General und English Literature, wurden nur geringe Leistungsunterschiede festgestellt. Das Fach mit der höchsten Effektstärke zugunsten von Frauen war Spanish Literature (d = 0.14). Stumpf und Stanley geben an, dass im Rahmen der ebenfalls 1996 durchgeführten Scholastic Assessment Test (SAT) II Subject Tests für 21 verschiedene Fächer insgesamt ähnliche Geschlechterdifferenzen festgestellt wurden. Hierbei zeigte sich jedoch in keinem Fach eine Effektstärke zugunsten von Frauen.

Für das Phänomen, dass Geschlechterdifferenzen in Tests des Allgemeinen Wissens einseitig zugunsten von Männern ausfallen, können unterschiedliche Erklärungsansätze herangezogen werden. Eine Reihe von Erklärungen wird im Folgenden erläutert. Hierbei besteht selbstverständlich kein Anspruch auf Vollständigkeit.

3.2 Erklärungsansätze für Geschlechterdifferenzen des Allgemeinen Wissens

Potentielle Erklärungsansätze für Geschlechterdifferenzen in Tests des Allgemeinen Wissens können in zwei Gruppen unterteilt werden. Bei der ersten Gruppe werden

Leistungsunterschiede von Frauen und Männern als eine Abbildung der Realität betrachtet. Hier wird davon ausgegangen, dass Frauen im Durchschnitt tatsächlich über ein geringeres Wissen als Männer verfügen. Diese als gegeben angesehene Tatsache ist somit erklärungsbedürftig. Bei der zweiten Gruppe von Ansätzen wird die Abbildung der Realität durch die unterschiedlichen Leistungen in Tests des Allgemeinen Wissens von Frauen und Männern in Zweifel gezogen. Es wird die Frage gestellt, ob die in standardisierten psychologischen Tests gemessenen Geschlechterdifferenzen möglicherweise ein verzerrtes Bild der tatsächlichen Geschlechterdifferenzen darstellen. Es ist hier nicht etwa ein (möglicherweise) real existierender Unterschied im Allgemeinen Wissen, der erklärt werden soll, sondern vielmehr die einseitige Verzerrung, die bei der Messung des Allgemeinen Wissens (möglicherweise) entsteht. Diese Differenzierung sollte bei den nachstehenden Erklärungsansätzen bedacht werden.

Im Folgenden werden zunächst Erklärungsmöglichkeiten beschrieben, bei denen die Ursache für die Leistungsunterschiede in den formalen Eigenschaften der Wissenstests, beispielsweise im Itemformat, liegt. Darauf folgen Erläuterungen von bisherigen Befunden zu Geschlechterdifferenzen in der Intelligenz. Nach der Investmenttheorie hängt Fluide Intelligenz, wie in Kapitel 2.3.3 erläutert, mit dem Allgemeinen Wissen zusammen. Sofern hier Geschlechterdifferenzen existieren, könnte darin ein Grund für Unterschiede im Wissen liegen. Zuletzt werden Erklärungsansätze beschrieben, die sich auf Persönlichkeitsmerkmale beziehen, und zu denen Befunde sowohl zu Geschlechterdifferenzen als auch zum korrelativen Zusammenhang mit Wissen vorliegen. Die jeweiligen Wirkmechanismen, durch die sich die Persönlichkeitsmerkmale auf das tatsächliche Wissen bzw. auf die Leistung im Wissenstest auswirken, werden hier ebenfalls erläutert. Hierunter fallen auch die beiden Erklärungsmöglichkeiten, die den Forschungsgegenstand der vorliegenden Arbeit darstellen.

Diese werden hier jedoch nur kurz umrissen. Eine ausführliche Behandlung erfolgt in den Kapiteln 4.1 und 5.1.

3.2.1 Formale Testeigenschaften

Messinvarianz ist ein psychometrisches Merkmal eines Tests, das erforderlich ist, um von Gruppenunterschieden in Testergebnissen auf tatsächliche Unterschiede in dem gemessenen Merkmal schließen zu können. Lubke et al. (2003) betonten, dass "to render the group comparisons meaningful it is necessary to address the issue of measurement invariance (MI) and demonstrate that a given test measures the same underlying factors across groups" (S. 544). Hierfür werden häufig Mehrgruppen-Messmodelle verwendet, bei denen Parameter, wie beispielsweise Faktorladungen, Fehlervarianzen und Achsenabschnitte, für sämtliche Gruppen gleichgesetzt werden. Sofern die Modellpassung durch die Restriktionen erheblich geringer ausfällt, ist das Kriterium der Messinvarianz nicht erfüllt. Es existieren unterschiedliche Abstufungen der Messinvarianz, die sich hinsichtlich der Parameter unterscheiden, die für die verschiedenen Gruppen gleichgesetzt werden. So setzt metric invariance beispielsweise lediglich identische Faktorladungen für die Gruppen voraus, während für die Prüfung von scalar invariance zusätzlich die Achsenabschnitte gleichgesetzt werden. Einen anwendungsorientierten Überblick über Messinvarianz und die Unterschiede zwischen den verschiedenen Abstufungen geben beispielsweise Steenkamp und Baumgartner (1998). Eine methodenorientierte Erläuterung der Messinvarianz findet sich bei Meredith (1993). In Kapitel 5.3.1 wird vertieft auf den Begriff eingegangen, da ein neuer Wissenstest entwickelt und auf Messinvarianz geprüft wurde. Bezüglich der Messinvarianz publizierter Tests des Allgemeinen Wissens, deren Geschlechterdifferenzen hier beschrieben wurden, ist dem Autor nur eine Untersuchung bekannt, in der diese geprüft wurde. Steinmayr, Beauducel und Spinath (2010) prüften den I-S-T 2000 R auf Messinvarianz, wobei in dem hier

getesteten Modell Faktorladungen, Faktorkorrelationen (das Modell beinhaltete den gesamten Test), Residualvarianzen und Achsenabschnitte der manifesten Variablen für Frauen und Männer gleichgesetzt waren. Steinmayr et al. fanden eine akzeptable Passung für das Modell. Für den BOWIT, den GKT und die anderen genannten Wissenstests liegen aktuell keine Untersuchungen vor, in denen deren Messinvarianz geprüft wurde. Eine Schlussfolgerung von gemessenen Geschlechterdifferenzen des Wissens auf wahre Unterschiede im Wissen ist somit nur unter Vorbehalt möglich, da Messinvarianz hierfür eine Voraussetzung darstellt. Lediglich für den Wissenstest des I-S-T ist diese Voraussetzung in bisher einer einzelnen veröffentlichten Untersuchung geprüft und bestätigt worden.

Eine weitere Erklärungsmöglichkeit für verzerrte Geschlechterdifferenzen in Wissenstests, die sich auf formale Testeigenschaften bezieht, ist das Itemformat. Die Mehrheit der in Kapitel 3.1 beschriebenen Wissenstests enthält Items im Multiple-Choice-Format (MC). Mondak und Canache (2004) analysierten die Daten von 12 Wissensitems des International Social Survey Program (ISSP) von 1993, welche als Antwortoptionen die drei Möglichkeiten *true, false* und *don't know (DK)* enthielten. "On average, men offered only 0.17 fewer incorrect responses than did women, versus 0.49 fewer DKs" (Mondak & Canache, 2004, S. 545). Der Befund weist auf Geschlechterdifferenzen im Rateverhalten hin, die sich auch in Wissenstests, die auf MC-Items basieren, zugunsten von Männern auswirken könnten. Während nicht beantwortete Items in der Regel als falsch gewertet werden, kann durch Raten zufällig die richtige Antwort gewählt werden. Wissenstests mit Items im MC-Format könnten daher zu einer systematischen Überschätzung des Wissens von Männern führen, sofern Frauen im Vergleich zu Männern Items bei Nicht-Wissen häufiger unbeantwortet lassen. Einen weiteren Beleg für Unterschiede zwischen Frauen und Männern im Rateverhalten lieferten Ben-Shakhar und Sinai (1991): Sie setzten verschiedene MC-Tests ein und stellten fest, dass "correction for guessing changed the pattern of the differences in

favor of females on all subtests" (S. 32), wobei der Effekt der Korrektur für die verschiedenen Tests zwischen 0.01 und 0.15 Standardabweichungseinheiten lag und damit gering ausfiel. Es wurde in der Vergangenheit außerdem über unterschiedliche Geschlechterdifferenzen in Abhängigkeit vom Itemformat berichtet. Mazzeo, Schmitt und Bleistein (1993) analysierten die Ergebnisse von Advanced Placement Prüfungen in den Fächern United States History, Biology, Chemistry und English Language and Composition. Die Prüfungen enthielten Tests im MC-Format und "constructed-response"(CR)-Tests, wobei letztere unterschiedliche Formate umfassten, wie beispielsweise Lückentexte und das Erstellen eigener Aufsätze. In allen vier Fächern und bei sämtlichen ethnischen Gruppen zeigte sich, dass die Geschlechterdifferenzen zugunsten von Männern bei MC-Tests größer ausfielen als bei CR-Tests. Ausschließlich bei CR-Tests zeigten vereinzelt Frauen die durchschnittlich besseren Leistungen. Willingham und Cole (1997) gaben einen Überblick über Geschlechterdifferenzen bei MC- und CR-Tests aus Advanced Placement Tests unterschiedlicher Fächer. Hierbei stellten sie fest, dass "the differences between the two formats vary considerably across subject fields" (S. 174). Auch Willingham und Cole berichteten über substantiell bessere Leistungen von Frauen ausschließlich bei Aufgaben im freien Antwortformat. Jedoch kamen die Autoren nach einer zusammenfassenden Betrachtung von fünf Studien, in denen der Effekt des Itemformats auf Geschlechterdifferenzen geprüft und nicht bestätigt wurde, zu dem Schluss, dass die Ergebnisse "considerably weaken the argument that MC format, per se, is a significant source of gender difference in test results" (Willingham und Cole, 1997, S. 276). Einschränkend muss hier allerdings hinzugefügt werden, dass in vier der fünf Studien Tests verwendet wurden, in denen Fähigkeiten im mathematischen Bereich geprüft wurden. Die thematische Bandbreite ist hier somit deutlich geringer als die der verschiedenen Advanced Placement Tests, über die Willingham und Cole ebenfalls berichteten und bei denen die

Geschlechterdifferenzen zwischen den Formaten deutlich variierten. Koivula, Hassmén und Hunt (2001) verglichen die Geschlechterdifferenzen in zwei Subtests des schwedischen Scholastic Aptitude Tests (SweSAT) bei Verwendung des MC-Formats mit den Geschlechterdifferenzen, die bei Verwendung eines Itemformats entstanden, bei welchem die Personen auch angeben sollten, wie sicher sie sich ihrer Antwort waren. Je höher die Sicherheit, umso mehr zusätzliche Punkte wurden bei korrekter Antwort vergeben beziehungsweise bei falscher Antwort abgezogen. Koivula et al. stellten zwischen beiden Itemformaten keine Unterschiede in den Geschlechterdifferenzen fest. In einer ähnlichen Untersuchung von Hassmén und Hunt (1994) profitierten Frauen von dem alternativen Antwortformat deutlich. Zusammenfassend kann keine eindeutige Antwort auf die Erklärung von Geschlechterdifferenzen durch das Itemformat gegeben werden, da hierzu widersprüchliche Ergebnisse vorliegen. Die Annahme einer geringen Bedeutsamkeit dieses Erklärungsansatzes wird auch durch die Studie von Evans et al. (2002) unterstützt. Hier wurden 12 Items im offenen Format verwendet und wie in Kapitel 3.1.4 berichtet, zeigten Männer bei sämtlichen Items die besseren Leistungen. Sofern ein für Frauen nachteiliger Effekt des Itemformats vorliegt, dürfte er vermutlich insgesamt gering ausfallen.

3.2.2 Fluide Intelligenz

Wie in Kapitel 2.3.3 erläutert, beinhaltet die Investmenttheorie von Cattell (1971/1987) die Annahme, dass die Fluide Intelligenz von maßgeblicher Bedeutung für die Entwicklung der Kristallinen Intelligenz und damit auch für die Entwicklung von Wissen ist. Da gf den Wissenserwerb erleichtert, könnten Geschlechterdifferenzen in diesem Konstrukt als Erklärung für Unterschiede im Wissen dienen. Die bisherige Forschung zu Geschlechterdifferenzen in der Intelligenz ist weit umfangreicher als die Forschung, die sich auf Unterschiede im Wissen konzentriert. Maccoby und Jacklin (1974) gaben einen

ausführlichen Überblick über damalige Studien zu Geschlechterdifferenzen in kognitiven Fähigkeiten und fassten die Befunde zusammen. Sie wiesen darauf hin, dass Mädchen und Jungen in verschiedenen Bereichen der Intelligenz unterschiedliche Stärken und Schwächen aufweisen. Bei diesen Bereichen handelt es sich um verbale, numerische und figurale Fähigkeiten. In Tests verbaler Fähigkeiten, die beispielsweise die schnelle Auffassung verbal ausgedrückter logischer Zusammenhänge und Textverständnis umfassten, wurden in zahlreichen Studien bessere Leistungen bei Mädchen festgestellt. Sofern auch Tests des Allgemeinen Wissens als Tests der verbalen Fähigkeiten definiert wurden, näherten sich die Leistungen von Mädchen und Jungen deutlich an. Bezüglich mathematischer Fähigkeiten berichteten Maccoby und Jacklin über Befunde, die konsistent bei Jungen höhere Leistungen als bei Mädchen zeigten. Für figurale Fähigkeiten, im Englischen häufig als *visual-spatial abilities* bezeichnet, berichteten Maccoby und Jacklin über steigende Geschlechterdifferenzen zugunsten von Männern ab dem frühen Erwachsenenalter. Die Differenzierung zwischen verbalen, numerischen und figuralen Fähigkeiten und die von Maccoby und Jacklin berichteten, hier sehr knapp zusammengefassten Geschlechterdifferenzen wurden häufig repliziert. So gaben beispielsweise Halpern (2012), Lippa (2005) und Hyde (2005) jeweils einen Überblick über Geschlechterdifferenzen in kognitiven Fähigkeiten und erwähnten in diesem Zusammenhang Studien, in welchen die von Maccoby und Jacklin zusammengefassten Befunde bestätigt wurden[2].

[2]Verbale, numerische und figurale Fähigkeiten können jeweils wiederum in mehrere Bereiche unterteilt werden, die hinsichtlich der Geschlechterdifferenzen Unterschiede aufweisen. So differenzieren beispielsweise Linn und Peterson (1985) zwischen den drei Faktoren *spatial perception, mental rotation* und *spatial visualization*, welche figuralen Fähigkeiten zugerechnet werden können. Halpern (2012) fügt hier noch die beiden Bereiche

Die Frage, inwieweit Geschlechterdifferenzen in der Fluiden Intelligenz bestehen, hängt somit wesentlich davon ab, auf welche Weise die Ausprägung in dem Konstrukt erfasst wird. Forschung zu Geschlechterdifferenzen der Intelligenz ist durch das Problem gekennzeichnet, dass "the sexes can be made to differ in either direction, or to be the same, depending on the mix of items included in a test" (Maccoby & Jacklin, 1974, S. 68). So berichteten beispielsweise Lynn und Irwing (2004) in einer Meta-Analyse über signifikante Geschlechterdifferenzen zugunsten von Männern bei Ravens Matrizentest. Ackerman et al. (2001) verwendeten zur Messung von gf fünf Tests, von denen lediglich einer eindeutig zur Erfassung verbaler Fähigkeiten diente. Ein weiterer Test beinhaltete Textaufgaben aus der Mathematik und die übrigen dienten zur Erfassung numerischer oder figuraler Fähigkeiten. Ackerman et al. fanden ebenfalls signifikant bessere Leistungen bei Männern. Dagegen verweisen beispielsweise Kaufman, Kaufman, Liu und Johnson (2009) auf unbedeutende Geschlechterdifferenzen, die sich bei der Messung der Fluiden Intelligenz mit dem Kaufman Brief Intelligence Test (K-BIT) von Kaufman und Kaufman (zitiert nach Kaufman et al., 2009) ergaben. Der Test beinhaltet einen verbalen und einen figuralen Untertest. In der Normierungsstichprobe des I-S-T 2000 R, der verbale, numerische und figurale Untertests enthält, zeigten sich keine bedeutsamen Geschlechterdifferenzen in der Fluiden Intelligenz. Die Antwort auf die Frage, ob Geschlechterdifferenzen in der Fluiden Intelligenz existieren, hängt somit von der Operationalisierung dieses Konstrukts ab. Sofern gf durch Aufgaben figuralen Inhalts operationalisiert wird, können Geschlechterdifferenzen angenommen werden, welche als potentielle Erklärung für die Unterschiede zwischen Frauen und Männern

---

spatiotemporal ability und *generation and maintenance of a spatial image* hinzu und berichtet über Geschlechterdifferenzen, die zwischen den fünf Bereichen stark variieren. Ähnlich heterogen sind verbale und numerische Fähigkeiten.

im Allgemeinen Wissen dienen. Beauducel, Brocke und Liepmann (2001) kritisierten die in der Literatur häufig anzutreffende, durch Faktorenanalysen begründete Gleichsetzung von gf mit figuralen und von gc mit verbalen Fähigkeiten. Nach Meinung von Beauducel et al. (2001) "this view leads away from the original conceptualization of gf and gc not as (figural or verbal) content factors, but as dependent on the degree of acculturation" (S. 980). Die Autoren schlagen daher sowohl für gf als auch für gc die Operationalisierung durch Aufgaben verbalen, numerischen und figuralen Inhalts vor, was im I-S-T 2000 R von Liepmann et al. (2007) umgesetzt wurde. Bei dieser Auffassung von Fluider Intelligenz muss die Frage nach Geschlechterdifferenzen verneint werden, womit gf als Erklärungsmöglichkeit für Geschlechterdifferenzen im Allgemeinen Wissen nicht in Frage kommt.

3.2.3 Persönlichkeitsmerkmale

Persönlichkeitsmerkmale, von denen angenommen wird, dass sie mit Wissenserwerb in Zusammenhang stehen und dass Frauen und Männer hierin Unterschiede aufweisen, stellen ebenfalls potentielle Erklärungen für Geschlechterdifferenzen in Tests des Allgemeinen Wissens dar. Hossiep und Schulte (2008) weisen beispielsweise auf Geschlechterdifferenzen in den Strategien des Wissenserwerbs hin: "Während für Männer vor allem Tageszeitungen und Zeitschriften als Wissensquellen herangezogen werden, sind für Frauen neben persönlichen Gesprächen vor allem Bücher von Bedeutung. Hierbei überwiegen im Vergleich vor allem Bücher belletristischer Art" (S. 53). Ein Beleg für die Annahmen wird jedoch nicht gegeben, weshalb sie als Spekulationen eingeordnet werden müssen.

Verschiedene Persönlichkeitsmerkmale werden mit Neugier in Zusammenhang gebracht und könnten daher für den Erwerb von Wissen, vor allem unter Annahme der in Kapitel 2.3.3 beschriebenen Investmenttheorie, relevant sein. Hierunter fallen beispielsweise die Konstrukte Typical Intellectual Engagement (Goff und Ackerman, 1992) und Offenheit für

Erfahrungen. Nach Ostendorf und Angleitner (2004) beschreiben sich Personen mit einer stark ausgeprägten Offenheit für Erfahrungen "als vielfältig interessiert, wissenshungrig, schöpferisch und interessiert an Theorie und am kulturellen Geschehen, als geneigt, bestehende Normen und Wertvorstellungen kritisch zu hinterfragen und als bereit, sich mit neuen ethischen, politischen und sozialen Themen und Orientierungen zu beschäftigen" (S. 42). Die Messung von Typical Intellectual Engagement dient nach Goff und Ackerman (1992) "to differentiate among individuals in their typical expression of a desire to engage and understand their world, their interest in a wide variety of things, and their preference for a complete understanding of a complex topic or problem, a need to know" (S. 539). Stärkere Ausprägungen in diesen Konstrukten beinhalten somit eine höhere intellektuelle Neugier. Sofern die intellektuelle Neugier bei Männern stärker als bei Frauen sein sollte, könnten Männer einen höheren Bedarf an Wissen aufweisen und andere Verhaltensweisen beim Erwerb von Wissen zeigen. Die Spekulationen von Hossiep und Schulte (2008) zu geschlechtsspezifischen Strategien des Wissenserwerbs wären ein mögliches Beispiel dafür. Mit den Begriffen von Cattell (1971/1987) würde man dies als ein stärkeres Investment bezeichnen. Die Beschreibungen der Konstrukte von Ostendorf und Angleitner (2004) bzw. Goff und Ackerman (1992) verdeutlichen außerdem, dass die Annahme, dass Personen mit hohen Ausprägungen in den beschriebenen Persönlichkeitsfaktoren über ein höheres Wissen verfügen, auch im Einklang mit der in Kapitel 2.3.5 beschriebenen Theorie der Verarbeitungstiefe nach Craik und Lockhart (1972) steht. Auf korrelativer Ebene wurde mehrfach über einen positiven Zusammenhang zwischen Typical Intellectual Engagement und Kristalliner Intelligenz berichtet (z.B. Ackerman, 2000a; Ackerman & Heggestad, 1997; Wilhelm, Schulze, Schmiedeck & Süß, 2003). Chamorro-Pemuzic, Furnham und Ackerman (2006) fanden eine geringe, aber signifikante positive Korrelation zwischen Offenheit für Erfahrungen und Allgemeinem Wissen, das über den GKT von Irwing et al. (2001) erfasst

worden war. Von Stumm und Ackerman (2013) publizierten eine Meta-Analyse zum Zusammenhang zwischen Intellekt und Persönlichkeitsmerkmalen, die sie als *investment traits* bezeichneten. In den von ihnen herangezogenen Studien dienten beispielsweise Kristalline Intelligenz, akademische Leistung und Wissen als Indikatoren des Intellekts. Beispiele für investment traits sind Need for Cognition, Openness to Experience, Sensation Seeking und Typical Intellectual Engagement (siehe von Stumm & Ackerman, 2013). Die Korrelation zwischen Indikatoren des Intellekts und den verschiedenen investment traits betrug durchschnittlich .30. Bezüglich Wissen fanden von Stumm und Ackerman hauptsächlich Studien, in denen über den Zusammenhang mit Typical Intellectual Engagement berichtet wurde. Die Korrelation zwischen diesen beiden Merkmalen betrug durchschnittlich .38. Die Ergebnisse bestätigen den beschriebenen Mechanismus als potentielle Erklärung für Geschlechterdifferenzen im Wissen. Voraussetzung hierfür wäre allerdings, dass neugierbezogene Persönlichkeitsmerkmale bei Männern durchschnittlich stärker ausgeprägt sind. Wilhelm et al. (2003) fanden jedoch höhere Ausprägungen für Typical Intellectual Engagement bei Frauen. Bezüglich Offenheit für Erfahrungen liegen widersprüchliche Ergebnisse vor. Beispielsweise untersuchten Gjerde und Cardilla (2009) 102 Personen im Alter zwischen 3 und 23 Jahren und fanden keine Geschlechterdifferenzen. Lehmann, Denissen, Allemand und Penke (2013) konnten auf die Daten von insgesamt 19022 Personen zwischen 16 und 60 Jahren zurückgreifen. Hier zeigte sich bei Männern eine höhere Ausprägung der Offenheit für Erfahrungen bei sämtlichen Altersstufen. Misra (2003) verglich die Ausprägungen bei Frauen und Männern anhand einer Stichprobe von insgesamt 156 Personen im Alter zwischen 19 und 26 Jahren. Hier waren die Ausprägungen der Frauen im Durchschnitt höher. Insgesamt ist die Befundlage für die Erklärung der Geschlechterdifferenzen durch intellektuelle Neugier somit nicht eindeutig. Zwar dürfte der positive Zusammenhang zwischen Offenheit für Erfahrungen bzw. Typical Intellectual

Engagement einerseits und Wissen andererseits kaum anzuzweifeln sein, die Frage nach Geschlechterdifferenzen der intellektuellen Neugier kann jedoch nicht eindeutig beantwortet werden, da hierzu sehr widersprüchliche Ergebnisse vorliegen, insbesondere in Bezug auf die Offenheit für Erfahrungen. Bezüglich Typical Intellectual Engagement lassen die bisherigen empirischen Befunde aber Zweifel an dem Erklärungsansatz aufkommen, da dieses Persönlichkeitsmerkmal bei Frauen vermutlich durchschnittlich höher ausgeprägt ist.

Extraversion ist ein weiteres Persönlichkeitsmerkmal, bei dem Frauen und Männer im Mittel Differenzen aufweisen und für das korrelative Zusammenhänge mit Allgemeinem Wissen gefunden wurden. Ostendorf und Angleitner (2004) nennen zur Beschreibung von Extraversion beispielsweise die Adjektive "freundschaftlich, geschwätzig, gesellig, gesprächig" (S. 39) sowie "personenorientiert, redselig, sich behauptend" (S. 39). Die genannten Adjektive sind für sich genommen vermutlich weder förderlich noch hinderlich für den Erwerb von Wissen. Bedenkt man jedoch, dass Zeit eine begrenzte Ressource ist und dass hoch Extravertierte in hohem Ausmaß in soziale Kontakte investieren und zu den genannten Eigenschaften tendieren, dann liegt es aus Perspektive der Investmenttheorie nahe, dass dies auf Kosten zeitlicher Ressourcen geht, die für die Investition in intellektuelle Tätigkeiten zur Verfügung stehen. Extraversion dürfte sich somit eher negativ als positiv auf den Erwerb von Wissen auswirken. Empirische Befunde bestätigen die Vermutung: So berichteten beispielsweise Ackerman et al. (2001) und Chamorro-Premuzic et al. (2006) über einen negativen Zusammenhang zwischen Extraversion und Allgemeinem Wissen. Bezüglich Geschlechterdifferenzen in der Extraversion fanden beispielsweise Lehmann et al. (2013) für sämtliche Altersgruppen bei Frauen eine durchschnittlich höher ausgeprägte Extraversion als bei Männern.

Für Persönlichkeitsmerkmale, die im Zusammenhang mit der Emotion der Angst stehen, wurden ebenfalls höhere Ausprägungen bei Frauen und negative Korrelationen mit

Leistungen in Wissenstests gefunden. Beispielsweise berichteten Ackerman et al. (2001) über negative Zusammenhänge zwischen Allgemeinem Wissen und Kristalliner Intelligenz einerseits und *emotionality* und *worry* andererseits. Es fanden sich in der bisherigen Forschung wiederholt Hinweise darauf, dass angstbezogene Persönlichkeitsmerkmale negativ mit Leistungen in Wissenstests korrelieren (z.B. Ackerman et al., 2001; Chamorro-Premuzic et al., 2006) und dass Frauen höhere Ausprägungen angstbezogener Persönlichkeitsmerkmale aufweisen (z.B. Ackerman et al., 2001; Kanfer und Ackerman, 2000). Kanfer und Heggestad (1999) berichteten, dass angstbezogene Emotionen und ein Mangel an Fähigkeit, solche Emotionen zu reduzieren "effectively reduce the individual's available cognitive resources for the task" (S. 297). Hierbei bezogen sie sich jedoch nicht auf Aufgaben aus Wissenstests, sondern auf "complex, but consistent and learnable" (S. 297) Aufgaben, wie z.B. die Simulation der Tätigkeit von Fluglotsen. Sofern dieser Wirkmechanismus auch auf Aufgaben aus Wissenstests übertragen werden kann, könnten sich derartige Emotionen negativ auf die Leistung von Frauen in Wissenstests auswirken.

Zwei weitere Erklärungsansätze, die mit der Persönlichkeit in Zusammenhang stehen, bilden die zentralen Inhalte der vorliegenden Arbeit und werden in den beiden folgenden Kapiteln detailliert beschrieben. Hierbei handelt es sich zum einen um Geschlechterdifferenzen in der Einschätzung des eigenen Wissens und zum anderen um Interessenunterschiede zwischen Frauen und Männern. Ähnlich wie eine höhere Ausprägung von angstbezogenen Persönlichkeitsmerkmalen könnte auch eine durchschnittlich geringere Einschätzung des eigenen Wissens zu einer stärkeren emotionalen und kognitiven Belastung in der Testsituation führen, was möglicherweise geringere Leistungen in den Wissenstests zur Folge hätte. In Kapitel 4.1 wird dieser Mechanismus näher erörtert. Die Investmenttheorie lässt annehmen, dass Personen über ein umso größeres Wissen über ein Thema verfügen, je stärker sie sich dafür interessieren. Sofern Frauen und Männer Unterschiede in ihren

Interessen aufweisen, könnten die Geschlechterdifferenzen in Wissenstests darauf basieren, dass hierin hauptsächlich das Wissen zu Themen erfasst wird, für die Männer sich durchschnittlich stärker interessieren. In Kapitel 5.1 wird dieser Erklärungsansatz thematisiert.

Selbsteinschätzung

In diesem Kapitel wird die Untersuchung beschrieben, in welcher der Frage nachgegangen wurde, ob Unterschiede zwischen Frauen und Männern in der Einschätzung ihres eigenen Wissens ursächlich für Leistungsunterschiede in Wissenstests sein könnten. Nach der Erläuterung der relevanten theoretischen Hintergründe und bisherigen empirischen Befunde wird zunächst ein möglicher Mechanismus beschrieben, durch den Selbsteinschätzung auf die Leistung in Wissenstests wirken könnte. Daraus wird die Hypothese abgeleitet, dass die Einschätzung des eigenen Wissens von ursächlicher Bedeutung für Geschlechterdifferenzen in Wissenstests ist. Im Anschluss erfolgen die Beschreibungen der methodischen Vorgehensweise und der Ergebnisse des Experiments, durch das die Hypothese geprüft wurde. Zuletzt werden die Schlussfolgerungen aus der Untersuchung gezogen.

4.1 Theorie

Es existieren in der Psychologie zahlreiche Konzepte, die mit der Selbsteinschätzung in Zusammenhang stehen. Hierbei wird jedoch nicht einheitlich der Begriff "Selbsteinschätzung" verwendet. Weitere Begriffe, die hiermit in Zusammenhang stehen, sind beispielsweise *self-concept*, *self-esteem* und *self-efficacy*. Eine Abgrenzung findet sich beispielsweise bei Valentine, DuBois und Cooper (2004). Nach diesen Autoren teilen sämtliche Begriffe "a common emphasis on an individual's beliefs about his or her attributes and abilities as a person" (S. 112). Die Einschätzung eigener Fähigkeiten ist somit ein gemeinsamer Bestandteil. Als Oberbegriff verwendeten Valentine et al. den Ausdruck *self-belief*. In ihrer Meta-Analyse prüften die Autoren die Beziehung verschiedener Konzepte, die unter self-belief zusammengefasst wurden und akademischer Leistung. Als Maß der

Effektstärke verwendeten sie den standardisierten Regressionskoeffizienten, der bei einer Regression von akademischer Leistung auf self-belief berechnet wurde. Valentine et al. konstatierten, dass die verschiedenen Konzepte des self-belief sich nicht signifikant in ihren Zusammenhängen mit der späteren akademischen Leistung unterschieden. Insgesamt variierten die Effektstärken in den 60 Untersuchungen, die der Meta-Analyse zugrunde lagen, zwischen -0.12 und 0.36. Der Mittelwert betrug 0.08 und das 95%-Konfidenzintervall umfasste nicht den Wert 0, weshalb Valentine et al. auf einen positiven, aber geringen Zusammenhang zwischen self-belief und akademischer Leistung schlossen.

Freund und Kasten publizierten 2012 eine Meta-Analyse, in der die Validität der Einschätzung eigener kognitiver Fähigkeiten überprüft wurde. Der Begriff der Selbsteinschätzung (*self-estimation*) wurde hier definiert als "a person's perception of her or his own abilities. Self-estimation is a process that is based on and involves repeated assessments in a variety of different concrete situations. Accordingly, this leads to domain specific ability self-estimates" (S. 299). Für die vorliegende Untersuchung wird diese Definition übernommen. Die Fähigkeit, deren Einschätzung Gegenstand der vorliegenden Arbeit ist, ist das deklarative semantische Wissen. Die Domänenspezifität drückt sich hier in den verschiedenen inhaltlichen Themengebieten aus. So mag beispielsweise die Selbsteinschätzung des Wissens im Bereich Wirtschaft von der Selbsteinschätzung des Wissens im Bereich Kunst abweichen.

Bei der Einschätzung eigener kognitiver Fähigkeiten wurden in der Vergangenheit häufig Geschlechtsunterschiede festgestellt. In der großen Mehrzahl der Studien wurde über höhere Selbsteinschätzungen bei Männern als bei Frauen berichtet. In einem Überblicksartikel beschrieb Furnham (2001) acht Untersuchungen, in denen Geschlechtsunterschiede in der Einschätzung der eigenen allgemeinen Intelligenz überprüft wurden. In sieben der acht Untersuchungen wurden signifikante Unterschiede festgestellt, wobei Männer in allen sieben

Untersuchungen die durchschnittlich höheren Selbsteinschätzungen abgaben. Über Ergebnisse, die in eine ähnliche Richtung weisen, berichteten beispielsweise Rammstedt und Rammsayer (2000), von Stumm, Chamorro-Premuzic und Furnham (2009) sowie Yuen und Furnham (2005). Bezüglich der Validität der Selbsteinschätzung kognitiver Leistungsfähigkeit stellten Freund und Kasten (2012) in ihrer Meta-Analyse keine Geschlechterdifferenzen fest: "Relying exclusively on either female or male samples did not significantly influence the validity of self-estimates of intelligence, compared to the standard case of using mixed samples" (S. 310). Die durchschnittliche Korrelation zwischen Selbsteinschätzung und intellektueller Leistung, die Freund und Kasten in ihrer Meta-Analyse berechneten, betrug r = .33 und lag damit sehr nahe an dem Befund, über den Mabe und West (1982) in ihrer Meta-Analyse zur gleichen Forschungsfrage berichteten (r = .34).

Zum Zeitpunkt der Erstellung dieser Arbeit konnten keine Studien ermittelt werden, in denen Geschlechterdifferenzen in der Selbsteinschätzung des Allgemeinen Wissens erfasst wurden. Es lagen im Vorfeld der Untersuchung somit keine Belege dafür vor, dass Männer ihr eigenes Wissen im Durchschnitt höher als Frauen einschätzen. In der bisherigen Forschung wurde jedoch über Geschlechterdifferenzen in der Selbsteinschätzung von kognitiven Fähigkeiten berichtet, die mit Wissen zusammenhängen. Ein Beispiel hierfür bilden Geschlechterdifferenzen in der Selbsteinschätzung der allgemeinen Intelligenz nach Spearman (z.B. von Stumm et al., 2009). Allgemeine Intelligenz umfasst, wie in Kapitel 2.3.3 beschrieben, auch auf Erfahrungen basierende intellektuelle Fähigkeiten. Ein empirischer Beleg für die Annahme, dass Männer auch explizit das eigene Wissen höher als Frauen einschätzen, steht jedoch aus und soll ebenfalls durch diese Untersuchung geliefert werden.

Für die Hypothese, dass geringere Selbsteinschätzungen des Wissens von Frauen negativ auf die Leistung in Wissenstests wirken, wird folgender Wirkmechanismus angenommen: geringere Einschätzung des eigenen Wissens führt zu Emotionen, die im Zusammenhang mit

Angst stehen, und damit zur Reduktion von kognitiven Ressourcen, die bei der Bearbeitung des Wissenstests verfügbar sind. Da Frauen eine geringere Selbsteinschätzung des Wissens aufweisen als Männer, werden hiermit die bei der Testbearbeitung zur Verfügung stehenden kognitiven Ressourcen von Frauen im Durchschnitt reduziert, was zu geringeren durchschnittlichen Leistungen in den Wissenstests führt. Wie bereits in Kapitel 3.2.3 beschrieben, wurde ein negativer korrelativer Zusammenhang zwischen kristalliner Intelligenz und Emotionen, die mit Angst zusammenhängen, bereits mehrfach nachgewiesen (z.B. Ackerman et al., 2001; Chamorro-Premuzic et al., 2006). Kanfer und Ackerman (1989) führten eine Untersuchung durch, bei der die Probandinnen und Probanden an einer Aufgabe teilnahmen, welche die Simulation der Tätigkeit von Fluglotsen beinhaltete. Kanfer und Ackerman (1996) berichteten über die Aussagen von Teilnehmenden der Studie zu emotionalen Empfindungen während der Aufgabenbearbeitung. Teilnehmende mit geringeren Leistungen berichteten beispielsweise über "thinking about doing more poorly than others on the task, dissatisfaction and anger with oneself for making mistakes, and feelings of unhappiness. In contrast, higher performing trainees reported having significantly fewer of these thoughts during task engagement" (Kanfer & Ackerman, 1996, S. 161). Hieraus schlossen Kanfer und Ackerman (1996) auf die negativen Auswirkungen angstbezogener Emotionen auf kognitive Leistungen. Es stellt sich die Frage, welche kognitiven Ressourcen konkret von negativen Emotionen betroffen sind. Eysenck und Calvo (1992) postulierten in ihrer processing efficiency theory, dass das Arbeitsgedächtnis, insbesondere die zentrale Exekutive aus dem Arbeitsgedächtnismodell von Baddeley (1986), durch angstbezogene Emotionen belastet wird. Ashcraft und Kirk (2001) untersuchten die Zusammenhänge zwischen Mathematik-Angst, Arbeitsgedächtnis und Leistung in Mathematikaufgaben. In Anknüpfung an Eysenck und Calvo schlussfolgerten Ashcraft und Kirk aus ihren Ergebnissen, dass die Angstreaktion die Kapazitäten des Arbeitsgedächnisses beeinträchtigt:

"The draining of resources implies continued, inappropriate (and self-defeating) attention to the cognitive components of the math-anxiety reaction and to intrusive thoughts, worry, preoccupation with performance evaluation, and the like" (S. 236). Bonnot und Croizet (2007) verglichen die Leistungen in Mathematikaufgaben von Frauen mit hoher und niedriger Einschätzung der eigenen mathematischen Fähigkeiten. Sie stellten fest, dass die Leistungsunterschiede am stärksten ausfielen, wenn das Arbeitsgedächtnis durch die Bearbeitung schwieriger Aufgaben in hohem Maße beansprucht wurde. Unter Bezugnahme auf Ashcraft und Kirk schlugen Bonnot und Croizet als eine mögliche Erklärung für den Effekt vor, dass bei Frauen mit geringer Einschätzung ihrer eigenen mathematischen Fähigkeiten diesbezügliche negative Emotionen ausgelöst werden, welche die Kapazitäten des Arbeitsgedächtnisses einschränken.

Es muss betont werden, dass in keiner der Studien, die zur Beschreibung der postulierten Wirkungsweise von Selbsteinschätzung auf die Leistung in Wissenstests herangezogen wurden, Testmaterial verwendet wurde, das zur Erfassung des Allgemeinen Wissens dient. In der Untersuchung von Kanfer und Ackerman (1989) sollten die Teilnehmenden den simulierten Flugverkehr regeln. Die Fähigkeit, die Aufgabe zu meistern umfasste nach den Autoren (a) den Erwerb neuen deklarativen Wissens, (b) die Integration der kognitiven und motorischen Fähigkeiten und (c) deren Transfer in prozedurales Wissen. Der Abruf deklarativen semantischen Wissens war somit kein Bestandteil der Aufgabe. Sowohl Ashcraft und Kirk (2001) als auch Bonnot und Croizet (2007) setzten in ihren Untersuchungen Mathematikaufgaben ein. Die von den Probandinnen und Probanden zu lösenden Aufgaben wie auch die negativen Emotionen bezogen sich hier ausschließlich auf Mathematik. Bei dem in der vorliegenden Arbeit angenommenen Wirkmechanismus werden die in den beschriebenen Studien gewonnenen Erkenntnisse auf Wissenstests übertragen. Dieser Transfer ist jedoch nicht selbsterklärend, sondern bildet eine eigene Forschungsfrage. So

wäre beispielsweise eine Voraussetzung hierfür, dass bei der Bearbeitung von Wissenstests das Arbeitsgedächtnis beansprucht wird. Diese Forschungsfrage wird im Rahmen der vorliegenden Untersuchung jedoch nicht weiter betrachtet. Gegenstand ist hier die Prüfung der Hypothese, dass die Einschätzung des eigenen Wissens von ursächlicher Bedeutung für die in Wissenstests erbrachte Leistung ist. Sollte sich die Hypothese bestätigen und sollten Frauen eine durchschnittlich geringere Einschätzung des eigenen Wissens aufweisen, könnte hieraus geschlussfolgert werden, dass die durchschnittlich geringere Leistung von Frauen in Wissenstest – zumindest teilweise – auf eine durchschnittlich geringere Einschätzung des eigenen Wissens zurückzuführen ist.

Ein Phänomen, welches in engem Zusammenhang mit der Einschätzung eigener Fähigkeiten und deren Auswirkungen auf die Leistung in Testsituationen steht, ist der *stereotype threat* nach Steele (1997). Der stereotype threat fand in der Forschung viel Beachtung und soll daher auch hier kurz umrissen werden: Sofern für eine Gruppe von Menschen ein negativer Stereotyp vorherrscht und diesem Stereotyp in der Testsituation Relevanz beigemessen wird, kann es unter den Angehörigen der betroffenen Gruppe zur Leistungsabnahme in der Testsituation kommen. Eine Voraussetzung hierfür ist, dass die Fähigkeit, auf die sich der Stereotyp bezieht, für die Person von persönlicher Bedeutung ist. In Bezug auf Geschlechter wurde der Effekt beispielsweise für die Leistungen in Mathematikaufgaben unter Studierenden festgestellt (Spencer, Steele & Quinn, 1999). In der Forschungsliteratur konnten jedoch keine Hinweise auf die Existenz des Stereotyps, dass Frauen über ein geringeres Allgemeines Wissen als Männer verfügen, gefunden werden. Daher wird im Folgenden kein weiterer Bezug auf diese Theorie genommen.

4.2 Methoden

Das Ziel der im Folgenden beschriebenen Untersuchung bestand in der Überprüfung der Hypothese, dass die Selbsteinschätzung einer Person über ihr Allgemeines Wissen von maßgeblicher Bedeutung für die erbrachte Leistung im Test ist. Zunächst werden die Datenerhebung und das dabei verwendete Testmaterial beschrieben. Im Anschluss daran erfolgt die Beschreibung der verwendeten Analysemethode.

4.2.1 Datenerhebung

Das experimentelle Design umfasste zwei Schritte: zunächst sollte die Einschätzung des eigenen Allgemeinen Wissens durch Manipulation in der einen Gruppe gesteigert und in der anderen Gruppe verringert werden. Anschließend bearbeiteten alle Teilnehmenden den Wissenstest des I-S-T 2000 R (Liepmann et al., 2007). Es wurde erwartet, dass die Mitglieder der ersten Gruppe, deren Selbsteinschätzung positiv beeinflusst worden war, eine durchschnittlich bessere Leistung in dem Wissenstest aufweisen würden als die Mitglieder der zweiten Gruppe.

Die Datenerhebungen wurden an einem Gymnasium und an einer Berufsschule für soziale Berufe durchgeführt. Die Erhebung an dem Gymnasium fand im Rahmen einer Diplomarbeit statt. Insgesamt wurden die Daten von 333 Personen erfasst, von denen jedoch einige aussortiert werden mussten, da die Teilnehmenden beispielsweise seit weniger als 10 Jahren Deutsch sprachen oder bei der Variable Geschlecht einen fehlenden Wert aufwiesen. Die Analysen wurden mit einer Stichprobe von insgesamt 316 Personen durchgeführt. Es lagen die Daten von 144 Schülerinnen und Schülern des Gymnasiums und 172 Schülerinnen und Schülern der Berufsschule vor. Unter den Teilnehmenden waren 228 Frauen im Alter von 16 bis 47 Jahren (M = 18.82, SD = 2.95). Eine weitere Frau hatte keine Angaben zu

ihrem Alter gemacht. Die 87 Männer waren zwischen 16 und 38 Jahre alt (M = 19.91, SD = 4.18).

Die Manipulation erfolgte in Form von zwei verschiedenen Versionen eines Fragebogens. In der einen Version wurden neun leichte und in der anderen Version neun schwierige Wissensitems präsentiert, anhand derer die Personen ihr eigenes Wissen einschätzen sollten. Beide Versionen des Fragebogens können als Supplement vom Autor angefordert werden.

Zu Beginn wurde in Form von schriftlichen Instruktionen die Verteilung des Wissens in der Bevölkerung erläutert. Hierzu wurde eine Normalverteilung grafisch dargestellt. Die Abszisse der Abbildung war mit *Wissen* beschriftet. Die Skala enthielt die Eintragungen *sehr niedrig*, *niedrig*, *mittel*, *hoch* und *sehr hoch*. Aus dem Text ging hervor, dass anhand der Grafik zu erkennen ist, dass die meisten Personen ein mittleres Wissen aufweisen, während wenige Personen ein niedriges beziehungsweise hohes und sehr wenige Personen ein sehr niedriges beziehungsweise sehr hohes Wissen haben. Die Teilnehmenden wurden darauf hingewiesen, dass sie im Folgenden angeben sollten, wie hoch sie ihr eigenes Wissen *im Vergleich zu Personen des gleichen Alters* einschätzen. Da die Angaben in Form von Prozenträngen erfolgen sollten, wurde außerdem eine Normalverteilung mit Markierungen von Prozenträngen in 5er-Abständen dargestellt und anhand eines beispielhaften Eintrags die Bearbeitung des Fragebogens erläutert.

Anschließend wurde die Manipulation der Einschätzung des eigenen Wissens durch die Präsentation von neun Beispielitems vorgenommen. Grundlage hierfür waren die drei Domänen (verbal, numerisch, figural) und die sechs Themen (Geographie/Geschichte, Wirtschaft, Kunst/Kultur, Mathematik, Naturwissenschaften, Alltag) des Wissenstests des I-S-T 2000 R. Jede Domäne und jedes Thema wurde kurz inhaltlich beschrieben. Auf jede Beschreibung folgte die Präsentation eines beispielhaften Items aus der jeweiligen Kategorie

mit der Information, dass etwa 50 von 100 Personen das Item korrekt bearbeiten könnten. Die erste Version des Fragebogens enthielt neun leichte Items und die zweite Version neun schwierige Items. Nach der Präsentation jedes einzelnen Items wurden die Teilnehmenden gefragt, wie schwierig sie persönlich das gezeigte Item einschätzen würden. Die Antwortoptionen reichten hierbei auf einer 5-stufigen Likert-Skala von *sehr leicht* bis *sehr schwierig*. Anschließend sollten die Teilnehmenden angeben, wie hoch sie den Anteil der Personen ihres Alters einschätzen, der in der jeweiligen Domäne beziehungsweise in dem jeweiligen Thema ein geringeres Wissen als sie hat. Zur Erinnerung wurde hier jeweils erneut eine Normalverteilung präsentiert, bei der die Prozentränge in 5er-Abständen markiert waren. Die Selbsteinschätzung erfolgte somit auf Basis eines sozialen Vergleichsprozesses. Bei erfolgreicher Manipulation sollten Teilnehmende, denen leichte Items vorgelegt worden waren, höhere Prozentränge angeben als Teilnehmende, denen schwierige Beispielitems vorgelegt worden waren. Die beschriebene Vorgehensweise der Manipulation durch leichte beziehungsweise schwierige Beispielitems wurde beispielsweise im dritten Experiment von Moore und Kim (2003) erfolgreich angewandt.

In ihrer Meta-Analyse untersuchten Freund und Kasten (2012) in Anlehnung an Mabe und West (1982) die Effekte einiger Moderatorvariablen auf die Stärke des Zusammenhangs zwischen Selbsteinschätzung der kognitiven Fähigkeiten und deren Messung. Eine Moderatorvariable ist beispielsweise die Methode der Selbstbeurteilung: Freund und Kasten nahmen an, dass eine hohe Validität der Selbsteinschätzung vor allem in Studien erreicht würde, bei denen die Selbsteinschätzung auf einem sozialen Vergleichsprozess basierte. Hierbei sollte nach Möglichkeit eine Referenzgruppe benannt werden. Die Erläuterung und Erklärung einer Normalverteilungskurve oder die Verwendung einer Skala, die mit relativen Begriffen gekennzeichnet ist (beispielsweise "überdurchschnittlich"), wären nach den Annahmen von Freund und Kasten ebenfalls förderlich für die Validität. Diese Annahmen

wurden in der Meta-Analyse im Wesentlichen bestätigt. Außerdem wiesen Freund und Kasten darauf hin, dass "an aggregate score covers a variety of different sources and is therefore much more diverse than a one-item measure, which explicitly demands cognitive integration of a number of concrete experiences" (S. 315). Im vorliegenden Experiment wurden (a) eine Bezugsgruppe benannt, (b) die Normalverteilung dargestellt und beschrieben sowie (c) als Skala Prozentränge verwendet. Außerdem (d) erfolgte die Manipulation und Messung der Selbsteinschätzung nicht durch ein einzelnes Item zur Fähigkeit des Allgemeinen Wissens, sondern durch neun themen- und domänenspezifische Items. Insgesamt wurde in dem hier vorgestellten Experiment also eine sehr valide Methode der Messung der Selbsteinschätzung verwendet und eine zielgerichtete, weil sehr spezifische, Manipulation der Selbsteinschätzung vorgenommen.

Die beiden Versionen des Fragebogens wurden unabhängig vom Geschlecht zufällig unter allen Teilnehmenden verteilt. Sowohl Frauen als auch Männer sollten teilweise positiv und teilweise negativ in ihrer Selbsteinschätzung des Allgemeinen Wissens manipuliert werden.

Die für die Manipulation verwendeten Items wurden dem BOWIT (Hossiep & Schulte, 2008) entnommen. Aus dem Wissenstest des I-S-T 2000 R wurden keine Items zur Manipulation eingesetzt, da dieser Test im weiteren Verlauf des Experiments zur Erfassung des Wissens verwendet wurde. Es wurden Items ausgewählt, die sich möglichst eindeutig den einzelnen Themen und Domänen zuordnen ließen. Anhand der Itemschwierigkeiten, die im Manual des BOWIT angegeben werden, wurden schwierige und leichte Items ausgewählt. In Tabelle 4 sind die für die einzelnen Kategorien ausgewählten Items des BOWIT (Form A) und zugehörigen Schwierigkeiten in der Normierungsstichprobe aufgelistet.

Tabelle 4

Zur Manipulation der Selbsteinschätzung eingesetzte Items des BOWIT (Form A) und zugehörige Schwierigkeiten

	Itemnummer (Schwierigkeit)	
Kategorie	Leichte Items	Schwierige Items
Verbales Wissen	33 (.88)	146 (.16)
Numerisches Wissen	49 (.78)	135 (.20)
Geographie/ Geschichte	04 (.93)	147 (.26)
Wirtschaft	44 (.85)	154 (.32)
Kunst/ Kultur	30 (.91)	144 (.34)
Mathematik	62 (.67)	139 (.44)
Naturwissenschaften	24 (.63)	145 (.29)
Alltag	58 (.83)	151 (.25)

Da der BOWIT keine Items mit figuralem Material enthält, wurde der Versuch unternommen, hierzu ein schwieriges und ein leichtes Item zu konstruieren. In Anlehnung an zwei figurale Items des I-S-T wurden zwei ähnliche Items erstellt. Da im Manual des I-S-T für die einzelnen Items nicht die Schwierigkeiten aufgeführt sind, die sich aus der Normierungsstichprobe ergaben, wurde auf die Daten der Arbeit von Engelberg (2008) zurückgegriffen. Die Stichprobe umfasste hier 273 Studierende (63.0% Frauen) der Universitäten in Münster und Wuppertal. Hier wies das Item mit der Nummer 246 eine Schwierigkeit von .70 und das Item mit der Nummer 285 eine Schwierigkeit von .52 auf, wobei sich beide Itemnummern auf Form A des Tests beziehen. Die an diese Items angelehnten, neu erstellten Items beinhalteten das Erkennen des Grundrisses einer gotischen Kirche (schwierig) und der geometrischen Figur eines Trapezes (leicht). Durch die Frage an die Probandinnen und Probanden, wie schwierig sie persönlich die einzelnen Items einschätzten, konnte im Anschluss an die Datenerhebung überprüft werden, ob das Ziel der Entwicklung eines schwierigen beziehungsweise leichten Items gelungen war. Ein *t*-Test

zeigte hier mit t_{299} = -4.288 (p < .001) ein hoch signifikantes Ergebnis in die gewünschte Richtung. Das Item, das die Abbildung des Grundrisses einer gotischen Kirche enthielt, wurde als schwieriger wahrgenommen als das Item, bei dem die geometrische Figur eines Trapezes erkannt werden sollte.

Direkt im Anschluss an die Manipulation der Einschätzung des eigenen Allgemeinen Wissens durch den im vorigen Abschnitt beschriebenen Fragebogen bearbeiteten die Teilnehmenden den Wissenstest des I-S-T 2000 R. Anhand dieser Daten wurde der ursächliche Zusammenhang zwischen der Einschätzung des eigenen Wissens und der Testleistung überprüft.

4.2.2 Analysen

Der für die Analysen verwendete Datensatz enthielt die Daten von 152 Personen (71.7% Frauen), deren Selbsteinschätzung negativ beeinflusst werden sollte und von 164 Personen (73.2% Frauen), deren Selbsteinschätzung erhöht werden sollte. Der Erfolg der experimentellen Manipulation der Selbsteinschätzung durch den Fragebogen und die Auswirkungen der Manipulation auf die Leistung in dem bearbeiteten Wissenstest wurden anhand eines Verfahrens geprüft, das von Sörbom (1978) als eine Alternative zur Kovarianzanalyse vorgeschlagen wurde. Das Verfahren beinhaltet die Erstellung von Strukturgleichungsmodellen, welche für mehrere Gruppen simultan getestet werden und bietet somit substanzielle Vorteile gegenüber der Kovarianzanalyse. So ermöglicht es die Gleichsetzung von Modellparametern, beispielsweise Regressionskoeffizienten, Erwartungswerten oder Achsenabschnitten, für verschiedene Gruppen. Durch den Vergleich mehrerer Strukturgleichungsmodelle, die unterschiedliche Restriktionen für bestimmte Parameter enthalten, können diese Restriktionen auf Haltbarkeit überprüft werden. Da außerdem die Verwendung von Strukturgleichungsmodellen die Zerlegung gemessener Werte

in einen Messfehler und einen oder mehrere Faktorwerte einschließt, können hiermit für Gruppenvergleiche die Erwartungswerte latenter Variablen anstelle der Mittelwerte manifester Variablen herangezogen werden. Konfidenzintervalle der einzelnen Parameterschätzungen erlauben auch hier die inferenzstatistische Absicherung.

Für das beschriebene Verfahren nach Sörbom (1978) wurden die Teilnehmenden nach dem Geschlecht und nach der Richtung der experimentellen Manipulation klassifiziert. Für beide Klassifizierungen wurde jeweils ein Strukturgleichungsmodell für zwei Gruppen (Frauen und Männer bzw. positiv und negativ manipulierte Selbsteinschätzung) getestet, das die beiden latenten Variablen Selbsteinschätzung und Wissen beinhaltete, wobei die Selbsteinschätzung die exogene und das Wissen die endogene Variable bildete. Die Selbsteinschätzung wurde durch drei Parcels erfasst, die jeweils aus einer zufälligen Auswahl von drei der neun Angaben der Personen zu den eingeschätzten Prozenträngen des eigenen Wissens berechnet wurden. Vier Parcels, die jeweils aus der Anzahl der richtigen Antworten von 21 zufällig ausgewählten Items der 84 Items des I-S-T 2000 R Wissenstests bestanden, wurden als Indikatoren des Wissens verwendet.

Die von Sörbom (1978) beschriebene Analyse beinhaltet die Testung von Strukturgleichungsmodellen, die sich in ihren Restriktionen unterscheiden. Für die vorliegende Arbeit wurden zwei Modelle erstellt: Im Modell des ersten Schrittes wurden sämtliche Regressionsgewichte (inklusive der Regression von Wissen auf Selbsteinschätzung) und die Achsenabschnitte aller manifesten Variablen für beide Gruppen (Geschlechter bzw. Experimentelle Bedingungen) gleichgesetzt. Die Varianzen, der Erwartungswert der exogenen und der Achsenabschnitt der endogenen Variable konnten zwischen den Gruppen variieren, wobei die beiden letzteren in einer zufällig ausgewählten Gruppe auf 0 fixiert und in der anderen Gruppe frei geschätzt wurden. Die Gruppenunterschiede im Wissen konnten somit direkt aus dem Achsenabschnitt der

entsprechenden latenten Variable abgelesen werden. Die einzige Änderung im Modell des zweiten Schrittes bestand in der Fixierung des Achsenabschnittes der latenten Variablen des Wissens auf 0 in beiden Gruppen. Das zweite Modell beinhaltete also die Annahme, dass das Wissen von Frauen und Männern beziehungsweise von den Teilnehmenden in beiden experimentellen Bedingungen gleich ist. In Abbildung 2 ist das Strukturgleichungsmodell mit den Restriktionen der beiden latenten Variablen dargestellt. Die Modelle des ersten und zweiten Schrittes wurden jeweils miteinander verglichen. Da das zweite Modell eine restriktivere Variante des ersten Modells war, konnte auch die Differenz der Modellpassungen auf Signifikanz geprüft werden. Alle Modelltestungen wurden mit drei verschiedenen Parcelgruppen wiederholt.

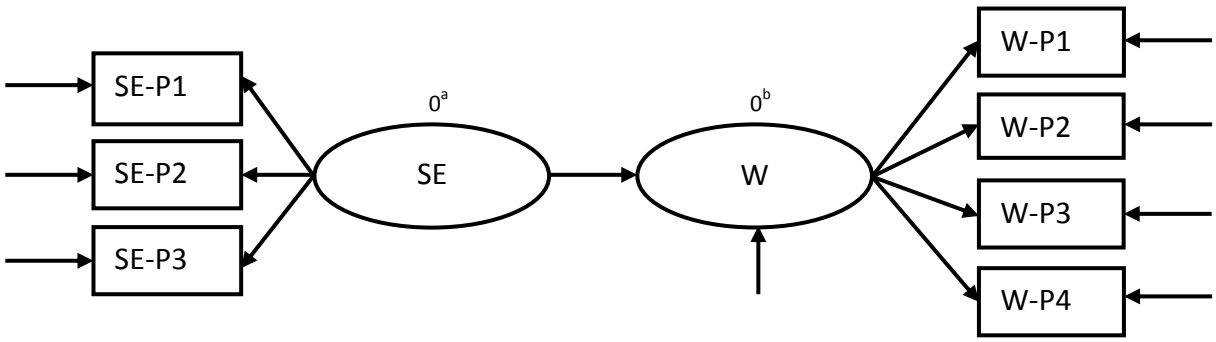

Abbildung 2. Strukturgleichungsmodell zur Erklärung von Wissen durch Selbsteinschätzung. SE = Selbsteinschätzung; W = Wissen; SE-P = Selbsteinschätzung Parcel; W-P = Wissen Parcel.

[a]Erwartungswert in Schritt 1 und in Schritt 2 in einer zufällig ausgewählten Gruppe fixiert, in der anderen Gruppe frei geschätzt.

[b]Achsenabschnitt in Schritt 1 in einer zufällig ausgewählten Gruppe fixiert, in der anderen Gruppe frei geschätzt. In Schritt 2 in beiden Gruppen fixiert.

Zuletzt wurde das Strukturgleichungsmodell für vier Gruppen getestet, wobei die Gruppen durch das Geschlecht und die beiden experimentellen Bedingungen bestimmt

waren. Der Erwartungswert der Selbsteinschätzung und der Achsenabschnitt des Wissens wurden hierbei in einer zufällig ausgewählten Gruppe auf 0 fixiert und in allen anderen Gruppen frei geschätzt. Dies ermöglichte die Vergleiche sowohl der Selbsteinschätzung als auch des Wissens zwischen allen vier Gruppen in einem Modell.

Die Überprüfungen der multivariaten Normalverteilung fielen in den drei Datensätzen, die sich durch die Parcelbildung ergaben, unterschiedlich aus. So lagen die critical ratios von Mardia‘s Koeffizient für multivariate Kurtosis für Frauen zwischen 1.36 und 2.37, für Männer zwischen 0.43 und 1.78. Bei Manipulation durch leichte Items reichten sie von 0.69 bis 4.82 und bei Manipulation durch schwierige Items von 0.13 bis 1.00. Die multivariate Normalverteilung der manifesten Variablen, welche Voraussetzung für die Verwendung des Maximum Likelihood Schätzverfahrens ist, war somit nicht erfüllt. Nach Lei und Wu (2012) tendiert der Maximum Likelihood Schätzer bei Verletzung der multivariaten Normalverteilung zu geringen Verzerrungen der Schätzungen, jedoch zu Überschätzung von χ^2-Werten und Unterschätzung von Standardfehlern. Aus diesem Grund wurden die mean-adjusted version des Maximum Likelihood Schätzers (MLM) und die von Satorra und Bentler (zitiert nach Muthén, 1998-2004) vorgeschlagene Korrektur der entsprechenden χ^2-Statistik verwendet. Die Robustheit des MLM-Schätzers gegenüber Verletzungen der Annahme der multivariaten Normalverteilung wurde in der Vergangenheit in Simulationsstudien wiederholt nachgewiesen (z.B. Asparouhov, 2005; Lei & Wu, 2012). Sämtliche Analysen wurden mit Mplus in der Version 7.1 durchgeführt. Neben der χ^2-Statistik wurden, den Empfehlungen von Beauducel und Wittmann (2005) entsprechend, der Comparative Fit Indes (CFI), der Root Mean Square Error of Approximation (RMSEA) und das Standardized Root Mean Square Residual (SRMR) als Maße der Modellpassung verwendet.

4.3 Ergebnisse

Im Folgenden werden die Strukturgleichungsmodelle mit den im vorigen Abschnitt beschriebenen Restriktionen und die zugehörigen Fit-Maße beschrieben. Zunächst wird das Modell behandelt, bei welchem die Gruppenzugehörigkeit durch das Geschlecht definiert war. Darauffolgend wird auf das Modell eingegangen, bei dem die Teilnehmenden nach der experimentellen Bedingung klassifiziert wurden. Zum Abschluss erfolgt die Darstellung des 4-Gruppen-Modells. Es werden hier lediglich die zentralen Ergebnisse der Strukturgleichungsmodelle beschrieben, bei denen die manifesten Variablen durch die erste Parcelgruppe gebildet wurden. In Anhang B sind die Ergebnisse für die erste Parcelgruppe ausführlich dargestellt. Die Anhänge C und D enthalten die Ergebnisse der zweiten und dritten Parcelgruppe.

4.3.1 Unterschiede in Selbsteinschätzung und Wissen zwischen den Geschlechtern

Wie in Kapitel 4.2.2 beschrieben, wurden im ersten Schritt die Erwartungswerte der Selbsteinschätzung und die Achsenabschnitte des Wissens bei Frauen und Männern nicht gleichgesetzt. Männer wurden als Referenzgruppe gewählt, bei der beide Parameter auf 0 fixiert waren. Für Frauen wurden die Parameter frei geschätzt. Der Fit des Modells des ersten Schrittes war mit CFI = .95 und RMSEA = .08 ausreichend. Für Frauen wurde der Achsenabschnitt des Wissens auf -0.91 und der Erwartungswert der Selbsteinschätzung auf -0.99 geschätzt.

Im Modell des zweiten Schrittes wurde die Restriktion des gleichen Achsenabschnittes für Frauen und Männer im Faktor Wissen hinzugefügt. Der Erwartungswert des Faktors Selbsteinschätzung wurde bei Frauen auf -1.09 geschätzt. Die Passung fiel hier mit CFI = .92

und RMSEA = .10 nicht zufriedenstellend aus. Der χ^2-Test auf die Differenz der Passung beider Modelle[3] ergab ein auf dem 1‰-Niveau signifikantes Ergebnis.

Tabelle 5

Auszug aus standardisierten Parameterschätzungen und Model-Fit der 2-Gruppen-Modelle, mit Klassifizierung nach Geschlecht (Parcelgruppe 1)

	Modell 1		Modell 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert SE	-0.99	0[a]	-1.09	0[a]
[95%-KI]	[-1.27; -0.70]		[-1.38; -0.79]	
Achsenabschnitt W	-0.91	0[a]	0[a]	0[a]
[95%-KI]	[-1.27; -0.54]			
Model-Fit				
χ^2_{df}	74.27_{37}		99.64_{38}	
p	< .001		< .001	
SCF	1.00		1.00	
CFI	.95		.92	
SRMR	.06		.12	
RMSEA	.08		.10	
[90%-KI]	[.05; .11]		[.08; .13]	
χ^2-Diff-Test: χ^2_{df}; *p*	21.02_1; < .001			

Anmerkungen. SE = Selbsteinschätzung; KI = Konfidenzintervall; W = Wissen; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Modell 1 vs. Modell 2.

[a]Parameter wurde vorab fixiert.

[3]Für diesen Test wurden die Satorra-Bentler-skalierten χ^2-Werte beider Modelle berechnet. Eine Beschreibung der Vorgehensweise findet sich bei Muthén (1998-2004) und auf der Mplus-Website unter http://www.statmodel.com/chidiff.shtml.

Eine Auswahl der standardisierten Parameterschätzungen und die Model-Fit-Maße der beiden 2-Gruppen-Modelle, bei denen die Klassifikation nach dem Geschlecht erfolgte, sind in Tabelle 5 dargestellt. Ausführliche Ergebnisse der Analysen finden sich in Tabelle B1.

4.3.2 Unterschiede in Selbsteinschätzung und Wissen zwischen den Gruppen der experimentellen Bedingungen

Für die 2-Gruppen-Modelle, bei denen die Klassifikation nach den experimentellen Bedingungen erfolgte, wurde als Referenzgruppe die Gruppe ausgewählt, der schwierige Beispielitems vorgelegt worden waren und deren Selbsteinschätzung damit reduziert werden sollte. Für diese Gruppe wurden der Erwartungswert der Selbsteinschätzung und der Achsenabschnitt des Wissens im ersten Modell auf 0 gesetzt. Das Modell des ersten Schrittes zeigte mit CFI = .99 und RMSEA = .05 eine gute Passung. Der Erwartungswert der Selbsteinschätzung fiel in der Gruppe, denen leichte Beispielitems präsentiert worden waren, mit 0.39 höher aus als in der anderen Gruppe. Das Wissen wurde hier jedoch mit einem Achsenabschnitt von -0.35 geringer eingeschätzt.

Im zweiten Schritt, bei dem das Wissen der Probanden beider Gruppen gleichgesetzt war, wurde der Erwartungswert des Faktors Selbsteinschätzung in der Gruppe der Teilnehmenden, deren Selbsteinschätzung erhöht werden sollte, auf 0.34 geschätzt. Die Passung des Modells war mit CFI = .98 und RMSEA = .06 etwas geringer als im ersten Schritt. Der χ^2-Test auf die Differenz der Passung beider Modelle[2] ergab hier χ^2_1=10.70 (p = .001). Eine Auswahl der standardisierten Parameterschätzungen und die Model-Fit-Maße beider Modelle sind in Tabelle 6 dargestellt. Die ausführlichen Ergebnisse für die Parcelgruppe 1 finden sich in Tabelle B2.

Tabelle 6

Auszug aus standardisierten Parameterschätzungen und Model-Fit der 2-Gruppen-Modelle, mit Klassifizierung nach experimenteller Bedingung (Parcelgruppe 1)

	Modell 1		Modell 2	
	SE hoch	SE niedrig	SE hoch	SE niedrig
Erwartungswert SE	0.39	0[a]	0.34	0[a]
[95%-KI]	[0.12; 0.66]		[0.08; 0.61]	
Achsenabschnitt W	-0.35	0[a]	0[a]	0[a]
[95%-KI]	[-0.58; -0.11]			
Model-Fit				
χ^2_{df}	51.49_{37}		60.46_{38}	
p	.057		.012	
SCF	0.98		0.98	
CFI	.99		.98	
SRMR	.04		.06	
RMSEA	.05		.06	
[90%-KI]	[.00; .08]		[.03; .09]	
χ^2-Diff-Test: χ^2_{df}; *p*	10.70_1; .001			

Anmerkungen. SE = Selbsteinschätzung; KI = Konfidenzintervall; W = Wissen; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Modell 1 vs. Modell 2.

[a]Parameter wurde vorab fixiert.

4.3.3 Vier-Gruppen-Modell

Die Gruppe der Männer, deren Selbsteinschätzung verringert werden sollte, wurde in dem letzten Modell als Gruppe für die Fixierungen ausgewählt. In dieser Gruppe wurden der Erwartungswert der Selbsteinschätzung und der Achsenabschnitts des Wissens auf 0 gesetzt. In den anderen Gruppen wurden beide Parameter frei geschätzt.

Am höchsten fiel die Selbsteinschätzung mit 0.87 in der Gruppe der Männer aus, die durch leichte Beispielitems manipuliert wurden. In der Gruppe der Frauen, deren

Selbsteinschätzung erhöht wurde, betrug die Schätzung -0.47. In der Gruppe der Frauen mit verringerter Selbsteinschätzung wurde der Erwartungswert auf -0.82 geschätzt. Die Schätzungen des Achsenabschnitts der latenten Variable Wissen betrugen -1.39 für Frauen, deren Selbsteinschätzung erhöht werden sollte, -1.03 für Frauen, deren Selbsteinschätzung verringert werden sollte und -0.42 für Männer, deren Selbsteinschätzung erhöht werden sollte. Das Modell zeigte mit CFI = .94 und RMSEA = .08 eine geringe Passung. Tabelle 7 enthält einen Auszug der Ergebnisse des 4-Gruppen-Modells. Ausführliche Ergebnisse der Analysen sind in Tabelle B3 dargestellt.

Tabelle 7

Auszug aus standardisierten Parameterschätzungen und Model-Fit des 4-Gruppen-Modells, mit Klassifizierung nach Geschlecht und experimenteller Bedingung (Parcelgruppe 1)

	Frauen, SE hoch	Frauen, SE niedrig	Männer, SE hoch	Männer, SE niedrig
Erwartungswert SE	-0.47	-0.82	0.87	0[a]
[95%-KI]	[-0.87; -0.07]	[-1.24; -0.40]	[0.23; 1.50]	
Achsenabschnitt W	-1.39	-1.03	-0.42	0[a]
[95%-KI]	[-1.88; -0.91]	[-1.47; -0.59]	[-0.83; -0.02]	
Model-Fit				
χ^2_{df}		129.38_{85}		
p		.001		
CFI		.94		
SRMR		.10		
RMSEA		.08		
[90%-KI]		[.05; .11]		

Anmerkungen. SE = Selbsteinschätzung; W = Wissen; KI = Konfidenzintervall.

[a]Parameter wurde vorab fixiert.

4.4 Diskussion

Die unterschiedlichen Leistungen von Frauen und Männern in Wissenstests wurden auch in der vorliegenden Untersuchung bestätigt. Frauen und Männer zeigten hier jedoch auch signifikante Altersunterschiede. Ein Levene-Test auf Varianzgleichheit des Alters bei Frauen und Männern fiel hochsignifikant aus ($f_{1,313}$ = 14.43, p < .001). Der anschließende t-Test zeigte ebenfalls ein signifikantes Ergebnis ($t_{120.10}$ = 2.23, p = .028). Die Annahme, dass Wissen mit zunehmendem Alter ansteigt, ist kongruent mit der in Kapitel 2.3.3 beschriebenen Investmenttheorie. Daher wurde zusätzlich eine partielle Korrelation zwischen Geschlecht und Wissen berechnet, wobei die Variable Alter kontrolliert wurde. Diese Korrelation betrug r = -.42 (p < .001). Das negative Vorzeichen weist hier auf einen höheren Mittelwert des Wissens bei Männern hin. Damit zeigt sich, dass die vorgefundenen Geschlechterdifferenzen des Wissens nicht auf Altersunterschieden von Frauen und Männern beruhen. Es ist hier ein weiteres Mal das Phänomen der Geschlechterdifferenzen in Wissenstests aufgetreten, das der zentrale Gegenstand der vorliegenden Arbeit ist. Auf der Ebene latenter Variablen zeigte sich das gleiche Ergebnis: In dem Strukturgleichungsmodell des ersten Schrittes war der Achsenabschnitt des Faktors Wissen bei Männern auf 0 fixiert und wurde bei Frauen auf -0.91 geschätzt. Das 95%-Konfidenzintervall umfasste nicht den Wert 0. Das Modell des zweiten Schrittes, bei welchem der Parameter bei beiden Gruppen auf 0 fixiert war, zeigte eine hoch signifikant schlechtere Passung.

Wie aus Tabelle 7 hervorgeht, war die Manipulation der Selbsteinschätzung durch die unterschiedlichen experimentellen Bedingungen erfolgreich. Die geschätzten Erwartungswerte des Faktors Selbsteinschätzung lassen sich bei dem 4-Gruppen-Modell in die folgende Rangordnung bringen: Frauen/schwierige Items < Frauen/leichte Items < Männer/schwierige Items < Männer/leichte Items. Mit Ausnahme der Gruppen Frauen/schwierige Items und Frauen/leichte Items überlagerten sich die 95%-

Konfidenzintervalle der Gruppen nicht. Sowohl unter den Teilnehmenden, denen schwierige Beispielitems vorgelegt worden waren als auch unter den Teilnehmenden, die leichte Beispielitems bearbeitet hatten, zeigte sich jeweils ein signifikanter Unterschied zwischen Frauen und Männern. Für die Männer zeigten die experimentellen Bedingungen ebenfalls einen signifikanten Effekt auf die Selbsteinschätzung. Bei den Frauen war dies nicht der Fall. Bei den in Kapitel 4.3.2 beschriebenen Modellen, bei denen das Geschlecht nicht zur Gruppierung der Personen verwendet wurde, konnte der gewünschte Effekt der experimentellen Bedingungen auf die Selbsteinschätzung inferenzstatistisch abgesichert werden. Das Geschlecht hatte jedoch einen stärkeren Effekt als die Manipulation.

Wie in den Tabellen der Anhänge B, C und D zu erkennen ist, wiesen alle Strukturgleichungsmodelle für sämtliche Gruppen einen positiven β-Koeffizienten für die Regression von Wissen auf die Selbsteinschätzung auf. Die β-Koeffizienten der ersten Parcelgruppe reichten insgesamt von 0.23 bis 0.51. Insgesamt wurde somit die Vermutung, dass die Selbsteinschätzung in einem positiven Zusammenhang mit der Leistung im Wissenstest steht, bestätigt.

Die Ergebnisse der beiden 2-Gruppen-Modelle, bei denen die Gruppierung nach den experimentellen Bedingungen erfolgte, widersprechen jedoch der in Abschnitt 4.1 hergeleiteten Hypothese. Der erwartete Effekt der Manipulation der Selbsteinschätzung auf die Leistungen im Wissenstest wurde nicht bestätigt. Obwohl die Manipulation der Selbsteinschätzung erfolgreich verlief und ein positiver Zusammenhang zwischen Selbsteinschätzung und Leistung bestätigt wurde, zeigten die Personen mit erhöhter Selbsteinschätzung durchschnittlich keine besseren Leistungen als die Personen mit verringerter Selbsteinschätzung. Die Ergebnisse der Strukturgleichungsmodelle, bei denen die Gruppierung auf den experimentellen Bedingungen beruhte, weisen vielmehr in die umgekehrte Richtung: Im ersten Modell wurde der Achsenabschnitt des Faktors Wissen in

der Gruppe der Personen, deren Selbsteinschätzung verringert wurde, auf 0 fixiert und in der Gruppe der Personen, deren Selbsteinschätzung erhöht wurde, auf -0.35 geschätzt. Das 95%-Konfidenzintervall umfasste nicht den Wert 0. Das zweite Modell, bei welchem der Achsenabschnitt für beiden Gruppen gleichgesetzt war, zeigte eine schlechtere Passung. Ähnliche Ergebnisse zeigte das 4-Gruppen-Modell. So lassen sich die Achsenabschnitte der vier Gruppen in folgende Rangordnung bringen: Frauen/leichte Items < Frauen/schwierige Items < Männer/leichte Items < Männer/schwierige Items. Unter Männern zeigte die Gruppe mit erhöhter Selbsteinschätzung signifikant geringere Leistungen als die Gruppe mit verringerter Selbsteinschätzung. Bei den Frauen trat der Effekt deskriptiv ebenfalls auf, ließ sich allerdings nicht inferenzstatistisch absichern. Jedoch zeigten unter beiden experimentellen Bedingungen Männer signifikant bessere Leistungen als Frauen.

Insgesamt wurden die für diese Untersuchung getroffenen Vorhersagen somit teilweise bestätigt. Die in der bisherigen Forschung häufig berichteten Geschlechterdifferenzen in Wissenstest wurden repliziert. Des Weiteren ließ sich eine erfolgreiche Manipulation der Selbsteinschätzung durch die Vorlage von schwierigen bzw. leichten Beispielitems nachweisen. Sämtliche Strukturgleichungsmodelle bestätigten außerdem den positiven Zusammenhang zwischen Selbsteinschätzung und Wissen. Das in Abschnitt 4.1 beschriebene Ziel, einen Beleg für den positiven Zusammenhang zwischen der Einschätzung des eigenen Wissens und der gemessenen Leistung in einem Wissenstest zu liefern, wurde erreicht. Die zentrale Hypothese der Untersuchung – die Erhöhung beziehungsweise Verringerung der Leistung im Wissenstest durch Erhöhung beziehungsweise Verringerung der Selbsteinschätzung – wurde jedoch nicht bestätigt. Es zeigte sich ein gegenteiliger Effekt: Teilnehmende, denen schwierige Beispielitems vorgelegt worden waren, brachten durchschnittlich bessere Leistungen im Wissenstest als die Teilnehmenden, denen leichte Beispielitems vorgelegt worden waren, wobei auch hier der Effekt des Geschlechts deutlicher

war. Wie den Anhängen C und D zu entnehmen ist, bestätigten sich die Ergebnisse bei allen drei Parcelgruppen.

Wie bereits in Kapitel 4.1 erwähnt, basierte die Herleitung der Hypothese, dass Geschlechterdifferenzen in Wissenstests durch Geschlechterdifferenzen der Selbsteinschätzung des Wissens verursacht werden, teilweise auf Annahmen, für die Indizien vorliegen, die jedoch noch der empirischen Überprüfung bedürfen. So wurde in der bisherigen Forschung über Geschlechterdifferenzen der Selbsteinschätzung kognitiver Fähigkeiten berichtet, jedoch nicht explizit in Bezug auf das Allgemeine Wissen. In der vorliegenden Untersuchung sollte jedoch das eigene Allgemeine Wissen eingeschätzt werden und es zeigten sich auch hierfür Ergebnisse, die der allgemeinen Tendenz zur höheren Selbsteinschätzung von Männern entsprechen. Eine weitere Annahme des beschriebenen Wirkmechanismus ist, dass durch die angstbezogenen Emotionen kognitive Ressourcen beansprucht werden. Hierfür liegen zahlreiche Belege vor. Jedoch muss die Frage gestellt werden, inwieweit die Beanspruchung kognitiver Ressourcen, konkret des Arbeitsgedächtnisses, zu Leistungseinbußen bei Wissenstests führt. Wie in Kapitel 4.1 beschrieben, beinhalteten die Studien zur Beanspruchung kognitiver Ressourcen keine Wissenstests sondern Aufgaben aus der Mathematik und das Erlernen der simulierten Tätigkeit von Fluglotsen. Inwieweit bei Wissenstests Leistungseinbußen durch eine zusätzliche Beanspruchung des Arbeitsgedächtnisses entstehen, ist ungewiss. Möglicherweise beansprucht der Abruf von deklarativem semantischem Wissen das Arbeitsgedächtnis in einem deutlich geringeren Ausmaß.

Sofern das Arbeitsgedächtnis bei Wissenstests kaum beansprucht würde, läge die Annahme nahe, dass die Leistung in Wissenstests weitgehend unabhängig von der Manipulation der Selbsteinschätzung ist. In der vorliegenden Untersuchung zeigte sich jedoch keine Unabhängigkeit zwischen Selbsteinschätzung und Leistung, sondern eine

Abhängigkeit in entgegengesetzter Richtung zur Hypothese: Personen mit verringerter Selbsteinschätzung zeigten bessere Leistungen als Personen mit erhöhter Selbsteinschätzung. Eine mögliche Erklärung hierfür wäre eine zu kurz anhaltende wirksame Manipulation der Selbsteinschätzung. Das experimentelle Design umfasste lediglich die Messung zu einem einzelnen Zeitpunkt – unmittelbar vor der Bearbeitung des Wissenstests – und wurde nicht im Anschluss an den Wissenstest wiederholt. Im Nachhinein stellt sich daher die Frage, ob die Manipulation möglicherweise nur über einen kurzen Zeitraum erfolgreich war und nach Beginn der Bearbeitung des Wissenstests nachließ. Da die erwarteten Auswirkungen der experimentellen Bedingungen auf die Testleistungen nicht nur ausblieben, sondern sogar gegenteilig ausfielen, wäre es beispielsweise denkbar, dass die Teilnehmenden nach Bearbeitung einiger weniger Items einschätzen konnten, wie erfolgreich sie hierbei waren. So könnten während der Testbearbeitung Korrekturen der Selbsteinschätzung vorgenommen worden sein, obwohl keine Informationen über die Schwierigkeit der einzelnen Items des Tests vorgegeben wurden. Bei diesen Korrekturen dürfte es sich bei Personen, deren Selbsteinschätzung vorab verringert worden war, um Verbesserungen und bei Personen, deren Selbsteinschätzung vorab erhöht worden war, um Verringerungen handeln. In mehreren Testungen wurde protokolliert, dass einzelne Teilnehmende während der Bearbeitung des Wissenstests fragten, ob sie im Nachhinein noch Änderungen an den eingetragenen Prozenträngen vornehmen dürften, was von der jeweiligen Leitung selbstverständlich verneint wurde. Für folgende Studien empfiehlt sich somit die wiederholte Kontrolle der Selbsteinschätzung, um den Erfolg der Manipulation während der gesamten Testbearbeitung zu überprüfen. So ist es auch auf Basis der vorliegenden Daten durchaus denkbar, dass die in Kapitel 4.1 hergeleiteten Hypothesen wahr sind, dass sie in der hier beschriebenen Studie jedoch nicht nachgewiesen werden konnten, da der Effekt der Manipulation auf die Selbsteinschätzung im Laufe der Testdurchführung nachließ, eventuell sogar umgekehrt

wurde. Letzteres wäre auch eine Erklärung dafür, dass in allen Strukturgleichungsmodellen ein positiver Zusammenhang der Selbsteinschätzung mit Wissen vorgefunden wurde, obwohl die Teilnehmenden, deren Selbsteinschätzung negativ manipuliert wurde, die durchschnittlich besseren Leistungen zeigten.

Auch wenn die manipulierte Selbsteinschätzung des Wissens im Hinblick auf die Leistungen der Teilnehmenden keine Wirkung gezeigt hat, ist nicht auszuschließen, dass dennoch Geschlechterdifferenzen in der Selbsteinschätzung von zentraler Bedeutung für Geschlechterdifferenzen in Wissenstests sind. Die hier vorgenommene Manipulation der Selbsteinschätzung konnte sich ausschließlich in der direkt darauf folgenden Testsituation auswirken. Langfristige Auswirkungen auf den Erwerb des Wissens, beispielsweise durch Motivation oder Attribution (siehe beispielsweise Dweck & Leggett, 1988), konnten hier nicht überprüft werden, da dieses Experiment keine Lernphase enthielt, in der neues Wissen erworben wurde. Es soll hier noch einmal an die Gegenüberstellung der beiden Gruppen von potentiellen Erklärungen für Geschlechterdifferenzen in Wissenstests erinnert werden, die zu Beginn des Kapitels 3.2 erläutert wurde: Während die eine Gruppe Erklärungen für gemessene Unterschiede des Wissens beinhaltete, enthielt die andere Gruppe Erklärungen für bestehende Unterschiede. In der hier vorgestellten Untersuchung wurde Selbsteinschätzung als Ursache für eventuelle Verzerrungen bei der Messung des Wissens zuungunsten von Frauen geprüft. Bei langfristigen Auswirkungen der Selbsteinschätzung würde diese nicht mehr als Erklärung für gemessene sondern für reale Unterschiede des Wissens zwischen Frauen und Männern dienen. Es wäre also ein anderer Sachverhalt, der hiermit eventuell begründet werden könnte.

Eine weitere Erklärung, unter der (a) die in allen Modellen positiven Koeffizienten für die Regression von Wissen auf Selbsteinschätzung und (b) das Ausbleiben der Wirkung der experimentellen Bedingungen auf die Testleistungen vereinbar sind, wäre eine Kausalität in

umgekehrter Richtung: Eventuell ist geringere Selbsteinschätzung des Wissens durch tatsächlich geringeres Wissen begründet.

Es bleiben somit verschiedene Erklärungen für die vorgefundenen Ergebnisse denkbar. Die Hypothese, dass Geschlechterdifferenzen in Wissenstests durch Geschlechterdifferenzen in der Einschätzung des eigenen Wissens bedingt sind, muss auf Basis der vorliegenden Daten vorläufig verworfen werden.

.

Interessen

In diesem Kapitel wird eine Gruppe von drei Untersuchungen beschrieben, durch die überprüft wurde, inwieweit Geschlechterdifferenzen in Interessen von ursächlicher Bedeutung für Geschlechterdifferenzen in Tests des Allgemeinen Wissens sein könnten. Zunächst wird erläutert, warum Interessenunterschiede zwischen Frauen und Männern eine potentielle Erklärung für die Leistungsunterschiede zwischen den Geschlechtern in Wissenstests darstellen (Kapitel 5.1). Im Anschluss werden die drei Studien beschrieben. In der ersten Untersuchung (Kapitel 5.2) wurde der Frage nachgegangen, welche Geschlechterdifferenzen in Interessen bestehen. Nach der Klassifikation von Themen in Interessengebiete von Frauen (IvF), Interessengebiete von Männern (IvM) und Neutrale Interessengebiete (NI) wurden zwei deutschsprachige Wissenstests dahingehend überprüft, ob mit ihnen in gleichem Ausmaß das Wissen zu Themen der Kategorien IvF und IvM erfasst wird. Die zweite Untersuchung (Kapitel 5.3) umfasste die Entwicklung eines neuen Wissenstests, der ausschließlich Items zu Themen enthielt, für die Frauen sich durchschnittlich stärker als Männer interessieren. Nach der Erstellung eines Itempools erfolgte eine Datenerhebung. Auf Basis der Ergebnisse wurde eine Itemselektion durchgeführt. Im Anschluss erfolgte die Überprüfung einiger psychometrischer Kennwerte des Tests, und die Leistungen von Frauen und Männern wurden verglichen. In der dritten Untersuchung (Kapitel 5.4) bearbeiteten die Probandinnen und Probanden den neuen Test zusammen mit dem Wissenstest des I-S-T 2000 R (Liepman et al., 2007). Im Mittelpunkt standen hier die faktorielle Struktur der Testbatterie und die Leistungsunterschiede zwischen Frauen und Männern, die sich hierbei zeigten.

5.1 Theorie

Wie bereits in Kapitel 3.1 beschrieben, variieren verschiedene Themengebiete in Wissenstests im Ausmaß und in der Richtung der Leistungsunterschiede zwischen Frauen und Männern. Ein möglicher Grund hierfür könnten Differenzen der Interessen von Frauen und Männer sein. Die einseitige Erfassung des Wissens in Themen, für die Männer sich durchschnittlich stärker als Frauen interessieren, könnte außerdem zu den berichteten Geschlechterdifferenzen in den Gesamtscores zugunsten von Männern führen.

Im Folgenden wird zunächst der Begriff des Interesses aus Sicht der Differenziellen Psychologie erläutert. Im Anschluss daran wird ein Überblick über den Forschungsstand zu Interessenunterschieden zwischen Frauen und Männern gegeben. Hieraus leitet sich die Hypothese ab, dass eine fehlende Balance zwischen Items zu Themen, für die Frauen bzw. Männer sich durchschnittlich stärker interessieren, eine potentielle Erklärung für Geschlechterdifferenzen in Wissenstests darstellt. Sofern Tests des Allgemeinen Wissens hauptsächlich Items zu Interessengebieten von Männern beinhalten, würden hiermit die höheren Mittelwerte in den Testscores von Männern begünstigt.

5.1.1 Definition des Begriffs Interesse

Der Begriff *Interesse* wird in der Psychologie im Allgemeinen aus zwei verschiedenen Perspektiven betrachtet. Zum einen kann Interesse situationsspezifisch aufgefasst werden, wofür auch der Begriff *situational interest* verwendet wird. Schraw und Lehman (2001) heben als zentrale Merkmale das spontane Entstehen und Abklingen und die örtliche Spezifität des situational interest hervor. Diese Perspektive wird häufig in der Pädagogischen Psychologie eingenommen. Ein Beispiel für Forschungsinhalte ist die Unterrichtsgestaltung zur Maximierung des Interesses von Schülerinnen und Schülern (z.B. Bergin, 1999). In der

Differenziellen Psychologie dominiert hingegen die Auffassung von Interesse als stabile Eigenschaft, wofür auch der Begriff *dispositional interest* verwendet wird.

Die Betrachtung von Interesse als stabiler Eigenschaft liegt auch der Definition von Rounds und Su (2014) zugrunde, welche in der vorliegenden Arbeit übernommen wurde: "We define interests as traitlike preferences for activities, contexts in which activities occur, or outcomes associated with preferred activities that motivate goal-oriented behaviors and orient individuals toward certain environments" (S. 98). Als zentrale Elemente der Definition heben Rounds und Su (a) die zeitliche Stabilität, (b) die Gegenstandsbezogenheit und (c) die motivationalen Auswirkungen von Interessen hervor.

Die zeitliche Stabilität von beruflichen Interessen wurde empirisch von Low, Yoon, Roberts und Rounds (2005) bestätigt. In einer Meta-Analyse wurde die zeitliche Stabilität von beruflichen Interessen überprüft und mit der Stabilität von Persönlichkeitsmerkmalen verglichen. Stabilität wurde hierbei in Form von Rangkorrelationen erfasst. Low et al. stellten für Interessen Korrelationen zwischen .55 für 12- bis 17.9-Jährige und .70 für 22- bis 29-Jährige fest. Die Korrelationen für Persönlichkeitsmerkmale lagen zwischen .47 für 12- bis 17.9-Jährige und .62 für 30- bis 39-Jährige. In drei der vier Altersgruppen fielen die Korrelationskoeffizienten für Interessen signifikant höher aus als die Koeffizienten für Persönlichkeitsmerkmale, woraus Low et al. auf eine höhere Stabilität beruflicher Interessen im Vergleich zu Persönlichkeitsmerkmalen schlossen. Zudem wurde festgestellt, dass die Stabilität von beruflichen Interessen im Verlauf des Lebens früher das Maximum erreichte als die Stabilität von Persönlichkeitsmerkmalen.

Die Definition von Interesse im Sinne von Rounds und Su (2014) beinhaltet ferner den Bezug auf ein bestimmtes Objekt oder einen bestimmten Kontext. Interesse ist kein Merkmal, das für sich genommen hoch oder niedrig ausgeprägt ist, worin Rounds und Su eine wichtige Abgrenzung zu Persönlichkeitsmerkmalen und kognitiven Fähigkeiten sehen. Allerdings

kann das Ausmaß der Spezifität, mit der ein Kontext beschrieben wird, variieren. So stellt das bereits in Kapitel 3.2.3 beschriebene Konstrukt Typical Intellectual Engagement (Goff & Ackerman, 1992) einen sehr breiten, allgemeinen Kontext dar, welcher sich jedoch kontinuierlich weiter spezifizieren lässt bis hin zu sehr eng gefassten Interessengebieten, wie beispielsweise der "biology of monarch butterflies" (Schmidt, 2014, S. 213). Zwischen diesen Polen des Kontinuums ist das weit verbreitete RIASEC-Modell für Interessen von Holland (1959, 1997) angesiedelt, welches sechs Interessentypen umfasst. Das R steht dabei für die praktisch-technische Orientierung (Realistic), I für intellektuell-forschende Orientierung (Investigative), A für künstlerisch-sprachliche Orientierung (Artistic), S für soziale Orientierung (Social), E für unternehmerische Orientierung (Enterprising) und C für konventionelle Orientierung (Conventional) (Bergmann & Eder, 2005). Nye, Su, Rounds und Drasgow (2012) geben folgende prägnante Umschreibungen für die verschiedenen Interessentypen:

> Realistic individuals are interested in working with things, gadgets, or in the outdoors; investigative individuals are interested in science, including mathematics, physical and social sciences, and the biological and medical sciences; artistic individuals prefer creative expression, including writing and the visual and performing arts; social individuals enjoy helping people; enterprising individuals like working in leadership or persuasive roles directed toward achieving economic objectives; and conventional individuals are interested in working in well-structured environments, especially business settings. (Nye et al., 2012, S. 385)

Die verschiedenen Orientierungen sind nicht unabhängig voneinander, sondern können in einem hexagonalen Modell abgebildet werden (siehe Abbildung 3), in dem durch die Nähe der Interessentypen zueinander die Ähnlichkeit veranschaulicht wird. Nebeneinander angeordnete Typen überschneiden sich inhaltlich stärker als weiter voneinander entfernte

Typen. Wichtig ist hierbei, dass mit Interessenfragebögen, die sich an dem RIASEC-Modell orientieren, wie beispielsweise dem Allgemeinen Interessen-Struktur-Test-Revision (AIST-R) von Bergmann und Eder (2005), keine Zuordnung des Individuums zu einem einzelnen Interessentyp bezweckt wird, wie der Begriff "Typ" suggerieren mag. Vielmehr ergibt sich durch die Ausprägungen der sechs Interessentypen für jede Person ein individuelles Interessenprofil.

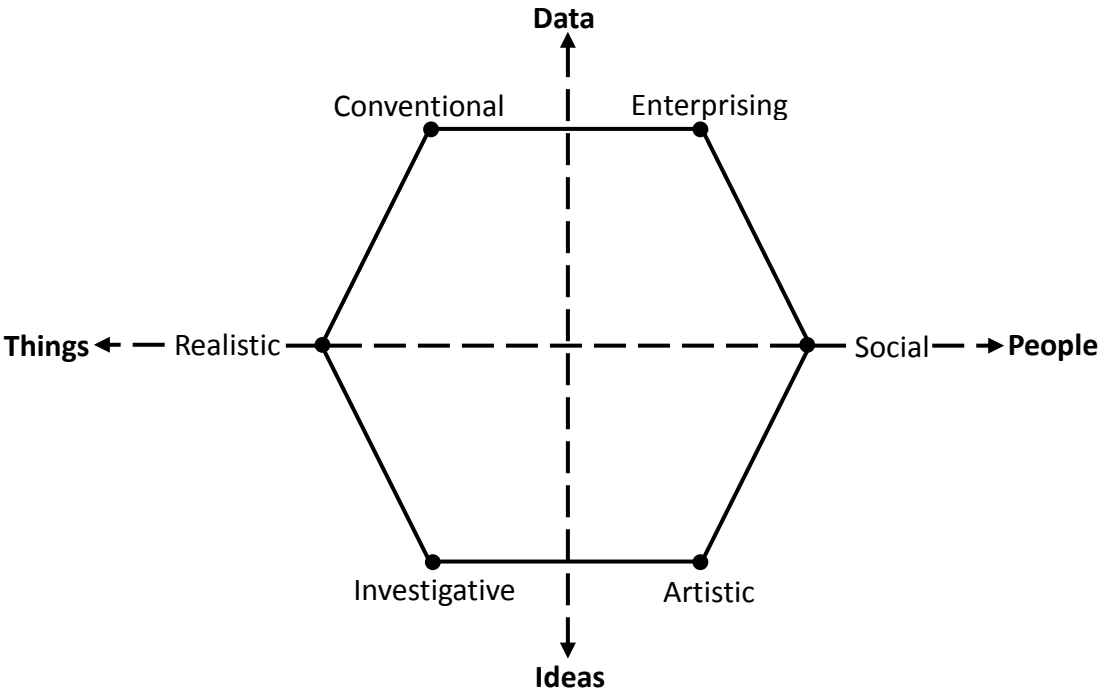

Abbildung 3. RIASEC-Modell nach Holland (1959, 1997), erweitert um Predigers (1982) People-Things und Data-Ideas Dimensionen.

Nye et al. (2012) verwiesen auf einige Studien, die das hexagonale Modell von Holland (1959, 1997) bestätigten und kamen zu dem Schluss, dass das Modell sowohl das Kriterium der Sparsamkeit erfüllt als auch empirische Unterstützung findet. Prediger (1982) schlug eine Erweiterung des Modells um zwei theoriebasierte bipolare Dimensionen vor, die ebenfalls empirisch bestätigt wurde. Hierbei handelt es sich um die people-things- und die data-ideas-Dimensionen. Wie in Abbildung 3 zu erkennen ist, werden die Pole der people-things-

Dimension durch die Interessentypen Social (*people*) und Realistic (*things*) gebildet. Die Pole der data-ideas-Dimension werden nicht durch einzelne Interessentypen repräsentiert, jedoch können vier der sechs Typen tendenziell einem der beiden Pole zugeordnet werden.

Bezüglich der motivationalen Komponente der Definition von Interesse nach Rounds und Su (2014) kann die in Kapitel 2.3.3 erläuterte Investmenttheorie von Cattell (1971/1987) herangezogen werden. Wie dort bereits erläutert, bildet dieser Theorie zufolge das Interesse an intellektuellen Tätigkeiten einen wichtigen Faktor, durch den bestimmt wird, in welchem Ausmaß eine Person in intellektuelle Tätigkeiten investiert, und in welchen Themenbereichen sie Wissen erlangt. Hierdurch bildet sich im Laufe der Zeit die kristalline Intelligenz. Interessen sind somit sowohl in quantitativer als auch in qualitativer Hinsicht für den Erwerb der kristallinen Intelligenz relevant. Wie in Kapitel 3.2.3 berichtet, wurde in den Arbeiten von Ackerman (2000a), Ackerman und Heggestad (1997) und Wilhelm et al. (2003) über positive Korrelationen zwischen Typical Intellectual Engagement und dem Ausmaß der kristallinen Intelligenz berichtet. Für die thematische Ausrichtung können spezifischere Inhaltsbereiche herangezogen werden, wie beispielsweise berufliche Orientierungen oder die sechs Interessentypen des RIASEC-Modells von Holland (1959, 1997). So fanden Nye et al. (2012) in ihrer Meta-Analyse zu Zusammenhängen zwischen Interessen und Leistungen im beruflichen beziehungsweise akademischen Kontext positive Korrelationen von r = .20 beziehungsweise r = .23[4]. Nye et al. postulierten ein Modell, demzufolge Motivation für

[4]Die Zusammenhänge fielen höher aus, wenn nicht Korrelationen zwischen Interessenscores und Leistung berechnet wurden, sondern die Kongruenz zwischen den Interessenprofilen der Person und dem beruflichen (akademischen) Kontext berücksichtigt wurde. Diese Kongruenz wurde durch einen systematischen Vergleich der Profile der Person und des Kontexts erfasst

Leistung relevant ist, da sie die Ziele, die sich eine Person setzt, sowie die Anstrengung und die Ausdauer, mit der sie ihre Ziele verfolgt, bestimmt. Interessen wiederum "should be related to performance because they affect all three aspects of the motivational process. In other words, interests are not the same as motivation but can influence motivational processes" (Nye et al., 2012, S. 386). Schmidt (2011) untersuchte Interessenunterschiede zwischen Frauen und Männern als Begründung für höhere Fähigkeiten von Männern im Vergleich zu Frauen im Bereich der Technik. Er postulierte, dass allgemeine kognitive Fähigkeiten beider Geschlechter gleich sind und dass höhere technische Fähigkeiten von Männern durch mehr Erfahrung in diesem Kontext erklärbar sind. Höhere Erfahrungen im Bereich der Technik ließen sich Schmidts Hypothese nach wiederum durch ein durchschnittlich höheres Interesse von Männern im Vergleich zu Frauen an Technik begründen. Schmidt stützte sich bei der Herleitung seiner Annahmen auf die Investmenttheorie und fand auf korrelativer Ebene Bestätigung für seine Hypothese. Auf Interessen beruhende Motivation ist somit – in Abgrenzung zu Motivation, die Gegenstand kognitiver Motivationstheorien ist, – inhaltlich spezifisch ausgerichtet (Hidi, Renninger & Krapp, 2004). Wie in Kapitel 2.3.5 erläutert, lässt sich auch die Theorie der Verarbeitungstiefe nach Craik und Lockhart (1972) mit der Investmenttheorie nach Cattell verknüpfen. Höheres Interesse an einem Thema – und damit stärkeres Investment in dieses Thema – dürften auch zu einer tieferen Verarbeitung der Inhalte und damit zu einem erfolgreicheren Abruf der entsprechenden Informationen zu einem späteren Zeitpunkt führen.

---

und bildete eine eigene Variable. Die Korrelation zwischen dieser Variablen und der beruflichen (akademischen) Leistung betrug $r = .36$ ($r = .32$).

5.1.2 Interessenunterschiede zwischen Frauen und Männern

Geschlechterdifferenzen von Interessen wurden beispielsweise von Su, Rounds und Armstrong (2009) in einer Meta-Analyse untersucht, die sich auf das bereits erwähnte RIASEC-Modell von Holland (1959, 1997) in Kombination mit den bipolaren Dimensionen von Prediger (1982) stützt. Hierbei wurden die Normierungsstichproben von insgesamt 47 Interessenfragebögen verwendet. Da nicht alle Fragebögen sechs Skalen enthielten, die zur Erfassung der sechs Interessentypen nach Holland dienten, sondern teilweise auf anderen Modellen beruhten, wurden soweit notwendig die Skalen dieser Tests sowohl von Rounds als auch von Su den Interessentypen von Holland zugeordnet. Beide Autoren erreichten hierbei eine hochgradige Übereinstimmung in ihren Urteilen (Su et al., 2009). Anschließend ermittelten die Autoren für alle sechs Interessentypen des RIASEC-Modells die Effektstärken der Geschlechterdifferenzen. Die größten Effektstärken wurden auf der people-things-Dimension festgestellt. Frauen zeigten eine deutlich höhere Ausprägung des Interessentyps Social (d = -0.68), während bei Männern das Interesse im Bereich Realistic stärker ausgeprägt war (d = 0.84). Für die übrigen Interessentypen wurden vergleichsweise geringe Effektstärken berechnet (Investigative: d = 0.26; Artistic: d = -0.35; Conventional: d = -0.33; Enterprising: d = 0.04). Bezüglich der data-ideas-Dimension zeigten sich somit sehr geringe Geschlechtsunterschiede. Auch in einer früheren Untersuchung von Lippa (1998) zu interessenbezogenen Geschlechterdifferenzen wurde bereits über hohe Geschlechterdifferenzen auf der people-things- und geringe Geschlechterdifferenzen auf der data-ideas-Dimension berichtet. Fouad (1999) wies auf die zeitliche Kontinuität der Geschlechterdifferenzen in Interessen seit mehreren Jahrzehnten hin "despite social changes that have increased women's participation in the workforce" (S. 200). Ausnahmen bildeten Interesseninventare, bei deren Entwicklung von vornherein Geschlechterdifferenzen

ausgeschlossen wurden. Der Entwicklung dieser Fragebögen liegt eine Kontroverse zugrunde, die im Folgenden näher beschrieben wird.

1974 wurden die sogenannten Guidelines for Assessment of Sex Bias and Sex Fairness in Career Interest Inventories des National Institute of Education (NIE) veröffentlicht. Einzelne Skalen von Interessenfragebögen sollten diesen Richtlinien zufolge in gleicher Anzahl Items aufweisen, die von Frauen beziehungsweise Männern bevorzugte Aktivitäten beschrieben "within the limitations imposed by validity considerations" (National Institute of Education, 1974, S. xxv). Die Forschergruppe des American College Testing Program (ACT) befürwortete die Entwicklung von Interesseninventaren, die keine Geschlechterdifferenzen aufweisen. Die Forschergruppe um John Holland stand dem hingegen kritisch gegenüber, da sie durch die Elimination von Geschlechterdifferenzen in den Interesseninventaren eine Abnahme der Validität befürchteten (siehe Su et al., 2009). In der Folge kam es zu einer langanhaltenden Debatte zwischen einerseits Vertretern des *opportunity approach to validation*, welche die Minimierung von Geschlechterdifferenzen befürworteten, und auf der anderen Seite denen des *socialization approach to validation*, welche diese Minimierung ablehnten. Prediger und Cole (1975) traten für den opportunity approach to validation ein. Sie kritisierten die Verwendung von Geschlechterverteilungen in verschiedenen Berufsfeldern als Validitätskriterium, da hierdurch Geschlechterstereotype aufrechterhalten würden. "By this standard, interest inventories of 1850s would have suggested farming to nearly all males and homemaking to nearly all females" (Prediger & Cole, 1975, S. 244). Ihrer Ansicht nach liegt ein wesentliches Ziel von Interessenfragebögen darin, die Möglichkeiten zur Exploration des Berufslebens sowohl für Frauen als auch für Männer zu maximieren, indem Geschlechterdifferenzen in den Verteilungen der Testscores minimiert werden. Die genannten Richtlinien und der opportunity approach to validation führten in Folge bei der Entwicklung einzelner Testverfahren zur gezielten Elimination von Items, die

Geschlechterdifferenzen aufwiesen (Su et al., 2009). Wichtige Vertreter des socialization approach to validation sind Gottfredson und Holland (1978). Sie betonten, dass die Sozialisierungen, die Frauen und Männer erfahren, für die Interessen einer Person relevant sind und daher nicht ignoriert werden sollten. Sofern ein Interessenfragebogen eingesetzt wird, um die aktuelle berufliche Neigung einer Person zu erfassen, ist ihrer Meinung nach davon auszugehen, dass durch die Elimination von Geschlechterdifferenzen die Konstruktvalidität des Fragebogens gemindert wird, "because the personality constructs measured are, in theory, dependent upon a person's experience, and at the present time this experience is usually different for men and women" (Gottfredson & Holland, 1978, S. 44).

Da bei der Konstruktion von Interessenfragebögen unterschiedliche Perspektiven eingenommen wurden, variieren die Geschlechterdifferenzen, die mit diesen Messinstrumenten gefunden wurden, stark. Su et al. (2009) stellten fest, dass nicht bei allen Interessentypen des RIASEC-Modells die Geschlechterdifferenzen in gleichem Ausmaß von der Haltung der Testentwickler betroffen sind. Für die Skala Realistic wurden die stärksten Variationen der Geschlechterdifferenzen in verschiedenen Test gefunden. So berichteten Su et al. (2009) über einen Test, bei dessen Entwicklung gezielt Items mit Effektstärken, die größer als 0.15 waren, eliminiert wurden. Der Test zeigte eine Effektstärke von $d = 0.15$ für die Skala Realistic. In der jüngsten Ausgabe eines anderen Tests, bei dessen Entwicklung auch Items mit hohen Geschlechterdifferenzen beibehalten wurden, betrug die Effektstärke der entsprechenden Skala hingegen $d = 1.16$. Eine solche Differenz in den Effektstärken wirft Fragen auf, wie beispielsweise ob Interesseninventare, deren Geschlechterdifferenzen sich in einem solchen Ausmaß unterscheiden, das gleiche Konstrukt messen, und ob ein Test, dessen Itemselektion unter anderem nach dem Kriterium der Vermeidung von Geschlechterdifferenzen erfolgte, thematisch hinreichend heterogen ist, um die Interessen von Probandinnen und Probanden abzubilden. In Bezug auf die vorliegende Arbeit ist

eindeutig der socialization approach to validation zu befürworten. Für die Beantwortung der Frage, ob gemessene Geschlechterdifferenzen des Wissens durch Geschlechterdifferenzen von Interessen erklärt werden können, bedarf es der Erfassung tatsächlicher Interessen von Frauen und Männern. Inwieweit bestehende Interessenunterschiede durch Sozialisation oder andere Faktoren begründet werden können, ist hierfür irrelevant. Auch spielt die Maximierung der beruflichen Exploration von Frauen und Männern in diesem Rahmen keine Rolle[5]. Wie bereits erwähnt, wurde in der Meta-Analyse von Su et al. (2009) über hohe Geschlechterdifferenzen für die Interessentypen Social und Realistic berichtet. Es sollte hierbei zusätzlich berücksichtigt werden, dass die Meta-Analyse auch Interesseninventare einschloss, bei deren Entwicklung der opportunity approach to validitation vertreten wurde, bei denen also das Ziel einer Verminderung von Geschlechterdifferenzen verfolgt wurde. Bei ausschließlicher Berücksichtigung von Fragebögen, deren Entwickler die Position des socialisation approach to validation einnahmen, würden die gefundenen Geschlechterdifferenzen höchstwahrscheinlich noch höher ausfallen.

Einschränkend bleibt jedoch anzumerken, dass bei der Überprüfung der Fairness von Interessentests auf psychometrischer Ebene Hinweise darauf gefunden wurden, dass bisher gemessene Geschlechterdifferenzen auch durch Verzerrungen begründet sind. So überprüften Wetzel und Hell (2013) die Items des AIST-R (Bergmann & Eder, 2005) auf Differential Item Functioning (DIF) für die Geschlechter und fanden vor allem bei den Skalen Realistic,

[5]Es sei erwähnt, dass auch unabhängig von der Fragestellung der Autor dieser Arbeit die Perspektive des socialization approach to validation bevorzugt. Gottfredson und Holland (1978) formulierten treffend: "When explicit attempts are made to influence a person to abandon traditionally held sex-role preferences, vocational interventions might not need to include an assessment of a person's current status at all" (S. 44).

Social und Enterprising Items mit DIF. Die Geschlechterdifferenzen der Skala Realistic wurden durch die Entfernung der entsprechenden Items reduziert (d = 0.37 statt 0.54). Bei anderen Skalen blieben nennenswerte Änderungen der Effektstärken durch die Entfernung von Items mit DIF aus. Einarsdóttir und Rounds (2009) überprüften den Strong Interest Inventory (SII) von Harmon, Hansen, Borgen und Hammer (zitiert nach Einarsdóttir & Rounds, 2009) auf DIF für Frauen und Männer. Bei etwa zwei Drittel der Items wurde DIF bestätigt. Insbesondere die Items der Skalen Realistic und Investigative waren hiervon betroffen. Nach Entfernung der Items mit DIF wurden die Geschlechterdifferenzen in diesen beiden Skalen reduziert (d = 0.64 statt 0.86 für Realistic; d = 0.14 statt 0.30 für Investigative). Geschlechterdifferenzen anderer Skalen änderten sich durch die Entfernung von Items mit DIF nicht in nennenswertem Ausmaß. Einarsdóttir und Rounds weisen darauf hin, dass sich die Arbeitswelt seit der Entwicklung des RIASEC-Modells verändert habe, beispielsweise hinsichtlich des Anteils von Frauen, die bezahlten Tätigkeiten nachgehen. Die Autoren fordern eine Überarbeitung der Definitionen der Interessentypen des RIASEC-Modells. Als Beispiel nannten sie den Interessentyp Realistic, dessen aktuelle Definition zu eng und auf für Männer typische Tätigkeiten der Arbeitswelt begrenzt sei, während "experiences women have (e.g. gardening, making clothes, driving children to school, cooking)" (Einarsdóttir & Rounds, 2009, S. 306) nicht berücksichtigt würden.

Forschung zu Geschlechterdifferenzen in Interessen ist nicht auf das RIASEC-Modell beschränkt. Als weiteres Beispiel wird hier eine Studie von Bennett und Hogarth (2009) angeführt, in der die Interessen von Schülerinnen und Schülern für verschiedene Naturwissenschaften erfasst wurden. Während mehr Mädchen als Jungen sich für Biologie interessierten ($p < .01$), äußerten mehr Jungen als Mädchen Interesse für Physik ($p < .001$) und Chemie ($p < .05$). Dieses Beispiel weist auch darauf hin, dass Interessen mit einem

möglichst hohen Grad an Spezifität erfragt werden sollten, wenn das Ziel in der Feststellung von Geschlechterdifferenzen liegt.

Insgesamt können geschlechtsspezifische Tendenzen zu bestimmten Interessengebieten als empirisch nachgewiesen betrachtet werden, wobei die Größe der Effekte aus den im vorigen Abschnitt beschriebenen Gründen offen ist. Die Befunde für die Richtung der Effekte sind jedoch eindeutig: Legt man das RIASEC-Modell von Holland (1959, 1997) zugrunde, findet man bei Frauen ein größeres Interesse für den Interessentyp Social als bei Männern. Umgekehrt tendieren Männer stärker zum Interessentyp Realistic als Frauen. Diese beiden Interessentypen bilden die Pole der people-things-Dimension nach Prediger (1982). Wie bereits erläutert, wird angenommen, dass das Interesse einer Person für einen bestimmten Themenbereich unmittelbar die Motivation der Person zur Beschäftigung mit dem entsprechenden Bereich und damit mittelbar auch das Wissen, das sie in dem Themenbereich erwirbt, beeinflusst. Demzufolge sollten Frauen und Männer in unterschiedlichen Themen ein hohes Wissen aufweisen. Eine fehlende Balance zwischen Items zu Interessengebieten von Frauen und Items zu Interessengebieten von Männern wäre somit eine potentielle Erklärung für die bestehenden Geschlechterdifferenzen in Tests des Allgemeinen Wissens. In dem Zusammenhang sei auch an die Befunde von Low et al. (2005) zur zeitlichen Stabilität von (beruflichen) Interessen erinnert. Wie in Kapitel 2.3.3 erläutert, differenzierte Cattell (1971/1987) im Rahmen der Investmenttheorie explizit zwischen früheren und zum Zeitpunkt der Messung von gc aktuellen Interessen. Für die vorliegende Arbeit ist diese Unterscheidung jedoch von untergeordneter Bedeutung, da Interessen – somit auch Geschlechterdifferenzen in Interessen – zeitlich stabil sein dürften. Im Folgenden wird die erste Untersuchung beschrieben, in welcher der Frage nachgegangen wurde, ob bestehende Wissenstests möglicherweise eine größere Anzahl an Items zu Themengebieten enthalten, für die Männer

ein höheres Interesse aufweisen, während das Wissen zu Interessengebieten von Frauen in geringerem Ausmaß erfasst wird.

5.2 Studie 1 - Balance der Interessen von Frauen und Männern in Wissenstests

Die Ziele der ersten Studie bestanden darin, (a) einen möglichst umfassenden Überblick über die Interessengebiete von Frauen und Männern zu gewinnen und (b) die Items des BOWIT (Hossiep & Schulte, 2008) und des Erweiterungsmoduls des I-S-T 2000 R (Liepmann et al., 2007) den Interessengebieten zuzuordnen, um damit die Fragen beantworten zu können, (c) inwieweit in diesen Tests eine Balance zwischen Items zu Interessengebieten von Frauen und Interessengebieten von Männern besteht und (d) inwieweit die Annahme tragbar ist, dass Interessenunterschiede zwischen Frauen und Männern von ursächlicher Bedeutung für die Leistungsunterschiede sein könnten. Nach der Strukturierung von Interessengebieten, also der Unterteilung von Interessen in verschiedene Inhaltsbereiche, wurden auf Basis hiervon die Interessen von Frauen und Männern erhoben sowie die Items des I-S-T Wissenstests und des BOWIT thematisch geordnet. Die so gewonnenen Daten ermöglichten verschiedene Analysen, die zur Beantwortung der genannten Fragen dienten. Im Folgenden werden die methodischen Vorgehensweisen, die Ergebnisse der Datenanalysen und die hieraus gezogenen Schlüsse näher beschrieben.

5.2.1 Methoden

Wie in Kapitel 5.1 bereits erläutert wurde, basiert die Forschung zu Interessenunterschieden von Frauen und Männern zu einem großen Teil auf dem RIASEC-Modell von Holland (1959, 1997). In den Manualen der beiden Wissenstests, die für die vorliegende Studie verwendet wurden, findet sich jedoch keine Zuordnung der Items zu den sechs Interessentypen dieses Modells. Stattdessen werden hier die Themen aufgelistet, zu denen die Tests Items enthalten. Die Items des BOWIT werden in 11 Bereiche unterteilt, für die jeweils eine eigene Skala

gebildet werden kann. Hierbei handelt es sich um die Bereiche Bildende Kunst, Biologie/Chemie, Ernährung/Bewegung/Gesundheit, Geographie/Verkehr, Geschichte/ Archäologie, Gesellschaft/Zeitgeschehen/Politik, Mathematik/Physik, Philosophie/Religion, Sprache/Literatur, Technik/EDV und Wirtschaft/Recht. Für den Wissenstest des I-S-T wurden Items zu den sechs Bereichen Geographie/Geschichte, Wirtschaft, Kunst/Kultur, Alltag, Naturwissenschaften und Mathematik entwickelt. Es werden jedoch keine themenspezifischen Skalen gebildet. Auch geht aus dem Manual nicht hervor, welches Item welchem Bereich zuzurechnen ist. Es wird somit deutlich, dass die Gruppierung von Themen in den beiden Tests sehr unterschiedlich ausfällt. Während im I-S-T beispielsweise eine Kategorie mit Geographie/Geschichte bezeichnet wird, existieren im BOWIT die Kategorien Geographie/Verkehr und Geschichte/Archäologie. Eine Voraussetzung für die vorliegende Untersuchung war eine einheitliche Strukturierung von Themengebieten. Es bedurfte einer Homogenisierung der Unterteilungen von Interessen in unterschiedliche Inhaltsbereiche, um zu einer Struktur zu gelangen, auf deren Basis im weiteren Verlauf der Studie sowohl die Interessen von Frauen und Männern erfasst als auch die Items des I-S-T und des BOWIT thematisch zugeordnet werden konnten.

Strukturierung von Interessengebieten

Der erste Schritt für die Entwicklung einer Struktur war die Erstellung einer Liste von Themen, die sämtliche Themengebiete abdeckte, zu denen im I-S-T und im BOWIT Items vorliegen. Da auch überprüft werden sollte, bei welchen Themenbereichen in den untersuchten Wissenstests möglicherweise ein Mangel an Items vorliegt, bedurfte es einer Erweiterung der Liste um zusätzliche Themen. Als Orientierung diente hierbei die Gliederung der Studienfächer, die auch in der Studentenstatistik des Jahres 2009 vom Statistischen Bundesamt verwendet wurde. Interessenunterschiede zwischen Frauen und

Männern waren bei der Erstellung der Themenliste hingegen nicht handlungsleitend. So wurde beispielsweise nicht das Ziel einer ausgewogenen Repräsentation der Interessentypen Social und Realistic des RIASEC-Modells verfolgt. Die Gliederung der Studienfächer in der Studentenstatistik des Jahres 2009 beinhaltet die Zusammenfassung von Studienfächern in Studienbereiche und die Zusammenfassung von Studienbereichen in 10 Fächergruppen. Hierbei handelt es sich um

1. Sprach- und Kulturwissenschaften
2. Sport
3. Rechts-, Wirtschafts-, und Sozialwissenschaften
4. Mathematik, Naturwissenschaften
5. Humanmedizin/Gesundheitswissenschaften
6. Veterinärmedizin
7. Agrar-, Forst- und Ernährungswissenschaften
8. Ingenieurwissenschaften
9. Kunst, Kunstwissenschaft
10. Sonstige Fächer und ungeklärt

Diese Einteilung erschien für das vorliegende Ziel der Erstellung einer Themenliste zu allgemein. Die Differenzierung auf Ebene der Studienbereiche erschien jedoch zu detailliert. So setzt sich beispielsweise die Fächergruppe Agrar-, Forst- und Ernährungswissenschaften aus den Studienbereichen (a) Landespflege, Umweltgestaltung, (b) Agrarwissenschaften, Lebensmittel- und Getränketechnologie, (c) Forstwissenschaft, Holzwirtschaft und (d) Ernährungs- und Haushaltswissenschaften zusammen.

Bei dem Versuch, einen Kompromiss zwischen einer sehr allgemeinen und einer sehr detaillierten Unterteilung der Interessengebiete zu finden, wurden im Ergebnis manche Fächergruppen relativ undifferenziert in die Liste aufgenommen. So wird beispielsweise die

Fächergruppe Agrar-, Forst- und Ernährungswissenschaften in der Themenliste lediglich durch die Interessengebiete Natur und Ernährung abgebildet. Die Fächergruppe Veterinärmedizin, die ausschließlich einen gleichnamigen Studienbereich beinhaltet, wurde nicht als separates Thema in die Liste aufgenommen, da diese durch die Themengebiete Medizin und Biologie hinreichend vertreten erschien. Für andere Fächergruppen wurde eine relativ präzise Differenzierung bevorzugt. Ein Beispiel hierfür ist die Fächergruppe Kunst, Kunstwissenschaft, welche in der Studentenstatistik des Statistischen Bundesamtes von 2009 in die Studienbereiche (a) Kunst, Kunstwissenschaft allgemein, (b) Bildende Kunst, (c) Gestaltung, (d) Darstellende Kunst, Film und Fernsehen, Theaterwissenschaft und (e) Musik, Musikwissenschaft unterteilt wird. Die für die vorliegende Untersuchung erstellte Liste enthält unter anderem die Themen Darstellende Kunst, Bildende Kunst, Musik, Modedesign und Raumdesign, womit nahezu die gesamte Fächergruppe abgedeckt ist.

Die Erstellung der Themenliste war somit stark subjektiv geprägt. Es wäre eine unbegrenzte Anzahl weiterer Unterteilungen denkbar, bei denen die Detailliertheit der thematischen Abgrenzungen anders ausfallen würde. Außerdem weisen viele Themen auch Überschneidungen auf und sind nicht immer eindeutig trennbar. Ein Beispiel hierfür sind die Themen Soziale Arbeit und Politik. Diese unvermeidbaren Umstände müssen bei der vorliegenden Untersuchung und auch bei den sich anschließenden Studien, die darauf aufbauen, berücksichtigt werden. Die vollständige Themenliste umfasste 36 Interessengebiete, welche die Grundlage für die Erstellung eines Interessenfragebogens und für die Zuordnung der Items des I-S-T Wissenstests und des BOWIT bildeten.

Erstellung des Interessenfragebogens

Der Interessenfragebogen beinhaltete 36 Items. Anhang E enthält den vollständigen Fragebogen, aus dem auch die Themen ersichtlich sind. Für jedes Thema aus der Liste wurde

ein Item erstellt. Um sicherzustellen, dass verschiedene Personen, die den Fragebogen bearbeiteten, jedes Item möglichst ähnlich interpretierten, wurde über die Benennung des Themas hinaus eine kurze beispielhafte Umschreibung in Stichpunkten gegeben. So wurde beispielsweise das Thema Modedesign durch die Begriffe Entwerfen und Gestalten, Kleidermode und Accessoires präzisiert. Die Umschreibungen dienten außerdem dazu, dass die Benennungen der Themen nicht ausschließlich im beruflichen Kontext interpretiert wurden. So sollte zum Beispiel das Thema Sport auch das Verfolgen von Sportsendungen im Fernsehen und das Betreiben von Sport als Hobby umfassen. Für jedes Thema sollten die Probandinnen und Probanden auf einer 5-stufigen Likert-Skala von 0 bis 4 angeben, wie stark ihr jeweiliges Interesse ist. Die Pole der Skala waren dabei umschrieben mit *Kein Interesse* für 0 und *Sehr großes Interesse* für 4. In Abbildung 4 sind beispielhaft zwei Items des Interessenfragebogens dargestellt.

	Kein Interesse				Sehr großes Interesse
Sport (Aktiv Sport treiben, Sportsendungen im TV sehen, …)	0	1	2	3	4
Ernährung (Kochen, Esskultur, …)	0	1	2	3	4

Abbildung 4. Auszug aus Interessenfragebogen.

Der Interessenfragebogen wurde von 150 Schülerinnen und Schülern (59.3% weiblich) aus der 11. und 12. Stufe eines Gymnasiums im Abstand von vier Monaten zweimal bearbeitet. Auf Basis dieser Daten wurde die Retest-Reliabilität des Interessenfragebogens überprüft.

Ein wichtiges Kriterium für die Verwertbarkeit der Ergebnisse des Fragebogens war Vollständigkeit. Der Fragebogen sollte dazu dienen, einen umfassenden Überblick über die

Interessen der Personen zu geben. Um die Vollständigkeit zu überprüfen, wurde am Ende des Fragebogens die Frage gestellt, ob die Probandin oder der Proband Interesse an einem oder mehreren Themen habe, die nicht in der Liste aufgeführt waren. Die Teilnehmenden konnten die Frage durch Ankreuzen bejahen oder verneinen. Falls die Person die Frage bejahte, sollten die entsprechenden Themen benannt werden.

Erfassung der Interessen von Frauen und Männern

Um eine Kategorisierung der Themen in Neutrale Interessengebiete (NI), Interessengebiete von Frauen (IvF) und Interessengebiete von Männern (IvM) vornehmen zu können, wurde dieser Fragenbogen n = 507 (72.0% Frauen) Personen vorgelegt. Da die Bearbeitung des Fragebogens weniger als 5 Minuten in Anspruch nahm, konnte er bei zahlreichen Datenerhebungen eingebunden werden. Sofern die Datenerhebung auch einen Wissenstest umfasste, bearbeiteten die Teilnehmenden den Interessenfragebogen vor dem Wissenstest. Bei den Probandinnen und Probanden handelte es sich um Studierende unterschiedlicher Fächer an den Universitäten in Potsdam und Wuppertal und um Schülerinnen und Schüler eines Berufskollegs in Xanten. In Potsdam wurden außerdem Personen vor einem Supermarkt angesprochen und gebeten, den Fragebogen direkt vor Ort auszufüllen. Die Altersspanne der Frauen reichte von 16 bis 70 Jahren (M = 23.45, SD = 9.28). Die Männer waren ebenfalls zwischen 16 und 70 Jahre alt (M = 27.35, SD = 9.71).

Die Kategorisierung der 36 Themen in NI, IvF und IvM erfolgte auf Basis des Effektstärkemaßes d nach Cohen (1988). Für beide Geschlechter wurden separat die Mittelwerte und Standardabweichungen der Interessenscores berechnet. Für jedes Thema wurde Cohens d berechnet, wobei die unterschiedlichen Standardabweichungen von Frauen und Männern entsprechend den Größen der Teilstichproben gewichtet wurden. Ein positiver Wert für Cohens d repräsentierte ein durchschnittlich stärkeres Interesse von Frauen an dem

Thema, während eine Effektstärke mit negativem Vorzeichen ein durchschnittlich stärkeres Interesse von Männern darstellte. Themen, für die $d < -0.10$ war, wurden als IvM kategorisiert. Themen mit $d > 0.10$ wurden als IvF kategorisiert. Die übrigen Themen wurden der Kategorie NI zugeordnet.

Zuordnung der Items des I-S-T und des BOWIT

Die Zuordnung der Items des Wissenstests des I-S-T und des BOWIT wurden durch neun Psychologie-Studierende vorgenommen. Es handelte sich hierbei um sieben Frauen und zwei Männer, die sich am Ende des vierten oder eines höheren Semesters des Bachelorstudiengangs der Psychologie an der Universität Wuppertal befanden. Die Personen bekamen den BOWIT und den Wissenstest des I-S-T (jeweils Form A) vorgelegt sowie die Themenliste inklusive der näheren Umschreibungen, die auch bei dem Interessenfragebogen zur Präzisierung dienten. Jedes Item sollte einem Thema der Liste zugeordnet werden.

In den folgenden Analysen wurde nicht mehr zwischen den 36 Themen sondern lediglich zwischen den drei Kategorien IvF, NI und IvM differenziert. So war es für die folgenden Analysen beispielsweise irrelevant, ob die Personen ein Item dem Thema Computer oder dem Thema Informatik zugeordnet hatten, da beide Themen in die Kategorie IvM fielen. Daher wurden die Zuordnungen der Items zu Themen in Zuordnungen zu den drei Kategorien transformiert. Um für die weiteren Analysen jedes Item einer Kategorie zuzuordnen, wurde für jedes einzelne Item ausgezählt, wie häufig es welcher Kategorie zugeordnet worden war. Die Kategorie, zu deren Themen das Item von den 9 Studierenden am häufigsten zugeordnet worden war, wurde ausgewählt.

Im Anschluss erfolgte für sämtliche Items die Berechnung der Beurteilerübereinstimmung. Als Koeffizient wurde hier Krippendorffs α (Krippendorff, 2004) gewählt, da dieser Koeffizient auch für den Vergleich der Urteile von mehr als zwei

Beurteilern geeignet ist. Der Wertebereich von Krippendorffs α reicht von 0 bis 1, wobei 0 bei völligem Fehlen von Übereinstimmung und 1 bei perfekter Übereinstimmung berechnet wird. Krippendorff (2004) schlug den Wert .667 als Minimum vor. Die weiteren hier beschriebenen Analysen erfolgten daher nur unter Einbeziehung der Items, für die bei der Zuordnung auf Ebene der Kategorien eine Beurteilerübereinstimmung von $\alpha \geq .667$ berechnet wurde. Hayes und Krippendorff (2007) bieten ein Makro für die Berechnung von Krippendorffs α an, das hier mit SPSS 22 verwendet wurde.

Balance der Kategorien im I-S-T und im BOWIT

Wie im vorigen Abschnitt beschrieben, wurden die Zuordnungen der Items des I-S-T und des BOWIT zu den 36 Themen in Zuordnungen zu den 3 Kategorien transformiert. Anschließend wurden für beide Tests die Häufigkeiten der Items in den Kategorien ausgezählt. Diese Ergebnisse gaben Informationen über die Ausgewogenheit der drei Kategorien im I-S-T und im BOWIT.

Kategorienspezifische Leistungsunterschiede

Auf Basis von zwei Datensätzen, die im Rahmen von Abschlussarbeiten erhoben und bei denen unter anderem auch der Wissenstest des I-S-T bzw. der BOWIT verwendet worden waren, wurden für die Kategorien NI, IvF und IvM die Geschlechterdifferenzen der Leistungen berechnet. Der für den I-S-T verwendete Datensatz umfasste 273 Studierende (63.0% Frauen) der Universitäten in Münster und Wuppertal. Der Datensatz des BOWIT umfasste 134 Studierende (60.4% Frauen) der Wuppertaler Universität. Neben den Berechnungen der Geschlechterdifferenzen im gesamten Test wurden für alle drei Kategorien jeweils die Summenscores der Teilnehmenden aus den Items gebildet, die mit einem Krippendorffs $\alpha \geq .667$ der jeweiligen Kategorie zugeordnet worden waren. Darauf folgten

für alle Kategorien die Berechnungen der Leistungsunterschiede von Frauen und Männern, welche jeweils durch einen *t*-Tests inferenzstatistisch geprüft wurden.

5.2.2 Ergebnisse

Die Ergebnisse der Prüfung der Retest-Reliabilität des Interessenfragebogens sind in Tabelle 8 wiedergegeben. Wie in Kapitel 5.2.1 beschrieben, erfolgte die Überprüfung anhand der Daten einer separaten Stichprobe von n = 150 Personen. Für alle 36 Interessengebiete wurden – sowohl geschlechtsspezifisch als auch für die Gesamtstichprobe – die Korrelationen zwischen dem ersten und dem zweiten Messzeitpunkt berechnet. Für die Gesamtstichprobe wurde ein Mittelwert aller 36 Korrelationen von M_r = .74 berechnet. Die Standardabweichung dieser Korrelationen betrug SD_r = .07. Zum Vergleich sei hier eine frühere Version des bereits in Kapitel 5.1 erwähnten AIST-R von Bergmann und Eder (2005) herangezogen. Die sechs Skalen des AIST von Bergmann und Eder (1992) wiesen über einen Zeitraum von vier Monaten Stabilitäten zwischen r = .61 (Conventional) und r = .74 (Artistic und Social) auf, wobei der Mittelwert der sechs Korrelationskoeffizienten M_r = .69 betrug. Die Retest-Reliabilität des neu entwickelten Interessenfragebogens ist somit sehr zufriedenstellend.

Von den insgesamt 507 Teilnehmenden, deren Interessen erfasst wurden, nutzten 44 Personen die Möglichkeit, weitere Themengebiete zu benennen, für die sie sich interessierten und die nicht im Fragebogen genannt wurden. Am häufigsten wurden die Begriffe „Tiere“ und "Haustiere" eingetragen. Insgesamt sieben Personen (1.4 %) gaben eines dieser Interessengebiete an. An zweiter Stelle standen die Begriffe „Sex“ oder „Sexualität“, die zusammen von fünf Personen genannt wurden (1.0 %). Die Begriffe „Freunde“ oder „Soziale Kontakte“ wurden insgesamt von vier Personen angegeben (0.8 %). Tabelle E1 enthält sämtliche Einträge der 44 Personen, die zusätzliche Themen benannt haben. Die Mehrheit der

Antworten auf die offene Frage kann durchaus einem der 36 Themen zugeordnet werden. So fällt beispielsweise der Begriff Tiere in die Kategorien Biologie und Natur. Der Begriff Sexualität bildet ebenfalls ein Teilgebiet der Biologie. Das für den Interessenfragebogen gesetzte Ziel der Vollständigkeit wurde somit erreicht.

Tabelle 8

Retest-Reliabilität des Interessenfragebogens

Interessengebiet	r_{Frauen}	$r_{Männer}$	r_{Gesamt}
Sport	.73	.69	.73
Ernährung	.72	.51	.63
Medizin	.85	.75	.81
Gesundheit	.66	.75	.70
Literatur	.74	.79	.78
Sprachwissenschaft	.68	.66	.69
Fremdsprachen	.72	.71	.74
Fremde Kulturen	.65	.47	.60
Elektrotechnik	.77	.65	.81
Computer	.68	.40	.66
Informatik	.80	.78	.82
Maschinen	.76	.81	.85
Biologie	.85	.81	.83
Chemie	.71	.77	.74
Physik	.77	.77	.80
Mathematik	.81	.83	.83
Archäologie	.65	.55	.60
Modedesign	.76	.64	.77
Raumdesign	.75	.66	.75
Architektur	.61	.73	.67
Bildende Kunst	.72	.70	.72
Darstellende Kunst	.68	.49	.68

(wird fortgesetzt)

Interessengebiet	r_{Frauen}	$r_{Männer}$	r_{Gesamt}
Musik	.70	.60	.65
Philosophie	.78	.70	.75
Religion	.78	.78	.77
Wirtschaft	.73	.83	.82
Politik	.84	.76	.82
Gesellschaft	.69	.56	.64
Psychologie	.77	.74	.76
Pädagogik	.72	.71	.74
Soziale Arbeit	.77	.60	.74
Natur	.77	.82	.79
Verkehr	.68	.73	.74
Geographie	.81	.84	.84
Geschichte	.81	.70	.77
Recht	.63	.48	.58
Mittelwert	.74	.69	.74
Standardabweichung	.06	.12	.07

Interessen von Frauen und Männern

Das Histogramm in Abbildung 5 gibt die Verteilung der Effektstärken der Interessenunterschiede bei den 36 Themen wieder, die mit dem Interessenfragebogen berechnet wurden. In Tabelle 9 findet sich eine Auflistung der Themen, sortiert nach Effektstärken. Unter den 36 Themen waren 11 Themen, für die Effektstärken von d > 0.10 berechnet wurden, und die somit als IvF kategorisiert wurden. Bei acht Themen ergaben sich Effektstärken von $-0.10 \leq d \leq 0.10$. Diese wurden als NI kategorisiert. Bei den übrigen 17 Themen ergaben sich Effektstärken von $d < -0.10$, weshalb sie als IvM klassifiziert wurden.

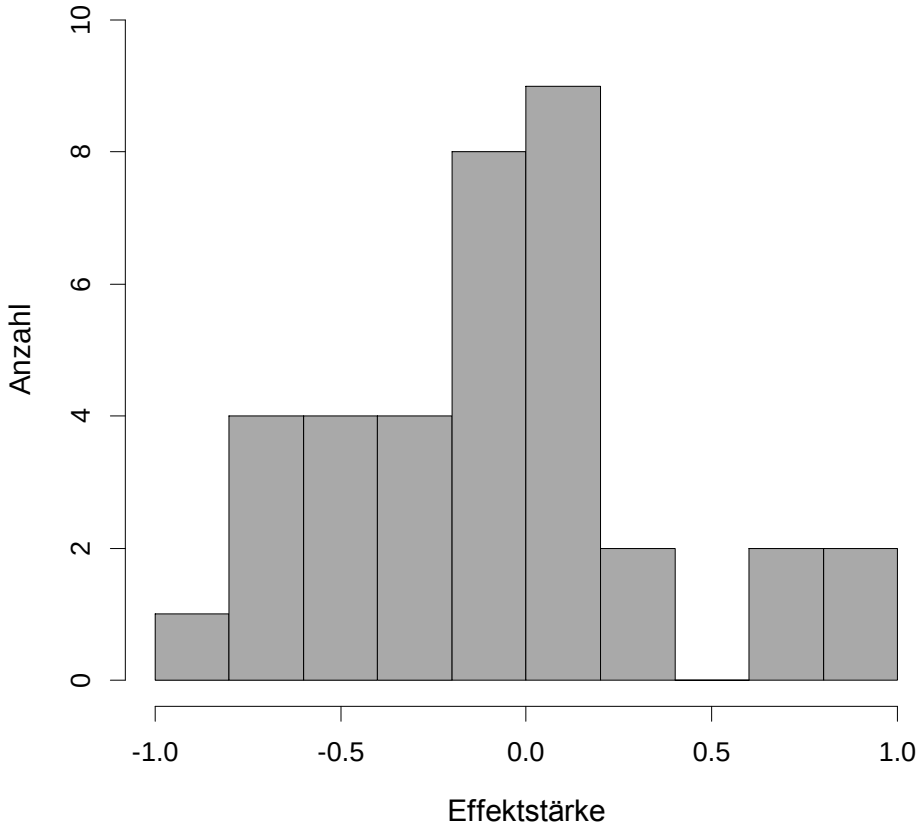

Abbildung 5. Histogramm der Effektstärken der Geschlechterdifferenzen des Interesses für 36 Themen (Cohens *d*). Positives Vorzeichen = Höheres Interesse von Frauen. Negatives Vorzeichen = Höheres Interesse von Männern.

Tabelle 9

Effektstärken der Geschlechterdifferenzen des Interesses für 36 Themen

Thema	Cohens *d*	Kategorie
Raumdesign	0.90	
Modedesign	0.89	Interessengebiete
Soziale Arbeit	0.71	von Frauen
Pädagogik	0.61	(IvF)
Psychologie	0.24	

(wird fortgesetzt)

Thema	Cohens *d*	Kategorie
Gesundheit	0.24	Interessengebiete von Frauen (IvF)
Biologie	0.20	
Natur	0.18	
Darstellende Kunst	0.18	
Medizin	0.18	
Ernährung	0.12	
Fremdsprachen	0.08	Neutrale Interessengebiete (NI)
Literatur	0.08	
Musik	0.06	
Bildende Kunst	0.05	
Religion	-0.01	
Mathematik	-0.02	
Sprachwissenschaft	-0.06	
Verkehr	-0.08	
Recht	-0.14	Interessengebiete von Männern (IvM)
Fremde Kulturen	-0.15	
Chemie	-0.15	
Gesellschaft	-0.16	
Architektur	-0.24	
Archäologie	-0.25	
Computer	-0.27	
Geographie	-0.38	
Sport	-0.42	
Philosophie	-0.42	
Geschichte	-0.54	
Informatik	-0.59	
Wirtschaft	-0.68	
Politik	-0.68	
Elektrotechnik	-0.74	
Physik	-0.78	
Maschinen	-0.80	

Zuordnung der Items des I-S-T und des BOWIT

Von den 84 Items des Wissenstests des I-S-T 2000 R wurde für 22 Items ein Krippendorffs $\alpha < .667$ berechnet (26.2 %). Beim BOWIT fiel Krippendorffs α bei 28 der 154 Items (18.2 %) in den nicht-akzeptablen Bereich. In Tabelle 10 sind die Häufigkeiten der Items angegeben, die in den überprüften Wissenstests den drei Kategorien IvF, NI und IvM zugeordnet wurden. Es fanden in die Tabelle lediglich die Items Eingang, für die mindestens eine akzeptable Beurteilerübereinstimmung ($\alpha \geq .667$) ermittelt wurde. In den Tabellen F1 und F2 sind für alle einzelnen Items des I-S-T und des BOWIT die Werte für Krippendorffs α und die Kategorien, denen sie mehrheitlich zugeordnet wurden, aufgelistet.

Tabelle 10

Absolute (relative) Häufigkeiten der Items im I-S-T 2000 R Wissenstest und im BOWIT zu Interessenkategorien

Kategorie	Anzahl Items I-S-T	Anzahl Items BOWIT
IvF	1 (.02)	16 (.13)
NI	23 (.37)	26 (.21)
IvM	38 (.61)	84 (.67)

Anmerkungen. Es wurden lediglich die Items berücksichtigt, für die ein Krippendorffs $\alpha \geq .667$ berechnet wurde. IvF = Interessengebiete von Frauen; NI = Neutrale Interessengebiete; IvM = Interessengebiete von Männern.

Balance der Kategorien im I-S-T und im BOWIT

Wie aus Tabelle 10 ersichtlich wird, wurde von den 62 Items des I-S-T, die ein $\alpha \geq .667$ aufwiesen, lediglich ein Item der Kategorie IvF zugeordnet, während 38 Items der Kategorie IvM zugeordnet wurden. Von den 126 Items des BOWIT, die eine akzeptable Beurteilerübereinstimmung aufwiesen, wurden der Kategorie IvF 16 Items und der Kategorie IvM 84 Items zugeordnet.

Kategorienspezifische Leistungsunterschiede

In den Tabellen 11 und 12 sind die kategorienspezifischen Leistungsunterschiede zwischen Frauen und Männern im Wissenstest des I-S-T und im BOWIT dargestellt. Unter Einbeziehung sämtlicher Items, auch derjenigen mit Krippendorffs $\alpha < .667$, zeigten sich in den beiden Tests für die Gesamtscores Geschlechterdifferenzen mit Effektstärken von -0.97 beim I-S-T und -1.16 beim BOWIT. Mit Ausnahme der Kategorie IvF beim I-S-T zeigten Männer bei allen Leistungsvergleichen bessere Ergebnisse. Für die Items der Kategorie IvF wurden Effektstärken von 0.05 beim I-S-T und -0.23 beim BOWIT berechnet. In der Kategorie NI ergaben sich Werte von -0.38 (I-S-T) und -0.59 (BOWIT). Die Effektstärken der Leistungsunterschiede in der Kategorie IvM betrugen -1.18 (I-S-T) und -1.39 (BOWIT).

Tabelle 11

Kategorienspezifische Leistungsunterschiede zwischen Frauen und Männern im Wissenstest des I-S-T 2000 R

	Gesamt		IvF		NI		IvM	
	F	M	F	M	F	M	F	M
M	50.76	59.35	0.88	0.86	16.21	17.27	19.62	25.19
SD	9.06	8.50	0.33	0.35	2.94	2.57	4.53	5.02
d	-0.97		0.05		-0.38		-1.18	
t_{df}	7.73_{271}		-0.39_{271}		3.01_{271}		9.42_{271}	
p	< .001		.700		.003		< .001	

Anmerkungen. IvF = Interessengebiete von Frauen; NI = Neutrale Interessengebiete; IvM = Interessengebiete von Männern; F = Frauen; M = Männer.

Tabelle 12

Kategorienspezifische Leistungsunterschiede zwischen Frauen und Männern im BOWIT

	Gesamt		IvF		NI		IvM	
	F	M	F	M	F	M	F	M
M	61.33	82.30	6.79	7.40	11.94	14.09	31.56	45.81
SD	15.83	21.40	2.73	2.60	3.40	4.11	9.17	11.94
d	-1.16		-0.23		-0.59		-1.39	
t_{df}	$6.12_{88.52}$		1.28_{132}		3.30_{132}		7.80_{132}	
p	< .001		.203		.001		< .001	

Anmerkungen. IvF = Interessengebiete von Frauen; NI = Neutrale Interessengebiete; IvM = Interessengebiete von Männern; F = Frauen; M = Männer.

Für die *t*-Tests und für die Levene-Tests zur Prüfung der Homoskedastizität wurde das 5%-Signifikanzniveau mit der Bonferroni-Methode auf 1.25% korrigiert. Nur der Levene-Test für die Varianzgleichheit des Gesamtscores für beide Geschlechter beim BOWIT ergab einen *p*-Wert < .0125. Sowohl beim I-S-T als auch beim BOWIT fielen die *t*-Tests für den gesamten Test und für die Leistungen in den Bereichen NI und IvM signifikant aus. In der Kategorie IvF ergaben sich für beide Tests *p*-Werte > .0125.

5.2.3 Diskussion

Die Entwicklung eines Fragebogens zur Erfassung der Interessen von Frauen und Männern ist zufriedenstellend verlaufen. Die Retest-Reliabilität mit einer durchschnittlichen Korrelation von .74 ist sehr gut. Die geringe Anzahl an zusätzlichen Eintragungen bei der Frage nach weiteren Interessengebieten, die nicht in der Liste aufgeführt sind, lässt die Schlussfolgerung zu, dass mit dem Fragebogen die Interessen der Probandinnen und

Probanden nahezu vollständig erfasst wurden. Ein Kritikpunkt bleibt die Subjektivität bei der Erstellung der Themenliste. Es wäre leicht möglich, einige der Themen des Interessenfragebogens zu einem einzelnen Thema zusammenzufassen. So könnten z.B. die Themen Elektrotechnik, Computer und Maschinen unter dem Begriff Technik zusammengefasst werden. Hierin zeigt sich auch das Problem, dass viele Themen nicht eindeutig voneinander abgrenzbar sind. Ebenso gibt es in dem Fragebogen Themen, die auch präziser hätten betrachtet werden können. Ein Beispiel hierfür ist die Darstellende Kunst. Diese ließe sich beispielsweise unterteilen in Theater, Film und Fernsehen, Tanz und Ballett. Aufgrund der Subjektivität wurde hier auch davon abgesehen, die Häufigkeiten der Themen, die als IvF und als IvM klassifiziert wurden, zu vergleichen. Die Tatsache, dass 17 Themen als IvM und lediglich 11 Themen als IvF kategorisiert wurden, dürfte wesentlich durch das Ausmaß der Differenziertheit bedingt sein. Allein die beiden oben genannten Beispiele für alternative Aufteilungen der Interessengebiete würden diese Häufigkeiten vermutlich verändern.

Die Beurteilerübereinstimmung bei der Zuordnung der einzelnen Items des BOWIT und des I-S-T zu den 3 Kategorien (nach der beschriebenen Transformation) fiel zufriedenstellend aus. Die beschriebenen Ergebnisse basierten auf insgesamt 73.8 % der Items des I-S-T und 81.8 % der Items des BOWIT. Bei beiden untersuchten Wissenstests wurde festgestellt, dass der Anteil der Items, die Themen zugeordnet wurden, für die Männer sich durchschnittlich stärker als Frauen interessieren, erheblich größer ist als der Anteil der Items, die Themen zugeordnet wurden, die der Kategorie IvF angehören. Nur ein einziges Item des I-S-T wurde mit hinreichender Beurteilerübereinstimmung der Kategorie IvF zugeordnet. Dagegen fielen 38 Items in die Kategorie IvM. Beim BOWIT war mit 16 (IvF) zu 84 (IvM) Items die gleiche Tendenz zu erkennen, wenn auch nicht in gleichem Ausmaß wie beim I-S-T. Eine Balance der Kategorien IvF und IvM ist weder beim I-S-T noch beim BOWIT gegeben. Sollten die

Geschlechterdifferenzen in den Gesamtscores der Tests durch eine fehlende Balance der Interessengebiete von Frauen und Männern verursacht werden, so wäre anzunehmen, dass die Leistungsunterschiede von Frauen und Männern in den drei Kategorien unterschiedlich ausfallen.

Diese Annahme wurde bestätigt. In beiden Tests zeigten sich sehr große Effektstärken in den Interessengebieten der Männer und die geringsten Effektstärken in den Interessengebieten von Frauen. Die Größe der Effektstärken der Kategorie NI lag in beiden Tests dazwischen. Es ist auffallend, dass auch in der Kategorie NI deutliche Leistungsunterschiede zugunsten von Männern auftraten. Auch bei Themen, für die Frauen und Männer durchschnittlich in etwa vergleichbare Interessen zeigen, schnitten Männer in den beiden überprüften Wissenstests deutlich besser ab. Des Weiteren ergab sich beim BOWIT sogar in der Kategorie IvF eine mittlere Effektstärke zugunsten von Männern. Beim I-S-T fiel die Effektstärke hier minimal zugunsten von Frauen aus. Es muss allerdings für beide Tests berücksichtigt werden, dass die Effektstärken der Kategorie IvF jeweils auf einer kleinen Anzahl an Items beruhen. Beim BOWIT handelte es sich um 16, beim I-S-T um ein einziges Item.

Zusammenfassend liefern die Ergebnisse dieser Untersuchung eine Unterstützung der Hypothese, dass Geschlechterdifferenzen in Wissenstests maßgeblich durch Interessenunterschiede und durch eine fehlende Balance an Items zu Interessengebieten von Frauen und Männern zustande kommen könnten. Die beiden überprüften Wissenstests sind hinsichtlich der Interessengebiete von Frauen und Männern nicht ausgewogen. Zu Themen, für die Männer sich durchschnittlich stärker als Frauen interessieren, finden sich deutlich mehr Items als zu Themen, für die Frauen sich durchschnittlich stärker als Männer interessieren. Zudem variieren die Leistungsunterschiede zwischen Frauen und Männern in den drei Kategorien IvF, NI und IvM sehr stark. Es wird ebenfalls deutlich, dass

Interessenunterschiede zwischen den Geschlechtern zwar bedeutsam, jedoch nicht die alleinige Ursache für Leistungsunterschiede sein dürften. In diesem Fall würden die Leistungsunterschiede in der Kategorie NI nahe 0 liegen und in der Kategorie IvF würden Frauen höhere Leistungen als Männer zeigen. Beides war in der hier vorgestellten Untersuchung nicht der Fall. Da die bisher erfassten Leistungsunterschiede in der Kategorie IvF allerdings bei beiden Tests auf wenigen Items beruhen, wurde in der Folgeuntersuchung der Frage nachgegangen, wie Geschlechterdifferenzen in einem umfangreichen Wissenstest ausfallen, der ausschließlich Items zu Themen enthält, für die Frauen sich durchschnittlich stärker interessieren.

5.3 Studie 2 - Entwicklung eines Wissenstests zu Interessengebieten von Frauen

Das Ziel der zweiten Untersuchung bestand in einer verlässlichen Beantwortung der Frage, in welchem Ausmaß Geschlechterdifferenzen in Wissenstests bei Fragen zu Themen bestehen, für die Frauen sich durchschnittlich stärker als Männer interessieren. Den Ergebnissen der ersten Untersuchung zufolge wären Effektstärken nahe 0 zu erwarten. Eine Replikation dieses Befundes würde die Bedeutsamkeit der Geschlechterdifferenzen in Interessen für die Leistungsunterschiede in bestehenden Wissenstests bestätigen. Gleichzeitig würde damit jedoch auch die Vermutung gestützt, dass Geschlechterdifferenzen in Interessen nicht die alleinige Erklärung für Leistungsunterschiede in Wissenstests sind. Da die Ergebnisse zu Geschlechterdifferenzen des Wissens in Themen der Kategorie IvF in der ersten Untersuchung auf einer sehr geringen Auswahl an Items aus dem BOWIT beziehungsweise dem Wissenstest des I-S-T 2000 R beruhten, ist ihre Verlässlichkeit stark eingeschränkt. Daher wurde in der zweiten Untersuchung ein neuer Wissenstest entwickelt, der ausschließlich Fragen zu Themen der Kategorie IvF enthielt, und der eine Prüfung der Geschlechterdifferenzen auf Basis einer deutlich größeren Anzahl an Items ermöglichte.

5.3.1 Methoden

Der neu entwickelte Wissenstest enthielt ausschließlich Items zu Themen, bei denen die Geschlechterdifferenzen der Interessen in der ersten Untersuchung (Kapitel 5.2) eine Effektstärke von $d > 0.10$ aufwiesen, und die damit als IvF kategorisiert wurden. Wie aus Tabelle 9 hervorgeht, handelte es sich hier um die Themen Biologie, Darstellende Kunst, Ernährung, Gesundheit, Medizin, Modedesign, Natur, Pädagogik, Psychologie, Raumdesign und Soziale Arbeit. Für alle 11 Themen wurden in einem ersten Schritt jeweils mindestens 30 Items entwickelt. Anschließend beurteilte jeweils mindestens eine Expertin oder ein Experte des jeweiligen Faches sämtliche Items eines Themas auf ihre Korrektheit und die Inhaltsvalidität. In Gesprächen und anhand von Emails gaben die Expertinnen und Experten zu den Items Rückmeldungen. Nach der Umsetzung der Hinweise, die hierbei gegeben wurden, lag ein Pool von insgesamt 348 Items vor, welche Gymnasialschülerinnen und -schülern vorgelegt wurden. Auf Basis dieses Datensatzes erfolgte die Selektion der Items, die den neuen Wissenstest bildeten. Die Auswahl erfolgte anhand der folgenden Kriterien: (a) Cronbachs α sollte in einem ähnlichen Bereich wie bei bestehenden Wissenstests liegen. (b) Zu sämtlichen Themen sollte die gleiche Anzahl an Items vorliegen. (c) Die durchschnittliche Schwierigkeit sämtlicher Items sollte in etwa .50 betragen. Zusätzlich wurden ökonomische Kriterien berücksichtigt: Der Test sollte maximal 14 Items pro Thema umfassen, da er damit insgesamt 154 Items beinhalten würde, was dem Umfang des BOWIT entsprechen würde. Nach Auswahl der Items wurden die Kriteriumsvalidität und die Leistungsunterschiede von Frauen und Männern überprüft. Zuletzt wurden die faktorielle Struktur und die Messinvarianz des Tests für die Geschlechter getestet. Im Folgenden werden sämtliche Schritte von der Itementwicklung bis zur Prüfung der Messinvarianz im Detail beschrieben.

Itementwicklung

Die Erstellung von Items erfolgte unter Verwendung von Sachbüchern, Lexika und umfangreicher Internetrecherche. Für die Themen Biologie und Pädagogik wurden Curricula der Schulfächer und für die Themen Medizin und Psychologie Curricula der Studienfächer herangezogen, um bei der Itementwicklung einen möglichst umfassenden Überblick über die jeweiligen Teilbereiche der Fächer zu haben. Gegenstand für die Generierung von Ideen für Iteminhalte waren beispielsweise Einführungstexte für die einzelnen Themen (oder Teilbereiche davon), oder die gezielte Suche nach bedeutenden Persönlichkeiten des jeweiligen Gebietes per Internet. Das Itemdesign war an das Itemdesign des Wissenstests des I-S-T 2000 R angelehnt. Auf eine Frage folgten fünf Optionen, von denen eine korrekt war. So wurde sichergestellt, dass Geschlechterdifferenzen, die mit diesem Test erfasst wurden und die eventuell von den Geschlechterdifferenzen der bestehenden Tests abweichen würden, nicht auf das Itemformat zurückgeführt werden konnten. Insgesamt beteiligten sich vier Frauen und zwei Männer an der Erstellung von Items.

Um für die einzelnen Themen die Inhaltsvalidität der Items zu überprüfen, wurden nach der Entwicklung von Itementwürfen Gespräche mit Expertinnen und Experten des jeweiligen Faches geführt. Hierbei wurden ausschließlich Personen kontaktiert, die in dem jeweiligen Gebiet erfolgreich ein Studium absolviert hatten und in diesem Bereich beruflich tätig waren. Für die Bereiche Raum- und Modedesign wurde ein Professor für Farbtechnik/ Raumgestaltung/Oberflächentechnik kontaktiert. Für die Bereiche Natur und Gesundheit wurden Expertinnen und Experten verschiedener Fachgebiete befragt, da diese beiden Begriffe weit gefasst sind. Bei dem Bereich Gesundheit handelte es sich beispielsweise um einen Arzt und einen Sportmediziner. Für den Bereich Natur wurde beispielsweise mit einem Geographen ein Gespräch geführt. In den Gesprächen mit den Expertinnen und Experten wurde die Korrektheit, Eindeutigkeit und Verständlichkeit der einzelnen Items thematisiert.

Vorschläge für Umformulierungen und die Elimination von Items wurden umgesetzt. Soweit möglich, wurde explizit die Frage gestellt, ob es einzelne Teilgebiete des jeweiligen Faches gibt, zu denen noch keine Items vorliegen, die aber von zentraler Bedeutung wären. Bei den Fächern Biologie, Raumdesign, Pädagogik und Soziale Arbeit ergaben sich hierbei auch konkrete Vorschläge für Inhalte weiterer Items, die dann realisiert wurden. Nach Abschluss der Gespräche lagen für sämtliche Themen 30 Items vor. Ausnahmen bildeten die Themen Biologie (39 Items), Pädagogik (36 Items) und Natur (33 Items). Die Expertinnen und Experten des jeweiligen Faches hatten für die Items, die letztendlich in der Datenerhebung verwendet wurden, die Frage "Wären diese Fragen geeignet, um das Wissen in diesem Fach bei einer Person zu erfassen, die keine Ausbildung hierin erhalten hat?" mit "Ja" beantwortet. Für die einzelnen Themen kann die Inhaltsvalidität daher als gegeben angesehen werden. Ausnahmen bilden hier die Themen Natur und Gesundheit, da diese Begriffe zu allgemein sind, um einer bestimmten Berufsgruppe zugeordnet zu werden. Dementsprechend konnten hier keine Expertenurteile für alle Items der beiden Themen durch einzelne Personen vorgenommen werden. Es muss daher festgehalten werden, dass die Auslegung und die konkrete Umsetzung in Iteminhalte bei den Begriffen Natur und Gesundheit subjektiv geprägt ist. Hier dienten die Gespräche mit den Expertinnen und Experten demnach auch lediglich zur Gewährleistung der Korrektheit der Items, nicht zur Überprüfung der Inhaltsvalidität.

Datenerhebung

Nach Entwicklung der Items lag ein Itempool von insgesamt 348 Items vor. Im Anschluss erfolgte eine Datenerhebung, um auf Basis von psychometrischen Kennwerten der Items eine Itemselektion für den Test durchführen zu können. Die Datenerhebung fand an zwei Gymnasien statt. Die Schülerinnen und Schüler der 10. und 11. Stufe der beiden Schulen

nahmen an der Untersuchung teil. Für eine einzelne Testung standen insgesamt 95 Minuten (zwei Schulstunden und eine dazwischenliegende 5-minütige Pause) zur Verfügung. Neben den neu entwickelten Items wurden den Personen auch ein demographischer Fragebogen und der Interessenfragebogen aus Studie 1 vorgelegt.

Der Interessenfragebogen ermöglichte die Berechnung von Zusammenhängen zwischen Interessen und Leistungen in den einzelnen Themen. Mit dem demographischen Fragebogen wurde unter anderem die Durchschnittsnote des letzten Zeugnisses erfasst, da Schulleistung ein sehr häufig verwendetes Kriterium für die Überprüfung der Kriteriumsvalidität von Wissenstests darstellt. So wird in den Manualen des I-S-T 2000 R und des BOWIT ebenfalls über Korrelationen mit Schulnoten als Evidenz für Kriteriumsvalidität berichtet (siehe Kapitel 2.3.2). Die Möglichkeit zur Berechnung der Korrelationen zwischen der Leistung in der Schule und der Leistung im vorgelegten Wissenstest war in dieser Untersuchung somit auch gegeben. Sämtliche 348 Items, die in der Datenerhebung verwendet wurden, können als Supplement vom Autor angefordert werden.

Geplante fehlende Werte

Es war davon auszugehen, dass für die Bearbeitung sämtlicher im vorigen Abschnitt beschriebener Items 95 Minuten nicht ausreichen würden. Daher wurde ein Versuchsdesign mit geplanten fehlenden Werten entworfen, das dem sogenannten 3-Form Design nach Graham, Taylor, Olchowski und Cumsille (2006) entspricht. Hierbei wurden die 348 Wissensitems in vier Testabschnitte aufgeteilt: Abschnitt X umfasste neun Items zu jedem Thema, insgesamt also 99 Items. Die übrigen Items jedes Themas wurden gleichmäßig auf die Abschnitte A, B und C verteilt. Dazu zwei Beispiele: Zur Biologie waren 39 Items entworfen worden. Neun Items wurden Abschnitt X zugeordnet, während A, B und C jeweils 10 Biologie-Items erhielten. Zur Darstellenden Kunst waren 30 Items entwickelt worden.

Auch hier wurden neun Items Abschnitt X zugeordnet. A, B und C enthielten jeweils sieben Items zur Darstellenden Kunst. Die Abschnitte A, B und C umfassten insgesamt jeweils 83 Items. Für jedes Thema erfolgte die Zuordnung der Items zu den vier Testabschnitten zufällig.

Tabelle 13

Schematische Darstellung des Versuchsdesigns mit geplanten fehlenden Werten

	Testabschnitt (Itemanzahl)			
Testversion	X (99)	A (83)	B (83)	C (83)
I	1	1	1	0
II	1	1	0	1
III	1	0	1	1

Anmerkungen. 1 = Testabschnitt war in Testversion enthalten;
0 = Testabschnitt war nicht in Testversion enthalten.

In Tabelle 13 werden die drei verschiedenen Testversionen I, II und III veranschaulicht, die in der Untersuchung verwendet wurden. Wie hier dargestellt wird, beruhten die Testversionen auf unterschiedlichen Kombinationen der Testabschnitte X, A, B und C. Alle drei Testversionen enthielten den Testabschnitt X und zwei der drei Abschnitte A, B und C. Durch die beschriebene Aufteilung der Items auf die vier Abschnitte war gewährleistet, dass sämtliche Probandinnen und Probanden die gleiche Anzahl an Items zu einem Thema bearbeiteten, unabhängig davon, welche Testversion ihnen vorgelegt wurde. Außerdem ermöglichte das 3-Form Design die Berechnungen der Korrelationen für alle Variablenpaare, was im späteren Verlauf die Verwendung von multiplen Imputationen als Schätzverfahren für die geplanten fehlenden Werte ermöglichte. Jede Person bearbeitete zunächst den demographischen Fragenbogen. Anschließend wurden die 265 Wissensitems aus einer der

drei Testversionen bearbeitet. Zuletzt machten die Teilnehmenden Angaben zu ihren Interessen.

Stichprobenbeschreibung

An den beiden Schulen wurden insgesamt 281 Personen in 14 Gruppen getestet. In jeder Gruppe wurden die drei Versionen des Wissenstests zufällig unter den Teilnehmenden verteilt. Die Daten von 33 Personen erwiesen sich als nicht verwertbar. Der häufigste Grund hierfür waren weniger als 10 Jahre Erfahrung mit der deutschen Sprache. Daneben gab es auch Teilnehmende, deren Antworten auf einzelne Fragen eindeutig auf eine nicht ernsthafte Bearbeitung des Testmaterials hinwiesen, wie beispielsweise "3 Jahre" als Altersangabe oder "Klingonisch" als Muttersprache. Der Datensatz, der im weiteren Verlauf für geschlechtsunspezifische Analysen verwendet wurde, umfasste die Daten von 248 Personen. Da eine Person keine Angabe zu ihrem Geschlecht gemacht hatte, umfasste der Datensatz für geschlechtsspezifische Analysen n = 247 Personen (57.1% Frauen). Das Durchschnittsalter der Frauen betrug 16.62 Jahre mit einem Median von 17.00 und einer Standardabweichung von 0.94. Die Männer waren durchschnittlich 16.72 Jahre alt, Median und Standardabweichung des Alters waren identisch mit den Werten der Frauen.

Multiple Imputation

Da sämtliche Probandinnen und Probanden der Voruntersuchung lediglich 265 von 348 Items bearbeitet hatten, war im Vorfeld der Itemanalysen eine Aufbereitung des Datensatzes erforderlich. Die geplanten fehlenden Werte wurden durch multiple Imputationen geschätzt. Gegenstand der Imputationen war dabei lediglich die Kategorisierung der Items als korrekt oder falsch, nicht die jeweilige Auswahl unter den fünf Optionen.

Fehlende Werte können als missing not at random (MNAR) missing at random (MAR) und missing completely at random (MCAR) klassifiziert werden (Graham et al., 2006).

„Statistically, the most problematic type of missing data are missing not at random (MNAR), or nonignorably missing data, for which missingness is related to the value that would have been observed“ (Sinharay, Stern & Russell, 2001, S. 318). Die Autoren geben folgendes Beispiel für MNAR: Sofern Personen mit hohem oder niedrigem Einkommen dazu tendieren, auf die Frage nach dem Einkommen nicht zu antworten, fallen die hierdurch entstehenden fehlenden Werte in die Kategorie MNAR. Die Kategorie MAR liegt vor, wenn fehlende Werte nicht aufgrund der Variablen, die gemessen werden soll, aber aufgrund anderer Störvariablen zustande kommen. Ein Beispiel für MAR würde vorliegen, wenn fehlende Werte bei der Frage nach dem Einkommen dadurch entstehen, dass Personen mit geringerer Bildung dazu tendieren, keine Information über ihr Einkommen zu geben (Sinharay et al., 2001). Bei Vorliegen von MCAR können Einflüsse eventueller Störvariablen, die zu fehlenden Werten führen würden, ausgeschlossen werden. Die Ursachen, die zu dem geplanten Fehlen des Wertes einer Variable führen, sind im vorliegenden Fall (a) die Zuordnung des Items zu einem der drei Testabschnitte und (b) der Umstand, dass dieser Abschnitt der Person nicht vorliegt. Da innerhalb der Themen die Zuordnung der Items zu den Testabschnitten A, B und C zufällig erfolgte und da die Verteilung der Testhefte an die Probanden zufällig erfolgte, können die geplanten fehlenden Werte der vorliegenden Studie als missing completely at random (MCAR) kategorisiert werden.

Fehlende Werte, die durch Nicht-Bearbeitung eines Items entstanden, das der Person vorlag, fallen in die Kategorie MNAR. Diese Items können dennoch eindeutig als falsch kategorisiert werden, da hier davon auszugehen ist, dass sie aufgrund fehlenden Wissens nicht beantwortet wurden.

Graham et al. (2006) empfehlen neben dem Maximum Likelihood Verfahren die multiple Imputation als Schätzverfahren für fehlende Werte, die als MCAR kategorisiert werden können. Maximum Likelihood ist als Schätzverfahren in diesem Fall ungeeignet, da hierbei

eine multivariate Normalverteilung der Variablen vorausgesetzt wird. Bei den Items des Wissenstests handelt es sich jedoch um kategoriale Variablen, wodurch die Voraussetzung der Normalverteilung nicht erfüllt ist. Daher empfiehlt sich hier die multiple Imputation als Schätzverfahren.

Die multiplen Imputationen wurden gesondert für die Items der 11 einzelnen Themen durchgeführt. Die Modelle, die den Imputationen zugrunde lagen, beinhalteten sämtliche Items des jeweiligen Themas und von den übrigen Themen jeweils vier Items aus jedem bearbeiteten Testabschnitt. Zur Verdeutlichung dient das folgende Beispiel: Das Modell der multiplen Imputationen der Psychologieitems aus Testabschnitt C enthielt sämtliche Psychologieitems aus den Abschnitten X, A, B und C (wobei letztere imputiert wurden). Zusätzlich wurden jeweils 4 Biologieitems aus den Testabschnitten X, A und B in das Modell aufgenommen. Auch für die übrigen neun Themen wurden aus den Testabschnitten X, A und B jeweils 4 Items herangezogen. Das Modell enthielt somit alle 30 Psychologieitems (von denen 7 imputiert wurden) sowie 10 (Themen) × 4 (zufällig ausgewählte Items) × 3 (Testabschnitte X, A und B) Items der übrigen Themen – insgesamt also 150 Variablen. Sämtliche anderen Modelle waren analog konzipiert. Die Anzahl der Variablen variierte nur durch die Anzahl der Items, die zu einem Thema entwickelt worden waren. So enthielt ein Modell für die Imputationen der Biologieitems beispielsweise 159 Variablen, da 39 Biologievariablen in der Testung verwendet wurden. Die zufällige Auswahl der vier Items zu jedem Thema aus jedem Testabschnitt erfolgte nur einmal und wurde nicht für jedes Model erneut getroffen. Für die Imputationen der Psychologieitems aus Testabschnitt C wurden also beispielsweise die gleichen 12 Biologieitems verwendet wie für die Imputationen der Pädagogikitems aus dem Testabschnitt.

Auf diesem Weg wurden für alle Items jeweils fünf Imputationen mit Mplus 7.1 durchgeführt. Als Schätzverfahren wurde weighted least squares with means and variance

adjusted (WLSMV) verwendet. Die fünf einzelnen Imputationen, die anschließend für jedes Item vorlagen, wurden jeweils mit den gemessenen Werten zu einem Datensatz zusammengefasst. So ergaben sich fünf vollständige Datensätze, die für jede Person 265 gemessene Werte und 83 imputierte Werte enthielten.

Um die Stabilität der Ergebnisse anschließender Datenanalysen zu überprüfen, wurden in einem zweiten Durchgang wiederum zufällige Auswahlen von vier Items zu jedem Thema und aus jedem Testabschnitt getroffen. Die Zusammenstellung von Items zu Modellen wurde erneut auf die beschriebene Weise durchgeführt, und sämtliche fehlenden Werte wurden erneut fünf Mal imputiert. Ein weiteres Mal wurden gemessene und imputierte Werte zu fünf vollständigen Datensätzen zusammengefasst. Insgesamt lagen somit 10 Datensätze vor, die im Folgenden mit den Zahlen 1.1 bis 1.5 beziehungsweise 2.1 bis 2.5 bezeichnet werden. Eine Liste, welche die zufällig ausgewählten Items für die Imputationen beinhaltet, kann vom Autor als Supplement angefordert werden.

Itemselektion

Die Itemselektion erfolgte nach psychometrischen Kriterien. Der im Folgenden beschriebene Prozess wurde mit dem Datensatz 1.1 durchgeführt, der das Ergebnis der ersten Imputationen war. Nach Abschluss der Itemselektion wurden die Item- und Skalenkennwerte des Tests auch mit den übrigen neun Datensätzen berechnet, um die Stabilität der Ergebnisse zu überprüfen.

Zunächst wurde separat für alle 11 Themen jeweils eine Selektion an Items durchgeführt, wobei hauptsächlich die Trennschärfen der Items ausschlaggebend waren. In mehreren Schritten wurden die Items eines Themas, beginnend mit den Items mit der niedrigsten Trennschärfe, eliminiert. Nach jeder Elimination von einem oder mehreren Items wurden die Trennschärfen der Items der aktuellen Skala neu berechnet. Ein fester Cut-Off-Wert der

Trennschärfe wurde hierbei nicht gesetzt, da ein wichtiges Ziel die gleiche Anzahl an Items zu jedem Thema war, die Trennschärfen der Items zwischen den einzelnen Themen jedoch deutlich variierten (siehe Tabelle G1). Ein weiteres Ziel der Itemselektion bestand in der Zusammenstellung von themenspezifischen Skalen, deren Items eine durchschnittliche Schwierigkeit von etwa .50 aufwiesen. Nachdem auf Basis der Trennschärfe zu jedem Thema 14 Items ausgewählt waren, wurden daher einzelne Items gegen schwierigere beziehungsweise leichtere Items ausgetauscht, je nachdem, ob die durchschnittliche Schwierigkeit der aktuellen Auswahl an Items größer oder kleiner als .50 war. Sofern notwendig, wurden hierbei auch Items mit geringerer Trennschärfe auf der jeweiligen Skala zugelassen. Nach Abschluss der Auswahl von 14 Items für jedes Thema entsprach die Länge des Tests der Länge des BOWIT.

Aus ökonomischen Gründen wurde der Test anschließend weiter gekürzt. Um die thematische Ausgewogenheit des Tests beizubehalten, wurde jeweils ein Item aus jedem Thema eliminiert. Auch hierbei gaben die Schwierigkeiten und Trennschärfen der Items auf themenspezifischer Ebene den Ausschlag. Wie zu Beginn des Abschnitts 5.3.1 beschrieben, bestand ein wichtiges Ziel der Testentwicklung in einem Wert für Cronbachs α, der mit den Werten bestehender Wissenstests vergleichbar war. Daher wurde nach jeder Elimination von 11 Items erneut Cronbachs α für den gesamten Test berechnet. Dieser Prozess wurde drei Mal wiederholt. Weitere Kürzungen führten zu Werten in Cronbachs α, die unter .90 lagen, weshalb keine weiteren Items entfernt wurden. Die endgültige Version des Tests enthielt somit 11 Items pro Thema, insgesamt also 121 Items.

Korrelation mit Schulleistung

Im Anschluss erfolgte die Überprüfung der Kriteriumsvalidität des Tests. Hierzu wurden für den Gesamtscore und für sämtliche Skalen des Tests (auf Basis der Itemauswahl aus den

vorherigen Schritten) die Korrelationen mit der Durchschnittsnote des letzten Zeugnisses berechnet. Hierbei handelte es sich um das Halbjahreszeugnis, das die Schülerinnen und Schüler im Winter 2012/2013 erhalten hatten. Die Angaben zur Durchschnittsnote konnten in Form von Noten oder Punkten erfolgen. Angaben in Noten wurden später bei der Aufbereitung der Daten in Punkte umgerechnet.

Geschlechterdifferenzen

Nach Abschluss der Testentwicklung wurden für sämtliche Teilnehmenden die Summenscores für die einzelnen Themen und für den gesamten Test berechnet. Durch *t*-Tests erfolgte eine Überprüfung der Leistungen von Frauen und Männern auf Mittelwertunterschiede. Diese wurden auch in Form von Cohens *d* berechnet, wobei die unterschiedlichen Standardabweichungen der Geschlechter durch die Größen der Teilstichproben gewichtet wurden. Die Verwendung des Interessenfragebogens, der in der vorhergehenden Studie entwickelt wurde, ermöglichte außerdem, die Leistungsunterschiede zu den Interessenunterschieden für die verschiedenen Themen in Beziehung zu setzen.

Faktorenstruktur

Zuletzt wurden die faktorielle Struktur und die Messinvarianz des Tests überprüft. Die manifesten Variablen sämtlicher Faktorenanalysen waren die 11 Summenscores der Skalen der einzelnen Themen. Für alle 10 Datensätze wurden die entsprechenden 11 Summenscores berechnet. Da anhand des neu entwickelten Tests das gleiche Merkmal wie mit bestehenden Wissenstests erfasst werden sollte – die Fähigkeit des Allgemeinen Wissens – wurde eine 1-faktorielle Struktur angenommen, die durch konfirmatorische Faktorenanalysen bei sämtlichen Datensätzen überprüft wurde. Es sei an dieser Stelle jedoch an die Faktorenanalysen anderer Tests des Allgemeinen Wissens erinnert, bei denen nicht auf einen einzelnen Faktor des Allgemeinen Wissens geschlossen werden konnte: Wie in Kapitel 2.3.2

berichtet, wies der BOWIT in einer exploratorischen Faktorenanalyse eine 2-Faktorenlösung auf. Ein Faktor wurde von Hossiep und Schulte (2008) als gesellschafts- und geisteswissenschaftliches Wissen und der andere als naturwissenschaftlich-technisches Wissen interpretiert. In Kapitel 3.1.1 wurde über Analysen des GKT (Irwing, et al., 2001) von Lynn und Irwing (2002) berichtet. Die Autoren entwickelten ein hierarchisches Modell, das sechs Faktoren erster und einen Faktor zweiter Ordnung enthielt. Die Überprüfung in Form eines MIMIC-Modells für Frauen und Männer erforderte neben Ladungen des Geschlechts auf dem Faktor zweiter Ordnung auch direkte Ladungen auf drei der sechs Faktoren erster Ordnung, um eine zufriedenstellende Passung aufzuweisen. Derartige Ergebnisse führen zu Schwierigkeiten in der Interpretation von Geschlechterdifferenzen und zu der Frage, ob das Konstrukt Allgemeines Wissen möglicherweise nicht in Form einer einzelnen, sondern angemessener durch mehrere latente Variablen dargestellt werden kann. Möglicherweise kann zwischen verschiedenen Arten von Wissen im Sinne von Wissen in unterschiedlichen Themengebieten differenziert werden, wobei die Geschlechterdifferenzen hier unterschiedlich ausfallen mögen. Daher wurde für alle Datensätze jeweils eine Parallelanalyse berechnet, um die Anzahl der latenten Variablen für eventuelle alternative Messmodelle zu bestimmen. In jeder Parallelanalyse wurde auf Basis von 1000 simulierten Datensätzen der Mittelwert und das 95%-Quantil der Eigenwerte der einzelnen Faktoren berechnet. Hierbei wurde die von O'Connor (2000) erstellte Syntax mit SPSS 22 verwendet, welche die Generierung von Daten ermöglicht, deren Verteilung den empirischen Daten entspricht. Da sich bei einzelnen Analysen Hinweise auf eine 2-faktorielle Struktur fanden, wurden für die entsprechenden Datensätze exploratorische Faktorenanalysen durchgeführt, um auf Basis der Faktorladungen die manifesten Variablen den latenten Variablen zuzuordnen. Die exploratorischen Faktorenanalysen wurden mit Mplus 7.1 durchgeführt. Für die Rotation wurde das oblique Oblimin-Verfahren verwendet, was Korrelationen zwischen

den Faktoren ≠ 0 ermöglichte. Hieraus ergab sich ein 2-faktorielles Messmodell, das im Anschluss mit sämtlichen Datensätzen durch konfirmatorische Faktorenanalysen getestet wurde.

Wie in Kapitel 3.2.1 beschrieben, ist die Messinvarianz eine Voraussetzung, um von Gruppenunterschieden in Testscores auf Unterschiede in den gemessenen Merkmalen schließen zu können. Die Überprüfung der Messinvarianz des für diese Studie entwickelten Tests erfolgte in Anlehnung an ein von Lubke et al. (2003) beschriebenes Verfahren für Mehr-Gruppen-Modelle, bei dem in zwei Schritten unterschiedliche Parameter des Messmodells gleichgesetzt werden, "in order to distinguish between the causes of misfit" (S. 554). In beiden Schritten wurden die Varianzen der Faktoren in beiden Gruppen auf 1 fixiert. Die Varianzen der Residuen wurden gleichgesetzt und deren Ladungen auf 1 fixiert. Die Erwartungswerte der Residuen und die Kovarianzen zwischen ihnen wurden auf 0 gesetzt. Die Korrelationen der Faktoren und die Faktorladungen wurden gleichgesetzt. Die beiden Schritte des Verfahrens von Lubke et al. unterscheiden sich lediglich hinsichtlich der Restriktionen für die Achsenabschnitte der manifesten Variablen und für die Erwartungswerte der Faktoren. Im ersten Schritt wurden die Erwartungswerte der Faktoren für alle Gruppen auf 0 fixiert. Die Achsenabschnitte wurden für die Gruppen frei geschätzt. Dieses Modell diente zur Überprüfung der Frage, ob Unterschiede *innerhalb* der Gruppen jeweils auf die gleichen Faktoren zurückgeführt werden konnten (Lubke et al., 2003). Im Anschluss erfolgte der zweite Schritt. Hierbei wurden die Achsenabschnitte für die Gruppen gleichgesetzt. Die einzigen Parameter, die sich zwischen den Gruppen unterscheiden konnten, waren hier somit die Erwartungswerte der Faktoren. Diese wurden im zweiten Schritt bei einer Gruppe auf 0 fixiert und in der anderen Gruppe frei geschätzt. Es dienten hier also ausschließlich Gruppenunterschiede auf latenter Ebene als Erklärung für eventuelle Gruppenunterschiede auf manifester Ebene. Das Modell des zweiten Schrittes diente somit

zur Überprüfung der Frage, ob Unterschiede *zwischen* den Gruppen auf die gleichen Faktoren zurückgeführt werden konnten wie Unterschiede *innerhalb* der Gruppen (Lubke et al., 2003). Sofern auch dieses Modell eine zufriedenstellende Passung aufweist, kann die Messinvarianz als gegeben angenommen werden. Sämtliche konfirmatorischen Faktorenanalysen wurden mit Mplus 7.1 und unter Verwendung der mean-adjusted version des Maximum Likelihood Schätzers (MLM) durchgeführt, der bereits in der Untersuchung zur Selbsteinschätzung (siehe Kapitel 4.2.2) Verwendung fand. Außerdem wurde auch hier die von Satorra und Bentler (zitiert nach Muthén, 1998-2004) vorgeschlagene Korrektur der χ^2-Statistik eingesetzt.

5.3.2 Ergebnisse

Im Folgenden werden zunächst die Ergebnisse der Itemselektion berichtet. Für den Wissenstest, der hieraus entstand, werden anschließend die Ergebnisse der Prüfung der Kriteriumsvalidität und die berechneten Geschlechterdifferenzen dargestellt. Zuletzt erfolgt eine ausführliche Beschreibung der Ergebnisse der verschiedenen Analysen zur Überprüfung der Faktorenstruktur des Wissenstests.

Itemselektion

Nach der schrittweisen Reduktion der Itemanzahl auf 14 Items pro Thema und der anschließenden Elimination von 33 weiteren Items betrug Cronbachs α für den gesamten Test im ersten Datensatz .91. Die Cronbachs α-Werte der einzelnen Skalen betrugen im Durchschnitt .55. Tabelle 14 enthält für sämtliche Skalen die Minima, Maxima und Mittelwerte der Schwierigkeiten und der korrigierten Trennschärfen sowie Cronbachs α nach Abschluss der Itemselektion.

Die Item- und Skalenkennwerte wurden mit den übrigen neun Datensätzen erneut berechnet. Für alle 10 Datensätze liegen Tabellen vor, welche die Schwierigkeiten und

Trennschärfen sämtlicher Items vor Beginn und nach Abschluss der Itemselektion enthalten. Diese Tabellen können vom Autor als Supplement angefordert werden. Drei der beibehaltenen Items wiesen in jeweils einem der 10 Datensätze eine geringe negative Trennschärfe auf der jeweiligen themenspezifischen Skala auf, was jedoch nicht als hinreichend gewichtig erschien, um weitere Änderungen in der Itemauswahl zu treffen. In den Tabellen G1 bis G10 finden sich für alle 10 Datensätze die Cronbachs α-Werte der einzelnen Skalen sowie die durchschnittlichen Schwierigkeiten und korrigierten Trennschärfen der Items vor Beginn und nach Beendigung der Itemselektion.

Tabelle 14

Item- und Skalenkennwerte des neuen Wissenstests (Studie 2, Datensatz 1.1)

	Schwierigkeit			Korr. Trennschärfe			
Thema	Min	Max	*M*	Min	Max	*M*	α
Biologie	.35	.78	.56	.29	.48	.39	.75
Darstellende Kunst	.21	.81	.49	.11	.41	.19	.48
Ernährung	.21	.88	.51	.13	.27	.20	.50
Gesundheit	.21	.90	.51	.19	.28	.23	.55
Medizin	.32	.83	.52	.15	.42	.29	.63
Modedesign	.30	.72	.50	.10	.38	.24	.56
Natur	.33	.69	.52	.11	.38	.23	.55
Pädagogik	.31	.90	.62	.13	.30	.22	.52
Psychologie	.24	.67	.41	.07	.38	.20	.50
Raumdesign	.26	.78	.49	.13	.25	.17	.45
Soziale Arbeit	.23	.82	.49	.12	.30	.21	.51
Gesamt	.21	.90	.51	.07	.50	.26	.91

Anmerkungen. Korr. Trennschärfe = Trennschärfeberechnungen nach part-whole-Korrektur; Min = Minimum; Max = Maximum; α = Cronbachs α.

Es zeigte sich, dass mit der getroffenen Auswahl an Items bei allen Datensätzen die für die Itemselektion gesetzten Ziele erreicht wurden. Die endgültige Auswahl, die aus 121 Items

bestand, ergab in allen 10 Berechnungen ein Cronbachs α von .90 oder .91 für den gesamten Test. Die durchschnittlichen Schwierigkeiten aller Items des Tests betrugen in allen Datensätzen .51.

Korrelationen mit Schulleistung

Nach der Itemselektion wurden für alle zehn Datensätze die Korrelationen zwischen der Schulleistung (in Punkten) einerseits und den 11 einzelnen Skalen sowie dem Gesamtscore des Wissenstests andererseits berechnet. 195 Personen (57,9 % Frauen) hatten Angaben zu der Durchschnittsnote ihres letzten Zeugnisses gemacht. Tabelle 15 enthält für jede Skala das Minimum, das Maximum und den Mittelwert der Korrelationen aus den 10 Datensätzen.

Tabelle 15

Korrelationen zwischen Skalen des Wissenstests und Schulleistung in Punkten aus allen 10 Datensätzen

Thema	Minimum	Maximum	Mittelwert
Biologie	.27	.34	.30
Darstellende Kunst	.24	.30	.27
Ernährung	.15	.20	.18
Gesundheit	.27	.31	.29
Medizin	.16	.24	.19
Modedesign	.23	.31	.26
Natur	.16	.22	.18
Pädagogik	.28	.32	.30
Psychologie	.20	.27	.23
Raumdesign	.20	.25	.22
Soziale Arbeit	.20	.27	.24
Gesamtscore	.35	.38	.37

Es zeigten sich erwartungsgemäß ausschließlich positive Korrelationen, wobei im Durchschnitt die Zusammenhänge mit den Summenscores der Skalen Ernährung und Natur

(.18) sowie Medizin (.19) am niedrigsten ausfielen. Die höchsten durchschnittlichen Zusammenhänge bestanden mit dem Gesamtscore (.37) und mit den Skalen Biologie und Pädagogik (.30).

Geschlechterdifferenzen

Für den fertigen Test erfolgten die Berechnungen der Geschlechterdifferenzen der Summenscores für die einzelnen Skalen und für die Gesamtskala. In Tabelle 16 sind die Effektstärken der Geschlechterdifferenzen und die Ergebnisse der *t*-Tests am Beispiel des Datensatzes 2.3 dargestellt. Das Signifikanzniveau war auf 5% festgelegt. Da 12 Tests durchgeführt wurden, erfolgte eine Bonferroni-Korrektur auf 0.4%. Unter allen Levene-Tests auf Homoskedastizität lieferte der Test der Skala Gesundheit mit $p = .035$ den niedrigsten *p*-Wert. Für sämtliche *t*-Tests konnte daher Homoskedastizität angenommen werden. Die *t*-Werte beziehen sich somit alle auf die *t*-Verteilung mit 245 Freiheitsgraden.

Mit Ausnahme des Themas Biologie wiesen die Effektstärken der Interessenunterschiede in die gleiche Richtung wie in der vorangegangenen Untersuchung. Für alle Themen bis auf die Biologie, bei der nahezu kein Effekt festgestellt wurde, zeigten Frauen ein größeres Interesse. Unter diesen Themen wurde die kleinste Effektstärke der Interessenunterschiede für Medizin festgestellt ($d = 0.15$), während sie beim Modedesign am höchsten ausfiel ($d = 1.40$). Die Effektstärken der Leistungsunterschiede variierten sehr stark. Für Ernährung, Gesundheit und Raumdesign zeigten sich geringe Leistungsunterschiede zwischen den Geschlechtern. Bei den Themen Biologie, Medizin, Natur, Psychologie, Soziale Arbeit und im Gesamtscore schnitten die Männer besser ab, wobei die Effektstärken hier von -0.11 (Gesamtscore) bis -0.60 (Medizin) reichten. In den Themen Darstellende Kunst, Modedesign und Pädagogik erreichten Frauen im Durchschnitt höhere Summenscores. Die Effektstärken

lagen hier zwischen 0.19 (Darstellende Kunst) und 0.61 (Modedesign). Die Effektstärken der Interessen und der Summenscores im Wissenstest korrelierten miteinander zu .53.

Tabelle 16

Geschlechterdifferenzen der Interessen und Leistungen

	Effektstärke Interessen	Effektstärke Leistung	*t*-Test Leistung	
Thema	Cohens *d*	Cohens *d*	t_{245}	*p* (2-seitig)
Biologie	-0.05	-0.34	-2.67	.008
Darstellende Kunst	0.78	0.19	1.48	.139
Ernährung	0.34	0.02	0.14	.891
Gesundheit	0.32	0.10	0.81	.421
Medizin	0.15	-0.60	-4.64	< .001
Modedesign	1.40	0.61	4.72	< .001
Natur	0.22	-0.15	-1.19	.234
Pädagogik	0.87	0.24	1.84	.068
Psychologie	0.40	-0.32	-2.45	.015
Raumdesign	0.94	0.07	0.54	.592
Soziale Arbeit	1.12	-0.57	-4.43	< .001
Gesamtscore		-0.11	-0.84	.400

Anmerkung. Effektstärken und *t*-Werte mit negativem Vorzeichen weisen auf höhere Durchschnittswerte bei Männern hin.

Die *t*-Tests der leistungsbezogenen Mittelwertdifferenzen für die Skalen Medizin, Modedesign und Soziale Arbeit ergaben *p*-Werte, die .004 unterschritten. In den Bereichen Medizin und Soziale Arbeit zeigten Männer im Durchschnitt die besseren Leistungen. Im Bereich Modedesign erreichten Frauen die durchschnittlich höheren Scores. In den Tabellen H1 bis H9 sind die Effektstärken der Geschlechterdifferenzen und die Ergebnisse der *t*-Tests für die übrigen 9 Datensätze dargestellt. Die Ergebnisse fielen hier insgesamt vergleichbar

aus. In fünf Datensätzen wurden zusätzlich in den *t*-Tests für das Thema Psychologie *p*-Werte berechnet, die .004 unterschritten.

Faktorenstruktur

Die Überprüfung der faktoriellen Struktur des neu entwickelten Tests erfolgte zunächst anhand konfirmatorischer Faktorenanalysen mit Messmodellen, bei denen sämtliche manifesten Variablen auf demselben Faktor luden. Diese Faktorenanalysen wurden mit allen zehn Datensätzen durchgeführt. Im Folgenden werden ausschließlich die Ergebnisse der Analysen eines zufällig ausgewählten Datensatzes beschrieben. Hierbei handelt es sich um den Datensatz 2.3. In Tabelle 17 sind die wichtigsten Ergebnisse des 1-faktoriellen Modells dargestellt.

Tabelle 17

Auszug aus standardisierten Parameterschätzungen und Model-Fit der 1-faktoriellen Messmodelle (Datensatz 2.3)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Wissen	0[a]		-0.16	0[a]
[95%-KI]			[-0.42; 0.11]	
Achsenabschnitte				
Biologie	2.07	2.42	2.25	
Darstellende Kunst	2.78	2.59	2.74	
Ernährung	2.69	2.68	2.75	
Gesundheit	2.65	2.55	2.67	
Medizin	2.43	3.02	2.64	
Modedesign	2.79	2.18	2.47	
Natur	2.47	2.62	2.59	
Pädagogik	3.36	3.12	3.28	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Psychologie	1.98	2.30	2.16	
Raumdesign	2.61	2.54	2.63	
Soziale Arbeit	2.66	3.23	2.85	
Model-Fit				
χ^2_{df}	156.28_{110}		291.32_{120}	
p	.003		< .001	
SCF	0.97		0.97	
CFI	.95		.82	
SRMR	.08		.11	
RMSEA	.06		.11	
[90%-KI]	[.04; .08]		[.09; .12]	
χ^2-Diff-Test: χ^2_{df}; p	132.37_{10}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.
[a]Parameter wurde vorab fixiert.

Die Modellpassung fiel hier bei Gleichsetzung der Erwartungswerte des Faktors für Frauen und Männer und freier Schätzung der Achsenabschnitte für beide Gruppen (Schritt 1) mit CFI = .95 und RMSEA = .06 gut aus. Im Anschluss wurden die Achsenabschnitte der manifesten Variablen für Frauen und Männer gleichgesetzt, während der Erwartungswert des Faktors lediglich für die Männer auf 0 fixiert wurde (Schritt 2). Bei den Frauen wurde der Erwartungswert auf -0.16 geschätzt. Da die Varianz des Faktors auf 1 fixiert war, kann die Differenz der Schätzungen der Erwartungswerte auch als Effektstärke interpretiert werden. Das Modell des zweiten Schrittes wies mit CFI = .82 und RMSEA=.11 eine deutlich schlechtere Passung auf als das Modell, das in Schritt 1 geprüft wurde, was sich auch in dem χ^2-Test auf die Differenz der Passung beider Modelle bestätigte (p < .001). In den Tabellen I1

bis I10 sind die Ergebnisse der Messmodelle für alle 10 Datensätze ausführlich dargestellt. Die Ergebnisse fielen ähnlich aus: Der Erwartungswert der latenten Variablen wurde im zweiten Schritt bei Frauen durchgehend kleiner -0.128 geschätzt. Das 95%-Konfidenzintervall umfasste bei allen Datensätzen den Wert 0. Die Differenz der Passungen der beiden Modelle aus Schritt 1und Schritt 2 war jeweils auf dem 1‰-Niveau signifikant.

Aufgrund der widersprüchlichen Ergebnisse zur faktoriellen Struktur von anderen Tests des Allgemeinen Wissens (siehe Kapitel 5.3.1) wurden im Anschluss für sämtliche Datensätze Parallelanalysen durchgeführt, um die Anzahl der Faktoren des neu entwickelten Wissenstests zu prüfen. Auch die unzureichende Passung des 1-faktoriellen Messmodells wies auf die Notwendigkeit hin, die Anzahl der Faktoren zu überprüfen. In den Tabellen J1 bis J10 sind die Ergebnisse aller Parallelanalysen ausführlich dargestellt. Acht der 10 Analysen wiesen entweder auf eine 1- oder eine 2-faktorielle Struktur hin. Bei zwei Datensätzen fiel auch der Mittelwert der Eigenwerte des dritten Faktor kleiner als der Eigenwert des dritten Faktors der Rohdaten aus, was jedoch nicht als hinreichender Grund betrachtet wurde, eine 3-faktorielle Struktur zu überprüfen. Hinsichtlich des zweiten Faktors waren die Ergebnisse jedoch nicht eindeutig. Das 95%-Quantil der Eigenwerte des zweiten Faktors der simulierten Daten lag bei sieben der 10 Analysen oberhalb des Eigenwertes des zweiten Faktors der Rohdaten, was auf eine 1-faktorielle Struktur hindeutete. Bei sieben der 10 Parallelanalysen fiel jedoch der Mittelwert der Eigenwerte des zweiten Faktors auf Basis der simulierten Daten geringer aus als der Eigenwert des zweiten Faktors der Rohdaten. In drei dieser sieben Analysen lag auch das 95%-Quantil der Eigenwerte unter dem Eigenwert des zweiten Faktors auf Basis der Rohdaten. Tabelle 18 enthält die Ergebnisse der zehn Parallelanalysen für den zweiten Faktor.

Tabelle 18

Ergebnisse der Parallelanalysen der zehn Datensätze für den zweiten Faktor

Datensatz	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1.1	0.21	0.29	0.37
1.2	0.21	0.29	0.37
1.3	0.30	0.30[a]	0.37
1.4	0.30	0.29[a]	0.37
1.5	0.35	0.29[a]	0.37
2.1	0.29	0.29	0.37
2.2	0.40	0.29[a]	0.37[a]
2.3	0.34	0.30[a]	0.37
2.4	0.39	0.29[a]	0.37[a]
2.5	0.39	0.30[a]	0.38[a]

[a]Wert ist kleiner als der Eigenwert des aus den Rohdaten stammenden Faktors.

Für die Datensätze 2.2, 2.4 und 2.5 wurden daher im Anschluss exploratorische Faktorenanalysen berechnet, bei denen die Anzahl der Faktoren auf 2 fixiert war. Auf Basis der berechneten Faktorladungen wurde versucht, jede manifeste Variable einem der beiden Faktoren zuzuordnen. Die Tabelle K1 enthält die Faktorladungen und die Korrelationen zwischen den Faktoren, die sich bei den drei exploratorischen Faktorenanalysen nach der Oblimin-Rotation ergaben. Die manifesten Variablen Biologie, Medizin, Natur und Soziale Arbeit konnten eindeutig einem gemeinsamen Faktor zugeordnet werden, da sie bei allen drei Datensätzen nur auf einem gemeinsamen Faktor Ladungen aufwiesen, die auf dem 5%-Niveau signifikant waren. Dieser Faktor wird im Folgenden mit Wissen1 bezeichnet. Auf die gleiche Weise konnten die Variablen Darstellende Kunst, Gesundheit, Pädagogik und Raumdesign dem anderen Faktor zugeordnet werden, der im weiteren Verlauf mit Wissen2 bezeichnet wird. Auch die Variable Modedesign konnte eindeutig dem Faktor Wissen2

zugewiesen werden, da sie in zwei Datensätzen ausschließlich auf diesem Faktor eine signifikante Ladung aufwies und im dritten Datensatz auf Wissen2 eine signifikante positive und auf Wissen1 eine signifikante negative Ladung zeigte. Für die Variablen Ernährung und Psychologie war keine eindeutige Zuordnung möglich. Beide Variablen zeigten bei zwei der drei Analysen signifikante Ladungen auf beiden Faktoren. Wie aus Tabelle K1 hervorgeht, lagen die Schätzungen der Korrelationen zwischen beiden Faktoren zwischen .62 und .75. Aus den Ergebnissen der exploratorischen Faktorenanalysen ergab sich ein 2-faktorielles Messmodell, das anschließend mit allen zehn Datensätzen durch konfirmatorische Faktorenanalysen geprüft wurde. Dieses Messmodell ist in Abbildung 6 dargestellt.

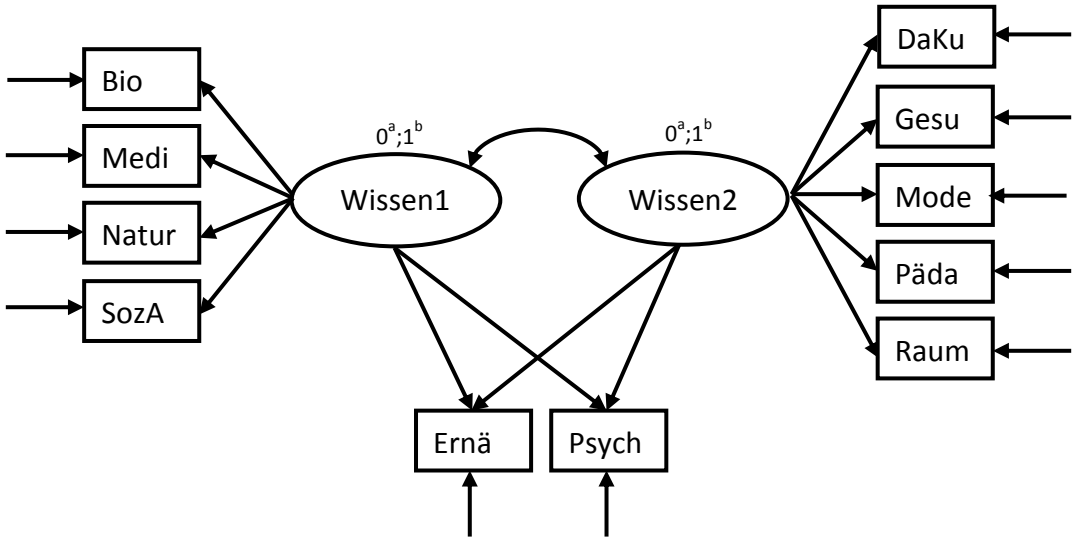

Abbildung 6. Messmodell mit zwei Faktoren – Studie 2. Bio = Biologie; Medi = Medizin; SozA = Soziale Arbeit; DaKu = Darstellende Kunst; Gesu = Gesundheit; Mode = Modedesign; Päda = Pädagogik; Raum = Raumdesign; Ernä = Ernährung; Psych = Psychologie.

[a]Erwartungswert in Schritt 1 in beiden Gruppen fixiert. In Schritt 2 in einer zufällig ausgewählten Gruppe fixiert, in der anderen Gruppe frei geschätzt.

[b]Varianz in Schritt 1 und in Schritt 2 in beiden Gruppen fixiert.

Von den Ergebnissen der konfirmatorischen Faktorenanalysen, mit denen die 2-faktorielle Struktur geprüft wurde, werden hier nur diejenigen des Datensatzes 2.3

beschrieben, dessen Ergebnisse im Rahmen der 1-faktoriellen konfirmatorischen Faktorenanalysen ebenfalls ausführlich beschrieben wurden (siehe Tabelle 17). Ein Auszug aus den standardisierten Parameterschätzungen der 2-faktoriellen Faktorenanalysen ist in Tabelle 19 dargestellt. Die Korrelation zwischen beiden Faktoren betrug im ersten Schritt .90. Die Passung des Modells war mit CFI = .96 und RMSEA = .05 gut, womit die Annahme, dass die Unterschiede innerhalb der Gruppen auf die gleichen Faktoren zurückgeführt werden können, als bestätigt angesehen werden kann. Im zweiten Schritt korrelierten die Faktoren zu .93. Die Erwartungswerte beider Faktoren waren für die Gruppe der Männer auf 0 fixiert. Die Schätzungen bei den Frauen betrugen -0.66 bei Wissen1 und 0.43 bei Wissen2. Die Varianzen beider Faktoren waren auch hier auf 1 fixiert, was eine Interpretation der Differenzen als Effektstärken ermöglicht. Dieses Modell wies mit CFI = .93 und RMSEA = 0.07 ebenfalls eine zufriedenstellende Passung auf. Der χ^2-Test auf die Differenz der Passung beider Modelle ergab $\chi^2_9 = 33.89$ ($p < .001$).

Tabelle 19

Auszug aus standardisierten Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 2.3)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.66	0[a]
[95%-KI]			[-0.97; -0.36]	
Wissen2	0[a]		0.43	0[a]
[95%-KI]			[0.14; 0.72]	
Korrelation				
Wissen1 – Wissen2	.90		.93	
[95%-KI]	[.84; .96]		[.87; .99]	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Achsenabschnitte				
Biologie	2.07	2.42	2.47	
Darstellende Kunst	2.78	2.59	2.55	
Ernährung	2.69	2.68	2.70	
Gesundheit	2.65	2.55	2.48	
Medizin	2.43	3.02	2.92	
Modedesign	2.79	2.18	2.27	
Natur	2.47	2.62	2.82	
Pädagogik	3.36	3.12	3.12	
Psychologie	1.98	2.30	2.27	
Raumdesign	2.61	2.54	2.46	
Soziale Arbeit	2.66	3.23	3.07	
Model-Fit				
χ^2_{df}	146.07_{107}		180.80_{116}	
p	.007		< .001	
SCF	0.97		0.98	
CFI	.96		.93	
SRMR	.08		.09	
RMSEA	.05		.07	
[90%-KI]	[.03; .08]		[.05; .09]	
χ^2-Diff-Test: χ^2_{df}; *p*	33.89_9; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.
[a]Parameter wurde vorab fixiert.

Die manifesten Variablen Ernährung und Psychologie, die auf Basis der exploratorischen Faktorenanalysen nicht eindeutig einem Faktor zugeordnet werden konnten, zeigten auch bei den konfirmatorischen Faktorenanalysen starke Variation in ihren Faktorladungen auf den

beiden Faktoren in Schritt 1 und Schritt 2. Wie aus Tabelle L8 ersichtlich wird, fiel beispielsweise in dem hier näher betrachteten Datensatz 2.3 im ersten Schritt die Schätzung der Ladung der Variable Psychologie auf dem Faktor Wissen1 geringer aus als die Schätzung der Ladung auf dem Faktor Wissen2. Im zweiten Schritt verhielt es sich umgekehrt: hier wurde die Ladung auf Wissen1 höher geschätzt als die Ladung auf Wissen2. Aus den übrigen Tabellen in Anhang L geht hervor, dass sich diese geringe Konsistenz für die Variablen Psychologie und Ernährung bei nahezu sämtlichen Datensätzen zeigte.

5.3.3 Diskussion

Das Ziel der zweiten Studie war die Erfassung von Geschlechterdifferenzen des Wissens bei Themen, für welche Frauen sich durchschnittlich stärker als Männer interessieren. Hierbei sollte ein reliables, valides Messinstrument verwendet werden. Es wurde ein neuer Wissenstest entwickelt, der das Wissen zu Themengebieten erfasst, für die Frauen den Ergebnissen der vorigen Studie zufolge durchschnittlich höher ausgeprägte Interessen aufweisen. Wie bereits in Kapitel 5.3.1 erläutert, führte die thematische Breite der Begriffe Gesundheit und Natur zwangsläufig dazu, dass die subjektive Auffassung dieser Begriffe ausschlaggebend für die Inhalte der Items war und dass eine Beurteilung der Inhaltsvalidität nicht möglich war. Für alle übrigen Themen bestätigten die Expertinnen und Experten die Inhaltsvalidität der Items. Nach der Datenerhebung wurden auf Basis von den Kennwerten der einzelnen Items und der 11 Skalen der verschiedenen Themen insgesamt 121 Items (11 pro Thema) ausgewählt, die den neuen Wissenstest bildeten. Trennschärfen und Schwierigkeiten der Items des Tests waren in sämtlichen Datensätzen zufriedenstellend. Bezüglich der Inhaltsvalidität ist einschränkend zu erwähnen, dass die Expertinnen und Experten die 30 bis 39 entwickelten Items auf ihre Inhaltsvalidität beurteilten, nicht jedoch

die nach der Itemselektion weiter verwendeten 11 Items der einzelnen Themen. Eindeutigkeit und Korrektheit können jedoch auch bei den ausgewählten Items angenommen werden.

Die Konsistenzanalysen der einzelnen Skalen zeigten Ergebnisse, die über die verschiedenen Themengebiete stark variierten. Mit einem Mittelwert von .55 lag Cronbachs α hier durchschnittlich deutlich unter den α-Koeffizienten, die für die einzelnen Facetten des BOWIT berechnet wurden. Aus den Angaben, die sich im Manual des Tests finden, lässt sich ein Mittelwert von .74 (Form A) beziehungsweise .75 (Form B) berechnen. Hier sollte allerdings beachtet werden, dass es sich bei einem Großteil der Facetten des BOWIT um Themen handelt, die auch als Schulfächer existieren. Dies ermöglicht die Orientierung an Curricula, womit die internen Konsistenzen der einzelnen Skalen vermutlich deutlich steigen. Beispiele hierfür sind die Themen Physik, Wirtschaft und Geschichte. In dem neu entwickelten Wissenstest hingegen liegen Skalen zu zahlreichen Themen vor, zu denen keine curricular fundierte Definition eines Itemuniversums bestand. Beispiele hierfür sind die Themen Gesundheit, Natur und Modedesign. Diese Erklärung für die geringeren internen Konsistenzen der Skalen des hier vorgestellten Wissenstests wird durch den Vergleich der α-Koeffizienten von einzelnen Themen, zu denen sowohl im BOWIT als auch im neu entwickelten Wissenstest eigene Skalen vorliegen, gestützt: für den Bereich Biologie wurde beim BOWIT über ein Cronbachs α von .70 (Form A) beziehungsweise .77 (Form B) berichtet. Die Skala des Themas Biologie erreichte beim vorliegenden Test ein α von .75. Für den Bereich Gesundheit gibt es ebenfalls in beiden Tests eine eigene Facette. Cronbachs α lag mit .55 im vorliegenden Test nur geringfügig unter den Koeffizienten der entsprechenden Facette beim BOWIT, die .60 (Form A) beziehungsweise .62 (Form B) betrugen. Die geringeren α-Koeffizienten des vorliegenden Tests erklären sich somit vermutlich wesentlich durch die Themengebiete, zu denen das Wissen abgefragt wird. Ein zweiter Grund dürfte die etwas geringere Anzahl an Items für die einzelnen Facetten sein. Während im BOWIT jede

Facette aus 14 Items besteht, setzen sich die Skalen des neu entwickelten Tests lediglich aus jeweils 11 Items zusammen. Eine Erhöhung der Itemanzahl auf 14 würde nach Berechnungen mit der Spearman-Brown-Formel beispielsweise die Reliabilität der Skala Gesundheit von .55 auf .61 erhöhen. Für die Skala Biologie würde die Reliabilität von .75 auf .79 gesteigert. Die Cronbachs α-Werte des gesamten Tests betrugen in allen 10 Datensätzen .90 oder .91 und lagen damit nur geringfügig unter den Werten, über die in den Manualen des I-S-T und des BOWIT berichtet wird. Die Konsistenzanalysen ergaben beim I-S-T α = .93 (Liepmann et al., 2007) und beim BOWIT α = .95 (Hossiep & Schulte, 2008). Zusammenfassend ist die interne Konsistenz des neu entwickelten Wissenstests somit zufriedenstellend.

Die Korrelationen der Summenscores mit den Schulnoten der Schülerinnen und Schüler lieferten Evidenz für die Kriteriumsvalidität. Die Korrelationen des BOWIT mit den durchschnittlichen Noten der Abiturfächer in der Sekundarstufe II betrugen -.32 (Form A) und -.31 (Form B) (Hossiep & Schulte, 2008). Das Manual des I-S-T enthält keine Information über Korrelationen zwischen dem Wissenstest und Schulnoten. Es wird jedoch über Korrelationen zwischen kristalliner Intelligenz, die maßgeblich über den Wissenstest erfasst wird, und Schulnoten berichtet. Diese liegen zwischen -.29 für das Schulfach Mathematik und -.38 für das Schulfach Erdkunde (Liepmann et al, 2007). Die negativen Vorzeichen, die die Korrelationen des BOWIT und des I-S-T aufweisen, erklären sich dadurch, dass hier die Zusammenhänge mit Schulnoten und nicht mit Schulleistungen in Form eines Punktesystems berichtet werden. Die Korrelation der Gesamtskala des neuen Wissenstests mit der Durchschnittsleistung des zuletzt erhaltenen Zeugnisses (in Punkten) betrug .37 und fiel damit insgesamt höher aus als die meisten für den BOWIT und den I-S-T berichteten Korrelationen.

Auch in dem neu entwickelten Wissenstest erreichten Männer die durchschnittlich höheren Scores. Das Ausmaß der Leistungsunterschiede variierte über die Themen jedoch

deutlich. Wie aus Tabelle 16 hervorgeht, wiesen Männer bei fünf Themen und Frauen bei drei Themen bessere Leistungen auf. Bei drei Themen zeigten sich keine bedeutsamen Leistungsunterschiede. Für den gesamten Test fiel die Geschlechterdifferenz mit d = -0.11 deutlich niedriger aus als in bestehenden Wissenstests. Allerdings bestätigten die Effektstärken auch in dieser Untersuchung die höhere Leistung von Männern in Wissenstests.

Es ist zu beachten, dass die Schülerinnen und Schüler der Stichprobe in Studie 2 nicht den gleichen Interessenunterschied für Biologie aufwiesen wie die Teilnehmenden der Untersuchung, deren Daten in Studie 1 für die Zuordnung der Themen in die Kategorien IvF, NI und IvM verwendet wurden. Hier hatte sich für das Interesse an der Biologie eine Effektstärke der Geschlechterdifferenz von d = 0.20 ergeben. In der vorliegenden Stichprobe wurde eine Effektstärke von -0.05 berechnet. Demnach handelt es sich bei der Biologie um ein Thema, das auf Basis der Daten aus Studie 2 nicht als IvF, sondern als NI klassifiziert worden wäre. Für alle weiteren Themen, die in Studie 1 als IvF kategorisiert wurden, gaben Frauen jedoch auch in Studie 2 im Durchschnitt stärkere Interessen als Männer an. Somit kann die Differenz der Effektstärken des Themas Biologie in den beiden Untersuchungen nicht als Erklärung dafür dienen, dass auch in den neu entwickelten Wissenstest Männer die durchschnittlich höheren Summenscores erzielten.

Die Ergebnisse dieser Untersuchung weisen somit in die gleiche Richtung wie die Ergebnisse der ersten Studie. Während in der ersten Studie die geringen Effektstärken in der Kategorie IvF noch auf ein Item (I-S-T) bzw. auf 16 Items (BOWIT) zurückzuführen waren, zeigen sich ähnliche Ergebnisse mit dem neu entwickelten Wissenstest auf Basis von 121 Items. Interessenunterschiede zwischen Frauen und Männern sind scheinbar von zentraler Bedeutung für Leistungsunterschiede in Tests des Allgemeinen Wissens. Auch die Korrelation zwischen den Effektstärken der Interessen- und Leistungsunterschiede von .53 unterstützt diese Vermutung. Offenbar müssen die einseitigen Geschlechterdifferenzen in

Wissenstests jedoch auch auf weitere Ursachen zurückzuführen sein, denn andernfalls würden Frauen in dem hier vorgestellten Test deutlich höhere Leistungen als Männer zeigen, was sich jedoch in der vorliegenden Studie nicht bestätigte. Die Konzentration auf Themen, für die Frauen sich stärker interessieren als Männer, hat nicht zu einer Umkehrung der Geschlechterdifferenzen geführt, sondern lediglich zu einer deutlichen Verringerung.

Durch die Überprüfung der faktoriellen Struktur des Tests durch Parallelanalysen und exploratorische und konfirmatorische Faktorenanalysen rückt ein weiteres Problem in den Fokus, das mit den einseitigen Geschlechterdifferenzen in bestehenden Wissenstests in Zusammenhang stehen könnte: Unter der Annahme, dass es sich bei Wissen um ein Konstrukt handelt, das eine einzelne Fähigkeit beinhaltet, war im Vorfeld für den vorliegenden Test von einem 1-faktoriellen Messmodell ausgegangen worden. Wie bereits dargelegt, bestehen jedoch aufgrund von Analysen bestehender Wissenstests Zweifel an diesem Modell. Die Überprüfung des Messmodells mit einem einzelnen Faktor durch konfirmatorische Faktorenanalysen war auch für den neu entwickelten Test nicht zufriedenstellend. Die Restriktionen des zweiten Schrittes, welche die Annahme enthalten, dass Unterschiede innerhalb der Gruppen und zwischen den Gruppen auf denselben Faktor zurückzuführen sind, brachten erhebliche Einbußen in den Fit-Maßen mit sich. Für das 1-faktorielle Modell kann die Messinvarianz daher nicht als gegeben angenommen werden. Zusätzlich fanden sich bei mehreren Parallelanalysen Hinweise auf eine 2-faktorielle Struktur (siehe Tabelle 18). Die Summenscores der Themen Biologie, Natur, Medizin und Soziale Arbeit wurden dem Faktor Wissen1 zugeordnet. Die Skalen Gesundheit, Darstellende Kunst, Modedesign, Raumdesign und Pädagogik wurden dem Faktor Wissen2 zugeordnet. Es stellt sich somit die Frage, worin sich die beiden Faktoren inhaltlich unterscheiden.

Ein auffallender Unterschied zwischen den Faktoren ist, dass die Themen des Faktors Wissen1 im Vergleich zu den Themen des Faktors Wissen2 stärkere Überschneidung mit den

Inhalten bestehender Wissenstests zeigen. Zu den Themen Darstellende Kunst, Raumdesign, Modedesign und Pädagogik finden sich wenige oder keine Items im BOWIT und im I-S-T. Lediglich zum Thema Gesundheit existiert im BOWIT eine eigene Facette. Die Themen des Faktors Wissen1 – Biologie, Natur, Medizin und Soziale Arbeit – sind hingegen zu einem großen Teil auch in den bestehenden Wissenstests wiederzufinden. Der BOWIT enthält eine Facette Biologie. Die Autoren des I-S-T fassen eine Gruppe von Items unter dem Begriff Naturwissenschaften zusammen. Die Items des neuen Wissenstests zum Thema Natur könnten theoretisch auch einzelnen Disziplinen, wie z.B. der Biologie, der Physik oder der Geographie, zugeordnet werden. Ebenso wurden zahlreiche Items des BOWIT und des I-S-T der Geographie und der Physik zugeordnet. Die Fächer Soziale Arbeit und Medizin liegen weder im BOWIT noch im I-S-T vor. Allerdings finden sich beim BOWIT Facetten zu Gesellschaft/Politik und zu Wirtschaft/Recht, was thematisch auch Überschneidungen mit der Sozialen Arbeit aufweisen dürfte. Obwohl bei der Entwicklung des hier vorgestellten Tests nicht die Absicht vorlag, einen hinsichtlich der Interessen von Frauen und Männern ausgewogenen Test zu schaffen, ist dies möglicherweise eingetreten. Hier spiegeln sich zwei bereits in Studie 1 erwähnte Probleme wieder: (a) Bei der Erstellung der Themenliste ist der gewählte Detailgrad bei der Differenzierung verschiedener Themen stark subjektiv gewichtet. (b) Die einzelnen Themen der Liste sind teilweise nicht abzugrenzen, sondern überlappen sich. Möglicherweise sind einzelne Themen daher ungeeignet, um als Oberbegriff für eine Gruppe von Items zu dienen. So wurde beispielsweise das Thema Natur in der ersten Studie als IvF kategorisiert, während es sich alternativ auch in mehrere einzelne Themen aufteilen lässt, welche in Studie 1 teilweise nicht als IvF kategorisiert wurden, wie beispielsweise Physik und Geographie. Unter dieser Perspektive ist der entwickelte Test nicht ausschließlich auf die Interessen von Frauen zugeschnitten, sondern umfasst ebenfalls Themen, die Männer durchschnittlich stärker interessieren und welche hier dem Faktor Wissen1 zugeordnet

wurden. Möglicherweise liegt der inhaltliche Unterschied zwischen beiden Faktoren somit in geschlechtsspezifischen Interessenschwerpunkten. Die konfirmatorischen Faktorenanalysen weisen auf eine gute Passung des in Abbildung 6 dargestellten Messmodells hin. Auch die Messinvarianz des Tests wurde unter der Annahme, dass zwei verschiedene Merkmale mit dem Test erfasst werden, bestätigt. Unklar bleibt jedoch die Zuordnung der Themen Psychologie und Ernährung. Weder die exploratorischen noch die konfirmatorischen Faktorenanalysen weisen die manifesten Variablen dieser Themen eindeutig einem der beiden Faktoren zu.

Zusammenfassend kann festgehalten werden, dass es gelungen ist, einen neuen Wissenstest zu entwickeln, der thematisch von bestehenden Wissenstests abweicht, hinsichtlich der Gütekriterien jedoch mit dem Wissenstest des I-S-T und dem BOWIT vergleichbar ist. Die interne Konsistenz und die Kriteriumsvalidität des neu entwickelten Tests sind sehr zufriedenstellend. Die Leistungsunterschiede von Frauen und Männern weisen nach wie vor in die gleiche Richtung: Männer zeigten auch in diesem Test bessere Leistungen. Die Unterschiede fallen jedoch deutlich geringer aus. Die gängige Praxis, Gruppenunterschiede des Wissens in Form von Mittelwertdifferenzen der Summenscores in entsprechenden Tests zu berichten, ist allerdings problematisch, sofern die Eindimensionalität des Messinstruments nicht sichergestellt ist, da Summenscores in diesem Fall nicht interpretierbar sind. Das Problem ergibt sich beispielsweise beim BOWIT, für den eine exploratorische Faktorenanalyse eine 2-faktorielle Lösung ergab (siehe Kapitel 2.3.2). In der vorliegenden Untersuchung wurden dennoch Mittelwertdifferenzen von Summenscores berichtet, um einen direkten Vergleich mit den in Kapitel 3.1 berichteten Effektstärken zu ermöglichen. Die gute Passung der 2-faktoriellen Messmodelle weist allerdings darauf hin, dass sich Wissen möglicherweise aus zwei verschiedenen Merkmalen zusammensetzt und der entwickelte Test somit nicht eindimensional ist. Die beiden Merkmale korrelieren zwar sehr

hoch miteinander, jedoch fallen die Geschlechterdifferenzen unterschiedlich aus: ein Merkmal könnte dem Wissen entsprechen, das auch mit bestehenden Wissenstests erfasst wird, die einseitige Geschlechterdifferenzen zugunsten von Männern aufweisen. Daneben könnte ein zweites Merkmal existieren, hier als Wissen2 bezeichnet, das in bestehenden Wissenstests nicht erfasst wird, jedoch in den hier vorgestellten Test Eingang gefunden hat. In Wissen2 zeigten Frauen die durchschnittlich höheren Ausprägungen. Eine konkrete inhaltliche Interpretation der Begriffe Wissen1 und Wissen2 steht an dieser Stelle aus. Die Ergebnisse weisen jedoch auf Zusammenhänge mit Interessengebieten von Männern (Wissen1) beziehungsweise Frauen (Wissen2) hin. Sollten etablierte Wissenstests tatsächlich in erster Linie Items zu Wissen1 enthalten, während das Merkmal Wissen2 nicht, oder nur geringfügig, für die Leistung von Bedeutung ist, so würde hieraus eine Überschätzung der Geschlechterdifferenzen des Wissens zugunsten von Männern folgen. In der dritten Studie wurde daher die faktorielle Struktur einer Testbatterie überprüft, die sich aus dem in Studie 2 entwickelten Wissenstest und dem Wissenstest des I-S-T 2000 R zusammensetzte.

5.4 Studie 3 – Überprüfung der faktoriellen Struktur einer thematisch ausbalancierten Wissenstestbatterie

Die Ergebnisse der vorangegangen Studie deuteten darauf hin, dass das Konstrukt Wissen nicht als eine einzelne latente Variable, sondern angemessener in Form zweier verschiedener Variablen abgebildet werden kann. Im Hinblick auf die Fragestellung der vorliegenden Arbeit ist hierbei außerdem von zentraler Bedeutung, dass in Studie 2 Frauen und Männer in den beiden Variablen gegenteilige Geschlechterdifferenzen aufgewiesen hatten. In den Faktorenanalysen eines Wissenstests, der mit der Absicht entwickelt worden war, thematisch einseitig die Themen der Kategorie IvF abzudecken, wurde Frauen für die mit Wissen1 bezeichnete Variable ein geringerer Erwartungswert als Männern attestiert, während es sich

bei der mit Wissen2 bezeichneten Variable umgekehrt verhielt. Aus den in Kapitel 5.3.3 genannten Gründen sind jedoch Zweifel an der Einseitigkeit der thematischen Ausrichtung des Tests berechtigt. Die Verteilung der verschiedenen Themen auf die beiden latenten Variablen ließen Überschneidungen von Wissen1 mit den Themen erkennen, die in etablierten Wissenstests dominieren, während Wissen2 einen Faktor bildet, der nicht oder kaum Bestandteil bekannter Wissenstests ist.

Es stellt sich die Frage, welche faktorielle Struktur eine thematisch breit gefasste Testbatterie aufweist, die sowohl einen etablierten Wissenstest enthält, der hauptsächlich Items zu Themen der Kategorie IvM umfasst, als auch den neu entwickelten Wissenstest, der einen zweiten Faktor aufweist, dessen Themen der Kategorie IvF zuzuordnen sind. Die Ergebnisse aus den beiden vorangegangenen Studien legen folgende Annahmen nahe: Eine entsprechend zusammengesetzte Testbatterie weist eine 2-faktorielle Struktur auf. Einer der beiden Faktoren entspricht dem Faktor Wissen1 aus Studie 2, welcher die Inhalte des etablierten Wissenstests und die Inhalte des neuen Wissenstests umfasst, die auch in Studie 2 dem Faktor Wissen1 zugeordnet wurden. Der geschätzte Erwartungswert dieses Faktors ist bei Frauen kleiner als bei Männern. Der Faktor Wissen1 bildet somit das Merkmal ab, das auch zentraler Inhalt der Mehrzahl bestehender Wissenstests ist, die zur Erfassung von Wissen auf einer breiten Ebene dienen sollen. Männer haben hierin die höhere Ausprägung. Die in Studie 3 verwendete Testbatterie umfasst jedoch auch einen zweiten Faktor, Wissen2, dem im Wesentlichen die gleichen Inhalte wie in Studie 2 zugeordnet werden können. Wissenstests, über die ausführlich in Kapitel 3.1 berichtet wurde, weisen keine (oder nur zu einem geringen Anteil) Inhalte auf, die Wissen2 zugeordnet werden können. Für diesen Faktor wird der Erwartungswert von Frauen höher geschätzt als der Erwartungswert von Männern. Wissen2 ist somit bei Frauen höher ausgeprägt. Das Ziel der dritten Studie bestand in der Überprüfung dieser Annahmen. Unterstützende Ergebnisse würden die Hypothesen

bestätigen, dass (a) in bestehenden Wissenstests einseitig das Wissen zu Themen der Kategorie IvM abgefragt wird und dass (b) hierin eine Erklärung für die häufig anzutreffenden Geschlechterdifferenzen in Tests des sogenannten Allgemeinen Wissens zugunsten von Männern liegt. Der Ausdruck des "sogenannten" Allgemeinen Wissens wird hier verwendet, weil eine Bestätigung der beschriebenen Hypothesen auch Zweifel an der Operationalisierung des Konstrukts des Allgemeinen Wissens aufkommen lassen würde. Dieses Thema wird in der abschließenden Diskussion aufgegriffen.

5.4.1 Methoden

Zu Beginn der Untersuchung wurde allen Personen der bereits in Studie 2 verwendete demographische Fragebogen und der in Studie 1 vorgestellte Interessenfragebogen vorgelegt. Die Batterie aus Wissenstests umfasste den im Rahmen von Studie 2 entwickelten Wissenstest und den Wissenstest des I-S-T 2000 R. Der neu entwickelte Wissenstest umfasste 121 Items, die gleichmäßig über die 11 Themen verteilt waren, welche im Rahmen von Studie 1 als IvF kategorisiert worden waren. Von jenen 11 Themen wurden in Studie 2 Biologie, Medizin, Natur und Soziale Arbeit ausschließlich dem Faktor Wissen1 zugeordnet. Darstellende Kunst, Gesundheit, Modedesign, Pädagogik und Raumdesign wurden ausschließlich Wissen2 zugeordnet. Für die Themen Ernährung und Psychologie war keine eindeutige Zuordnung möglich, da sie bedeutsame Ladungen auf beiden Faktoren aufwiesen. Der Wissenstest des I-S-T wurde in den Kapiteln 2.3.2 und 5.2 näher beschrieben. Das für die im Folgenden beschriebene Untersuchung wichtige Merkmal dieses Tests war die deutliche Dominanz von Items zu Themen der Kategorie IvM gegenüber Items zu Themen der Kategorie IvF. Es sei daran erinnert, dass bei der Zuordnung der Items zu den Kategorien in Studie 1 ausschließlich jene mit einer Beurteilerübereinstimmung von Krippendorffs $\alpha \geq .667$ berücksichtigt wurden. In Studie 3 wurden jedoch sämtliche 84 Items des Tests verwendet.

Die Teilnehmenden bearbeiteten somit insgesamt 205 Wissensitems. Sämtliche Items stimmten hinsichtlich ihres Formats überein. Auf den Itemstamm folgten jeweils 5 Antwortoptionen, von denen eine einzige die korrekte Antwort darstellte. Alle 205 Items wurden bei der Planung der Untersuchung in zufälliger Reihenfolge sortiert und anschließend in einem einzelnen Testheft entsprechend angeordnet. Die Tatsache, dass es sich hierbei um zwei verschiedene Wissenstests handelte, war für die Teilnehmenden somit nicht salient.

An den Datenerhebungen nahmen Studierende unterschiedlicher Fächer der Universität Wuppertal und Schülerinnen und Schüler zweier Bochumer Schulen teil. Aus dem privaten Umfeld von studentischen Hilfskräften konnten ebenfalls einige Probandinnen und Probanden rekrutiert werden. Für eine einzelne Testung standen 95 Minuten zur Verfügung. Als Anreiz für die Teilnehmenden aus dem privaten Umfeld der Hilfskräfte und aus der Universität diente die Verlosung von 5 × 50 und 1 × 250 Euro. Studierende des Faches Psychologie nahmen nicht an der Verlosung teil, konnten sich jedoch die Teilnahme an der Untersuchung bescheinigen lassen. Sämtliche Personen konnten außerdem in schriftlicher Form Rückmeldung zur individuellen Leistung im Wissenstest des I-S-T erhalten. Bei den Schülerinnen und Schülern handelte es sich um die 10. Jahrgangsstufe einer Realschule, die im Mai 2014 an der Untersuchung teilnahm, sowie um die 11. Jahrgangsstufe eines Gymnasiums, die im August 2014 teilnahm. Obwohl die Schülerinnen und Schüler der beiden Schulen aus zwei verschiedenen Jahrgangsstufen stammten, unterschieden sie sich hinsichtlich des Alters somit nur geringfügig. Die gesamte Stichprobe umfasste 216 Personen. Aufgrund des offensichtlichen Fehlens von Ernsthaftigkeit bei der Bearbeitung der Aufgaben (beispielsweise Auswahl der gleichen Option über weite Teile des Wissenstests, Abgabe der Testunterlagen nach 30 Minuten) und da einige Teilnehmende seit weniger als 10 Jahren Deutsch sprachen, wurden die Daten von insgesamt 14 Personen nicht weiter

verwertet. Der für die folgenden Analysen verwendete Datensatz umfasste somit insgesamt n = 202 Personen (52.5% Frauen).

Vor der Durchführung der Analysen, die zur Prüfung der beschriebenen Hypothesen dienten, erfolgte zunächst eine Überprüfung der Kennwerte der Items und Skalen des neuen Wissenstests. Itemschwierigkeiten und -trennschärfen, Cronbachs α-Werte der einzelnen Skalen sowie die Kriteriumsvalidität in Form der Korrelation zwischen Leistung im Wissenstest und Schulleistung (in Punkten) wurden anhand der für diese Studie erhobenen Daten erneut berechnet und mit den Ergebnissen aus Studie 2 verglichen. Da in Studie 3 sowohl ein etablierter Wissenstest als auch der neu entwickelte Wissenstest verwendet wurden, konnten die Korrelationen der Summenscores beider Tests mit der Schulleistung verglichen werden.

Im Vorfeld der Faktorenanalysen von Studie 3 wurden manifeste Variablen für den neuen Wissenstest und für den Wissenstest des I-S-T gebildet. Für den neuen Wissenstest wurden – analog zu Studie 2 – jeweils die Summenscores der 11 Items der einzelnen Themen berechnet, um die Themen den einzelnen Faktoren zuordnen zu können, sofern ein passendes Messmodell mehrere Faktoren enthalten würde. Die 84 Items des I-S-T wurden in vier Parcels aufgeteilt, wobei die Zuordnung der Items zufällig erfolgte. Für jeden Parcel wurde die Summe der jeweiligen 21 Items berechnet. Die so erstellten vier manifesten Variablen des I-S-T bildeten eine Parcelgruppe. Ein einzelner Datensatz umfasste insgesamt 15 manifeste Variablen: die 11 Variablen des neuen Wissenstests und die vier Variablen der Parcelgruppe des I-S-T. Die zufällige Zusammenstellung von Parcels wurde dreimal wiederholt. Hieraus ergaben sich drei Parcelgruppen und damit drei verschiedene Datensätze. Sämtliche im Folgenden beschriebenen Analysen erfolgten mit allen drei Datensätzen.

In einem ersten Schritt wurde zur Bestimmung der Anzahl der Faktoren eine Parallelanalyse durchgeführt. Den beschriebenen Hypothesen zufolge war eine 2-faktorielle

Struktur zu erwarten. Der nächste Schritt bestand in einer exploratorischen Faktorenanalyse für die Zuordnung der 15 manifesten Variablen zu den 2 Faktoren. Bezüglich des in Studie 2 entwickelten Wissenstests wurde erwartet, dass die Verteilung der 11 Themen dieses Tests auf die Faktoren Wissen1 und Wissen2 weitgehend mit der Verteilung aus Studie 2 übereinstimmen würde. Auf Basis der Faktorladungen wurde ein Messmodell erstellt, das anschließend anhand von konfirmatorischen Faktorenanalysen geprüft wurde. Sämtliche angewandten Methoden entsprachen hierbei den in Kapitel 3 beschriebenen Verfahren. Für die Parallelanalyse wurde die von O'Connor (2000) erstellte Syntax mit SPSS 22 verwendet. In der exploratorischen Faktorenanalyse wurde die Oblimin-Rotation verwendet. Die konfirmatorischen Faktorenanalysen beinhalteten auch die Überprüfung der Messinvarianz nach dem in Studie 2 verwendeten Verfahren. Im ersten Schritt wurden zunächst die Erwartungswerte der Faktoren für Frauen und Männer gleichgesetzt und die Achsenabschnitte der manifesten Variablen in beiden Gruppen frei geschätzt. Im zweiten Schritt wurden die Achsenabschnitte gleichgesetzt und die Erwartungswerte der Faktoren in einer Gruppe auf 0 fixiert und in der anderen Gruppe frei geschätzt. Die Restriktionen aller übrigen Parameter stimmten in beiden Analysen überein. Die exploratorischen und konfirmatorischen Faktorenanalysen wurden mit Mplus 7.1 durchgeführt.

5.4.2 Ergebnisse

In Tabelle 20 sind die Item- und Skalenparameter des neuen Wissenstests zusammengefasst, die anhand der Daten aus Studie 3 berechnet wurden. Diese können mit den entsprechenden Ergebnissen aus Studie 2, die in Tabelle 14 zusammengefasst wurden, verglichen werden. Mit einer durchschnittlichen Schwierigkeit aller Items von .55 war der Test in Studie 3 insgesamt etwas leichter. Vor allem die Items der Skalen Ernährung, Medizin und Modedesign wiesen hier im Durchschnitt geringere Schwierigkeiten auf. Bei drei der 11

themenspezifischen Skalen wurde im Datensatz dieser Studie jeweils ein Item mit negativer Trennschärfe identifiziert. Die Gesamtskala enthielt drei Items mit negativer Trennschärfe. Deutliche Verringerungen der internen Konsistenz traten bei den Themen Biologie, Gesundheit und Soziale Arbeit auf, während die Cronbachs α-Werte der Skalen Psychologie und Raumdesign auffallend zunahmen. Die Cronbachs α-Werte aller Skalen betrugen im Durchschnitt .52. Die interne Konsistenz der Gesamtskala des neu entwickelten Tests blieb mit .91 gegenüber Studie 2 unverändert. Für den Wissenstest des I-S-T wurde Cronbachs α = .86 berechnet.

Tabelle 20

Item- und Skalenkennwerte des neuen Wissenstests (Studie 3)

	Schwierigkeit			Korr. Trennschärfe			
Thema	Min	Max	*M*	Min	Max	*M*	α
Biologie	.25	.71	.52	.04	.39	.21	.51
Darstellende Kunst	.16	.76	.51	-.09	.35	.19	.47
Ernährung	.30	.87	.58	.06	.34	.21	.51
Gesundheit	.27	.84	.53	.00	.21	.13	.36
Medizin	.32	.84	.62	.18	.48	.31	.66
Modedesign	.34	.80	.57	.16	.33	.24	.57
Natur	.42	.72	.55	.13	.35	.24	.57
Pädagogik	.27	.96	.62	-.07	.35	.20	.49
Psychologie	.26	.71	.46	.20	.42	.31	.67
Raumdesign	.33	.71	.51	.10	.41	.29	.63
Soziale Arbeit	.26	.88	.54	-.03	.23	.11	.30
Gesamt	.16	.96	.55	-.10	.53	.27	.91

Anmerkungen. Korr. Trennschärfe = Trennschärfeberechnungen nach part-whole-Korrektur; Min = Minimum; Max = Maximum; α = Cronbachs α.

Für die Korrelationen mit der Schulleistung konnte nur auf eine Stichprobe von 161 Personen (52.8% Frauen) zurückgegriffen werden, da die übrigen Teilnehmenden keine

Angaben zu ihrer Gesamtleistung in der Schule gemacht hatten. Der Wissenstest des I-S-T 2000 R korrelierte mit der Gesamtleistung in der Schule (in Punkten) zu .31. Der im Rahmen von Studie 2 entwickelte Wissenstest wies eine Korrelation mit der Schulleistung von .49 auf.

Alle drei Parallelanalysen, die mit den Datensätzen durchgeführt wurden, wiesen eindeutig auf eine 2-faktorielle Lösung hin. Wie aus den Tabellen M1 bis M3 hervorgeht, lagen in allen drei Datensätzen sowohl der Mittelwert als auch das 95%-Quantil der aus den simulierten Daten stammenden Eigenwerte des zweiten Faktors unterhalb des Eigenwerts des aus den empirischen Daten stammenden zweiten Faktors. In keinem der drei Datensätze wurden Hinweise auf einen dritten Faktor festgestellt.

Im Anschluss wurden drei exploratorische Faktorenanalysen durchgeführt, bei denen die Anzahl der Faktoren auf 2 fixiert war. In Tabelle 21 sind für alle drei Datensätze die Faktorladungen der manifesten Variablen auf den beiden Faktoren nach der Oblimin-Rotation dargestellt. Folgende Schlüsse wurden hieraus gezogen: Drei der vier Parcels der aus dem I-S-T stammenden Items und die Skala Biologie wiesen in allen drei Faktorenanalysen ausschließlich auf dem ersten Faktor auf dem 5%-Niveau signifikante Ladungen auf. Ein Parcel des I-S-T zeigte in zwei Datensätzen ebenfalls nur auf dem ersten Faktor und im dritten Datensatz zusätzlich auf dem zweiten Faktor eine signifikante Ladung, welche jedoch gering ausfiel und ein negatives Vorzeichen aufwies. Die Skala Medizin zeigte zwar in allen drei Datensätzen signifikante Ladungen auf beiden Faktoren, allerdings war die Ladung auf dem ersten Faktor durchgehend größer als die Ladung auf dem zweiten Faktor. Diese fünf manifesten Variablen konnten somit eindeutig dem ersten Faktor zugeordnet werden. Die Skalen Darstellende Kunst, Ernährung, Raumdesign und Soziale Arbeit zeigten in allen drei Datensätzen ausschließlich auf dem zweiten Faktor signifikante Ladungen. Die Themen Gesundheit, Modedesign, Pädagogik und Psychologie zeigten in den unterschiedlichen Datensätzen entweder nur auf dem zweiten Faktor signifikante Ladungen, oder die Ladungen

auf beiden Faktoren waren signifikant, diejenige auf dem zweiten Faktor jedoch höher als die auf dem ersten Faktor. Diese acht manifesten Variablen konnten somit eindeutig dem zweiten Faktor zugeordnet werden.

Tabelle 21

Faktorladungen der manifesten Variablen der Testbatterie in den exploratorischen Faktorenanalysen

Manifeste Variable	Parcelgruppe 1		Parcelgruppe 2		Parcelgruppe 3	
	Faktor 1	Faktor 2	Faktor 1	Faktor 2	Faktor 1	Faktor 2
I-S-T Parcel 1	**.70***	.03	**.75***	.06	**.89***	-.14*
I-S-T Parcel 2	**.87***	-.04	**.71***	.08	**.70***	.17
I-S-T Parcel 3	**.71***	.09	**.76***	.03	**.73***	.10
I-S-T Parcel 4	**.77***	-.04	**.89***	-.07	**.59***	.19
Biologie	**.45***	.04	**.44***	.05	**.43***	.07
Medizin	**.52***	.28*	**.49***	.32*	**.47***	.34*
DaKu	.11	**.58***	.08	**.61***	.09	**.60***
Ernährung	.16	**.57***	.16	**.57***	.16	**.57***
Gesundheit	.28*	**.44***	.23*	**.49***	.27*	**.46***
Modedesign	-.17*	**.83***	-.17*	**.83***	-.15*	**.81***
Pädagogik	.28*	**.49***	.25*	**.52***	.23	**.54***
Psychologie	.26*	**.53***	.24*	**.55***	.23*	**.56***
Raumdesign	.03	**.78***	.00	**.80***	-.01	**.82***
Soziale Arbeit	-.03	**.50***	.00	**.47***	-.04	**.50***
Natur	**.40***	.37*	.35*	**.42***	.35*	**.43***

Anmerkungen. Faktorladungen > .40 sind fett gedruckt. DaKu = Darstellende Kunst.
*$p < .05$.

Die Variable Natur zeigte in allen drei Datensätzen signifikante Ladungen auf beiden Faktoren. Hier war eine eindeutige Zuordnung nicht möglich, da bei einem Datensatz die Ladung auf dem ersten Faktor und bei zwei Datensätzen die Ladung auf dem zweiten Faktor

höher ausfiel. Daher wurde diese Variable beiden Faktoren zugeordnet. Wie aus Tabelle 21 ersichtlich wird, zeigte sich das gleiche Muster der Zuordnung, wenn hierfür die Höhe der Faktorladungen herangezogen und als Grenzwert .40 festgelegt wurde. Die Korrelation der beiden Faktoren betrug in den Datensätzen 1 und 2 jeweils .71 und im dritten Datensatz .68.

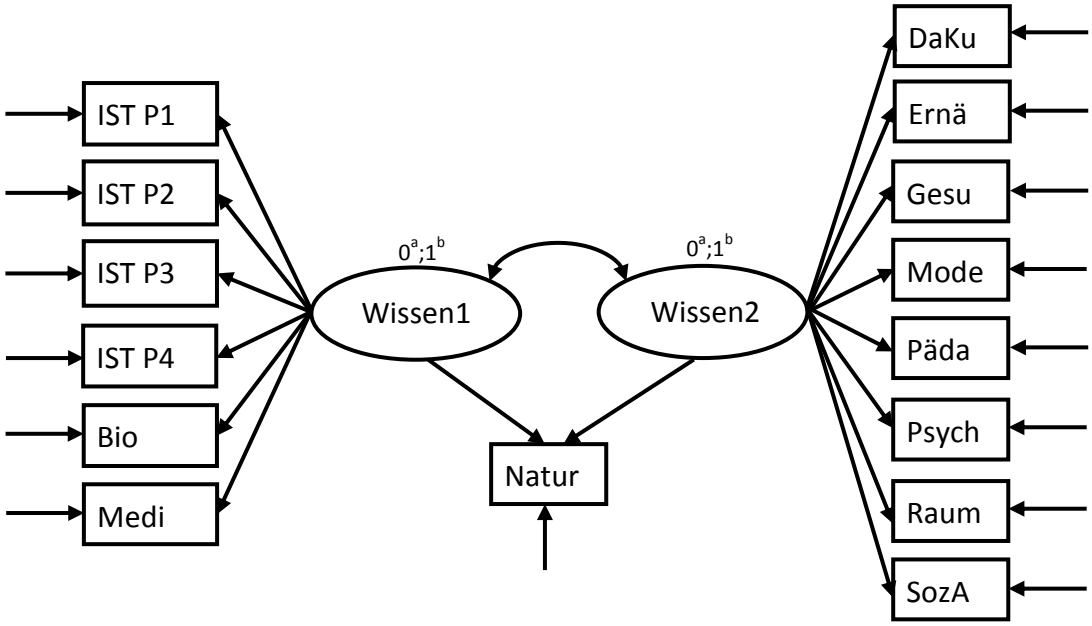

Abbildung 7. Messmodell mit zwei Faktoren – Studie 3. IST P= Wissenstest des I-S-T 2000 R - Parcel; Bio = Biologie; Medi = Medizin; DaKu = Darstellende Kunst; Ernä = Ernährung; Gesu = Gesundheit; Mode = Modedesign; Päda = Pädagogik; Psych = Psychologie; Raum = Raumdesign; SozA = Soziale Arbeit.

[a]Erwartungswert in Schritt 1 in beiden Gruppen fixiert. In Schritt 2 in einer zufällig ausgewählten Gruppe fixiert, in der anderen Gruppe frei geschätzt.

[b]Varianz in Schritt 1 und in Schritt 2 in beiden Gruppen fixiert.

In Abbildung 7 ist das Messmodell dargestellt, das anhand der exploratorischen Faktorenanalysen entworfen und in der Folge durch konfirmatorische Faktorenanalysen geprüft wurde. Als Bezeichnungen für die beiden Faktoren wurden hier die Begriffe Wissen1 und Wissen2 aus Studie 2 übernommen. Die Zuordnungen der Skalen des neuen Wissenstests zu den beiden Faktoren stimmten in Studie 2 und in Studie 3 nicht exakt überein. Für die

Themen Natur, Ernährung und Psychologie wurden Ladungen auf einem Faktor hinzugenommen (Natur) oder eliminiert (Ernährung und Psychologie). Das Thema Soziale Arbeit wechselte von Wissen1 zu Wissen2. Die übrigen 7 Skalen des neuen Wissenstests konnten dem gleichen Faktor wie in Studie 2 zugeordnet werden. Somit war in beiden Studien eine weitgehende Übereinstimmung der Inhalte von Wissen1 und Wissen2 zu erkennen. In Tabelle 22 ist ein Auszug der Parameterschätzungen der beiden konfirmatorischen Faktorenanalysen dargestellt, die mit einem zufällig ausgewählten Datensatz durchgeführt wurden. Hierbei handelte es sich um den zweiten Datensatz. In den Tabellen N1 bis N3 sind die Ergebnisse der Analysen für die drei Datensätze ausführlich dargestellt.

Tabelle 22

Auszug aus standardisierten Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 2)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.46	0[a]
[95%-KI]			[-0.75; -0.16]	
Wissen2	0[a]		0.54	0[a]
[95%-KI]			[0.24; 0.84]	
Korrelation				
Wissen1 – Wissen2	.94		.94	
[95%-KI]	[.91; .97]		[.91; .97]	
Achsenabschnitte				
I-S-T Parcel 1	3.22	3.61	3.58	
I-S-T Parcel 2	4.02	4.39	4.37	
I-S-T Parcel 3	3.37	3.75	3.72	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
I-S-T Parcel 4	2.89	3.32	3.27	
Biologie	2.64	2.67	2.78	
Darstellende Kunst	2.90	2.66	2.63	
Ernährung	3.20	2.89	2.88	
Gesundheit	3.30	3.14	3.08	
Medizin	2.74	2.97	3.05	
Modedesign	3.29	2.54	2.61	
Natur	2.63	2.56	2.56	
Pädagogik	3.51	3.30	3.25	
Psychologie	2.19	1.84	1.82	
Raumdesign	2.64	2.09	2.13	
Soziale Arbeit	3.53	3.22	3.24	
Model-Fit				
χ^2_{df}	273.80_{208}		301.32_{221}	
p	.002		< .001	
SCF	0.98		0.98	
CFI	.96		.95	
SRMR	.08		.09	
RMSEA	.06		.06	
[90%-KI]	[.04; .07]		[.04; .08]	
χ^2-Diff-Test: χ^2_{df}; *p*	27.51_{13}; .011			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. I-S-T = Wissenstest des I-S-T 2000 R; KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Die Passung des Modells des ersten Schrittes fiel mit CFI = .96 und RMSEA = .06 gut aus. Im zweiten Schritt wurden die Erwartungswerte der Faktoren für Männer auf 0 fixiert. Die Schätzungen der Erwartungswerte für Frauen betrugen -0.46 (Wissen1) und 0.54

(Wissen2). Auch dieses Modell wies mit CFI = .95 und RMSEA = .06 eine gute Passung auf. Der χ^2-Test auf die Differenz der Passung beider Modelle ergab $\chi^2_{13} = 27.51$ ($p = .011$). In allen Datensätzen lag für Frauen die Schätzung des Erwartungswertes von Wissen1 unter 0 und die Schätzung des Erwartungswertes von Wissen2 über 0 (siehe Tabellen in Anhang N), wobei die 95%-Konfidenzintervalle den Wert 0 nicht umfassten. Da die Varianzen in sämtlichen Modellen auf 1 fixiert waren, können die Differenzen der Erwartungswerte auch als Effektstärken interpretiert werden. Die Korrelation der beiden Faktoren betrug in allen Modellen .94 oder .95.

5.4.3 Diskussion

Die Überprüfung der Item- und Skalenkennwerte des im Rahmen von Studie 2 entwickelten Wissenstests sind größtenteils zufriedenstellend. Die Mehrzahl der Item- und Skalenparameter des Tests ist vergleichbar mit den Parametern, die aus den Daten der Studie 2 berechnet wurden. So ist die interne Konsistenz des Tests in den für die Studie 3 erhobenen Daten unverändert, und die durchschnittliche Schwierigkeit aller Items unterscheidet sich nur geringfügig. Es ist jedoch auch eine Reihe von Unterschieden festzustellen: Die internen Konsistenzen der Skalen Biologie, Gesundheit und Soziale Arbeit sind deutlich geringer ausgefallen als in Studie 2. Für die Biologie lässt sich das eventuell damit erklären, dass die aktuelle Stichprobe auch Realschülerinnen und -schüler enthielt, die zum Zeitpunkt der Testung weniger Unterricht im Fach Biologie erhalten hatten als die ausschließlich aus Gymnasien stammenden Schülerinnen und Schüler aus Studie 2. Allerdings wäre demnach auch zu erwarten, dass die Anteile korrekt gelöster Items dieses Themas deutlich abnehmen würden, was jedoch nicht eingetreten ist. Die Differenz der durchschnittlichen Schwierigkeiten der Biologie-Items beträgt lediglich .04. Für die Skalen Psychologie und Raumdesign wurde in Studie 3 eine deutliche Erhöhung der internen Konsistenzen

festgestellt. Für das Thema Psychologie lässt sich das möglicherweise durch die Teilnahme von Psychologie-Studierenden in dieser Studie erklären. Die Differenz der durchschnittlichen Schwierigkeiten der Items geht auch hier in die erwartete Richtung, fällt allerdings ebenfalls gering aus. Von den 121 Items des neuen Wissenstests wiesen in der vorliegenden Untersuchung drei Items negative Trennschärfen auf der jeweiligen Skala des entsprechenden Themas auf. Eines dieser drei Items sowie zwei weitere zeigten negative Trennschärfen auf der Gesamtskala. Aufgrund der geringen Anzahl an betroffenen Items erscheint dies jedoch unproblematisch. Hinsichtlich der Kriteriumsvalidität ist der Test in der aktuellen Version im Vergleich zu bestehenden Wissenstests hervorragend. Die Korrelation zwischen Gesamtscore und Schulleistung beträgt .49 und liegt damit deutlich über der Korrelation zwischen dem Gesamtscore des I-S-T Wissenstests und der Schulleistung sowie auch über der entsprechenden Korrelation des BOWIT, über die im Manual des Tests berichtet wird (siehe Kapitel 5.3.3). Insgesamt weisen somit auch die Ergebnisse aus Studie 3 auf eine hohe psychometrische Qualität des neu entwickelten Wissenstests hin.

Die Wissenstestbatterie, die aus dem neu entwickelten Wissenstest und dem Wissenstest des I-S-T bestand, wies in den Parallelanalysen der verschiedenen Datensätze durchgehend eine 2-faktorielle Struktur auf. Während in Studie 2 die Tendenz zur Ausgewogenheit des neuen Wissenstests hinsichtlich der Interessen von Frauen und Männern noch unbeabsichtigt eingetreten war, bestand hierin ein explizites Ziel in Studie 3, das durch die Hinzunahme des I-S-T verfolgt wurde. Die Ergebnisse zeigen, dass durch die Wissenstestbatterie zwei Faktoren gemessen werden, wodurch die Vermutung bekräftigt wird, dass eine einzelne latente Variable eine unzureichende Abbildung des Allgemeinen Wissens darstellt. Dies entspricht den zu Beginn des Kapitels 5.4 geschilderten Annahmen.

In den exploratorischen Faktorenanalysen zeigten die vier Parcels des I-S-T Wissenstests ausschließlich signifikante positive Ladungen auf dem Faktor Wissen1, während die

Schätzungen der Ladungen auf dem Faktor Wissen2 nahe 0 lagen. Die Zuordnung der Skalen des neuen Wissenstests zu den beiden Faktoren war für einige Themen nicht so eindeutig möglich wie für die Parcels des I-S-T, da einige Skalen auf beiden Faktoren signifikante Ladungen zeigten und die Differenzen der Ladungen auf beiden Faktoren nicht so groß ausfielen wie bei den Parcels des I-S-T. Es wurde jedoch versucht, jede Skala nur einem der beiden Faktoren zuzuordnen[6]. Wie in Kapitel 5.4.2 gezeigt, ergaben sich (a) beim Vergleich der beiden Ladungen der jeweiligen Skala auf den Faktoren und anschließender Auswahl des Faktors mit der höheren Ladung sowie (b) bei ausschließlicher Beachtung von Ladungen > .40 die gleichen Zuordnungen (siehe Tabelle 21). Biologie und Medizin konnten eindeutig Wissen1 zugeordnet werden. Darstellende Kunst, Ernährung, Gesundheit, Modedesign, Psychologie, Pädagogik, Raumdesign und Soziale Arbeit konnten eindeutig Wissen2 zugeordnet werden. Für die Variable Natur wurden Ladungen auf beiden Faktoren zugelassen, da hier keine eindeutige Zuordnung möglich war. Diese Verteilung der manifesten Variablen auf die beiden Faktoren erscheint auf Basis der Ergebnisse von Studie 2 größtenteils plausibel. Von den 11 manifesten Variablen des neuen Wissenstests wurden sieben dem gleichen Faktor wie in der vorangegangenen Untersuchung zugeordnet. Für vier Variablen ergaben sich Unterschiede. Für die Themen Psychologie und Ernährung wurde jeweils die Ladung auf dem Faktor Wissen1 eliminiert. Während das Thema Natur in Studie 2 noch Wissen1 zugeordnet wurde, erforderte die beschriebene Vorgehensweise der Zuordnung in Studie 3 eine zusätzliche Ladung auf Wissen2. Das Thema Soziale Arbeit

[6]Es sollte hierbei nicht außer Acht gelassen werden, dass auch alternative Messmodelle gute Passungen aufweisen könnten, in denen auf eine Einfachstruktur der Faktorladungen verzichtet wird, wie beispielsweise hierarchische Modelle. Die Interpretation solcher Modelle ist jedoch aus verschiedenen Gründen häufig problematisch (siehe Schulze, 2005).

wurde in Studie 2 noch Wissen1 zugeordnet und wechselte in Studie 3 zu Wissen2. Da die Testbatterie aus Studie 3 thematisch vielseitiger ausgerichtet war und das Wissen zu Themen der Kategorie IvM eingehender erfasst wurde, sind die Ergebnisse der exploratorischen Faktorenanalyse von Studie 3 den Ergebnissen von Studie 2 vorzuziehen. In Studie 3 wurde das Wissen deutlich umfassender gemessen, da hier auch der Wissenstest des I-S-T verwendet wurde.

Das auf Basis der exploratorischen Faktorenanalysen erstellte Messmodell wurde durch die konfirmatorischen Faktorenanalysen in vollem Umfang bestätigt. Die Analysen beider Schritte ergaben jeweils eine gute Passung, womit auch die Messinvarianz dieser Testbatterie belegt wurde. Die Schätzungen der Erwartungswerte von Wissen1 und Wissen2 für Frauen und Männer entsprachen vollständig den zu Beginn des Kapitels 5.4 hergeleiteten Annahmen: Während Frauen in Wissen1 die durchschnittlich geringeren Leistungen zeigten, wurde ihr Erwartungswert in Wissen2 höher als bei Männern geschätzt. Die Effektstärken lagen bei -0.46 für Wissen1 und 0.54 für Wissen2.

Die für die Studie 3 getroffenen Annahmen wurden somit bestätigt. Für den Wissenstest des I-S-T wurde festgestellt, dass hiermit ein Merkmal erfasst wird, das nur geringfügig die Interessengebiete von Frauen abdeckt, und das bei Männern durchschnittlich höher ausgeprägt ist. Mit dem im Rahmen von Studie 2 entwickelten Wissenstest wird primär ein anderes, stärker an die Interessengebiete von Frauen angelehntes Merkmal erfasst, das bei Frauen durchschnittlich höher ausgeprägt ist. Da die Korrelation zwischen dem neu entwickelten Wissenstest und der Schulleistung höher ausfiel als die Korrelation zwischen dem Wissenstest des I-S-T und der Schulleistung, wurde nachträglich überprüft, wie hoch die beiden latenten Variablen des Messmodells jeweils mit Schulleistung korrelieren. Dem in Abbildung 7 dargestellten Messmodell wurde die manifeste Variable der Schulleistung (in Punkten) hinzugefügt, welche mit beiden latenten Variablen korrelierte. Für dieses

Messmodell wurde anhand der Stichprobe von 161 Personen (52.8% Frauen), deren Schulleistungen erfasst worden waren, nachträglich eine konfirmatorische Faktorenanalyse durchgeführt. In Tabelle O1 sind die Schätzungen der Korrelationen für alle drei Datensätze aufgelistet. In dem Datensatz der zweiten Parcelgruppe betrug die Schätzung der Korrelation zwischen Wissen1 und Schulleistung für die Gesamtstichprobe .38 [.25; .51][7]. Die Korrelation zwischen Wissen2 und Schulleistung wurde auf .49 [.39; .60] geschätzt. Bei geschlechtsspezifischer Berechnung der Korrelationen wurden sowohl für Wissen1 als auch für Wissen2 durchgehend höhere Werte bei Frauen als bei Männern berechnet. Außerdem fällt auf, dass für die Gesamtstichprobe die Punktschätzungen der Korrelation zwischen Wissen2 und Schulleistung höher ausfielen als die Punktschätzungen der Korrelation zwischen Wissen1 und Schulleistung, während bei geschlechtsspezifischer Betrachtung Wissen1 höher als Wissen2 mit der Schulleistung korrelierte. Sämtliche Abweichungen waren jedoch nicht signifikant, wie aus den Konfidenzintervallen in Tabelle O1 hervorgeht. Demnach handelt es sich bei Wissen2 um ein mit der hier verwendeten Testbatterie – nicht jedoch mit dem Wissenstest des I-S-T – erfasstes Merkmal, dessen Zusammenhang mit Schulleistung mit dem Zusammenhang zwischen Wissen1 und Schulleistung vergleichbar ist. Dieser Befund ist durchaus erstaunlich, da die Themen des Faktors Wissen1 wesentlich stärkere Überschneidungen mit den Curricula weiterführender Schulen zeigen als die dem Faktor Wissen2 zugeordneten Themen. So finden sich beispielsweise die Themen Physik und Mathematik, zu denen der I-S-T zahlreiche Items enthält, und das Thema Biologie aus dem neu entwickelten Wissenstest als Schulfächer wieder. Unter den Themen des Faktors Wissen2 gilt dies lediglich für die Pädagogik.

[7]Grenzen des 95%-Konfidenzintervalls.

Die Ergebnisse der dritten Studie können folgendermaßen zusammengefasst werden: (a) Es ist die Entwicklung eines Wissenstests gelungen, der sehr gute psychometrische Eigenschaften aufweist und dessen Korrelation mit der durchschnittlichen Schulleistung die entsprechenden Korrelationen des I-S-T Wissenstests und des BOWIT – teilweise deutlich – übertrifft. (b) Es wurde eine thematisch umfangreiche Batterie aus 2 Wissenstests vorgelegt, die vorrangig auf unterschiedliche Interessengebiete ausgerichtet sind. Diese Testbatterie dient zur Erfassung von 2 Merkmalen, die sehr hoch miteinander korrelieren – zumindest unter Annahme der in den hier dargestellten Messmodellen postulierten Einfachstruktur. (c) Frauen und Männer weisen in den beiden Merkmalen gegenteilige Geschlechterdifferenzen auf. Die Beträge der Effektstärken betragen bei beiden Variablen in etwa .50. (d) Der Wissenstest des I-S-T dient ausschließlich zur Erfassung des Merkmals, das Geschlechterdifferenzen zugunsten von Männern aufweist. Die Ergebnisse weisen somit darauf hin, dass Themengebiete, für welche Frauen sich durchschnittlich stärker als Männer interessieren und in denen sie über ein durchschnittlich höheres Wissen als Männer verfügen, nicht im Wissenstest des I-S-T enthalten sind. Die Korrelation zwischen der Schulleistung und diesen Wissensgebieten, die hier unter Wissen2 zusammengefasst werden, ist jedoch mit der Korrelation zwischen Wissen1 und Schulleistung vergleichbar. Eine Ergänzung des Wissenstests des I-S-T um die Themen des Faktors Wissen2 dürfte nach den vorliegenden Ergebnissen zu einer deutlichen Verringerung der Geschlechterdifferenzen führen, ohne dass bezüglich der Kriteriumsvalidität Einbußen hingenommen werden müssten. Für weitere Wissenstests steht eine entsprechende Überprüfung aus. Im Fall des BOWIT ist jedoch ein ähnliches Ergebnis anzunehmen, da auch dieser Test, wie in Studie 1 gezeigt wurde, hauptsächlich auf Themen der Kategorie IvM ausgerichtet ist.

5.5 Einschränkungen der Studien 1-3

Eine Reihe von Kritikpunkten der in Kapitel 5 präsentierten Studien wird im Folgenden diskutiert. Wie in Abschnitt 5.2.3 erläutert, war die Erstellung der Themenliste für den Interessenfragebogen von hoher Subjektivität geprägt. Für die Auswahl der Bezeichnungen der einzelnen Themen, für den Grad der Spezifität und für die verwendeten beispielhaften Beschreibungen zur Erläuterung der einzelnen Themen sind unbegrenzt weitere Alternativen denkbar. Wie in Kapitel 5.2.2 begründet, kann für den im Rahmen von Studie 1 entwickelten Fragebogen jedoch zumindest der Anspruch auf eine umfassende Erfassung von Interessen erhoben werden.

Die methodische Vorgehensweise bei der Kategorisierung von Themen in IvF, NI und IvM erscheint im Nachhinein ebenfalls verbesserungsfähig: Wie in Kapitel 5.2.1 beschrieben, wurden hierfür die Effektstärken der Mittelwertunterschiede zwischen Frauen und Männern in den einzelnen Themen herangezogen. In Abbildung 5 ist zu erkennen, dass die Themen mehrheitlich als Interessengebiete von Männern klassifiziert wurden. Es stellt sich die Frage, ob eine ähnliche Verteilung der Themen auf die drei Kategorien auch erfolgen würde, wenn anstatt des Vergleichs der absoluten Interessen von Frauen und Männern für das jeweilige Thema die Relationen der Interessen innerhalb der Geschlechtergruppen für die Kategorisierung verwendet worden wären. Nye et al. (2012) weisen in ihrer Untersuchung darauf hin, dass aus hohem Interesse an einem Themengebiet nicht zwangsläufig hohe Investitionen (im Sinne der Investmenttheorie) in dieses Gebiet erfolgen, da möglicherweise höhere Interessen an anderen Themen bestehen und die für die Investitionen zur Verfügung stehenden Ressourcen begrenzt sind. Nach Abschluss der in Kapitel 5 beschriebenen Untersuchungen wurden separat für Frauen und Männer sämtliche Themengebiete des Interessenfragebogens auf Basis der Daten aus Studie 1 in eine Rangordnung gebracht. Die 11 Themen, die auf Basis der Effektstärken als IvF klassifiziert wurden, nehmen bei Frauen

folgende Ränge ein: 2, 3, 4, 5, 8, 9, 11, 12, 15, 17 und 18. Bei Männern sind durch die entsprechenden Themen folgende Rangplätze belegt: 3, 7, 9, 11, 14, 15, 17, 18, 21, 31 und 36. Insgesamt besetzen die 11 Themen somit bei Frauen höhere Rangplätze als bei Männern, was die Kategorisierung als IvF für die Mehrzahl unterstützt. Näher betrachtet werden sollen hier die vier Themen, deren Zuordnung zu den Faktoren Wissen1 und Wissen2 in den Studien 2 und 3 nicht eindeutig möglich war. Hierbei handelte es sich um die Themen Psychologie, Natur, Ernährung und Soziale Arbeit. Das Thema Psychologie nahm bei Frauen Rang 2 und bei Männern Rang 3 ein. Das Thema Natur nahm bei Frauen Rang 8 und bei Männern Rang 9 ein. Ernährung lag bei Frauen auf Rangplatz 5 und bei Männern auf Rangplatz 7. Aufgrund der geringen Abstände in den Rängen wären somit auch Klassifikationen als NI gut begründbar. Soziale Arbeit wurde bei Frauen dem 4. Rang und bei Männern dem 17. Rang zugeordnet, was die Kategorisierung als IvF bestätigt. In Tabelle P1 sind die Rangordnungen der 36 Themen für beide Geschlechter aufgeführt. Es sollte auch bedacht werden, dass die ersten 18 Rangplätze der Themen (die obere Hälfte aller Rangplätze) bei Frauen sieben Themen enthalten, die nicht als IvF klassifiziert wurden. Diese sieben Themen erhielten jedoch alle bei Männern den gleichen Rangplatz (Musik, Literatur, Fremdsprachen) oder einen höheren Rangplatz (Fremde Kulturen, Sport, Gesellschaft, Computer) als bei Frauen. Sämtliche Themen der oberen 18 Rangplätze bei Frauen, die nicht als IvF klassifiziert wurden, wären somit auch der hier beschriebenen alternativen Methodik zufolge nicht als solche klassifiziert worden. Für die unteren 18 Rangplätze der Themen bei Frauen erscheint es ebenfalls fraglich, ob hierfür eine Klassifizierung als IvF angemessen wäre. Zwar finden sich darunter Themen, die in der Gruppe der Männer einen niedrigeren Rangplatz erhalten, wie beispielsweise Religion, jedoch dürfte man nach der Argumentation von Nye et al. nicht davon ausgehen, dass Frauen viel in diese Themen investieren. Die in Studie 1 als IvF klassifizierten Themen liegen alle auf den oberen 18 Rangplätzen der

Themen bei Frauen. Es zeigt sich insgesamt, dass sowohl nach der Auswertung auf Basis von Effektstärken als auch nach der Auswertung auf Basis von Rängen weitgehend dieselben Themen als IvF klassifiziert werden. Daher bietet der Kritikpunkt zwar eine wünschenswerte Verbesserung für künftige Untersuchungen, rechtfertigt jedoch keine Zweifel an der Brauchbarkeit der in Studie 1 präsentierten Ergebnisse.

Ein weiterer Kritikpunkt ist die problematische Zuordnung einzelner Themengebiete des neuen Wissenstests zu den beiden Faktoren. Hier bedarf es der Erhebung größerer Stichproben, um zu verlässlichen Ergebnissen zu gelangen. Auch stellt sich die Frage, wie die Zuordnungen der Skalen bei Verwendung anderer Wissenstests ausfallen würden. In der hier vorgestellten Untersuchung wurde lediglich der Wissenstest des I-S-T in die Testbatterie aufgenommen, um die Frage nach der faktoriellen Struktur des Wissens zu beantworten. Für eine zusätzliche Bestätigung der beschriebenen Befunde bedarf es der Replikation mit weiteren Wissenstests, wie beispielsweise dem BOWIT. Dieser Test würde außerdem die Bildung manifester Variablen in Form von thematischen Facetten erlauben, was aufschlussreich sein könnte, da hiermit eine differenziertere Zuweisung von Themen – auch von Themen der Kategorien NI und IvM – ermöglicht würde. Die Ergebnisse von Studie 3 lassen vermuten, dass die Mehrzahl der Facetten des BOWIT dem hier als Wissen1 bezeichneten Faktor zugeordnet würde, während die Skalen des neu entwickelten Wissenstests mehrheitlich auf dem anderen Faktor, der hier als Wissen2 bezeichnet wurde, höhere Ladungen aufweisen dürften.

Schlussfolgerungen

In der vorliegenden Arbeit wurden zwei potentielle Erklärungsmöglichkeiten für Geschlechterdifferenzen in der Messung des Allgemeinen Wissens überprüft. Hierbei handelte es sich um Unterschiede zwischen Frauen und Männern in Interessengebieten sowie in der Einschätzung des eigenen Wissens. Im Folgenden werden zunächst die Ergebnisse aller Untersuchungen und die entsprechenden Befunde zusammengefasst. Daran anschließend erfolgt eine nähere Betrachtung des Problems der Operationalisierung von Allgemeinem Wissen. Dieses Thema rückte im Verlauf der in Kapitel 5 beschriebenen Untersuchungen in den Mittelpunkt. Zuletzt wird ein Fazit aus allen Untersuchungen der Arbeit gezogen.

6.1 Zusammenfassung der Ergebnisse

Im Rahmen des Experiments, welches Gegenstand von Kapitel 4 war, wurde die Selbsteinschätzung der Probandinnen und Probanden manipuliert und anschließend ein Wissenstest vorgelegt. Die Hypothese lautete, dass Frauen ihr eigenes Wissen geringer einschätzen als Männer und dass geringere Selbsteinschätzung des Wissens von ursächlicher Bedeutung für geringere Leistung im Wissenstest ist. Es wurde postuliert, dass die Bearbeitung von Wissenstests mit einer Beanspruchung des Arbeitsgedächtnisses einhergeht und dass geringere Selbsteinschätzung zu Leistungsminderung führt, da hiermit verbundene negative Emotionen die zum Testzeitpunkt verfügbaren Kapazitäten des Arbeitsgedächtnisses einschränken. Die Annahme der geringeren Selbsteinschätzung des Wissens von Frauen und ein positiver Zusammenhang zwischen Selbsteinschätzung des Wissens und Leistung im Wissenstest wurden durch die Ergebnisse unterstützt. Jedoch erreichten die Teilnehmenden, deren Selbsteinschätzung durch Manipulation verringert wurde, durchschnittlich höhere

Punktwerte in dem Wissenstest als die Teilnehmenden, deren Selbsteinschätzung manipulativ erhöht wurde. In Kapitel 4.4 wurden zahlreiche Erklärungsmöglichkeiten für das vorliegende Ergebnis beschrieben, wie beispielsweise eine zu kurze Wirksamkeit der Manipulation der Selbsteinschätzung, eine Kausalität in umgekehrter Richtung oder die Ungewissheit, inwieweit die Bearbeitung von Tests des Allgemeinen Wissens das Arbeitsgedächtnis beansprucht. Für die vorliegende Arbeit wurde die Hypothese, dass gemessene Geschlechterdifferenzen des Allgemeinen Wissens durch Geschlechterdifferenzen der Selbsteinschätzung des Allgemeinen Wissens verzerrt werden, verworfen.

Die im Rahmen des fünften Kapitels geprüfte Hypothese lautete, dass Interessenunterschiede zwischen Frauen und Männern eine Erklärung für die einseitigen Geschlechterdifferenzen in bestehenden Wissenstests sind. Es wurde postuliert, dass bestehende Wissenstests, die deutliche Geschlechterdifferenzen zugunsten von Männern aufweisen, vorrangig Items zu Interessengebieten enthalten, für die Männer sich durchschnittlich stärker als Frauen interessieren, während das Wissen zu Themen, an denen Frauen durchschnittlich mehr als Männer interessiert sind, in deutlich geringerem Umfang erfasst wird. Unter Annahme der Investmenttheorie liegt der Schluss nahe, dass diese Unausgewogenheit zu Leistungsunterschieden zwischen Frauen und Männern führt. In Studie 1 wurde die Hypothese geprüft, dass zwei deutschsprachige Wissenstests, der BOWIT und der Wissenstest des I-S-T 2000 R, in welchen Männer durchschnittlich höhere Scores erreichen als Frauen, eine mangelnde Balance in der Anzahl an Items zu den geschlechtsspezifischen Interessengebieten aufweisen. Die Ergebnisse zeigten zum einen, dass beide Tests deutlich mehr Items zur Kategorie IvM als zur Kategorie IvF enthalten, und zum anderen, dass die Leistungsunterschiede zwischen Frauen und Männern in den verschiedenen Kategorien deutlich variieren. In der Kategorie IvM fielen sie am stärksten und in der Kategorie IvF am geringsten aus. Allerdings beruhten die Ergebnisse der

Kategorie IvF in beiden Tests jeweils auf sehr wenigen Items. Daher wurde in Studie 2 ein neuer Wissenstest entwickelt, der ausschließlich Items zu Themen dieser Kategorie enthielt. Nach der Auswahl von 121 Items lag ein Test mit guten psychometrischen Kennwerten vor. Ergebnisse von Faktorenanalysen ließen Zweifel daran aufkommen, dass lediglich ein einzelnes psychologisches Merkmal für die Leistung in diesem Test relevant ist. Ein alternatives Modell beinhaltete zwei latente Variablen, die gegenteilige Geschlechterdifferenzen aufwiesen. In Studie 3 wurde die 2-faktorielle Struktur anhand einer thematisch breit gefassten Batterie aus zwei Wissenstests – dem neu entwickelten Wissenstest und dem Wissenstest des I-S-T 2000 R – überprüft. Die 2-faktorielle Struktur und die umgekehrten Geschlechterdifferenzen auf beiden Faktoren wurden hier repliziert. Außerdem wurde festgestellt, dass für die Leistung im I-S-T ausschließlich das Merkmal relevant ist, für das Männern ein höherer Erwartungswert als Frauen attestiert wurde. Die Ergebnisse der drei Untersuchungen bestätigen die Annahme, dass Geschlechterdifferenzen in Wissenstests teilweise durch die Unausgewogenheit der Interessengebiete erklärt werden können.

6.2 Operationalisierung des Konstrukts Wissen

Zu Beginn des Kapitels 3.2 wurde auf die Differenzierung zwischen zwei verschiedenen Gruppen von Erklärungsansätzen hingewiesen. Hierbei handelt es sich einerseits um Erklärungen für eventuelle systematische Verzerrungen bei der Messung des Wissens und andererseits um Erklärungen für als gegeben angenommene Geschlechterdifferenzen im Wissen. Die in Kapitel 5 beschrieben Ergebnisse weisen auf eine weitere Erklärungsmöglichkeit hin, die keiner dieser beiden Kategorien eindeutig zugeordnet werden kann. Vielmehr besteht Anlass, die Operationalisierung des Konstruktes Wissen kritisch zu hinterfragen. Der Wissenstest des I-S-T erfasst offensichtlich ein Merkmal zuverlässig und

valide, für das Männer eine durchschnittlich höhere Ausprägung vorweisen. Eine Verzerrung bei der Messung findet hier nicht statt, da es den Ergebnissen der hier vorgestellten Untersuchungen zufolge offensichtlich auch bei einer thematisch sehr heterogenen und hinsichtlich der Interessen von Frauen und Männern ausgewogenen Diagnostik des Wissens relevant ist. Die Ergebnisse bestätigen die Annahme, dass Männer eine durchschnittlich höhere Ausprägung dieses Merkmals aufweisen. Das Problem liegt jedoch in der Interpretation: die Schlussfolgerung, dass Männer im Vergleich zu Frauen über ein durchschnittlich besseres Allgemeines Wissen verfügen, ist problematisch, da die Untersuchungen gezeigt haben, dass bei der Bearbeitung thematisch vielfältiger Wissenstests auch ein zweites Merkmal relevant ist, bei dem Frauen im Durchschnitt die höhere Ausprägung aufweisen, und das bei der Bearbeitung des I-S-T Wissenstests nicht bedeutsam ist. Im Vorfeld der Studien 2 und 3 war erwartet worden, dass die Hinzunahme von Themen der Kategorie IvF die Geschlechterdifferenzen im Wissen, was als eine einzelne Fähigkeit betrachtet wurde, reduzieren würde. Die Ergebnisse zeigen jedoch, dass die Operationalisierung des Wissens in Form eines einzelnen Merkmals angezweifelt werden muss. Der Wissenstest des I-S-T misst valide das hier als Wissen1 bezeichnete Merkmal, welches jedoch lediglich einen Teilbereich des Wissens abdeckt. Die thematisch breitgefasste Testbatterie, die in Studie 3 verwendet wurde, erfasst offensichtlich zwei Merkmale, die sehr hoch miteinander korrelieren.

Sofern man an der Auffassung des Allgemeinen Wissens als einer einzelnen Fähigkeit festhält, liegt es nahe, auch der Testbatterie aus Studie 3 ein 1-faktorielles Modell zugrunde zu legen. Um die Geschlechterdifferenzen in dem Faktor zu überprüfen, wurde nachträglich eine konfirmatorische Faktorenanalyse für Frauen und Männer durchgeführt, bei der das Messmodell lediglich einen einzelnen Faktor umfasste. Die vier Parcels des I-S-T und die 11 Summenscores der Themen des neu entwickelten Wissenstests bildeten die manifesten

Variablen. Sämtliche Parameter des Modells, mit Ausnahme des Erwartungswerts des Faktors, waren für Frauen und Männer gleichgesetzt, und die Varianz des Faktors war auf 1 fixiert. Der Eigenwert des Faktors wurde für Männer auf 0 fixiert und bei Frauen frei geschätzt. Die Schätzungen lagen in den drei Datensätzen zwischen 0.04 und 0.06, wobei das 95%-Konfidenzintervall jeweils die 0 umfasste. Alle drei Modelle zeigten jedoch erwartungsgemäß mit CFI $\leq$ .85 und RMSEA $\geq$ .10 eine unzureichende Modellpassung. In Anhang Q sind sämtliche standardisierten Parameterschätzungen und die Ergebnisse der Prüfungen des Model-Fits für dieser Faktorenanalysen aufgeführt.

Die Differenzierung zwischen zwei verschiedenen Fähigkeiten, die beide als Wissen bezeichnet werden können, wurde bereits von Lynn und Irwing (2002) angedeutet, in deren Analysen aber nicht konsequent berücksichtigt. So enthielt das Modell trotz der Hinweise auf mehrere für die Leistung im GKT relevante Merkmale, die von den Autoren als verschiedene Arten des Allgemeinen Wissens bezeichnet wurden (siehe Kapitel 3.1.1), lediglich einen einzelnen Faktor höherer Ordnung, der als general knowledge bezeichnet wurde. Diese Interpretation erscheint aufgrund der vorliegenden Ergebnisse jedoch fragwürdig. Wie in Kapitel 2.3.3 geschildert, sah Cattell (1971/1987) Schwierigkeiten in der Diagnostik der Kristallinen Intelligenz nach Abschluss der Schule, die in der thematischen Vielfalt von gc begründet waren. Das gleiche Problem zeigt sich auch bei der Interpretation von Wissenstests als Messinstrumente des Allgemeinen Wissens beziehungsweise der Kristallinen Intelligenz. In diesem Zusammenhang sei an das Problem der Inhaltsvalidität – und damit auch an das Problem der Interpretationsobjektivität – von Wissenstests erinnert, die in Kapitel 2.3.2 näher beschrieben wurden. Auf welchem Weg könnte ein Itemuniversum für einen Wissenstest bestimmt werden, der zur Erfassung des Allgemeinen Wissens eingesetzt werden soll? Die hier berichteten Ergebnisse weisen darauf hin, dass die Verwendung der curricularen Validität keine Lösung des Problems darstellt, sondern zu einer Einschränkung der Themen

führt, zu denen das Wissen erfasst wird. Der im Rahmen der vorliegenden Arbeit entwickelte Wissenstest war nur zu einem geringen Anteil an Curricula angelehnt (siehe Kapitel 5.3.1). Dennoch ist die Qualität des Tests mit der Qualität etablierter Wissenstests vergleichbar.

Die verwendete Wissenstestbatterie weist eine 2-faktorielle Struktur auf, was als Grund für Zweifel an der Konstruktvalidität aufgefasst werden könnte. So mag argumentiert werden, dass die Testbatterie neben Allgemeinem Wissen (Wissen1) ein weiteres, nicht definiertes Merkmal (Wissen2) erfasst. Der hier als Wissen2 bezeichnete Faktor eignet sich, wie in Kapitel 5.4.3 beschrieben wurde, jedoch ebenso gut zur Vorhersage schulischer Leistungen wie Allgemeines Wissen. Aus diesem Grund erscheint die beschriebene Kritik an der Konstruktvalidität der Testbatterie kaum überzeugend. Nicht die Konstruktvalidität der Testbatterie ist in Frage zu stellen, sondern vielmehr die Auffassung des Konstrukts Wissen als einzelnes psychologisches Merkmal. Es wurden in den hier beschriebenen Untersuchungen Hinweise darauf gefunden, dass die Auffassung von Allgemeinem Wissen als einzelne Fähigkeit keine hinreichende Abbildung der Realität ist. Angemessener erscheint die Differenzierung zwischen zwei verschiedenen Fähigkeiten, welche als Wissen in unterschiedlichen Themengebieten aufgefasst werden können. Die beiden Fähigkeiten korrelieren sehr hoch miteinander, sind aufgrund der gegenteiligen Geschlechterdifferenzen jedoch nicht redundant. Somit bleibt festzuhalten, dass die in Kapitel 3.1 beschriebenen Belege für ein durchschnittlich geringeres Allgemeines Wissen von Frauen mit Vorsicht interpretiert werden sollten.

Der Befund der 2-faktoriellen Struktur wird auch durch die exploratorische Faktorenanalyse des BOWIT unterstützt. Wie in Kapitel 2.3.2 beschrieben, wurden die zwei Faktoren des BOWIT von Hossiep und Schulte (2008) als naturwissenschaftlich-technisches Wissen und als gesellschafts- und geisteswissenschaftliches Wissen interpretiert. Die hier vorgestellten Faktoren Wissen1 und Wissen2 können jedoch nicht entsprechend bezeichnet

werden, obwohl sich vereinzelt Hinweise auf derartige Interpretationsmöglichkeiten finden. So umfasst der I-S-T zahlreiche Items zum Thema Physik. Auch die Biologie ist eine Naturwissenschaft. Die Themen Soziale Arbeit, Pädagogik und Psychologie können unter dem Begriff der Geisteswissenschaften zusammengefasst werden. Für andere Themengebiete, wie beispielsweise Darstellende Kunst oder Modedesign ist dieser Oberbegriff jedoch unpassend. Aus den Normtabellen des BOWIT lässt sich außerdem schließen, dass die Erwartungswerte beider Faktoren bei Frauen geringer ausfallen als bei Männern, was ebenfalls ein Argument gegen eine entsprechende Interpretation der Faktoren Wissen1 und Wissen2 darstellt. Vielmehr sind deutliche Tendenzen zu interessenspezifischen Faktoren erkennbar. So wurden lediglich zwei als IvF-Themen kategorisierte Skalen dem Faktor Wissen1 zugeordnet, während acht Skalen derselben Kategorie einen eigenen Faktor bildeten.

Weitere Untersuchungen zur Struktur des Wissens sollten der Frage nachgehen, ob durch die Hinzunahme von Items zu weiteren Themengebieten in Wissenstests weitere Faktoren aufgedeckt würden, oder ob die Zweidimensionalität beibehalten würde. Sollte sich die enge Verknüpfung zwischen einerseits Interessenschwerpunkten von Frauen und Männern und andererseits psychologischen Merkmalen, die als Themengebiete des Wissens aufgefasst werden können, weiterhin bestätigen, so wäre eine interessante Forschungsfrage, inwieweit die beiden Merkmale des Wissens mit den Polen der people-things-Dimension nach Prediger (1982) zusammenhängen. Die people-things-Dimension stammt aus dem Bereich der Interessenforschung und es erscheint widersprüchlich, dass die Begriffe people und things die Pole einer einzelnen Dimension bilden, während für das Konstrukt Wissen die Eindimensionalität aufgegeben werden sollte. Entsprechend ist auch die hohe positive Korrelation zwischen Wissen1 und Wissen2 ebenfalls schwierig mit der people-things-Dimension in Einklang zu bringen. Möglicherweise bieten Persönlichkeitsmerkmale, die mit Wissenserwerb in Zusammenhang stehen, wie beispielsweise Typical Intellectual

Engagement, hierfür eine Erklärung. So wäre etwa bei hoher Ausprägung dieses Merkmals ein stark ausgeprägter Wissenserwerb in zwei verschiedenen Themenbereichen denkbar, die im Hinblick auf Interessen die Pole einer einzelnen Dimension bilden. Es sei hier jedoch betont, dass es sich hierbei lediglich um Spekulationen handelt, die weiterer Forschung bedürfen.

6.3 Fazit

Die Ergebnisse der in dieser Arbeit vorgestellten Untersuchungen lassen folgendes Fazit zu: Die Annahme, dass Unterschiede zwischen Frauen und Männern in der Einschätzung des eigenen Wissens von kausaler Bedeutung für gemessene Geschlechterdifferenzen des Allgemeinen Wissens sind, muss vorläufig verworfen werden. Die Annahme, dass die in Tests des Allgemeinen Wissens vorzufindenden Geschlechterdifferenzen teilweise durch Interessenunterschiede zwischen den Geschlechtern und einen Mangel an Ausgewogenheit der entsprechenden Themen in den Tests erklärt werden können, fand hier Unterstützung. Eine ausbalancierte Prüfung von Wissen in Themengebieten, für die Frauen beziehungsweise Männer vergleichsweise höhere Interessen angeben, beinhaltet die Erfassung von zwei Merkmalen, die sehr hoch miteinander korrelieren, in denen die Geschlechter jedoch gegenteilige Geschlechterdifferenzen aufweisen. Während das eine Merkmal bei Frauen durchschnittlich höher ausgeprägt ist als bei Männern, verhält es sich bei dem anderen Merkmal umgekehrt. Ein Messinstrument, mit dem beide Merkmale erfasst werden, erwies sich als ebenso guter Prädiktor der Schulleistung wie ein etablierter Wissenstest, in dem Männer im Durchschnitt die höheren Scores erreichen. Geht man davon aus, dass die hier als Wissen1 und Wissen2 bezeichneten Merkmale jeweils das Wissen in speziellen Themenbereichen umfassen, und nimmt man außerdem an, dass beide Merkmale im Hinblick auf Allgemeines Wissen in gleichem Ausmaß bedeutsam sind, dann weisen Frauen und

Männer in ihrem Allgemeinen Wissen den Befunden der hier vorgestellten Untersuchungen zufolge keine nennenswerten Unterschiede auf. Ob allerdings Wissen1 und Wissen2 in gleichem Ausmaß bedeutsam sind, mag aus verschiedenen Gründen angezweifelt werden. Wie in Kapitel 2.3.3 dargelegt, bezog sich Cattell (1971/1987) mit der Investmenttheorie im Wesentlichen auf Investitionen während der Schulzeit und sah drei Alternativen, Kristalline Intelligenz bei Erwachsenen nach der Schulzeit zu messen. Die dritte Alternative beinhaltete den fortgesetzten Bezug auf Wissensinhalte, welche im Schulalter erworben wurden. Aus dieser Perspektive erscheint es fraglich, die Faktoren Wissen1 und Wissen2 in gleichem Ausmaß in ihrer Bedeutsamkeit zu gewichten, da Wissen1 wesentlich stärker an Lehrpläne von Schulen angelehnt ist. Andererseits betonte beispielsweise Horn (1988) die thematische Heterogenität der Kristallinen Intelligenz (siehe Kapitel 2.3.3). Auch der in Studie 3 berichtete Befund, dass Wissen1 und Wissen2 in vergleichbarer Höhe mit der Schulleistung korrelieren, legt nahe, bei der Wissensdiagnostik auf eine Beschränkung der Inhalte auf Wissen1 zu verzichten. Allerdings bedarf dieser Befund nachdrücklich der Replikation, da er nicht plausibel erscheint.

Es muss außerdem abgewogen werden, ob unter dem Begriff des Allgemeinen Wissens lediglich eine möglichst große Breite an Themen aufgefasst wird, oder ob zusätzlich Wichtigkeit, gesellschaftliche Relevanz und gesellschaftliche Anerkennung, die manchen Themengebieten mehr als anderen zukommen dürften, bei der Auslegung des Begriffs berücksichtigt werden. In der vorliegenden Arbeit wurde hierauf verzichtet, da das Ziel in der Erweiterung der thematischen Bandbreite auf Basis der Interessen von Frauen und Männern lag.

Eine weitere Voraussetzung für eine vergleichbare Bedeutsamkeit von Wissen1 und Wissen2 wäre ein vergleichbare thematische Bandbreite der beiden Merkmale. Sofern sich Wissen1 und Wissen2 hierin unterscheiden, würde die Reduktion der Themen auf zwei

Faktoren und die gleiche Gewichtung der beiden Faktoren für das sogenannte "Allgemeine Wissen" eine artifizielle Reduktion der Geschlechterdifferenzen mit sich bringen. Unabhängig davon, ob hiermit Einschränkungen der Kriteriumsvalidität im Hinblick auf Schulleistung einhergingen, würde dies eine Verzerrung der Realität implizieren. Die hier vorgestellten Untersuchungen lassen keinen Schluss darüber zu, inwieweit Wissen1 und Wissen2 eine vergleichbare thematische Bandbreite aufweisen.

Aus den genannten Gründen ist die Frage, ob Wissen1 und Wissen2 in gleichem Ausmaß bedeutsam sind, derzeit nicht zu beantworten. Allerdings zeigen die Ergebnisse, dass die Verwendung des Begriffs "Allgemeines Wissen" kritisch hinterfragt werden sollte. Handelt es sich bei Allgemeinem Wissen um eine einzelne Fähigkeit? Sofern man sich an der Höhe der Korrelation zwischen Wissen1 und Wissen2 orientiert, erscheint eine Differenzierung zwischen zwei Merkmalen nicht sinnvoll. Die Relationen beider Merkmale zur schulischen Leistung sind dementsprechend vergleichbar. Sofern man sich an den Geschlechterdifferenzen der Erwartungswerte von Wissen1 und Wissen2 orientiert, ist eine Unterscheidung jedoch nutzbringend.

Unabhängig davon, ob Allgemeines Wissen als mehrdimensional oder als ein einzelnes Merkmal zu betrachten ist, weisen die in der vorliegenden Arbeit dargestellten Ergebnisse darauf hin, dass Interessenunterschiede zwischen Frauen und Männern und eine hohe Gewichtung der Themengebiete, für die Männer sich durchschnittlich stärker interessieren, von maßgeblicher Bedeutung für die Geschlechterdifferenzen in Tests des Allgemeinen Wissens sind. Auch dieser Befund bedarf der Replikation mit weiteren Tests. Inwieweit die hohe Gewichtung der Interessengebiete von Männern und damit die einseitigen Geschlechterdifferenzen in Tests des Allgemeinen Wissens zugunsten von Männern eine Verzerrung der Realität beinhalten, hängt wesentlich davon ab, wie der Begriff des Allgemeinen Wissens definiert wird.

Literaturverzeichnis

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*, 227-257. doi:10.1016/S0160-2896(96)90016-1

Ackerman, P. L. (2000a). Domain-specific knowledge as the "dark matter" of adult intelligence: Gf/Gc, personality and interest correlates. *Journal of Gerontology: Psychological Sciences, 55B*, P69-P84. doi:10.1093/geronb/55.2.P69

Ackerman, P. L. (2000b). Traits and knowledge as determinants of learning and individual differences: Putting it all together. In P. L. Ackerman, P. C. Kyllonen & R. D. Roberts (Hrsg.), *Learning and individual differences: Process, trait, and content determinants* (S. 437-462). Washington, DC: American Psychological Association.

Ackerman, P. L. & Beier, M. E. (2005). Knowledge and intelligence. In O. Wilhelm & R. W. Engle (Hrsg.), *Handbook of understanding and measuring intelligence* (S. 125-140). Thousand Oaks, CA: Sage.

Ackerman, P. L., Bowen, K. R., Beier, M. E. & Kanfer, R. (2001). Determinants of individual differences and gender differences in knowledge. *Journal of Educational Psychology, 93*, 797-825. doi:10.1037/0022-0663.93.4.797

Ackerman, P. L. & Heggestad, E. D. (1997). Intelligence, personality, and interest: Evidence for overlapping traits. *Psychological Bulletin, 121*, 219-245. doi:10.1037/0033-2909.121.2.219

Ambrosini, V. & Bowman, C. (2001). Tacit knowledge: Some suggestions for operationalization. *Journal of Management Studies, 38,* 811-829. doi:10.1111/1467-6486.00260

Amelang, M. & Bartussek, D. (2001). *Differentielle Psychologie und Persönlichkeitsforschung* (5. Aufl.). Stuttgart: Kohlhammer.

Anderson, J. R. (2001). *Kognitive Psychologie* (3. Aufl.) (R. Graf & J. Grabowski, Übers.). Heidelberg: Spektrum. (Original erschienen 2000: Cognitive psychology and its implications)

Ashcraft, M. H. & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General, 130,* 224-237. doi:10.1037/0096-3445.130.2.224

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12,* 411-434. doi:10.1207/s15328007sem1203_4

Baddeley, A. D. (1968). A 3 min reasoning test based on grammatical transformation. *Psychonomic Science, 10*, 341-342. doi:10.3758/BF03331551

Baddeley, A. D. (1986). *Working memory*. New York: Clarendon Press.

Baddeley, A. D. (1997). *Human memory: Theory and practice* (Rev. ed.). Hove: Psychology Press.

Baltes, P. B. & Smith, J. (1990). Toward a psychology of wisdom and its ontogenesis. In R. J. Sternberg (Hrsg.), *Wisdom: Its nature, origins, and development* (S. 87-120). Cambridge University Press.

Baumann, P. (2006). *Erkenntnistheorie* (2. Aufl.). Stuttgart: Metzler.

Beauducel, A., Brocke, B. & Liepmann, D. (2001). Perspectives on fluid and crystallized intelligence: Facets for verbal, numerical, and figural intelligence. *Personality and Individual Differences, 30,* 977-994. doi:10.1016/S0191-8869(00)00087-8

Beauducel, A. & Kersting, M. (2002). Fluid and crystallized intelligence and the Berlin Model of Intelligence Structure (BIS). *European Journal of Psychological Assessment, 18*, 97-112. doi:10.1027//1015-5759.18.2.97

Beauducel, A. & Süß, H.-M. (2011). Wissensdiagnostik: Allgemeine und spezielle Wissenstests. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Leistungs-, Intelligenz- und Verhaltensdiagnositk* (Enzyklopädie der Psychologie, Serie Psychologische Diagnostik, Bd. 3, S. 235-273). Göttingen: Hogrefe.

Beauducel, A. & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*, 41-75. doi:10.1207/s15328007sem1201_3

Bellinger, G., Castro, D. & Mills, A. (2004). *Data, information, knowledge, and wisdom*. Verfügbar unter http://www.systems-thinking.org/dikw/dikw.htm

Bennet, J. & Hogarth, S. (2009). Would you want to talk to a scientist at a party? High school students' attitudes to school science and to science. *International Journal of Science Education, 31*, 1975-1998. doi:10.1080/09500690802425581

Ben-Shakhar, G. & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*, 23-35. doi:10.1111/j.1745-3984.1991.tb00341.x

Bergin, D. A. (1999). Influences on classroom interest. *Educational Psychologist, 34*, 87-98. doi:10.1207/s15326985ep3402_2

Bergmann, C. & Eder, F. (1992). *Allgemeiner Interessen-Struktur-Test. Umwelt-Struktur-Test*. (2., korrigierte Aufl.). Weinheim: Beltz.

Bergmann, C. & Eder, F. (2005). *Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R) - Revision*. Göttingen: Beltz Test.

Bonnot, V. & Croizet, J.-C. (2007). Stereotype internalization and women's math performance: The role of interference in working memory. *Journal of Experimental Social Psychology, 43*, 857-866. doi:10.1016/j.jesp.2006.10.006

Broadbent, D. E., FitzGerald, P. & Broadbent, M. H. P. (1986). Implicit and explicit knowledge in the control of complex systems. *British Journal of Psychology, 77*, 33-50. doi:10.1111/j.2044-8295.1986.tb01979.x

Buchner, A. & Brandt, M. (2002). Gedächtniskonzeptionen und Wissensrepräsentationen. In J. Müsseler & W. Prinz (Hrsg.), *Allgemeine Psychologie* (S. 493-543). Heidelberg: Spektrum.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*, 153-193. doi:10.1037/h0059973

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22. doi:10.1037/h0046743

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action* (Rev. ed.). Amsterdam: North Holland. (Original veröffentlicht 1971)

Chamorro-Premuzic, T., Furnham, A. & Ackerman, P. L. (2006). Ability and personality correlates of general knowledge. *Personality and Individual Differences, 41*, 419-429. doi:10.1016/j.paid.2005.11.036

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, A. M. & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*, 240-247. doi:10.1016/S0022-5371(69)80069-1

Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684. doi:10.1016/S0022-5371(72)80001-X

Dweck, C. S. & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256-273. doi:10.1037/0033-295X.95.2.256

Einarsdóttir, S. & Rounds, J. (2009). Gender bias and construct validity in vocational interest measurement: Differential item functioning in the Strong Interest Inventory. *Journal of Vocational Behavior, 74*, 295-307. doi:10.1016/j.jvb.2009.01.003

Engelberg, P. M. (2008). *Fairness des I-S-T 2000 R Wissenstests bei Frauen und Männern* (Unveröffentlichte Diplomarbeit). Bergische Universität Wuppertal.

Evans, E. M., Schweingruber, H. & Stevenson, H. W. (2002). Gender differences in interest and knowledge acquisition: The United States, Taiwan, and Japan. *Sex Roles, 47*, 153-167. doi:10.1023/A:1021047122532

Eysenck, M. W. & Calvo, M. G. (1992). Anxiety and performance: The processing efficiency theory. *Cognition and Emotion, 6*, 409-434. doi:10.1080/02699939208409696

Feingold, A. (1993). Cognitive gender differences: A developmental perspective. *Sex Roles, 29*, 91-112. doi:10.1007/BF00289998

Fouad, N. A. (1999). Validity evidence for interest inventories. In M. L. Savickas & A. R. Spokane (Hrsg.), *Vocational interests: Meaning, measurement, and counseling use* (S. 193-209). Palo Alto, CA: Davies-Black.

Freund, P. A. & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin, 138*, 296-321. doi:10.1037/a0026556

Furnham, A. (2001). Self-estimates of intelligence: Culture and gender difference in self and other estimates of both general (g) and multiple intelligences. *Personality and Individual Differences, 31*, 1381-1405. doi:10.1016/S0191-8869(00)00232-4

Furnham, A., Christopher, A. N., Garwood, J. & Martin, G. N. (2007). Approaches to learning and the acquisition of general knowledge. *Personality and Individual Differences, 43*, 1563-1571. doi:10.1016/j.paid.2007.04.013

Furnham, A., Monsen, J. & Ahmetoglu, G. (2009). Typical intellectual engagement, Big Five personality traits, approaches to learning and cognitive ability predictors of academic performance. *British Journal of Educational Psychology, 79*, 769-782. doi:10.1348/978185409X412147

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis, 23*, 121-123. doi:10.2307/3326922

Gjerde, P. F. & Cardilla, K. (2009). Developmental implications of openness to experience in preschool children: Gender differences in young adulthood. *Developmental Psychology, 45*, 1455-1464. doi:10.1037/a0016714

Godden, D. R. & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66*, 325-331. doi:10.1111/j.2044-8295.1975.tb01468.x

Goff, M. & Ackerman, P. L. (1992). Personality-intelligence relations: Assessment of typical intellectual engagement. *Journal of Educational Psychology, 84*, 537-552. doi:10.1037/0022-0663.84.4.537

Gottfredson, G. D. & Holland, J. L. (1978). Toward beneficial resolution of the interest inventory controversy. In C. K. Tittle & D. Zytowski (Hrsg.), *Sex-fair interest measurement: Research and implications* (S. 43-51). Washington, DC: National Institute of Education.

Graham, J. W., Taylor, B. J., Olchowski, A. E. & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11*, 323-343. doi:10.1037/1082-989X.11.4.323

Halpern, D. (2012). *Sex differences in cognitive abilities* (4th ed.). New York: Psychology Press.

Hassmén, P. & Hunt, D. P. (1994). Human self-assessment in multiple-choice testing. *Journal of Educational Measurement, 31*, 149-160. doi:10.1111/j.1745-3984.1994.tb00440.x

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77-89. doi:10.1080/19312450709336664

Hidi, S., Renninger, K. A. & Krapp, A. (2004). Interest, a motivational variable that combines affective and cognitive functioning. In Y. D. Dai & R. L. Sternberg (Hrsg.), *Motivation, emotion and cognition* (S. 89-115). Mahwah, NJ: Erlbaum.

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6*, 35-45. doi:10.1037/h0040767

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Ressources.

Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade & R. B. Cattell. (Hrsg.), *Handbook of multivariate experimental psychology* (2nd ed., S. 645-685). New York: Plenum Press.

Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*, 253-270. doi:10.1037/h0023816

Horn, J. L. & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26*, 107-129. doi:10.1016/0001-6918(67)90011-X

Horn, J. L. & Cattell, R. B. (1982). Whimsy and misunderstanding of gf-gc theory: A comment on Guilford. *Psychological Bulletin, 91*, 623-633. doi:10.1037/0033-2909.91.3.623

Horn, J. L. & Noll, J. (1997). Human cognitive abilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Hrsg.), *Contemporary intellectual assessment: Theories, tests and issues* (S. 53-91). New York: Guilford Press.

Hossiep, R. & Schulte, M. (2008). *Bochumer Wissenstest*. Göttingen: Hogrefe.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*, 581-592. doi:10.1037/0003-066X.60.6.581

Hyde, T. S. & Jenkins, J. J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behavior, 12*, 471-480. doi:10.1016/S0022-5371(73)80027-1

Irwing, P., Cammock, T. & Lynn, R. (2001). Some evidence for the existence of a general factor of semantic memory and its components. *Personality and Individual Differences, 30*, 857-871. doi:10.1016/S0191-8869(00)00078-7

Jäger, A. O., Süß, H.-M. & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test. BIS-Test, Form 4*. Göttingen: Hogrefe.

Johnson-Laird, P. N., Herrmann, D. J. & Chaffin, R. (1984). Only connections: A critique of semantic networks. *Psychological Bulletin, 96*, 292-315. doi:10.1037/0033-2909.96.2.292

Kanfer, R. & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*, 657-690. doi:10.1037/0021-9010.74.4.657

Kanfer, R. & Ackerman, P. L. (1996). A self-regulatory skills perspective to reducing cognitive interference. In I. G. Sarason, G. R. Pierce & B. R. Sarason (Hrsg.), *Cognitive interference: Theories, methods, and findings* (S. 153-171). Hillsdale, NJ: Erlbaum.

Kanfer, R. & Ackerman, P. L. (2000). Individual differences in work motivation: Further explorations of a trait framework. *Applied Psychology, 49*, 470-482. doi:10.1111/1464-0597.00026

Kanfer, R. & Heggestad, E. D. (1999). Individual differences in motivation: Traits and self-regulatory skills. In P. L. Ackerman, P. C. Kyllonen & R. D. Roberts (Hrsg.), *Learning and individual differences: Process, trait, and content determinants* (S. 293-313). Washington, DC: American Psychological Association.

Kaufman, A. S., Kaufman, J. C., Liu, X. & Johnson, C. K. (2009). How do educational attainment and gender relate to fluid intelligence, crystallized intelligence, and academic skills at ages 22-90 years? *Archives of Clinical Neuropsychology, 24*, 153-163. doi:10.1093/arclin/acp015

Klix, F. (1988). Gedächtnis und Wissen. In H. Mandl & H. Spada (Hrsg.), *Wissenspsychologie* (S. 19-54). München: Psychologie Verlags Union.

Kluwe, R. H. (1988). Methoden der Psychologie zur Gewinnung von Daten über menschliches Wissen. In H. Mandl & H. Spada (Hrsg.), *Wissenspsychologie* (S. 359-385). München: Psychologie Verlags Union.

Koivula, N., Hassmén, P. & Hunt, D. P. (2001). Performance on the Swedish Scholastic Aptitude Test: Effects of self-assessment and gender. *Sex Roles, 44*, 629-645. doi:10.1023/A:1012203412708

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Lehmann, R., Denissen, J. J. A., Allemand, M. & Penke, L. (2013). Age and gender differences in motivational manifestations of the Big Five from Age 16 to 60. *Developmental Psychology, 49*, 365-383. doi:10.1037/a0028277

Lei, P.-W. & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Hrsg.), *Handbook of structural equation modeling* (S. 164-179). New York: Guilford Press.

Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R* (2., erweiterte und überarbeitete Aufl.). Göttingen: Hogrefe.

Linn, M. C. & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*, 1479-1498. doi:10.2307/1130467

Lippa, R. (1998). Gender-related individual differences and the structure of vocational interests: The importance of the people-things dimension. *Journal of Personality and Social Psychology, 74*, 996-1009. doi:10.1037/0022-3514.74.4.996

Lippa, R. A. (2005). *Gender, nature, and nurture* (2nd ed.). Mahwah, NJ: Erlbaum.

Liu, O. L., Lee, H.-S. & Linn, M. C. (2011). Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching, 48*, 1079-1107. doi:10.1002/tea.20441

Lockhart, R. S. & Craik, F. I. M. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology, 44*, 87-112. doi:10.1037/h0084237

Loftus, E. F. & Loftus, G. R. (1980). On the permanence of stored information in the human brain. *American Psychologist, 35*, 409-420. doi:10.1037/0003-066X.35.5.409

Low, K. S. D., Yoon, M., Roberts, B. W. & Rounds, J. (2005). The stability of vocational interests from early adolescence to middle adulthood: A quantitative review of

longitudinal studies. *Psychological Bulletin, 131*, 713-737. doi:10.1037/0033-2909.131.5.713

Lubke, G. H., Dolan, C. V., Kelderman, H. & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*, 543-566. doi:10.1016/S0160-2896(03)00051-5

Lynn, R. & Irwing, P. (2002). Sex differences in general knowledge, semantic memory and reasoning ability. *British Journal of Psychology, 93*, 545-556. doi:10.1348/000712602761381394

Lynn, R. & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence, 32*, 481-498. doi:10.1016/j.intell.2004.06.008

Lynn, R., Irwing, P. & Cammock, T. (2002). Sex differences in general knowledge. *Intelligence, 30*, 27-39. doi:10.1016/S0160-2896(01)00064-2

Lynn, R., Wilberg, S. & Margraf-Stiksrud, J. (2004). Sex differences in general knowledge in German high school students. *Personality and Individual Differences, 37*, 1643-1650. doi:10.1016/j.paid.2004.02.018

Lynn, R., Wilberg-Neidhardt, S. & Margraf-Stiksrud, J. (2005). Sex differences in general knowledge in German and Northern Irish university students. *Sexualities, Evolution & Gender, 7*, 277-285. doi:10.1080/14616660500477755

Mabe, P. A. & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*, 280-296. doi:10.1037/0021-9010.67.3.280

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

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling, 1*, 5-34. doi:10.1080/10705519409539960

Mazzeo, J., Schmitt, A. P. & Bleistein, C. A. (1993). *Sex-related performance differences on constructed-response and multiple-choice sections of Advanced Placement Examinations* (College Board Report No. 92-7; ETS RR No. 93-5). Verfügbar unter Wiley Online Library Website: http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.1993.tb01516.x/pdf

McNamara, T. P., Hardy, J. K. & Hirtle, S. C. (1989). Subjective hierarchies in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 211-227. doi:10.1037/0278-7393.15.2.211

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543. doi:10.1007/BF02294825

Misra, I. (2003). Openness to experience: Gender differences and its correlates. *Journal of Personality and Clinical Studies, 19*(1), 141-151.

Mondak, J. J. & Canache, D. (2004). Knowledge variables in cross-national social inquiry. *Social Science Quarterly, 85*, 539-558. doi:10.1111/j.0038-4941.2004.00232.x

Moore, D. A. & Kim, T. G. (2003). Myopic social prediction and the solo comparison effect. *Journal of Personality and Social Psychology, 85*, 1121-1135. doi:10.1037/0022-3514.85.6.1121

Morris, C. D., Bransford, J. D. & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519-533. doi:10.1016/S0022-5371(77)80016-9

Mplus (Version 7.1) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. (1998-2004). Mplus Technical Appendices. Los Angeles, CA: Muthén & Muthén.

National Institute of Education. (1974). Guidelines for assessment of sex bias and sex fairness in career interest inventories. In E. E. Diamond (Hrsg.), *Issues of sex bias and sex fairness in career interest measurement* (S. xxiii-xxix). Washington, DC: U.S. Government Printing Office.

Neidhardt-Wilberg, S. (2005). Weibliches und männliches Allgemeinwissen? In S. R. Schilling, J. R. Sparfeldt & C. Pruisken (Hrsg.), *Aktuelle Aspekte pädagogisch-psychologischer Forschung: Detlev. H. Rost zum 60. Geburtstag* (S. 145-158). Münster: Waxmann.

Nye, C. D., Su, R., Rounds, J. & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives on Psychological Science, 7*, 384-403. doi:10.1177/1745691612449021

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers, 32*, 396-402. doi:10.3758/BF03200807

Ostendorf, F. & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae - Revidierte Fassung*. Göttingen: Hogrefe.

Oswald, W. D., Rupprecht, R. & Hagen, B. (1997). Aspekte der kognitiven Leistungsfähigkeit bei 62-64jährigen aus Ost- und Westdeutschland. *Zeitschrift für Gerontopsychologie und -psychiatrie, 10*(4), 213-229.

Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations? *Journal of Vocational Behavior, 21*, 259-287. doi:10.1016/0001-8791(82)90036-7

Prediger, D. J. & Cole, N. S. (1975). Sex-role socialization and employment realities: Implications for vocational interest measures. *Journal of Vocational Behavior, 7*, 239-251. doi:10.1016/0001-8791(75)90064-0

Rammstedt, B. & Rammsayer, T. H. (2000). Sex differences in self-estimates of different aspects of intelligence. *Personality and Individual Differences, 29*, 869-880. doi:10.1016/S0191-8869(99)00238-X

Rolfhus, E. L. & Ackerman, P. L. (1996). Self-report knowledge: At the crossroads of ability, interest, and personality. *Journal of Educational Psychology, 88*, 174-188. doi:10.1037/0022-0663.88.1.174

Rolfhus, E. L. & Ackerman, P. L. (1999). Assessing individual differences in knowledge: Knowledge, intelligence, and related traits. *Journal of Educational Psychology, 91*, 511-526. doi:10.1037/0022-0663.91.3.511

Rounds, J. & Su, R. (2014). The nature and power of interests. *Current Directions in Psychological Science, 23*, 98-103. doi:10.1177/0963721414522812

Ryle, G. (1949). *The concept of mind*. New York: Barnes & Noble.

Schmidt, F. L. (2011). A theory of sex differences in technical aptitude and some supporting evidence. *Perspectives on Psychological Science, 6*, 560-573. doi:10.1177/1745691611419670

Schmidt, F. L. (2014). A general theoretical integrative model of individual differences in interests, abilities, personality traits, and academic and occupational achievement: A

commentary on four recent articles. *Perspectives on Psychological Science, 9*, 211-218. doi:10.1177/1745691613518074

Schmitt, M., Wahl, H.-W. & Kruse, A. (2008). *Interdisziplinäre Längsschnittstudie des Erwachsenenalters (ILSE): Abschlussbericht anlässlich der Fertigstellung des dritten Messzeitpunkts*. Verfügbar unter https://www.bmfsfj.de/RedaktionBMFSFJ/Abteilung3/Pdf-Anlagen/abschlussbericht-laengsschnittstudie-ilse.pdf

Schraw, G. & Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. *Educational Psychology Review, 13*, 23-52. doi:10.1023/A:1009004801455

Schulze, R. (2005). Modeling structures of intelligence. In O. Wilhelm & R. W. Engle (Hrsg.), *Handbook of understanding and measuring intelligence* (S. 241-263). Thousand Oaks, CA: Sage.

Sinharay, S., Stern, H. S. & Russell, D. (2001). The use of multiple imputation for analysis of missing data. *Psychological Methods, 6*, 317-329. doi:10.1037/1082-989X.6.4.317

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. *Psychometrika, 43*, 381-396. doi:10.1007/BF02293647

Spada, H. & Mandl, H. (1988). Wissenspsychologie: Einführung. In H. Mandl & H. Spada (Hrsg.), *Wissenspsychologie* (S. 1-16). München: Psychologie Verlags Union.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology, 15*, 201-293. doi:10.2307/1412107

Spencer, S. J., Steele, C. M. & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 4-28. doi:10.1006/jesp.1998.1373

Statistisches Bundesamt. (2009). *Bildung und Kultur. Studierende an Hochschulen* (Fachserie 11, Reihe 4.1). Wiesbaden: Autor.

Staudinger, U. M., Smith, J. & Baltes, P. B. (1992). Wisdom-related knowledge in a life review task: Age differences and the role of professional specialization. *Psychology and Aging, 7*, 271-281. doi:10.1037/0882-7974.7.2.271

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*, 613-629. doi:10.1037/0003-066X.52.6.613

Steenkamp J.-B. E. M. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78-90. doi:10.1086/209528

Steinmayr, R. & Amelang, M. (2006). Erste Untersuchungen zur Kriteriums-Validität des I-S-T 2000R an Erwachsenen beiderlei Geschlechts. *Diagnostica, 52,* 181-188. doi:10.1026/0012-1924.52.4.181

Steinmayr, R., Beauducel, A. & Spinath, B. (2010). Do sex differences in a faceted model of fluid and crystallized intelligence depend on the method applied? *Intelligence, 38,* 101-110. doi:10.1016/j.intell.2009.08.001

Sternberg, R. J. (1990). Wisdom and its relations to intelligence and creativity. In R. J. Sternberg (Hrsg.), *Wisdom: Its nature, origins, and development* (S. 142-159). Cambridge University Press.

Sternberg, R. J. (1998). A balance theory of wisdom. *Review of General Psychology, 2*, 347-365. doi:10.1037/1089-2680.2.4.347

Sternberg, R. J. (2000). Intelligence and wisdom. In R. J. Sternberg (Hrsg.), *Handbook of intelligence* (S. 631-649). Cambridge University Press.

Stumpf, H. & Stanley, J. C. (1998). Stability and change in gender-related differences on the College Board Advanced Placement and Achievement Tests. *Current Directions in Psychological Science, 7*, 192-196. doi:10.1111/1467-8721.ep10836927

Su, R., Rounds, J. & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin, 135*, 859-884. doi:10.1037/a0017364

Süß, H.-M. (1996). *Intelligenz, Wissen und Problemlösen*. Göttigen: Hogrefe.

Thurstone, L. L. (1938). *Primary mental abilities*. University of Chicago Press.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Hrsg.), *Organization of memory* (S. 381-403). New York: Academic Press.

Valentine, J. C., DuBois, D. L. & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39*, 111-133. doi:10.1207/s15326985ep3902_3

von Stumm, S. & Ackerman, P. L. (2013). Investment and intellect: A review and meta-analysis. *Psychological Bulletin, 139*, 841-869. doi:10.1037/a0030746

von Stumm, S., Chamorro-Premuzic, T. & Furnham, A. (2009). Decomposing self-estimates of intelligence: Structure and sex differences across 12 nations. *British Journal of Psychology, 100*, 429-442. doi:10.1348/000712608X357876

Wetzel, E. & Hell, B. (2013). Gender-related differential item functioning in vocational interest measurement: An analysis of the AIST-R. *Journal of Individual Differences, 34*, 170-183. doi:10.1027/1614-0001/a000112

Wilhelm, O., Schroeders, U. & Schipolowski, S. (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe*. Göttingen: Hogrefe.

Wilhelm, O., Schulze, R., Schmiedeck, F. & Süß, H.-M. (2003). Interindividuelle Unterschiede im typischen intellektuellen Engagement. *Diagnostica, 49*, 49-60. doi:10.1026//0012-1924.49.2.49

Willingham, W. W. & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

Wissen. (1999). In *Duden - Das große Wörterbuch der deutschen Sprache* (Band 10, S. 4537 f.). Mannheim: Duden-Verlag.

Yuen, M. & Furnham, A. (2005). Sex differences in self-estimation of multiple intelligences among Hong Kong Chinese adolescents. *High Ability Studies, 16*, 187-199. doi:10.1080/13598130600618009

Zeleny, M. (1987). Management support systems: Towards integrated knowledge management. *Human Systems Management, 7*, 59-70. doi:10.3233/HSM-1987-7108

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology, 58*, 479-493. doi:10.1002/asi.20508

Anhang

Anhang A: Liste der 19 Themen des General Knowledge Tests, geordnet nach den sechs Faktoren

Current Affairs

- Politics
- History
- Finance
- Geography
- Discovery and Exploration

Fashion

- Fashion
- Popular Music
- Film

Arts

- Film
- Classical Music
- Literature
- Art
- Jazz and Blues

Family

- Cookery
- Medicine

Physical Health and Recreation

- Games
- Sport
- Biology

Science

- General Science
- History of Science

Anhang B: Ergebnisse der Studie zur Selbsteinschätzung (Erste Parcelgruppe)

Tabelle B1

Standardisierte Parameterschätzungen und Model-Fit der 2-Gruppen-Modelle, mit Klassifizierung nach Geschlecht (Parcelgruppe 1)

	Modell 1		Modell 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Selbsteinschätzung	-0.99	0[a]	-1.09	0[a]
[95%-KI]	[-1.27; -0.70]		[-1.38; -0.79]	
Achsenabschnitte				
Wissen	-0.91	0[a]	0[a]	0[a]
[95%-KI]	[-1.27; -0.54]			
SE-Parcel 1	5.07	4.80	5.15	4.85
SE-Parcel 2	4.67	5.25	4.73	5.30
SE-Parcel 3	4.87	5.31	4.95	5.38
W-Parcel 1	3.51	3.08	2.96	2.42
W-Parcel 2	5.30	4.65	4.73	3.96
W-Parcel 3	4.53	4.09	3.98	3.34
W-Parcel 4	4.02	3.38	3.43	2.75
β-Koeffizienten (SE)				
SE-Parcel 1	.70[a]	.67[a]	.69[a]	.66[a]
SE-Parcel 2	.64	.72	.62	.71
SE-Parcel 3	.74	.80	.72	.80
β-Koeffizienten (W)				
W-Parcel 1	.76[a]	.82[a]	.78[a]	.85[a]
W-Parcel 2	.71	.76	.74	.82
W-Parcel 3	.73	.81	.74	.83
W-Parcel 4	.76	.79	.78	.83
Selbsteinschätzung	.36	.30	.48	.36

(wird fortgesetzt)

	Modell 1		Modell 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
SE-Parcel 1	.51	.56	.52	.56
SE-Parcel 2	.59	.49	.62	.50
SE-Parcel 3	.46	.36	.48	.36
W-Parcel 1	.42	.33	.40	.29
W-Parcel 2	.50	.42	.46	.33
W-Parcel 3	.47	.35	.45	.31
W-Parcel 4	.42	.38	.39	.31
Wissen	.87	.91	.77	.87
Model-Fit				
χ^2_{df}	74.27_{37}		99.64_{38}	
p	< .001		< .001	
SCF	1.00		1.00	
CFI	.95		.92	
SRMR	.06		.12	
RMSEA	.08		.10	
[90%-KI]	[.05; .11]		[.08; .13]	
χ^2-Diff-Test: χ^2_{df}; *p*	21.02_1; < .001			

Anmerkungen. Sämtliche (nicht standardisierten) Achsenabschnitte der manifesten Variablen und Regressionskoeffizienten waren jeweils für beide Gruppen gleichgesetzt. KI = Konfidenzintervall; SE = Selbsteinschätzung; W = Wissen; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Modell 1 vs. Modell 2.

[a]Parameter wurde vorab fixiert.

Tabelle B2

Standardisierte Parameterschätzungen und Model-Fit der 2-Gruppen-Modelle, mit Klassifizierung nach experimenteller Bedingung (Parcelgruppe 1)

	Modell 1		Modell 2	
	SE hoch	SE niedrig	SE hoch	SE niedrig
Erwartungswert				
Selbsteinschätzung	0.39	0[a]	0.34	0[a]
[95%-KI]	[0.12; 0.66]		[0.08; 0.61]	
Achsenabschnitte				
Wissen	-0.35	0[a]	0[a]	0[a]
[95%-KI]	[-0.58; -0.11]			
SE-Parcel 1	4.13	4.27	4.14	4.28
SE-Parcel 2	4.32	3.83	4.33	3.84
SE-Parcel 3	4.04	4.23	4.04	4.24
W-Parcel 1	2.65	2.49	2.51	2.35
W-Parcel 2	4.18	4.30	4.06	4.17
W-Parcel 3	3.65	3.44	3.52	3.31
W-Parcel 4	3.02	2.89	2.89	2.76
β-Koeffizienten (SE)				
SE-Parcel 1	.71[a]	.73[a]	.71[a]	.73[a]
SE-Parcel 2	.75	.67	.75	.67
SE-Parcel 3	.74	.78	.75	.78
β-Koeffizienten (W)				
W-Parcel 1	.81[a]	.81[a]	.81[a]	.82[a]
W-Parcel 2	.75	.82	.74	.82
W-Parcel 3	.79	.79	.78	.79
W-Parcel 4	.80	.81	.79	.81
Selbsteinschätzung	.51	.48	.47	.44
Fehlervarianzen				
SE-Parcel 1	.50	.47	.50	.46
SE-Parcel 2	.44	.56	.44	.56
SE-Parcel 3	.45	.40	.44	.39

(wird fortgesetzt)

	Modell 1		Modell 2	
	SE hoch	SE niedrig	SE hoch	SE niedrig
W-Parcel 1	.35	.34	.34	.34
W-Parcel 2	.44	.33	.45	.33
W-Parcel 3	.38	.37	.39	.38
W-Parcel 4	.37	.34	.37	.34
Wissen	.74	.78	.78	.81
Model-Fit				
χ^2_{df}	51.49_{37}		60.46_{38}	
p	.057		.012	
SCF	0.98		0.98	
CFI	.99		.98	
SRMR	.04		.06	
RMSEA	.05		.06	
[90%-KI]	[.00; .08]		[.03; .09]	
χ^2-Diff-Test: χ^2_{df}; *p*	10.70_1; .001			

Anmerkungen. Sämtliche (nicht standardisierten) Achsenabschnitte der manifesten Variablen und Regressionskoeffizienten waren jeweils für beide Gruppen gleichgesetzt. SE = Selbsteinschätzung; KI = Konfidenzintervall; W = Wissen; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Modell 1 vs. Modell 2.

[a]Parameter wurde vorab fixiert.

Tabelle B3

Standardisierte Parameterschätzungen und Model-Fit des 4-Gruppen-Modells, mit Klassifizierung nach Geschlecht und experimenteller Bedingung (Parcelgruppe 1)

	Frauen, SE hoch	Frauen, SE niedrig	Männer, SE hoch	Männer, SE niedrig
Erwartungswert				
Selbsteinschätzung	-0.47	-0.82	0.87	0[a]
[95%-KI]	[-0.87; -0.07]	[-1.24; -0.40]	[0.23; 1.50]	
Achsenabschnitte				
Wissen	-1.39	-1.03	-0.42	0[a]
[95%-KI]	[-1.88; -0.91]	[-1.47; -0.59]	[-0.83; -0.02]	
SE-Parcel 1	4.90	4.76	4.54	4.86
SE-Parcel 2	4.79	4.19	6.06	4.57
SE-Parcel 3	4.49	4.81	5.59	4.78
W-Parcel 1	3.80	3.48	3.19	3.30
W-Parcel 2	5.48	5.37	4.30	5.54
W-Parcel 3	4.96	4.40	4.09	4.43
W-Parcel 4	4.31	4.01	3.33	3.77
β-Koeffizienten (SE)				
SE-Parcel 1	.73[a]	.69[a]	.53[a]	.80[a]
SE-Parcel 2	.70	.59	.69	.73
SE-Parcel 3	.70	.74	.69	.83
β-Koeffizienten (W)				
W-Parcel 1	.74[a]	.77[a]	.85[a]	.77[a]
W-Parcel 2	.67	.74	.71	.81
W-Parcel 3	.73	.73	.82	.78
W-Parcel 4	.74	.78	.78	.78
Selbsteinschätzung	.39	.34	.23	.36
Fehlervarianzen				
SE-Parcel 1	.47	.52	.72	.37
SE-Parcel 2	.52	.65	.53	.47
SE-Parcel 3	.51	.45	.53	.32

(wird fortgesetzt)

	Frauen, SE hoch	Frauen, SE niedrig	Männer, SE hoch	Männer, SE niedrig
W-Parcel 1	.45	.41	.29	.40
W-Parcel 2	.56	.45	.50	.34
W-Parcel 3	.47	.47	.34	.39
W-Parcel 4	.45	.40	.40	.39
Selbsteinschätzung	.85	.89	.95	.87
Model-Fit				
χ^2_{df}		129.38_{85}		
p		.001		
CFI		.94		
SRMR		.10		
RMSEA		.08		
[90%-KI]		[.05; .11]		

Anmerkungen. Sämtliche (nicht standardisierten) Achsenabschnitte der manifesten Variablen und Regressionskoeffizienten waren für alle Gruppen gleichgesetzt. SE = Selbsteinschätzung; KI = Konfidenzintervall; W = Wissen.

[a]Parameter wurde vorab fixiert.

Anhang C: Ergebnisse der Studie zur Selbsteinschätzung (Zweite Parcelgruppe)

Tabelle C1

Standardisierte Parameterschätzungen und Model-Fit der 2-Gruppen-Modelle, mit Klassifizierung nach Geschlecht (Parcelgruppe 2)

	Modell 1		Modell 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Selbsteinschätzung	-1.08	0[a]	-1.23	0[a]
[95%-KI]	[-1.37; -0.78]		[-1.53; -0.92]	
Achsenabschnitte				
Wissen	-0.80	0[a]	0[a]	0[a]
[95%-KI]	[-1.16; -0.43]			
SE-Parcel 1	5.14	4.95	5.27	5.12
SE-Parcel 2	4.90	4.72	5.01	4.74
SE-Parcel 3	4.54	4.52	4.60	4.55
W-Parcel 1	3.79	2.91	3.37	2.52
W-Parcel 2	4.58	4.18	4.09	3.67
W-Parcel 3	4.67	4.39	4.18	3.83
W-Parcel 4	3.71	3.32	3.26	2.86
β-Koeffizienten (SE)				
SE-Parcel 1	.72[a]	.65[a]	.71[a]	.67[a]
SE-Parcel 2	.77	.70	.74	.67
SE-Parcel 3	.52	.49	.50	.47
β-Koeffizienten (W)				
W-Parcel 1	.71[a]	.66[a]	.72[a]	.66[a]
W-Parcel 2	.75	.81	.78	.85
W-Parcel 3	.75	.84	.78	.86
W-Parcel 4	.74	.79	.76	.81
Selbsteinschätzung	.36	.29	.48	.38

(wird fortgesetzt)

	Modell 1		Modell 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
SE-Parcel 1	.49	.58	.49	.55
SE-Parcel 2	.41	.51	.46	.55
SE-Parcel 3	.73	.76	.76	.78
W-Parcel 1	.49	.57	.48	.57
W-Parcel 2	.44	.34	.40	.28
W-Parcel 3	.44	.29	.40	.26
W-Parcel 4	.46	.38	.43	.35
Wissen	.87	.92	.77	.86
Model-Fit				
χ^2_{df}	85.72_{37}		104.14_{38}	
p	< .001		< .001	
SCF	1.02		1.03	
CFI	.93		.90	
SRMR	.08		.12	
RMSEA	.09		.11	
[90%-KI]	[.07; .12]		[.08; .13]	
χ^2-Diff-Test: χ^2_{df}; *p*	16.35_1; < .001			

Anmerkungen. Sämtliche (nicht standardisierten) Achsenabschnitte der manifesten Variablen und Regressionskoeffizienten waren jeweils für beide Gruppen gleichgesetzt. KI = Konfidenzintervall; SE = Selbsteinschätzung; W = Wissen; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Modell 1 vs. Modell 2.
[a]Parameter wurde vorab fixiert.

Tabelle C2

Standardisierte Parameterschätzungen und Model-Fit der 2-Gruppen-Modelle, mit Klassifizierung nach experimenteller Bedingung (Parcelgruppe 2)

	Modell 1		Modell 2	
	SE hoch	SE niedrig	SE hoch	SE niedrig
Erwartungswert				
Selbsteinschätzung	0.36	0[a]	0.32	0[a]
[95%-KI]	[0.10; 0.63]		[0.06; 0.58]	
Achsenabschnitte				
Wissen	-0.32	0[a]	0[a]	0[a]
[95%-KI]	[-0.55; -0.09]			
SE-Parcel 1	4.13	4.18	4.19	4.19
SE-Parcel 2	3.83	3.97	3.83	3.99
SE-Parcel 3	4.24	3.66	4.26	3.67
W-Parcel 1	2.88	2.62	2.76	2.50
W-Parcel 2	3.64	3.61	3.52	3.48
W-Parcel 3	3.83	3.63	3.71	3.50
W-Parcel 4	2.92	2.76	2.80	2.64
β-Koeffizienten (SE)				
SE-Parcel 1	.71[a]	.70[a]	.71[a]	.70[a]
SE-Parcel 2	.78	.80	.79	.81
SE-Parcel 3	.62	.53	.62	.53
β-Koeffizienten (W)				
W-Parcel 1	.76[a]	.73[a]	.75[a]	.72[a]
W-Parcel 2	.80	.84	.81	.84
W-Parcel 3	.82	.82	.82	.82
W-Parcel 4	.78	.78	.79	.78
Selbsteinschätzung	.51	.47	.47	.44
Fehlervarianzen				
SE-Parcel 1	.50	.52	.50	.52
SE-Parcel 2	.39	.37	.38	.35
SE-Parcel 3	.61	.72	.61	.72

(wird fortgesetzt)

	Modell 1		Modell 2	
	SE hoch	SE niedrig	SE hoch	SE niedrig
W-Parcel 1	.43	.47	.43	.48
W-Parcel 2	.36	.30	.35	.29
W-Parcel 3	.33	.33	.33	.33
W-Parcel 4	.39	.39	.38	.39
Wissen	.74	.78	.78	.81
Model-Fit				
χ^2_{df}	42.32_{37}		49.54_{38}	
p	.252		.100	
SCF	1.04		1.04	
CFI	.99		.99	
SRMR	.04		.06	
RMSEA	.03		.04	
[90%-KI]	[.00; .07]		[.00; .08]	
χ^2-Diff-Test: χ^2_{df}; *p*	7.97_1; .005			

Anmerkungen. Sämtliche (nicht standardisierten) Achsenabschnitte der manifesten Variablen und Regressionskoeffizienten waren jeweils für beide Gruppen gleichgesetzt. SE = Selbsteinschätzung; KI = Konfidenzintervall; W = Wissen; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Modell 1 vs. Modell 2.

[a]Parameter wurde vorab fixiert.

Tabelle C3

Standardisierte Parameterschätzungen und Model-Fit des 4-Gruppen-Modells, mit Klassifizierung nach Geschlecht und experimenteller Bedingung (Parcelgruppe 2)

	Frauen, SE hoch	Frauen, SE niedrig	Männer, SE hoch	Männer, SE niedrig
Erwartungswert				
Selbsteinschätzung	-0.56	-0.85	1.00	0[a]
[95%-KI]	[-0.95; -0.17]	[-1.24; -0.46]	[0.27; 1.74]	
Achsenabschnitte				
Wissen	-1.24	-0.88	-0.38	0[a]
[95%-KI]	[-1.71; -0.77]	[-1.29; -0.46]	[-0.77; 0.01]	
SE-Parcel 1	4.96	4.79	5.09	4.62
SE-Parcel 2	4.78	4.45	4.31	5.05
SE-Parcel 3	4.49	4.25	6.26	3.56
W-Parcel 1	4.16	3.64	3.05	2.90
W-Parcel 2	4.75	4.58	4.05	4.55
W-Parcel 3	5.01	4.54	4.31	4.69
W-Parcel 4	4.01	3.59	3.30	3.59
β-Koeffizienten (SE)				
SE-Parcel 1	.73[a]	.69[a]	.52[a]	.69[a]
SE-Parcel 2	.80	.74	.51	.87
SE-Parcel 3	.53	.50	.52	.43
β-Koeffizienten (W)				
W-Parcel 1	.71[a]	.72[a]	.73[a]	.58[a]
W-Parcel 2	.70	.79	.84	.78
W-Parcel 3	.72	.76	.87	.79
W-Parcel 4	.72	.75	.82	.75
Selbsteinschätzung	.39	.34	.20	.35
Fehlervarianzen				
SE-Parcel 1	.48	.52	.73	.52
SE-Parcel 2	.36	.45	.74	.25
SE-Parcel 3	.72	.75	.73	.81

(wird fortgesetzt)

	Frauen, SE hoch	Frauen, SE niedrig	Männer, SE hoch	Männer, SE niedrig
W-Parcel 1	.50	.48	.47	.67
W-Parcel 2	.51	.38	.30	.39
W-Parcel 3	.48	.42	.25	.38
W-Parcel 4	.49	.44	.33	.44
Selbsteinschätzung	.85	.89	.96	.88
Model-Fit				
χ^2_{df}		146.52_{85}		
p		< .001		
CFI		.91		
SRMR		.12		
RMSEA		.10		
[90%-KI]		[.07; .12]		

Anmerkungen. Sämtliche (nicht standardisierten) Achsenabschnitte der manifesten Variablen und Regressionskoeffizienten waren für alle Gruppen gleichgesetzt. SE = Selbsteinschätzung; KI = Konfidenzintervall; W = Wissen.

[a]Parameter wurde vorab fixiert.

Anhang D: Ergebnisse der Studie zur Selbsteinschätzung (Dritte Parcelgruppe)

Tabelle D1

Standardisierte Parameterschätzungen und Model-Fit der 2-Gruppen-Modelle, mit Klassifizierung nach Geschlecht (Parcelgruppe 3)

	Modell 1		Modell 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Selbsteinschätzung	-0.86	0[a]	-1.08	0[a]
[95%-KI]	[-1.16; -0.57]		[-1.39; -0.77]	
Achsenabschnitte				
Wissen	-0.95	0[a]	0[a]	0[a]
[95%-KI]	[-1.32; -0.58]			
SE-Parcel 1	4.06	3.26	4.19	3.45
SE-Parcel 2	4.56	4.57	4.70	4.66
SE-Parcel 3	5.23	5.97	5.35	6.05
W-Parcel 1	4.68	3.91	4.10	3.30
W-Parcel 2	3.82	3.54	3.30	2.90
W-Parcel 3	4.27	3.80	3.76	3.20
W-Parcel 4	3.98	3.49	3.44	2.87
β-Koeffizienten (SE)				
SE-Parcel 1	.62[a]	.48[a]	.62[a]	.50[a]
SE-Parcel 2	.71	.68	.67	.65
SE-Parcel 3	.63	.69	.58	.64
β-Koeffizienten (W)				
W-Parcel 1	.73[a]	.76[a]	.75[a]	.80[a]
W-Parcel 2	.71	.81	.72	.84
W-Parcel 3	.69	.76	.72	.80
W-Parcel 4	.73	.79	.75	.82
Selbsteinschätzung	.39	.31	.51	.38

(wird fortgesetzt)

	Modell 1		Modell 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
SE-Parcel 1	.62	.77	.61	.75
SE-Parcel 2	.50	.53	.55	.58
SE-Parcel 3	.61	.53	.66	.59
W-Parcel 1	.46	.43	.43	.36
W-Parcel 2	.50	.35	.48	.30
W-Parcel 3	.52	.43	.48	.35
W-Parcel 4	.47	.38	.44	.33
Wissen	.85	.91	.74	.86
Model-Fit				
χ^2_{df}	48.44_{37}		73.61_{38}	
p	.099		< .001	
SCF	1.05		1.06	
CFI	.98		.94	
SRMR	.06		.11	
RMSEA	.04		.08	
[90%-KI]	[.00; .08]		[.05; .10]	
χ^2-Diff-Test: χ^2_{df}; *p*	19.87_1; < .001			

Anmerkungen. Sämtliche (nicht standardisierten) Achsenabschnitte der manifesten Variablen und Regressionskoeffizienten waren jeweils für beide Gruppen gleichgesetzt. KI = Konfidenzintervall; SE = Selbsteinschätzung; W = Wissen; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Modell 1 vs. Modell 2.
[a]Parameter wurde vorab fixiert.

Tabelle D2

Standardisierte Parameterschätzungen und Model-Fit der 2-Gruppen-Modelle, mit Klassifizierung nach experimenteller Bedingung (Parcelgruppe 3)

	Modell 1		Modell 2	
	SE hoch	SE niedrig	SE hoch	SE niedrig
Erwartungswert				
Selbsteinschätzung	0.41	0[a]	0.35	0[a]
[95%-KI]	[0.13; 0.70]		[0.07; 0.62]	
Achsenabschnitte				
Wissen	-0.37	0[a]	0[a]	0[a]
[95%-KI]	[-0.62; -0.13]			
SE-Parcel 1	3.16	3.40	3.17	3.42
SE-Parcel 2	4.08	3.64	4.10	3.65
SE-Parcel 3	4.94	4.60	4.95	4.62
W-Parcel 1	3.66	3.46	3.53	3.32
W-Parcel 2	2.96	2.91	2.83	2.77
W-Parcel 3	3.43	3.22	3.31	3.08
W-Parcel 4	3.14	2.88	3.00	2.74
β-Koeffizienten (SE)				
SE-Parcel 1	.56[a]	.60[a]	.56[a]	.60[a]
SE-Parcel 2	.78	.69	.78	.69
SE-Parcel 3	.69	.64	.69	.64
β-Koeffizienten (W)				
W-Parcel 1	.79[a]	.78[a]	.78[a]	.78[a]
W-Parcel 2	.75	.78	.75	.78
W-Parcel 3	.77	.76	.77	.76
W-Parcel 4	.80	.77	.80	.77
Selbsteinschätzung	.53	.50	.48	.45
Fehlervarianzen				
SE-Parcel 1	.68	.64	.68	.64
SE-Parcel 2	.40	.53	.39	.53
SE-Parcel 3	.53	.60	.52	.59

(wird fortgesetzt)

	Modell 1		Modell 2	
	SE hoch	SE niedrig	SE hoch	SE niedrig
W-Parcel 1	.38	.39	.39	.39
W-Parcel 2	.43	.39	.43	.39
W-Parcel 3	.40	.42	.40	.42
W-Parcel 4	.36	.40	.36	.40
Wissen	.72	.75	.77	.80
Model-Fit				
χ^2_{df}	29.46_{37}		38.59_{38}	
p	.807		.443	
SCF	1.04		1.04	
CFI	1		1	
SRMR	.04		.06	
RMSEA	.00		.01	
[90%-KI]	[.00; .04]		[.00; .06]	
χ^2-Diff-Test: χ^2_{df}; *p*		9.57_1; .002		

Anmerkungen. Sämtliche (nicht standardisierten) Achsenabschnitte der manifesten Variablen und Regressionskoeffizienten waren jeweils für beide Gruppen gleichgesetzt. SE = Selbsteinschätzung; KI = Konfidenzintervall; W = Wissen; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Modell 1 vs. Modell 2.

[a]Parameter wurde vorab fixiert.

Tabelle D3

Standardisierte Parameterschätzungen und Model-Fit des 4-Gruppen-Modells, mit Klassifizierung nach Geschlecht und experimenteller Bedingung (Parcelgruppe 3)

	Frauen, SE hoch	Frauen, SE niedrig	Männer, SE hoch	Männer, SE niedrig
Erwartungswert				
Selbsteinschätzung	-0.38	-0.75	1.00	0[a]
[95%-KI]	[-0.78; 0.03]	[-1.17; -0.33]	[0.20; 1.80]	
Achsenabschnitte				
Wissen	-1.39	-1.03	-0.44	0[a]
[95%-KI]	[-1.85; -0.94]	[-1.45; -0.61]	[-0.85; -0.03]	
SE-Parcel 1	3.77	4.01	3.19	3.24
SE-Parcel 2	4.51	4.21	5.70	3.78
SE-Parcel 3	5.10	5.01	8.16	5.02
W-Parcel 1	4.99	4.56	3.91	4.24
W-Parcel 2	4.01	3.82	3.47	4.03
W-Parcel 3	4.69	4.08	3.67	4.48
W-Parcel 4	4.26	3.92	3.65	3.67
β-Koeffizienten (SE)				
SE-Parcel 1	.61[a]	.64[a]	.32[a]	.53[a]
SE-Parcel 2	.74	.67	.58	.63
SE-Parcel 3	.64	.61	.64	.64
β-Koeffizienten (W)				
W-Parcel 1	.72[a]	.74[a]	.78[a]	.68[a]
W-Parcel 2	.69	.73	.82	.76
W-Parcel 3	.71	.68	.76	.74
W-Parcel 4	.72	.74	.85	.68
Selbsteinschätzung	.43	.38	.20	.40
Fehlervarianzen				
SE-Parcel 1	.63	.60	.90	.72
SE-Parcel 2	.46	.55	.66	.61
SE-Parcel 3	.60	.63	.59	.60

(wird fortgesetzt)

	Frauen, SE hoch	Frauen, SE niedrig	Männer, SE hoch	Männer, SE niedrig
W-Parcel 1	.48	.46	.39	.54
W-Parcel 2	.53	.47	.33	.43
W-Parcel 3	.50	.53	.42	.45
W-Parcel 4	.48	.46	.28	.54
Selbsteinschätzung	.81	.86	.96	.84
Model-Fit				
χ^2_{df}		128.16_{85}		
p		.002		
CFI		.93		
SRMR		.11		
RMSEA		.08		
[90%-KI]		[.05; .11]		

Anmerkungen. Sämtliche (nicht standardisierten) Achsenabschnitte der manifesten Variablen und Regressionskoeffizienten waren für alle Gruppen gleichgesetzt. SE = Selbsteinschätzung; KI = Konfidenzintervall; W = Wissen.

[a]Parameter wurde vorab fixiert.

Anhang E: Interessenfragebogen

Wir möchten gerne von Ihnen erfahren, für welche Themen und Aktivitäten Sie sich interessieren. Hierbei kann es sich um Interessen jeglicher Art handeln – also z.B. im Zusammenhang mit Ihrem Beruf, mit Familie oder Hobby. Zu diesem Zweck werden Ihnen 36 verschiedene Themenbereiche genannt, zu denen Sie bitte jeweils angeben, wie stark Sie sich dafür interessieren.

Verwenden Sie hierzu bitte die Skala mit den Werten **0 (kein Interesse)** bis **4 (sehr großes Interesse)**. Markieren Sie bei jedem Thema bitte eine der 5 Zahlen. Am Ende des Fragebogens haben Sie noch die Möglichkeit zur Benennung eines oder mehrerer Themen, für die Sie sich interessieren, die jedoch nicht in dem Fragebogen aufgeführt sind.

Unter der Benennung der Themen finden Sie jeweils ein paar exemplarische Inhalte des Bereiches. Es kommt durchaus vor, dass die Interessenbereiche nicht eindeutig voneinander abgrenzbar sind, so z.B. Computer und Informatik. Stören Sie sich daran nicht. **Antworten sie bitte zügig und spontan. Lassen Sie bitte keinen Themenbereich unbeantwortet.**

Viel Spaß!

	Kein Interesse				Sehr großes Interesse
Sport (Aktiv Sport treiben, Sportsendungen im TV sehen, …)	0	1	2	3	4
Ernährung (Kochen, Esskultur, …)	0	1	2	3	4
Medizin (Forschung, Physiologie, Funktionsweise des Menschen, ...)	0	1	2	3	4
Gesundheit (Gesund leben, Gesundheitsvorsorge, …)	0	1	2	3	4
Literatur (Bücher, Autoren, Gedichte und Poesie, …)	0	1	2	3	4
Sprachwissenschaft (Fremdworte, Satzbau, Grammatik, …)	0	1	2	3	4
Fremdsprachen (Wissen über Sprachen, Kommunikation in Fremdsprachen, …)	0	1	2	3	4
Fremde Kulturen (Reisen, Reiseberichte, Eigenarten von Kulturen, …)	0	1	2	3	4
Elektrotechnik (Hardware, Elektrischer Modellbau, …)	0	1	2	3	4

(wird fortgesetzt)

	Kein Interesse				Sehr großes Interesse
Computer (Nutzen des Computers, Surfen im Internet, …)	0	1	2	3	4
Informatik (Programmieren, Netzwerke einrichten, …)	0	1	2	3	4
Maschinen (Maschinen bedienen, Konstruktion von Maschinen, …)	0	1	2	3	4
Biologie (Umwelt und Organismen kennen und erforschen, …)	0	1	2	3	4
Chemie (Chemische Reaktionen, Laborarbeit, …)	0	1	2	3	4
Physik (Naturlehre, Wechselwirkungen Materie und Energie, …)	0	1	2	3	4
Mathematik (Untersuchung geometrischer Figuren, Rechnen, ..)	0	1	2	3	4
Archäologie (Kulturelle Entwicklung der Menschheit, Ausgrabungen, …)	0	1	2	3	4
Modedesign (Entwerfen und Gestalten, Kleidermode, Accessoires, …)	0	1	2	3	4
Raumdesign (Verschönern gegebener Räumlichkeiten, Inneneinrichtung, ..)	0	1	2	3	4
Architektur (Tektonik, Bautechnik, Baukunst, …)	0	1	2	3	4
Bildende Kunst (Visuell gestaltende Künste, Gemälde, Maler, …)	0	1	2	3	4
Darstellende Kunst (Theater, Tanz, Filmkunst, Schauspieler, …)	0	1	2	3	4
Musik (Musikrichtungen, Musik hören, Musizieren, …)	0	1	2	3	4
Philosophie (Deuten der Welt und der menschlichen Existenz, …)	0	1	2	3	4
Religion (Glaube, Glaubensgemeinschaft, Heilige Schriften…)	0	1	2	3	4
Wirtschaft (Aktien, Geschehen am Finanzmarkt, Börse, Unternehmen …)	0	1	2	3	4
Politik (Steuerung von Staat und Gesellschaft, Wahlprogramme, …)	0	1	2	3	4
Gesellschaft (Soziologie, Formen des Zusammenlebens, …)	0	1	2	3	4
Psychologie (Menschliches Verhalten und Erleben, …)	0	1	2	3	4
Pädagogik (Erziehung, Lehren von Kindern und Jugendlichen, …)	0	1	2	3	4
Soziale Arbeit (Familien-, Jugendhilfe, Unterstützung sozialer Einrichtungen, ...)	0	1	2	3	4
Natur (Spaziergänge im Wald, Tiere beobachten, Pflanzen züchten …)	0	1	2	3	4

(wird fortgesetzt)

	Kein Interesse				Sehr großes Interesse
Verkehr (Auto, Logistik, Verkehrsregeln, Bus, Züge, …)	0	1	2	3	4
Geographie (Lage von Ländern, Flüssen, Bergen, Hauptstädten …)	0	1	2	3	4
Geschichte (Menschheitsgeschichte, Wissen über Vergangenes, ..)	0	1	2	3	4
Recht (Gesetze, Rechtsprechung, Rechte und Pflichten, …)	0	1	2	3	4

Interessieren Sie sich für Themen, die nicht in dieser Liste aufgeführt sind?

Nein ❒

Ja ❒, nämlich: _

Tabelle E1

Einträge der 44 Personen, welche die Frage nach weiteren Interessengebieten beantworteten

Person	Eintrag 1	Eintrag 2	Eintrag 3	Eintrag 4
1	Alchemie			
2	Angeln			
3	Anthropologie			
4	Astrologie			
5	Autos	Waffen		
6	Bin gläubiger Christ			
7	Design	Kultur		
8	Einige			
9	Familien-angelegenheiten	Problemlösungen		
10	Fotografie			
11	Fotographie			
12	Freunde			
13	Freunde	Familie		
14	Gender	Sexualität		
15	Geschichten Schreiben			
16	Handarbeiten			
17	Handwerkliches	Kreatives		
18	Haustiere			
19	Japanische Kultur	Anime und Manga		
20	Journalismus			
21	Kampfsport			
22	Kochen	Backen	Kochsendungen, -bücher	
23	Kosmetik			
24	Kreatives Schreiben			
25	Marken			
26	Medien			
27	Mode	Autos	Fahrzeuge	
28	Musikgeschichte			
29	PC- & Videospiele			
30	Sex			
31	Sex	Film		

(wird fortgesetzt)

Person	Eintrag 1	Eintrag 2	Eintrag 3	Eintrag 4
32	Sex	Reisen		
33	Sex	Waffen		
34	Soziale Kontakte	Menschenrechte	Krieg	Zustand anderer Länder
35	Tiere			
36	Tiere			
37	Tiere			
38	Tiere			
39	Tiere			
40	Tiere	Freunde		
41	Umweltschutz	Kommunikation		
42	Universum	SiFi	Extremsport	
43	Unterwasserbiologie			
44	Weltall	Sternbilder		

Anhang F: Kategorisierung der Items des I-S-T 2000 R Wissenstests und des Bochumer Wissenstests

Tabelle F1

Zuordnungen und Krippendorffs α-Werte der Items des I-S-T 2000 R Wissenstests (Form A)

Itemnr.	Kategorie	Krippendorffs α	Itemnr.	Kategorie	Krippendorffs α
204	IvM	1	231	NI	.16
205	IvM	1	232	IvM	1
206	NI	.44	233	IvM	1
207	IvM	1	234	IvM	.68
208	IvM	1	235	IvF	.08
209	IvM	1	236	IvM	.68
210	IvF	1	237	NI	1
211	IvM	1	238	NI	1
212	NI	.28	239	IvM	.44
213	NI	.68	240	IvM	1
214	IvM	1	241	IvM	.68
215	IvM	.68	242	IvM	1
216	IvM	1	243	NI	1
217	IvM	.68	244	IvM	.44
218	NI	1	245	IvM	1
219	IvM	1	246	NI	.68
220	IvM	1	247	IvM	.68
221	NI	.08	248	NI	1
222	IvM	.83	249	IvM	.68
223	IvM	1	250	NI	.68
224	NI	.44	251	NI	1
225	IvM	1	252	IvM	.68
226	IvM	.68	253	NI	.68
227	IvM	1	254	IvM	.68
228	IvM	1	255	IvM	.20
229	IvM	1	256	IvM	.68
230	IvM	1	257	NI	.68

(wird fortgesetzt)

Itemnr.	Kategorie	Krippendorffs α	Itemnr.	Kategorie	Krippendorffs α
258	IvM	.68	273	IvM	.83
259	NI	.44	274	NI	.40
260	NI	1	275	IvM	.20
261	IvM	.68	276	IvM	.20
262	NI	.68	277	IvM	1
263	NI	1	278	IvM	.44
264	IvM	.53	279	NI	.68
265	NI	1	280	IvM	.44
266	IvM	1	281	NI	.28
267	IvM	.68	282	NI	.44
268	NI	.68	283	NI	1
269	NI	.28	284	NI	.44
270	IvM	.28	285	IvM	.68
271	IvM	1	286	NI	.68
272	IvM	.54	287	IvM	.68

Anmerkungen. IvF = Interessengebiete von Frauen; NI = Neutrale Interessengebiete; IvM = Interessengebiete von Männern.

Tabelle F2

Zuordnungen und Krippendorffs α-Werte der Items des Bochumer Wissenstests (Form A)

Itemnr.	Kategorie	Krippendorffs α	Itemnr.	Kategorie	Krippendorffs α
1	NI	.68	31	NI	.28
2	IvM	1	32	IvM	1
3	IvM	1	33	IvM	1
4	IvM	1	34	IvM	1
5	IvM	1	35	IvF	1
6	IvM	.28	36	IvM	.28
7	IvM	1	37	IvM	1
8	NI	.68	38	IvM	1
9	IvM	.20	39	IvM	.28
10	IvM	.68	40	IvM	1
11	IvM	1	41	IvM	1
12	NI	1	42	NI	.40
13	IvM	1	43	IvM	1
14	IvF	.68	44	IvM	1
15	IvM	1	45	IvM	1
16	IvM	1	46	IvF	1
17	IvM	.68	47	IvF	1
18	IvM	1	48	IvM	1
19	NI	1	49	IvM	1
20	NI	.28	50	IvM	1
21	IvM	1	51	IvM	.44
22	IvM	1	52	NI	.28
23	NI	1	53	NI	1
24	IvF	.83	54	IvM	1
25	IvF	1	55	IvM	1
26	IvM	1	56	IvM	.68
27	IvM	1	57	IvM	1
28	IvM	1	58	IvF	1
29	IvM	1	59	IvM	1
30	NI	.68	60	NI	.20

(wird fortgesetzt)

Itemnr.	Kategorie	Krippendorffs α	Itemnr.	Kategorie	Krippendorffs α
61	IvM	1	93	IvM	1
62	NI	1	94	IvF	1
63	IvM	.68	95	IvM	.68
64	IvF	.44	96	NI	1
65	NI	.28	97	NI	.20
66	IvM	.83	98	IvM	1
67	IvM	.44	99	IvM	1
68	IvM	1	100	NI	.68
69	IvF	.44	101	IvM	.68
70	IvM	1	102	IvF	1
71	IvM	1	103	IvM	1
72	IvM	1	104	IvM	1
73	IvM	1	105	IvM	1
74	NI	1	106	IvM	.19
75	NI	.68	107	NI	1
76	IvM	.28	108	NI	.68
77	IvM	1	109	IvM	1
78	NI	1	110	IvM	1
79	IvF	.68	111	NI	.20
80	IvF	1	112	IvF	.44
81	IvM	1	113	IvF	.68
82	IvM	1	114	IvM	1
83	IvM	1	115	IvM	.28
84	NI	1	116	IvM	.28
85	NI	1	117	IvM	1
86	NI	1	118	NI	1
87	IvM	1	119	NI	1
88	IvM	1	120	IvM	.20
89	IvF/NI[a]	.04	121	IvM	1
90	IvM	.44	122	IvM	.32
91	IvF	1	123	IvF	1
92	IvM	1	124	IvF	1

(wird fortgesetzt)

Itemnr.	Kategorie	Krippendorffs α	Itemnr.	Kategorie	Krippendorffs α
125	IvM	1	140	NI	.68
126	IvM	1	141	NI	.28
127	IvM	1	142	IvM	1
128	IvM	1	143	IvM	1
129	NI	.68	144	NI	1
130	NI	1	145	IvM	1
131	IvM	1	146	IvF	1
132	IvM	1	147	IvM	1
133	IvM	.68	148	IvM	1
134	IvM	1	149	IvM	1
135	IvF	.44	150	IvM	1
136	IvM	1	151	IvM	.32
137	IvM	1	152	NI	1
138	IvM	1	153	IvM	1
139	NI	1	154	IvM	1

Anmerkungen. IvF = Interessengebiete von Frauen; NI = Neutrale Interessengebiete; IvM = Interessengebiete von Männern.

[a]Dieses Item wurde beiden Kategorien jeweils 4mal zugeordnet.

Anhang G: Skalenparameter des neu entwickelten Wissenstests

Tabelle G1

Skalenparameter des neuen Wissenstests (Datensatz 1.1)

	Vor Itemselektion			Nach Itemselektion		
Thema	M_S	M_{TS}	α	M_S	M_{TS}	α
Biologie	.43	.21	.74	.56	.39	.75
Darstellende Kunst	.38	.05	.22	.49	.19	.48
Ernährung	.43	.05	.27	.51	.20	.50
Gesundheit	.46	.13	.52	.51	.23	.55
Medizin	.43	.14	.56	.52	.29	.63
Modedesign	.38	.10	.46	.50	.24	.56
Natur	.40	.12	.51	.52	.23	.55
Pädagogik	.50	.08	.38	.62	.22	.52
Psychologie	.31	.08	.37	.41	.20	.50
Raumdesign	.42	.05	.27	.49	.17	.45
Soziale Arbeit	.36	.04	.22	.49	.21	.51
Gesamt	.41	.13	.88	.51	.26	.91

Anmerkungen. M_S = Mittelwert der Itemschwierigkeiten; M_{TS} = Mittelwert der Trennschärfen nach part-whole-Korrektur; α = Cronbachs α.

Tabelle G2

Skalenparameter des neuen Wissenstests (Datensatz 1.2)

	Vor Itemselektion			Nach Itemselektion		
Thema	M_S	M_{TS}	α	M_S	M_{TS}	α
Biologie	.43	.21	.73	.56	.37	.73
Darstellende Kunst	.38	.03	.16	.49	.19	.48
Ernährung	.43	.05	.24	.51	.16	.43
Gesundheit	.45	.13	.51	.50	.21	.51
Medizin	.43	.16	.59	.53	.30	.64
Modedesign	.38	.10	.46	.49	.24	.57
Natur	.40	.11	.47	.52	.23	.54
Pädagogik	.50	.10	.43	.61	.21	.51
Psychologie	.31	.07	.35	.41	.17	.45
Raumdesign	.42	.04	.24	.50	.15	.41
Soziale Arbeit	.36	.04	.18	.49	.20	.49
Gesamt	.41	.12	.87	.51	.25	.90

Anmerkungen. M_S = Mittelwert der Itemschwierigkeiten; M_{TS} = Mittelwert der Trennschärfen nach part-whole-Korrektur; α = Cronbachs α.

Tabelle G3

Skalenparameter des neuen Wissenstests (Datensatz 1.3)

	Vor Itemselektion			Nach Itemselektion		
Thema	M_S	M_{TS}	α	M_S	M_{TS}	α
Biologie	.43	.21	.74	.55	.42	.77
Darstellende Kunst	.38	.04	.21	.50	.18	.47
Ernährung	.43	.06	.28	.51	.18	.47
Gesundheit	.45	.13	.52	.50	.21	.51
Medizin	.42	.15	.58	.54	.30	.64
Modedesign	.39	.10	.45	.50	.24	.56
Natur	.40	.12	.50	.53	.21	.52
Pädagogik	.50	.08	.36	.61	.21	.51
Psychologie	.31	.08	.36	.41	.18	.46
Raumdesign	.42	.05	.25	.50	.17	.44
Soziale Arbeit	.36	.03	.18	.49	.20	.50
Gesamt	.41	.12	.87	.51	.25	.90

Anmerkungen. M_S = Mittelwert der Itemschwierigkeiten; M_{TS} = Mittelwert der Trennschärfen nach part-whole-Korrektur; α = Cronbachs α.

Tabelle G4

Skalenparameter des neuen Wissenstests (Datensatz 1.4)

	Vor Itemselektion			Nach Itemselektion		
Thema	M_S	M_{TS}	α	M_S	M_{TS}	α
Biologie	.43	.21	.75	.55	.41	.76
Darstellende Kunst	.38	.04	.22	.49	.18	.46
Ernährung	.44	.07	.33	.51	.20	.49
Gesundheit	.45	.12	.48	.50	.19	.49
Medizin	.43	.14	.56	.53	.29	.63
Modedesign	.38	.09	.44	.50	.23	.55
Natur	.40	.12	.51	.52	.20	.50
Pädagogik	.50	.09	.41	.61	.22	.52
Psychologie	.31	.09	.41	.40	.21	.52
Raumdesign	.41	.04	.24	.50	.17	.45
Soziale Arbeit	.36	.04	.21	.49	.22	.52
Gesamt	.41	.13	.88	.51	.26	.90

Anmerkungen. M_S = Mittelwert der Itemschwierigkeiten; M_{TS} = Mittelwert der Trennschärfen nach part-whole-Korrektur; α = Cronbachs α.

Tabelle G5

Skalenparameter des neuen Wissenstests (Datensatz 1.5)

	Vor Itemselektion			Nach Itemselektion		
Thema	M_S	M_{TS}	α	M_S	M_{TS}	α
Biologie	.43	.22	.74	.55	.38	.74
Darstellende Kunst	.38	.04	.20	.49	.20	.50
Ernährung	.44	.06	.27	.51	.18	.47
Gesundheit	.45	.14	.52	.50	.24	.57
Medizin	.42	.14	.55	.53	.29	.63
Modedesign	.39	.10	.47	.50	.25	.58
Natur	.40	.13	.54	.51	.23	.55
Pädagogik	.50	.09	.41	.61	.22	.52
Psychologie	.31	.10	.42	.40	.22	.53
Raumdesign	.42	.04	.23	.50	.18	.46
Soziale Arbeit	.35	.06	.30	.48	.22	.53
Gesamt	.41	.13	.88	.51	.26	.91

Anmerkungen. M_S = Mittelwert der Itemschwierigkeiten; M_{TS} = Mittelwert der Trennschärfen nach part-whole-Korrektur; α = Cronbachs α.

Tabelle G6

Skalenparameter des neuen Wissenstests (Datensatz 2.1)

	Vor Itemselektion			Nach Itemselektion		
Thema	M_S	M_{TS}	α	M_S	M_{TS}	α
Biologie	.43	.22	.75	.54	.39	.74
Darstellende Kunst	.38	.06	.29	.49	.18	.48
Ernährung	.44	.05	.24	.51	.17	.44
Gesundheit	.45	.13	.52	.50	.22	.53
Medizin	.42	.13	.53	.53	.26	.59
Modedesign	.38	.10	.45	.50	.22	.53
Natur	.40	.12	.52	.53	.24	.56
Pädagogik	.50	.10	.43	.62	.23	.54
Psychologie	.31	.09	.40	.40	.19	.49
Raumdesign	.42	.04	.23	.50	.16	.42
Soziale Arbeit	.35	.02	.11	.48	.18	.46
Gesamt	.41	.12	.87	.51	.25	.90

Anmerkungen. M_S = Mittelwert der Itemschwierigkeiten; M_{TS} = Mittelwert der Trennschärfen nach part-whole-Korrektur; α = Cronbachs α.

Tabelle G7

Skalenparameter des neuen Wissenstests (Datensatz 2.2)

	Vor Itemselektion			Nach Itemselektion		
Thema	M_S	M_{TS}	α	M_S	M_{TS}	α
Biologie	.42	.23	.77	.54	.38	.74
Darstellende Kunst	.39	.06	.29	.50	.18	.47
Ernährung	.44	.05	.23	.50	.18	.46
Gesundheit	.45	.12	.47	.49	.19	.49
Medizin	.42	.14	.55	.53	.27	.60
Modedesign	.38	.11	.48	.50	.21	.52
Natur	.40	.12	.53	.52	.24	.57
Pädagogik	.50	.10	.41	.61	.21	.51
Psychologie	.31	.10	.42	.40	.20	.50
Raumdesign	.42	.06	.30	.50	.18	.46
Soziale Arbeit	.35	.04	.21	.48	.17	.44
Gesamt	.41	.13	.88	.51	.26	.90

Anmerkungen. M_S = Mittelwert der Itemschwierigkeiten; M_{TS} = Mittelwert der Trennschärfen nach part-whole-Korrektur; α = Cronbachs α.

Tabelle G8

Skalenparameter des neuen Wissenstests (Datensatz 2.3)

	Vor Itemselektion			Nach Itemselektion		
Thema	M_S	M_{TS}	α	M_S	M_{TS}	α
Biologie	.43	.23	.77	.55	.38	.74
Darstellende Kunst	.38	.06	.29	.49	.19	.48
Ernährung	.44	.06	.27	.50	.19	.48
Gesundheit	.44	.14	.53	.49	.22	.53
Medizin	.42	.14	.54	.53	.27	.60
Modedesign	.38	.11	.48	.49	.23	.56
Natur	.40	.13	.53	.52	.22	.53
Pädagogik	.50	.10	.42	.61	.23	.53
Psychologie	.31	.09	.38	.40	.20	.50
Raumdesign	.42	.04	.23	.49	.19	.47
Soziale Arbeit	.36	.04	.19	.49	.14	.39
Gesamt	.41	.13	.88	.51	.25	.90

Anmerkungen. M_S = Mittelwert der Itemschwierigkeiten; M_{TS} = Mittelwert der Trennschärfen nach part-whole-Korrektur; α = Cronbachs α.

Tabelle G9

Skalenparameter des neuen Wissenstests (Datensatz 2.4)

	Vor Itemselektion			Nach Itemselektion		
Thema	M_S	M_{TS}	α	M_S	M_{TS}	α
Biologie	.43	.22	.75	.56	.37	.73
Darstellende Kunst	.38	.06	.29	.49	.17	.46
Ernährung	.44	.06	.28	.51	.20	.49
Gesundheit	.45	.12	.49	.49	.20	.51
Medizin	.43	.14	.55	.54	.26	.59
Modedesign	.38	.11	.50	.50	.25	.58
Natur	.40	.13	.54	.52	.22	.54
Pädagogik	.50	.10	.44	.61	.21	.50
Psychologie	.31	.09	.39	.40	.20	.49
Raumdesign	.42	.05	.27	.49	.17	.45
Soziale Arbeit	.35	.02	.14	.49	.16	.42
Gesamt	.41	.13	.88	.51	.25	.90

Anmerkungen. M_S = Mittelwert der Itemschwierigkeiten; M_{TS} = Mittelwert der Trennschärfen nach part-whole-Korrektur; α = Cronbachs α.

Tabelle G10

Skalenparameter des neuen Wissenstests (Datensatz 2.5)

	Vor Itemselektion			Nach Itemselektion		
Thema	M_S	M_{TS}	α	M_S	M_{TS}	α
Biologie	.42	.24	.77	.55	.39	.74
Darstellende Kunst	.38	.07	.31	.50	.19	.48
Ernährung	.44	.07	.31	.51	.20	.49
Gesundheit	.44	.13	.51	.48	.21	.53
Medizin	.42	.15	.56	.53	.28	.62
Modedesign	.38	.11	.48	.49	.21	.52
Natur	.40	.13	.54	.52	.24	.57
Pädagogik	.50	.10	.42	.61	.21	.50
Psychologie	.31	.08	.38	.40	.18	.47
Raumdesign	.42	.04	.24	.50	.18	.46
Soziale Arbeit	.35	.04	.20	.49	.17	.44
Gesamt	.41	.13	.88	.51	.25	.90

Anmerkungen. M_S = Mittelwert der Itemschwierigkeiten; M_{TS} = Mittelwert der Trennschärfen nach part-whole-Korrektur; α = Cronbachs α.

Anhang H: Geschlechterdifferenzen im neu entwickelten Wissenstest

Tabelle H1

Effektstärken und t-Tests der Geschlechterdifferenzen im neu entwickelten Wissenstest (Datensatz 1.1)

Thema	Cohens *d*	t_{245}	*p* (2-seitig)
Biologie	-0.34	-2.61	.010
Darstellende Kunst	0.13	1.01	.315
Ernährung	-0.02	-0.15	.880
Gesundheit	0.10	0.78	.434
Medizin	-0.65	-5.06	< .001
Modedesign	0.53	4.09	< .001
Natur	-0.12	-0.95	.342
Pädagogik	0.23	1.78	.077
Psychologie	-0.31	-2.41	.017
Raumdesign	0.04	0.28	.779
Soziale Arbeit	-0.46	-3.61	< .001
Gesamtscore	-0.14	-1.06	.292

Anmerkung. Effektstärken und *t*-Werte mit negativem Vorzeichen weisen auf höhere Durchschnittswerte bei Männern hin.

Tabelle H2

Effektstärken und t-Tests der Geschlechterdifferenzen im neu entwickelten Wissenstest (Datensatz 1.2)

Thema	Cohens *d*	t_{245}	*p* (2-seitig)
Biologie	-0.35	-2.69	.008
Darstellende Kunst	0.05	0.37	.714
Ernährung	0.07	0.55	.583
Gesundheit	0.10	0.78	.434
Medizin	-0.69	-5.33	< .001
Modedesign	0.64	4.98	< .001
Natur	-0.05	-0.36	.716
Pädagogik	0.20	1.57	.118
Psychologie	-0.32	-2.53	.012
Raumdesign	0.15	1.17	.245
Soziale Arbeit	-0.53	-4.11	< .001
Gesamtscore	-0.11	-0.88	.379

Anmerkung. Effektstärken und *t*-Werte mit negativem Vorzeichen weisen auf höhere Durchschnittswerte bei Männern hin.

Tabelle H3

Effektstärken und t-Tests der Geschlechterdifferenzen im neu entwickelten Wissenstest (Datensatz 1.3)

Thema	Cohens *d*	t_{245}	*p* (2-seitig)
Biologie	-0.34	-2.66	.008
Darstellende Kunst	0.13	1.03	.305
Ernährung	0.06	0.43	.664
Gesundheit	0.12	0.90	.367
Medizin	-0.71	-5.67[a]	< .001
Modedesign	0.55	4.30	< .001
Natur	-0.16	-1.26	.208
Pädagogik	0.25	1.95	.053
Psychologie	-0.39	-3.03	.003
Raumdesign	0.10	0.81	.421
Soziale Arbeit	-0.50	-3.89	< .001
Gesamtscore	-0.14	-1.09	.276

Anmerkungen. Effektstärken und *t*-Werte mit negativem Vorzeichen weisen auf höhere Durchschnittswerte bei Männern hin.

[a]Dieser *t*-Wert bezieht sich auf die *t*-Verteilung mit *df* = 224.25, da der Levene-Test auf Homoskedastizität hier ein signifikantes Ergebnis lieferte (*p* = .002).

Tabelle H4

Effektstärken und t-Tests der Geschlechterdifferenzen im neu entwickelten Wissenstest (Datensatz 1.4)

Thema	Cohens *d*	t_{245}	*p* (2-seitig)
Biologie	-0.34	-2.45	.015
Darstellende Kunst	0.13	1.16	.246
Ernährung	0.06	0.23	.816
Gesundheit	0.12	1.17	.242
Medizin	-0.71	-5.10	< .001
Modedesign	0.55	4.49	< .001
Natur	-0.16	-1.37	.173
Pädagogik	0.25	2.21	.028
Psychologie	-0.39	-3.29	.001
Raumdesign	0.10	0.91	.366
Soziale Arbeit	-0.50	-3.67	< .001
Gesamtscore	-0.14	0.93	.354

Anmerkung. Effektstärken und *t*-Werte mit negativem Vorzeichen weisen auf höhere Durchschnittswerte bei Männern hin.

Tabelle H5

Effektstärken und t-Tests der Geschlechterdifferenzen im neu entwickelten Wissenstest (Datensatz 1.5)

Thema	Cohens *d*	t_{245}	*p* (2-seitig)
Biologie	-0.34	-2.62	.009
Darstellende Kunst	0.12	0.93	.354
Ernährung	0.01	0.10	.917
Gesundheit	0.06	0.48	.631
Medizin	-0.62	-4.83	< .001
Modedesign	0.59	4.61	< .001
Natur	-0.22	-1.74	.084
Pädagogik	0.16	1.22	.222
Psychologie	-0.50	-3.85	< .001
Raumdesign	0.07	0.58	.564
Soziale Arbeit	-0.48	-3.77	< .001
Gesamtscore	-0.17	-1.31	.190

Anmerkung. Effektstärken und *t*-Werte mit negativem Vorzeichen weisen auf höhere Durchschnittswerte bei Männern hin.

Tabelle H6

Effektstärken und t-Tests der Geschlechterdifferenzen im neu entwickelten Wissenstest (Datensatz 2.1)

Thema	Cohens *d*	t_{245}	*p* (2-seitig)
Biologie	-0.35	-2.74	.007
Darstellende Kunst	0.12	0.96	.337
Ernährung	0.00	0.03	.974
Gesundheit	0.04	0.31	.760
Medizin	-0.58	-4.53	< .001
Modedesign	0.73	5.71	< .001
Natur	-0.12	-0.96	.336
Pädagogik	0.18	1.39	.164
Psychologie	-0.40	-3.13	.002
Raumdesign	0.04	0.28	.782
Soziale Arbeit	-0.47	-3.66	< .001
Gesamtscore	-0.13	-0.98	.326

Anmerkung. Effektstärken und *t*-Werte mit negativem Vorzeichen weisen auf höhere Durchschnittswerte bei Männern hin.

Tabelle H7

Effektstärken und t-Tests der Geschlechterdifferenzen im neu entwickelten Wissenstest (Datensatz 2.2)

Thema	Cohens *d*	t_{245}	*p* (2-seitig)
Biologie	-0.33	-2.56	.011
Darstellende Kunst	0.21	1.60	.111
Ernährung	-0.03	-0.25	.799
Gesundheit	0.05	0.39	.695
Medizin	-0.58	-4.51	< .001
Modedesign	0.72	5.58	< .001
Natur	-0.18	-1.37	.172
Pädagogik	0.25	1.91	.057
Psychologie	-0.33	-2.56	.011
Raumdesign	0.07	0.54	.588
Soziale Arbeit	-0.55	-4.27	< .001
Gesamtscore	-0.11	-0.86	.388

Anmerkung. Effektstärken und *t*-Werte mit negativem Vorzeichen weisen auf höhere Durchschnittswerte bei Männern hin.

Tabelle H8

Effektstärken und t-Tests der Geschlechterdifferenzen im neu entwickelten Wissenstest (Datensatz 2.4)

Thema	Cohens *d*	t_{245}	*p* (2-seitig)
Biologie	-0.37	-2.89	.004
Darstellende Kunst	0.18	1.44	.152
Ernährung	0.05	0.36	.723
Gesundheit	0.07	0.51	0.613
Medizin	-0.63	-4.88	< .001
Modedesign	0.67	5.23	< .001
Natur	-0.08	-0.61	.542
Pädagogik	0.16	1.25	.213
Psychologie	-0.38	-2.96	.003
Raumdesign	0.10	0.77	.442
Soziale Arbeit	-0.57	-4.44	< .001
Gesamtscore	-0.11	-0.89	.374

Anmerkung. Effektstärken und *t*-Werte mit negativem Vorzeichen weisen auf höhere Durchschnittswerte bei Männern hin.

Tabelle H9

Effektstärken und t-Tests der Geschlechterdifferenzen im neu entwickelten Wissenstest (Datensatz 2.5)

Thema	Cohens *d*	t_{245}	*p* (2-seitig)
Biologie	-0.36	-2.77	.006
Darstellende Kunst	0.17	1.29	.198
Ernährung	0.03	0.21	.830
Gesundheit	0.10	0.78	.437
Medizin	-0.68	-5.27	< .001
Modedesign	0.75	5.83	< .001
Natur	-0.17	-1.31	.192
Pädagogik	0.21	1.61	.108
Psychologie	-0.33	-2.54	.012
Raumdesign	0.08	0.60	.546
Soziale Arbeit	-0.55	-4.26	< .001
Gesamtscore	-0.12	-0.93	.352

Anmerkung. Effektstärken und *t*-Werte mit negativem Vorzeichen weisen auf höhere Durchschnittswerte bei Männern hin.

Anhang I: Überprüfung des neu entwickelten Wissenstests auf 1-faktorielle Struktur

Tabelle I1

Standardisierte Parameterschätzungen und Model-Fit der 1-faktoriellen Messmodelle (Datensatz 1.1)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Wissen	0[a]		-0.16	0[a]
[95%-KI]			[-0.43; 0.10]	
β-Koeffizienten				
Biologie	.67		.67	
Darstellende Kunst	.61		.59	
Ernährung	.71		.71	
Gesundheit	.59		.58	
Medizin	.71		.70	
Modedesign	.70		.64	
Natur	.64		.64	
Pädagogik	.47		.45	
Psychologie	.67		.68	
Raumdesign	.58		.58	
Soziale Arbeit	.50		.50	
Achsenabschnitte				
Biologie	2.07	2.41	2.25	
Darstellende Kunst	2.77	2.64	2.77	
Ernährung	2.66	2.68	2.74	
Gesundheit	2.68	2.58	2.69	
Medizin	2.25	2.91	2.48	
Modedesign	2.73	2.20	2.49	
Natur	2.43	2.55	2.54	
Pädagogik	3.43	3.20	3.36	
Psychologie	2.04	2.35	2.21	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Raumdesign	2.65	2.62	2.69	
Soziale Arbeit	2.44	2.91	2.62	
Fehlervarianzen				
Biologie	.55		.55	
Darstellende Kunst	.63		.65	
Ernährung	.50		.50	
Gesundheit	.65		.66	
Medizin	.49		.52	
Modedesign	.51		.59	
Natur	.59		.59	
Pädagogik	.78		.80	
Psychologie	.55		.54	
Raumdesign	.66		.67	
Soziale Arbeit	.75		.75	
Model-Fit				
χ^2_{df}	121.52_{110}		242.36_{120}	
p	.21		< .001	
SCF	0.96		0.96	
CFI	.99		.86	
SRMR	.07		0.10	
RMSEA	.03		0.09	
[90%-KI]	[.00; .06]		[.07; .11]	
χ^2-Diff-Test: χ^2_{df}; *p*	116.95_{10}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle I2

Standardisierte Parameterschätzungen und Model-Fit der 1-faktoriellen Messmodelle (Datensatz 1.2)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Wissen	0[a]		-0.13	0[a]
[95%-KI]			[-0.39; 0.14]	
β-Koeffizienten				
Biologie	.65		.65	
Darstellende Kunst	.65		.64	
Ernährung	.68		.67	
Gesundheit	.59		.59	
Medizin	.70		.67	
Modedesign	.70		.62	
Natur	.64		.65	
Pädagogik	.50		.49	
Psychologie	.71		.71	
Raumdesign	.60		.59	
Soziale Arbeit	.49		.49	
Achsenabschnitte				
Biologie	2.16	2.51	2.33	
Darstellende Kunst	2.71	2.66	2.73	
Ernährung	2.85	2.78	2.87	
Gesundheit	2.79	2.69	2.79	
Medizin	2.26	2.95	2.47	
Modedesign	2.76	2.12	2.42	
Natur	2.48	2.53	2.55	
Pädagogik	3.42	3.22	3.36	
Psychologie	2.13	2.46	2.30	
Raumdesign	2.84	2.69	2.81	
Soziale Arbeit	2.48	3.01	2.65	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
Biologie	.58		.58	
Darstellende Kunst	.58		.59	
Ernährung	.54		.55	
Gesundheit	.66		.66	
Medizin	.51		.55	
Modedesign	.51		.61	
Natur	.59		.58	
Pädagogik	.75		.76	
Psychologie	.49		.50	
Raumdesign	.65		.66	
Soziale Arbeit	.76		.76	
Model-Fit				
χ^2_{df}	157.66_{110}		170.17_{120}	
p	.002		< .001	
SCF	0.95		0.96	
CFI	.95		.81	
SRMR	.09		.11	
RMSEA	.06		.11	
[90%-KI]	[.04; .08]		[.10; .13]	
χ^2-Diff-Test: χ^2_{df}; *p*	142.16_{10}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle I3

Standardisierte Parameterschätzungen und Model-Fit der 1-faktoriellen Messmodelle (Datensatz 1.3)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Wissen	0[a]		-0.16	0[a]
[95%-KI]			[-0.43; 0.10]	
β-Koeffizienten				
Biologie	.66		.67	
Darstellende Kunst	.61		.59	
Ernährung	.65		.65	
Gesundheit	.59		.58	
Medizin	.67		.65	
Modedesign	.68		.62	
Natur	.65		.66	
Pädagogik	.52		.50	
Psychologie	.68		.68	
Raumdesign	.56		.55	
Soziale Arbeit	.46		.47	
Achsenabschnitte				
Biologie	1.99	2.33	2.17	
Darstellende Kunst	2.80	2.67	2.79	
Ernährung	2.75	2.69	2.79	
Gesundheit	2.76	2.65	2.76	
Medizin	2.28	2.99	2.51	
Modedesign	2.77	2.21	2.50	
Natur	2.51	2.68	2.64	
Pädagogik	3.43	3.18	3.34	
Psychologie	2.05	2.44	2.25	
Raumdesign	2.75	2.65	2.76	
Soziale Arbeit	2.48	2.98	2.66	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
Biologie	.56		.56	
Darstellende Kunst	.63		.65	
Ernährung	.57		.58	
Gesundheit	.66		.67	
Medizin	.55		.57	
Modedesign	.53		.62	
Natur	.57		.56	
Pädagogik	.73		.75	
Psychologie	.54		.54	
Raumdesign	.69		.70	
Soziale Arbeit	.79		.78	
Model-Fit				
χ^2_{df}	165.58_{110}		304.46_{120}	
p	< .001		< .001	
SCF	0.95		0.96	
CFI	.94		.79	
SRMR	.09		.12	
RMSEA	.06		.11	
[90%-KI]	[.04; .08]		[.10; .13]	
χ^2-Diff-Test: χ^2_{df}; *p*	133.43_{10}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle I4

Standardisierte Parameterschätzungen und Model-Fit der 1-faktoriellen Messmodelle (Datensatz 1.4)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Wissen	0[a]		-0.16	0[a]
[95%-KI]			[-0.42; 0.11]	
β-Koeffizienten				
Biologie	.64		.64	
Darstellende Kunst	.61		.60	
Ernährung	.68		.68	
Gesundheit	.61		.59	
Medizin	.70		.68	
Modedesign	.69		.62	
Natur	.63		.64	
Pädagogik	.52		.50	
Psychologie	.71		.71	
Raumdesign	.59		.57	
Soziale Arbeit	.53		.53	
Achsenabschnitte				
Biologie	2.04	2.35	2.21	
Darstellende Kunst	2.80	2.65	2.78	
Ernährung	2.70	2.67	2.75	
Gesundheit	2.86	2.71	2.84	
Medizin	2.30	2.96	2.52	
Modedesign	2.81	2.23	2.52	
Natur	2.53	2.71	2.66	
Pädagogik	3.41	3.12	3.30	
Psychologie	1.93	2.36	2.13	
Raumdesign	2.74	2.63	2.74	
Soziale Arbeit	2.42	2.89	2.60	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
Biologie	.59		.59	
Darstellende Kunst	.63		.65	
Ernährung	.53		.54	
Gesundheit	.63		.65	
Medizin	.51		.54	
Modedesign	.53		.62	
Natur	.60		.59	
Pädagogik	.73		.75	
Psychologie	.50		.50	
Raumdesign	.66		.67	
Soziale Arbeit	.72		.72	
Model-Fit				
χ^2_{df}	141.09_{110}		283.00_{120}	
p	.024		< .001	
SCF	0.97		0.97	
CFI	.97		.82	
SRMR	.08		.11	
RMSEA	.05		.11	
[90%-KI]	[.02; .07]		[.09; .12]	
χ^2-Diff-Test: χ^2_{df}; *p*	138.78_{10}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle I5

Standardisierte Parameterschätzungen und Model-Fit der 1-faktoriellen Messmodelle (Datensatz 1.5)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Wissen	0[a]		-0.22	0[a]
[95%-KI]			[-0.49; 0.05]	
β-Koeffizienten				
Biologie	.65		.66	
Darstellende Kunst	.62		.60	
Ernährung	.68		.67	
Gesundheit	.59		.59	
Medizin	.71		.70	
Modedesign	.69		.60	
Natur	.65		.66	
Pädagogik	.49		.47	
Psychologie	.68		.68	
Raumdesign	.58		.56	
Soziale Arbeit	.49		.50	
Achsenabschnitte				
Biologie	2.08	2.42	2.28	
Darstellende Kunst	2.72	2.60	2.75	
Ernährung	2.78	2.76	2.86	
Gesundheit	2.59	2.53	2.64	
Medizin	2.29	2.91	2.54	
Modedesign	2.76	2.17	2.49	
Natur	2.38	2.61	2.55	
Pädagogik	3.35	3.19	3.34	
Psychologie	1.90	2.40	2.14	
Raumdesign	2.71	2.63	2.75	
Soziale Arbeit	2.35	2.84	2.55	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
Biologie	.58		.57	
Darstellende Kunst	.61		.64	
Ernährung	.54		.56	
Gesundheit	.65		.66	
Medizin	.50		.51	
Modedesign	.52		.64	
Natur	.58		.57	
Pädagogik	.76		.78	
Psychologie	.54		.54	
Raumdesign	.67		.69	
Soziale Arbeit	.76		.75	
Model-Fit				
χ^2_{df}	170.73_{110}		304.73_{120}	
p	< .001		< .001	
SCF	0.96		0.96	
CFI	.93		.80	
SRMR	.09		.12	
RMSEA	.07		.11	
[90%-KI]	[.05; .09]		[.10; .13]	
χ^2-Diff-Test: χ^2_{df}; *p*	129.95_{10}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle I6

Standardisierte Parameterschätzungen und Model-Fit der 1-faktoriellen Messmodelle (Datensatz 2.1)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Wissen	0[a]		-0.19	0[a]
[95%-KI]			[-0.45; 0.08]	
β-Koeffizienten				
Biologie	.63		.63	
Darstellende Kunst	.64		.62	
Ernährung	.70		.69	
Gesundheit	.63		.63	
Medizin	.73		.72	
Modedesign	.67		.57	
Natur	.66		.66	
Pädagogik	.45		.44	
Psychologie	.70		.70	
Raumdesign	.56		.56	
Soziale Arbeit	.52		.53	
Achsenabschnitte				
Biologie	2.00	2.36	2.19	
Darstellende Kunst	2.77	2.65	2.78	
Ernährung	2.75	2.75	2.83	
Gesundheit	2.69	2.66	2.75	
Medizin	2.41	2.99	2.64	
Modedesign	2.99	2.25	2.58	
Natur	2.48	2.60	2.60	
Pädagogik	3.36	3.18	3.32	
Psychologie	1.97	2.38	2.18	
Raumdesign	2.75	2.71	2.79	
Soziale Arbeit	2.50	2.97	2.69	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
Biologie	.60		.60	
Darstellende Kunst	.59		.61	
Ernährung	.52		.52	
Gesundheit	.60		.60	
Medizin	.47		.49	
Modedesign	.55		.67	
Natur	.57		.57	
Pädagogik	.80		.81	
Psychologie	.51		.51	
Raumdesign	.68		.69	
Soziale Arbeit	.73		.72	
Model-Fit				
χ^2_{df}	135.04_{110}		272.79_{120}	
p	.053		< .001	
SCF	0.98		0.99	
CFI	.97		.83	
SRMR	.08		.11	
RMSEA	.04		.10	
[90%-KI]	[.00; .07]		[.09; .12]	
χ^2-Diff-Test: χ^2_{df}; *p*	136.50_{10}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle I7

Standardisierte Parameterschätzungen und Model-Fit der 1-faktoriellen Messmodelle (Datensatz 2.2)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Wissen	0[a]		-0.17	0[a]
[95%-KI]			[-0.44; 0.09]	
β-Koeffizienten				
Biologie	.66		.66	
Darstellende Kunst	.66		.64	
Ernährung	.69		.69	
Gesundheit	.63		.63	
Medizin	.74		.73	
Modedesign	.66		.57	
Natur	.65		.66	
Pädagogik	.50		.48	
Psychologie	.69		.70	
Raumdesign	.59		.58	
Soziale Arbeit	.56		.56	
Achsenabschnitte				
Biologie	2.05	2.38	2.23	
Darstellende Kunst	2.85	2.64	2.82	
Ernährung	2.70	2.74	2.79	
Gesundheit	2.74	2.69	2.79	
Medizin	2.38	2.97	2.61	
Modedesign	2.99	2.27	2.59	
Natur	2.40	2.58	2.54	
Pädagogik	3.43	3.18	3.35	
Psychologie	1.98	2.31	2.17	
Raumdesign	2.68	2.61	2.71	
Soziale Arbeit	2.51	3.06	2.71	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
Biologie	.57		.57	
Darstellende Kunst	.56		.59	
Ernährung	.53		.53	
Gesundheit	.60		.61	
Medizin	.45		.47	
Modedesign	.56		.68	
Natur	.58		.57	
Pädagogik	.75		.77	
Psychologie	.52		.52	
Raumdesign	.66		.67	
Soziale Arbeit	.69		.69	
Model-Fit				
χ^2_{df}	139.76_{110}		286.62_{120}	
p	.029		< .001	
SCF	0.96		0.97	
CFI	.97		.83	
SRMR	.08		.11	
RMSEA	.05		.11	
[90%-KI]	[.02; .07]		[.09; .12]	
χ^2-Diff-Test: χ^2_{df}; *p*	142.85_{10}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle I8

Standardisierte Parameterschätzungen und Model-Fit der 1-faktoriellen Messmodelle (Datensatz 2.3)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Wissen	0[a]		-0.16	0[a]
[95%-KI]			[-0.42; 0.11]	
β-Koeffizienten				
Biologie	.62		.62	
Darstellende Kunst	.63		.62	
Ernährung	.65		.65	
Gesundheit	.64		.63	
Medizin	.71		.69	
Modedesign	.67		.60	
Natur	.65		.66	
Pädagogik	.52		.50	
Psychologie	.72		.72	
Raumdesign	.57		.56	
Soziale Arbeit	.57		.56	
Achsenabschnitte				
Biologie	2.07	2.42	2.25	
Darstellende Kunst	2.78	2.59	2.74	
Ernährung	2.69	2.68	2.75	
Gesundheit	2.65	2.55	2.67	
Medizin	2.43	3.02	2.64	
Modedesign	2.79	2.18	2.47	
Natur	2.47	2.62	2.59	
Pädagogik	3.36	3.12	3.28	
Psychologie	1.98	2.30	2.16	
Raumdesign	2.61	2.54	2.63	
Soziale Arbeit	2.66	3.23	2.85	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
Biologie	.62		.62	
Darstellende Kunst	.60		.62	
Ernährung	.58		.58	
Gesundheit	.60		.61	
Medizin	.50		.52	
Modedesign	.55		.64	
Natur	.57		.57	
Pädagogik	.73		.75	
Psychologie	.49		.49	
Raumdesign	.68		.69	
Soziale Arbeit	.68		.68	
Model-Fit				
χ^2_{df}	156.28_{110}		291.32_{120}	
p	.003		< .001	
SCF	0.97		0.97	
CFI	.95		.82	
SRMR	.08		.11	
RMSEA	.06		.11	
[90%-KI]	[.04; .08]		[.09; .12]	
χ^2-Diff-Test: χ^2_{df}; *p*	132.37_{10}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle I9

Standardisierte Parameterschätzungen und Model-Fit der 1-faktoriellen Messmodelle (Datensatz 2.4)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Wissen	0[a]		-0.15	0[a]
[95%-KI]			[-0.41; 0.12]	
β-Koeffizienten				
Biologie	.63		.63	
Darstellende Kunst	.66		.64	
Ernährung	.65		.65	
Gesundheit	.65		.64	
Medizin	.69		.68	
Modedesign	.69		.61	
Natur	.65		.65	
Pädagogik	.52		.51	
Psychologie	.71		.71	
Raumdesign	.59		.58	
Soziale Arbeit	.55		.55	
Achsenabschnitte				
Biologie	2.15	2.52	2.33	
Darstellende Kunst	2.83	2.65	2.80	
Ernährung	2.70	2.66	2.74	
Gesundheit	2.70	2.64	2.73	
Medizin	2.46	3.09	2.67	
Modedesign	2.84	2.16	2.47	
Natur	2.52	2.60	2.61	
Pädagogik	3.36	3.20	3.33	
Psychologie	1.97	2.35	2.16	
Raumdesign	2.66	2.56	2.67	
Soziale Arbeit	2.62	3.19	2.80	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
Biologie	.61		.60	
Darstellende Kunst	.57		.59	
Ernährung	.57		.58	
Gesundheit	.58		.59	
Medizin	.52		.54	
Modedesign	.53		.63	
Natur	.58		.58	
Pädagogik	.73		.75	
Psychologie	.50		.50	
Raumdesign	.65		.66	
Soziale Arbeit	.69		.70	
Model-Fit				
χ^2_{df}	145.68_{110}		297.18_{120}	
p	.013		< .001	
SCF	0.96		0.97	
CFI	.96		.81	
SRMR	.07		.11	
RMSEA	.05		.11	
[90%-KI]	[.03; .07]		[.09; .13]	
χ^2-Diff-Test: χ^2_{df}; *p*	146.93_{10}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle I10

Standardisierte Parameterschätzungen und Model-Fit der 1-faktoriellen Messmodelle (Datensatz 2.5)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswert				
Wissen	0[a]		-0.14	0[a]
[95%-KI]			[-0.41; 0.13]	
β-Koeffizienten				
Biologie	.62		.62	
Darstellende Kunst	.68		.66	
Ernährung	.67		.67	
Gesundheit	.66		.66	
Medizin	.71		.69	
Modedesign	.67		.58	
Natur	.64		.65	
Pädagogik	.50		.49	
Psychologie	.68		.68	
Raumdesign	.57		.57	
Soziale Arbeit	.50		.50	
Achsenabschnitte				
Biologie	2.04	2.40	2.22	
Darstellende Kunst	2.80	2.63	2.77	
Ernährung	2.69	2.66	2.74	
Gesundheit	2.67	2.57	2.68	
Medizin	2.32	3.00	2.53	
Modedesign	2.98	2.23	2.54	
Natur	2.41	2.57	2.52	
Pädagogik	3.44	3.23	3.37	
Psychologie	2.06	2.38	2.23	
Raumdesign	2.67	2.59	2.68	
Soziale Arbeit	2.56	3.11	2.74	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Fehlervarianzen				
Biologie	.62		.62	
Darstellende Kunst	.55		.57	
Ernährung	.55		.55	
Gesundheit	.56		.57	
Medizin	.50		.53	
Modedesign	.55		.67	
Natur	.59		.58	
Pädagogik	.75		.76	
Psychologie	.54		.54	
Raumdesign	.67		.68	
Soziale Arbeit	.75		.75	
Model-Fit				
χ^2_{df}	149.40_{110}		311.98_{120}	
p	.007		< .001	
SCF	0.96		0.96	
CFI	.96		.80	
SRMR	.08		.11	
RMSEA	.05		.11	
[90%-KI]	[.03; .08]		[.10; .13]	
χ^2-Diff-Test: χ^2_{df}; *p*	158.11_{10}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Anhang J: Parallelanalysen des neu entwickelten Wissenstests

Tabelle J1

Ergebnisse der Parallelanalyse des neu entwickelten Wissenstests (Datensatz 1.1)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	4.18	0.40	0.51
2	0.21	0.29	0.37
3	0.17	0.21	0.28
4	0.04	0.14	0.20
5	0.02	0.08	0.13
6	0.00	0.02	0.07
7	-0.06	-0.03	0.01
8	-0.09	-0.09	-0.05
9	-0.13	-0.14	-0.10
10	-0.16	-0.19	-0.15
11	-0.22	-0.26	-0.21

Tabelle J2

Ergebnisse der Parallelanalyse des neu entwickelten Wissenstests (Datensatz 1.2)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	4.18	0.40	0.51
2	0.21	0.29	0.37
3	0.17	0.21	0.28
4	0.04	0.14	0.20
5	0.02	0.08	0.13
6	0.00	0.02	0.07
7	-0.06	-0.03	0.01
8	-0.09	-0.09	-0.05
9	-0.13	-0.14	-0.10
10	-0.16	-0.19	-0.15
11	-0.22	-0.26	-0.21

Tabelle J3

Ergebnisse der Parallelanalyse des neu entwickelten Wissenstests (Datensatz 1.3)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	4.02	0.40	0.50
2	0.30	0.30	0.37
3	0.22	0.21	0.28
4	0.15	0.14	0.20
5	0.06	0.08	0.13
6	-0.01	0.02	0.07
7	-0.09	-0.03	0.01
8	-0.11	-0.09	-0.05
9	-0.14	-0.14	-0.10
10	-0.18	-0.19	-0.16
11	-0.24	-0.26	-0.21

Tabelle J4

Ergebnisse der Parallelanalyse des neu entwickelten Wissenstests (Datensatz 1.4)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	4.20	0.40	0.51
2	0.30	0.29	0.37
3	0.19	0.21	0.28
4	0.09	0.14	0.20
5	0.04	0.08	0.13
6	0.01	0.02	0.07
7	-0.04	-0.03	0.01
8	-0.09	-0.08	-0.05
9	-0.13	-0.14	-0.10
10	-0.18	-0.19	-0.15
11	-0.25	-0.26	-0.21

Tabelle J5

Ergebnisse der Parallelanalyse des neu entwickelten Wissenstests (Datensatz 1.5)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	4.14	0.40	0.50
2	0.35	0.29	0.37
3	0.19	0.21	0.28
4	0.10	0.14	0.20
5	0.08	0.08	0.13
6	-0.01	0.02	0.07
7	-0.07	-0.03	0.01
8	-0.08	-0.09	-0.05
9	-0.14	-0.14	-0.10
10	-0.17	-0.19	-0.16
11	-0.25	-0.26	-0.21

Tabelle J6

Ergebnisse der Parallelanalyse des neu entwickelten Wissenstests (Datensatz 2.1)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	4.22	0.40	0.51
2	0.29	0.29	0.37
3	0.19	0.21	0.28
4	0.07	0.15	0.21
5	0.05	0.08	0.13
6	-0.03	0.02	0.07
7	-0.05	-0.03	0.01
8	-0.07	-0.09	-0.05
9	-0.12	-0.14	-0.10
10	-0.16	-0.19	-0.15
11	-0.26	-0.26	-0.21

Tabelle J7

Ergebnisse der Parallelanalyse des neu entwickelten Wissenstests (Datensatz 2.2)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	4.36	0.40	0.51
2	0.40	0.29	0.37
3	0.17	0.22	0.28
4	0.07	0.14	0.20
5	0.02	0.08	0.13
6	-0.04	0.02	0.06
7	-0.07	-0.03	0.00
8	-0.11	-0.09	-0.05
9	-0.14	-0.14	-0.10
10	-0.17	-0.19	-0.16
11	-0.20	-0.26	-0.22

Tabelle J8

Ergebnisse der Parallelanalyse des neu entwickelten Wissenstests (Datensatz 2.3)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	4.24	0.40	0.51
2	0.34	0.30	0.37
3	0.10	0.21	0.27
4	0.10	0.14	0.20
5	0.02	0.08	0.13
6	-0.04	0.02	0.07
7	-0.08	-0.03	0.01
8	-0.09	-0.08	-0.04
9	-0.13	-0.14	-0.10
10	-0.16	-0.20	-0.15
11	-0.21	-0.26	-0.22

Tabelle J9

Ergebnisse der Parallelanalyse des neu entwickelten Wissenstests (Datensatz 2.4)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	4.28	0.40	0.51
2	0.39	0.29	0.37
3	0.15	0.21	0.28
4	0.09	0.14	0.20
5	0.08	0.08	0.13
6	-0.06	0.02	0.07
7	-0.09	-0.03	0.01
8	-0.09	-0.09	-0.05
9	-0.14	-0.14	-0.10
10	-0.15	-0.19	-0.15
11	-0.23	-0.26	-0.21

Tabelle J10

Ergebnisse der Parallelanalyse des neu entwickelten Wissenstests (Datensatz 2.5)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	4.20	0.40	0.51
2	0.39	0.30	0.38
3	0.25	0.21	0.28
4	0.06	0.14	0.20
5	0.00	0.08	0.13
6	-0.01	0.02	0.07
7	-0.05	-0.03	0.01
8	-0.10	-0.09	-0.05
9	-0.14	-0.14	-0.10
10	-0.20	-0.19	-0.15
11	-0.22	-0.26	-0.22

Anhang K: Ergebnisse der exploratorischen Faktorenanalysen mit neuem Wissenstest

Tabelle K1

Auszug aus Parameterschätzungen der Analysen mit 2 Faktoren nach Oblimin-Rotation

	Datensatz 2.2		Datensatz 2.4		Datensatz 2.5	
Parameter	Faktor 1	Faktor 2	Faktor 1	Faktor 2	Faktor 1	Faktor 2
Faktorladungen						
Biologie	.67*	.02	.06	.62*	.55*	.14
Medizin	.82*	-.06	.07	.66*	.87*	-.08
Natur	.70*	-.02	-.05	.76*	.49*	.22
Soziale Arbeit	.58*	.01	-.03	.62*	.43*	.12
Darstellende Kunst	.04	.69*	.80*	-.09	.20	.54*
Gesundheit	.21	.47*	.58*	.12	.19	.54*
Modedesign	-.06	.72*	.66*	.01	-.12*	.80*
Pädagogik	-.03	.57*	.47*	.08	.11	.44*
Raumdesign	.12	.52*	.44*	.19	.13	.50*
Ernährung	.53*	.20	.42*	.28*	.36*	.39*
Psychologie	.52*	.22	.33*	.42*	.47*	.28*
Faktorkorrelation	.75*		.73*		.62*	

*$p < .05$.

Anhang L: Überprüfung des neu entwickelten Wissenstests auf 2-faktorielle Struktur

Tabelle L1

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 1.1)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.65	0[a]
[95%-KI]			[-0.95; -0.34]	
Wissen2	0[a]		0.40	0[a]
[95%-KI]			[0.11; 0.69]	
Korrelation				
Wissen1 – Wissen2	.96		.99	
[95%-KI]	[.90; 1.02]		[.93; 1.04]	
β-Koeffizienten (W1)				
Biologie	.68		.67	
Ernährung	.51		.29	
Medizin	.73		.73	
Natur	.65		.60	
Psychologie	-.25		.55	
Soziale Arbeit	.50		.51	
β-Koeffizienten (W2)				
Darstellende Kunst	.61		.60	
Ernährung	.21		.42	
Gesundheit	.60		.58	
Modedesign	.71		.72	
Pädagogik	.47		.47	
Psychologie	.93		.13	
Raumdesign	.58		.57	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Achsenabschnitte				
Biologie	2.07	2.41	2.48	
Darstellende Kunst	2.77	2.64	2.59	
Ernährung	2.66	2.68	2.68	
Gesundheit	2.68	2.58	2.52	
Medizin	2.25	2.91	2.75	
Modedesign	2.73	2.20	2.28	
Natur	2.43	2.55	2.75	
Pädagogik	3.43	3.20	3.22	
Psychologie	2.04	2.35	2.34	
Raumdesign	2.65	2.62	2.52	
Soziale Arbeit	2.44	2.91	2.80	
Fehlervarianzen				
Biologie	.54		.56	
Darstellende Kunst	.63		.64	
Ernährung	.49		.50	
Gesundheit	.65		.66	
Medizin	.47		.47	
Modedesign	.50		.49	
Natur	.58		.64	
Pädagogik	.78		.78	
Psychologie	.52		.54	
Raumdesign	.66		.68	
Soziale Arbeit	.75		.74	
Model-Fit				
χ^2_{df}	118.82_{107}		148.93_{116}	
p	.205		.021	
SCF	0.95		0.96	
CFI	.99		.96	
SRMR	.07		.08	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
RMSEA	.03		.05	
[90%-KI]	[.00; .06]		[.02; .07]	
χ^2-Diff-Test: χ^2_{df}; p	28.63_9; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle L2

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 1.2)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.68	0[a]
[95%-KI]			[-0.98; -0.37]	
Wissen2	0[a]		0.43	0[a]
[95%-KI]			[0.14; 0.71]	
Korrelation				
Wissen1 – Wissen2	.94		.98	
[95%-KI]	[.88; 1.00]		[.93; 1.03]	
β-Koeffizienten (W1)				
Biologie	.67		.64	
Ernährung	.47		.20	
Medizin	.72		.72	
Natur	.66		.59	
Psychologie	-.24		.56	
Soziale Arbeit	.49		.51	
β-Koeffizienten (W2)				
Darstellende Kunst	.65		.63	
Ernährung	.22		.48	
Gesundheit	.60		.58	
Modedesign	.70		.71	
Pädagogik	.50		.50	
Psychologie	.97		.16	
Raumdesign	.60		.59	
Achsenabschnitte				
Biologie	2.16	2.51	2.58	
Darstellende Kunst	2.71	2.66	2.56	
Ernährung	2.85	2.78	2.78	
Gesundheit	2.79	2.69	2.63	
Medizin	2.26	2.95	2.77	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Modedesign	2.76	2.12	2.22	
Natur	2.48	2.53	2.77	
Pädagogik	3.42	3.22	3.22	
Psychologie	2.13	2.46	2.45	
Raumdesign	2.84	2.69	2.64	
Soziale Arbeit	2.48	3.01	2.85	
Fehlervarianzen				
Biologie	.56		.59	
Darstellende Kunst	.58		.60	
Ernährung	.54		.54	
Gesundheit	.65		.67	
Medizin	.49		.48	
Modedesign	.50		.50	
Natur	.56		.66	
Pädagogik	.75		.75	
Psychologie	.45		.49	
Raumdesign	.64		.65	
Soziale Arbeit	.76		.74	
Model-Fit				
χ^2_{df}	152.16_{107}		201.06_{116}	
p	.003		< .001	
SCF	0.95		0.96	
CFI	.95		.91	
SRMR	.09		.10	
RMSEA	.06		.08	
[90%-KI]	[.04; .08]		[.06; .10]	
χ^2-Diff-Test: χ^2_{df}; *p*	46.13_9; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle L3

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 1.3)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.70	0[a]
[95%-KI]			[-1.01; -0.39]	
Wissen2	0[a]		0.44	0[a]
[95%-KI]			[0.15; 0.73]	
Korrelation				
Wissen1 – Wissen2	.95		.98	
[95%-KI]	[.89; 1.02]		[.93; 1.04]	
β-Koeffizienten (W1)				
Biologie	.68		.65	
Ernährung	.36		.20	
Medizin	.68		.70	
Natur	.67		.61	
Psychologie	-.16		.60	
Soziale Arbeit	.47		.48	
β-Koeffizienten (W2)				
Darstellende Kunst	.61		.60	
Ernährung	.31		.45	
Gesundheit	.59		.57	
Modedesign	.69		.70	
Pädagogik	.53		.53	
Psychologie	.85		.08	
Raumdesign	.56		.55	
Achsenabschnitte				
Biologie	1.99	2.33	2.41	
Darstellende Kunst	2.80	2.67	2.61	
Ernährung	2.75	2.69	2.69	
Gesundheit	2.76	2.65	2.59	
Medizin	2.28	2.99	2.79	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Modedesign	2.77	2.21	2.29	
Natur	2.51	2.68	2.88	
Pädagogik	3.43	3.18	3.18	
Psychologie	2.05	2.44	2.44	
Raumdesign	2.75	2.65	2.59	
Soziale Arbeit	2.48	2.98	2.85	
Fehlervarianzen				
Biologie	.54		.58	
Darstellende Kunst	.63		.64	
Ernährung	.57		.57	
Gesundheit	.65		.67	
Medizin	.54		.52	
Modedesign	.52		.51	
Natur	.55		.62	
Pädagogik	.72		.72	
Psychologie	.52		.53	
Raumdesign	.69		.70	
Soziale Arbeit	.78		.77	
Model-Fit				
χ^2_{df}	163.10_{107}		194.61_{116}	
p	< .001		< .001	
SCF	0.95		0.96	
CFI	.94		.91	
SRMR	.09		.10	
RMSEA	.07		.07	
[90%-KI]	[.04; .09]		[.06; .09]	
χ^2-Diff-Test: χ^2_{df}; *p*	30.22_9; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle L4

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 1.4)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.67	0[a]
[95%-KI]			[-0.98; -0.37]	
Wissen2	0[a]		0.47	0[a]
[95%-KI]			[0.17; 0.76]	
Korrelation				
Wissen1 – Wissen2	.96		.99	
[95%-KI]	[.89; 1.02]		[.93; 1.04]	
β-Koeffizienten (W1)				
Biologie	.65		.63	
Ernährung	.70		.26	
Medizin	.71		.72	
Natur	.64		.60	
Psychologie	.20		.66	
Soziale Arbeit	.52		.54	
β-Koeffizienten (W2)				
Darstellende Kunst	.61		.60	
Ernährung	.00		.43	
Gesundheit	.61		.60	
Modedesign	.70		.70	
Pädagogik	.52		.53	
Psychologie	.52		.05	
Raumdesign	.59		.57	
Achsenabschnitte				
Biologie	2.04	2.35	2.44	
Darstellende Kunst	2.80	2.65	2.59	
Ernährung	2.70	2.67	2.67	
Gesundheit	2.86	2.71	2.65	
Medizin	2.30	2.96	2.80	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Modedesign	2.81	2.23	2.31	
Natur	2.53	2.71	2.88	
Pädagogik	3.41	3.12	3.14	
Psychologie	1.93	2.36	2.35	
Raumdesign	2.74	2.63	2.56	
Soziale Arbeit	2.42	2.89	2.80	
Fehlervarianzen				
Biologie	.58		.60	
Darstellende Kunst	.62		.64	
Ernährung	.52		.53	
Gesundheit	.62		.65	
Medizin	.50		.49	
Modedesign	.51		.50	
Natur	.59		.64	
Pädagogik	.73		.73	
Psychologie	.49		.49	
Raumdesign	.65		.67	
Soziale Arbeit	.73		.71	
Model-Fit				
χ^2_{df}	139.10_{107}		164.93_{116}	
p	.020		.002	
SCF	0.97		0.97	
CFI	.97		.95	
SRMR	.08		.09	
RMSEA	.05		.06	
[90%-KI]	[.02; .07]		[.04; .08]	
χ^2-Diff-Test: χ^2_{df}; *p*	25.19_9; .003			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle L5

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 1.5)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.68	0[a]
[95%-KI]			[-0.98; -0.37]	
Wissen2	0[a]		0.40	0[a]
[95%-KI]			[0.12; 0.69]	
Korrelation				
Wissen1 – Wissen2	.90		.95	
[95%-KI]	[.83; .97]		[.89; 1.00]	
β-Koeffizienten (W1)				
Biologie	.67		.65	
Ernährung	.58		.27	
Medizin	.73		.73	
Natur	.68		.63	
Psychologie	.10		.68	
Soziale Arbeit	.50		.51	
β-Koeffizienten (W2)				
Darstellende Kunst	.64		.62	
Ernährung	.12		.41	
Gesundheit	.61		.58	
Modedesign	.71		.72	
Pädagogik	.50		.50	
Psychologie	.59		.01	
Raumdesign	.59		.57	
Achsenabschnitte				
Biologie	2.08	2.42	2.49	
Darstellende Kunst	2.72	2.60	2.54	
Ernährung	2.78	2.76	2.78	
Gesundheit	2.59	2.53	2.45	
Medizin	2.29	2.91	2.80	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Modedesign	2.76	2.17	2.27	
Natur	2.38	2.61	2.76	
Pädagogik	3.35	3.20	3.18	
Psychologie	1.90	2.40	2.36	
Raumdesign	2.71	2.63	2.56	
Soziale Arbeit	2.35	2.84	2.72	
Fehlervarianzen				
Biologie	.56		.58	
Darstellende Kunst	.59		.61	
Ernährung	.53		.54	
Gesundheit	.63		.66	
Medizin	.47		.46	
Modedesign	.49		.49	
Natur	.54		.60	
Pädagogik	.75		.76	
Psychologie	.53		.53	
Raumdesign	.65		.67	
Soziale Arbeit	.76		.74	
Model-Fit				
χ^2_{df}	159.17_{107}		195.52_{116}	
p	< .001		< .001	
SCF	0.96		0.96	
CFI	.94		.91	
SRMR	.09		.10	
RMSEA	.06		.08	
[90%-KI]	[.04; .08]		[.06; .09]	
χ^2-Diff-Test: χ^2_{df}; *p*	35.24_9; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle L6

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 2.1)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.62	0[a]
[95%-KI]			[-0.93; -0.32]	
Wissen2	0[a]		0.42	0[a]
[95%-KI]			[0.13; 0.71]	
Korrelation				
Wissen1 – Wissen2	.93		.97	
[95%-KI]	[.86; .99]		[0.91; 1.03]	
β-Koeffizienten (W1)				
Biologie	.64		.63	
Ernährung	.43		.29	
Medizin	.75		.75	
Natur	.68		.63	
Psychologie	.14		.66	
Soziale Arbeit	.53		.54	
β-Koeffizienten (W2)				
Darstellende Kunst	.65		.64	
Ernährung	.28		.42	
Gesundheit	.65		.62	
Modedesign	.68		.69	
Pädagogik	.46		.46	
Psychologie	.57		.05	
Raumdesign	.57		.55	
Achsenabschnitte				
Biologie	2.00	2.36	2.39	
Darstellende Kunst	2.77	2.65	2.58	
Ernährung	2.75	2.75	2.75	
Gesundheit	2.69	2.66	2.55	
Medizin	2.41	2.99	2.89	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Modedesign	2.99	2.25	2.37	
Natur	2.48	2.60	2.79	
Pädagogik	3.36	3.18	3.18	
Psychologie	1.97	2.38	2.36	
Raumdesign	2.75	2.71	2.62	
Soziale Arbeit	2.50	2.97	2.86	
Fehlervarianzen				
Biologie	.59		.60	
Darstellende Kunst	.58		.60	
Ernährung	.52		.52	
Gesundheit	.58		.62	
Medizin	.44		.44	
Modedesign	.53		.53	
Natur	.54		.61	
Pädagogik	.79		.79	
Psychologie	.51		.51	
Raumdesign	.67		.70	
Soziale Arbeit	.72		.71	
Model-Fit				
χ^2_{df}	130.17_{107}		176.84_{116}	
p	.063		< .001	
SCF	0.98		0.99	
CFI	.97		.93	
SRMR	.08		.09	
RMSEA	.04		.07	
[90%-KI]	[.00; .07]		[.05; .08]	
χ^2-Diff-Test: χ^2_{df}; *p*	45.79_9; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle L7

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 2.2)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.63	0[a]
[95%-KI]			[-0.93; -0.33]	
Wissen2	0[a]		0.45	0[a]
[95%-KI]			[0.16; 0.73]	
Korrelation				
Wissen1 – Wissen2	.88		.90	
[95%-KI]	[.82; .95]		[.84; .96]	
β-Koeffizienten (W1)				
Biologie	.67		.66	
Ernährung	.63		.38	
Medizin	.76		.77	
Natur	.68		.64	
Psychologie	.31		.55	
Soziale Arbeit	.57		.58	
β-Koeffizienten (W2)				
Darstellende Kunst	.70		.68	
Ernährung	.08		.32	
Gesundheit	.66		.63	
Modedesign	.68		.69	
Pädagogik	.52		.52	
Psychologie	.41		.16	
Raumdesign	.61		.59	
Achsenabschnitte				
Biologie	2.05	2.38	2.44	
Darstellende Kunst	2.85	2.64	2.61	
Ernährung	2.70	2.74	2.77	
Gesundheit	2.74	2.69	2.59	
Medizin	2.38	2.97	2.88	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Modedesign	2.99	2.27	2.38	
Natur	2.40	2.58	2.75	
Pädagogik	3.43	3.18	3.19	
Psychologie	1.98	2.31	2.27	
Raumdesign	2.68	2.61	2.52	
Soziale Arbeit	2.51	3.06	2.90	
Fehlervarianzen				
Biologie	.55		.56	
Darstellende Kunst	.52		.54	
Ernährung	.52		.54	
Gesundheit	.57		.61	
Medizin	.42		.41	
Modedesign	.54		.53	
Natur	.54		.59	
Pädagogik	.73		.73	
Psychologie	.52		.52	
Raumdesign	.63		.65	
Soziale Arbeit	.68		.66	
Model-Fit				
χ^2_{df}	123.19_{107}		169.30_{116}	
p	.136		< .001	
SCF	0.96		0.97	
CFI	.98		.95	
SRMR	.08		.09	
RMSEA	.04		.06	
[90%-KI]	[.00; .06]		[.04; .08]	
χ^2-Diff-Test: χ^2_{df}; *p*	44.67_9; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle L8

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 2.3)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.66	0[a]
[95%-KI]			[-0.97; -0.36]	
Wissen2	0[a]		0.43	0[a]
[95%-KI]			[0.14; 0.72]	
Korrelation				
Wissen1 – Wissen2	.90		.93	
[95%-KI]	[.84; .96]		[.87; .99]	
β-Koeffizienten (W1)				
Biologie	.63		.62	
Ernährung	.56		.28	
Medizin	.73		.73	
Natur	.68		.63	
Psychologie	.28		.54	
Soziale Arbeit	.57		.59	
β-Koeffizienten (W2)				
Darstellende Kunst	.66		.65	
Ernährung	.11		.38	
Gesundheit	.65		.63	
Modedesign	.69		.70	
Pädagogik	.53		.53	
Psychologie	.46		.18	
Raumdesign	.59		.57	
Achsenabschnitte				
Biologie	2.07	2.42	2.47	
Darstellende Kunst	2.78	2.59	2.55	
Ernährung	2.69	2.68	2.70	
Gesundheit	2.65	2.55	2.48	
Medizin	2.43	3.02	2.92	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Modedesign	2.79	2.18	2.27	
Natur	2.47	2.62	2.82	
Pädagogik	3.36	3.12	3.12	
Psychologie	1.98	2.30	2.27	
Raumdesign	2.61	2.54	2.46	
Soziale Arbeit	2.66	3.23	3.07	
Fehlervarianzen				
Biologie	.60		.61	
Darstellende Kunst	.57		.58	
Ernährung	.57		.58	
Gesundheit	.58		.60	
Medizin	.47		.46	
Modedesign	.53		.52	
Natur	.53		.60	
Pädagogik	.72		.72	
Psychologie	.49		.49	
Raumdesign	.66		.68	
Soziale Arbeit	.67		.65	
Model-Fit				
χ^2_{df}	146.07_{107}		180.80_{116}	
p	.007		< .001	
SCF	0.97		0.98	
CFI	.96		.93	
SRMR	.08		.09	
RMSEA	.05		.07	
[90%-KI]	[.03; .08]		[.05; .09]	
χ^2-Diff-Test: χ^2_{df}; *p*	33.89_9; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle L9

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 2.4)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.68	0[a]
[95%-KI]			[-0.98; -0.37]	
Wissen2	0[a]		0.41	0[a]
[95%-KI]			[0.13; 0.70]	
Korrelation				
Wissen1 – Wissen2	.88		.92	
[95%-KI]	[.81; .94]		[.86; .98]	
β-Koeffizienten (W1)				
Biologie	.65		.64	
Ernährung	.24		.22	
Medizin	.72		.72	
Natur	.70		.62	
Psychologie	.20		.58	
Soziale Arbeit	.57		.58	
β-Koeffizienten (W2)				
Darstellende Kunst	.68		.67	
Ernährung	.43		.44	
Gesundheit	.66		.64	
Modedesign	.70		.71	
Pädagogik	.53		.52	
Psychologie	.53		.14	
Raumdesign	.60		.59	
Achsenabschnitte				
Biologie	2.15	2.52	2.57	
Darstellende Kunst	2.83	2.65	2.61	
Ernährung	2.70	2.66	2.66	
Gesundheit	2.70	2.64	2.54	
Medizin	2.46	3.09	2.96	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Modedesign	2.84	2.16	2.27	
Natur	2.52	2.60	2.84	
Pädagogik	3.36	3.20	3.18	
Psychologie	1.97	2.35	2.32	
Raumdesign	2.66	2.56	2.50	
Soziale Arbeit	2.62	3.19	3.03	
Fehlervarianzen				
Biologie	.58		.60	
Darstellende Kunst	.53		.55	
Ernährung	.58		.57	
Gesundheit	.56		.59	
Medizin	.49		.48	
Modedesign	.51		.50	
Natur	.51		.62	
Pädagogik	.72		.73	
Psychologie	.50		.49	
Raumdesign	.64		.65	
Soziale Arbeit	.67		.67	
Model-Fit				
χ^2_{df}	131.49_{107}		180.48_{116}	
p	.054		< .001	
SCF	0.96		0.97	
CFI	.97		.93	
SRMR	.07		.08	
RMSEA	.04		.07	
[90%-KI]	[.00; .07]		[.05; .09]	
χ^2-Diff-Test: χ^2_{df}; *p*	47.68_9; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle L10

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 2.5)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.73	0[a]
[95%-KI]			[-1.03; -0.42]	
Wissen2	0[a]		0.44	0[a]
[95%-KI]			[0.16; 0.73]	
Korrelation				
Wissen1 – Wissen2	.90		.92	
[95%-KI]	[.84; .96]		[.86; .98]	
β-Koeffizienten (W1)				
Biologie	.64		.63	
Ernährung	.26		.25	
Medizin	.75		.75	
Natur	.68		.61	
Psychologie	.09		.50	
Soziale Arbeit	.51		.53	
β-Koeffizienten (W2)				
Darstellende Kunst	.69		.68	
Ernährung	.43		.43	
Gesundheit	.67		.65	
Modedesign	.68		.69	
Pädagogik	.51		.51	
Psychologie	.60		.19	
Raumdesign	.58		.57	
Achsenabschnitte				
Biologie	2.04	2.40	2.48	
Darstellende Kunst	2.80	2.63	2.58	
Ernährung	2.69	2.66	2.67	
Gesundheit	2.67	2.57	2.48	
Medizin	2.32	3.00	2.87	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Modedesign	2.98	2.23	2.35	
Natur	2.41	2.57	2.78	
Pädagogik	3.44	3.23	3.22	
Psychologie	2.06	2.38	2.35	
Raumdesign	2.67	2.59	2.51	
Soziale Arbeit	2.56	3.11	2.97	
Fehlervarianzen				
Biologie	.59		.61	
Darstellende Kunst	.52		.54	
Ernährung	.55		.55	
Gesundheit	.55		.57	
Medizin	.44		.44	
Modedesign	.53		.52	
Natur	.54		.62	
Pädagogik	.74		.74	
Psychologie	.53		.53	
Raumdesign	.66		.68	
Soziale Arbeit	.74		.72	
Model-Fit				
χ^2_{df}	137.89_{107}		187.49_{116}	
p	.024		< .001	
SCF	0.96		0.96	
CFI	.97		.92	
SRMR	.08		.09	
RMSEA	.05		.07	
[90%-KI]	[.02; .07]		[.05; .09]	
χ^2-Diff-Test: χ^2_{df}; *p*	48.20_9; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Anhang M: Parallelanalysen der Testbatterie, bestehend aus neuem Wissenstest und Wissenstests des I-S-T 2000 R

Tabelle M1

Ergebnisse der Parallelanalyse der Testbatterie (Datensatz 1)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	6.80	0.58	0.70
2	0.67	0.46	0.54
3	0.28	0.37	0.44
4	0.25	0.29	0.36
5	0.13	0.22	0.28
6	0.06	0.16	0.21
7	0.04	0.10	0.15
8	0.01	0.04	0.09
9	-0.03	-0.01	0.03
10	-0.07	-0.06	-0.02
11	-0.09	-0.11	-0.07
12	-0.12	-0.16	-0.13
13	-0.14	-0.21	-0.17
14	-0.19	-0.27	-0.23
15	-0.23	-0.33	-0.28

Tabelle M2

Ergebnisse der Parallelanalyse der Testbatterie (Datensatz 2)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	6.88	0.58	0.70
2	0.70	0.46	0.54
3	0.28	0.37	0.44
4	0.28	0.29	0.36
5	0.11	0.22	0.28
6	0.11	0.16	0.21
7	0.03	0.10	0.15
8	0.00	0.04	0.08
9	-0.02	-0.01	0.03
10	-0.03	-0.06	-0.02
11	-0.07	-0.11	-0.07
12	-0.10	-0.16	-0.13
13	-0.16	-0.21	-0.18
14	-0.20	-0.26	-0.23
15	-0.24	-0.33	-0.28

Tabelle M3

Ergebnisse der Parallelanalyse der Testbatterie (Datensatz 3)

Faktor	Rohdaten	Simulierte Daten	
	Eigenwert	Mittelwert der Eigenwerte	95%-Quantil der Eigenwerte
1	6.84	0.58	0.69
2	0.70	0.46	0.54
3	0.31	0.37	0.45
4	0.29	0.29	0.36
5	0.11	0.22	0.28
6	0.09	0.15	0.21
7	0.05	0.10	0.15
8	0.00	0.04	0.09
9	-0.03	-0.01	0.03
10	-0.05	-0.06	-0.02
11	-0.10	-0.11	-0.08
12	-0.14	-0.16	-0.13
13	-0.17	-0.21	-0.18
14	-0.17	-0.27	-0.23
15	-0.22	-0.32	-0.28

Anhang N: Überprüfung der Testbatterie, bestehend aus neu entwickeltem Wissenstest und Wissenstest des I-S-T 2000 R, auf 2-faktorielle Struktur

Tabelle N1

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 1)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.45	0[a]
[95%-KI]			[-0.75; -0.16]	
Wissen2	0[a]		0.54	0[a]
[95%-KI]			[0.24; 0.84]	
Korrelation				
Wissen1 – Wissen2	.95		.95	
[95%-KI]	[.91; .98]		[.91; .98]	
β-Koeffizienten (W1)				
I-S-T Parcel 1	.73		.74	
I-S-T Parcel 2	.81		.81	
I-S-T Parcel 3	.77		.77	
I-S-T Parcel 4	.72		.72	
Biologie	.47		.45	
Medizin	.76		.75	
Natur	.47		.34	
β-Koeffizienten (W2)				
Darstellende Kunst	.65		.64	
Ernährung	.69		.69	
Gesundheit	.67		.65	
Modedesign	.64		.66	
Natur	.26		.39	
Pädagogik	.72		.70	
Psychologie	.74		.74	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Raumdesign	.77		.78	
Soziale Arbeit	.43		.44	
Achsenabschnitte				
I-S-T Parcel 1	3.17	3.66	3.53	
I-S-T Parcel 2	3.12	3.54	3.49	
I-S-T Parcel 3	3.28	3.55	3.61	
I-S-T Parcel 4	3.71	4.07	4.04	
Biologie	2.64	2.67	2.77	
Darstellende Kunst	2.90	2.66	2.63	
Ernährung	3.20	2.89	2.88	
Gesundheit	3.30	3.14	3.08	
Medizin	2.74	2.97	3.05	
Modedesign	3.29	2.54	2.61	
Natur	2.63	2.56	2.57	
Pädagogik	3.51	3.30	3.25	
Psychologie	2.19	1.84	1.82	
Raumdesign	2.64	2.09	2.13	
Soziale Arbeit	3.53	3.22	3.24	
Fehlervarianzen				
I-S-T Parcel 1	.47		.46	
I-S-T Parcel 2	.35		.35	
I-S-T Parcel 3	.41		.41	
I-S-T Parcel 4	.48		.48	
Biologie	.78		.80	
Darstellende Kunst	.58		.59	
Ernährung	.52		.53	
Gesundheit	.55		.57	
Medizin	.43		.44	
Modedesign	.59		.56	
Natur	.49		.49	
Pädagogik	.49		.51	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Psychologie	0.45		0.46	
Raumdesign	0.41		0.40	
Soziale Arbeit	0.81		0.81	
Model-Fit				
χ^2_{df}	252.03_{208}		283.02_{221}	
p	.020		.003	
SCF	0.96		0.96	
CFI	.97		.96	
SRMR	.08		.09	
RMSEA	.05		.05	
[90%-KI]	[.02; .07]		[.03; .07]	
χ^2-Diff-Test: χ^2_{df}; *p*	30.91_{13}; .003			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; I-S-T = Wissenstest des I-S-T 2000 R; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle N2

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 2)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.46	0[a]
[95%-KI]			[-0.75; -0.16]	
Wissen2	0[a]		0.54	0[a]
[95%-KI]			[0.24; 0.84]	
Korrelation				
Wissen1 – Wissen2	.94		.94	
[95%-KI]	[.91; .97]		[.91; .97]	
β-Koeffizienten (W1)				
I-S-T Parcel 1	.78		.79	
I-S-T Parcel 2	.75		.76	
I-S-T Parcel 3	.77		.77	
I-S-T Parcel 4	.80		.80	
Biologie	.47		.45	
Medizin	.75		.74	
Natur	.41		.33	
β-Koeffizienten (W2)				
Darstellende Kunst	.65		.64	
Ernährung	.69		.69	
Gesundheit	.67		.65	
Modedesign	.65		.66	
Natur	.32		.39	
Pädagogik	.72		.70	
Psychologie	.74		.74	
Raumdesign	.77		.78	
Soziale Arbeit	.43		.44	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Achsenabschnitte				
I-S-T Parcel 1	3.22	3.61	3.58	
I-S-T Parcel 2	4.02	4.39	4.37	
I-S-T Parcel 3	3.37	3.75	3.72	
I-S-T Parcel 4	2.89	3.32	3.27	
Biologie	2.64	2.67	2.78	
Darstellende Kunst	2.90	2.66	2.63	
Ernährung	3.20	2.89	2.88	
Gesundheit	3.30	3.14	3.08	
Medizin	2.74	2.97	3.05	
Modedesign	3.29	2.54	2.61	
Natur	2.63	2.56	2.56	
Pädagogik	3.51	3.30	3.25	
Psychologie	2.19	1.84	1.82	
Raumdesign	2.64	2.09	2.13	
Soziale Arbeit	3.53	3.22	3.24	
Fehlervarianzen				
I-S-T Parcel 1	.39		.38	
I-S-T Parcel 2	.43		.43	
I-S-T Parcel 3	.41		.41	
I-S-T Parcel 4	.37		.36	
Biologie	.78		.80	
Darstellende Kunst	.58		.59	
Ernährung	.52		.53	
Gesundheit	.55		.58	
Medizin	.44		.45	
Modedesign	.58		.56	
Natur	.49		.49	
Pädagogik	.49		.51	
Psychologie	.45		.46	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Raumdesign	.41		.40	
Soziale Arbeit	.81		.80	
Model-Fit				
χ^2_{df}	273.80_{208}		301.32_{221}	
p	.002		< .001	
SCF	0.98		0.98	
CFI	.96		.95	
SRMR	.08		.09	
RMSEA	.06		.06	
[90%-KI]	[.04; .07]		[.04; .08]	
χ^2-Diff-Test: χ^2_{df}; *p*	27.51_{13}; .011			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; I-S-T = Wissenstest des I-S-T 2000 R; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Tabelle N3

Standardisierte Parameterschätzungen und Model-Fit der 2-faktoriellen Messmodelle (Datensatz 3)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Erwartungswerte				
Wissen1	0[a]		-0.44	0[a]
[95%-KI]			[-0.73; -0.14]	
Wissen2	0[a]		0.54	0[a]
[95%-KI]			[0.24; 0.84]	
Korrelation				
Wissen1 – Wissen2	.94		.95	
[95%-KI]	[.91; .98]		[.91; .98]	
β-Koeffizienten (W1)				
I-S-T Parcel 1	.72		.73	
I-S-T Parcel 2	.81		.81	
I-S-T Parcel 3	.77		.77	
I-S-T Parcel 4	.75		.75	
Biologie	.47		.45	
Medizin	.75		.75	
Natur	.45		.34	
β-Koeffizienten (W2)				
Darstellende Kunst	.65		.64	
Ernährung	.69		.69	
Gesundheit	.67		.65	
Modedesign	.65		.66	
Natur	.28		.38	
Pädagogik	.72		.70	
Psychologie	.74		.74	
Raumdesign	.77		.78	
Soziale Arbeit	.44		.44	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Achsenabschnitte				
I-S-T Parcel 1	2.87	3.47	3.23	
I-S-T Parcel 2	3.30	3.56	3.63	
I-S-T Parcel 3	3.14	3.47	3.48	
I-S-T Parcel 4	4.07	4.45	4.41	
Biologie	2.64	2.67	2.77	
Darstellende Kunst	2.90	2.66	2.63	
Ernährung	3.20	2.89	2.88	
Gesundheit	3.30	3.14	3.08	
Medizin	2.74	2.97	3.04	
Modedesign	3.29	2.54	2.61	
Natur	2.63	2.56	2.57	
Pädagogik	3.51	3.30	3.25	
Psychologie	2.19	1.84	1.82	
Raumdesign	2.64	2.09	2.13	
Soziale Arbeit	3.53	3.22	3.24	
Fehlervarianzen				
I-S-T Parcel 1	.48		.47	
I-S-T Parcel 2	.34		.35	
I-S-T Parcel 3	.40		.40	
I-S-T Parcel 4	.45		.44	
Biologie	.78		.79	
Darstellende Kunst	.58		.59	
Ernährung	.52		.53	
Gesundheit	.55		.58	
Medizin	.43		.44	
Modedesign	.58		.56	
Natur	.49		.49	
Pädagogik	.49		.51	
Psychologie	.45		.46	

(wird fortgesetzt)

	Schritt 1		Schritt 2	
	Frauen	Männer	Frauen	Männer
Raumdesign	.41		.40	
Soziale Arbeit	.81		.80	
Model-Fit				
χ^2_{df}	274.60_{208}		310.60_{221}	
p	.001		< .001	
SCF	0.98		0.98	
CFI	.96		.94	
SRMR	.08		.09	
RMSEA	.06		.06	
[90%-KI]	[.04; .07]		[.05; .08]	
χ^2-Diff-Test: χ^2_{df}; *p*	35.97_{13}; < .001			

Anmerkungen. Parameter, die für Frauen und Männer gleichgesetzt waren, werden in einer einzelnen Spalte angezeigt. KI = Konfidenzintervall; W1 = Wissen1; I-S-T = Wissenstest des I-S-T 2000 R; W2 = Wissen2; SCF = Scaling Correction Factor (siehe Muthén 1998-2004); χ^2-Diff-Test = χ^2-Difference Test: Schritt 1 vs. Schritt 2.

[a]Parameter wurde vorab fixiert.

Anhang O: Korrelationen von Wissen1 und Wissen2 mit Schulleistungen

Tabelle O1

Korrelationen von Wissen1 und Wissen2 mit der Schulleistung in Punkten [95%-KI], unterteilt nach Datensätzen und (Teil-)Stichproben

Stichprobe	Wissen1	Wissen2
	Datensatz 1	
Frauen	.61 [.49; .72]	.52 [.38; .65]
Männer	.42 [.22; .61]	.38 [.22; .54]
Gesamt	.38 [.26; .51]	.49 [.39; .60]
	Datensatz 2	
Frauen	.61 [.50; .72]	.52 [.38; .65]
Männer	.42 [.22; .61]	.37 [.21; .53]
Gesamt	.38 [.25; .51]	.49 [.39; .60]
	Datensatz 3	
Frauen	.61 [.49; .72]	.52 [.38; .65]
Männer	.42 [.22; .62]	.37 [.21; .54]
Gesamt	.39 [.26; .51]	.49 [.39; .60]

Anmerkung. KI = Konfidenzintervall.

Anhang P: Rangordnungen der Interessengebiete für Frauen und Männer

Tabelle P1

Interessengebiete, geordnet nach Rangplätzen für Frauen und Männer

Rang	Frauen	Männer
1	Musik	Musik
2	Psychologie	Sport
3	Pädagogik	Psychologie
4	Soziale Arbeit	Fremde Kulturen
5	Ernährung	Gesellschaft
6	Fremde Kulturen	Computer
7	Sport	Ernährung
8	Natur	Geschichte
9	Gesundheit	Natur
10	Gesellschaft	Philosophie
11	Raumdesign	Gesundheit
12	Darstellende Kunst	Politik
13	Literatur	Literatur
14	Computer	Darstellende Kunst
15	Medizin	Pädagogik
16	Fremdsprachen	Fremdsprachen
17	Biologie	Soziale Arbeit
18	Modedesign	Medizin
19	Geschichte	Sprachwissenschaft
20	Bildende Kunst	Recht
21	Sprachwissenschaft	Biologie
22	Philosophie	Geographie
23	Recht	Wirtschaft
24	Religion	Architektur
25	Politik	Bildende Kunst
26	Mathematik	Archäologie
27	Architektur	Religion
28	Verkehr	Physik

(wird fortgesetzt)

Rang	Frauen	Männer
29	Geographie	Mathematik
30	Archäologie	Verkehr
31	Wirtschaft	Raumdesign
32	Chemie	Maschinen
33	Physik	Elektrotechnik
34	Informatik	Informatik
35	Elektrotechnik	Chemie
36	Maschinen	Modedesign

Anhang Q: Überprüfung der Testbatterie, bestehend aus neu entwickeltem Wissenstest und Wissenstest des I-S-T 2000 R, auf 1-faktorielle Struktur

Tabelle Q1

Standardisierte Parameterschätzungen und Model-Fit des 1-faktoriellen Messmodells

	Datensatz 1	Datensatz 2	Datensatz 3
Erwartungswert			
Wissen	0.06	0.04	0.06
[95%-KI]	[-0.23; 0.34]	[-0.25; 0.32]	[-0.23; 0.34]
β-Koeffizienten			
I-S-T Parcel 1	.68	.75	.66
I-S-T Parcel 2	.76	.72	.78
I-S-T Parcel 3	.75	.73	.74
I-S-T Parcel 4	.68	.75	.71
Biologie	.46	.46	.46
Darstellende Kunst	.64	.63	.63
Ernährung	.68	.67	.68
Gesundheit	.67	.67	.66
Medizin	.75	.75	.75
Modedesign	.60	.59	.60
Natur	.72	.72	.72
Pädagogik	.71	.71	.71
Psychologie	.73	.73	.73
Raumdesign	.73	.73	.73
Soziale Arbeit	.42	.42	.43
Achsenabschnitte			
I-S-T Parcel 1	3.29	3.33	3.00
I-S-T Parcel 2	3.22	4.11	3.37
I-S-T Parcel 3	3.36	3.47	3.23
I-S-T Parcel 4	3.80	3.02	4.16
Biologie	2.64	2.65	2.64
Darstellende Kunst	2.75	2.76	2.75
Ernährung	3.00	3.01	3.00
Gesundheit	3.19	3.20	3.19

(wird fortgesetzt)

	Datensatz 1	Datensatz 2	Datensatz 3
Medizin	2.81	2.82	2.81
Modedesign	2.73	2.74	2.73
Natur	2.58	2.58	2.58
Pädagogik	3.37	3.38	3.37
Psychologie	1.97	1.98	1.97
Raumdesign	2.28	2.28	2.28
Soziale Arbeit	3.33	3.34	3.33
Fehlervarianzen			
I-S-T Parcel 1	.53	.44	.57
I-S-T Parcel 2	.43	.48	.39
I-S-T Parcel 3	.44	.47	.45
I-S-T Parcel 4	.54	.44	.49
Biologie	.79	.79	.79
Darstellende Kunst	.60	.60	.60
Ernährung	.54	.55	.54
Gesundheit	.56	.56	.56
Medizin	.44	.44	.44
Modedesign	.65	.65	.65
Natur	.49	.49	.49
Pädagogik	.50	.50	.50
Psychologie	.46	.47	.46
Raumdesign	.47	.47	.47
Soziale Arbeit	.82	.82	.82
Model-Fit			
χ^2_{224}	455.32	481.04	478.78
p	< .001	< .001	< .001
CFI	.85	.84	.83
SRMR	.10	.10	.10
RMSEA	.10	.11	.11
[90%-KI]	[.09; .11]	[.09; .12]	[.09; .12]

Anmerkungen. Sämtliche Parameter mit Ausnahme des Erwartungswertes des Faktors Wissen wurden für Frauen und Männer gleichgesetzt. Der Erwartungswert des Faktors Wissen wurde bei Männern auf 0 fixiert. In der Tabelle finden sich die Schätzungen für Frauen. KI = Konfidenzintervall.