universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

Kalibrierung eines Tests zur Angewandten Raumvorstellung unter Berücksichtigung unterschiedlicher Facetten der Raumvorstellung

Verfasser

Stefan Haberstroh

angestrebter akademischer Grad

Magister der Naturwissenschaften (Mag. rer. nat.)

Wien, am 20. April 2012

Studienkennzahl lt. Studienblatt:	A 298
Studienrichtung lt. Studienblatt:	Psychologie
Betreuer:	Univ.-Prof. Mag. Dr. Klaus D. Kubinger

Für die Betreuung der Diplomarbeit möchte ich mich besonders bei MMag. Lisbeth Weitensfelder bedanken und ebenso bei Univ.-Prof. Mag. Dr. Klaus D. Kubinger.
Des Weiteren danke ich den DirektorInnen und LehrerInnen der Schulen und natürlich ihren SchülerInnen für die Teilnahme an der Kalibrierung.

Zusammenfassung

Die Dimensionalität des Merkmals Raumvorstellung ist bis heute ungeklärt. Verschiedenste Faktoren wurden bisher identifiziert, die sich jedoch nicht klar voneinander abgrenzen und sich in ihren Definitionen ähneln. Studien zu Geschlechtsunterschieden ergeben daher ein widersprüchliches Bild, je nachdem welche Faktoren angenommen und welche Tests im Weiteren verwendet werden. Daher wurde ein Test entwickelt, der *Test zur Angewandten Raumvorstellung* (TARV; Weitensfelder (2012)), der verschiedene Facetten der Raumvorstellung beinhaltete. Diese wurden als die notwendigen mentalen Operationen zum Lösen der Items definiert, ohne eigene Faktoren darzustellen. Die Kalibrierung erfolgte zudem mit dem Ziel, den Test im Rahmen der Wiener Self-Assessments für die Studiengänge Architektur und Maschinenbau zu verwenden. Er wurde von 298 österreichischen SchülerInnen (216 männlich, 82 weiblich) zweier verschiedener Schultypen (145 AHS, 153 HTL) bearbeitet. Für die Items galt das Rasch-Modell. Das heißt, der Test maß eindimensional. Der Nachweis der Konstruktvalidität bzw. der Facetten mittels LLTM misslang. Die Geschlechter sowie SchülerInnen verschiedener Schulen unterschieden sich nicht statistisch signifikant voneinander. Die Items erwiesen sich jedoch als zu schwierig für die SchülerInnen, was den fehlenden Unterschied begünstigt haben könnte. Für die weitere Entwicklung sind daher die Itemschwierigkeit zu senken und die Facetten weiter aufzugliedern, um neben der Eindimensionalität auch Konstruktvalidität zu erreichen. Vorschläge hierfür wurden gegeben.

Abstract

The dimensionality of spatial ability is still unclear. Various factors had been identified yet, but cannot be separated clearly from each other and are similiar by definition. As a consequence, studies concerning gender-related differences in spatial abilities are inconsistent in results depending on what factor they adpot and therefore what test they use. Hence a new test was developed, the *Test of Applied Relations and Visuo-spatial abilities* (TARV; Weitensfelder (2012)), which consisted of differenet facets of spatial ability. A facet wasn't seen as a factor, instead it was defined as a necessary mental operation to solve an item. Items were also calibrated for a further use of the test as part of self-assessments for studies in Engineering and Architecture. In total 298 Austrian pupils (216 male, 82 female) of two different types of schools (145 AHS, 153 HTL) completed the test. Rasch-model conformity was given, indicating one-dimensional measurement. Evaluation of construct validity respectively the concept of different facets by means of LLTM failed. Also neither males and females nor pupils of different types of schools differed significantly from each other. However items showed to be too difficult, which could abet the missing difference. As a result, further development tends to decrease item difficulty and to subdivide the facets in order to confirm the test's construct validity in addition to its one-dimensionality. Appropriate proposals are given.

Inhaltsverzeichnis

I.
Einleitung

Ziel dieser Arbeit war die Kalibrierung eines neuen Raumvorstellungstests, des *Tests zur Angewandten Raumvorstellung* (im Weiteren als *TARV* bezeichnet) (Weitensfelder, Grubesic, Kubinger & Gittler, 2010), der bereits in Kurzversionen im Rahmen der Wiener Self-Assessments[1] für die Studiengänge Architektur und Maschinenbau angeboten wird. Der TARV unterschied sich in seinem Testkonzept wesentlich von gängigen psychologisch-diagnostischen Verfahren, sodass zuerst eine Auseinandersetzung mit der umfangreichen Forschung zur Raumvorstellung erfolgte.

Diese lässt sich gemäß Linn und Petersen (1985) in vier Richtungen einteilen. Die differentielle Richtung, die Unterschiede in der Raumvorstellung in verschiedenen Populationen (z. B. Männer und Frauen) erfasst, die psychometrische Richtung, der es darum geht, verschiedene Faktoren der Raumvorstellung zu identifizieren, die kognitive Richtung, die kognitive Prozesse bei der Bearbeitung von Raumvorstellungsaufgaben untersucht sowie die strategische Richtung, die sich auf die Identifikation verschiedener Bearbeitungsstrategien konzentriert. Alle vier Richtungen befinden sich in einem Spannungsfeld, in dem sie einander sowohl ergänzen als auch widersprechen. Als relevant erwiesen sich die ersten beiden Richtungen, die differentielle wie psychometrische, weswegen eine Vertiefung in sie erfolgte. Einen Überblick über die strategische sowie kognitive Richtung, aber auch über die anderen Richtungen, findet sich beispielsweise bei Shah und Miyake (2005).

Zu Beginn der Arbeit wird erörtert, ob Raumvorstellung innerhalb der psychometrischen Forschungsrichtung ein ein- oder multidimensionales Merkmal ist, welches sich aus verschiedenen Konstrukten (Fähigkeiten) zusammensetzt. Im weiteren Verlauf werden die Ergebnisse der differentiellen Forschungsrichtung zu Unterschieden in der Raumvorstellungsleistung zwischen den Geschlechtern und zwischen Personengruppen mit verschiedenen Ausbildungen dargestellt. Auch wird auf Test- und Itemcharakteristika eingegangen,

[1] „Zumeist beabsichtigen (psychologische) *Self-Assessments* (...): Über die Stärken und Schwächen des Bewerbers in Bezug auf die angestrebte (Ausbildungs-)Stelle soll informiert werden. Keinesfalls wird eine Selektionsentscheidung getroffen.“ (Kubinger, 2009, S. 156)

die den Geschlechtsunterschied beeinflussen können. Darauf aufbauend wird das Testkonzept des TARV und seine Konstruktion geschildert und in Bezug zu den Erkenntnissen aus der Literatur gestellt. Im Weiteren erfolgt die Beschreibung des methodischen Vorgehens. Dies inkludiert die Darstellung des Testdesigns, der Stichprobe sowie der für die Kalibrierung notwendigen speziellen theoriebildenden Verfahren, nämlich das Dichotom Logistische Test-Modell von G. Rasch (1960) sowie das Linear Logistische Test-Modell von Fischer (1973).

Die Gestaltung dieser Arbeit folgte zudem zwei Bestrebungen des Autors. Zum einen stelle sich schnell heraus, dass Raumvorstellung aufgrund jahrzehntelanger Forschung ein äußerst vielfältiges Themengebiet mit einander widersprechenden Erkenntnissen ist. In dieser Arbeit wurde versucht, diesem Umstand gerecht zu werden. Aus diesem Grund finden sich zu beiden dargestellten Forschungsrichtungen kritische Auseinandersetzungen. Insbesondere die psychometrische Forschungsrichtung, der es um die Anzahl und Art vom Raumvorstellungsfaktoren geht, stellte einen theoretischen Schwerpunkt der Arbeit dar. Hier wurde bewusst versucht, möglichst viele Studien zu berücksichtigen, um die Kontroversen in dieser Richtung zu verdeutlichen und daraus Erkenntnisse für den TARV zu gewinnen.

Die zweite Bestrebung dieser Arbeit betraf die Verständlichkeit der Methodik. So konnte kaum Literatur gefunden werden, die sich mit der Itemkalibrierung mit einem unvollständigen Testdesign bei kleineren Stichproben befasste. Dies stellte eine Herausforderung in dieser Arbeit dar, da durch den Umfang der Stichprobe, der Anzahl an Items und der begrenzten Bearbeitungszeit nur wenige Kombinationen an Anzahl von Testheften und Items je Testheft möglich waren. Das methodische Vorgehen wurde daher versucht, verständlich genug zu schildern, sodass es womöglich auch einem/einer zukünftigen LeserIn helfen kann.

II.
Theoretischer Teil

Vor der theoretischen Auseinandersetzung zum Thema Raumvorstellung ist es notwendig, einige Begriffe vorab zu präzisieren:

Grundsätzliches Problem ist, dass keine allgemein akzeptierte Definition zur Raumvorstellung existiert (Eliot & Macfarlane Smith, 1983). Schon der Begriff *Raumvorstellung* kann ebenso „als räumliches Vorstellungsvermögen, (...) als Fähigkeit zur Vorstellung räumlicher Relationen, „Raum-Lage-Orientierung“ oder, gleich englisch, als *Spatial Ability* bezeichnet [werden]“ (Kubinger, 2009, S. 199f.). In dieser Arbeit wird der Begriff Raumvorstellung verwendet. Unterscheiden sich Personen in ihrer Raumvorstellung, meint dies den Unterschied, der sich in entsprechenden „psychologischen Test[s]“ (Kubinger, Rasch & Yanagida, 2011, S. 29) bzw. „psychologisch-diagnostisches Verfahren“ (Kubinger, 2009, S. 197) ergibt.

Personen, die einen Test bearbeiten, werden als *Testpersonen* oder kurz *Tpn* bezeichnet. Ein Test besteht aus verschiedenen *Items* bzw. *Testaufgaben*, das sind die „Bestandteile[n] eines Tests, die eine Reaktion oder Antwort hervorrufen sollen“ (Rost, 2004, S. 18). Die Leistung, die eine Testperson bei allen Items erzielt, ist ihr *Testwert*.

Die Begriffe *Faktor* sowie *Dimension* bedeuten latente Variablen, die bestimme Konstrukte repräsentieren. Die Konstrukte spiegeln Merkmale wider (hier: das Merkmal *Raumvorstellung*), die durch die jeweiligen Tests erst erfasst werden. Dabei wird unterschieden, ob es sich aus einem oder mehreren Konstrukten zusammensetzt bzw. es sich um ein ein- oder multidimensionales Merkmal handelt (Jonkisz, Moosbrugger & Brandt, 2012). „Mehrere Faktoren (oder Dimensionen) der Raumvorstellung“ bedeutet daher z. B., dass sich das Merkmal Raumvorstellung aus verschiedenen Konstrukten zusammensetzt.

1. Raumvorstellung: Psychometrische Forschungsrichtung

Raumvorstellungstests können laut Hegarty, Montello, Richardson, Ishikawa und Lovelace (2006) gemäß der Konstruktion ihrer Items grob klassifiziert werden in „small-scale“ (S. 151) und „larger-scale“ bzw. „environmental“ (S. 152) Tests. Während es bei den Ersteren darum geht, kleinere Objekte oder Strukturen wahrzunehmen und gedanklich zu bearbeiten, so ist das Ziel bei den Letzteren, sich in einer vollständigen Umgebung (z. B. einem Stadtplan) zurechtzufinden (Hegarty et al., 2006). In dieser Arbeit wurden ausschließlich „small-scale“ Tests berücksichtigt.

1.1. Das Merkmal Raumvorstellung

Bereits zu Beginn des 20. Jahrhunderts beschäftigten sich ForscherInnen mit Raumvorstellung und versuchten, einen eigenen Faktor Raumvorstellung und damit ein solches Merkmal als festen Bestandteil der Intelligenz zu identifizieren. Exemplarisch sei daher kurz auf die Arbeiten von (in zeitlicher Reihenfolge) Thorndike (1921), McFarlane (1925), Kelley (1928), Cox (1928), El Koussy (1935), Alexander (1935) und Thurstone (1938/1969) eingegangen, die jede für sich einen solchen Faktor identifizieren konnten. Eine ausführliche Betrachtung zu den Anfängen der Raumvorstellungsforschung findet sich bei Macfarlane Smith (1964) sowie Eliot und Macfarlane Smith (1983).

Trotz ähnlichem Resultat der Studien unterschieden sie sich insbesondere in ihren Forschungsfragen. So stand bei El Koussy (1935) die dezidierte Suche nach einem Faktor Raumvorstellung im Vordergrund, während sich andere (Cox, 1928; McFarlane, 1925) mit dem Konzept einer dem Menschen inhärenten mechanischen oder praktischen Begabung beschäftigten. Anderen ForscherInnen ging es hingegen generell um die Auseinandersetzung mit und die „study of the nature and scope of mental traits“ (Kelley, 1928, S. 1). Gleich war ihnen allerdings, dass sie im Rahmen ihrer Studien die Testpersonen Items bearbeiten ließen, durch die sich in der statistischen Auswertung der Daten ein Faktor Raumvorstellung herauskristallisierte. Tabelle 1 gibt einen Überblick über die in den Studien ermittelten Faktoren mitsamt ihrer Beschreibung.

Tabelle 1
Termini eines Raumvorstellungsfaktors in Studien bis 1938

	Benennung	Definition
Thorndike (1921)	Spatial Relations as content	-
McFarlane (1925)	Practical Ability	„involves analysis and synthesis, judgment and conception (...) about concrete spatial situations“ (S. 56)
Kelley (1928)	Spatial 1	„sensing and retention of geometric forms“ (S. 148)
	Spatial 2	„manipulation of spatial relationships“ (S. 148f.)
Cox (1928)	M	„apprehending of (...) spacially arranged items, (...) education of their space relations (...) [and] to deal mentally with mechanical movements“ (S. 161, 185)
El Koussy (1935)	K	„ability to obtain and the facility for utilising visual spatial imagery“ (S. 84)
Alexander (1935)	F, Practical factor	„ability measured by performance tests, (...) peculiar to the sphere of things“ (S. 123)
Thurstone (1938/1969)	S	„facility in spatial and visual imagery“ (S. 80)

Die Definition des Faktors erfolgte zumeist operational. Das heißt, es wurden die gedanklichen Schritte beschrieben, die zum Lösen der jeweiligen Items identifiziert wurden. Dies rührte nicht zuletzt daher, da zum Teil theoriebildend geforscht wurde und ein Konzept der Raumvorstellung als Ausgangspunkt nicht vorhanden war. Auch blieb beispielsweise Thorndike (1921) eine klare Definition schuldig und verwies lediglich auf den gefunden statistischen Zusammenhang inhaltlich zusammengehöriger Tests. Eine Besonderheit stellte die Studie von Kelley (1928) dar, da in dieser bereits die Möglichkeit zweier voneinander abgrenzbarer Raumvorstellungsfaktoren angedeutet, und folglich Raumvorstellung

als ein multidimensionales Merkmal betrachtet wurde. Damit nahm Kelley die zukünftige Forschung, die darauf abzielte, verschiedenste Raumvorstellungsfaktoren zu identifizieren, vorweg. Zuletzt sei noch auf die Studien von El Koussy (1935) und Thurstone (1938/1969) hingewiesen, da sie ihre Daten mithilfe einer Faktorenanalyse analysierten. Diese statistische Auswertungsmethode erlaubt es, die Anzahl und Art von Faktoren zu identifizieren, „die zur statistischen Erklärung einer größeren Anzahl korrelierender Variablen ausreichen“ (Kubinger, 2009, S. 58) und zu deren Entwicklung Thurstone maßgeblich beitrug. Sie öffnete schließlich der Suche nach weiteren Faktoren Tür und Tor.
Diesen ersten Studien gelang es damit, ein eigenes, wenngleich auch unterschiedlich bezeichnetes und definiertes, Merkmal Raumvorstellung als Bestandteil der Intelligenz zu etablieren.

1.2. Raumvorstellung als ein multidimensionales Merkmal

Die spätere Forschung verfolgte das Ziel, verschiedene Faktoren des Merkmals Raumvorstellung zu identifizieren, oder wie Michael, Guilford, Fruchter und Zimmerman (1957) es formulierten: „The question as to whether a single unitary spatial ability or several different spatial abilities are necessary to describe the intellectual processes of spatial thinking, has probably not been answered to the satisfaction“ (S. 185). Da der TARV von diesen Studien beeinflusst ist, erfolgt ihre Beschreibung ausführlicher. Insgesamt wurde eine Vielzahl an Faktoren formuliert, weswegen im Folgenden nur die für den TARV relevanten beschrieben werden. Eine ausführliche Abhandlung aller Faktoren findet sich bei Carroll (1993), Hegarty und Waller (2005) und Lohman (1988).
Im Laufe der Zeit kristallisierten sich mehrere Faktoren heraus, die Raumvorstellung als ein multidimensionales Merkmal konstituierten. Drei für den TARV relevante Faktoren wurden dabei in verschiedenen Studien immer wieder repliziert. Diese sind (nach der Bezeichnung von Lohman (1979)): *Spatial Relations*, *Spatial Orientation* und *Visualization*. Tabelle A liefert einen Überblick über die verschiedenen Faktoren, die in ausgewählten Studien formuliert wurden. Dabei wiesen die einzelnen Studien auf bestimmte

Tabelle 2
Gruppierung der Raumvorstellungsfaktoren in Studien ab 1947 nach marker tests

	Spatial Relations	Spatial Orientation and Perception	Visualization	Kinesthetic factor	Length Estimation
Guilford und Lacey (1947)		Spatial Relations	Visualization	S_2	Length Estimation
Thurstone (1949, 1950)	S_1	S_3	S_2	K	
French (1951, Michael et al. (1957))	Space, Spatial Orientation		Visualization		Length Estimation
Zimmerman (1953)	Spatial Relations		Visualization		
Michael et al. (1957)	Spatial Relations and Orientation		Visualization	Kinesthetic Imagery	
Ekstrom, French, Harman und Dermen (1976)	Spatial Orientation		Visualization		
McGee (1979)	Spatial Orientation		Spatial Visualization		
Lohman (1979)	Spatial Relations	Spatial Orientation	Visualization	K	
Linn und Petersen (1985)	Mental Rotation	Spatial Perception	Spatial Visualization		
Lohman (1988)	Speeded Rotation	Spatial Orientation	General Visualization	Kinesthetic factor	
Carroll (1993)	Spatial Relations	Visualization			Length Estimation
Colom, Contreras, Botella und Santacreu (2001)	General Visualization				
Hegarty und Waller (2004)	Spatial Visualization	Spatial Orientation	[a]		
W. Johnson und Bouchard Jr. (2005)		Spatial factor	Image Rotation factor		

Anmerkung. Wenn aufgrund verschiedener Tests eine klare Zuordnung nicht möglich war, so wurde der Faktor ausgewählt, der von der Mehrzahl der Tests erfasst wurde.
[a]Es wurden nur Tests zum Erfassen des Faktors Spatial Relations verwendet. Ein eigener Faktor Visualization wurde von den AutorInnen jedoch nicht ausgeschlossen.

Tests hin, sogenannte „marker tests“ (Colom et al., 2001, S. 904), die den jeweiligen Faktor bestmöglich erfassten. In Tabelle 2 wurden die Faktoren nach diesen Tests gruppiert. Zudem gibt diese Tabelle Hinweise auf die widersprüchlichen Forschungsergebnisse der vergangenen Jahrzehnte. Folgende inhaltliche Übersicht über die Faktoren stützte sich dabei auf die Ergebnisse dieser Gruppierung und fasst die Definitionen von Tabelle A (siehe *Anhang*) zusammen:

Spatial Relations bedeutet, ein zwei- oder dreidimensionales Objekt gedanklich zu rotieren. Items, die auf diesen Faktor laden, zeichnet vor allem ihre geringe Komplexität aus. Das heißt, es müssen *nicht* mehrere Schritte gedanklich vollzogen werden, wie z. B. das vorherige Zerlegen des Objekts in verschiedene Teile. Im Kern geht es darum, zu überprüfen, ob ein Objekt durch einfache Rotation einem anderen Objekt entspricht. Tests dieses Faktors sind beispielsweise der Mental Rotations Test (Vandenberg & Kuse, 1978, Abbildung 1) mit dreidimensionalen Objekten sowie

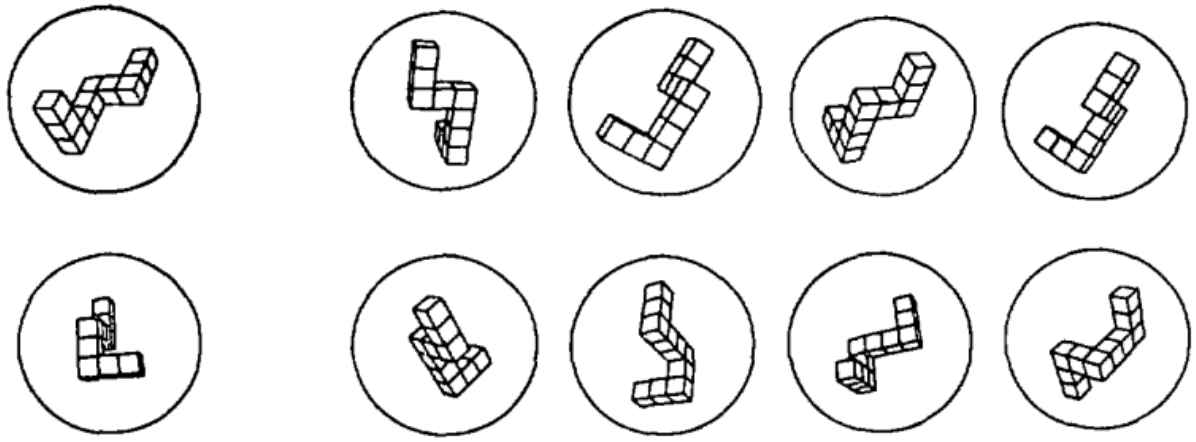


Abbildung 1. Beispielitem: Mental Rotations Test (Vandenberg & Kuse, 1978).
Testpersonen müssen erkennen, welche zwei der vier rechts dargestellten Objekte eine rotierte Darstellung des linken Objekts sind. Für die obere Reihe sind es das erste und vierte Objekt, für die untere Reihe das zweite und dritte. (Beispielitem aus Vandenberg & Kuse, 1978.)

der Card Rotations Test mit zweidimensionalen Objekten (Ekstrom et al., 1976, Abbildung 2). Uneinigkeit bei diesem Faktor gibt es allerdings hinsichtlich der Notwendigkeit einer Zeitbeschränkung, also einer Speed-Komponente für die Bearbeitung der einzelnen Items. Es bleibt daher ungeklärt, ob es nur um die Fähigkeit

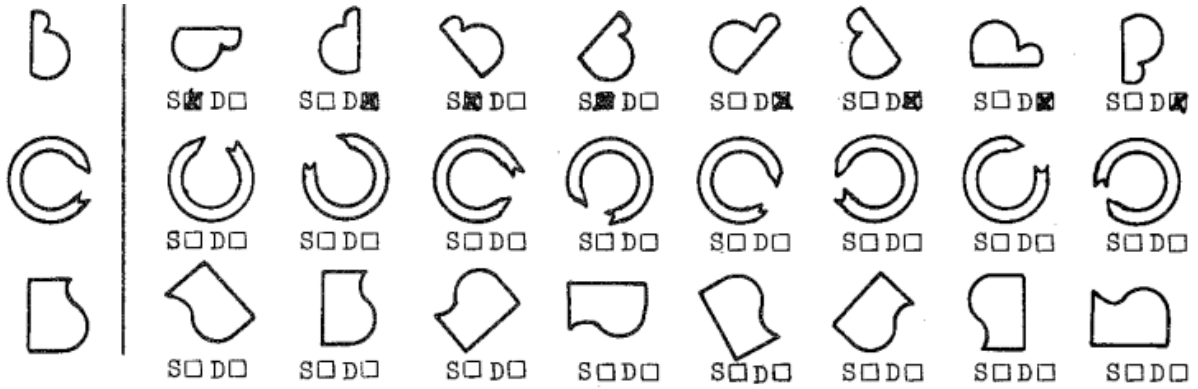


Abbildung 2. Beispielitems: Card Rotations Test (Ekstrom et al., 1976).
Testpersonen müssen erkennen, welche der sechs rechts dargestellten Objekte eine zweidimensional rotierte Darstellung des linken Objekts sind. Für die obere Reihe sind die Lösungen gekennzeichnet. *S* bedeutet „same“ und *D* „different“. (Beispielitem aus Ekstrom et al., 1976.)

geht, mental rotieren zu können (z. B. W. Johnson & Bouchard Jr., 2005) oder ob im Vordergrund steht, Items möglichst schnell durch Rotation (z. B. Lohman, 1988) zu lösen.

Spatial Orientation and Perception bedeutet, bei der Wahrnehmung eines Objekt oder einer Zusammenstellung mehrerer Objekte den eigenen Standpunkt zu berücksichtigen bzw. ändern zu können. Zum einen geht es darum, sich ein Objekt aus einer anderer Perspektive vorzustellen. Wesentlich hier ist, seinen eigenen Standpunkt zu variieren und *nicht* die Objekte gedanklich entsprechend zu rotieren oder manipulieren. Dies ist der zentrale Unterschied zum Faktor Spatial Relations. Erfasst wird dies beispielsweise mit dem Spatial Orientation Test (Guilford & Lacey, 1947, Abbildung 3) sowie dem Test Schlauchfiguren (Stumpf & Fay, 1983, Abbildung 4). Dieser ist allerdings als strittig zu betrachten, da er keine Maßnahmen trifft, um das Lösen der Items durch Rotation zu vermeiden. Beide Lösungswege, Perspektivenwechsel wie Rotation oder eine Kombination beider, liegen somit für Testpersonen auf der Hand. Während Eliot und Macfarlane Smith (1983) ihn noch zu den „perspective tasks“ (S. 370) zuordneten, die beinhalteten, andere Standpunkte neben dem eigenen zu berücksichtigen, so wiesen ihn Stumpf und Klieme

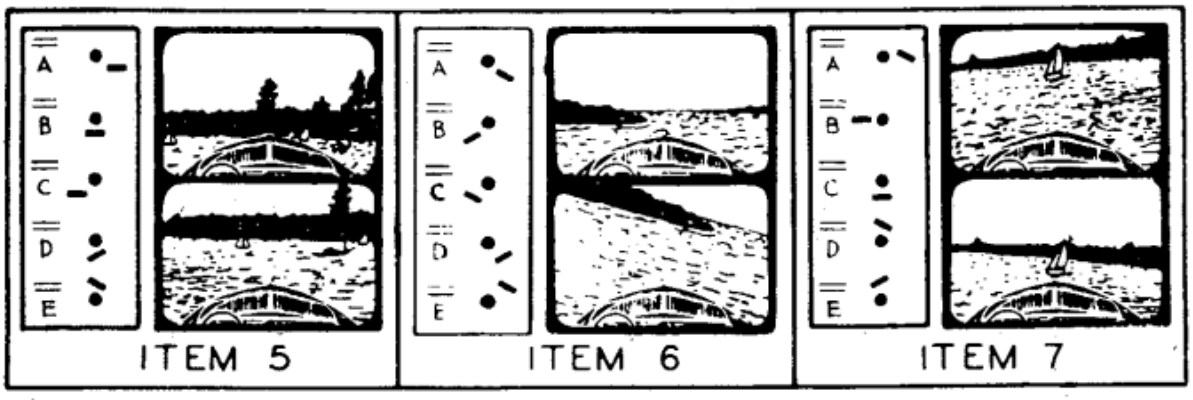


Abbildung 3. Beispielitem: Spatial Orientation Test (Guilford & Lacey, 1947). Testpersonen müssen erkennen, in welche Richtung sich der Bug des dargestellten Bootes bewegt hat. Der Punkt bei den Antwortmöglichkeiten links ist dabei die Spitze des Bugs in der oberen Darstellung. Der schwarze Balken gibt die Richtung wieder, in die sich das Boot vom oberen zum unteren Bild in Bezug zum Punkt bewegt hat. Die Lösungen sind C für Item 5, B für Item 6 und E für Item 7. (Beispielitem aus Eliot & Macfarlane Smith, 1983.)

(1989) später als Test für den Faktor Spatial Visualization aus und andere wiederum als Test für den Faktor Spatial Relations (Peters et al., 1995).

Zum anderen bedeutet dieser Faktor, die Positionen von Objekten und deren räumliche Beziehung zueinander in Bezug zur eigenen Position bestimmen zu können, demzufolge also die Orientierung zu behalten. Testbeispiele hierfür sind der Water Level Task (Piaget & Inhelder, 1967), der Object Perspective Taking Test (Kozhevnikov & Hegarty, 2001) sowie der Pictures Test (Hegarty & Waller, 2004). Insbesondere McGee (1979) betonte, dass das Lösen von Tests beider Bereiche, Spatial Orientation sowie Spatial Perception, die gleiche Fähigkeit benötigte.

Visualization bedeutet, mehrere gedankliche Operationen bei komplexen Objekten oder Zusammenstellungen solcher Objekte anzuwenden (z. B. rotieren, Standpunkt verändern, in Einzelteile zerlegen und zusammensetzen, die Muster eines Objekts identifizieren). Im Gegensatz zum Faktor Spatial Relations steht hier bei Items ihre Komplexität im Vordergrund. Es sind mehrere Schritte notwendig, um zur Lösung zu gelangen (z. B. ob zwei Objekte einander entsprechen). Ebenso ist die Bearbeitungszeit nicht limitiert. Tests dieses Faktors sind zum Beispiel der Dreidimensio-

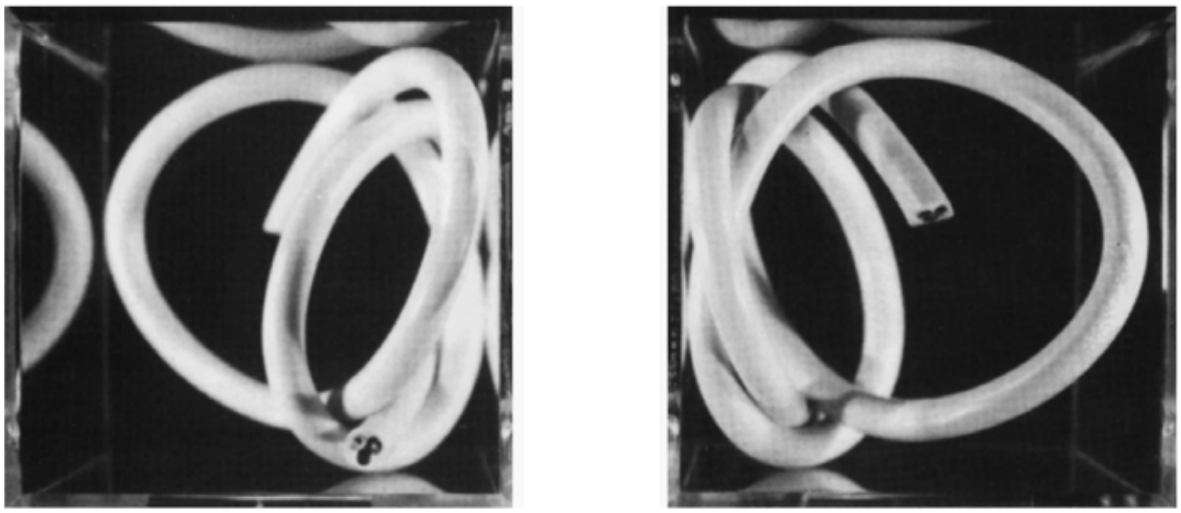


Abbildung 4. Beispielitem: Schlauchfiguren (Stumpf & Fay, 1983).
Testpersonen müssen erkennen, „von welcher Position der einen Ansicht des Würfels aus die zweite entstehen würde: Von rechts, von links, von unten, von oben oder von hinten“ (Kubinger, 2009, S. 200). Die Lösung ist „von hinten“. (Beispielitem aus Eliot & Macfarlane Smith, 1983.)

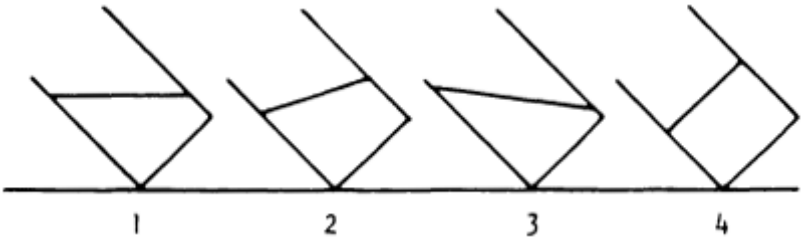


Abbildung 5. Beispielitem: Water Level Task (Piaget & Inhelder, 1967).
Testpersonen müssen erkennen, in welchem Glas der Wasserstand korrekt dargestellt ist oder sie müssen den korrekten Wasserstand selbst einzeichnen. Die Lösung ist Glas Nummer 1. Die Linie des Wassers verläuft stets horizontal. (Beispielitem aus Linn & Petersen, 1985.)

nale Würfeltest (3DW) (Gittler, 1990) sowie nach Ekstrom et al. (1976) der Paper Folding Test (Abbildung 9), Surface Development Test (Abbildung 10) und Form Board Test (Abbildung 11).

Neben diesen drei oftmals replizierten Faktoren existieren noch eine Vielzahl weiterer. Zwei für den TARV relevante sind *Length Estimation* (Carroll, 1993) und der *Kinesthetic factor* (Lohman, 1988):

Kinesthetic factor bedeutet, vom eigenen Standpunkt aus zwischen links und rechts

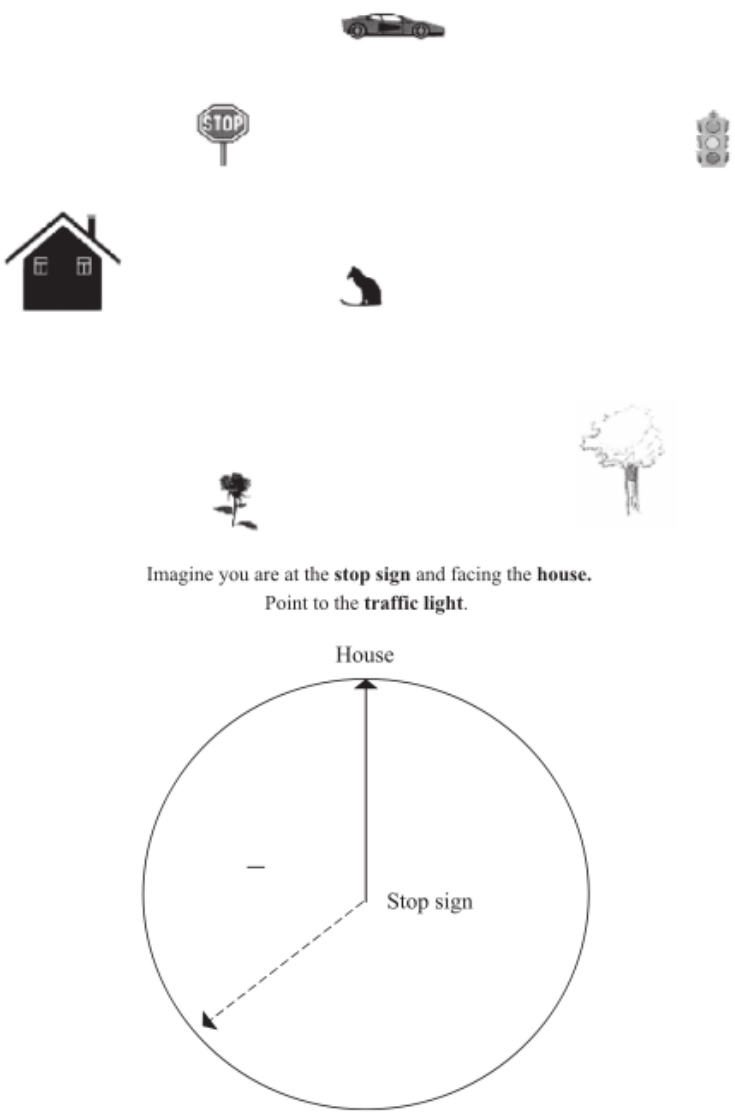


Abbildung 6. Beispielitem: Object Perspective Taking Test (Kozhevnikov & Hegarty, 2001). Testpersonen müssen mit einer gestrichelten Linie (hier bereits eingezeichnet) kennzeichnen, wo sich ein entsprechendes Objekt von ihrem Standpunkt und ihrer vorgegebenen Perspektive aus befindet. (Beispielitem aus Hegarty & Waller, 2004.)

zu unterscheiden, demnach im weitesten Sinne auch Links-Rechts-Vertauschungen von Objekten zu erkennen. Diskordanz herrscht auch hier wiederum hinsichtlich der Notwendigkeit einer vorgegebenen Bearbeitungszeit. Testbeispiele sind der Hands Test (Thurstone, 1938/1969, Abbildung 12) sowie der Right-Left Discrimination Test (Ofte & Hugdahl, 2002, Abbildung 13).

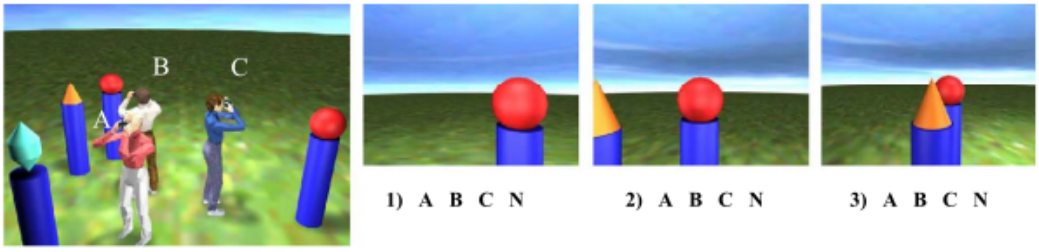


Abbildung 7. Beispielitem: Pictures Test (Hegarty & Waller, 2004).
Testpersonen müssen angeben, welche der drei FotografInnen (A, B, C) das jeweilige Foto geschossen hat. Auch kann es keiner von ihnen gewesen sein (N). Die Lösungen sind: 1) C, 2) B und 3) N. (Beispielitem aus Hegarty & Waller, 2004.)

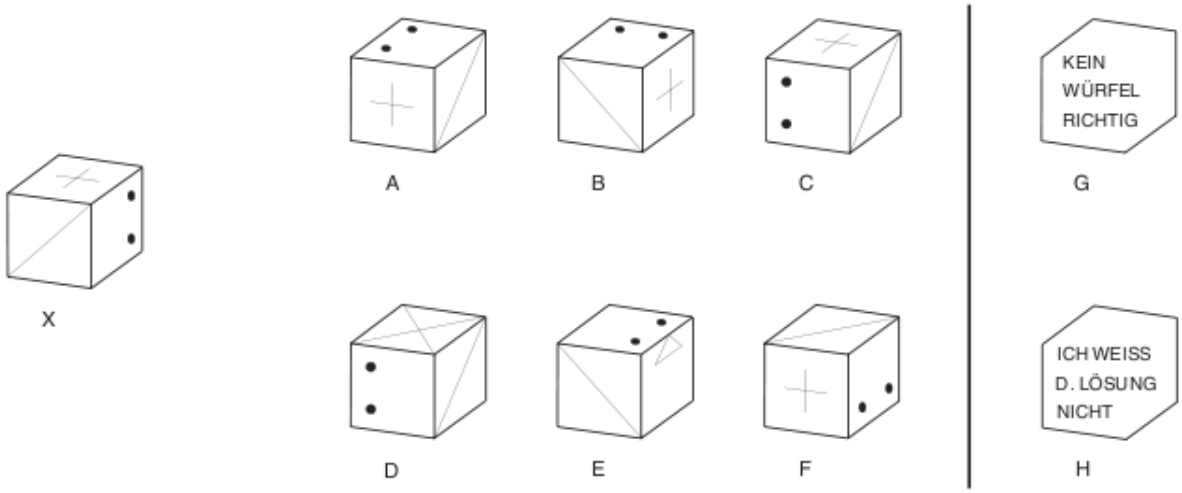


Abbildung 8. Beispielitem: Dreidimensionaler Würfeltest (Gittler, 1990).
„Jeder einzelne Würfel hat sechs verschiedene Muster. Prüfen Sie, ob einer der Würfel A-F den Würfel X in veränderter Lage darstellen kann, oder ob die Antwort G zutrifft“ (Gittler, 1990). Die Lösung ist D. (Beispielitem aus Gittler & Glück, 1998.)

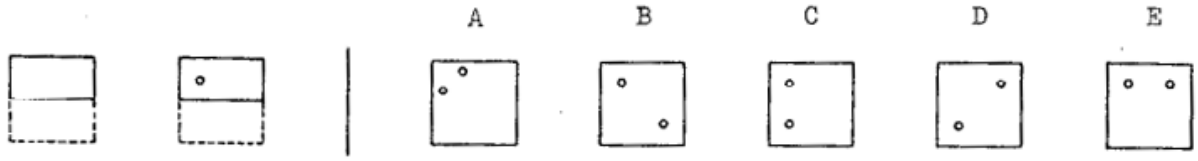


Abbildung 9. Beispielitem: Paper Folding Test (Ekstrom et al., 1976).
Testpersonen müssen erkennen, welches der fünf rechts dargestellten Papiere dem linken entspricht, wenn dieses, wie angegeben, gefaltet und einmal durchbohrt wird (Hegarty & Waller, 2005). Die Lösung ist Papier C. (Beispielitem aus Ekstrom et al., 1976.)

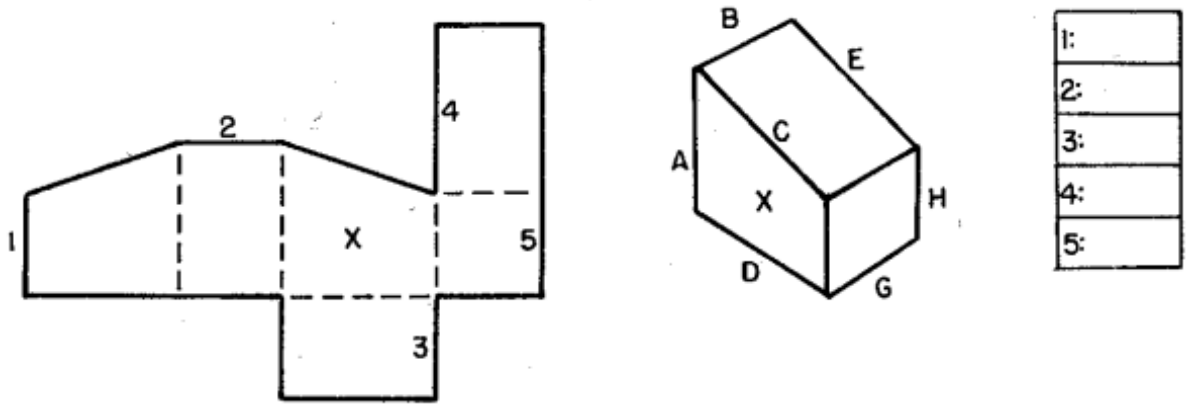


Abbildung 10. Beispielitem: Surface Development Test (Ekstrom et al., 1976).
Das linke Papier kann zum rechten Objekt gefaltet werden. Die Knicklinien sind dabei gestrichelt eingezeichnet. X markiert bei beiden Darstellungen die gleiche Seite und dient als Orientierung. Die Testpersonen müssen erkennen, welche nummerierten Papierränder oder gestrichelten Linien in der linken Darstellung den mit Buchstaben gekennzeichneten Rändern beim rechten Objekt entsprechen (Ekstrom et al., 1976). Die richtige Zuordnung ist 1:H, 2:B, 3:G, 4:C, 5:H. (Beispielitem aus Ekstrom et al., 1976.)

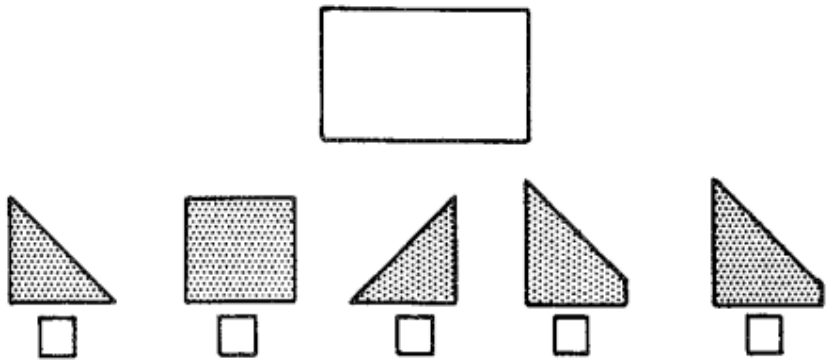


Abbildung 11. Beispielitem: Form Board Test (Ekstrom et al., 1976).
Testpersonen müssen erkennen, welche der dargestellten Einzelteile so zusammengelegt werden können, um die darüber gezeichneten Form auszufüllen. Die Lösung ist (von links nach rechts) Einzelteil eins, drei, vier und fünf. (Beispielitem aus Ekstrom et al., 1976.)

Length Estimation bedeutet, die Länge eines Objekts oder den Abstand zwischen Objekten richtig einzuschätzen. Der Faktor geriet bereits kurz nach seiner Entdeckung ins Abseits der Forschung, so dass aktuell keine Tests existieren, die ihn erfassen würden. Ein heute nicht mehr publizierter Test wäre der Estimation of Length Test (Guilford & Lacey, 1947; French, Ekstrom & Price, 1963). Die Aufgabe für die Testpersonen bestand bei diesem Test darin, Linien zu identifizieren, die gleich oder

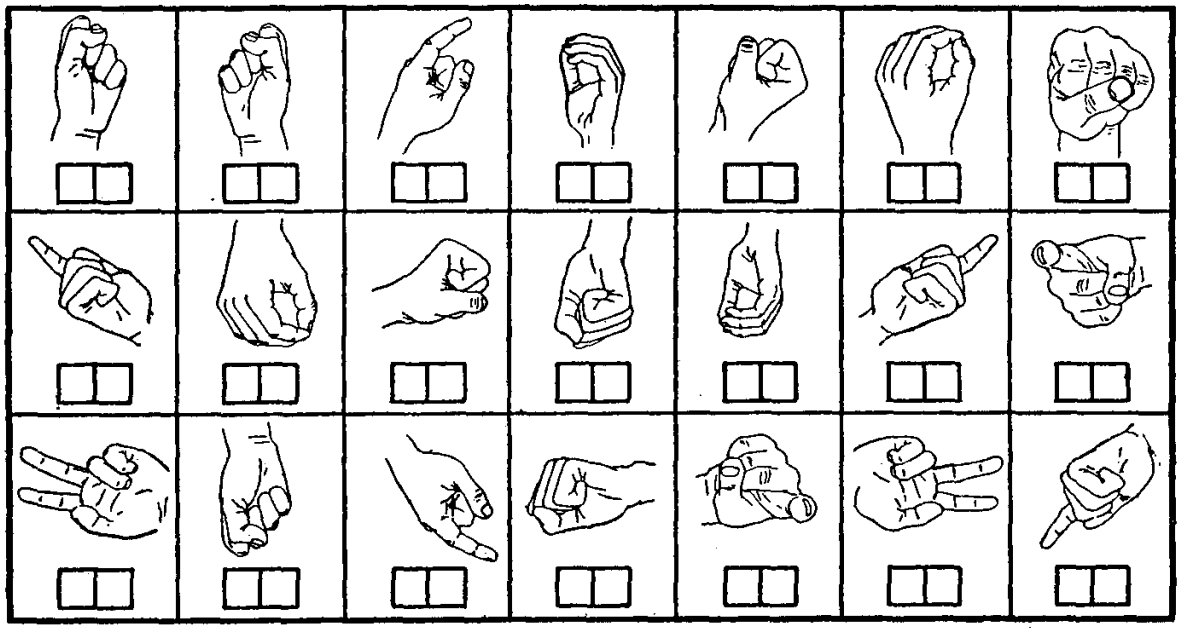


Abbildung 12. Beispielitem: Hands Test (Thurstone, 1938/1969).
Testpersonen müssen erkennen, ob eine rechte oder linke Hand dargestellt ist und dementsprechend das linke oder rechte Kästchen ankreuzen. Die erste Darstellung links oben wäre beispielsweise eine linke Hand. Das Kreuzchen wäre im linken Kästchen zu setzen. (Beispielitem aus Thurstone, 1938/1969.)

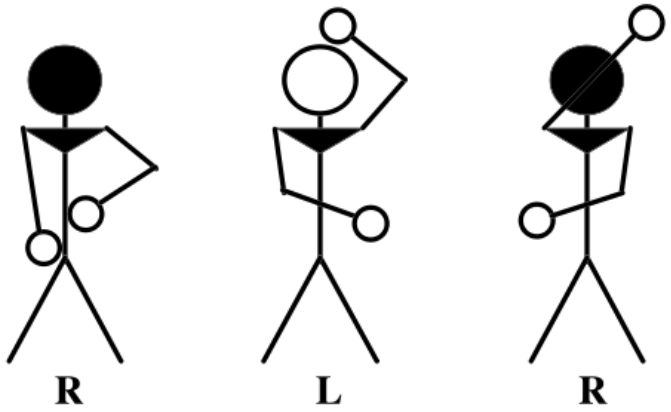


Abbildung 13. Beispielitem: Right-Left Discrimination Test (Ofte & Hugdahl, 2002).
Testpersonen müssen die rechte oder linke Hand der dargestellten Figur markieren, je nachdem ob ein R (für rechts) oder ein L (für links) unter der Figur steht. Ein weißer Kopf der Figur bedeutet, dass die Figur von vorne gesehen wird und vica versa. Die korrekten Markierungen (von links nach rechts) sind: rechter Kreis, oberer Kreis, linker Kreis. (Beispielitem aus Ofte & Hugdahl, 2002.)

doppelt so lang waren wie eine andere Linie.

1.2.1. Kritische Auseinandersetzung hinsichtlich mehrerer Faktoren der Raumvorstellung

Bisher ist es nicht vollkommen gelungen, verschiedene Raumvorstellungsfaktoren inhaltlich wie statistisch klar voneinander zu separieren, und damit bleibt die Frage, ob und wenn ja, wie viele verschiedene Dimensionen der Raumvorstellung es gibt, offen. So besteht weiterhin Disput, ob eine Speed-Komponente entscheidend ist für den Faktor Spatial Relations. Deutlich wird dies bei Lohmans Studien. Während er 1979 noch den Aspekt der mentalen Rotation betonte, so legte er 1988 den Schwerpunkt auf die Geschwindigkeit, mit der Rotationsaufgaben gelöst wurden. In letzter Konsequenz bedeutet dies, dass der Faktor inhaltlich entweder die Fähigkeit zur mentalen Rotation oder die Bearbeitungsgeschwindigkeit bei entsprechenden Items darstellt.
Ein weiterer Kritikpunkt besteht in der Trennung dieses Faktors vom Faktor Visualization, die in erster Linie auf die Komplexität der Items fußt (Carroll, 1993; Lohman, 1988; Zimmerman, 1953). Problem ist hier, dass der Faktor Visualization nur vage definiert wird (z. B. Ekstrom et al., 1976; McGee, 1979; Thurstone, 1949), verschiedenste mentale Operationen anspricht (z. B. Guilford & Lacey, 1947; Lohman, 1988) und es damit nicht ersichtlich ist, wie die geforderte Komplexität von Items zu erreichen bzw. zu modifizieren ist. Somit ist auch eine Trennung der Faktoren Spatial Relations und Visualization nicht immer möglich (Linn & Petersen, 1985). Lohman (1988) sah den Mental Rotations Test beispielsweise als marker test für den Faktor Visualization.
Hinsichtlich des Faktors Spatial Orientation and Perception zeigen sich ähnliche Probleme. Zum einen konnte er in einigen Studien grundsätzlich nicht nachgewiesen werden (z. B. Ekstrom et al., 1976; Zimmerman, 1953), zum anderen fiel es schwer, ihn von anderen Faktoren zu trennen (z. B. Carroll, 1993; McGee, 1979). Michael et al. (1957) definierten so zum Beispiel *einen* gemeinsamen Faktor für Spatial Relations und Spatial Orientation and Perception. Auf den hohen Zusammenhang mit dem Faktor Visualizati-

on wiesen Hegarty und Waller (2005) hin, was darauf zurückzuführen ist, dass die Items dieses Faktors auch durch Fähigkeiten, die dem Faktor Visualization zuzuordnen sind, gelöst werden können (D'Oliviera, 2004; Lohman, 1988). Das heißt, anstatt beispielsweise den eigenen Standpunkt gedanklich zu verändern, kann eine Testperson auch das Objekt entsprechend manipulieren (Hegarty & Waller, 2004).

Dass die Faktoren kaum nachgewiesen wurden oder schwer von anderen zu trennen sind, trifft auch den Kinesthetic factor sowie dem Faktor Length Estimation zu. Lohman (1979) wies für den Kinesthetic factor darauf hin, dass er auch Teil des Faktors Spatial Orientation and Perception sein könnte, insbesondere bei komplexen Items. Die Existenz des Faktors ergab sich laut Lohman (1979) nur bei Items mit Zeitbeschränkung, was wiederum zu ähnlichen inhaltlichen Problemen führt, die beim Faktor Spatial Relations bereits erläutert wurden. Das grundsätzliche Problem für beide, Kinesthetic factor und dem Faktor Length Estimation, besteht jedoch darin, dass kaum Tests existieren, die ihre Existenz nachweisen konnten (Carroll, 1993; Lohman, 1988). Beim Length Estimation Faktor zweifelte Carroll (1993), auch wenn er ihn als Raumvorstellungsfaktor führte, sogar seine Nützlichkeit an.

Die Schwierigkeit, mögliche Faktoren klar voneinander abzugrenzen, ist allerdings auch auf die Operationalisierung der Fragestellung sowie der statistischen Methode zurückzuführen (Hegarty & Waller, 2005). Wie bereits erwähnt, hat die Faktorenanalyse die Raumvorstellungsforschung entscheidend geprägt und hier insbesondere die explorative Faktorenanalyse. Diese erlaubt es, ohne ein theoriebasiertes Modell, lediglich aufgrund der Korrelationen der Variablen miteinander (in diesem Falle der verwendeten Tests) auf einzelne, wenige Faktoren zu schließen (Kubinger et al., 2011). Dieser Ansatz birgt jedoch Probleme, die in den dargestellten Studienergebnissen wiederzufinden sind:

Anzahl der Faktoren: Bei einer explorativen Faktorenanalyse werden mehr Faktoren ermittelt, als letztendlich inhaltlich sinnvoll interpretiert werden können. Es existieren jedoch keine eindeutigen Kriterien, lediglich Faustregeln, wie viele Faktoren ausgewählt werden (Kubinger et al., 2011).

Auswahl der Tests: Der fehlende Nachweis eines Faktors bedeutet nicht zwangsweise dessen Nicht-Existenz. Vielmehr kann ein Faktor nur dann ermittelt werden, wenn auch entsprechende Tests vorgegeben wurden, die diesen Faktor ansprechen. Für die Faktorenanalyse bedeutet dies, dass die jeweiligen Tests hoch auf diesen Faktor laden (Hegarty & Waller, 2005).

Aufgrund der fehlenden Theorie, welche Dimensionen der Raumvorstellung a priori angenommen wurden, gab es auch keine Bestimmungen, welche Tests dementsprechend vorgegeben wurden. Daher ist die gefundene Faktorenkonstellation auch das Ergebnis der Testauswahl der jeweiligen Studien (D'Oliviera, 2004).

Ebenso hängt die Identifizierung der Faktoren von der Verfügbarkeit von relevanten Tests ab, was für den Faktor Length Estimation und dem Kinesthetic factor sowie für den Faktor Spatial Orientation and Percepetion (Carroll, 1993) problematisch ist.

Missachtung individueller Unterschiede: Die Faktorenanalyse führt zu dem Ergebnis, dass ein bestimmter Test hoch auf einen Faktor lädt, was im Umkehrschluss bedeutet, dass dieser Faktor (z. B. die Fähigkeit zur mentalen Rotation) wesentlich zur Lösung der Items dieses Tests beiträgt. Damit missachtet sie allerdings individuelle Unterschiede, sodass Testpersonen zum Beispiel unterschiedliche Strategien zum Lösen der Items verwenden können (Carroll, 1993; Hegarty & Waller, 2005; Lohman, 1988). Dies kann die Identifizierung der Faktoren beeinflussen und letztendlich dazu führen, dass sich ein hoher statistischer Zusammenhang zwischen Faktoren ergibt, obwohl diese unabhängig voneinander sein sollten. Ein Beispiel hierfür ist die mangelnde Differenzierbarkeit zwischen den Faktoren Visualization und Spatial Relations zum Faktor Spatial Orientation and Percepetion. Das bedeutet, dass Testpersonen die Tests verschiedener Faktoren auf die gleiche Art und Weise lösen können.

An dieser Stelle sei erwähnt, dass Colom et al. (2001), Hegarty und Waller (2004) und W. Johnson und Bouchard Jr. (2005) mit konfirmatorischen Faktorenanalysen arbeite-

ten, die im Gegensatz zur explorativen Faktorenanalyse versuchen, eine theoriebasierte Faktorenstruktur nachzuweisen (Kubinger et al., 2011). Ihre Ergebnisse (siehe Tabelle 2) sind im Vergleich zur bisherigen Forschung sowie untereinander höchst widersprüchlich, was unter anderem die Annahme eines multidimensionalen Merkmals Raumvorstellung in Frage stellt. So gelang es Colom et al. (2001) nicht, mehrere Raumvorstellungsfaktoren zu identifizieren, was sie zu der Schlussfolgerung führte, Raumvorstellung als eindimensionale Fähigkeit, ähnlich den Studien zu Beginn des 20. Jahrhunderts (mit Ausnahme von Kelley (1928)), zu betrachten. Sie wiesen in ihrer Studie auf die hohe Test- und Item-Spezifität der bisherigen Raumvorstellungsfaktoren hin. Das heißt, dass für verschiedenen Faktoren spezifische marker tests existieren, die sie bestmöglich erfassen. Eine Kritik, die bereits Lohman (1988) formulierte, indem er monierte, dass Faktoren statt einer latenten Dimension der Raumvorstellung auch nur die Fähigkeit zur Lösung sehr spezifischer Items, nämlich derjenigen des jeweiligen Tests, darstellen könnten. Indem Colom et al. (2001) im Weiteren (absichtlich oder unabsichtlich) größtenteils Tests vorgaben, die nicht den gängigen entsprachen und es nicht gelang, verschiedene Raumvorstellungsfaktoren zu identifizieren, führten sie damit auch einen empirischen Beleg für diese Kritik.

Von besonderer Relevanz für den TARV ist eine Studie von Stumpf und Eliot (1995). In ihr wurden zwei Testbatterien[2] mit Raumvorstellungstests erstellt. Beide Batterien enthielten zwar unterschiedliche Tests, jedoch wurde angenommen, dass die Tests marker tests für die gleichen Faktoren waren. Das bedeutet, jede Testbatterie sollte die gleiche Faktorenstruktur ermitteln, allerdings konnte für beide mittels explorativer Faktorenanalyse jeweils nur ein Faktor identifiziert werden. In Anlehnung an Kelley (1928) wurde dieser als k-Faktor bezeichnet. Stumpf und Eliot (1995) vermuteten daher ein hierarchisches Modell der Raumvorstellung mit einem allgemeinem Faktor, dem k-Faktor, an der Spitze und die beispielsweise in Kapitel *1.2. Raumvorstellung als ein multidimensionales Merkmal* genannten Faktoren diesem untergeordnet. Ebenso gingen auch Gittler und Arendasy (2003) von einem hierarchischen Model aus, allerdings ohne einen allgemeinen k-Faktor, stattdessen erfolgte eine Aufteilung in mehrere Haupt- und Nebenfaktoren.

[2]Mehrere Untertests werden dabei zu einer Testbatterie zusammengefasst (Kubinger, 2009).

Das Resultat dieser divergenten Ergebnisse zur Dimensionalität bzw. zur Anzahl der Faktoren der Raumvorstellung ist, dass es der Forschung bis dato nicht gelungen ist, eine allgemein akzeptierte Definition des Merkmals Raumvorstellung zu formulieren. Voyer, Voyer und Bryden (1995) brachten dies wie folgt auf den Punkt:

> To define spatial abilities, one must initially determine whether spatial ability is a unitary concept or involves a number of diverse components. Unfortunately, there is little agreement among authors as to how spatial abilities should be classified. At first glance, it seems that this lack of agreement could be avoided by the use of a factor analytic approach in order to group and define spatial abilities. However, even this approach does not necessarily produce converging definitions in the literature. (...) The lack of a universally accepted definition of spatial ability may be due to the large variety of tests used in the psychometric studies or to the lack of replicability of the factor structures found when several tests are used. (S. 251)

Ähnliche Kritik, insbesondere an der Methode, kam auch von Gittler und Arendasy (2003):

> Welche und wie viele Faktoren jedoch notwendig sind, um die dimensionale Binnenstruktur dieses Fähigkeitsbereichs adäquat zu beschreiben, ist bislang nicht einheitlich geklärt worden. Diese durchaus unbefriedigende Situation bezüglich Faktorenidentifikation kann vor dem Hintergrund grundlegender Kritik an der Faktorenanalyse (...) als ein Methodenproblem gewertet werden. (S. 164)

Zusammengefasst führte demnach ein streckenweise theoriebildender Zugang, gefördert durch die Verwendung explorativer Faktorenanalysen, zu einer unsystematischen Auswahl und/oder Konstruktion an bzw. von Tests mit der Folge einer unterschiedlichen Menge an identifizierten Faktoren. Die Frage jedoch, ob Raumvorstellung ein ein- oder multidimensionales Merkmal ist und sich womöglich ein hierarchisches Modell ergibt, blieb unbeantwortet.

2. Raumvorstellung: Differentielle Forschungsrichtung

2.1. Experimente, Zufallsauswahl und Randomisierung

In den folgenden Kapiteln wird auf die Unterschiede zwischen den Geschlechtern sowie zwischen Personen mit verschiedenen Ausbildungen in ihrer Raumvorstellung eingegangen. Vorab ist dazu allerdings eine Anmerkung hinsichtlich der Interpretation der geschilderten Ergebnisse notwendig.

Ein Versuch oder ein Experiment bedeutet, Personen einer Stichprobe werden „systematisch gewissen „Einwirkungen"/Bedingungen (...) ausgesetzt - etwa zuerst *zufällig* in Gruppen eingeteilt und dann je Gruppe einer anderen (psychologischen) „Behandlung" unterzogen - und schließlich (psychologisch) untersucht" (Kubinger et al., 2011, S. 52). Nur ein Experiment erlaubt kausale Schlussfolgerungen z. B. in der Form, dass Unterschiede zwischen den Gruppen auf die jeweilige „Behandlung" zurückzuführen sind. Notwendige Voraussetzungen für ein Experiment sind Zufallsauswahl und Randomisierung. Das heißt, dass die „Versuchseinheiten [ausgewählte Personen] (...) zufällig der Grundgesamtheit [aller in Frage kommenden Personen] entnommen wurden" (Kubinger et al., 2011, S. 158) und anschließend die „zufällige Zuordnung der Versuchseinheiten [ausgewählte Personen] zu den Versuchsbedingungen erfolgt" (Kubinger et al., 2011, S. 158).

Betrachtet man Unterschiede zwischen den Geschlechtern sowie zwischen Personen, die verschiedene Ausbildungen absolvieren bzw. absolviert haben, so handelt es sich hierbei nicht um Experimente. Eine zufällige Zuteilung zu den verschiedenen Bedingungen, Mann bzw. Frau bei Geschlecht und z. B. unterschiedliche Studiengänge bei der Ausbildung, ist nicht möglich bzw. nicht vertretbar. Die Personen teilen sich also von selbst in verschiedene Gruppen ein, weswegen nicht kausal argumentiert werden kann, dass Unterschiede zwischen den Gruppen auf die jeweilige Gruppenzugehörigkeit zurückzuführen sind. Für den konkreten Fall bedeutet dies: Festgestellte Unterschiede in der Raumvorstellung zwischen den Geschlechtern sowie zwischen Personen mit verschiedenen Ausbildungen können nicht einzig und allein durch das Geschlecht bzw. die absolvierte Ausbildung erklärt werden.

Der Fokus dieser Arbeit liegt auf der Kalibrierung der Items des TARV, um einen fairen[3] Test zu kreieren, der beiden Geschlechtern sowie Personen mit verschiedenen Ausbildungen ähnlich schwer fällt und systematische Benachteiligung vermeidet. Die Gründe zu identifizieren, weswegen sich Personen in ihrer Raumvorstellung unterscheiden, ist daher nicht das Ziel.

2.2. Unterschiede in der Raumvorstellung: Geschlecht

Die offene Frage zur Anzahl möglicher Raumvorstellungsfaktoren findet bei der Erforschung von Geschlechtsunterschieden ihren Niederschlag. Die Ergebnisse unterscheiden sich dabei, je nachdem, welche Faktoren a priori angenommen und mit welchen Tests demzufolge Raumvorstellung zu erfassen versucht wurde. Die Existenz bzw. die Größe des Geschlechtsunterschieds hängt folglich vom jeweiligen Test ab, der verwendet wurde (Halpern, 2000).
Für eine Rekapitulation der umfangreichen Literatur zu Geschlechtsunterschieden in der Raumvorstellung wäre es daher notwendig, von der Ebene der Faktoren auf die Ebene einzelner Tests zu gehen und deren Items konkret zu erläutern. Davon wird in dieser Arbeit abgesehen. Ein, wenn auch mittlerweile veralteter, Überblick über die verschiedenen Raumvorstellungstests findet sich beispielsweise bei Eliot und Macfarlane Smith (1983).
Die folgende Betrachtung liefert zuerst einen Überblick über die Geschlechtsunterschiede in den einzelnen Faktoren. Auf die Rolle einzelner Tests wird eingegangen, allerdings wird sie, wie soeben erwähnt, nur ansatzweise skizziert. Vielmehr liegt der Schwerpunkt auf den Ursachen für mögliche Geschlechtsunterschiede, die im Bereich der Test- und Itemkonstruktion liegen und damit Relevanz für den TARV besitzen.

Einen ersten Überblick über die Ergebnisse verschiedener Studien zu Geschlechtsunterschieden in der Raumvorstellung liefern die Metaanalysen von Linn und Petersen (1985)

[3] „Die resultierenden Testwerte [führen] zu keiner systematischen Diskriminierung bestimmter Testpersonen zum Beispiel aufgrund ihrer ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppenzugehörigkeit.“ (Kubinger, 2009, S. 123)

sowie Voyer et al. (1995). Linn und Petersen (1985) definierten hierfür vorab drei Faktoren der Raumvorstellung (siehe Tabelle A), die größtenteils den geschilderten Faktoren in Kapitel *1.2. Raumvorstellung als ein multidimensionales Merkmal* entsprechen, und ordneten ihnen die Testergebnisse der in ihre Metaanalyse einbezogenen Studien entsprechend zu. Voyer et al. (1995) fuhren mit diesem Schema fort, standen dem Vorgehen allerdings kritisch gegenüber (siehe Kapitel *2.2.1. Kritik an der differentiellen Forschungsrichtung zu Geschlechtsunterschieden*). Für die in Kapitel *1.2. Raumvorstellung als ein multidimensionales Merkmal* beschriebenen Faktoren ergab sich dabei folgendes Bild.

Spatial Relations: In beiden Metanalysen erzielten Männer bessere Leistungen als Frauen mit einer durchschnittlichen Effektgröße[4] von $d = 0{,}73$ (Linn & Petersen, 1985) bzw. $d = 0{,}56$ (Voyer et al., 1995). Die Höhe der Effektgröße bedingte insbesondere der Mental Rotations Test mit $d = 0{,}94$ (Linn & Petersen, 1985) respektive $d = 0{,}67$ (Voyer et al., 1995), während die anderen diesem Faktor zugeordneten Tests zwar ebenso eine bessere Leistung der Männer ermittelten, jedoch mit Effektgrößen von maximal $d = 0{,}26$ (Linn & Petersen, 1985) und $d = 0{,}44$ (Voyer et al., 1995). Im Verlauf der vergangenen Jahre konnte die hohe Effektgröße des Mental Rotations Test mehrfach repliziert werden (z. B. Masters & Sanders, 1993; Peters, 2005; Kaufman, 2007; Geiser, Lehmann & Eid, 2008; Voyer, 2011), eine Erklärung für die Ursache des großen Geschlechtsunterschieds fehlt allerdings bis dato (Bors & Vigneau, 2011).

Ein differenzierteres Bild ergibt sich durch Studien, die keine durchwegs statistisch signifikanten Unterschiede ermitteln konnten (Collins & Kimura, 1997; Rilea, 2008; Weiss, Kemmler, Deisenhammer, Fleischhacker & Delazer, 2003). So unterschieden sich beispielsweise bei Olson und Eliot (1986) im Gegensatz zur Metaanalyse von Voyer et al. (1995) Männer und Frauen beim Card Rotations Test nicht voneinander. Gründe hierfür können in der Test- und Itemkonstruktion gefunden werden, auf die in Kapitel *2.2.2. Test- und Itemcharakteristika als mögliche Ursachen des*

[4] „(Mittelwerts-)Unterschied in Einheiten der Standardabweichung." (Kubinger et al., 2011, S. 372)

Geschlechtsunterschieds (Anzahl der Dimensionen) eingegangen wird.

Spatial Orientation and Perception: Dieser Faktor wurde in beiden Metaanalysen (Linn & Petersen, 1985; Voyer et al., 1995) sehr eng gefasst (siehe die entsprechende Definition „Spatial Perception“ in Tabelle A) und bezog sich mehr darauf, die räumliche Beziehung von Objekten zueinander in Bezug zur eigenen Position korrekt zu bestimmen, anstatt in der Lage zu sein, sich ein Objekt aus einer anderen Perspektive vorzustellen. Diese Einschränkung führte dazu, dass der Faktor in beiden Metaanalysen nur durch zwei Tests, unter anderem dem Water Level Task, repräsentiert wurde. Für beide ergaben sich statistisch signifikante Unterschiede zugunsten der Männer mit einer durchschnittlichen Effektgröße von jeweils $d = 0{,}44$ (Linn & Petersen, 1985; Voyer et al., 1995). Zwar existieren auch Studien, die keine statistisch signifikanten Unterschiede ermittelten, die generelle Tendenz, dass Männer in diesen beiden Tests bessere Leistungen erbringen, überwiegt jedoch (Halpern, 2000; Rilea, 2008).

Für Tests, die ihren Schwerpunkt bei der Bearbeitung der Items auf einen Perspektivenwechsel legen, zeigte sich ein inkonsistentes Bild. Eliot und Macfarlane Smith (1983) und Moffat, Hampson und Hatzipantelis (1998) berichteten beim Spatial Orientation Test von einer besseren Leistung der Männer, während Hegarty et al. (2006) beim Object Perspective Taking Test keine Unterschiede feststellen konnten. Die Ergebnisse könnten allerdings davon beeinflusst sein, dass die Items dieser Tests ebenso durch gedankliche Manipulation der dargestellten Objekte anstatt mit einem Wechsel der eigenen Perspektive zu lösen sind (Hegarty & Waller, 2004). Für den Pictures Test wurden keine Angaben zu Geschlechtsunterschieden gemacht.

Die Vorgehensweise der Testperson würde damit bestimmen, welche Faktoren erfasst und welche Unterschiede sich zwischen den Geschlechtern ergeben würden. Für den Spatial Orientation Test wird daher in Frage gestellt, ob er den gewünschten Perspektivenwechsel bei allen Testpersonen erzeugte (Hegarty & Waller, 2004; Kozhevnikov & Hegarty, 2001). Gleiches trifft auf den Schlauchfiguren-Test zu, in

dem Männer besser abschneiden als Frauen mit einer durchschnittlichen Effektgröße von $d = 0{,}48$ (Stumpf & Klieme, 1989). Das Problem, dass dieser Test auch anderen Faktoren zuzuordnen wäre, wurden bereits in Kapitel *1.2. Raumvorstellung als ein multidimensionales Merkmal* erläutert.

Visualization: Linn und Petersen (1985) ermittelten eine durchschnittliche Effektgröße von $d = 0{,}13$ und Voyer et al. (1995) von $d = 0{,}19$. Männer erbrachten in den Studien bessere Leistungen, der Unterschied zu den Frauen war jedoch nicht statistisch signifikant. Voyer et al. (1995) sahen diesen Faktor allerdings als problematisch an, da er in ihren Augen lediglich die Tests beinhaltete, die den zwei zuvor genannten Faktoren nicht zugeordnet werden konnten. Die Testauswahl in diesem Faktor hatte für sie eine hohe Heterogenität, betrachtet man die verschiedenen gedanklichen Operationen, die in den Tests durchzuführen waren. Als Indiz dafür zeigten sich bei Linn und Petersen (1985) in den einzelnen Studien eine Spanne der Effektgrößen von d = -0,91 bis 0,71. Hinsichtlich der *durchschnittlichen* Geschlechtsunterschiede je Test schwankten die Effektgrößen von $d = 0{,}12$ bis 0,27. Insbesondere der Paper Folding Test sowie Surface Development Test zeigten sich mit den niedrigsten Effektgrößen resistent gegenüber Geschlechtsunterschieden, was sich in weiteren Studien bestätigte (Halpern & Collaer, 2005; vgl. Feingold, 1988; Weiss et al., 2003). Für den Dreidimensionalen Würfeltest berichtete Gittler (1990) allerdings von besseren Leistungen der Männer.

Kinesthetic factor: In beiden Metaanalysen wurde dieser Faktor nicht berücksichtigt. Für den Right-Left Discrimination Test wurde in einigen Studien eine bessere Leistung der Männer nachgewiesen (Gormley, Dempster & Best, 2011; Ocklenburg, Hirnstein, Ohmann & Hausmann, 2011; Ofte, 2002; Ofte & Hugdahl, 2002), bei Gormley et al. (2011) mit einer Effektgröße von $d = 0{,}25$. Beim Hands Test, der bei SchülerInnen verschiedenen Alters durchgeführt wurde, zeigte sich bei E. S. Johnson und Meade (1987) ein statistisch signifikanter Geschlechtsunterschied zugunsten der männlichen Teilnehmer mit einer Effektgröße von bis zu $d = 0{,}60$. Teng und Lee

(1982) hingegen konnten keinen statistisch signifikanten Unterschied ermitteln. Bei einem Test, bei dem bestimmt werden musste, ob sich ein Objekt links oder rechts von einem anderen Objekt befand, unterschieden sich Männer und Frauen ebenso nicht statistisch signifikant voneinander (Jordan, Wüstenberg, Jaspers-Feyer, Fellbrich & Peters, 2006). Der Kinesthetic factor liefert somit ein sehr widersprüchliches Bild zwischen nicht vorhandenen und deutlichen Geschlechtsunterschieden. Ein Kritikpunkt an den Tests ist, dass sie mentale Rotation erfordern (Jordan et al., 2006), da zur Lösung von Items es zum Teil erst notwendig ist, ein Objekt entsprechen zu rotieren, bevor Richtungsangaben getätigt werden können. Damit würde der Test wiederum andere Faktoren ansprechen. Widerlegt wurde diese Aussage von Hirnstein, Ocklenburg, Schneider und Hausmann (2009), die nachweisen konnten, dass bei Aufgaben mit und ohne Rotation Männer eine bessere Leistung erbrachten.

Length Estimation: Dadurch, dass der Faktor wenig im Fokus der Wissenschaft stand, konnten keine Studien zu Geschlechtsunterschieden ermittelt werden.

Es lässt sich zusammenfassen, dass es bedeutsame Unterschiede bei der Fähigkeit zur mentalen Rotation von Objekten gibt (Spatial Relations), in denen Männer eine bessere Leistung erbringen. Das Gleiche gilt für die Wahrnehmung der Positionen von Objekten und ihrer räumlichen Beziehungen zueinander (Spatial Orientation and Perception). Bei Tests, in denen der Wechsel des eigenen Standpunkts bzw. der Perspektive zur Lösung der Items notwendig ist, können keine Aussagen getroffen werden. Den männlichen Vorteil, den manche Tests ermitteln, könnte durch verschiedene Bearbeitungsstrategien der Testpersonen bedingt sein. Das heißt, anstatt die Perspektive gedanklich zu wechseln, wird das jeweilige Objekt mental rotiert. Für Tests, die aus diesem Grund eine Rotation der Objekte auszuschließen versuchen, konnten keine Daten ermittelt werden. Hinsichtlich der Fähigkeit, mehrere gedankliche Operationen auszuführen, ohne dass der Test den Schwerpunkt nur auf eine Bearbeitungsstrategie legt (Visualization), zeigte sich kein Unterschied zwischen den Geschlechtern. Keine Aussagen können getroffen werden, wenn es darum geht, korrekte Richtungs- und Längenangaben zu tätigen (Kinesthetic factor,

Length Estimation), da im ersten Fall die Ergebnisse zu widersprüchlich sind und im zweiten Fall keine Daten vorliegen.
Ein generelles Fazit formulierte Halpern und Collaer (2005): "When sex differences are found, they favor males. (...) There are, however, imortant differences among these tasks in the size of the differences between the sexes"(S. 175).

2.2.1. Kritik an der differentiellen Forschungsrichtung zu Geschlechtsunterschieden

Wie soeben erwähnt, hängt die Größe des Geschlechtsunterschieds für Faktoren von den Tests und deren Zusammenstellung ab, zumal einzelne Tests nicht eindeutig den Faktoren zugeordnet werden können. Voyer et al. (1995) gelang es zwar, ähnliche Durchschnittswerte wie Linn und Petersen (1985) für die Faktoren zu ermitteln, sie wiesen jedoch auf die Unterschiede zwischen den Tests innerhalb der Faktoren hin. Dies bestätigte sich darin, dass in beiden Metaanalysen die Effektgrößen innerhalb der Faktoren nicht homogen waren. Streng genommen stellt dies die von Linn und Petersen (1985) definierte 3-Faktoren-Struktur des Merkmals Raumvorstellung in Frage, da anzunehmen sei, dass Tests innerhalb eines Faktors ähnlich große Geschlechtsunterschiede ermitteln würden. Beispielhaft erzielten so E. S. Johnson und Meade (1987) mit einer etwas anderen Testauswahl als Linn und Petersen (1985) auch andere Werte für die Geschlechtsunterschiede. Am auffälligsten war dies beim Faktor Visualization, bei dem Männer mit einer durchschnittlichen Effektgröße von $d = 0{,}40$ besser abschnitten. Zurückzuführen ist das darauf, dass sie keinen der Tests vorgaben, die bei Linn und Petersen (1985) die geringsten Effektgrößen erzielten (Surface Development Test und Paper Folding Test). Sie, wie Voyer et al. (1995), merkten daher an, dass Unterschiede nicht auf Faktorenebene, sondern auf einer niedrigeren, z. B. auf der Ebene einzelner Tests betrachtet werden sollten. Dies würde allerdings implizieren, dass jeder Test für sich eine andere Fähigkeit der Raumvorstellung erfasst (Voyer et al., 1995), was die Diskussion um die Anzahl der Raumvorstellungfaktoren und die damit gemeinten unterschiedlichen (Fähigkeits-)Dimensionen der

Raumvorstellung verschärfen würde.

Eine umfassende Kritik an der gesamten differentiellen Forschungsrichtung zu Geschlechtsunterschieden kam von Caplan, MacPherson und Tobin (1985). Zwei Kritikpunkte sind hierbei hervorzuheben. Zum einen, dass nicht jeder Test Raumvorstellung misst und zum anderen, dass Ergebnisse zu Geschlechtsunterschieden bei einzelnen Tests zu sehr ausgeweitet werden. Beide Punkte waren auf den Rod and Frame Test (Witkin et al., 1954/1972) bezogen, der in seiner Konzeption dem Water Level Task ähnelt, allerdings findet sich der Inhalt der Kritik auch in anderen Studien mit anderen Tests.

Kubinger (2009) wies hinsichtlich des ersten Kritikpunkts bereits darauf hin, dass Raumvorstellungsitems teilweise auch mit schlussfolgerndem Denken[5] gelöst werden könnten und somit Tests nicht bei allen Testpersonen Raumvorstellung messen würden. Bors und Vigneau (2011) gruppierten die Items des Mental Rotations Test in leichte und schwierige auf Grundlage des Ausmaß an mentaler Rotation, die zur Lösung der Items durchgeführt werden musste. Sie gingen davon aus, dass der Geschlechtsunterschied bei den schwierigen Items noch größer werden würde, da Männer bei diesem Test durchschnittlich deutlich bessere Werte (vgl. Linn & Petersen, 1985; Voyer et al., 1995) erzielen. Dies war jedoch nicht der Fall, was zur Frage führte, was den Geschlechtsunterschied bei diesem Test überhaupt bedingt, wenn es nicht das Ausmaß an gedanklicher Rotation ist, was der Test eigentlich zu messen beabsichtigt.

Der zweite Kritikpunkt von Caplan et al. (1985) war darauf bezogen, dass Geschlechtsunterschiede vom jeweiligen Test abhängig sind und ebenso verschwinden können, wenn sich die Gestaltung des Tests ändert. Daher sollten Aussagen zu Unterschieden nur in Bezug zum Test gesetzt und nicht darüber hinaus ausgeweitet werden. Beispiele für das Verschwinden von Geschlechtsunterschieden in Tests durch Änderung seiner Gestaltung finden sich im folgenden Kapitel bei „dimensionality crossing“, wo durch das Rotieren echter Objekte oder die Vorgabe mittels dreidimensionaler Projektionen sich Männer und Frauen in Rotationstests nicht mehr voneinander unterschieden.

[5] „Fähigkeit, Gesetzmäßigkeiten oder logisch zwingende Zusammenhänge erkennen und zweckentsprechend verwerten zu können.“ (Kubinger, 2009, S. 206)

2.2.2. Test- und Itemcharakteristika als mögliche Ursachen des Geschlechtsunterschieds

Im Folgenden werden Gründe dargestellt, weswegen Männer in manchen Tests besser abschneiden als Frauen. Das Augenmerk der meisten Studie lag dabei auf dem Unterschied im Bereich der mentalen Rotation und dort insbesondere beim Mental Rotation Test, der seit Jahren konstante Geschlechtsunterschiede ermittelt (Masters & Sanders, 1993). Wenngleich die Ergebnisse bezüglich dieses Tests nicht analog auf andere Test für mentale Rotation übertragbar sind, so können sie doch als wichtige Anhaltspunkte fungieren, die bei einer Testkonstruktion zu berücksichtigen sind, oder wie Peters et al. (1995) es formulierten: „The question is not: why are males better than females on spatial tasks, but rather, what tasks yield sex differences and why“ (S. 39)?
Generell ist die Forschung bezüglich der Ursachen von Geschlechtsunterschieden äußerst umfangreich. Einen Überblick über die gesamte Breite der Forschung zu diesem Thema liefert Halpern und Collaer (2005). Aus Gründen der Relevanz für diese Arbeit werden lediglich diejenigen Ursachen für den Geschlechtsunterschied dargestellt, die in der Itemkonstruktion sowie der generellen Gestaltung von Tests zu finden sind. Diese werden von Halpern und Collaer (2005) als „Performance Factors“ (S. 198) bezeichnet und meinen Variablen, die weitestgehend unabhängig vom Merkmal Raumvorstellung sind, aber den Geschlechtsunterschied zugunsten eines der Geschlechter begünstigen. Allerdings, so schränken sie auch ein: “It is unlikely that the large and consistent sex differences found across many decades of research are caused by performance factors.“ (S. 199).

Unmögliche Rotationen:

Kerkman, Wise und Harwood (2000) zeigten, dass sich ein statistisch signifikanter Geschlechtsunterschied zugunsten der Männer nur bei möglichen, nicht jedoch bei unmöglichen Rotationen ergab. Eine unmögliche Rotation bedeutet, dass ein Objekt in keiner Weise so rotiert werden kann, dass es einem anderen, zu vergleichenden Objekt entspricht, zum Beispiel weil dieses andere Objekt spiegelverkehrt dargestellt ist. Dies würde zu ei-

ner systematischen Benachteiligung von Frauen führen, je nachdem, ob ein Test von den Testpersonen verlangt, mögliche oder unmögliche Rotationen bzw. im weitere Sinne falsch rotierte Objekte oder Fehler zu erkennen.

Bearbeitungszeit insgesamt:

Goldstein, Haldane und Mitchel (1990) konnten beim Mental Rotations Test nachweisen, dass die Geschlechtsunterschiede lediglich bei begrenzter, nicht jedoch bei unbegrenzter Bearbeitungszeit vorlagen. Robert und Chevrier (2003) und Voyer (1997) replizierten dieses Ergebnis, was die Frage aufwarf, ob ein Zeitlimit die Ursache des Unterschieds sei. Eine Metaanalyse von Voyer (2011) kam zu dem Schluss, dass die Geschlechtsunterschiede bei Tests für mentale Rotation abnahmen, je mehr Zeit den Testpersonen zur Verfügung stand. Die Effektgröße des Unterschieds betrug bei zeitlich unbegrenzten Tests dennoch durchschnittlich $d = 0{,}51$. Glück und Fabrizii (2010) verglichen die Leistung von Männern und Frauen beim Mental Rotations Test mit und ohne Zeitbeschränkung. Nur bei der Bedingung mit Zeitbeschränkung zeigte sich ein statistisch signifikanter Unterschied. Sie führten diesen auf unterschiedliche Bearbeitungsstrategien zurück. Männer wandten in ihren Augen eine „quick-and-dirty“ (S. 109) Strategie an, um so möglichst viele Items in der vorgegebenen Zeit zu lösen, während Frauen zu viel Zeit dadurch verloren, dass sie ihre Antworten mehrmals überprüften.

Komplexität von Items:

Eine Studie von Rilea (2008) erbrachte das Ergebnis, dass die Zunahme der Itemkomplexität Auswirkungen auf den Geschlechtsunterschied hatte. Dabei unterschieden sich Männer und Frauen bei der Rotation zweidimensionaler Strichmännchen nicht statistisch signifikant, jedoch schon, wenn stattdessen komplexere, zweidimensionale Polygone rotiert werden mussten. Ein ähnliches Resultat lieferten Bors und Vigneau (2011). Bei ihnen nahm die Größe des Geschlechtsunterschieds zugunsten der Männer beim Mental Rotations Test mit zunehmender Itemkomplexität zu. Der Korrelationskoeffizient hierfür

betrug $r = 0{,}64$.

Anzahl der Dimensionen:

Intensiver beschäftigte sich die Forschung damit, ob Geschlechtsunterschiede auf die Dimensionalität der dargestellten Objekte sowie auf den Zustand des Testmaterials zurückzuführen sind. Das heißt, ob die gedankliche Rotation zwei- oder dreidimensionaler Objekte stärker zwischen Männern und Frauen differenziert oder aber auch, ob die Handhabung mit echten Objekten bzw. Materialien den Unterschied anders beeinflusst als die Durchführung am Computer oder als Papier-Bleistift-Verfahren.

zweidimensional versus dreidimensional dargestellte Objekte: Bereits in den Metaanalysen (Linn & Petersen, 1985; Voyer et al., 1995) zeigte sich, dass Tests mit ausschließlich mentaler Rotation von zweidimensionalen Objekten deutlich geringere Effektgrößen erzielten als entsprechende „dreidimensionale Tests“. Der Vergleich betrug bei Linn und Petersen (1985) $d = 0{,}26$ zu $d = 0{,}94$ und bei Voyer et al. (1995) zum Beispiel $d = 0{,}31$ für den Card Rotations Test zu $d = 0{,}67$ für den Mental Rotations Test. Bezogen auf einzelne Studien konnten beispielsweise Peters et al. (1995), Roberts und Bell (2003) sowie Uecker und Obrzut (1993) keinen statistisch signifikanten Unterschied bei zweidimensionaler Rotation feststellen. Insbesondere Peters et al. (1995) und Roberts und Bell (2003) sind hervorzuheben. Sie ließen in ihren Studien die selben Testpersonen Tests mit zwei- und dreidimensionalen Objekten bearbeiten. Die Effektgrößen betrugen exemplarisch bei Peters et al. (1995) $d = 0{,}03$ für den Card Rotations Test sowie $d = 0{,}89$ für den Mental Rotations Test. Konträre Ergebnisse kommen jedoch von Birenbaum, Kelly und Levi-Keren (1994) und Collins und Kimura (1997), die statistisch signifikante Geschlechtsunterschiede feststellen konnten. Collins und Kimura (1997) konnten dabei zeigen, dass sich ein Unterschied zwar nicht bei leichten, jedoch bei der Rotation schwieriger zweidimensionaler Objekte ergab. Die Effektgröße lag hierfür sogar höher, nämlich bei $d = 1{,}10$, als die eines Tests mit dreidimensionalen Objekten (Mental Rotati-

ons Test) mit $d = 0{,}86$. Ihre Schlussfolgerung war, dass der Geschlechtsunterschied weniger durch die Dimension als durch die Schwierigkeit der Items bedingt sei. Die Schwierigkeit ihres Tests lag unter Umständen in den Antwortmöglichkeiten, die die Testpersonen bei jedem Item zur Verfügung hatten. Bei den Tests von Peters et al. (1995) oder Roberts und Bell (2003) lag der Schwerpunkt darauf, die Antwortmöglichkeiten bzw. Objekte zu identifizieren, die einem vorgegebenen Objekt entsprachen oder umgekehrt, diejenigen Antwortmöglichkeiten auszuschließen, bei denen es unmöglich war, sie durch Rotation in Einklang mit der Vorgabe zu bringen. Bei Collins und Kimura (1997) hingegen entsprach jede Antwortmöglichkeit bzw. jedes Objekt durch Rotation dem vorgegebenen Objekt. Die Schwierigkeit lag darin, beim vorgegeben Objekt die jeweilige durchzuführende Rotation auszuwählen.

Die Ursachen zum großen Geschlechtsunterschied bei dreidimensionalen Objekten bleiben allerdings unklar. So haben sich, wie erwähnt, bereits etliche Studien mit dem Mental Rotations Test befasst und sind der Frage nachgegangen, weswegen Männer darin deutlich besser abschneiden als Frauen. Allerdings, so konstatieren Bors und Vigneau (2011): „We are still left with the problem of identifying the source of the robust sex differences on the MRT [Mental Rotations Test] and we are yet to confirm what it is actually measuring“ (S. 132).

dimensionality crossing: Dies bedeutet: „The ability to transform a spatial problem presented in two dimensions to a solution in three dimensions“ (Voyer et al., 1995, S. 262). Horan und Rosser (1984) brachten diesen Gedanken zuerst ins Spiel und konnten bei einer jedoch sehr jungen Stichprobe (4 bis 8 Jahre) nachweisen, dass männliche Testpersonen in ihrer Leistung von sich wechselnden Dimensionen profitierten. Das Konzept des „dimensionality crossing“ (Horan & Rosser, 1984, S. 408) ist jedoch nur sehr vage definiert und es ist unklar, ob es die Dimensionalität der dargestellten Objekte betrifft (d. h. den Vergleich zwischen einem zwei- und dreidimensional dargestellten Objekt) oder auch die „Dimensionalität“ des Testmaterials (d. h. die mentale Manipulation eines dreidimensionalen Objekts, dass auf einem

„zweidimensionalen“ Papier gezeichnet ist oder die tatsächliche Manipulation eines konkreten Objekts). Voyer et al. (1995) wiesen bereits darauf hin, dass der fehlende Geschlechtsunterschied in manchen Tests, die ein Wechseln der Dimensionen für die Lösung benötigten, diesem Konzept zuwiderläuft.

Die neuere Forschung griff dieses Thema allerdings wieder auf, um zu überprüfen, wie es sich mit dem Geschlechtsunterschied bei klassischen Papier-Bleistift-Verfahren im Vergleich zur Manipulation echter Objekte oder bei der Arbeit mit dreidimensionalen Projektionen verhält. Wurden die Items des Mental Rotations Test als echte Objekte (z. B. aus Holz) den Testpersonen vorgegeben, so konnten McWilliams, Hamilton und Muncer (1997) und Robert und Chevrier (2003) keine Unterschiede ermitteln, während bei Felix, Parker, Lee und Gabriel (2011) Männer weiterhin eine bessere Leistung zeigten. Laut Felix et al. (2011) war dies womöglich darauf zurückzuführen, dass es ein Zeitlimit je Item gab. Zudem konnten weitere Aspekte verantwortlich gewesen sein, die die einzelnen Items des Mental Rotations Test und die Auswertung der Daten betrafen. Nichtsdestotrotz zeigte sich in ihrer Studie, dass beide Geschlechter bei den echten Objekten mehr Items lösten. Wurden die Items (des Mental Rotations Test) als virtuelle Umgebungen am Computer erzeugt, bei denen es möglich war, die Objekte zu rotieren, anstatt sich die Rotation nur gedanklich vorzustellen (Larson et al., 1999; Parsons et al., 2004) oder wurden dreidimensionale Projektionen (Neubauer, Bergner & Schatz, 2010) verwendet, so verschwand der Geschlechtsunterschied. Neubauer et al. (2010) schlussfolgerten daraus: „The female disadvantage in solving mental rotation tasks might not lie in the process of mental rotation per se but in the derivation of a 3-dimensional representation from a 2-dimensional image“ (S. 537).

2.3. Unterschiede in der Raumvorstellung: Ausbildung

Unterschiede in der Raumvorstellung hinsichtlich der Ausbildung beziehen sich darauf, ob eine naturwissenschaftlich-technisch geprägte Ausbildung zu besseren Leistungen in

Raumvorstellungstests führt. Aus Relevanzgründen geht es hauptsächlich um die schulische und universitäre Ausbildung, z. B. durch entsprechende Schulzweige und -typen bzw. Studiengänge, da die für die Kalibrierung verwendete Stichprobe sich aus SchülerInnen verschiedener Schultypen zusammensetzte (siehe Kapitel *5.3. Stichprobe*).

Generell lassen sich Studien zu diesem Thema darin unterscheiden, ob sie eine Wechselwirkung zwischen den Ausbildungsrichtungen und dem Geschlecht der Testpersonen feststellten oder nicht. Dies bedeutet, dass Leistungsunterschiede entweder durch den alleinigen Effekt verschiedener Ausbildungen oder des Geschlechts der Testpersonen erklärt werden können, oder diese Unterschiede nur durch eine Kombination beider Effekte erklärbar sind (Kubinger et al., 2011). In den meisten Studien wurde ein Geschlechtsunterschied zugunsten der Männer festgestellt. Im Fokus steht jedoch, ob Leistungsdifferenzen neben dem Geschlecht auch auf die Art der Ausbildung zurückzuführen sind. Souvignier (2001) bezeichnete dies als „indirekte Förderung räumlicher Fähigkeiten“ (S. 301). Das bedeutet, eine Verbesserung der Raumvorstellung ist nicht primäres Ziel der Ausbildung ist, sie geht allerdings mit ihr einher.

Hinsichtlich dem Water Level Task zeigten bei Hammer, Hoffer und King (1995) Architekturstudierende bessere Leistungen als Kunststudierende, und bei Kalichman (1986) waren Studierende naturwissenschaftlicher Fächer besser als Studierende der Wirtschafts- wie Sozialwissenschaften. Jedoch ermittelten beide Studien besagten Wechselwirkungseffekt in der Form, dass Frauen *nicht*-naturwissenschaftlich-technischer Fächer die deutlich schlechtesten Testwerte aufwiesen und zudem bei Kalichman (1986) die Kombination aus männlicher Testperson und naturwissenschaftlichem Fach die höchsten Testwerte erbrachte.

Keine Interaktion entdeckten Nordvik und Amponsah (1998), in deren Arbeit Studierende naturwissenschaftlicher Fächer bessere Leistungen im Mental Rotations Test, Water Level Task, Surface Development Test sowie einem zweidimensionalen Rotationstest erlangten als Studierende sozialwissenschaftlicher Fächer. Da männliche Sozialwissenschaftsstudierende bis auf den Surface Development Test in allen anderen Tests besser waren als weib-

liche Naturwissenschaftsstudierende, wurde geschlussfolgert, dass das Geschlecht einen größeren Einfluss auf die Testleistung hat als der jeweilige Studiengang. Eine fehlende Wechselwirkung zwischen Geschlecht und Ausbildung gab es auch in den Studien von Peters et al. (1995) und Peters, Lehmann, Takahira, Takeuchi und Jordan (2006). Studierende, die einen Bachelor-of-Science-Abschluss anstrebten, demonstrierten beim Mental Rotations Test eine bessere Leistung als solche, deren geplanter Abschluss ein Bachelor of Arts war. Die Effektgröße des Unterschieds betrug bei Peters et al. (2006) $d = 0{,}65$.
Gittler und Glück (1998) versuchten, die Wirkung bestimmter Unterrichtsfächer auf die Raumvorstellung besser hervorzuheben, indem sie die SchülerInnen ihrer Stichprobe nach zwei Jahren abermals den Dreidimensionalen Würfeltest vorgaben. Während dieser Zeit hatte ein Teil der SchülerInnen Unterricht in Darstellender Geometrie. Eben jene zeigten schließlich zum zweiten Testzeitpunkt eine deutlich bessere Leistung als die SchülerInnen, die dieses Unterrichtsfach nicht besucht hatten. Gittler und Glück (1998) sahen dies als Beleg für die fördernde Wirkung dieser Beschulung auf die Raumvorstellung, zumal sich beide SchülerInnengruppen zum ersten Testzeitpunkt in ihrer Leistung nicht unterschieden hatten sowie keine Wechselwirkungen mit dem Geschlecht festgestellt werden konnten. In ihrer Studie profitierten insbesondere Frauen von dem Unterricht, so dass ebenso der Geschlechtsunterschied bei den SchülerInnen mit Darstellender Geometrie verschwand, während er beim ersten Testzeitpunkt noch vorzufinden war. Auch bei Hammer et al. (1995) und Quaiser-Pohl und Lehmann (2002) unterschieden sich naturwissenschaftlich-technisch ausgebildete Frauen nicht von den Männern, wenngleich es in diesen Studien nur einen Testzeitpunkt gab.
Der Effekt der Unterrichtung Darstellender Geometrie konnte ebenso bei Studierenden festgestellt werden, die einen entsprechenden Kurs ein Semester lang besuchten und anschließend eine bessere Raumvorstellungsleistung erbrachten (Górska, Sorby & Leopold, 1998; Górska, 2005; Leopold, Górska & Sorby, 2001). Männer und Frauen verbesserten sich hier allerdings verhältnismäßig ähnlich, so dass der Geschlechtsunterschied, im Gegensatz zu Gittler und Glück (1998), bestehen blieb. Auf die fördernde Wirkung ei-

ner naturwissenschaftlich-technischen Ausbildung auf die Raumvorstellung wiesen auch Burnett und Lane (1980) hin. Bei ihnen erbrachten Physik-, Mathematik-, Maschinenbau- und Biologiestudierende nach vier Semestern bessere Testwerte als solche der Geistes- und Sozialwissenschaften. Von besonderer Bedeutung für diese Arbeit ist noch das Ergebnis von Gittler (1990), dass Schüler (es wurden keine Schülerinnen getestet) einer naturwissenschaftlich-technisch geprägten Schule (HTL) bessere Leistungen beim Dreidimensionalen Würfeltest zeigten als SchülerInnen allgemeinbildender Schulen (AHS).
Es zeigt sich resümierend, dass die Raumvorstellung im Rahmen einer schulischen oder universitären Ausbildung mit naturwissenschaftlich-technischen Schwerpunkt gemäß Souvignier (2001) indirekt gefördert wird. Nichtsdestotrotz bleibt ungeklärt, ob und wenn, welche Rolle das Geschlecht dabei spielt. Das heißt, ob das Geschlecht oder die Ausbildung einen größeren Effekt auf die Raumvorstellung hat und ob eine Interaktion beider vorliegt. Während Gittler und Glück (1998) die Ausbildung als stärkeren Einflussfaktor auf die Raumvorstellung ansahen, so erklärte bei Peters et al. (1995) die Ausbildung einen geringeren Anteil der Varianz[6] als das Geschlecht (6,90 % zu 17,70 %).
Grundsätzliches Problem ist jedoch, dass es sich in allen Studien, wie in Kapitel *2.1. Experimente, Zufallsauswahl und Randomisierung* beschrieben und ebenso von Gittler und Glück (1998) erwähnt, um keine Experimente handelte, da keine Randomisierung stattfand. Dies bedeutet, die SchülerInnen wurden nicht zufällig den verschiedenen Gruppen (z. B.: Unterrichtsfächer, Studiengänge) zugewiesen, sondern wählten diese selbst, bzw. waren zum Zeitpunkt der Studie bereits in diesen Gruppen. Der kausale Rückschluss, dass der jeweilige Unterricht Raumvorstellung fördert, ist daher nicht zulässig. Es wäre ebenso möglich, dass Personen naturwissenschaftlich-technische Studiengänge bzw. Unterrichtsfächer wählten, eben weil sie sich selbst gut in ihrer Raumvorstellung einschätzten. Natürlich wäre weiterführend auch denkbar, dass die Auswahl bestimmter Fachgebiete mit weiteren Merkmalen einer Person einherging, die sich auf die Raumvorstellung auswirkten, und im Rahmen der erwähnten Studien nicht erfasst wurden.

[6] „Das ihm [dem Mittelwert] entsprechende Streuungsmaß, das (...) die Variabilität der beobachteten Werte beschreibt.“ (Kubinger et al., 2011, S. 97)

III.
Empirischer Teil

3. Test zur Angewandten Raumvorstellung (TARV)

3.1. Rahmenbedingungen des und Ansprüche an den Test

Rahmenbedingungen des Tests:

Die Entwicklung des TARV erfolgte im Rahmen einer laufenden Dissertation, wobei der Test in Kurzformen bereits Anwendung in den Wiener Self-Assessments Architektur und Maschinenbau findet. In dieser Anwendung soll der Test Studieninteressierten Auskunft über ihre Befähigung für diese Studiengänge im Bereich der Raumvorstellung geben, was somit auch den Geltungsbereich (Jonkisz et al., 2012) der aktuell eingesetzten Versionen des TARV darstellt. Sollte sich die TARV-Version, welche im Rahmen dieser Arbeit kalibriert wird, jedoch als ökonomischer[7] und zumutbarer[8] erweisen, ist auch von dieser Version ein späterer Einsatz in den Self-Assessments denkbar.

Die Bearbeitung des Tests ist seit 2010 über eine Online-Plattform (http://studienwahl.tuwien.ac.at) möglich, auf der Studierende Self-Assessments zu verschiedenen Studiengängen finden. Nachteilig wirkte sich diese Rahmenbedingungen für diese Arbeit in der Form aus, dass sämtliche Tests der Plattform in einem einheitlichen Bild präsentiert werden sollten und damit ein Aufgabenstamm (siehe Kapitel *3.4. Items*) lediglich eine begrenzte Größe (800 x 450 Pixel bei einer Auflösung von 72 dpi) haben durfte. Abbildung 14 zeigt einen Screenshot von der Online-Plattform.

[7] „Ein Test erfüllt das Gütekriterium *Ökonomie*, wenn er, gemessen am diagnostischen Informationsgewinn, relativ wenig Ressourcen (Zeit und Geld) beansprucht." (Kubinger, 2009, S. 98)

[8] „Ein Test erfüllt das Gütekriterium *Zumutbarkeit*, wenn er die Testpersonen absolut und relativ zu dem aus seiner Anwendung resultierenden Nutzen in zeitlicher, psychischer (insbesondere energetisch-motivationaler und emotionaler) sowie körperlicher Hinsicht schont." (Kubinger, 2009, S. 116)

Abbildung 14. Screenshot Online-Plattform.

Ansprüche an den Test:

Aus der beschriebenen Literatur ergaben sich vielfältige Ansprüche an einen Raumvorstellungstest. Die größte Schwierigkeit stellte sich hinsichtlich der Frage zur Ein- oder Multidimensionalität des Merkmals Raumvorstellung und damit in weiterer Folge auch zur Konstruktion der Items. Konkret ging es darum, als was Raumvorstellung definiert und demzufolge zu erfassen versucht wurde. Wie in Kapitel *1.2.1. Kritische Auseinandersetzung hinsichtlich mehrerer Faktoren der Raumvorstellung* beschrieben, gab es widersprüchliche Ergebnisse zu diesem Thema, die dazu führten, dass bis heute Uneinigkeit über die Anzahl an Raumvorstellungsfaktoren herrscht und über die Beziehung, in der sie

zueinander stehen. Des Weiteren spielte die Wahl der statistischen Methode eine entscheidende Rolle, mit der eine oder mehrere Dimensionen der Raumvorstellung nachzuweisen versucht wurden.

Die dargestellten Studien zur differentiellen Forschungsrichtung (siehe Kapitel *2. Raumvorstellung: Differentielle Forschungsrichtung*) zeigten, dass sich Männer und Frauen in ihrer Raumvorstellung unterschieden, dieser Unterschied jedoch stark von dem jeweils verwendeten Test und dessen Charakteristika abhing. Auch zeigte sich ein Unterschied zwischen Personen hinsichtlich ihrer Ausbildungsrichtung. Der Anspruch an einen Raumvorstellungstest musste demnach sein, a priori zu formulieren, als was Raumvorstellung im Rahmen des Tests betrachtet und zu erfassen versucht wurde. Dies stets im Hinblick darauf, eine systematische Diskriminierung von Personengruppen zu vermeiden (Kubinger, 2009), die sich aus der Konstruktion der Items ergeben hätte und dies gegebenenfalls a posteriori, durch den Ausschluss bestimmter Items, zu bewerkstelligen.

Diese Forderung stellt als Gütekriterium der *Fairness* eine generelle Forderung an jeden Test dar. Weitensfelder et al. (2010) wiesen besonders für die Raumvorstellung auf die Bedeutung dieses Kriteriums hin, da Unterschiede zwischen Personengruppen zwar durch eigene Eichtabellen für die jeweiligen Gruppen relativiert werden können, die faktischen Unterschiede im Test dennoch bestünden und bei der Eignungsdiagnostik, insbesondere bei der Auswahl der Besten, negative Konsequenzen nach sich zögen. Dies sei relevant für die Raumvorstellung, da die meisten Tests nur gemäß einem Raumvorstellungsfaktor konstruiert wurden und der Unterschied zwischen Personengruppen, wie dargestellt, je nach Faktor stark variieren konnte. Ebenso stellte sich in diesem Zusammenhang die Frage, ob das Erfassen eines Faktors genügte, um valide Aussagen zur Raumvorstellung einer Person zu treffen und womöglich auf dieser Basis die Eignung für bestimmte Berufe (z. B. PilotIn, ArchitektIn) festzustellen (Voyer et al., 1995; Weitensfelder et al., 2010).

Ein populäres Beispiel für den soeben beschriebenen Sachverhalt stellt der Eignungstest für das Medizinstudium (EMS) (Zentrum für Testentwicklung und Diagnostik (ZTD), 1995) dar. Diese Testbatterie enthält mit Schlauchfiguren einen Raumvorstellungstest,

bei dem Frauen auch im Rahmen dieser Testbatterie schlechtere Werte erzielen (Hänsgen & Spicher, 2011).

3.2. Ziele

Für ein besseres Verständnis des TARV ist vorwegzunehmen ist, dass der Test anstatt Faktoren einzelne Facetten[9] definierte. Diese stellten die notwendigen Denkoperationen zum Lösen von Items dar, ohne jedoch notwendigerweise eigene Dimensionen der Raumvorstellung zu sein (Näheres hierzu in Kapitel *3.3. Konzept*).
Anhand verschiedener Versionen des TARV wurde bisher versucht, dieses Testkonzept zu validieren, was erste Hinweise auf die Eindimensionalität erbrachte (Weitensfelder, 2011). Das heißt, dass die Facetten das gleiche Merkmal erfassten, demzufolge der Test eindimensional maß. Dabei wurde bei den Items ein sequentielles Antwortformat[10] verwendet. Dies hatte zur Folge, dass jedes Item nur gemäß einer Facette konstruiert werden konnte, bzw. ein Item maximal eine Facette erfassen konnte. Die Bearbeitung des TARV erwies sich in dieser Form allerdings als zeitaufwendig für die Testpersonen, was ihn daher unökonomisch machte.
Das vorrangige Ziel dieser Arbeit war es daher, eine neu erstellte TARV-Version mit Multiple-Choice-Antwortformat zu kalibrieren und mit einer Version mit dem herkömmlichen sequentiellen Antwortformat zu vergleichen. Durch ein Multiple-Choice-Antwortformat war es möglich, Items zu generieren, die sich aus mehreren Facetten zusammensetzten. Die so erstellten Items mussten von einer ausreichenden Menge an Testpersonen bearbeitet werden, um sie anschließend kalibrieren bzw. skalieren zu können. Dies entspricht dem Nachweis, dass der Test eindimensional, also bei allen Personen die gleiche Fähigkeit misst. Ebenso wird dadurch sichergestellt, dass sich die geschätzte Schwierig-

[9] Die Facetten von Weitensfelder (2012) sind nicht zu verwechseln mit Facetten im Rahmen eines Facettendesigns, „die die Dimensionen meinen, die bei der Itemkonstruktion miteinander gekreuzt werden“ (Rost, 2004, S. 251), z. B. Items, TestleiterInnen, Testzeitpunkte.

[10] „(...) wenn der Tp [Testperson] die vorgesehenen Antwortmöglichkeiten einzeln, nach einander vorgegeben werden, also ohne die übrigen Antwortmöglichkeiten gleichzeitig darzubieten, und die Tp dabei strikt der Reihe nach, pro Antwortmöglichkeit zu entscheiden hat, ob diese richtig oder falsch ist.“ (Kubinger, 2009, S. 140)

keit der Items zwischen verschiedenen Personengruppen (z. B. Männer und Frauen oder Personen verschiedener Ausbildungsrichtungen) nicht statistisch signifikant voneinander unterscheidet.

Zudem wurde gemäß des Testkonzepts versucht, die Konstruktvalidität[11] des TARV nachzuweisen. Das hieß, dass es sich bei der vom TARV erfassten Dimension auch um das Merkmal Raumvorstellung handeln würde, welches durch die Facetten beschrieben wurde. Somit bestand die Annahme, dass lediglich die Facetten die zugrunde liegenden Denkoperationen zum Lösen der Items darstellen würden. Dies wird allgemein als Itemkomponenten- und genauer als Schwierigkeitsmodell beschrieben, was bedeutet, dass die Itemmerkmale bzw. -komponenten (Facetten) die Itemschwierigkeit festlegen (Rost, 2004). Der Nachweis der Konstruktvalidität bedeutete daher für den TARV, dass zum Lösen eines Items diejenigen mentalen Operation notwendig waren, die durch die Facetten bei jedem Item beschrieben wurden (Rost, 2004). Da die Facetten verschiedene mentale Operationen der Raumvorstellung darstellten (z.B. mentales Rotieren), würde der TARV demnach das Merkmal Raumvorstellung erfassen.

Als Resultat der Kritik an den Methoden, mit denen Raumvorstellungsfaktoren ermittelt wurden (siehe Kapitel *1.2.1. Kritische Auseinandersetzung hinsichtlich mehrerer Faktoren der Raumvorstellung*), geschah die Kalibrierung und Überprüfung der Konstruktvalidität mittels spezieller theoriebildender Verfahren (siehe Kapitel *5.4. Theoriebildende Verfahren*).

Ebenso wurde geprüft, ob sich die verschiedenen Personengruppen in ihrer tatsächlichen Raumvorstellungsleistung voneinander unterschieden. Neben Personen verschiedener Ausbildungsrichtungen stand insbesondere das Geschlecht im Fokus, da ersichtlich wurde, dass der Unterschied je nach Test und je nachdem, welchen Raumvorstellungsfaktor dieser angab zu erfassen, verschieden groß sein konnte (siehe Kapitel *2. Raumvorstellung: Differentielle Forschungsrichtung*). Diesen Unterschied versuchte der TARV durch seinen

[11] „(...) wenn der Rückschluss vom Verhalten der Testperson innerhalb der Testsituation auf zugrunde liegende psychologische Persönlichkeitsmerkmale (»Konstrukte«, »latente Variablen«, »Traits«) wie Fähigkeiten, Dispositionen, Charakterzüge, Einstellungen wissenschaftlich fundiert ist.“ (Moosbrugger & Kelava, 2012, S. 16)

facettenorientierten Ansatz zu nivellieren. Das hieß, dadurch dass die Items mehrere Facetten beinhalteten, die letztendlich gemeinsam das eindimensionale Merkmal Raumvorstellung widerspiegeln sollten, wurde versucht, mehr Fairness zu erreichen und separate Eichtabellen für die Gruppen zu vermeiden.
Ein zusätzliches Ziel im Rahmen der Kalibrierung war der Vergleich von Items der Versionen mit Multiple-Choice und sequentiellem Antwortformat. Einerseits wurden die Bearbeitungszeiten verglichen, da die Multiple-Choice-Version darauf abzielte, geringere Bearbeitungszeiten je Item zu benötigen. Andererseits sollten mögliche Einflüsse des Antwortformats auf die Itemschwierigkeit aufgedeckt werden.
Zusammengefasst ergaben sich also folgende Ziele:

Skalierung:

- Kalibrierung der Items mit Multiple-Choice-Antwortformat in der Form, dass sie sich in ihrer geschätzten Schwierigkeit zwischen verschiedenen Personengruppen nicht unterscheiden. Damit ergibt sich der Nachweis, dass der Test bei allen Personen die gleiche Fähigkeit erfasst bzw. eindimensional misst.

Konstruktvalidität:

- Nachweis, dass die TARV-Version mit Multiple-Choice-Antwortformat eindimensional misst (siehe Skalierung).
- Nachweis, dass es sich dabei um das Merkmal Raumvorstellung handelt, das sich aus den Facetten zusammensetzt. Dabei stellen die Facetten die zugrunde liegenden Denkoperationen zum Lösen der Items dar.

Fairness:

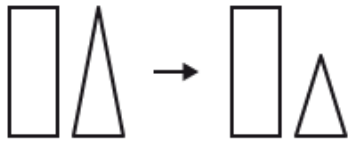
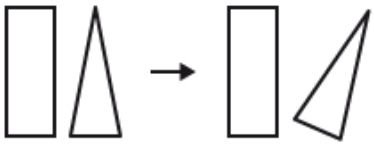
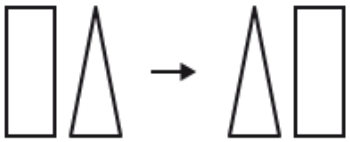
- Kalibrierung der Items mit Multiple-Choice-Antwortformat (siehe Skalierung).
- Überprüfung, ob es statistisch signifikante Unterschiede zwischen verschiedenen Personengruppen in ihrer Raumvorstellung gibt.

Vergleich der TARV-Versionen mit Multiple-Choice und sequentiellem Antwortformat dahingehend, welche Rolle das Antwortformat für die Itemschwierigkeit spielt und welche Unterschiede sich in den durchschnittlichen Bearbeitungszeiten je Item ergeben.

3.3. Konzept

Weitensfelder (2011) formulierte als Ziel einen Test, der mehrere, klar voneinander abgrenzbare Facetten der Raumvorstellung beinhaltete, womit der Test auch die Frage zur Ein- oder Multidimensionalität des Merkmals Raumvorstellung aufgriff und versuchte, sie durch seinen facettenorientierten Ansatz aus einem anderen Blickwinkel zu betrachten. Dabei wies er Ähnlichkeiten zu den erwähnten Ideen von Stumpf und Eliot (1995) und Gittler und Arendasy (2003) auf, die von einem hierarchischen Modell der Raumvorstellung ausgingen. Der TARV ging davon aus, dass die Facetten, die Ähnlichkeiten zu den in Kapitel *1.2. Raumvorstellung als ein multidimensionales Merkmal* erwähnten Faktoren aufwiesen, möglicherweise keine eigenen Dimensionen der Raumvorstellung darstellten. Eine Facette meinte demzufolge die gedanklichen Operationen, die zum Lösen eines Items notwendig waren. Sie bestimmten demnach nur die Schwierigkeit zur Lösung eines Items, ohne gleich eigenständige Dimensionen der Raumvorstellung darzustellen. Unterstützt wurde diese Annahme dadurch, dass erste Versionen des TARV auf die Eindimensionalität des Merkmals Raumvorstellung hinwiesen (Weitensfelder, 2011).
Tabelle 3 liefert einen Überblick über die von Weitensfelder (2012) formulierten Facetten. Zur besseren Verständlichkeit werden sie zudem in schematischen Abbildungen veranschaulicht, die jedoch noch nicht die endgültigen Items darstellen. Ebenso wird beschrieben, wie die jeweilige Facette als Item realisiert wird.
Ein wesentliches Merkmal des TARV mit Multiple-Choice Format war der Versuch, diese drei Facetten mit Items zu erfassen, die sich in ihrer Darstellung ähnelten. Das hieß, es wurden keine separaten marker tests für jede Facette erzeugt, stattdessen sollte ein Item auch mehrere Facetten erfassen können. Der Nachteil des TARV mit sequentiel-

Tabelle 3
Testkonzept: Test zur Angewandten Raumvorstellung (TARV)

Relationen	Rotation	Orientierung
Erkennen von Relationen, insbesondere Größen- und Abstandverhältnisse	Durchführen mentaler Rotationen	Einschätzen der Positionen verschiedener Objekte zueinander
 *Abbildung 12.* Facette Relationen: Größenverhältnisse.	 *Abbildung 13.* Facette Rotation: Winkelverhältnisse.	 *Abbildung 14.* Facette Orientierung: Objektpositionen.
 *Abbildung 15.* Facette Relationen: Abstände.		
Items beinhalten Fehler mit falschen Größen und/oder Abständen der Objekte zueinander.	Items beinhalten Fehler mit fehlerhaften Winkeln der Objekte zueinander.	Items beinhalten Fehler mit Richtungs- bzw. Positionsveränderungen der Objekte zueinander (z. B. Links-Rechts-Verwechslungen) sowie falschen Seitenangaben zu den Objekten.

lem Antwortformat, keine Facettenkombination pro Item zu ermöglichen, sollte so behoben werden. Auch war dies im Sinne der hier rezitierten Forschungsergebnisse, die die Abhängigkeit der ermittelten Faktorenstrukturen von den jeweiligen marker tests kritisierten.

Des Weiteren hatten Testpersonen beim TARV mit zwei- sowie dreidimensionalen Objekten zu tun und mussten von dem einen auf das andere schließen und vice versa. Laut (Weitensfelder et al., 2010) war dies z. B. relevant, um Pläne und Skizzen erstellen bzw.

lesen zu können.

3.4. Items

Gemäß Jonkisz et al. (2012) untergliedert sich ein Item in einen *Aufgabenstamm* sowie in das *Antwortformat*, wobei Ersteres die zu lösende Problemstellung meint und Letzteres die Art und Weise, mit der diese Lösung gegeben werden kann. Wie bereits erwähnt, war ein Ziel dieser Arbeit, eine neu erstellte Multiple-Choice-Version des TARV zu kalibrieren. Diese Version wurde unter anderem auch mit einer bereits existierenden Version mit sequentiellem Antwortformat verglichen. Es ergab sich daher für beide TARV-Versionen auch ein unterschiedlicher Aufgabenstamm. *MC* steht dabei im Folgenden für die Version mit Antwortformat Multiple-Choice und *SQ* für diejenige mit sequentiellem Antwortformat.

3.4.1. Aufgabenstamm & Antwortformat (SQ)

Vorab sei zur SQ-Version erwähnt, dass diese bereits existierte und für die Datenerhebung dieser Arbeit lediglich die Darstellung der Items mit den neu erstellten Items der MC-Version vereinheitlicht wurde (z. B. gleich dargestellte Überschriften und Trennlinien).
Die Problemstellung des Aufgabenstamms der SQ-Version bestand aus den Bedingungen „Darstellungsmodus“ sowie „2 x richtig-oder-falsch“. Der Darstellungsmodus gliederte sich wiederum in die Unterkategorien „2D auf 3D“ und „3D auf 2D“.

Darstellungsmodus: Die Testpersonen erhielten zweidimensionale Planansichten sowie dreidimensionale Gebildeansichten von Objekten, die sie miteinander verglichen. Eine Ansicht war dabei vorgegeben bzw. „korrekt“, für die andere Ansicht musste überprüft werden, ob sie der vorgegeben entsprach. Dazu war es notwendig, von der vorgegeben zuerst auf die zu überprüfende zu schließen. Als mögliche Ansichten fungierten die Vorder-, Rück-, Ober-, Unter- sowie linke und rechte Seite.

2D auf 3D: Testpersonen mussten von *zwei* zweidimensionalen Planansichten ei-

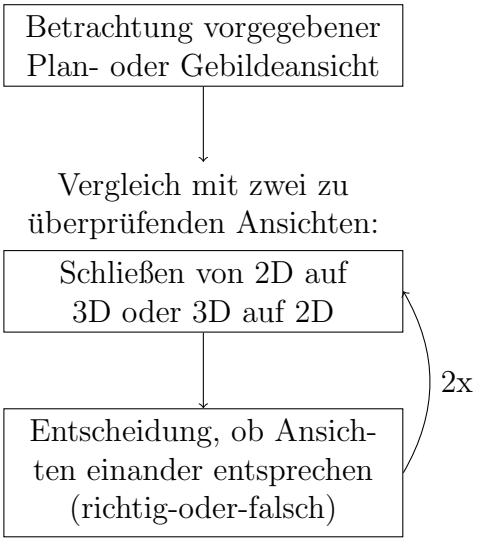


Abbildung 19. TARV, SQ: schematische Darstellung eines Items.

nes Objekts auf *eine* dreidimensionale Gebildeansicht dieses Objekts schließen.

3D auf 2D: Testpersonen mussten von *einer* dreidimensionalen Gebildeansicht eines Objekts auf *eine* zweidimensionale Planansicht des Objekts schließen.

2 x richtig-oder-falsch: Ein Item bestand aus *einer* vorgegebenen sowie *zwei* zu überprüfenden Ansichten, wobei der Darstellungsmodus (2D auf 3D und umgekehrt) ident blieb. Testpersonen mussten sich für jede einzelne zu überprüfende Ansicht entscheiden, ob sie der vorgegebenen entsprach. Ein Item war dabei nur gemäß *einer* Facette konstruiert. Das hieß, wenn die Planansicht und Gebildeansicht einander nicht entsprachen, so konnten Objekte die falsche Größe *und/oder* den falschen Abstand zueinander haben (Facette Relationen) *oder* im falschen Winkel zueinander stehen (Facette Rotation) *oder* sich auf den falschen Positionen befinden (Facette Orientierung).

Abbildung 19 zeigt den schematischen Aufbau eines Items der SQ-Version. Die Testperson erhielt dabei zuerst die vorgegebene Plan- oder Gebildeansicht ohne eine zu überprüfende Ansicht, um sich mit dem Item vertraut zu machen. Anschließend folgten zwei zu überprüfende Ansichten mit den soeben genannten Problemstellungen. Einen

konkreten Aufgabenstamm zeigt Abbildung 20. Bei diesem Item musste von zwei vorgegebenen zweidimensionalen Planansichten (linke Seite) auf eine zu überprüfende dreidimensionale Gebildeansicht (rechte Seite) geschlossen werden. Diese Gebildeansicht war in Bezug auf die Planansichten in diesem Beispiel falsch. Die Position eines Objekts im Gebilde war nicht korrekt (Facette Orientierung).

Hinsichtlich des Antwortformats handelte es sich, wie bereits erwähnt, um ein sequentielles Antwortformat. Die Testpersonen mussten nacheinander je zwei Plan- und Gebildeansichten vergleichen und bewerten, ob diese einander entsprachen. Konkret hatten sie folgende Antwortmöglichkeiten: „In Bezug auf das Gebilde ist der vorgegebene Plan richtig/falsch." bzw. „In Bezug auf die Pläne ist das vorgegebene Gebilde richtig/falsch". Mit „vorgegebenen Plan/Gebilde" waren hier der zu überprüfende Plan bzw. das zu überprüfende Gebilde gemeint. Ein Item galt als gelöst, wenn beide Darstellungen korrekt beantwortet wurden. Damit ergab sich eine a priori-Ratewahrscheinlichkeit für die Testpersonen, das hieß, das Item nur zufällig zu lösen, von $\frac{1}{4}$.

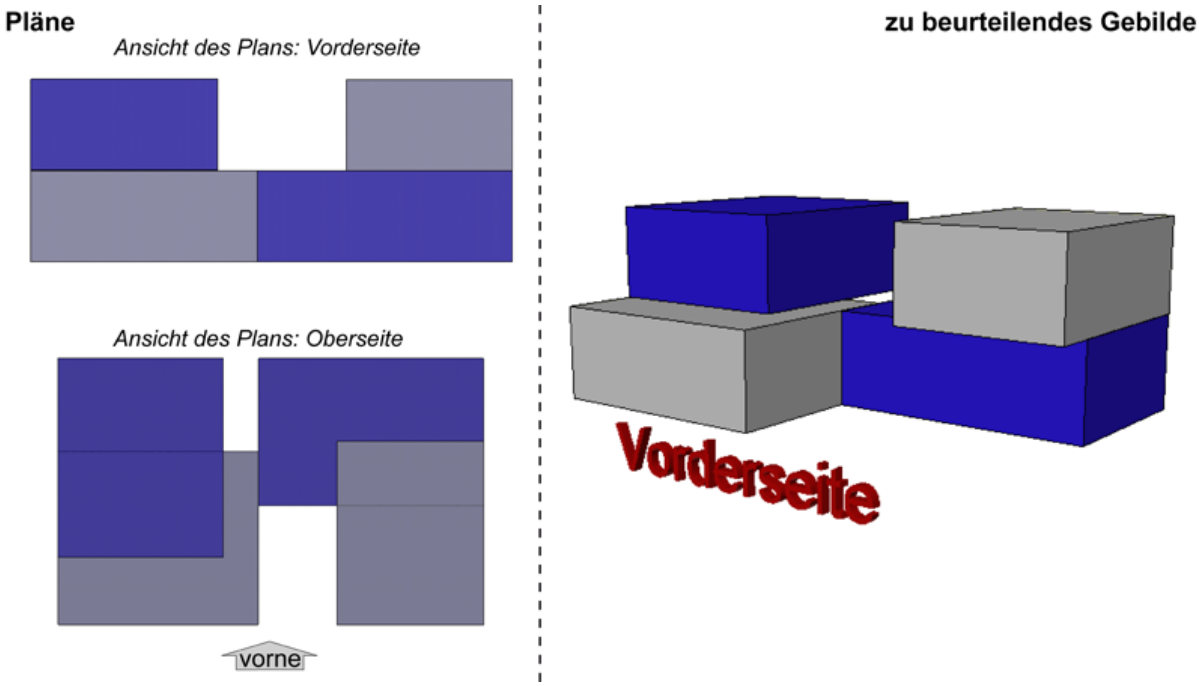


Abbildung 20. TARV, SQ: Aufgabenstamm (Facette Orientierung, 2D auf 3D).

3.4.2. Aufgabenstamm & Antwortformat (MC)

Ziel dieser Arbeit war es, die Items einer neu erstellten TARV-Version mit Multiple-Choice-Antwortformat zu kalibrieren. Bereits bestehende Items mit sequentiellem Antwortformat wurden dazu in ein Multiple-Choice-Antwortformat überführt sowie neue Items konstruiert. Neue Items wurden mit den Programmen Anim8or (Version 0.95c; Glanville, 2007) erstellt, sowie für eine einheitliche Darstellung mit Adobe Photoshop CS3 (Version 10.0) und Adobe Illustrator CS3 (Version 13.0.0) bearbeitet.
Die Problemstellung des Aufgabenstamms der MC-Version bestand aus zwei Bedingungen. Zum einen die Bedingung „Darstellungsmodus“, die vollkommen identisch zur SQ-Version ist und sich in „2D auf 3D“ bzw. „3D auf 2D“ untergliederte, zum anderen gab es die Bedingung „Fehleridentifikation“.

Darstellungsmodus:

2D auf 3D: Testpersonen mussten von *zwei* zweidimensionalen Planansichten eines Objekts auf *eine* dreidimensionale Gebildeansicht dieses Objekts schließen.

3D auf 2D: Testpersonen mussten von *einer* dreidimensionalen Gebildeansicht eines Objekts auf *eine* zweidimensionale Planansicht des Objekts schließen.

Fehleridentifikation: Ein Item bestand aus *einer* vorgegebenen sowie *einer* zu überprüfenden Ansicht. In der vorgegebenen Ansicht waren einzelne Objekte mit Buchstaben markiert, für die festzustellen war, ob sie den ihnen korrespondieren Objekten in der zu überprüfenden Ansicht entsprachen. Dabei stellten die Antwortmöglichkeiten Vorschläge dar, was der oder die Fehler in der zu überprüfenden Ansicht sein konnte/n und die Testpersonen mussten bewerten, ob dies zutraf.

Es gab zwei Unterschiede zur SQ-Version. Zum einen beinhaltete jede zu überprüfende Ansicht immer mindestens einen Fehler, worüber die Testpersonen informiert waren. Das hieß, im Gegensatz zur SQ-Version gab es kein Item, in dem beide Ansichten tatsächlich einander entsprachen. Zum anderen konnte ein Item Fehler einer bis allen dreien Facetten beinhalten. Das hieß, die Objekte konnten die

falsche Größe *und/oder* den falschen Abstand zueinander haben (Facette Relationen) *und/oder* im falschen Winkel zueinander stehen (Facette Rotation) *und/oder* sich auf den falschen Positionen befinden (Facette Orientierung).

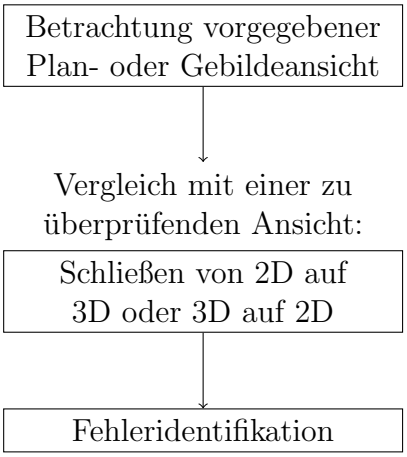


Abbildung 21. TARV, MC: schematische Darstellung eines Items.

Den schematischen Aufbau eines Items der MC-Version zeigt Abbildung 21. Die Testperson erhielt auch hier zuerst die vorgegebene Plan- oder Gebildeansicht ohne eine zu überprüfende Ansicht, um sich mit dem Item vertraut zu machen. Anschließend folgte, der Problemstellung entsprechend, eine zu überprüfende Ansicht. Abbildung 22 zeigt einen Aufgabenstamm. Testpersonen mussten hier von einer vorgegeben Gebildeansicht (linke Seite), in der einzelne Objekte markiert waren, auf eine zu überprüfende Planansicht (rechte Seite) schließen und die korrekten Fehler in der Planansicht identifizieren, die in den Antwortmöglichkeiten gegeben waren. In diesem Beispiel war in der Planansicht die Größe von Objekt C falsch dargestellt (Facette Relationen). Der Vollständigkeit halber zeigt Abbildung 23 einen Aufgabenstamm für die noch fehlende Facette Rotation. In der zu überprüfenden Ansicht war dabei die Neigung von Objekt B falsch dargestellt.

Das Antwortformat war das bereits erwähnte Multiple-Choice-Format. Pro Item gab es vier Antwortmöglichkeiten, die jeweils auf verschiedene mögliche Fehler der Objekte hinwiesen, von denen die Testpersonen den bzw. die für sie zutreffenden auswählten. Bei der

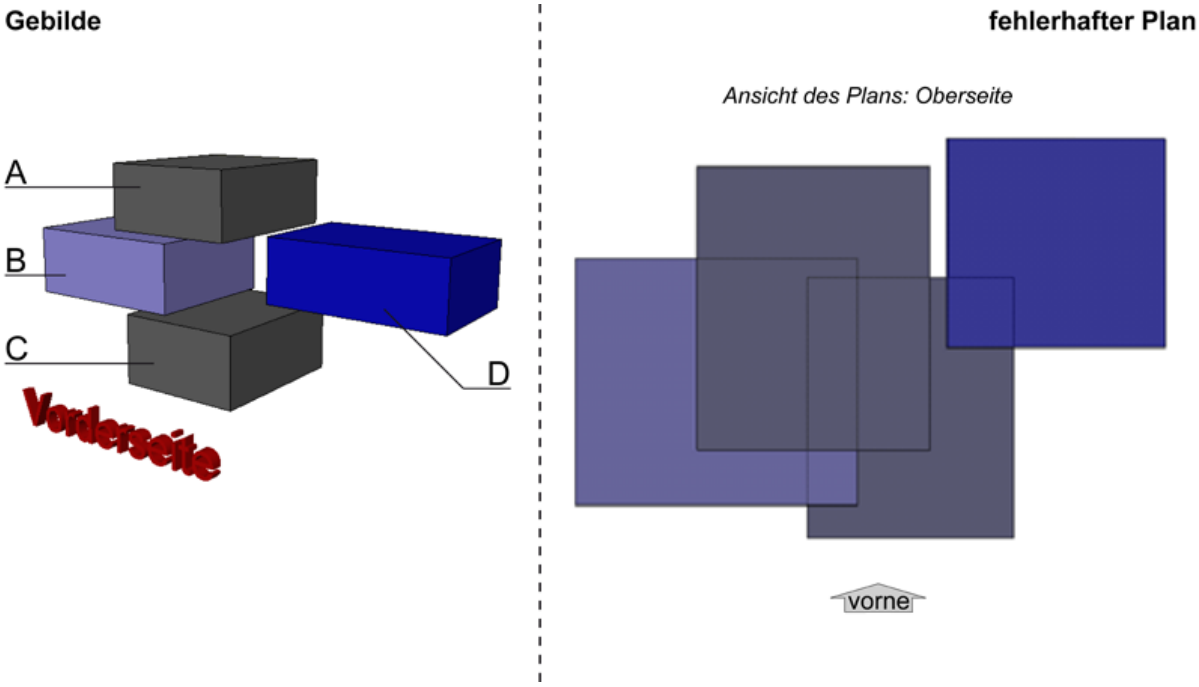


Abbildung 22. TARV, MC: Aufgabenstamm (Facette Relationen, 3D auf 2D).

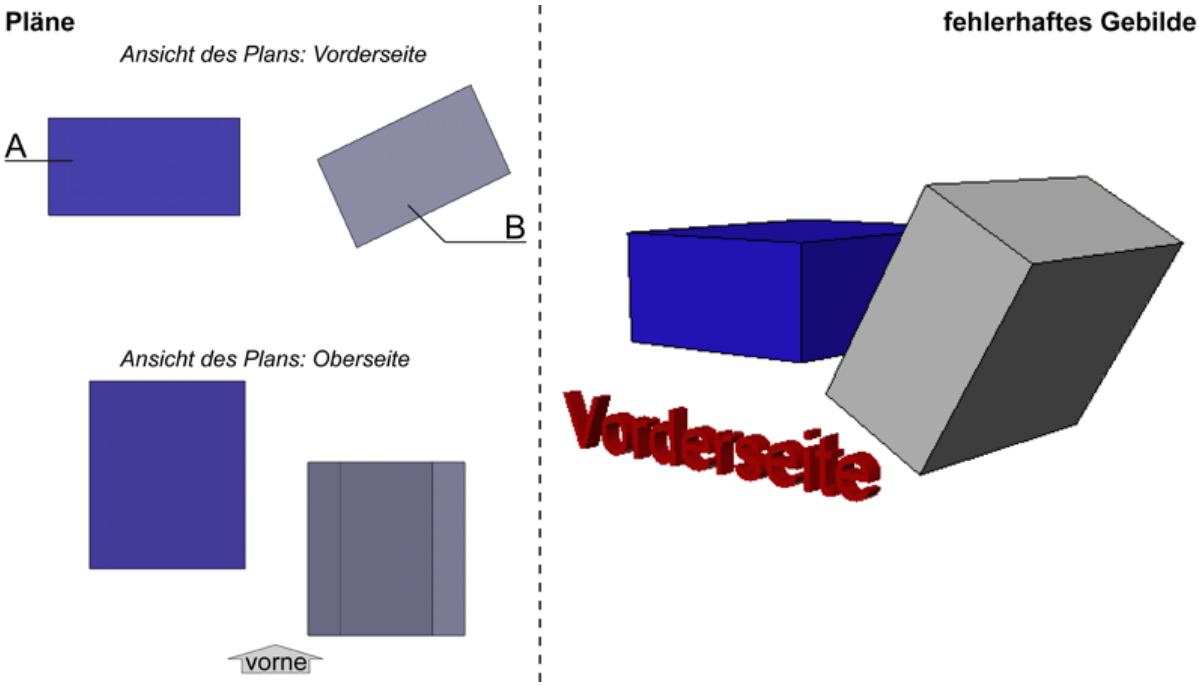


Abbildung 23. TARV, MC: Aufgabenstamm (Facette Rotation, 2D auf 3D).

Konstruktion neuer Items oder Überführung von Items der SQ- in die MC-Version zeigte sich, dass es nicht möglich war, Items mit vier korrekten Antwortmöglichkeiten zu erstellen, da so bei manchen Items verschiedene Kombinationen der Antwortmöglichkeiten zu einer Lösung geführt hätten. Die Disjunktheit der Antwortmöglichkeiten wäre somit nicht gewährleistet gewesen, da sie sich nicht gegenseitig ausgeschlossen hätten (Jonkisz et al., 2012). Das bedeutete folglich, dass eine bis maximal drei der vier gelisteten Antwortmöglichkeiten korrekt sein konnten, wobei die Testpersonen über die Anzahl der Fehler je Item nicht informiert wurden. Die a priori-Ratewahrscheinlichkeit war demnach vernachlässigbar klein.

Die Antwortmöglichkeiten selbst waren sprachlich formuliert und damit keine Grafiken, wie dies in Raumvorstellungstest üblich ist. Auf potentiell negative Folgen dadurch wird in Kapitel *3.8. Mögliche Probleme hinsichtlich der Kalibrierung* eingegangen. Die Entscheidung zu sprachlich formulierten Antwortmöglichkeiten war auf die Rahmenbedingung zurückzuführen, die nur eine Grafik und zwar für den Aufgabenstamm in einer Größe von 800 x 450 Pixel erlaubten. Die Antwortmöglichkeiten gemeinsam mit dem Aufgabenstamm in einer Grafik dieser Größe zu präsentieren, hätte die einzelnen Ansichten und Objekte zu klein und schwer erkennbar gemacht.

Hinsichtlich der Antwortmöglichkeiten wurde allerdings versucht, sie möglichst einheitlich zu gestalten. Fehler bezüglich der einzelnen Facetten wurden mit folgenden Antwortmöglichkeiten beschrieben:

Facette Relationen:

- Im Plan/Gebilde ist die Größe von Teil X (und Teil Y jeweils) falsch dargestellt.
- Im Plan/Gebilde ist der Abstand zwischen Teil X und Teil Y falsch dargestellt.

Facette Rotation:

- Im Plan/Gebilde ist der Winkel zwischen Teil X und Teil Y falsch dargestellt.
- Im Plan/Gebilde ist die Neigung von Teil X falsch dargestellt.
- Im Plan/Gebilde ist Teil X falsch gedreht dargestellt. (Nur ein Item wies diese

Antwortmöglichkeit auf, da die darüber genannten dieser Facette den Fehler nicht eindeutig identifiziert hätten.)

Facette Orientierung:

- Im Plan/Gebilde ist die Position von Teil X (zu Teil Y) falsch dargestellt.
- Der Plan/Das Gebilde zeigt nicht die ...-Seite sondern die ...-Seite.

Um sicherzustellen, dass die Items auch lediglich die gewünschten Facetten ansprachen, enthielt jedes Items nur Antwortmöglichkeiten derjenigen Facette/n, die zur Identifizierung des oder der Fehler notwendig war/en. Das hieß, wenn beispielsweise der Aufgabenstamm eines Items Objekte beinhaltete, die die falsche Größe hatten (Facette Relationen) sowie im falschen Winkel zueinander standen (Facette Rotationen), so standen nur Antwortmöglichkeiten dieser beiden Facetten (Relationen und Orientierung) zur Auswahl, nicht jedoch welche der Facette Orientierung.
Abbildung 24 zeigt zum besseren Verständnis ein vollständiges Item der MC-Version, das Fehler bezüglich aller drei Facetten beinhaltet. Die Lösungen waren hier die erste (Facette Orientierung), dritte (Facette Relationen) und vierte (Facette Rotation) Antwortmöglichkeit.

3.4.3. Items: Überblick

Tabelle 4 gibt einen Überblick über sämtliche Items und ihre Problemstellungen, die von den Testpersonen bearbeitet wurden. Insgesamt wurden 16 Items der MC-Version des TARV vorgegeben, ebenso wurde auf 14 Items der SQ-Version zurückgegriffen. Sowohl aufgrund einer begrenzten Gesamtbearbeitungszeit als auch der Zumutbarkeit war es nicht möglich, alle Items einer entsprechenden Anzahl an Testpersonen (ca. 200) vorzugeben, um sie kalibrieren zu können. Von den insgesamt 30 Items wurden daher 18 Items (MC1 bis MC14 und SQ1 bis SQ4) ausgewählt, die von ca. 200 Testpersonen bearbeitet wurden, während die restlichen 12 Items (MC15, MC16 und SQ5 bis SQ14) ca. 100 Testpersonen vorgegeben wurden. Auf Konsequenzen daraus bezüglich des Testdesigns wird

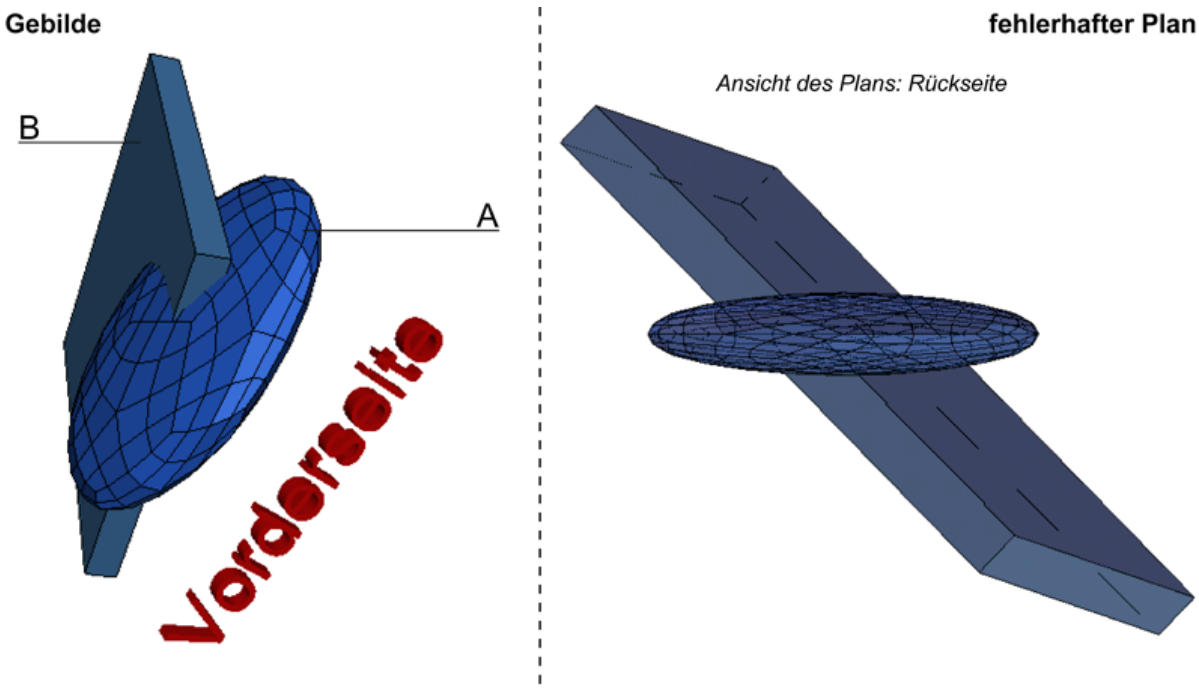


Abbildung 24. TARV, MC: Beispielitem (Facette Relationen, Rotation & Orientierung, 3D auf 2D).

in Kapitel *5.1. Testdesign* eingegangen.

Bei den von ca. 200 Testpersonen bearbeiteten 18 Items (MC1 bis MC14 und SQ1 bis SQ4) wurde in erster Linie darauf geachtet, dass die zu kalibrierenden Items der MC-Version sämtliche möglichen Kombinationen der Problemstellungen abdeckten. Da dies mit 14 MC-Items bewerkstelligt werden konnte, wurden MC15 und MC16 von der Kalibrierung mit 200 Testpersonen ausgeschlossen. Stattdessen wurden vier Items der SQ-Version verwendet, um für einen Vergleich mit der MC-Version einerseits Items der SQ-Version zu haben, die von einer ähnlichen Anzahl an Testpersonen bearbeitet wurden, sowie andererseits sicherzustellen, dass die einzelnen Testhefte auch mit Items der SQ-Version untereinander verknüpft waren (siehe Kapitel *5.1. Testdesign*). Bei der Auswahl der vier Items der SQ-Version wurde versucht, dass Items aller drei Facetten wenigstens ein Mal

vorgegeben wurden.

3.5. Instruktion

Die Testpersonen wurden in der Instruktion zuerst über die generelle Unterscheidung zwischen dreidimensionalen Gebilden und zweidimensionalen Plänen informiert und bearbeiteten anschließend *insgesamt* drei Beispielitems zu den beiden Darstellungsmodi (2D auf 3D, 3D auf 2D). Jedes Beispielitem war gemäß einer Facette konstruiert, womit die Testpersonen alle drei Facetten im Rahmen der Instruktion kennenlernten.
Aufgrund des Testdesigns erhielten die Testpersonen diese Instruktion doppelt, nämlich für die MC- sowie die SQ-Version des TARVS mit entsprechenden Änderungen bei den Beispielitems sowie den Erklärungen zum Lösen der Items (Fehleridentifikation oder 2x richtig-oder-falsch). Der Aufgabenstamm des Beispielitems für die Facette Rotation war bei der MC- wie SQ-Version der gleiche, die restlichen Aufgabenstämme unterschieden sich zwischen den Versionen.

3.6. Ablauf

Um Positionseffekte[12], je nachdem, ob die Testpersonen zuerst die MC- oder SQ-Version bearbeiteten, zu vermeiden, existierten zwei verschiedene Testbedingungen. In der ersten erhielten die Testpersonen zuerst die MC-Version und anschließend die SQ-Version, in der zweiten war es umgekehrt. Abbildung 25 veranschaulicht den Ablauf beider Testversionen.

3.7. Vergleich der Facetten des TARV mit den Raumvorstellungsfaktoren

Die Facetten sind laut Weitensfelder (2012) zumindest in der Bezeichnung angelehnt an die Raumvorstellungsfaktoren von Thurstone (1949, 1950) und Lohman (1979, 1988),

[12] „Veränderung der Schwierigkeit oder anderer Merkmale eines Items [hier: der jeweiligen Testversion] infolge seiner Platzierung im Test [hier: im Testdesign].“ (Rost, 2004, S. 69)

Tabelle 4
TARV: Überblick Problemstellungen je Item der MC- und SQ-Version

	Facette			Darstellungsmodus	
Item	Relationen	Rotation	Orientierung	2D auf 3D	3D auf 2D
MC1	X				X
MC2		X			X
MC3			X		X
MC4	X	X			X
MC5	X		X		X
MC6		X	X		X
MC7	X	X	X		X
MC8	X			X	
MC9		X		X	
MC10			X	X	
MC11	X	X		X	
MC12	X		X	X	
MC13		X	X	X	
MC14	X	X	X	X	
SQ1	X				X
SQ2		X			X
SQ3			X	X	
SQ4		X		X	
MC15	X				X
MC16			X	X	
SQ5	X				X
SQ6	X				X
SQ7		X			X
SQ8			X		X
SQ9			X		X
SQ10			X		X
SQ11	X			X	
SQ12	X			X	
SQ13		X		X	
SQ14			X	X	

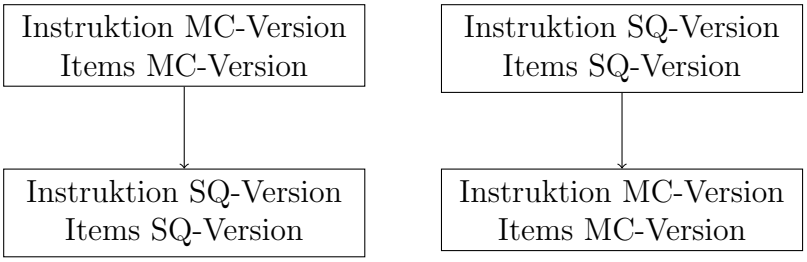


Abbildung 25. TARV: Ablauf MC- und SQ-Version.

wenngleich sich ebenso Ähnlichkeiten zu den sonstigen Raumvorstellungsfaktoren fanden, die in Kapitel *1.2. Raumvorstellung als ein multidimensionales Merkmal* beschrieben waren. Die Faktoren konnten dabei einzelnen Facetten zugeordnet werden, fanden sich aber teilweise auch in allen Items, unabhängig der Facetten, wieder.

Die Facette Relationen hatte bisher als Faktor keine vollkommene Entsprechung in der Literatur. Der bisher nur wenig nachgewiesene Faktor Length Estimation ähnelte der Anforderung dieser Facette, fehlerhafte Abstände von Objekten zueinander zu erkennen und im weitesten Sinne auch, falsche Größenverhältnisse zu erkennen, wenn angenommen wurde, dass sich mit verändernder Größe eines Objekts auch dessen dargestellten Seitenlängen änderten.

Die Facette Rotation ähnelte dem Faktor Spatial Relations in der Form, dass zur Lösung von Items die mentale Rotation von Objekten notwendig war. Der TARV machte jedoch keine Annahmen hinsichtlich der Komplexität von Items, da für den Faktor Spatial Relations die geringe Komplexität der Items ausschlaggebend ist, um ihn vom Faktor Visualization zu unterscheiden. Stattdessen wurde Rotation unabhängig der Itemkomplexität als eine für sich stehende mentale Operation gesehen. Der allgemeine Kritikpunkt jedoch, dass, anstatt mental zu rotieren, auch ein Perspektivenwechsel durchgeführt werden konnte (siehe Kapitel *1.2.1. Kritische Auseinandersetzung hinsichtlich mehrerer Faktoren der Raumvorstellung*), war auch für die Items der Facette Rotation nicht von der Hand zu weisen. Dies ist im TARV allerdings erschwert, dass bei dieser Facette als Rotation oft die Winkelverhältnisse von Objekten zueinander betrachtet werden und nicht ausschließlich nur einzelne oder isolierte Objekte rotiert werden müssen.

Die Facette Orientierung wies Gemeinsamkeiten mit den Faktor Spatial Orientation and Perception sowie dem Kinesthetic factor auf. Beim erstgenannten Faktor wurden insbesondere seine beide Aspekte, nämlich der Perspektivenwechsel sowie das Erkennen der korrekten Position von Objekten und ihrer räumlichen Beziehung zueinander, durch die Facette Orientierung angesprochen. Das Wechseln der Perspektive anstatt der Rotation der Objekte wurde im TARV dadurch forciert, dass sich sämtliche Seitenangaben auf die Vorderseite bezogen. Das hieß, die Angabe „Rückseite“ in der zu überprüfenden Ansicht in z. B. Abbildung 24 meinte die Rückseite der angegebenen „Vorderseite“ der vorgegebenen Ansicht. Während dies in diesem Beispiel trivial erscheint, da die Vorderseite gekennzeichnet ist, so war z. B. für Item MC6 (Abbildung 26) in der vorgegebenen Ansicht die Rückseite und in der zu überprüfenden Ansicht die linke Seite beschriftet. Testpersonen mussten daher stets in Bezug zur Vorderseite überprüfen, ob die Positionen der Objekte aus einer anderen Seitenansicht korrekt dargestellt waren. Nichtsdestotrotz war auch hier nicht ausgeschlossen, dass Testpersonen die Objekte rotieren würden, anstatt ihre Perspektive zu wechseln, was die Abgrenzung dieser Facette von der Facette Rotation erschweren kann. Des Weiteren kam hinzu, dass es bei allen Items im Zuge des Schließens von 2D auf 3D und umgekehrt zunächst notwendig war, die jeweilige Seite bezüglich der Vorderseite zu bestimmen, was die Differenzierung zusätzlich erschweren kann. In Kapitel *3.8. Mögliche Probleme hinsichtlich der Kalibrierung* wird darauf näher eingegangen.

Hinsichtlich der Positionsüberprüfung einzelner Objekte und ebenso dem generellen Erkennen, was z. B. die linke Seite der Vorderseite war, wenn nur die Rückseite gekennzeichnet ist, zeigten sich auch Ähnlichkeiten mit dem Kinesthetic factor. In Kapitel *1.2.1. Kritische Auseinandersetzung hinsichtlich mehrerer Faktoren der Raumvorstellung* wurde bereits darauf hingewiesen, dass dieser Faktor, insbesondere bei komplexen Items, im Faktor Spatial Orientation and Perception subsumiert sein könnte. Zentraler Bestandteil dieses Faktors ist jedenfalls, vom eigenen Standpunkt aus zwischen links und rechts differenzieren zu können. Im TARV war dies aus zweierlei Hinsicht relevant. Zum einen, wie

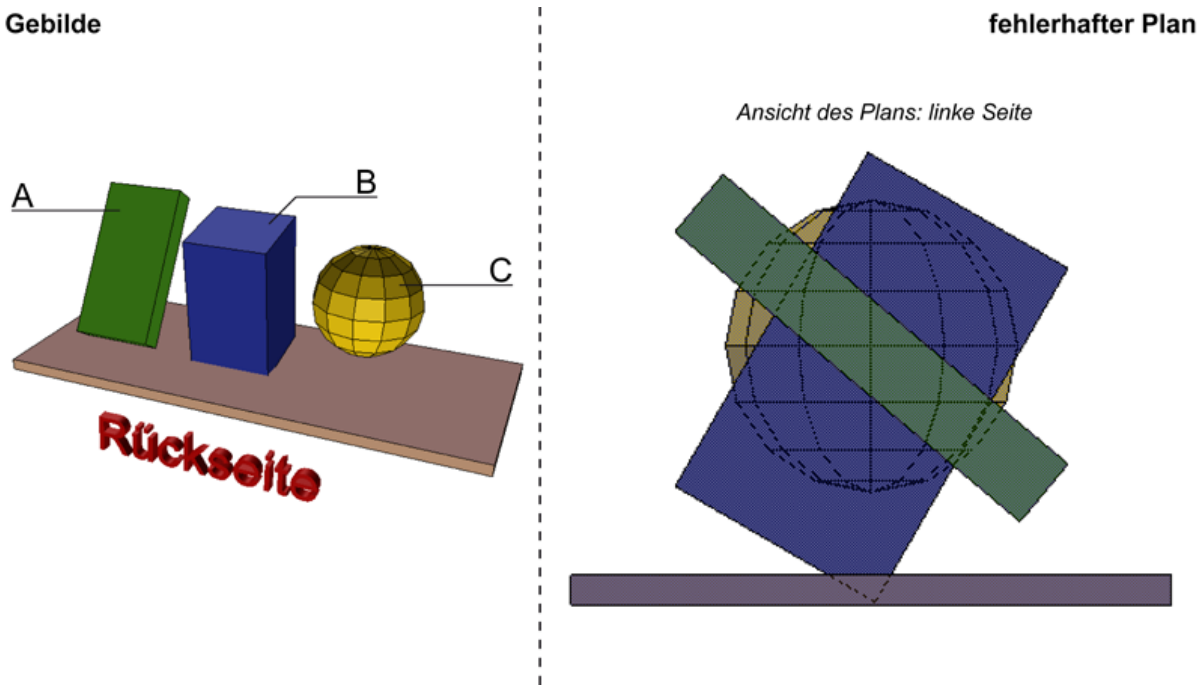


Abbildung 26. TARV: Aufgabenstamm Item MC6.

bereits erwähnt, um die korrekte Seitenansicht in Bezug zur Vorderseite zu identifizieren. Fehler in den Items der Facette Orientierung bezogen sich nicht nur auf falsche Positionen der Objekte zueinander, sondern ebenso auch auf eine falsch angegebene Seitenansicht. Abbildung 26 zeigt ein solches Item bzw. dessen Aufgabenstamm. Die Testpersonen mussten hier neben einer falschen Rotation von Objekt A und B zudem erkennen, dass nicht die linke sondern fälschlicherweise die rechte Seite in der zu überprüfenden Ansicht dargestellt war. Es war also notwendig, zuerst die Vorderseite zu identifizieren, um anschließend die linke und rechte Seite von ihr zu bestimmen. Zum anderen fand sich der Kinesthetic factor eben auch bei der Identifizierung falscher Objektpositionen wieder, wenn diese z. B. miteinander vertauscht waren oder sich Objekte in der zu überprüfenden Ansicht auf der spiegelverkehrten Position befanden. Diesen Aspekt zeigt Abbildung 27 bzw. der Aufgabenstamm von Item MC5. Neben einer falschen Größe von Objekt B mussten Testpersonen feststellen, dass sich Objekt A (Tür) fälschlicherweise rechts und nicht links der Bänke befand bzw. an der rechten und nicht linken Wand ist. Für Objekt D galt, dass es anstatt vor den rechten nun vor den linken Bänken war.

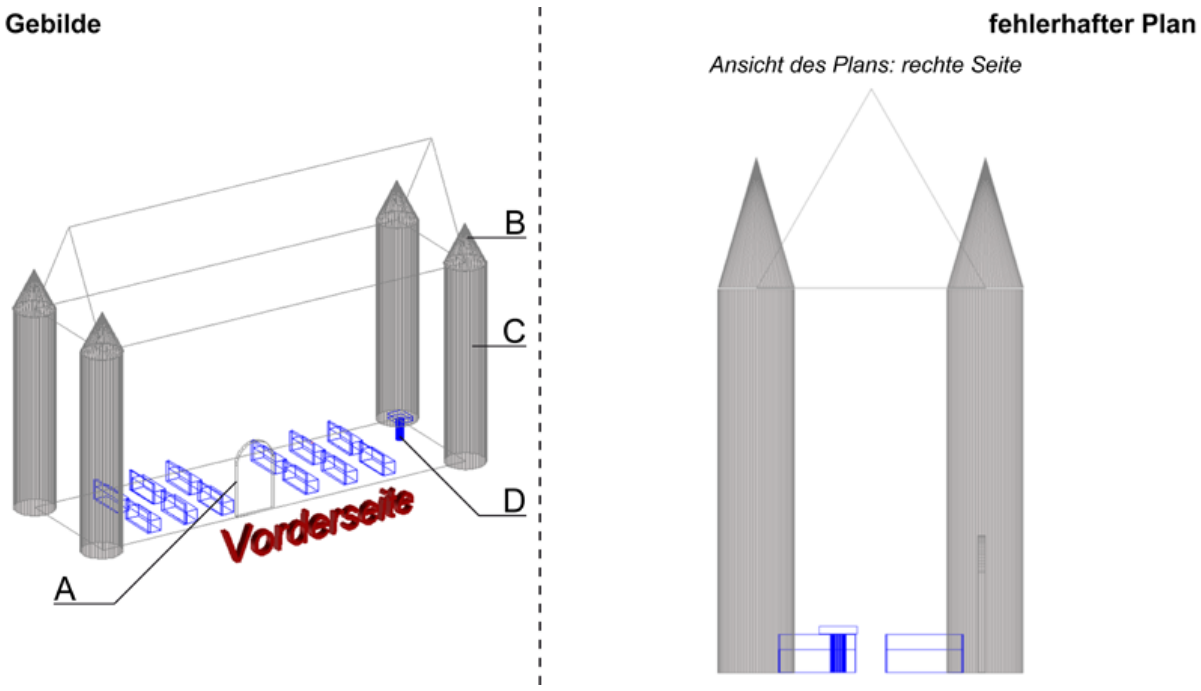


Abbildung 27. TARV: Aufgabenstamm Item MC5.

Ein Faktor, der hingegen bei allen Items wiederzufinden war, war Visualization. Sein wesentliches Merkmal, mehrere gedankliche Operationen bei komplexen Objekten oder Zusammenstellungen solcher Objekte anzuwenden, war Bestandteil sämtlicher Items des TARV. Zur Lösung eines Items war es in der MC- wie SQ-Version zuerst notwendig, je nach Darstellungsmodus, von 2D auf 3D oder umgekehrt zu schließen, um anschließend die vorgegebene und zu überprüfende Ansicht auf Entsprechung zu testen. Dabei konnten Fehler gemäß der drei Facetten auftreten, deren Identifikation wiederum diejenigen mentalen Operationen erforderten, die die jeweilige Facette beschreiben. Ebenso implizierte das Schließen von 2D auf 3D oder 3D auf 2D auch, wie gerade geschildert, sich eine andere Seite der Objekte vorzustellen. Dass der Faktor Visualization sich keiner Facette genau zuordnen ließ, war jedoch auch im Einklang mit der Literatur, die eben kritisierte, dass er zu allgemein definiert sei, anstatt genaue mentale Operationen zu beschreiben und somit in der Folge schwer von anderen Faktoren zu trennen oder in diesem Fall, einer genauen Facette zuzuordnen sei.

3.8. Mögliche Probleme hinsichtlich der Kalibrierung

Identifikation der Facetten:

Weitensfelder et al. (2010) wiesen bereits darauf hin, dass die Facetten durchaus Überschneidungen aufweisen, die einer klaren Abgrenzung zuwiderlaufen. Eine Veränderung der Neigung eines Objekts (Facette Rotation) kann sich auch auf den Abstand zu anderen Objekten (Facette Relationen) auswirken. Ein anderes Abstandsverhältnis kann wiederum die Positionen der Objekte (Facette Orientierung) beeinflussen. Diese beiden Aspekte wurden allerdings durch die Itemkonstruktion insofern vermieden, als dass die zu überprüfende Darstellung eine Seitenansicht war, bei der kein Zwiespalt zu anderen Facetten auftrat oder indem bei den Antwortmöglichkeiten diese „ungewollten“ Facetten ausgeschlossen wurden.

Nicht zu kontrollieren war jedoch die Tatsache, dass Testpersonen bei jedem Item die vorgegebene Ansicht rotieren oder einen Perspektivenwechsel durchführen mussten, um (gedanklich) zur zu überprüfenden (Seiten-)Ansicht zu gelangen. Im Weiteren ist es bei der Bearbeitung der Items der Facette Rotation, aber in gewissem Maße auch der Facette Orientierung, daher nicht möglich, zu bestimmen, ob die Objekte rotiert wurden, oder die Testperson stattdessen ihren Standpunkt veränderten. Das heißt, diese generelle Problematik der Raumvorstellungsforschung, die konkrete Bearbeitungsstrategie der Testpersonen nicht zu wissen, traf auch auf den TARV zu und erschwert die Differenzierung aller drei Facetten voneinander, insbesondere aber die der Facetten Rotation von der Facette Orientierung.

Die Identifikation der Facette Relationen ist insofern problematisch, als dass in der Literatur bisher kein entsprechender Raumvorstellungsfaktor nachgewiesen wurde und sich die Frage stellt, ob das Erkennen falscher Größen- und Abstandsverhältnisse im Vergleich zur Durchführung mentaler Rotationen nicht zu trivial ist. Für Items dieser Facette bedeutet dies unter Umständen, dass es den Testpersonen schwerer fiel, zur richtigen Seitenansicht (durch Rotieren oder Perspektivenwechsel) zu gelangen, als die Größe der Objekte und ihre Abstände zueinander zu erkennen. Das heißt, dass diese Items mehr der Facette

Rotation oder Orientierung zuzuweisen sind. Insbesondere die Facette Orientierung ist dabei hervorzuheben, da eine Veränderung der Größe oder des Abstandes zweier Objekte zueinander sich auch auf ihre Position auswirken kann.
Falls es daher nicht gelingen sollte, die Facetten als die dem TARV zugrunde liegenden Denkoperationen zu identifizieren, so müssen die soeben genannten Aspekte in Betracht gezogen werden, die dazu führen können, dass sich die Items überlappen, eine Differenzierung verhindern und keine Aussagen getroffen werden können, welche Denkoperationen bei welchem Item durchgeführt wurden.
Hinsichtlich des Nachweis, dass der TARV eindimensional misst, ist noch zu berücksichtigen, dass die Antwortmöglichkeiten sprachlich formuliert wurden. Das hieß, um Items (der MC-Version) zu lösen, war streng genommen neben der Raumvorstellung auch noch ein entsprechendes Sprachverständnis zur Identifizierung der Lösung notwendig.

Fairness:
Ein wesentlicher Anspruch des TARV war es, mehrere Aspekte der Raumvorstellung (Facetten) in einem Test zu vereinen, um eine systematische Diskriminierung zwischen den Geschlechtern zu vermeiden, da sich gezeigt hatte, dass der Geschlechtsunterschied je nach Test stark variieren bzw. auch verschwinden konnte. Da der Großteil der Studien jedoch über einen Leistungsunterschied zugunsten der Männer berichtete, bleibt es natürlich fraglich, ob der TARV diesen Unterschied mit seinem Konzept vollständig nivellieren konnte, zumal der Test verschiedene Charakteristika enthielt, die einen solchen begünstigten. Im Folgenden werden auf die in Kapitel *2.2.2. Test- und Itemcharakteristika als mögliche Ursachen des Geschlechtsunterschieds* erwähnten Aspekte eingegangen. Ihre Hauptaussagen sind zur besseren Verständlichkeit in Klammern kurz zusammengefasst.

Unmögliche Rotationen: *(Bei unmöglichen Rotationen vergrößert sich der Geschlechtsunterschied.)* Um ein Item in der MC-Version zu lösen, war es notwendig, Fehler zu erkennen. Das hieß, Items der Facette Rotation beinhalteten ausschließlich unmögliche Rotationen, da sie mindestens einen Fehler enthielten.

Bearbeitungszeit insgesamt: *(Bei einer begrenzten Gesamtbearbeitungszeit vergrößert sich der Geschlechtsunterschied.)* Die Bearbeitungszeit des TARV im Rahmen des Self-Assessments ist unbegrenzt. Für die Durchführung dieser Studie war aus organisatorischen Gründen (siehe Kapitel *5.2. Durchführung der Testung*) jedoch eine Beschränkung auf maximal 50 Minuten notwendig.

Komplexität von Items: *(Bei komplexeren Items vergrößert sich der Geschlechtsunterschied.)* Items des TARV bestanden aus zwei- wie dreidimensionalen Objekten, zu deren Lösung die Testpersonen verschiedene Denkoperationen gemäß des Darstellungsmodus sowie den jeweiligen Facetten durchführen mussten. Verglichen mit den Items von Rileas Studie (2008) waren diese als komplex zu betrachten.

Anzahl Dimensionen: *(Bei dreidimensionalen Objekten und beim Wechseln der Dimensionen (dimensionality crossing) vergrößert sich der Geschlechtsunterschied.)* Bei jedem Item des TARV mussten die Testpersonen die Dimensionen wechseln und je nach Darstellungsmodus eine zweidimensionale Ansicht einer dreidimensionalen Ansicht (3D auf 2D) oder eine dreidimensionale Ansicht anhand zweier zweidimensionalen Ansichten (2D auf 3D) gedanklich erzeugen. Damit hatten sie bei jedem Item mit dreidimensionalen Objekten zu tun.

4. Hypothesen im Rahmen der Kalibrierung

Neben den gängigen Hypothesen hinsichtlich der theoriebildenden Verfahren zum Nachweis der Konstruktvalidität sind die Hypothesen für die Überprüfung der Unterschiede zwischen den Personengruppen noch zu spezifizieren. Hier stellten sich drei Fragestellungen:

1. Gibt es einen Unterschied zwischen den Geschlechtern in ihrer Raumvorstellung?
2. Gibt es einen Unterschied zwischen Personen verschiedener Ausbildungen in ihrer Raumvorstellung?

3. Gibt es eine Wechselwirkung zwischen dem Geschlecht und der Ausbildung in Bezug zur Raumvorstellung?

Kapitel *2.2. Unterschiede in der Raumvorstellung: Geschlecht* und Kapitel *2.3. Unterschiede in der Raumvorstellung: Ausbildung* zeigten, dass sich zwischen Personen verschiedenen Geschlechts sowie Ausbildungen Unterschiede in ihrer Raumvorstellung auftun. Unklar blieb jedoch, welche der beiden, Geschlecht oder Ausbildung, eine größere Rolle für den Unterschied spielte, oder ob sogar eine Wechselwirkung beider vorherrschte. Für den TARV wurde davon ausgegangen, dass sich Unterschiede zwischen den Geschlechtern sowie Personen verschiedener Ausbildungen (hier: SchülerInnen verschiedener Schulen, siehe Kapitel *5.3. Stichprobe*) auftun, ohne dass eine Wechselwirkung nachgewiesen werden konnte. Das hieß, die Raumvorstellung der Testpersonen sollte bei naturwissenschaftlich-technischen geprägten Schulen besser sein als bei allgemeinbildenden Schulen. Innerhalb der einzelnen Schulen sollte sich ein Geschlechtsunterschied zugunsten der Männer ergeben, dessen Ausmaß sich zwischen den Schulen ähnlich verhalten sollte. Ein Wechselwirkungseffekt wurde nicht vermutet, womöglich war jedoch davon auszugehen, dass sich der Geschlechtsunterschied bei naturwissenschaftlich-technischen Schulen geringfügig verringern würde.
Ein Anliegen des TARV war es zwar, Unterschiede zwischen Personengruppen zu vermeiden, hinsichtlich der Tatsache jedoch, dass die Facetten bestimmen Raumvorstellungsfaktoren ähnlich waren (siehe Kapitel *3.7. Vergleich der Facetten des TARV mit den Raumvorstellungsfaktoren*), für die ihrerseits geringe bis sehr deutliche Geschlechtsunterschiede ermittelt wurden (siehe Kapitel *2.2. Unterschiede in der Raumvorstellung: Geschlecht*), war anzunehmen, dass sich auch beim TARV Unterschiede zu Gunsten der Männer zeigen würden. Des Weiteren beinhaltete der TARV Charakteristika, die einen solchen begünstigten (siehe Kapitel *3.8. Mögliche Probleme hinsichtlich der Kalibrierung*). Insbesondere der Aspekt, dass jedes Item beim Schließen von 2D auf 3D oder 3D auf 2D mentale Rotation (oder einen Perspektivenwechsel) benötigte, für die wiederum die größten Geschlechtsunterschiede zu Gunsten der Männer dokumentiert wurden, ließ auch

solche für den TARV vermuten.

Für Personen verschiedener Ausbildungen (hier: Schulen) wurde angenommen, dass SchülerInnen einer verstärkt naturwissenschaftlich-technischen ausgerichteten Schulen bessere Werte im TARV erzielen würden als solche einer allgemeinbildenden Schule. Dies war weitestgehend im Einklang mit der Literatur (siehe Kapitel *2.3. Unterschiede in der Raumvorstellung: Ausbildung*), die jedoch verstärkt den Unterschied auf universitärer Ebene, also zwischen verschiedenen Studiengängen, als auf schulischer Ebene untersucht hatte. Generell war jedoch zu vermuteten, dass der Unterschied weniger deutlich ausfiel. Zum einen, eben weil zwischen Studierenden bereits eine stärkere Förderung und damit Differenzierung hinsichtlich ihrer Fähigkeiten erfolgt als zwischen SchülerInnen. Zum anderen, da ein Vergleich zwischen Schulen die verschiedenen Schulzweige ignorierte. Das hieß, dass aufgrund von SchülerInnen, die einen verstärkt naturwissenschaftlich-technischen Schulzweig innerhalb einer allgemeinbildenden Schule oder womöglich vica versa, einen eher allgemeinbildenden Schulzweig einer naturwissenschaftlich-technischen Schule besuchten, sich die SchülerInnen in ihrer Raumvorstellung annähern könnten. Für den ersten Fall zeigten z. B. Gittler und Glück (1998) seine fördernde Wirkung für die Raumvorstellung. Ein Wechselwirkungseffekt zwischen Geschlecht und Ausbildung wurde aus zwei Gründen nicht vermutet. Erstens, aufgrund des soeben geschilderten Problems verschiedener Schulzweige innerhalb der Schulen sowie der Tatsache, dass die Differenzierung gemäß der Raumvorstellung auf schulischer Ebene noch geringer ausgefallen sein sollte. Zweitens, weil nur wenige Entsprechungen dafür in der Literatur gefunden werden konnten (siehe Kapitel *2.3. Unterschiede in der Raumvorstellung: Ausbildung*). Vielmehr stellte sich die Frage, ob das Geschlecht oder die Ausbildung eine größere Rolle für unterschiedliche Leistungen in der Raumvorstellung spielte. Aufgrund der immer wieder publizierten Geschlechtsunterschieden wurde daher vermutet, dass der Unterschied zwischen verschiedenen Schulen bestehen bleiben würde, die jeweilige schulische Ausbildung ihn jedoch geringfügig beeinflussen würde, ohne allerdings zu einem Wechselwirkungseffekt beider zu führen. Zusammengefasst hieß dies: Die Raumvorstellung sollte bei Männern und Frau-

en mit einer naturwissenschaftlich-technischen Ausbildung steigen, der Geschlechtsunterschied zu Gunsten der Männer würde jedoch bestehen bleiben. Es ergaben sich daher zwei einseitige Alternativhypothesen (H_A(A), H_A(B)) mit den entsprechenden Nullhypothesen (H_0), dass es keinen Unterschied zwischen den Geschlechtern (H_0(A)) sowie zwischen SchülerInnen verschiedener Schulen (H_0(B)) gibt. Ebenso ergab sich eine Nullhypothese H_0(C) bezüglich der Wechselwirkung zwischen Geschlecht und Ausbildung.

- H_A(A): Männer zeigen bei den kalibrierten Items des TARV eine bessere Raumvorstellung als Frauen.
- H_A(B): SchülerInnen naturwissenschaftlich-technischer Schulen zeigen bei den kalibrierten Items des TARV eine bessere Raumvorstellung als SchülerInnen allgemeinbildender Schulen.
- H_0(C): Es gibt keine Wechselwirkung zwischen Geschlecht und Ausbildung hinsichtlich ihrer Raumvorstellung.

5. Methode

5.1. Testdesign

Aufgrund der Begrenzung der Gesamtbearbeitungszeit auf eine Schulstunde bzw. 50 Minuten (siehe Kapitel *5.2. Durchführung der Testung*) war es nicht möglich, dass eine Testperson alle Items (siehe Tabelle 4) bearbeitete. Das hieß, es gab mehrere *Testhefte*, auf die die einzelnen Items verteilt waren, wobei eine Testperson nur eines der Testhefte bearbeitete. Wichtig dabei war, dass die Testhefte durch gemeinsame Items miteinander verbunden waren (Walter & Rost, 2011). Dadurch war es möglich, die Items als zusammenhängendes Ganzes auszuwerten bzw. zu kalibrieren und die Werte, die die Testpersonen in den jeweiligen Testheften erzielten, miteinander zu vergleichen (Kubinger et al., 2011). Ein Testdesign, bei dem die Items auf verschiedene Testhefte verteilt sind, wird als

Incomplete Block Design (Frey, Hartig & Rupp, 2009) oder *unvollständige Blockanlage* (Kubinger et al., 2011) bezeichnet. Verbinden die Items verschiedene Testhefte, so ist dies eine *unvollständige zusammenhängende Bockanlage* (Kubinger et al., 2011).

Die Konstruktion des Testdesigns für den TARV musste drei wesentliche Bedingungen erfüllen. Erstens durfte, wie bereits erwähnt, die Bearbeitung eines Testhefts durch eine Testperson die Gesamtbearbeitungszeit nicht überschreiten. Zweitens musste es so konstruiert sein, dass die zu kalibrierenden Items der MC-Version (MC1 bis MC14) von jeweils ca. 200 Testpersonen bearbeitet wurden, ohne dabei eine übermäßig große Stichprobe zu benötigen. Und drittens mussten die Testhefte durch Items der MC- wie SQ-Version miteinander verknüpft sein, um einen späteren Vergleich beider Versionen zu ermöglichen. Auch diente dies dazu, um für zukünftige Forschungsfragen beide Versionen separat voneinander kalibrieren zu können. Tabelle 5 gibt einen Überblick darüber, welche Items welchen Testheften zugeordnet wurden und wie die Testhefte dadurch miteinander verbunden waren.

Ziel dieser Studie war die Kalibrierung der Items MC1 bis MC14 und der Vergleich beider Versionen, der MC- wie SQ-Version, mit sämtlichen Items (MC1 bis SQ14). Es wurden drei Testhefte (A bis C) erzeugt, und die Items MC1 bis MC14 und SQ1 bis SQ4 kombinatorisch so auf die Testhefte verteilt, dass sichergestellt wurde, dass jedes Testheft gleich häufig mit den anderen Testheften durch gemeinsame Items verbunden war. Dies war wichtig, um zu vermeiden, dass Testhefte ihre Verbindung verlieren würden, falls im Rahmen der Kalibrierung Items entfernt werden mussten.

Die Items ab MC15 mussten aufgrund der beschränkten Gesamtbearbeitungszeit so verteilt werden, dass sie jeweils nur in einem Testheft waren. Die Testhefte teilten sich folglich diese Items nicht gemeinsam. Dennoch können auch diese Items für weitere Studien kalibriert werden, da über die vorher genannten Items (MC1 bis MC14 und SQ1 bis SQ4) die Testhefte ausreichend miteinander verknüpft sind.

Die Entscheidung für drei Testhefte beruhte darauf, dass für zwei Testhefte die begrenzte Bearbeitungszeit nicht eingehalten hätte werden können, da so mehr Items pro Testheft

Tabelle 5
TARV: Testdesign, Zuordnung der Items zu den Testheften

Item	Testheft A	Testheft B	Testheft C
MC1	X	X	
MC4	X	X	
MC7	X	X	
MC10	X	X	
MC11	X	X	
SQ1	X	X	
MC2	X		X
MC5	X		X
MC8	X		X
MC13	X		X
MC14	X		X
SQ3	X		X
MC3		X	X
MC6		X	X
MC9		X	X
MC12		X	X
SQ2		X	X
SQ4		X	X
SQ7	X		
SQ10	X		
SQ12	X		
SQ13	X		
MC16		X	
SQ6		X	
SQ8		X	
SQ11		X	
MC15			X
SQ5			X
SQ9			X
SQ14			X

notwendig gewesen wären. Für vier Testhefte hingegen wäre eine zu große Stichprobe vonnöten gewesen, damit die Items MC1 bis MC14 und SQ1 bis SQ4 von 200 Testpersonen bearbeitet worden wären. Das Testdesign war daher so konzipiert, dass jedes Testheft 100 Testpersonen vorgegeben wurde. Durch die gemeinsamen Items der Testhefte wurden so die Items MC1 bis MC14 und SQ1 bis SQ4 insgesamt von 200 Testpersonen und die restlichen Items (MC15, MC16 und SQ5 bis SQ14) von 100 Testpersonen bearbeitet. Demnach war eine Stichprobe von ca. 300 Testpersonen notwendig (siehe Kapitel *5.3. Stichprobe*).
Damit die Itemkalibierung mit einer unvollständigen Blockanlage funktionierte, war entscheidend, dass diese zusammenhängend war:

> „Eine unvollständige Blockanlage ist zusammenhängend, falls es für jedes Paar (A_k, A_l) von Behandlungen [Items] A_1,A_2,..., A_v eine Folge gibt, die mit A_k beginnt und mit A_l endet, so dass aufeinander folgende Behandlungen [Items] in dieser Folge in mindestens einem Block [Testheft] gemeinsam auftreten.“ (Kubinger et al., 2011, S. 168)

Ein Item sollten sich dabei stets zwei Testhefte teilen. Um die Items MC1 bis MC14 und SQ1 bis SQ4 somit gleichmäßig auf die Testhefte zu verteilen, wurde die Anzahl möglicher Kombinationen $\binom{n}{k}$, mit n als Anzahl der Testhefte und k als Anzahl der Testhefte, die sich ein Item teilen, berechnet. $\binom{3}{2}$ ergab drei mögliche Kombinationen. Gemäß diesem Schema wurden die 18 Items sukzessive auf die Testhefte verteilt, so dass sich in der Folge jedes Testheft sechs Items mit jedem anderen Testheft teilte. Dieser Ansatz, sämtliche Kombinationen bei der Itemverteilung wiederholt zu berücksichtigen, stellte vor allem sicher, dass die Blockanlage zusammenhängend war. Das hieß, dass es für jedes Itempaar eine entsprechende Folge von Itempaaren gab.
Zusammengefasst wurden demnach sämtliche 30 Items (MC1 bis SQ14) auf drei Testhefte verteilt. Ein Testheft bestand aus 16 Items, von denen ausgegangen wurde, dass sie in der vorgegebenen Zeit vollständig bearbeitet werden konnten. Da zu den Items keine Daten zur Schwierigkeit vorlagen, erfolgte die Zuweisung der Items zu den Testheften gemäß

ihrer Facetten und Darstellungsmodi (3D auf 2D, 2D auf 3D). Die zu kalibrierenden Items der MC-Version (MC1 bis MC14) wurden so auf die Testhefte aufgeteilt, dass in jedem Testheft diese Items zu gleichen Teilen aus solchen bestanden, die nur gemäß einer Facette konstruiert wurden, sowie aus solchen, die Facettenkombinationen beinhalteten. Ein weiteres Auswahlkriterium war, dass jedes Testheft Items jeder Facette hinsichtlich jedes Darstellungsmodus enthielt (z. B. Facette Rotation jeweils mit Darstellungsmodus 3D auf 2D und 2D auf 3D).

Bei den insgesamt 14 Items der SQ-Version wurden vier ausgewählt, die von jeweils zwei Testheften geteilt wurden und diese somit verbanden. Bei diesen Items wurde darauf geachtet, dass beide Darstellungsmodi sowie sämtliche Facetten berücksichtigt wurden. Die Aufteilung auf die Testhefte erfolgte gemäß der Darstellungsmodi. Das heißt, jedes Testheft beinhaltete Items der SQ-Version bei denen von 2D auf 3D und umgekehrt geschlossen werden musste. Eine zusätzliche Aufteilung nach den Facetten war bei nur vier Items und drei Testheften nicht möglich.

Es bestand die Annahme, dass die Schwierigkeit eines Items mit Anzahl der Facetten steigen würde, weswegen versucht wurde, in jedem Testheft Items einzelner Facetten wie Facettenkombinationen zu haben. Nichtsdestotrotz konnte vorab nicht ausgeschlossen werden, dass z. B. ein Testheft verhältnismäßig schwierigere Items als andere Testhefte enthielt.

Die Verteilung der Items innerhalb der MC- sowie SQ-Version erfolgte aufsteigend nach ihrer vom Autor dieser Arbeit vermuteten Schwierigkeit (z. B. Anzahl der Facetten, Deutlichkeit des zu erkennenden Fehlers). Zudem wurde darauf geachtet, dass sich beide Darstellungsmodi regelmäßig abwechselten und Testpersonen nicht zuerst alle Items eines und anschließend des anderen Darstellungsmodus bearbeiteten. Mögliche Positionseffekte konnten dadurch allerdings nicht ausgeschlossen werden und stellten ein Problem dar, das mit dem Testdesign einherging. Verschiedene Szenarien waren hier denkbar. Zum Beispiel, dass aufgrund von Ermüdungserscheinungen gegen Ende der Testbearbeitung die letzten Items eines Testhefts von weniger Testpersonen gelöst wurden (Frey et al.,

Tabelle 6
TARV: Testdesign, Position der Items je Testheft und Abfolge der MC- und SQ-Version

	Testheft A		Testheft B		Testheft C	
Item	Testheft 1 MC→SQ	Testheft 2 SQ→MC	Testheft 3 MC→SQ	Testheft 4 SQ→MC	Testheft 5 MC→SQ	Testheft 6 SQ→MC
MC1	1	7	2	8		
MC2	7	13			7	13
MC3			6	12	6	12
MC4	4	10	7	13		
MC5	3	9			9	15
MC6			1	7	5	11
MC7	9	15	4	10		
MC8	5	11			2	8
MC9			10	16	10	16
MC10	6	12	3	9		
MC11	8	14	5	11		
MC12			9	15	3	9
MC13	10	16			4	10
MC14	2	8			8	14
MC15					1	7
MC16			8	14		
SQ1	12	2	11	1		
SQ2			16	6	15	5
SQ3	11	1			11	1
SQ4			13	3	13	3
SQ5					16	6
SQ6			12	2		
SQ7	16	6				
SQ8			14	4		
SQ9					12	2
SQ10	14	4				
SQ11			15	5		
SQ12	13	3				
SQ13	15	5				
SQ14					14	4

2009). Auch stellte sich die Frage, welchen Einfluss es hatte, ob zuerst die MC- oder SQ-Version des TARV bearbeitet wurde, da beide verschiedene Bedingungen zur Lösung eines Items stellten (2 x richtig-oder-falsch, Fehleridentifikation). Dadurch, dass die Items nicht nach ihrer Schwierigkeit geordnet werden konnten, konnte beispielsweise ein schwieriges Item zu Beginn des Tests Frustration bewirken, was sich in der Folge auf die Bearbeitung der weiteren Items auswirkte. Ebenso konnten sich unterschiedliche Lerneffekte in Abhängigkeit davon einstellen, in welcher Reihenfolge die Items in jedem Testheft bearbeitet wurden. Lediglich mögliche Positionseffekte hinsichtlich der MC- oder SQ-Version des TARV wurden dadurch kontrolliert, dass jedes Testheft einmal mit der MC- und einmal mit der SQ-Version begann (siehe Abbildung 25). Streng genommen ergaben sich dadurch sechs Testhefte, je nachdem mit welcher Version begonnen wurde, wobei sich die Positionen der Items *innerhalb* der Version (MC oder SQ) nicht änderten. Tabelle 6 gibt einen Überblick über diese sechs Testhefte und die Position der Items in ihnen. Hinsichtlich möglicher Positionseffekte sei allerdings vorwegzunehmen, dass bei Geltung des Dichotom Logistischen Test-Modells von G. Rasch (1960) (siehe Kapitel *5.4.1. Dichotom Logistische Test-Modell von Rasch (1960)*) diese auszuschließen waren. Auf eine dezidierte Überprüfung von Positionseffekten mittels dem Linear Logistische Test-Modell von Fischer (1973) (siehe Kapitel *5.4.2. Linear Logistische Test-Modell von Fischer (1973)*) gemäß dem Vorgehen von Hohensinn et al. (2008) wurde verzichtet.

5.2. Durchführung der Testung

Die für die Kalibrierung notwendigen sechs Testhefte waren auf einer anderen Online-Plattform (http://www.tbst-assessments.at) zugänglich, als auf derjenigen, auf die der TARV als Self-Assessment angeboten wird. Der Test konnte allerdings unabhängig davon an einem Computer mit bestehender Internetverbindung bearbeitet werden.
Die Testungen erfolgten in insgesamt fünf Schulen in Wien, Niederösterreich und der Steiermark im Zeitraum von Dezember 2011 bis Jänner 2012. Um den Vergleich zwischen Personengruppen mit verschiedenen Ausbildungen zu ermöglichen, wurde der TARV zu

gleichen Teilen von SchülerInnen Allgemeinbildenden Höheren Schulen (AHS) sowie Berufsbildenden Höheren Schulen (BHS) bearbeitet. AHS haben dabei die „Vermittlung einer umfassenden und vertiefenden Allgemeinbildung“ (Bundesministerium für Unterricht, Kunst und Kultur, 2008) zum Ziel, während BHS neben der Allgemeinbildung noch eine zusätzliche berufliche Ausbildung ermöglichen, wodurch sich die Schulzeit um ein Schuljahr verlängert. Gemeinsam ist beiden, dass die SchülerInnen mit Abschluss der Schule die allgemeine Hochschulreife erhalten.
Beide Schulen untergliedern sich in verschiedene Schultypen und Fachrichtungen. Für die AHS können SchülerInnen zwischen unterschiedlichen Schultypen wählen. Die SchülerInnen, die den TARV bearbeitet haben, besuchten entweder ein Gymnasium oder Realgymnasium. Letzteres besitzt einen stärkeren naturwissenschaftlicheren Fokus, insbesondere die Unterrichtung in Darstellender Geometrie sei hier hervorgehoben (Bundesministerium für Unterricht, Kunst und Kultur, 2008). Bei den BHS wurden nur SchülerInnen Höherer Technischer Lehranstalten (HTL) in Betracht gezogen. Diese Schulen zielen insbesondere auf den Erwerb „höherer allgemeiner und fachlicher Bildung, die zur Ausübung eines höheren Berufs auf technischem [...] Gebiet in der industriellen [...] Wirtschaft befähigt“ (Bundesministerium für Unterricht, Kunst und Kultur, 2011). Wesentliches Element ist, dass die SchülerInnen sich für eine Fachrichtung (z. B. Maschineningenieurwesen, Elektrotechnik, Informationstechnologie, Innenraumgestaltung) entscheiden, in der die theoretische wie praktische (schulinterne Werkstätten, Pflichtpraktika) Vertiefung während der Schulzeit erfolgt.
Es wurde daher davon ausgegangen, dass SchülerInnen einer HTL eine stärkere naturwissenschaftlich-technische Ausbildung erhielten, als SchülerInnen einer AHS. Wie jedoch in Kapitel *4. Hypothesen im Rahmen der Kalibrierung* erwähnt, war nicht auszuschließen, dass sich beide Personengruppen in ihrer Ausbildung annäherten, wenn SchülerInnen einer AHS ein Realgymnasium besuchten und SchülerInnen einer HTL sich in eine weniger naturwissenschaftlich-technische Fachrichtung (z. B. Informatik, Kunst und Design) vertieften. Dieser Aspekt konnte insofern nicht kontrolliert werden, da die Auswahl, wel-

che SchülerInnen den TARV bearbeiten würden, die Schulen bzw. die DirektorInnen und LehrerInnen trafen. Der Grund lag darin, dass zwei wesentliche Bedingungen erfüllt sein mussten. Zum einen war die Bearbeitung des TARV nur an einem Computer mit Internetverbindung möglich, zum anderen mussten die SchülerInnen ein Mindestalter von 16 Jahren haben. Dem Spielraum zur Auswahl bestimmter Klassen war somit enge Grenzen gesetzt, zumal die Durchführbarkeit dieser Arbeit von den technischen Ressourcen (Computer) der Schule abhing. Dies führte dazu, dass sich die Stichprobe aus SchülerInnen unterschiedlicher Schultypen (AHS) und Fachrichtungen (HTL) zusammensetzte.
Daher wurde in weiterer Folge nicht zwischen diesen unterschieden, insbesondere da sich manche Fachrichtungen nur auf SchülerInnen bestimmter HTLs beschränkten und dadurch ebenso ein Effekt durch die jeweilige Schule und ihre Unterrichtung denkbar wäre. Das hieß, es erfolgte ausschließlich eine Unterscheidung zwischen SchülerInnen, die eine AHS oder eine HTL besuchten. Ein genauer Überblick über die Anzahl der SchülerInnen je Schule findet sich in Kapitel *5.3. Stichprobe*.
Vor dem Testtermin erhielten die SchülerInnen einen Elternbrief (siehe Kapitel *B. Elternbrief*), indem sie über das Ziel und dem Nutzen dieser Arbeit aufgeklärt wurden. Ebenso wurde auf die Anonymität hingewiesen und Kontaktmöglichkeiten für mögliche Fragen gegeben. Die Einverständniserklärungen der Eltern, sofern die SchülerInnen nicht selbstberechtigt waren, wurden vor der Testung eingesammelt. Getestet wurde schließlich klassenweise. Das heißt, die SchülerInnen bearbeiteten innerhalb ihrer Klasse den TARV im Rahmen einer Gruppentestung. Dabei war die Bearbeitung am Computer abhängig von der technischen Ausstattung der jeweiligen Schule. Dies stellte einen Nachteil dieser Arbeit dar, da somit aus technischer Sicht keine einheitlichen Testbedingungen für alle SchülerInnen gewährleistet werden konnte, insbesondere nicht hinsichtlich der Größe der Monitore. So stellten einige Schulen Computerräume für die Testungen zur Verfügung, während in anderen Schulen in *Laptop-Klassen*, also in Klassen, in denen jede/r SchülerIn ein eigenes Laptop für den Unterricht zur Verfügung hatte, getestet wurde. Ebenso variierte, je nach Klasse, deren Gruppengröße.

Vor Beginn der Testungen wurden die SchülerInnen von *einem/einer* TestleiterIn über das grundsätzliche Ziel dieser Arbeit (Diplomarbeit, Entwicklung eines Raumvorstellungstests im Rahmen der Wiener Self-Assessments für Architektur und Maschinenbau) und auf die Gewährleistung ihrer Anonymität durch individuelle Testcodes hingewiesen. Ebenso wurde erläutert, wie die Online-Plattform zu bedienen sei (Einloggen mit dem Testcode und Auswahl der jeweiligen Testhefts), dass den SchülerInnen die gesamte Schulstunde für die Bearbeitung zur Verfügung stünde und wann und wie sie ihre Testergebnisse erhalten würden.

Anschließend wurde von den TestleiterInnen jedem/jeder SchülerIn ein Zettel ausgehändigt, auf dem sein/ihr Testcode stand sowie die Nummer des jeweiligen Testhefts. Mit dem Testcode konnten sich die SchülerInnen auf der Online-Plattform einloggen, das entsprechende Testheft auswählen und mit der Bearbeitung des TARV beginnen. Das Austeilen der Zettel erfolgte nach einer Methode von Frey et al. (2009), indem sie reihenweise nach den sechs Testheften (Testheft 1, 3 und 5 beginnend mit der MC-Version, Testheft 2, 4 und 6 mit der SQ-Version; siehe Tabelle 6) ausgegeben wurden. Das heißt, SchülerIn 1 erhielt einen Testcode mit Testheft 1, SchülerIn 2 mit Testheft 2 usw. War eine vollständige Reihe an Testheften ausgegeben, so wurde wieder mit Testheft 1 begonnen, in diesem Beispiel also bei SchülerIn 7. In der nächsten Klasse wurde mit dem Testcode des Testhefts weitergemacht, der in der vorherigen Klasse der nächste gewesen wäre. Das heißt, erhielt z. B. der/die letzte SchülerIn einer Klasse Testheft 4, so wurde in der nächste Klasse mit Testheft 5 fortgefahren. Dieses Muster wurde schließlich auch auf Ebene der Schulen durchgeführt. In einer neuen Schule wurde mit dem Testheft begonnen, dass in der letzten Schule das nächste gewesen wäre.

Diese Methode barg drei Vorteile in sich. Erstens wurden so die Testhefte über alle Schulen und Klassen hinweg gleichmäßig oft bearbeitet, was womöglich nicht der Fall gewesen wäre, hätte man in jeder Klasse z. B. mit Testheft 1 begonnen. Ebenso ermöglichte sie es, dass sich die Testhefte auf alle SchülerInnen (verschiedener Schulen und verschiedenen Geschlechts) gleichmäßig verteilten, so dass z. B. ein Testheft nicht von deutlich

mehr Frauen als Männern bearbeitet wurde. Zweitens garantierte die Methode, dass die SchülerInnen bei der Bearbeitung des Tests nicht voneinander abschauen konnten, da aufgrund der verschiedenen Itempositionen innerhalb der Testhefte (je nachdem, ob mit der MC- oder SQ-Version begonnen wurde) die Wahrscheinlichkeit verringert wurde, dass benachbarte SchülerInnen das gleiche Item zur selben Zeit bearbeiteten. Drittens wurden so mögliche klassen- oder schulspezifische Effekte vermieden, die beispielsweise dadurch zu Stande gekommen wären, wenn jede Klasse oder Schule nur bestimmte Testhefte zugewiesen bekommen hätte.
Die Rückmeldung ihrer Ergebnisse erhielten die SchülerInnen auf einer separaten Website. Sie konnten sich hier mit ihren Testcodes einloggen und bekamen neben dem Ergebnis ausführliche Informationen zum TARV und dem Gesamtergebnis dieser Arbeit. Des Weiteren stand ihnen eine Kontaktmöglichkeit mit dem Autor zur Verfügung.

5.3. Stichprobe

Der TARV wurde von SchülerInnen fünf verschiedener Schulen in Wien, Niederösterreich und der Steiermark bearbeitet. Bei den Schulen handelte es sich um Allgemeinbildende Höhere Schulen (AHS) sowie Höhere Technische Lehranstalten (HTL). Die Akquirierung der Schulen erfolgte durch direkte Kontaktaufnahme mit den jeweiligen DirektorInnen, die, wie bereits erwähnt, anschließend die Auswahl trafen, welche SchülerInnen bzw. Klassen getestet werden konnten.
Insgesamt 315 SchülerInnen bearbeiteten den TARV. Für die weitere Auswertung der Daten mussten einige SchülerInnen aus dem Datensatz entfernt werden:

- 15 Testpersonen konnten den Test innerhalb einer Schulstunde nicht beenden. Dies lag hauptsächlich an zwei Gründen: Zum einen, dass diese Testpersonen zu spät in der Klasse erschienen und erst verspätet mit dem Test beginnen konnten. Zum anderen hatten manche Testpersonen Schwierigkeiten, eine Internetverbindung aufzubauen oder diese war bei ihnen aus unerklärlichen Gründen so langsam, dass es mehrere Minuten dauerte, bis eine Abbildung vollständig angezeigt wurde. Aller-

dings konnten auch einige Testpersonen den Test in der vorgegeben Zeit tatsächlich nicht beenden.

- Die durchschnittliche Bearbeitungszeit einer Testperson lag lediglich bei 15 Sekunden pro Item, weswegen davon ausgegangen wurde, dass die Items ausschließlich durch Raten zu lösen versucht wurden. Bei der Berechnung der Bearbeitungszeiten wurden ausschließlich diejenigen für die zu überprüfenden Ansichten verwendet. Die Bearbeitungszeiten für die vorgegebenen Ansichten ohne die zu überprüfenden, anhand derer sich die Testpersonen mit dem Item vertraut machen sollten, wurden von der Online-Plattform nicht für alle Testpersonen erfasst.

- Bei einer Testperson wurde ihre/seine markierten Antwortmöglichkeiten bei einem Item aus technischen Gründen nicht erfasst. Das Fehlen eines Items für diese Testperson führte in der weiteren Auswertung der Daten dazu, dass für diese Testperson automatisch angenommen wurde, sie hätte ein völlig anderes Testheft als die anderen SchülerInnen bearbeitet.

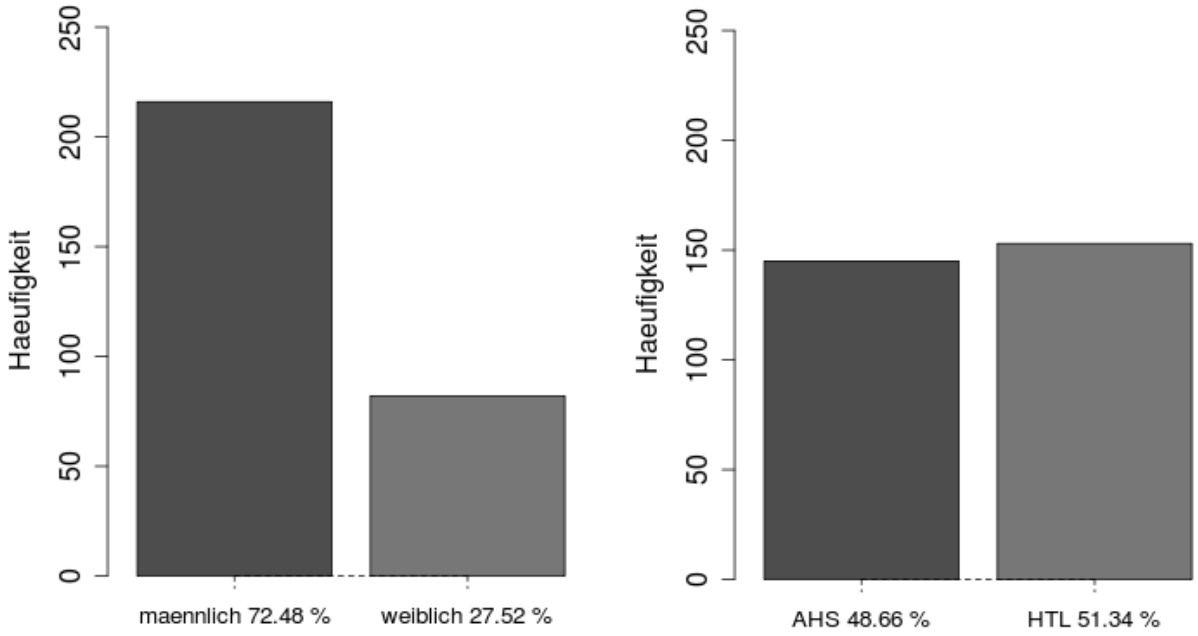
Aus den soeben genannten Gründen wurden diese 17 Testpersonen von der weiteren Auswertung der Daten ausgeschlossen, womit sich die endgültige Stichprobe auf 298 Testpersonen belief. Ihr Durchschnittsalter lag bei 17,22 Jahren mit einer Standardabweichung von 1,81 Jahren. Die jüngste Testperson war, gemäß dem vorgegebenen Mindestalter, 16 Jahre alt, die älteste 24 Jahre. Da dies für ein/e SchülerIn bereits ein gehobenes Alter war, wurde beim dem/der jeweiligen DirektorIn die Bestätigung eingeholt, dass 24 Jahre tatsächlich zutreffend waren. Aufgeteilt je nach Ausbildung (AHS oder HTL) hatten Testpersonen einer AHS ein Durchschnittsalter von 16,67 Jahren (Standardabweichung: 0,74 Jahre) und diejenigen einer HTL ein durchschnittliches Alter von 17,75 Jahren (Standardabweichung: 1,28 Jahre). Das höhere Durchschnittsalter von SchülerInnen einer HTL war auf die längere Schulzeit dieser Schule zurückzuführen, da auch SchülerInnen der höchsten Klassen getestet wurden.
Die absoluten Häufigkeiten hinsichtlich Geschlecht, Ausbildung und Testheft zeigt Tabelle 7. Ähnlich viele SchülerInnen der AHS wie HTL bearbeiteten den TARV, wobei

Tabelle 7
Stichprobe: Absolute Häufigkeiten nach Geschlecht, Ausbildung und Testheft

	AHS			HTL			AHS & HTL		
	m	w	g	m	w	g	m	w	g
Testheft A	25	28	53	46	6	52	71	34	105
Testheft 1	11	16	27	24	4	28	35	20	55
Testheft 2	14	12	26	22	2	24	36	14	50
Testheft B	23	22	45	41	7	48	64	29	93
Testheft 3	10	9	19	22	3	25	32	12	44
Testheft 4	13	13	26	19	4	23	32	17	49
Testheft C	32	15	47	49	4	53	81	19	100
Testheft 5	16	8	24	23	4	27	39	12	51
Testheft 6	16	7	23	26	0	26	42	7	49
	80	65	145	136	17	153	216	82	298

Anmerkung. m = männliche Testpersonen, w = weibliche Testpersonen, g = gesamt bzw. alle Testpersonen bzgl. Ausbildung und/oder Testheft.

hinsichtlich des Geschlechts deutlich mehr männliche Testpersonen, insbesondere bei der HTL, an der Untersuchung teilnahmen als weibliche. Dies entsprach jedoch dem allgemeinen Ungleichgewicht in der Geschlechterverteilung bei HTLs mit einem durchschnittlichen Frauenanteil von 11 % (Lassnigg & Vogtenhuber, 2009). Die relativen Häufigkeiten sind grafisch in Abbildung 28 veranschaulicht. Tabelle 7 gibt zudem die Verteilung der Testhefte in der Stichprobe wider. Diese waren aufgrund der Methode von Frey et al. (2009) gleichmäßig verteilt, so dass jedes Testheft insgesamt sowie zwischen den verschiedenen Ausbildungen ähnlich oft bearbeitet wurde. Testhefte A bis C wurden, wie intendiert und für die Kalibrierung notwendig, von jeweils ca. 100 Testpersonen bearbeitet. Beim Geschlecht ergab sich wegen der geringeren Anzahl an Schülerinnen ein entsprechendes Ungleichgewicht. Besonders deutlich ist dies bei Testheft 6, das 42 Schüler und lediglich 7 Schülerinnen bearbeitet haben, wovon keine eine HTL besuchte.



(a) Geschlecht (b) Ausbildung

Abbildung 28. Relative Häufigkeiten.

5.4. Theoriebildende Verfahren

5.4.1. Dichotom Logistische Test-Modell von Rasch (1960)

Das Dichotom Logistische Test-Modell von G. Rasch (1960) (im Weiteren als *Rasch-Modell* oder *RM* bezeichnet) „beschreibt die Wahrscheinlichkeit, dass [Testperson] v Item i löst („+“), in Abhängigkeit des Personenparameters ξ, das ist die (wahre) Fähigkeit von [Testperson] v, und des Itemparameters σ, das ist die (wahre) Schwierigkeit von [Item] i“ (Kubinger, 2009, S. 89):

$$P(+|\xi_v, \sigma_i) = \frac{e^{\xi_v - \sigma_i}}{1 + e^{\xi_v - \sigma_i}} \tag{1}$$

Das Rasch-Modell stellt demnach die Bedingung auf, dass das manifeste Antwortverhalten einer Testperson, welche Items sie also löst und nicht löst, ausschließlich von der jeweiligen dahinter liegende latente Variable bzw. der jeweiligen Fähigkeitsausprägung der

Testperson abhängt (Moosbrugger, 2012). Das heißt, der Item- oder Schwierigkeitsparameter σ sowie der Personen- oder Fähigkeitsparameter ξ sind eindimensional und drücken demzufolge „verschiedene Ausprägungsgrade auf derselben Dimension“ (Kubinger et al., 2011, S. 556) aus. Gilt für einen Test das Rasch-Modell, so ist dieser eindimensional.
Der Nachweis der Geltung erfolgt mittels dem *Likelihood-Quotienten-Test von Andersen* (1973; kurz: LQT), eines *Grafischen-Modell-Tests* sowie dem *z-Test* (Fischer & Scheiblechner, 1970). Alle testen die gleiche Annahme, nämlich dass sich die geschätzten Schwierigkeitsparameter der Items für verschiedene Teilstichproben nicht voneinander unterscheiden. Dies stellt eine überprüfbare Konsequenz des Rasch-Modells dar, dass bei dessen Geltung die Parameterschätzung unabhängig von der zugrundeliegenden Stichprobe ist (Kubinger et al., 2011).
Dazu werden, wie bereits erwähnt, Teilstichproben aus der Gesamtstichprobe gemäß bestimmter Teilungskriterien gebildet und die geschätzten Schwierigkeitsparameter dieser Gruppen miteinander verglichen. Als Teilungskriterien fungierten hier:

- **Rohscore:** Aufteilung der Stichprobe nach dem Median in gleich große Teilstichproben mit hohem und niedrigem Rohscore. Der Rohscore einer Testperson ist die Anzahl der gelösten Items.

- **Geschlecht:** Als Teilungskriterium relevant aufgrund der beschriebenen Unterschiede zwischen Männern und Frauen in Raumvorstellungstests.

- **Ausbildung:** Ebenso wie beim Geschlecht zeigten sich Unterschiede zwischen Personen verschiedener Ausbildungen, weswegen es als Teilungskriterium in Betracht gezogen wurde.

Das Rasch-Modell gilt, wenn die geschätzten Schwierigkeitsparameter bei allen Teilstichproben übereinstimmen bzw. große Ähnlichkeit besitzen. Ist dies nicht der Fall, werden sukzessive Items entfernt, bei denen sich die Teilstichproben in ihrer geschätzten Schwierigkeit unterscheiden. Die Auswahl dieser Items erfolgt mithilfe der Ergebnisse des Grafischen-Modell-Tests sowie des z-Tests. Der Nachweis der Modellgeltung wäre in

diesem Falle lediglich a posteriori.

Ein Nachteil des Likelihood-Quotienten-Tests von Andersen ist die Akkumulation des Typ I-Risikos. Das heißt, mit zunehmender Anzahl an Teilungskriterien und damit auch an Signifikanztests, die den Unterschied zwischen den Teilstichproben testen, erhöht sich die Wahrscheinlichkeit, fälschlicherweise zum Schluss zu gelangen, die Teilstichproben würden sich unterscheiden (siehe hierzu Kubinger et al., 2011). Aus diesem Grund wurde die von Kubinger (2005) vorgeschlagene *Bonferroni-Methode* verwendet, bei der das Risiko 1. Art ($\alpha = 0{,}05$) durch die Anzahl an Signifikanztests bzw. Teilungskriterien geteilt wird: $\frac{0{,}05}{3} = 0,017$. Da das gängige Signifikanzniveau in diesem Wertebereich bei 0,01 liegt, wurde für alle Likelihood-Quotienten-Tests daher dieses verwendet. Ebenso wurde für den z-Test ein Signifikanzniveau von 0,01 gewählt, da die Bonferroni-Methode in diesem Fall zu einem Signifikanzniveau führen würde ($\frac{0{,}05}{Anzahl\ an\ Items\ bzw.\ Signifikanztests\ (max.\ 30)} = mind.\ 0,002$) bzw. zur Folge gehabt hätte, dass sich bei fast allen geschätzten Itemparametern die Teilstichproben, im Widerspruch zum Likelihood-Quotienten-Test von Andersen und dem Grafischen-Modell-Test, nicht unterschieden hätten.

Eine (a priori oder a posteriori) Geltung des Rasch-Modells bedeutete für den TARV, dass der Test eindimensional ist, folglich der Test in allen Teilstichproben die gleiche Fähigkeit bzw. das gleiche Merkmal misst.

Die Überprüfung der Geltung des Rasch-Modells erfolgte in dieser Arbeit anhand von zwei Gruppen von Items. Einerseits an den 14 Items mit Multiple-Choice-Format (MC1 bis MC14), andererseits an sämtlichen Items (MC1 bis SQ14). Der Grund für diese Differenzierung war folgender: Der Nachweis der Konstruktvalidität, dass die Facetten die zugrunde liegenden Denkoperationen des TARV darstellen, war ausschließlich bei Items mit Multiple-Choice-Format (die von ca. 200 Testpersonen bearbeitet wurden) möglich, da dies bei ihrer Konstruktion intendiert wurde. Des Weiteren war eben die Kalibrierung des Tests mit diesen Items das Hauptziel dieser Arbeit, weswegen die Überprüfung sämtlicher Ziele hinsichtlich der Fairness und Konstruktvalidität des TARV mit diesen Items (MC1 bis MC14) erfolgte. Die erneute Überprüfung der Rasch-Modell-Geltung mit

sämtlichen Items war für den Vergleich beider Versionen, der MC- wie SQ-Version, notwendig.
Dies stellte natürlich keine optimale Vorgehensweise dar, allerdings wurde so vermieden, dass der Nachweis der Konstrukvalidität bei den Items MC1 bis MC14 durch die unterschiedlichen Antwortformate beider Versionen beeinträchtigt wurde. Schließlich war anzunehmen, dass das Antwortformat und deren inhärente unterschiedliche Ratewahrscheinlichkeit ihren eigenen Beitrag zur Itemschwierigkeit leistete, der allerdings nicht berücksichtigt werden konnte.

5.4.2. Linear Logistische Test-Modell von Fischer (1973)

Das Linear Logistische Test-Modell von Fischer (1973) (im Weiteren als *LLTM* bezeichnet) verallgemeinert das Rasch-Modell in der Form, dass hier der Item- bzw. Schwierigkeitsparameter in eine Linearkombination einzelner Itemkomponenten bzw. Basisparameter zerlegt wird (Kubinger et al., 2011). Die Kombination der jeweiligen Basisparameter η und deren Gewichtung q ergibt damit die Schwierigkeit σ eines Items:

$$\sigma_i = \sum_{j=1}^{p} q_{ij}\eta_j \tag{2}$$

Übertragen auf Gleichung 1 bedeutet dies für die Wahrscheinlichkeit, dass Testperson v mit Personenparameter ξ Item i mit Schwierigkeitsparameter σ löst: (c stellt eine Konstante dar, die der gewichteten Summe der Basisparameter für jedes Item hinzuzurechnen ist (Rost, 2004))

$$P(+|\xi_v, \sigma_i = \sum_{j=1}^{p} q_{ij}\eta_j) = \frac{e^{(\xi_v - \sum_{j=1}^{p} q_{ij}\eta_j - c)}}{1 + e^{(\xi_v - \sum_{j=1}^{p} q_{ij}\eta_j - c)}} \tag{3}$$

Da das LLTM die Schwierigkeit der gleichen Anzahl an Items durch weniger Parameter zu erklären versucht als das Rasch-Modell, das für jedes Item einen eigenen Schwierig-

keitsparameter definiert, ist die Geltung des Rasch-Modells eine notwendige Bedingung für die Geltung des LLTM (Moosbrugger, 2012). Ergo wird mittels eines Likelihood-Quotienten-Test überprüft, ob das LLTM die Daten nicht signifikant schlechter erklärt als das Rasch-Modell (Kubinger et al., 2011). Da hier nur *ein* Signifikanztest durchgeführt wird, wird ein Risiko 1. Art von $\alpha = 0,05$ eingegangen.

Ist der Likelihood-Quotienten-Test nicht signifikant, gilt das LLTM, was für diese Arbeit bedeutet, dass neben der Eindimensionalität des Tests nachgewiesen werden würde, dass die Facetten die zugrunde liegenden Denkoperationen zur Lösung der Items sind. Das heißt, dass die Dimension, die der TARV erfasst, das Merkmal Raumvorstellung ist und sich dieses aus den drei Facetten zusammensetzt.

Das LLTM setzt dazu gewisse Hypothesen voraus, welche Denkoperationen in welchem Ausmaß bzw. mit welchem Gewicht für die Lösung von Items notwendig sind (Rost, 2004). Dies wird durch eine Rechteckmatrix (*Q-Matrix*) dargestellt (Rost, 2004). Die Q-Matrix Q1 für die Items MC1 bis MC14 gibt Tabelle 8 wieder. Die Facetten wurden dabei

Tabelle 8
Q-Matrix Q1 für die Items MC1 bis MC14

	Basisparameter			
Item	Relationen	Rotation	Orientierung	Konstante
MC1	1	0	0	1
MC2	0	1	0	1
MC3	0	0	1	1
MC4	1	1	0	1
MC5	1	0	1	1
MC6	0	1	1	1
MC7	1	1	1	1
MC8	1	0	0	1
MC9	0	1	0	1
MC10	0	0	1	1
MC11	1	1	0	1
MC12	1	0	1	1
MC13	0	1	1	1
MC14	1	1	1	1

als Basisparameter formuliert. Die Gewichtung ging danach, ob ein Item diese Facette enthielt (1) oder nicht (0). Der Darstellungsmodus konnte vom LLTM nicht berücksichtigt werden, was auf eine Besonderheit der Q-Matrix zurückzuführen war, die Rost (2004) beschreibt: Diese darf keine Spalten beinhalten, die voneinander abhängig sind. Das heißt, keine Spalte darf sich durch die gewichtete Summe anderer Spalten ergeben. Wäre der Darstellungsmodus miteinbezogen worden, so wären dies zwei zusätzliche Spalten (3D auf 2D und 2D auf 3D) mit exakt entgegengesetzten Gewichten. 3D auf 2D hätte von Item MC1 bis MC7 Gewichtung 1 und von MC8 bis MC 14 Gewichtung 0. Für 2D auf 3D wäre es genau umgekehrt. Die Addition dieser beiden Spalten würde allerdings die Konstante ergeben, was den Bedingungen der Q-Matrix zuwiderlief.

6. Ergebnisse

Die statistische Auswertung der Daten erfolgte mit der Software *R* (R Development Core Team, 2011). Für die Überprüfung der Geltung des Rasch-Modells sowie des LLTM wurde das für diese Software verfügbare Zusatzpaket *eRm* (Mair, Hatzinger & Maier, 2011) verwendet. Die Berechnung des optimalen Stichprobenumfangs für den Vergleich verschiedener Personengruppen erfolgte mit dem Zusatzpaket *OPDOE* (D. Rasch, Pilz, Verdooren & Gebhardt, 2011).

6.1. Items MC1 bis MC14

6.1.1. Überprüfung der Geltung des Rasch-Modells (Skalierung)

Die Ergebnisse des Likelihood-Quotienten-Test von Andersen zeigt Tabelle 9. Zu beachten dabei ist, dass aufgrund inadäquater Antwortmuster beim Teilungskriterium Geschlecht Item MC12 und bei Teilungskriterium Ausbildung Item MC4 von den Analysen ausgeschlossen wurden. Das heißt, diese Itemparameter konnten für die jeweiligen Teilstichproben nicht geschätzt werden, da keine Schülerin Item MC12 und keine Testperson einer

HTL Item MC4 lösen konnte. Diese beiden Items wurden jedoch nicht aus dem Itempool, also der Menge aller Items, ausgeschlossen, da wegen ihrer unmöglichen Schätzung keine Aussagen hinsichtlich einer Modellverletzung getroffen werden konnten. Sie blieben daher bei der Itemanalyse hinsichtlich Rasch-Modell-Konformität außen vor, wurden aber bei der inhaltlichen Analyse berücksichtigt. Für alle drei Teilungskriterien galt allerdings

Tabelle 9
MC1 bis MC14: Geltung Rasch-Modell, LQ-Test von Andersen

Teilungskriterium	χ^2 (LQT)	*df*	$\chi^2_{\alpha=1\%}$	*p*-Wert		ohne Item
Rohscore	17,886	13	27,688	0,162	$> 0,01$ n. s.	-
Geschlecht	21,560	12	26,217	0,043	$> 0,01$ n. s.	MC12
Ausbildung	21,285	12	26,217	0,046	$> 0,01$ n. s.	MC4

Anmerkung. df = Freiheitsgrade, n. s. = nicht signifikant.

das Rasch-Modell, ohne dass, neben dem automatischen Nicht-Berücksichtigen von Items MC4 und MC12, Items ausgeschlossen werden mussten.

Die Ergebnisse der Grafischen-Modell-Tests mit angegebenen Konfidenzbändern zeigen Abbildung 29 bis 31. Erkennbar ist, dass die Items MC4 und MC12 auch beim Teilungskriterium Rohscore auffällig waren und lediglich wegen ihres großen Standardfehlers (Tabelle 10) zu keiner Verletzung der Rasch-Modell-Gültigkeit führten. Ähnliches war für Item MC4 beim Teilungskriterium Geschlecht anzunehmen. Item MC12 zeigte sich beim Teilungskriterium Ausbildung jedoch weitestgehend modellkonform.

Der Grafische-Modell-Test zeigte zudem bei den Teilungskriterien Rohscore und Ausbildung ein ähnliches Bild, verglichen mit Teilungskriterium Geschlecht. So befanden sich Items MC10 und MC3 beim Teilungskriterium Ausbildung und etwas weniger deutlich beim Teilungskriterium Rohscore an der Grenze zur Modellkonformität, während sie beim Teilungskriterium Geschlecht näher der Winkelhalbierenden (1. Mediane) waren. Das heißt, dass die Itemparamter dieser Items beim Teilungskriterium Geschlecht in den Teilstichproben ähnlich geschätzt wurden. Hingegen zeigte sich bei Teilungskriterium Geschlecht verglichen mit den anderen Teilungskriterien Item MC1 auffällig. Ein Ausschuss von Item MC10 würde den χ^2-Wert beider Likelihood-Quotienten-Tests von Andersen ver-

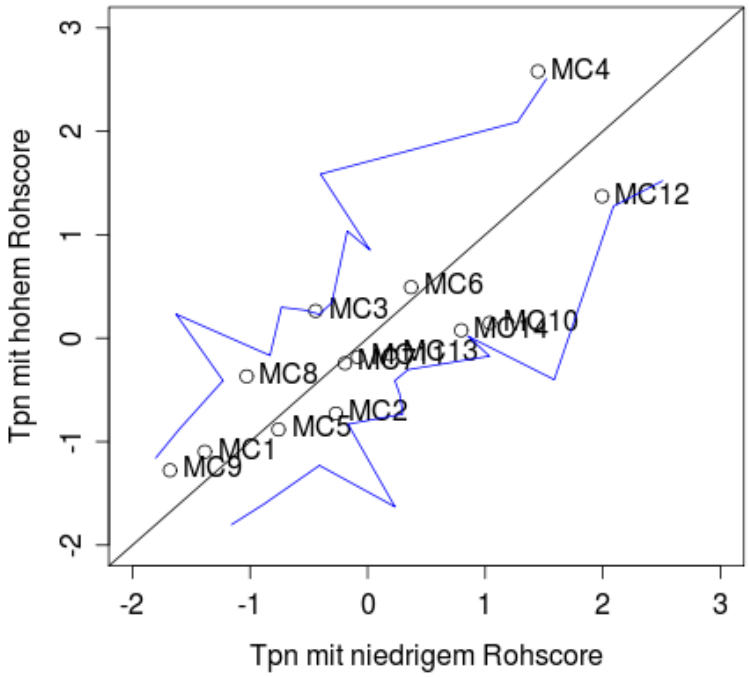


Abbildung 29. MC1 bis MC14: Grafischer-Modell-Test, Teilungskriterium Rohscore.

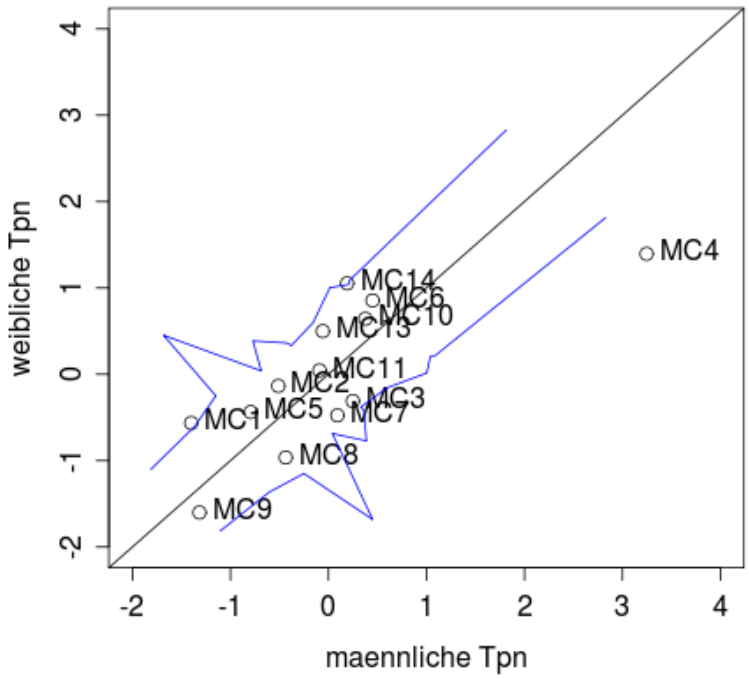


Abbildung 30. MC1 bis MC14: Grafischer-Modell-Test, Teilungskriterium Geschlecht.

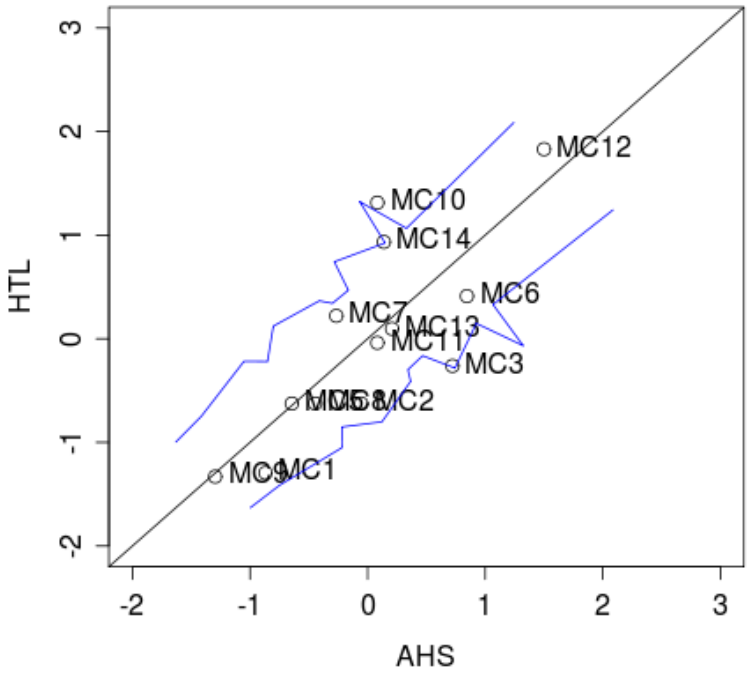


Abbildung 31. MC1 bis MC14: Grafischer-Modell-Test, Teilungskriterium Ausbildung.

ringern ($\chi^2_{Rohscore} = 15,643$ und $\chi^2_{Ausbildung} = 14,596$), während sich für das Geschlecht kaum Änderungen einstellen würden ($\chi^2_{Geschlecht} = 21,177$). Der Unterschied zwischen den Teilungskriterien war auch durch die unterschiedlichen Größen der Teilstichproben bedingt. Während die Teilungskriterien Rohscore und Ausbildung die Gesamtstichprobe in zwei ähnlich große Teilstichproben teilten, so war das Verhältnis zwischen dem Anteil an Männern und Frauen in der Gesamtstichprobe 72 % zu 28 %. Es war daher nicht auszuschließen, dass sich mit einer gleichmäßigeren Verteilung der Geschlechter die Grafischen-Modell-Tests aller drei Teilungskriterien mehr ähnelten.

Ein Item, das bei allen drei Teilungskriterien (bei Teilungskriterium Rohscore jedoch geringfügig) auffällig war, war Item MC14, dessen Ausschluss sich auf alle Teilungskriterien in Richtung einer besseren Modellkonformität auswirken würde ($\chi^2_{Rohscore} = 10,985$, $\chi^2_{Geschlecht} = 19,044$, $\chi^2_{Ausbildung} = 15,850$). Auf diesen Schritt wurde jedoch verzichtet und stattdessen für den bestehenden Itempool, ohne Ausschluss weiterer Items, die Modellgültigkeit angenommen, wie sie auch durch die Likelihood-Quotienten-Test von Andersen ausgewiesen wurde.

Tabelle 10
MC1 bis MC14: geschätzte Itemparameter, Standardfehler, Konfidenzintervall, rel. Lösungshäufigkeit

Item	Itemparameter	Standardfehler	Konfidenzintervall 95%	relative Lösungshäufigkeit
MC1	-1,223	0,163	[-1,543;-0,904]	35,35 %
MC2	-0,487	0,171	[-0,822;-0,151]	25,37 %
MC3	0,035	0,208	[-0,373;0,444]	15,03 %
MC4	2,183	0,477	[1,247;3,118]	2,02 %
MC5	-0,773	0,164	[-1,095;-0,452]	30,73 %
MC6	0,484	0,238	[0,017;0,951]	10,36 %
MC7	-0,197	0,196	[-0,581;0,187]	17,17 %
MC8	-0,673	0,166	[-0,999;-0,347]	28,78 %
MC9	-1,454	0,164	[-1,776;-1,133]	41,45 %
MC10	0,390	0,232	[-0,065;0,845]	10,61 %
MC11	-0,120	0,200	[-0,511;0,271]	16,16 %
MC12	1,502	0,349	[0,819;2,185]	4,15 %
MC13	0,010	0,190	[-0,362;0,383]	17,56 %
MC14	0,323	0,207	[-0,083;0,728]	13,66 %

Die Verteilung der geschätzten Personenparameter im Vergleich zu den geschätzten Itemparametern zeigt Abbildung 32. Die oberen Balken geben dabei die Häufigkeitsverteilung der geschätzten Personenparameter wieder. Die schwarzen Punkte stehen für die geschätzten Itemparameter. Es ist zu beachten, dass die Personenparameter dann am genauesten geschätzt werden, wenn sie den jeweiligen geschätzten Itemparametern entsprechen. Das bedeutet, wenn eine Person Items bearbeitet, deren Schwierigkeit ihrer Fähigkeit entsprechen, dann beträgt die Wahrscheinlichkeit, ein Item zu lösen, genau $p = 0,50$. Sind die Items deutlich zu schwierig oder zu leicht, so kann zwischen den Fähigkeitsausprägungen verschiedener Personen nicht ausreichend differenziert werden.
Dies war bei den Items MC1 bis MC14 der Fall. So differenzierten diese gut in einem Personenparameterbereich von -0,8 (MC5) bis +0,5 (MC6), allerdings befand sich der Großteil der geschätzten Personenparameter der Stichprobe (142 Testpersonen) unterhalb -1,5. Des Weiteren konnten, was in der Abbildung nicht berücksichtigt wird, für 49 Testpersonen keine Personenparameter geschätzt werden, da sie kein Item lösen konn-

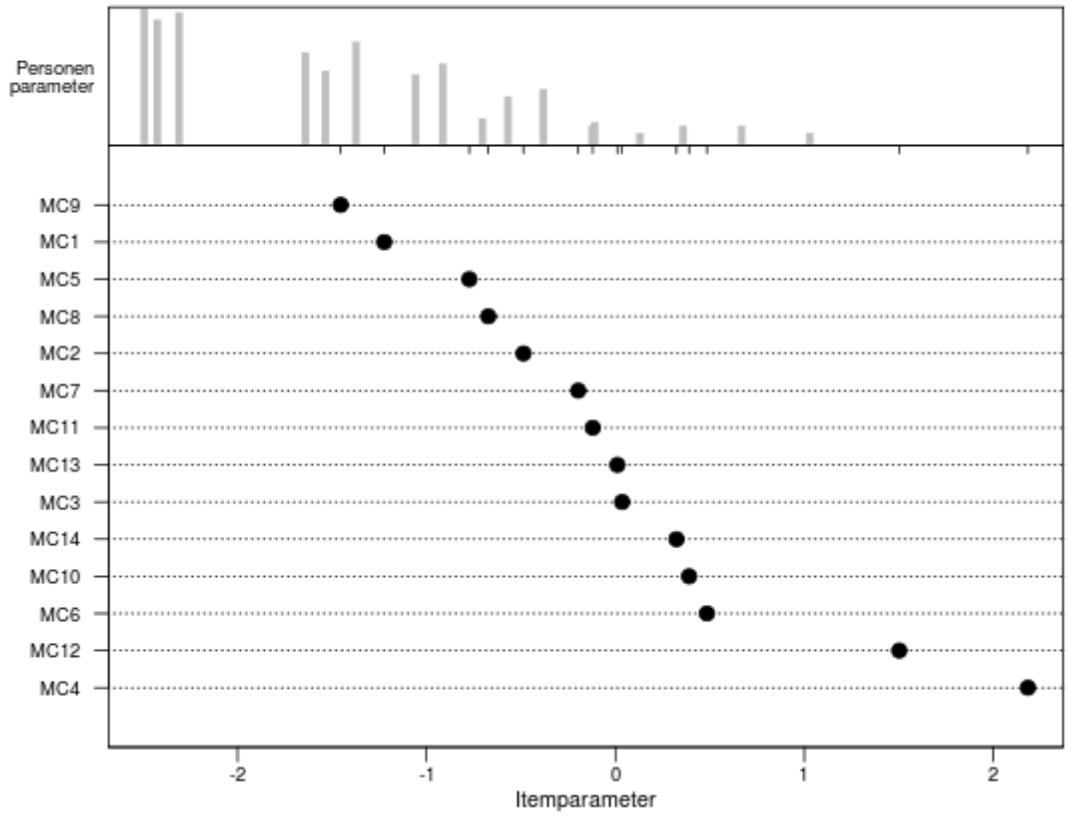


Abbildung 32. MC1 bis MC14: Vergleich geschätzte Personen-/Itemparameter.

ten. Die somit rechtsschiefe Verteilung der geschätzten Personenparameter ließ darauf schließen, dass die Items den Testpersonen insgesamt zu schwer fielen. Dies spiegelten auch die relativen Lösungshäufigkeiten der Items (Tabelle 10) wider, wonach kein Item von mehr als der Hälfte der Testpersonen, die es bearbeiteten, gelöst wurde. Item MC9 wurde dabei mit 41,45 % am häufigsten gelöst. Die Items MC4 und MC12, für die die Itemparameter für manche Teilstichproben nicht geschätzt werden konnten, befanden sich dementsprechend in einem Schwierigkeitsbereich ($> +1{,}5$), für den in dieser Stichprobe keine entsprechenden Personenparameter geschätzt wurden.

Zusammengefasst bedeutete dies für die Items, dass der Test, aufgrund der Geltung des Rasch-Modells, eindimensional war. Eine bessere Modellpassung wäre durch den Ausschluss von Items MC10 und MC14 gegeben, wenngleich sich dies zum größten Teil auf

die Teilungskriterien Rohscore und Ausbildung bezog. Die Items MC4 und MC12 wurden von der Itemanalyse ausgeschlossen, da sie für nicht alle Teilungskriterien geschätzt werden konnten. Insbesondere für Item MC4 wurde in den verschiedenen Teilstichproben je nach Teilungskriterium zu unterschiedliche Itemparameter geschätzt, weswegen dieses Item als kritisch zu betrachten war. Beide, MC4 wie MC12 zeigten zudem die höchste Schwierigkeit, weswegen eine Umgestaltung angebracht wäre. Generell erwiesen sich die meisten Items als zu schwierig, so dass in dem Bereich, in dem sie die Personenparameter am genausten schätzen würden, nur wenige Testpersonen die entsprechende Fähigkeit besaßen.

6.1.2. Überprüfung der Geltung des LLTM (Konstruktvalidität)

Wie soeben gezeigt, galt für die Items MC1 bis MC14 das Rasch-Modell, weswegen die Geltung des LLTM überprüft werden konnte. Tabelle 11 gibt das Ergebnis wieder, das zeigte, dass das LLTM die Daten signifikant schlechter erklärte als das Rasch-Modell, da der errechnete χ^2-Wert des Likelihood-Quotienten-Tests deutlich größer als der kritische Wert war. Die Korrelationskoeffizient der geschätzten Itemparameter des Rasch-Modells

Tabelle 11
MC1 bis MC14: Geltung LLTM, LQ-Test, Q-Matrix Q1

L_{LLTM}	L_{RM}	χ^2 (LQT)	$\chi^2_{\alpha=5\%}$	*df*	
-826,999	-743,035	167,927	18,307	10	s.

Anmerkung. df = Freiheitsgrade, s. = signifikant.

und des LLTM betrug 0,38, was auf deutliche Abweichungen schließen ließ. In Abbildung 33, in der die geschätzten Itemparameter gegenübergestellt sind, zeigte sich insbesondere für Item MC4 und MC12, dass diese deutlich in der Form abwichen, dass beide vom LLTM als leichter geschätzt wurden. Auf die Notwendigkeit einer Umgestaltung beider Items aufgrund der Tatsache, dass sie kaum gelöst wurden und ihre Schwierigkeit folglich nicht bei allen Teilungskriterien geschätzt werden konnte, wurde bereits hingewiesen. Ein Ausschluss beider Items mit anschließender Neuberechnung des LLTM ermittelte

einen höheren Korrelationkoeffizienten der geschätzten Itemparameter (0,72) und führte zu einem deutlich niedrigeren χ^2-Wert des Likelihood-Quotienten-Tests von Andersen, der jedoch weiterhin statistisch signifikant blieb ($\chi^2_{LQT} = 65,697$, $\chi^2_{\alpha=5\%} = 15,507$, $df = 8$).

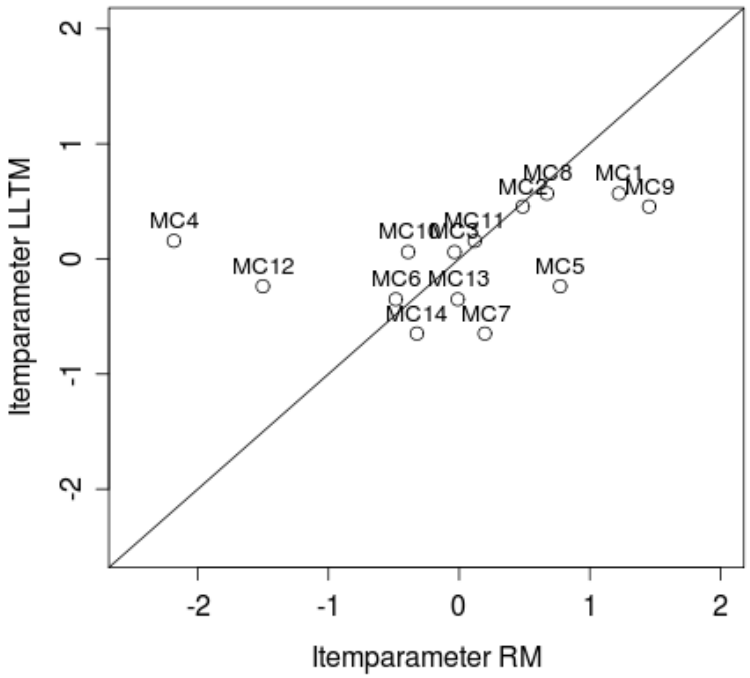


Abbildung 33. MC1 bis MC14: Vergleich geschätzte Itemparameter RM und LLTM.

In Kapitel *3.8. Mögliche Probleme hinsichtlich der Kalibrierung* wurde bereits darauf hingewiesen, dass die Items die Facetten womöglich nicht klar genug voneinander trennten. Zwei Möglichkeiten wurden hier in Betracht gezogen. Zum einen, dass Testpersonen Items der Facette Rotation wie Orientierung mit den gleichen Denkoperationen lösten. Das hieße zum Beispiel, dass alle Items mit mentaler Rotation gelöst wurden. Zum anderen, dass bei Items der Facette Relationen keine Größen- und Abstandsverhältnisse sondern Positionsveränderungen erkannt worden wären, somit das Item der Facette Orientierung zuzuordnen wäre. Bei der Q-Matrix des ersten Falles (Q2) wurden daher die Gewichte der Basisparameter Rotation und Orientierung zusammengeführt, bei der des zweiten Falles (Q3) die der Basisparameter Relationen und Orientierung. Tabelle C gibt einen Überblick über die so veränderten Q-Matrizen. In beiden Fällen wurden deutlich unterschiedliche Itemparameter für Item MC4 und MC12, verglichen mit denen des Rasch-

Modells geschätzt, weswegen der Likelihood-Quotienten-Test von Andersen ohne diese Items durchgeführt wurde. Es zeigte sich (siehe Tabelle 12), dass das LLTM für keine der drei Möglichkeiten die Daten besser erklärte als das Rasch-Modell sowie auch als das ursprüngliche LLTM.

Tabelle 12
MC1 bis MC14: Geltung LLTM, LQ-Test, Q-Matrizen Q2 bis Q4, mit und ohne MC4 und MC12

Items	Q2	Q3	$\chi^2_{\alpha=5\%}$	*df*	Q4	$\chi^2_{\alpha=5\%}$	*df*
alle	175,323	155,258	19,675	11	149,312	16,919	9
ohne MC4 und MC12	104,390	73,488	16,919	9	49,477	14,067	7

Eine weitere Möglichkeit, welche Denkoperationen die Schwierigkeit der Items beeinflussen könnte, spiegelt Q-Matrix Q4 wider, indem neben den drei ursprünglichen Basisparametern ein vierter erstellt wurde, der für die Anzahl der Objekte steht, die in jedem Aufgabenstamm gekennzeichnet waren. Diesem Basisparameter lag somit die Annahme zugrunde, dass die Items umso schwieriger waren, je mehr Objekte gleichzeitig berücksichtigt wurden und bei je mehr Objekten mentale Operationen durchgeführt werden mussten, um zur Lösung zu gelangen. Im weitesten Sinne beschrieb er damit die Rolle des Arbeitsgedächtnisses (siehe hierzu Baddeley & Hitch, 1974). Es zeigte sich (siehe Tabelle 12), dass mit diesem zusätzlichen Basisparametern die Daten durch das LLTM am besten erklärt wurden, allerdings weiterhin statistisch signifikant schlechter als durch das Rasch-Modell.

Zusammengefasst ließ sich somit die Schwierigkeit der Items nicht durch die gewichtete Kombination der Basisparameter allein erklären. Das heißt, obwohl der TARV eindimensional maß, waren nicht oder nicht ausschließlich die Facetten die notwendigen Denkoperationen zum Lösen der Items. Auch unter Ein- und Ausschluss verschiedener Basisparameter konnte keine Geltung des LLTM erreicht werden, womit die Konstruktvalidität nicht nachgewiesen werden konnte.

6.1.3. Vergleich zwischen verschiedenen Gruppen von Testpersonen (Fairness)

Die Überprüfung der Unterschiede zwischen Männern und Frauen sowie Testpersonen einer AHS und HTL erfolgte mittels einer zweifachen Varianzanalyse, die ebenso eine Interaktion beider Faktoren (Geschlecht und Ausbildung) testete. Zwar wurde eine solche Wechselwirkung durch die formuliertes Hypothesen ausgeschlossen bzw. nicht vermutet, dennoch konnte in Studien (siehe Kapitel *2.3. Unterschiede in der Raumvorstellung: Ausbildung*) eine solche festgestellt werden, weshalb sie bei der Hypothesenprüfung berücksichtigt wurde.

Um die 49 Testpersonen, für die kein Personenparameter geschätzt werden konnte, in die Hypothesentestung miteinzubeziehen, wurde folgendes Vorgehen gewählt: Da diese Testpersonen kein Item lösen konnten, wurde der geringste geschätzte Personenparameter der restlichen 249 Testpersonen gewählt (-2,493) und von diesem eine halbe Standardabweichung abgezogen ($\frac{0{,}827}{2}$). Der geschätzte Personenparameter dieser 49 Testpersonen betrug damit -2,907.

Für die Hypothesentestung wurde die Gesamtstichprobe verwendet. Das bedeutete, dass den Genauigkeitsanforderungen nicht entsprochen wurde, den Stichprobenumfang in Abhängigkeit des Risikos 1. Art[13] und des Risikos 2. Art[14] sowie dem inhaltlich relevanten Mittelwertsunterschied beider Personengruppe zu bestimmen. Die Zusammensetzung und Größe der Stichprobe folgte in erster Linie den Notwendigkeiten der Kalibrierung von ca. 200 Testpersonen je Testheft bzw. Item.

Eine a posteriori Schätzung der notwendigen Stichprobengröße, um signifikante sowie inhaltlich relevante Unterschiede zwischen den Gruppen festzustellen, ergab 236 Testpersonen mit 59 Testpersonen je Gruppe (weiblich/AHS, weiblich/HTL, männlich/AHS, männlich/HTL). Dabei würde ein Risiko 1. Art von $\alpha = 0,05$ und Risiko 2. Art von $\beta = 0,05$ eingegangen. So würde vermieden werden, dass der TARV im Rahmen des Self-Assessments angewandt zu einer falschen Studienentscheidung führt, da er fälschlicherwei-

[13]Die Wahrscheinlichkeit, die Nullhypothese fälschlicherweise zu verwerfen.
[14]Die Wahrscheinlichkeit, die Alternativhypothese fälschlicherweise zu verwerfen.

se zwischen Personen verschiedenen Geschlechts und Ausbildungen differenziert. Der relevante Mittelwertsunterschied (hier: zwischen Personen verschiedenen Geschlechts sowie Ausbildungen) würde gemäß Kubinger et al. (2011) mindestens $\frac{2}{3}$ einer angenommenen Standardabweichung von 1 betragen, demnach $\delta = 0,67$. Abgeleitet aus der Literatur wäre ein ähnlicher Mittelwertsunterschied anzunehmen, da der TARV mit seinen Itemcharakteristika sowie insbesondere der Facette Rotation, Elemente beinhaltete, die zu einem relevanten Unterschied führen könnten.

An Tabelle 7 ist zu erkennen, dass die Gesamtstichprobe die Voraussetzung von 59 Testpersonen je Gruppe für die Kombination weibliche Testperson/HTL nicht erfüllte. Die Gruppe männliche Testperson/HTL war dagegen mehr als doppelt so oft besetzt. Die Ergebnisse der zweifachen Varianzanalyse mussten daher vor dem Hintergrund ungleicher Gruppengrößen betrachtet werden.

Wichtige Voraussetzung einer zweifachen Varianzanalyse ist die Homogenität der Varianzen. Da kein geeignetes Verfahren zu Überprüfung dessen existierte, wurde eine Regel von Kubinger et al. (2011) verwendet. Diese besagt, dass das Verhältnis der größten zur kleinsten Standardabweichung einer Gruppe kleiner als 1,5 sein sollte. Das war hier der Fall: $\frac{s_{weiblich/HTL}}{s_{weiblich/AHS}} = \frac{0,967}{0,855} = 1,131 < 1,5$.

Tabelle 13 gibt die Mittelwerte und Standardabweichung der einzelnen Gruppen wieder. Auffallend dabei war, dass weibliche Testpersonen einer HTL die höchsten und männliche Testpersonen einer HTL die niedrigsten Werte aller Gruppen erzielten. Allerdings zeigte sich bereits auch, dass die deskriptiven Unterschiede zwischen den Geschlechtern und SchülerInnen verschiedener Schulen gering waren.

Die Ergebnisse der zweifachen Varianzanalyse mit einem Risiko 1. Art von $\alpha = 0,05$ zeigt Tabelle 14. Es kam weder zwischen männlichen und weiblichen Testpersonen noch zwischen SchülerInnen einer AHS oder HTL zu statistisch signifikanten Unterschieden. Beide erklärten zudem nur einen sehr kleinen Anteil an der Gesamtvarianz mit $\eta^2_{Geschlecht}$ = 0,014 bzw. 1,4 % und $\eta^2_{Ausbildung}$ = 0,071 bzw. 7,1 %. Des Weiteren zeigte sich kein Wechselwirkungseffekt zwischen den Gruppen.

Tabelle 13
MC1 bis MC14: Mittelwerte ($\overline{x}$) und Standardabweichungen (s) der geschätzten Personenparameter für Geschlecht und Ausbildung

Ausbildung, $\overline{x}$ (s)	Geschlecht, $\overline{x}$ (s)		
	männlich	weiblich	
AHS	-1,646 (0,954)	-1,828 (0,855)	-1,728 (0,912)
HTL	-1,892 (0,881)	-1,565 (0,967)	-1,856 (0,894)
	-1,801 (0,914)	-1,773 (0,880)	

Tabelle 14
*MC1 bis MC14: Zweifache Varianzanalyse, Geschlecht, Ausbildung, Geschlecht*Ausbildung*

	SQ	*df*	*MQ*	*F*	*p*-Wert	
Geschlecht	0,05	1	0,045	0,056	0,813	$> 0,05$ n. s.
Ausbildung	1,23	1	1,231	1,518	0,219	$> 0,05$ n. s.
Geschlecht * Ausbildung	2,760	1	2,759	3,401	0,066	$> 0,05$ n. s.
Rest	238,45	294	0,811			

Anmerkung. SQ = Summe der quadratischen Abweichung, df = Freiheitsgrade, MQ = mittlere quadratische Abweichung, $F = \frac{MQ}{MQ_{Rest}}$, n. s. = nicht signifikant.

Somit waren beide Alternativhypothesen $H_A(A)$ und $H_A(B)$ zu verwerfen und ihre entsprechenden Nullhypothesen, dass es keinen Unterschied gibt, anzunehmen. Die Nullhypothese $H_0(C)$, dass es zu keinem Wechselwirkungseffekt kommt, konnte beibehalten werden.

6.2. Items MC1 bis SQ14: Vergleich der MC- und SQ-Version des TARV

Für den Vergleich der Items der MC-Version mit denen der SQ-Version war der abermalige Nachweis notwendig, dass das Rasch-Modell galt, diesmal für die Items MC1 bis SQ14. Dies ermöglichte zum einen, die Schwierigkeit der Items einander gegenüberzustellen sowie zum anderen zu überprüfen, ob die Items trotz ihres unterschiedlichen Antwortformats die gleiche zugrunde liegende Fähigkeit erfassten. Das Vorgehen (Teilungskriterien, Risiko 1. Art) war dabei analog zu Kapitel *6.1.1. Überprüfung der Geltung des Rasch-*

Modells (Skalierung). Tabelle 15 zeigt die Ergebnisse des Likelihood-Quotienten-Tests von Andersen. Für Item SQ9 sowie abermals für die Items MC4 und MC12 konnten die Itemparameter nicht bei allen Teilungskriterien geschätzt werden, weswegen sie bei der weiteren Itemanalyse außen vor blieben. Insgesamt galt das Rasch-Modell mit allen Items für alle drei Teilungskriterien nicht.

Tabelle 15
MC1 bis SQ14: Geltung Rasch-Modell, LQ-Test von Andersen

Teilungskriterium	χ^2 (LQT)	*df*	$\chi^2_{\alpha=1\%}$	*p*-Wert		ohne Items
Rohscore	57,334	29	49,588	0,001	$\leq 0,01$ s.	-
Geschlecht	49,631	27	46,963	0,005	$\leq 0,01$ s.	MC12, SQ9
Ausbildung	58,645	28	48,278	0,001	$\leq 0,01$ s.	MC4
Rohscore	52,833	28	48,278	0,003	$\leq 0,01$ s.	SQ12
Geschlecht	42,026	26	45,642	0,024	$> 0,01$ n. s.	MC12, SQ9, SQ12
Ausbildung	55,434	27	46,963	0,001	$\leq 0,01$ s.	MC4, SQ12
Rohscore	42,718	27	46,963	0,028	$> 0,01$ n. s.	SQ4, SQ12
Geschlecht	38,777	25	44,314	0,039	$> 0,01$ n. s.	MC12, SQ4, SQ9, SQ12
Ausbildung	44,549	26	45,642	0,013	$> 0,01$ n. s.	MC4, SQ4, SQ12

Anmerkung. df = Freiheitsgrade, s. = signifikant, n. s. = nicht signifikant.

Beim Grafischen-Modell-Test (Abbildungen 39 bis 41, siehe Anhang) war Item SQ12 als gemeinsames Item auszumachen, dass bei allen Teilungskriterien von der Winkelhalbierenden (1. Mediane) abwich. Beim z-Test zeigten sich entsprechend niedrige *p*-Werte für dieses Item bei allen Teilungskriterien, die jedoch nicht statistisch signifikant wurden bzw. größer als das festgelegte Risiko 1. Art von $\alpha = 0,01$ waren: $p_{Rohscore} = 0,042$, $p_{Geschlecht} = 0,023$, $p_{Ausbildung} = 0,038$. Bei Ausschluss dieses Items galt das Rasch-Modell für das Teilungskriterium Geschlecht. Beim folgenden Grafischen-Modell-Test ohne Item SQ12 zeigte sich insbesondere beim Teilungskriterium Ausbildung Item SQ4 als auffällig (Abbildungen 42 bis 44, siehe Anhang). Ebenso unterschieden sich gemäß des z-Tests die geschätzten Itemparameter der Teilstichproben ($p_{Ausbildung} = 0,001$). Mit Ausschluss

dieses Items galt schließlich das Rasch-Modell für alle Teilungskriterien. Abbildungen 34 bis 36 geben den Grafischen-Modell-Test bei den Items MC1 bis SQ14 ohne die Items SQ4 und SQ12 wieder. Es zeigten sich bei den Teilungskriterien Geschlecht und Ausbildung noch auffällige Items, z. B. Item MC16 für beide Teilungskriterien oder SQ11 für das Teilungskriterium Ausbildung. Aufgrund der Geltung des Rasch-Modells für alle Teilungskriterien wurde dies jedoch nicht weiter berücksichtigt, zumal im Vordergrund nicht die Kalibrierung der Items sondern dessen Vergleich stand.

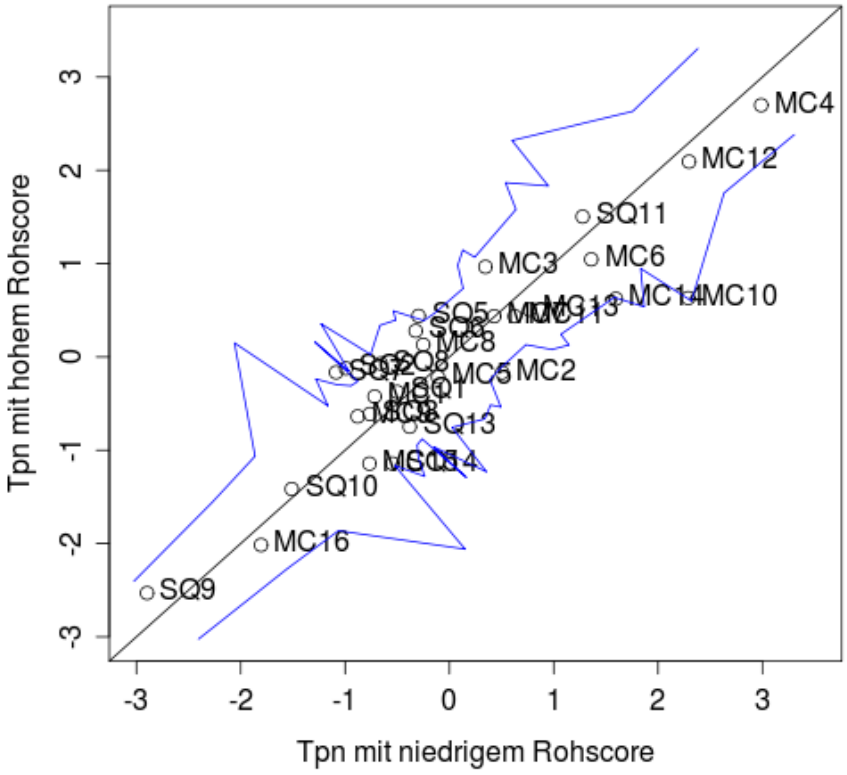


Abbildung 34. MC1 bis SQ14 ohne SQ4 und SQ12: Grafischer-Modell-Test, Teilungskriterium Rohscore.

Die durchschnittliche Bearbeitungszeit für die Items der MC-Version (MC1 bis MC16) lag bei 61,05 Sekunden mit einer Standardabweichung von 19,92 Sekunden. Für die verbliebenen Items der SQ-Version (SQ1 bis SQ14 ohne SQ4 und SQ12) betrug die Bearbeitungszeit im Durchschnitt 58,05 Sekunden. Die Standardabweichung lag bei 24,15 Sekunden. Der Korrelationskoeffizient zwischen den geschätzten Itemparametern aller Items der MC-

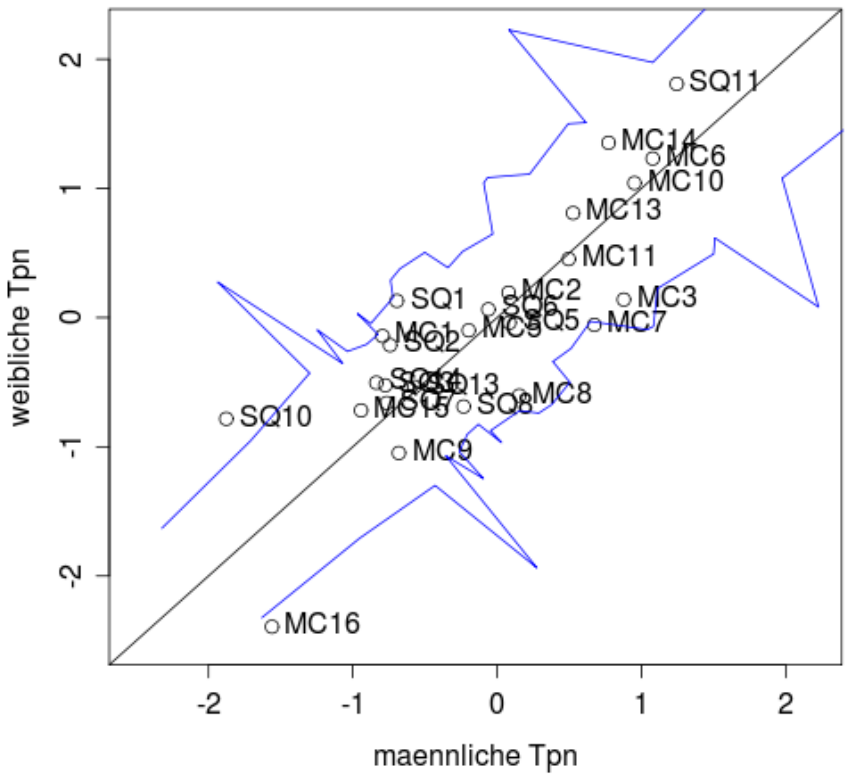


Abbildung 35. MC1 bis SQ14 ohne SQ4 und SQ12: Grafischer-Modell-Test, Teilungskriterium Geschlecht.

und SQ-Version mit Ausnahme der Items SQ4 und SQ12 und ihren durchschnittlichen Bearbeitungszeiten betrug -0,18.

Des Weiteren wurden die verbliebenen Items der SQ-Version den Items der MC-Version gegenübergestellt, die ihnen in den Facetten sowie Darstellungsmodi entsprachen (siehe Tabelle 4). Dabei musste berücksichtigt werden, dass durch die fehlende Geltung des LLTM ein Vergleich von Items gleicher Facetten nur bedingt aussagekräftig war. Tabelle 16 gibt einen Überblick über die geschätzten Itemparameter, relativen Lösungshäufigkeiten und den durchschnittlichen Bearbeitungszeiten der Items.

Trotz Geltung des Rasch-Modells fiel ein Vergleich aufgrund der verschiedenen Aufgabenstämme der Items schwer. Diese unterschieden sich nicht nur in den jeweils dargestellten Objekten, sondern auch in der Anzahl der Fehler je Aufgabenstamm. Die Aufgabenstämme der MC-Version konnten mehrere Fehler besitzen, auch wenn sie nur gemäß

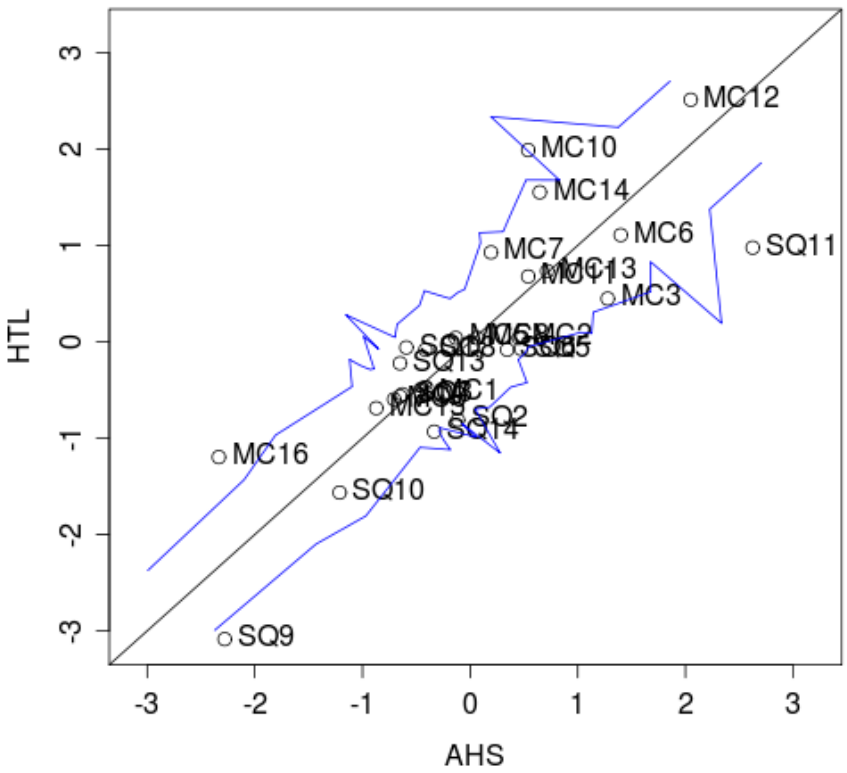


Abbildung 36. MC1 bis SQ14 ohne SQ4 und SQ12: Grafischer-Modell-Test, Teilungskriterium Ausbildung.

einer Facette konstruiert wurden. Durch das Antwortformat (Multiple-Choice) konnte identifiziert werden, ob eine Testperson die Fehler erkannte. Die Aufgabenstämme der SQ-Version hatten indes wegen des Antwortformats (sequentiell) stets nur einen Fehler, da nicht zwischen mehreren Fehlern differenziert werden konnte. Vor diesem Hintergrund fiel auf, dass Item MC2 und MC10, die je zwei Fehler beinhalteten, den Testpersonen schwerer fielen als ihre korrespondierenden Items der SQ-Version SQ2 und SQ7 bzw. SQ3 und SQ14. Auch innerhalb der MC-Version wurde für Item MC16, mit einem Fehler, ein niedrigerer Itemparameter geschätzt als für das korrespondierende Item MC10. Die verbliebenen Items der MC-Version beinhalteten je einen Fehler, was sich unter Umständen in einer ähnlichen und auch geringeren geschätzten Schwierigkeit (außer Item MC3) im Vergleich zu den korrespondierenden Items der SQ-Version niederschlug. Bei den durchschnittlichen Bearbeitungszeiten ergab sich bei den Items MC2 und insbesondere MC10

Tabelle 16
Vergleich Items der MC- und SQ-Version nach Facette und Darstellungsmodus

F. D.	Item	Itempa-rameter	Standard-fehler	Konfidenzin-tervall 95%	rel. Lösungs-häufigkeit	Bearbeitungszeit in Sek., $\overline{x}$ (s)
Relationen 3D/2D	MC1	-0,549	0,158	[-0,858;-0,240]	35,35 %	54,41 (22,72)
	MC15	-0,854	0,215	[-1,276;-0,433]	49,00 %	51,42 (24,39)
	SQ1	-0,408	0,161	[-0,723;-0,094]	32,32 %	51,89 (23,95)
	SQ5	0,104	0,236	[-0,358;0,566]	28,00 %	57,96 (33,38)
	SQ6	0,028	0,263	[-0,487;0,543]	21,50 %	53,47 (23,02)
Relationen 2D/3D	MC8	-0,028	0,163	[-0,348;0,291]	28,78 %	60,86 (26,28)
	SQ11	1,430	0,420	[0,607;2,252]	6,45 %	74,39 (44,89)
Rotation 3D/2D	MC2	0,154	0,169	[-0,177;0,484]	25,37 %	68,38 (34,18)
	SQ2	-0,570	0,158	[-0,879;-0,261]	37,82 %	56,61 (38,03)
	SQ7	-0,685	0,231	[-1,102;-0,268]	40,95 %	49,78 (30,95)
Rotation 2D/3D	MC9	-0,730	0,155	[-1,035;-0,425]	41,45 %	46,72 (34,05)
	SQ13	-0,517	0,216	[-0,939;-0,094]	37,14 %	59,53 (41,06)
Orien-tierung 3D/2D	MC3	0,726	0,206	[0,322;1,130]	15,03 %	61,62 (31,03)
	SQ8	-0,328	0,244	[-0,805;0,149]	27,96 %	48,67 (23,77)
	SQ9	-2,764	0,294	[-3,340;-2,189]	86,00 %	39,87 (27,22)
	SQ10	-1,454	0,214	[-1,873;-1,035]	59,05 %	57,76 (37,58)
Orien-tierung 2D/3D	MC10	1,029	0,232	[0,573;1,484]	10,61 %	92,67 (59,09)
	MC16	-1,776	0,231	[-2,228;-1,324]	60,00 %	70,95 (39,18)
	SQ3	-0,663	0,152	[-0,961;-0,366]	42,44 %	58,77 (29,82)
	SQ14	-0,727	0,216	[-1,150;-0,304]	40,00 %	69,22 (36,64)

Anmerkung. F. = Facette, D. = Darstellungsmodus, 3D/2D = 3D auf 2D, 2D/3D = 2D auf 3D.

ein ähnliches Bild, dass diese, im Vergleich zu den Items der SQ-Version, höher waren. Die durchschnittlichen Bearbeitungszeiten der restlichen Items ließen allerdings keine weiteren Schlussfolgerungen zu.

Für einen genaueren Vergleich beider Versionen boten sich Itemfit-Indizes an. Diese geben die Itemtrennschärfe wieder. Das heißt, sie geben Auskunft darüber, wie exakt die Items zwischen Personen mit hoher und niedriger Fähigkeitsausprägung trennen (Rost, 2004). Dabei wird zwischen einem *Item-Overfit* und einem *Item-Underfit* unterschieden. Ein Item-Overfit deutet darauf hin, dass das Antwortmuster, welche Testpersonen mit welcher Fähigkeit das Item lösen bzw. nicht lösen, zu wenig variiert, als bei Geltung

des Rasch-Modells angenommen wird (Bond & Fox, 2007). Das Antwortmuster ist folglich zu deterministisch. Da dass Rasch-Modell ein probabilistisches Modell ist, demnach auf Wahrscheinlichkeiten basiert, wird bei jedem Item eine gewisse Schwankung vorausgesetzt. Folglich sollten auch Personen mit geringerer Fähigkeitsausprägung schwierigere Items (per Zufall) lösen, was bei einem Item-Overfit nicht der Fall ist. Das Gegenteil stellt der Item-Underfit dar. Dieser weist darauf hin, dass das Antwortmuster zu zufällig ist. Das heißt, dass das Lösen eines Items weniger von der Fähigkeit der Person als vom Zufall abhängt (Rost, 2004), das Item also ungenügend zwischen Personen mit verschiedenen Fähigkeitsausprägungen trennt.

Tabelle 17 gibt einen Überblick über die Infit- und Outfit-Maße der mittleren quadratischen Abweichung sowie deren entsprechenden standardnormalverteilten t-Werte für die Items MC1 bis SQ14 ohne SQ4 und SQ12. Für genauere Erklärungen zu den Maßen und ihrer Berechnung sei auf Bond und Fox (2007) und Rost (2004) verwiesen. Bond und Fox (2007) geben Grenzwerte für die Infit- und Outfit-Maße an. Demnach weist ein $MeanSQ < 0,75$ und/oder ein t-Wert $< -2,00$ auf einen Overfit hin. Ein $MeanSQ > 1,30$ und/oder t-Wert $> 2,00$ geben Hinweis auf einen Underfit.

Bei den Items war so gut erkennbar, dass die der MC-Version tendenziell zu einem Overfit neigten. Insbesondere bei den Items MC4, MC10, MC12 und MC14 war ein solcher anhand der Outfit $MeanSQ$- und/oder Outfit t-Werte konkret festzustellen. Bei MC4 und MC12 war dieser vor allem dadurch erklärbar, dass die Items kaum gelöst und dementsprechend bei manchen Teilungskriterien des Likelihood-Quotienten-Tests von Andersen ausgeschlossen wurden. Bei den Items der SQ-Version zeigte sich vor allem für die Items SQ2, SQ3, SQ5, SQ7 und SQ11 ein umgekehrtes Bild in Form eines Underfits bezogen auf die Outfit $MeanSQ$- und/oder Outfit t-Werte. Diese Unterschiede mussten allerdings auch vor dem Hintergrund unterschiedlicher Ratewahrscheinlichkeiten für Items beider Versionen gesehen werden, da diejenige der SQ-Version höher war. Es war demnach wahrscheinlicher, ein Item der SQ-Version durch Raten zu lösen als eines der MC-Version.

Tabelle 17
MC1 bis SQ14 ohne SQ4 und SQ12: Itemfit-Indizes

Item	Outfit *MeanSQ*	Infit *MeanSQ*	Outfit t	Infit t
MC1	0,935	0,966	-1,12	-0,86
MC2	0,859	0,926	-1,75	-1,37
MC3	1,088	1,084	0,72	1,01
MC4	0,568	0,853	-0,97	-0,37
MC5	0,883	0,939	-1,77	-1,34
MC6	0,948	0,909	-0,23	-0,78
MC7	0,919	0,954	-0,66	-0,57
MC8	1,188	1,084	2,54	1,73
MC9	0,994	1,027	-0,09	0,74
MC10	0,659	0,883	-2,25	-1,05
MC11	0,736	0,878	-2,31	-1,54
MC12	0,523	0,810	-1,74	-0,91
MC13	0,803	0,887	-1,79	-1,55
MC14	0,737	0,857	-2,01	-1,62
MC15	0,889	0,923	-1,82	-1,63
MC16	0,984	0,999	-0,20	0,00
SQ1	1,020	1,006	0,34	0,16
SQ2	1,179	1,165	3,12	4,09
SQ3	1,117	1,033	2,45	0,95
SQ5	1,257	1,128	2,31	1,77
SQ6	1,054	1,024	0,46	0,31
SQ7	1,184	1,165	2,69	3,35
SQ8	1,019	1,021	0,23	0,33
SQ9	0,841	0,866	-0,86	-1,15
SQ10	0,972	0,987	-0,42	-0,25
SQ11	1,391	0,990	1,30	0,06
SQ13	0,985	0,994	-0,17	-0,10
SQ14	0,781	0,825	-3,60	-3,74

7. Diskussion

7.1. Skalierung und Konstrukvalidität

Für die Items MC1 bis MC14 des TARV gilt das Rasch-Modell, was bedeutet, dass die Items homogen bzw. eindimensional sind. Streng genommen kann nur aufgrund dessen

allerdings keine Aussage getroffen werden, welches Merkmal die Items erfassen. Hinsichtlich der Konstruktion der Items nach bestimmten Facetten, die ihrerseits Ähnlichkeiten zu Raumvorstellungsfaktoren aufweisen, ist jedoch im Sinne der Inhaltlichen Gültigkeit[15] zu vermuten, dass der TARV Raumvorstellung erfasst. Hervorzuheben ist zudem, dass die Geltung des Rasch-Modells ohne Ausschluss von Items gelang. Damit wäre eine Kreuz-Validierung[16] an einem neuen Datensatz, wie sie Kubinger et al. (2011) bei einer lediglich a posteriori Geltung vorschlägt, nicht von Nöten.

Der Test erfüllt damit zwar die als Skalierung definierten Ziele, jedoch erwiesen sich die Items der MC-Version als deutlich zu schwierig für die Testpersonen, was durch den Vergleich der geschätzten Item- und Personenparameter erkennbar wurde. Die Folge davon ist, dass die Items nur gering zwischen den verschiedenen Merkmalsausprägungen differenzieren und somit wenig Information liefern. Kurz gesagt führt die Anwendung der MC-Version des TARV wegen seiner schwierigen Items zu Bodeneffekten. Daher ist, trotz der Geltung des Rasch-Modells, eine Umgestaltung der Items angebracht, mit dem Ziel, ihre Schwierigkeit zu senken. Dies jedoch gestaltet sich aufgrund des fehlenden Nachweises der Konstruktvalidität problematisch, denn so ist nicht ersichtlich, was die Schwierigkeit der Items ausmacht. Es ist also unklar, welche gedanklichen Operationen bei der Itembearbeitung tatsächlich erfolgen und wie dementsprechend die Itemschwierigkeit moduliert werden kann.

Die fehlende Konstruktvalidität dahingehend, dass keine drei unterschiedlichen Facetten nachgewiesen werden konnten, stellen infrage, inwiefern die Konzeptumsetzung der drei Facetten in dem Test gelang. Daher kann die Frage, ob Raumvorstellung ein ein- oder multidimensionales Merkmal ist, in dieser Arbeit nicht beantwortet werden. Die Geltung des Rasch-Modells gibt natürlich deutliche Hinweise auf die Eindimensionalität des Merkmals, dennoch ist, ohne Nachweis der Konstruktvalidität, diese Schlussfolgerung nicht hinreichend. Es wäre daher ebenso möglich, dass Raumvorstellung doch ein multidimen-

[15] "Von *Inhaltlicher Gültigkeit* eines Tests ist zu sprechen, wenn dieser selbst, quasi definitionsgemäß, das optimale Kriterium des interessierenden Merkmals darstellt.„ (Kubinger, 2009, S. 55)

[16] Anhand der geschätzten Parameter zweier Datensätze wird versucht, „die Daten des jeweils anderen Teils zu beschreiben bzw. vorherzusagen.„ (Kubinger et al., 2011, S. 508)

sionales Merkmal ist und der TARV, trotz Formulierung verschiedener Facetten, nur eine Dimension davon erfasst. In Betracht kommt dabei vor allem der Faktor Visualization (siehe Kapitel *3.7. Vergleich der Facetten des TARV mit den Raumvorstellungsfaktoren*). Entscheidendes Kriterium dieses Faktors ist, dass Testpersonen mehrere gedankliche Operationen bei komplexen Objekten durchführen müssen, um zur Lösung zu gelangen. In Kapitel *3.7. Vergleich der Facetten des TARV mit den Raumvorstellungsfaktoren* wurde dieser Sachverhalt, dass jedes Item den Faktor Visualization beinhaltet, bereits ausführlich dargestellt.

Erschwerend hinsichtlich der Differenzierung der Facetten kommt bei manchen Items hinzu, dass die Fehler womöglich nicht unabhängig voneinander entdeckt werden konnten. Das würde demzufolge auch heißen, dass die Facetten voneinander abhängig sind und mehrere mentale Operationen zueinander in Bezug gesetzt werden müssen. Abbildung 26 zeigt ein solches Item bzw. einen solchen Aufgabenstamm. Um dieses Item zu lösen, müssen Testpersonen erst die falsche Seitenansicht erkennen, um anschließend die falschen Rotationen der Objekte zu identifizieren. Diese notwendige Vorgehensweise, mehrere mentale Operationen schrittweise auszuführen, würde im Übrigen ebenso wieder dem Faktor Visualization entsprechen.

Die verschiedenen Versuche, das LLTM zur Geltung zu bringen, legen zudem die Vermutung nahe, dass das Arbeitsgedächtnis bei der Bearbeitung der MC-Version des TARV eine nicht zu verachtende Rolle spielt. Dies steht allerdings auch im Einklang mit der Literatur, wonach sich Testpersonen mit guter und schlechter Raumvorstellung auch darin unterscheiden, wie gut sie die Ergebnisse ihrer mentalen Operationen im Gedächtnis behalten können (Hegarty & Waller, 2005). Dabei geht es zum einen um die Qualität und zum anderen um die Dauer dieser mentalen Repräsentationen bestimmter Objekte. Diese Annahme wäre auch auf die MC-Version des TARV übertragbar, schließlich hängt das Lösen eines Items davon ab, mehrere gedankliche Operationen durchzuführen. Eine Testperson muss, je nach Darstellungsmodus, zuerst von 3D auf 2D oder umgekehrt schließen und somit eine entsprechende mentale Repräsentation der Objekte erzeugen.

Für die folgende Fehleridentifikation ist es entscheidend, diese Repräsentation permanent aufrechtzuerhalten. Eine noch größere Rolle spielt das Arbeitsgedächtnis womöglich dann, wenn, wie vorhin beschrieben, mentale Operationen zueinander in Bezug gesetzt werden müssen. Da im TARV kein Item mit nur einer mentalen Operation gelöst werden kann, ist ein Anteil des Arbeitsgedächtnisses beim Lösen der Items demnach nicht von der Hand zu weisen. Allerdings sei erwähnt, dass bei den Items beide Ansichten, die vorgegebene und die zu überprüfende, gleichzeitig dargestellt sind, was einer Überbeanspruchung des Arbeitsgedächtnisses entgegenwirkt.
Auch kann die Geltung des LLTM misslungen sein, weil zu wenige Basisparameter formuliert wurden. Hier gibt es zwei Möglichkeiten. Einerseits könnten neben den bestehenden Basisparametern, nämlich den drei Facetten, weitere notwendig sein, wie z. B. das Arbeitsgedächtnis. Andererseits könnten ebenso die bestehenden Basisparameter zu ungenau sein und müssten dementsprechend aufgegliedert werden. Dies läuft natürlich abermals dem Testkonzept zuwider, erscheint allerdings aus mehreren Gründen logisch. Der wichtigste Grund ist, dass die Items des TARV keine einheitliche Darstellung besitzen. Das heißt, die Items sind in ihren Objekten hinsichtlich Anzahl, Farbe und Form völlig verschieden. So ist nicht auszuschließen, dass eine Testperson ein Item nur deswegen nicht lösen kann, weil sie Probleme mit der Darstellung hat. Zum Beispiel könnte es sein, dass der falsche Winkel zweier Objekte zueinander nicht erkannt wird, weil sich die Testperson von einem anderen Objekt irritieren lässt. Möglich wäre auch, dass es Testpersonen schwerer fallen könnte, mentale Operationen an runden anstatt an eckigen Objekten durchzuführen. Nichtsdestotrotz könnte dies dazu führen, dass Testpersonen Items der gleichen Facette aufgrund anderer Aspekte, die nichts mit der Facette zu tun haben, lösen bzw. nicht lösen können. Die Q-Matrix des LLTM mit nur drei Basisparametern (Facetten) und den Gewichten „vorhanden“ bzw. „nicht vorhanden“ wird diesem Umstand nicht gerecht. In letzter Konsequenz bedeutet dies: Ob eine Testperson z. B. einen Rotationsfehler erkennt, hängt davon ab, ob die Testperson aufgrund der Darstellung des Items überhaupt in der Lage ist, die notwendige mentale Operation (hier: Das Objekt zu rotieren und mit dem

vorgegebenen zu vergleichen.) durchzuführen.

Ein weiterer Grund für eine Aufgliederung der drei Basisparameter ergibt sich durch den Vergleich mit der Literatur. So wurde bisher einzig beim Dreidimensionalen Würfeltest (Gittler, 1990) der Versuch unternommen, die Schwierigkeit der Items durch eine gewichtete Kombination von Basisparametern zu erklären. Das LLTM galt für diesen Test nicht, obwohl (Gittler, 1990) fünf Hauptbasisparameter formulierte, die sich in insgesamt 18 Basisparameter aufgliederten. Diese waren z. B. die Anzahl der Lösungsschritte oder auch die Berücksichtigung der Symmetrieachsen. Zudem waren diese äußerst eng an die Items angelehnt. Das heißt, es existierten Basisparameter, deren Gewichtung nur auf einem Item lag. Auch wenn dieser Test dem TARV in seiner Darstellung nicht ähnelt, so wird deutlich, dass es für den Nachweis der Konstruktvalidität womöglich notwendig ist, die Basisparameter weiter und insbesondere itemspezifischer aufzugliedern. Allerdings sei erwähnt, dass weiterhin weniger Basisparameter formuliert werden müssen, als der TARV Items besitzt. Dies würde sonst der grundlegenden Annahme des LLTM zuwiderlaufen.

Weitere Gründe für den fehlenden Nachweis der Konstruktvalidität könnten auch die sprachlich formulierten Antwortmöglichkeiten sowie die fehlende Berücksichtigung des Darstellungsmodus sein. Ersteres würde bedeutet, dass das Lösen eines Items auch davon abhängt, die Antwortmöglichkeit verbal zu verstehen sowie die zutreffende/n zu identifizieren. Hier sei allerdings erwähnt, dass durch die Rahmenbedingungen der Online-Plattform visuelle Antwortmöglichkeiten nicht verwendet werden konnten. Letzteres, der Darstellungsmodus, konnte bei der Überprüfung der Konstruktvalidität wegen der Modellvoraussetzungen des LLTM (siehe Kapitel *5.4.2. Linear Logistische Test-Modell von Fischer (1973)* nicht berücksichtigt werden. Es ist damit völlig unklar, welchen Einfluss das Schließen von 2D auf 3D und umgekehrt auf die Itemschwierigkeit hat. Zu vermuten ist nur, dass es den Testpersonen schwerer fällt, sich die notwendige Seitenansicht eines Gebildes anhand von zwei Planansichten (2D auf 3D) mental zu repräsentieren, anstatt von einem Gebilde auf eine Planansicht (3D auf 2D) zu schließen. Im ersten Fall muss eine Testperson eine Ansicht aus zwei Ansichten mental erst konstruieren. Im zweiten Fall

muss eine Ansicht von einer anderen Ansicht eher abgeleitet werden. Ein Vergleich der geschätzten Itemparameter für Items beider Darstellungsmodi lässt jedoch diesen Schluss nicht zu. Hier ist allerdings erneut darauf hinzuweisen, dass die völlig verschiedenen Darstellungen der Items keine Aussagen ermöglichen. Es kann z. B. je nach Form der Objekte schwieriger sein, von einer Ansicht auf die andere zu schließen. Nichtsdestotrotz stellt die Nicht-Berücksichtigung der zwei Darstellungsmodi einen Kritikpunkt des methodischen Vorgehens dieser Arbeit dar.

7.2. Fairness

Ein wichtiges Anliegen des TARV ist es, ein fairer Test zu sein. Das heißt zum einen, dass sich die geschätzten Itemparameter zwischen den Teilstichproben nicht statistisch signifikant unterscheiden sollen. Dies trifft durch die Geltung des Rasch-Modells auf die Items zu. Zum anderen sollten im besten Falle für die Teilstichproben keine separaten Eichtabellen verwendet werden müssen. Zwar wurden, abgeleitet aus der Literatur, Hypothesen formuliert, die von einem Unterschied zwischen den Geschlechtern und Personen verschiedener Ausbildungen ausgingen, allerdings wurden diese aufgrund der Ergebnisse verworfen.

Das Gütekriterium der Fairness ist damit jedoch nicht vollständig erfüllt, da erneut auf die hohe geschätzte Schwierigkeit der Items hingewiesen werden muss. Dies führt dazu, dass sich die Wertebereiche der geschätzten Item- und Personenparameter kaum decken und somit die Items nicht zwischen den Merkmalausprägungen der Testpersonen differenzieren. Verdeutlicht wird dies dadurch, dass der geschätzte Personenparameter von 191 von insgesamt 298 Testpersonen unter -1,5 liegt, während der niedrigste geschätzte Itemparameter -1,454 ist. Schlussfolgerungen aus den Hypothesen sind damit nicht möglich, da zudem nicht auszuschließen ist, dass es mit leichteren Items zu Unterschieden zwischen den Geschlechtern und Personen verschiedener Ausbildungen gekommen wäre. Dies würde eher der bisherigen Forschung entsprechen, die Geschlechtsunterschiede bei verschiedenen Raumvorstellungstests und Unterschiede bei Personen verschiedener Ausbildungen fan-

den (siehe Kapitel *2. Raumvorstellung: Differentielle Forschungsrichtung*). Auch besitzen die Items des TARV Charakteristika, die einen Geschlechtsunterschied begünstigen (siehe Kapitel *2.2.2. Test- und Itemcharakteristika als mögliche Ursachen des Geschlechtsunterschieds*), weswegen die nicht ermittelten statistisch signifikanten Unterschiede in dieser Arbeit nicht endgültig interpretiert werden können.

Kritisch sei hier angemerkt, dass mehr Männer als Frauen den TARV bearbeiteten. Insbesondere in Kombination mit der Ausbildung zeigte sich dabei ein deutliches Ungleichgewicht (siehe Tabelle 7). So stellten Schülerinnen, die eine HTL besuchten, die kleinste Gruppe dar, die allerdings die höchsten Testwerte erzielte. Umgekehrt waren die Schüler einer HTL die größte Gruppe mit den geringsten Testwerten. Auch wenn, wie bereits erwähnt, wegen der hohen geschätzten Itemparameter kaum Aussagen getroffen werden können, so wird hier im Ansatz versucht, eine Erklärung für diese unerwarteten deskriptiven Unterschiede zu geben.

Bei den Frauen könnten dabei womöglich Selektionsprozesse eine Rolle gespielt haben. Es ist zu vermuten, dass der generell hohe Anteil an Männern an einer naturwissenschaftlich-technischen Schule auch auf entsprechende Stereotypen und Rollenbilder zurückzuführen ist, wonach diese Fachgebiete eher Männern liegen. Dies könnte dazu führen, dass Frauen, die die notwendige Eignung für diese Schule besitzen, sie nicht besuchen. Ebenso wäre vorstellbar, dass dadurch eher die Frauen eine HTL besuchen, die mehr als die notwendige Grundeignung dafür mitbringen. In Kapitel *2.3. Unterschiede in der Raumvorstellung: Ausbildung* wurde dargestellt, dass SchülerInnen einer naturwissenschaftlich-technischen Schule bessere Leistungen in Raumvorstellungstests zeigen. Auch wenn sich dies, womöglich durch die hohen geschätzten Itemparameter, in dieser Arbeit nicht ergab, so ist für die hier getesteten 17 Schülerinnen einer HTL vorstellbar, dass sie aufgrund von Selektionsprozessen eine durchschnittlich bessere Raumvorstellung besitzen als ihre männlichen Mitschüler. Eine Erklärung allerdings, weswegen die Gruppe männlicher Schüler einer HTL den niedrigsten durchschnittlichen geschätzten Personenparameter aller Gruppen aufweist, kann nicht gegeben werden. Eine mögliche Vermutung beträfe Un-

terschiede im Arbeitsstil bzw. den Bearbeitungsstrategien. Glück und Fabrizii (2010) wiesen darauf hin (siehe Kapitel *2.2.2. Test- und Itemcharakteristika als mögliche Ursachen des Geschlechtsunterschieds*), dass Männer bei begrenzter Bearbeitungszeit „quick-and-dirty“ (S. 109) Strategien verwendeten, um so möglichst viele Items bearbeiten zu können, während Frauen ihre Antworten hingegen mehrmals überprüften. Diese Vorgehensweise der Männer könnte sich bei den Items der MC-Version nachteilig ausgewirkt haben, die mehrere Fehler beinhalteten, aufgrund der Bearbeitungsstrategie jedoch nur einzelne erkannt wurden. Nichtsdestotrotz bleibt ungeklärt, weswegen sich bei SchülerInnen einer AHS das genau umgekehrte Bild hinsichtlich der geschätzten Personenparameter zeigt. Auch widersprechen die ähnlichen durchschnittlichen Gesamtbearbeitungszeiten des TARV für SchülerInnen einer AHS wie HTL der Annahme, dass weibliche und männliche Testpersonen unterschiedliche Bearbeitungsstrategien verwendeten, die unterschiedlich viel Zeit beanspruchen. Schülerinnen einer HTL bzw. AHS hatten eine durchschnittliche Gesamtbearbeitungszeit von 14,27 bzw. 16,39 Minuten mit einer Standardabweichung von 4,15 bzw. 4,45 Minuten. Schüler einer HTL bzw. AHS benötigten durchschnittlich 15,73 bzw. 16,43 Minuten mit einer Standardabweichung von 4,62 bzw. 4,84 Minuten. Zu erwähnen ist hier, dass ausschließlich die Bearbeitungszeiten berücksichtigt wurden, wenn die SchülerInnen die vorgegebene und die zu überprüfende Ansicht bei Items miteinander verglichen. Die Bearbeitungszeiten für die Instruktion sowie für die Darstellung der ausschließlich vorgegebenen Ansicht, um sich mit dem Item vertraut zu machen, wurden nicht bei allen Testpersonen erfasst und daher bei der Berechnung nicht berücksichtigt. Die tatsächliche durchschnittliche Gesamtbearbeitungszeit liegt somit höher.

7.3. Vergleich der MC- und SQ-Version des TARV

Ein Nachteil dieser Arbeit ist, dass aufgrund der begrenzten Gesamtbearbeitungszeit von einer Schulstunde nur vier SQ-Items von ca. 200 Testpersonen bearbeitet werden konnten. Des Weiteren war durch die fehlende Konstruktvalidität und der zu hohen Schwierigkeit der Items der MC-Version, die demzufolge erst eine Umgestaltung benötigen, ein Vergleich

beider Versionen nur bedingt möglich.

Hinsichtlich der Bearbeitungszeit zeigte sich, dass Testpersonen im Durchschnitt für ein Item der MC-Version länger benötigten als für eines der SQ-Version. Dies würde allerdings nicht dem Konstruktionsziel der MC-Version widersprechen, weniger zeitaufwändig zu sein. Die mangelnde Ökonomie der SQ-Version liegt auch darin, dass ein Item aktuell nur eine Facette beinhaltet. Diesen Umstand versuchte die MC-Version bekanntermaßen zu ändern. Auch wenn die Items der MC-Version dadurch mehr Zeit beansprucht hätten, so hätten durch ein Item mehrere Facetten erfasst werden können, weswegen im Vergleich zur SQ-Version auch insgesamt weniger Items benötigt worden wären. Durch die fehlende Konstruktvalidität der MC-Version muss allerdings konstatiert werden, dass dieses Konstruktionsziel nicht erreicht wurde. Zudem erfassen beide Versionen aufgrund der Geltung des Rasch-Modells dasselbe Merkmal, weswegen die MC-Version nicht als ökonomischer betrachtet werden kann.

Der Vergleich der geschätzten Itemparameter einander korrespondierender Items der MC- wie SQ-Version legt nahe, dass die Lösung von Items, die nur einen Fehler beinhalten, den Testpersonen leichter fallen. Auch wenn das LLTM nicht gilt und damit auch nicht die Facettenstruktur des Testkonzepts nachgewiesen werden konnte, so ist ansatzweise zu vermuten, dass Items, die mehrere Fehler (verschiedener Facetten) beinhalten, schwieriger zu lösen sind. Bei Items der SQ-Version sowie bei Items der MC-Version, die nur einen Fehler beinhalten, ist es zum Lösen des Items eben nur notwendig, diesen einen Fehler zu entdecken. Die Schwierigkeit der restlichen Items der MC-Version könnte darin liegen, mehrere Fehler zu erkennen bzw. gegebenenfalls in Bezug zueinander zu setzen. Dies führt zurück auf den bereits erwähnten Punkt, mehrere gedankliche Operationen zur Lösung von Items der MC-Version bewerkstelligen zu müssen.

Hinsichtlich der Itemfit-Maße weisen die Items der MC-Version darauf hin, dass sie tendenziell zu sehr zwischen den Merkmalsausprägungen der Testpersonen trennen. Die ist allerdings auch durch die geringen Lösungshäufigkeiten der Items dieser Version bedingt. Es bleibt daher ausstehend, ob sich dies mit Items geringerer Schwierigkeit verbessert.

Die Items der SQ-Version zeigen im Vergleich zu den Items der MC-Version ein weniger deterministisches Antwortmuster, womöglich auch durch die deutlich höhere Ratewahrscheinlichkeit beeinflusst.

7.4. Fazit

Die in dieser Arbeit überprüfte MC-Version des TARV sollte in dieser Form nicht als Beratungstest für technische Studiengänge verwendet werden. Zwar gilt für die Items das Rasch-Modell, dennoch haben sich die Items als zu schwierig erwiesen. Die Folge ist, dass die Items nur gering zwischen den Merkmalausprägungen der Testpersonen differenzieren. Die MC-Version wird damit ihrem Geltungsbereich, Informationen über die Eignung für ein technisches Studium zu geben, nicht gerecht. Die Schwierigkeit der Items ist insofern problematisch, da anzunehmen ist, dass gerade bei Beratungszwecken Studieninteressierte sich ihrer Stärken und Schwächen hinsichtlich eines Studiums unsicher sind. Über die Eignung für ein Studium können die Items derzeit allerdings nur bedingt Auskunft geben, da es zu große Bodeneffekte gäbe.
Das Ergebnis, dass sich keine statistisch signifikanten Unterschiede zwischen den Geschlechtern und Personen verschiedener Ausbildungen zeigten, lässt vor dem Hintergrund der hohen geschätzten Itemparameter kaum Rückschlüsse zu. Den Test als fairen Test aufgrund ähnlicher Itemschwierigkeiten und fehlender Unterschiede zwischen den Teilstichproben zu bezeichnen, ist aufgrund dieser Tatsache nicht zur Gänze möglich.
Des Weiteren konnte die Konstruktvalidität nicht nachgewiesen werden, was somit das Testkonzept in Frage stellt. Es bleibt daher streng genommen unklar, welches Merkmal die Items des TARV erfassen und wie sich die Schwierigkeit der Items modulieren lässt. Damit kann auch die Frage, ob Raumvorstellung ein ein- oder multidimensionales Merkmal ist, nicht beantwortet werden.
Der Vergleich beider Versionen kommt zum Schluss, dass die Items der SQ-Version geringfügig weniger Zeit benötigen, weniger deterministische Itemfit-Maße aufweisen sowie, verglichen mit Items der MC-Version mit mehreren zu identifizierenden Fehlern, leich-

ter sind. Vor dem Hintergrund, dass für beide Versionen das Rasch-Modell gilt und sie demnach das gleiche Merkmal messen, macht dies eine Umgestaltung der MC-Version dringend notwendig. Im Folgenden werden daher Ansätze gegeben, wie dies für weitere Arbeiten zu bewerkstelligen ist.

Positiv ist allerdings hervorzuheben, dass auch diese Arbeit den Vorteil theoriebildender Verfahren und dabei insbesondere den der Modellen der Item-Response-Theorie[17] zeigt. Zwar konnte mit dem LLTM die Konstruktvalidität des TARV nicht nachgewiesen werden, dennoch gilt für die Items das Rasch-Modell, was ungeachtet inhaltlicher Aspekte die Aussage erlaubt, dass der Test eindimensional ist. Dabei ist zu vermuten, dass es sich um die Dimension der Raumvorstellung handelt. Ein Nachweis diesbezüglich steht jedoch aus. Die in Kapitel *1.2.1. Kritische Auseinandersetzung hinsichtlich mehrerer Faktoren der Raumvorstellung* beschriebenen Probleme bei Anwendung einer explorativen Faktorenanalyse, nämlich die Anzahl der voneinander unabhängigen Faktoren zu bestimmen sowie diese inhaltlich zu erklären, ergeben sich somit nicht.

7.5. Ansätze zur Umgestaltung der Items der MC-Version

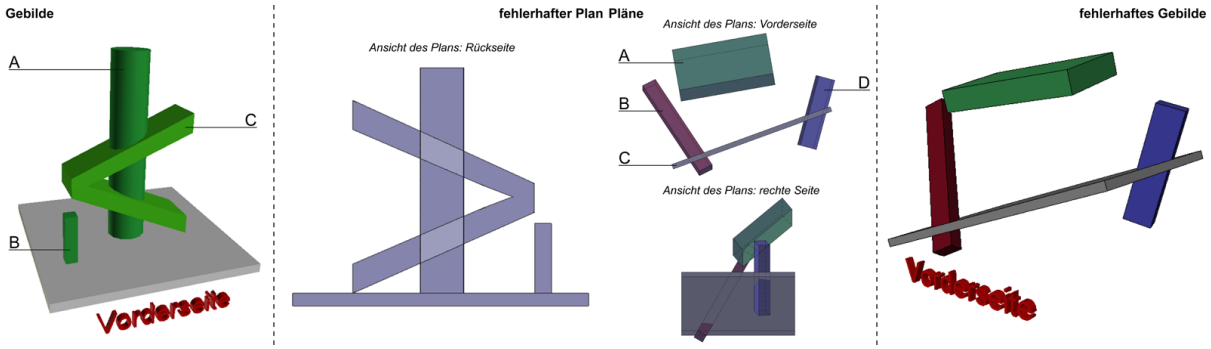
Durch die fehlende Konstruktvalidität ist nicht ersichtlich, was die Schwierigkeit der einzelnen Items der MC-Version ausmacht. Der Vergleich mit den Items der SQ-Version lässt vermuten, dass die Anzahl an zu identifizierenden Fehlern eine Rolle spielt. Dies könnte womöglich dazu führen, dass mehr mentale Operationen je Item durchzuführen sind, um zur Lösung zu gelangen. Dass ein Item mehrere Fehler beinhalten kann, ist jedoch der Grundgedanke der MC-Version, sodass die Itemschwierigkeit auf diesem Wege nicht gesenkt werden kann. Auch die Verwendung verbal formulierter Antwortmöglichkeiten ist aufgrund der Rahmenbedingungen der Online-Plattform nicht zu ändern.

Hauptaugenmerk für eine Änderung liegt somit in der Darstellung der Items, die, wie bereits beschrieben, sich für jedes Item deutlich unterscheidet. Es ist zu vermuten, dass die

[17] „Wahrscheinlichkeitsfunktion zur Beschreibung der Wahrscheinlichkeit für das Auftreten einer bestimmten Reaktionskategorie in Abhängigkeit von der fraglichen Eigenschaft einer Person und in Abhängigkeit von gewissen Charakteristika des Items.“ (Kubinger et al., 2011, S. 555)

Schwierigkeit, Fehler der jeweiligen Facette zu identifizieren, durch die Art und Weise der Darstellung beeinflusst wird. An dieser Stelle erfolgt daher ein Vergleich der am leichtesten mit den am schwierigsten geschätzten Items der MC-Version, um mögliche Aspekte zu identifizieren, die die Schwierigkeit beeinflussen könnten. Diese Aspekte können auch als mögliche Basisparameter verstanden werden, im Hinblick auf zukünftige Versuche, die Konstruktvalidität nachzuweisen.

Die zwei am leichtesten geschätzten Items sind die Items MC1 und MC9 und die am schwierigsten geschätzten Items MC4 und MC12. Die Items MC1 und MC9 beinhalten je einen Fehler. Bei Item MC1 hat Objekt A die falsche Größe (Facette Relationen) und bei Item MC9 ist Objekt A falsch geneigt (Facette Rotation). Bei den Items MC4 und MC12 sind mehrere Fehler zu identifizieren. Bei Item MC4 ist der Abstand zwischen Objekt B und C (Facette Relationen) sowie die Neigung von Objekt A (Facette Rotation) falsch. Für Item MC12 sind die Position von Objekt B (Facette Orientierung) sowie jeweils die Größe von Objekt A und C (Facette Relationen) falsch.

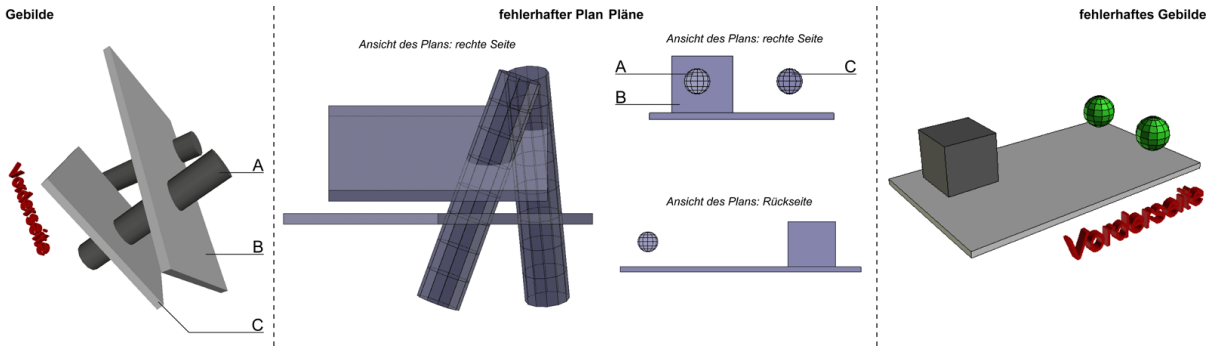


(a) MC1 (b) MC9

Abbildung 37. Items mit den niedrigsten geschätzten Itemparametern.

Bei Items MC1 und MC9 fällt auf, dass die fehlerhaften Objekte relativ freistehend sind, während sie vor allem bei Item MC4 ein zusammenhängendes Gebilde formen. Dadurch stehen wiederum die zu identifizierenden Fehler bei diesem Item zueinander in Bezug. Es ist zu vermuten, dass durch den falschen Abstand von Objekt B und C es einer Testperson schwerer fällt, die falsche Neigung von Objekt A zu erkennen oder eben umgekehrt. Auch

könnte die Neigung der hinteren nicht markierten Säule die Wahrnehmung der Neigung von Objekt A beeinflussen.



(a) MC4 (b) MC12

Abbildung 38. Items mit den höchsten geschätzten Itemparametern.

Des Weiteren geben die Items Aufschluss auf die Rolle der dargestellten Seite. So könnte für Items MC1 und MC9 die niedrige geschätzte Schwierigkeit auch davon beeinflusst sein, dass sich für beide Items das Schließen von der vorgegebenen auf die zu überprüfende Seitenansicht leichter gestaltet. Für Item MC1 ist dies womöglich der Fall, da die Objekte weniger ineinander verschlungen sind wie bei Item MC4. Noch einfacher könnte es sich bei Item MC9 gestalten, da bei den Planansichten bereits die Vorderansicht dargestellt ist, auf die in der Gebildeansicht zu schließen ist. Testpersonen könnten hier direkt die Objekte miteinander vergleichen und ihre Entscheidung, ob ein Objekt fehlerhaft dargestellt ist oder nicht, durch die zweite Planansicht gegenprüfen. Das heißt, für Item MC9 wäre es daher nicht nötig, eine tatsächliche mentale Repräsentation der Gebildeansicht durch die zwei Planansichten zu erstellen. Bei Item MC4 könnte weniger die Seitenansicht die Schwierigkeit darstellen, als die Tatsache, dass die Gebildeansicht rotiert ist. Testpersonen müssen daher nicht nur von der Vorderseite auf die rechte Seite schließen, sondern zusätzlich, je nach Bearbeitungsstrategie, entsprechend das Gebilde kippen bzw. sich gedanklich um das Gebilde bewegen.

Die Schwierigkeit für Item MC12 liegt hingegen wohl vielmehr am ungleichen Maßstab beider Planansichten. So sind die Objekte A und C in der oberen Planansicht größer als in

der unteren. Dies könnte es für die Testpersonen erschwert haben, festzustellen, dass die Objekte in der Gebildeansicht tatsächlich zu groß dargestellt sind. Eine Bestätigung dafür geben die absoluten Häufigkeiten der ausgewählten Antwortkategorien. So haben 120 von 193 Testpersonen, die dieses Item bearbeiten haben, die falsche Position von Objekt B erkannt, allerdings nur 50 von 193 Testpersonen die falschen Größen von Objekt A und C.

Der Vergleich dieser Items liefert allerdings nur bedingt Aufschluss über mögliche Einflussgrößen auf die Itemschwierigkeit. Für die hohen geschätzten Itemparameter für Item MC4 und MC12 ist allerdings anzunehmen, dass diese mehr durch die Darstellung und weniger durch die eigentlichen Fehler bedingt sind, gemäß derer die Items konstruiert wurden. Des Weiteren könnte die Schwierigkeit sämtlicher Items auch von den gegebenen Seiten in der Plan- wie Gebildeansicht abhängen. Zum Beispiel könnten Testpersonen Items leichter fallen, wenn in der vorgegeben Ansicht bereits die zu überprüfende zu sehen wäre. Dies wäre zum Beispiel auch der Fall, wenn die Vorderseite wie rechte Seite einer Gebildeansicht dargestellt wären und die rechte Seite in der Planansicht beurteilt werden müsste. Item MC12 ist im Prinzip ein solches, wenn das Gebilde nicht zusätzlich gedreht wäre. Auf den möglichen Einfluss zusammenhängender Objekte, die zudem verschiedene Fehler beinhalten, wurde bereits hingewiesen.

Für zukünftige Versuche, eine MC-Version des TARV zu erstellen, wird daher Folgendes empfohlen: Eine Analyse der Häufigkeiten der gegebenen Antwortmöglichkeiten je Item würde, insbesondere bei Items mit mehreren fehlerhaften Objekten, Aufschluss darüber geben, welche Fehler wie oft erkannt werden. Dies gäbe erste Anhaltspunkte, ob nicht die Darstellung der Items das Erkennen eines Fehlers verhindert. Auch könnte in diesem Zusammenhang hinsichtlich des Antwortformats geprüft werden, ob den Testpersonen die Lösung der Items leichter fällt, wenn die Anzahl an zu identifizierenden Fehlern je Item vorgegeben wird. Des Weiteren würde die Bearbeitung beider Versionen des TARV durch *lautes Denken*[18] Informationen geben, welche und vor allem wie viele mentalen Operatio-

[18] „[Testpersonen sollen] während der Testbearbeitung möglichst alle „inneren Vorgänge“, Überlegungen und Einfälle (...) verbalisieren – und dabei weniger auf die Testleistung z. B. im Sinne einer rasche

nen bei der Bearbeitung der Items durchzuführen sind. Auf Basis dieser Informationen könnte versucht werden, leichtere Items für die MC-Version zu erstellen bzw. die vorhandenen entsprechend zu ändern. Im Vordergrund sollte dabei stehen, die Darstellung der einzelnen Items stärker einander anzupassen. Dies wäre bewerkstelligt, indem in der vorgegebenen Ansicht bei einem Gebilde wie bei Plänen stets die gleichen Seitenansichten zu sehen sind. Zudem sollten diese Ansichten keine zusätzlichen Drehungen wie bei Item MC4 beinhalten, sofern diese nicht für die Facette Rotation beabsichtigt sind. Durch die gewonnenen Erkenntnisse bei der Testbearbeitung mittels lautem Denken könnten weitere Basisparameter formuliert werden. Dies würde auch Hinweise darauf geben, ob neben den Facetten weitere Basisparameter zu erstellen sind, oder ob die Facetten zusätzlich weiter aufgegliedert werden müssen. Auch könnte es so möglich sein, den Darstellungsmodus zu berücksichtigen. Dies wäre zum Beispiel möglich, wenn ermittelt werden würde, welche und wie viele mentale Operationen bei Schließen von 3D auf 2D und umgekehrt notwendig sind. Damit wären die Basisparameter sowie deren Gewichtung je Item gegeben.
Kann dadurch abermals mittels des LLTM die Konstruktvalidität nicht nachgewiesen werden, sollten zumindest im Sinne der Kriteriumsvalidität[19] Korrelationskoeffizienten mit anderen Raumvorstellungstests (Übereinstimmungsvalidität) oder mit einem anderem Außenkriterium (prognostische Validität, z. B. Studienerfolg in technischen Studiengängen) berechnet werden. Dies dient dem Nachweis, dass die Items des TARV im Falle, dass das Rasch-Modell gilt, auch Raumvorstellung erfassen bzw. im weitesten Sinne auch, welchen Faktor der Raumvorstellung sie erfassen. So könnte erreicht werden, dass die MC-Version des TARV schließlich ihren Geltungsbereich erfüllt, nämlich Studieninteressierten Auskunft über ihre Stärken und Schwächen zu geben und damit eine Entscheidungshilfe für sie zu sein.

Bearbeitung achten –, um eventuell validitätsmindernde Testeigenschaften zu erkennen.“ (Kubinger, 2009, S. 64)

[19] “Eine bestimmte als relevant angesehen Variable (sog. “Außenkriterium„) wird mit dem interessierenden Test korreliert.„ (Kubinger, 2009, S. 64)

8. Zusammenfassung

Kapitel *1. Raumvorstellung: Psychometrische Forschungsrichtung* zeigte, dass bis heute Unklarheit herrscht, ob Raumvorstellung ein ein- oder multidimensionales Merkmal ist. Verschiedene Faktoren der Raumvorstellung konnten bisher ermittelt und zum Teil mehrmals repliziert werden. Eine allgemeine akzeptierte Definition von Raumvorstellung hinsichtlich Art und Anzahl der Faktoren fehlt allerdings bis dato. Dies wurde unter anderem auch auf ein methodische Problem zurückgeführt (Kapitel *1.2.1. Kritische Auseinandersetzung hinsichtlich mehrerer Faktoren der Raumvorstellung*), nämlich der Anwendung explorativer Faktorenanalysen. Damit war es möglich, Faktoren ohne ein zugrunde liegendes theoriebasiertes Modell zu identifizieren. Die Folge davon war, dass die Faktoren inhaltlich nicht exakt voneinander getrennt werden konnten, damit also ungenügend definiert waren, und sie sich nur bei Verwendung bestimmter Tests ergaben.

Dieser Umstand führte dazu, dass die Ergebnisse von Studien zu Geschlechtsunterschieden in der Raumvorstellung von den verwendeten Tests bzw. den angenommenen Faktoren abhingen (Kapitel *2.2.1. Kritik an der differentiellen Forschungsrichtung zu Geschlechtsunterschieden*). Nichtsdestotrotz zeigte sich über alle Studien hinweg eine tendenziell bessere Leistung der Männer (Kapitel *2.2. Unterschiede in der Raumvorstellung: Geschlecht*). Des Weiteren wurde auf Test- und Itemcharakteristika eingegangen, die diesen Unterschied begünstigten (Kapitel *2.2.2. Test- und Itemcharakteristika als mögliche Ursachen des Geschlechtsunterschieds*).

Bezogen auf den Einfluss der Ausbildung zeigte sich, dass Personen mit einer naturwissenschaftlich-technisch geprägten Ausbildung bessere Testwerte erzielten (Kapitel *2.3. Unterschiede in der Raumvorstellung: Ausbildung*). Dies führte zum Schluss, dass Raumvorstellung durch die Ausbildung indirekt gefördert würde. Offen blieb jedoch, ob die Ausbildung oder das Geschlecht einen größeren Einfluss auf die Raumvorstellung besaß.

Aufbauend auf diesen Ergebnissen erfolgte die Beschreibung des TARV, der im Rahmen der Wiener Self-Assessments für die Studiengänge Architektur und Maschinenbau Verwendung finden sollte (Kapitel *3.1. Rahmenbedingungen des und Ansprüche an den Test*).

Der TARV definierte mehrere Facetten, die den Faktoren ähnelten (Kapitel *3.7. Vergleich der Facetten des TARV mit den Raumvorstellungsfaktoren*), jedoch möglicherweise keine eigenen Dimensionen der Raumvorstellung darstellten, sondern stattdessen die gedankliche Operationen, die zum Lösen der Items notwendig waren (Kapitel *3.3. Konzept*). Es wurde zum einen versucht, dieses Testkonzept zu validieren, und zum anderen, die Items der MC-Version für einen eventuell späteren Verwendungszweck als Self-Assessment zu kalibrieren (Kapitel *3.2. Ziele*). Für diesen Zweck wurde eine Version des TARV mit Multiple-Choice-Antwortformat (MC) überprüft, da sich die bereits bestehende mit sequentiellem Antwortformat (SQ) hinsichtlich ihrer Bearbeitungszeit als unökonmisch erwies (Kapitel *3.4. Items*). Beide Versionen wurden miteinander verglichen. Zusätzlich wurden in Anlehnung an die bisherigen Forschungsergebnisse Hypothesen formuliert, wonach Männer und Personen mit naturwissenschaftlicher-technischer Ausbildung bessere Testwerte im TARV erzielen würden (Kapitel *4. Hypothesen im Rahmen der Kalibrierung*).
Die Items beider Versionen wurden anschließend ausreichend zusammenhängend auf drei Testhefte verteilt (Kapitel *5.1. Testdesign*) und von insgesamt 298 Testpersonen aus fünf verschiedenen Schulen Österreichs bearbeitet (Kapitel *5.3. Stichprobe*). Die Testungen erfolgten als Gruppentestungen in den Klassenzimmern oder Computersälen der Schulen (Kapitel *5.2. Durchführung der Testung*). Die SchülerInnen teilten sich gleichmäßig hinsichtlich der beiden Schultypen (AHS und HTL) auf, allerdings wurden deutlich mehr Männer als Frauen getestet. Dies war zurückzuführen auf die ungleiche Geschlechterverteilung innerhalb des Schultyps HTL. Bei der statistischen Auswertung der Daten wurden aufgrund der Methodenkritik an der bisherigen Raumvorstellungsforschung spezielle theoriebildende Verfahren (Rasch-Modell, LLTM) verwendet (Kapitel *5.4. Theoriebildende Verfahren*).
Für die MC-Version galt das Rasch-Modell ohne Ausschluss von Items. Diese erwiesen sich allerdings als zu schwierig, was zur Folge hatte, dass die Items nur gering zwischen den Merkmalausprägungen der Testpersonen differenzierten (Kapitel *6.1.1. Überprüfung der Geltung des Rasch-Modells (Skalierung)*). Des Weiteren konnte die Geltung des LLTM

und damit die Konstruktvalidität nicht nachgewiesen werden (Kapitel *6.1.2. Überprüfung der Geltung des LLTM (Konstruktvalidität)*). Hinsichtlich der zusätzlichen Hypothesen zeigten sich keine statistisch signifikanten Unterschiede zwischen den Geschlechtern und zwischen Personen verschiedener Ausbildungen, weswegen beide Alternativhypothesen verworfen werden mussten (Kapitel *6.1.3. Vergleich zwischen verschiedenen Gruppen von Testpersonen (Fairness)*). Der Vergleich zwischen der MC- und SQ-Version führte zum Ergebnis, dass die geschätzten Itemparameter der Items der SQ-Version, verglichen mit Items der MC-Version mit mehreren zu identifizierenden Fehlern, geringer war, die Items weniger Zeit zur Bearbeitung benötigten sowie geringere deterministische Itemfit-Maße aufwiesen (Kapitel *16. Items MC1 bis SQ14: Vergleich der MC- und SQ-Version des TARV*).

Das Testkonzept konnte somit durch die fehlende Konstruktvalidität nicht nachgewiesen werden. Damit blieb auch unklar, welches Merkmal der TARV eindimensional durch die Geltung des Rasch-Modells maß, wenngleich aufgrund der Konstruktion der Items das Merkmal Raumvorstellung zu vermuten war. Die Frage, ob Raumvorstellung ein ein- oder multidimensionales Merkmal ist, konnte allerdings nicht beantwortet werden (Kapitel *7.1. Skalierung und Konstrukvalidität*). Aufgrund der hohen geschätzten Itemparamter waren auch kaum Aussagen über die nicht vorhandenen Unterschiede zwischen den Geschlechtern und Personen verschiedener Ausbildungen möglich (Kapitel *7.2. Fairness*). Der Vergleich beider Versionen, der MC- und SQ-Version, legte den Schluss nahe, dass eine Umgestaltung der Items der MC-Version notwendig sein würde, damit die MC-Version für das Self-Assessment verwendet werden könnte (Kapitel *7.3. Vergleich der MC- und SQ-Version des TARV*). Hierfür wurden erste Ansätze formuliert, die darauf abzielten, die Items der MC-Version in ihrer Darstellung stärker einander anzupassen sowie die Facetten des Testkonzepts im Sinne der Konstruktvalidität weiter aufzugliedern (Kapitel *7.5. Ansätze zur Umgestaltung der Items der MC-Version*).

Literatur

Alexander, W. P. (1935). Intelligence, concrete and abstract. A study in differential traits [Supplement]. *The British Journal of Psychology. Monograph Supplements*, *19*.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140. doi:10.1007/BF02291180

Baddeley, A. D. & Hitch, G. (1974). Working memory. *The Psychology of Learning and Motivation*, *8*, 47–89. doi:10.1016/S0079-7421(08)60452-1

Birenbaum, M., Kelly, A. E. & Levi-Keren, M. (1994). Stimulus features and sex differences in mental rotation test performance. *Intelligence*, *19*, 51–64. doi:10.1016/0160-2896(94)90053-1

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2. Aufl.). Mahwah, NJ: Lawrence Erlbaum.

Bors, D. A. & Vigneau, F. (2011). Sex differences on the mental rotation test: An analysis of item types. *Learning and Individual Differences*, *21*, 129–132. doi:10.1016/j.lindif.2010.09.014

Bundesministerium für Unterricht, Kunst und Kultur. (2008, 13. März). *Allgemeinbildende höhere Schulen (AHS).* Verfügbar unter `http://bmukk.gv.at/schulen/bw/abs/ahs.xml`

Bundesministerium für Unterricht, Kunst und Kultur. (2011, 19. Juli). *Technische, gewerbliche und kunstgewerbliche Schulen.* Verfügbar unter `http://bmukk.gv.at/schulen/bw/bbs/tgkg.xml`

Burnett, S. A. & Lane, D. M. (1980). Effects of academic instruction on spatial visualization. *Intelligence*, *4*, 233–242. doi:10.1016/0160-2896(80)90021-5

Caplan, P. J., MacPherson, G. M. & Tobin, P. (1985). Do sex-related differences in spatial abilities exist? A multilevel critique with new data. *American Psychologist*, *40*, 786–799. doi:10.1037/0003-066X.40.7.786

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

Collins, D. W. & Kimura, D. (1997). A large sex difference on a two-dimensional

mental rotation task. *Behavioral Neuroscience*, *111*, 845–849. doi:10.1037//0735-7044.111.4.845

Colom, R., Contreras, M. J., Botella, J. & Santacreu, J. (2001). Vehicles of spatial ability. *Personality and Individual Differences*, *32*, 903–912. doi:10.1016/S0191-8869(01)00095-2

Cox, J. W. (1928). *Mechanical aptitude: Its existence, nature and measurement.* London: Methuen & Co. LTD.

D'Oliviera, T. C. (2004). Dynamic spatial ability: An exploratory analysis and a confirmatory study. *The international journal of aviation psychology*, *14*, 19–38. Verfügbar unter `http://dx.doi.org/10.1207/s15327108ijap1401_2`

Ekstrom, R. B., French, J. W., Harman, H. H. & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests.* Princeton: Educational Testing Service.

El Koussy, A. A. H. (1935). An investigation into the factors in tests involving the visual perception of space [Supplement]. *The British Journal of Psychology. Monograph Supplements*, *20*.

Eliot, J. & Macfarlane Smith, I. (1983). *An international directory of spatial tests.* Windsor: Nfer-Nelson.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, *43*, 95–103. doi:10.1037/0003-066X.43.2.95

Felix, M. C., Parker, J. D., Lee, C. & Gabriel, K. I. (2011). Real three-dimensional objects: Effects on mental rotation. *Perceptual and Motor Skills*, *113*, 38–50. doi:10.2466/03.22.PMS.113.4.38-50

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374. doi:10.1016/0001-6918(73)90003-6

Fischer, G. H. & Scheiblechner, H. H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch. *Psychologische Beiträge*, *12*, 23–51.

French, J. W., Ekstrom, R. B. & Price, L. A. (1963). *Manual for kit of reference tests for cognitive factors.* Princeton: Educational Testing Service.

Frey, A., Hartig, J. & Rupp, A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice.

Educational Measurement: Issues and Practice, *28*, 39–53. doi:10.1111/j.1745-3992.2009.00154.x

Geiser, C., Lehmann, W. & Eid, M. (2008). A note on sex differences in mental rotation in different age groups. *Intelligence*, *36*, 556–563. doi:10.1016/j.intell.2007.12.003

Gittler, G. (1990). *Dreidimensionaler Würfeltest (3DW). Ein rasch-skalierter Test zur Messung des räumlichen Vorstellungsvermögens.* Weinheim: Beltz.

Gittler, G. & Arendasy, M. (2003). Endlosschleifen: Psychometrische Grundlagen des Aufgabentyps E^P. *Diagnostica*, *49*, 164–175. doi:10.1026//0012-1924.49.4.164

Gittler, G. & Glück, J. (1998). Differential transfer of learning: Effects of instruction in descriptive geometry on spatial test performance. *Journal for Geometry and Graphics*, *2*, 71–84. Verfügbar unter `http://www.heldermann-verlag.de/jgg/jgg01_05/jgg0208.pdf`

Glanville, S. (2007). Anim8or (Version 0.95c) [Software]. Verfügbar unter `http://www.anim8or.com/download/animv095c.zip`

Glück, J. & Fabrizii, C. (2010). Gender differences in the mental rotations test are partly explained by response format. *Journal of Individual Differences*, *31*, 106–109. doi:10.1027/1614-0001/a000019

Goldstein, D., Haldane, D. & Mitchel, C. (1990). Sex differences in visual-spatial ability: The role of performance factors. *Memory & Cognition*, *18*, 546–550. doi:10.3758/BF03198487

Gormley, G. J., Dempster, M. & Best, R. (2011). Right-left discrimination among medical students: Questionnaire and psychometric study. *British Medical Journal*, *337*, 1–4. doi:10.1136/bmj.a2826

Górska, R. (2005). Spatial imagination - an overview of the longitudinal research at Cracow University of Technology. *Journal of Geometry and Graphics*, *9*, 201–208. Verfügbar unter `http://www.heldermann-verlag.de/jgg/jgg09/j9h2gors.pdf`

Górska, R., Sorby, S. A. & Leopold, C. (1998). Gender differences in visualization skills - An international perspective. *Engineering Design Graphics Journal*, *62*, 9–18. Verfügbar unter `http://www.edgj.org/index.php/EDGJ/article/view/115/111`

Guilford, J. P. & Lacey, J. I. (1947). *Printed classification tests* (Army Air Forces Aviation Psychology Program Research Reports No. 5). Verfügbar unter Defense Technical Information Center website: `http://handle.dtic.mil/100.2/AD651781`

Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3. Aufl.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Halpern, D. F. & Collaer, M. L. (2005). Sex differences in visuospatial abilities. In P. Shah & A. Miyake (Hrsg.), *The Cambridge handbook of visuospatial thinking* (S. 170–212). Cambridge: Cambridge University Press.

Hammer, R. E., Hoffer, N. & King, W. L. (1995). Relationships among gender, cognitive style, academic major, and performance on the piaget water-level task. *Perceptual and Motor Skills*, *80*, 771–778. doi:10.2466/pms.1995.80.3.771

Hegarty, M., Montello, D. R., Richardson, A. E., Ishikawa, T. & Lovelace, K. (2006). Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, *34*, 151–176. doi:10.1016/J.Intell.2005.09.005

Hegarty, M. & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, *32*, 175–191. doi:10.1016/j.intell.2003.12.001

Hegarty, M. & Waller, D. A. (2005). Individual differences in spatial ability. In P. Shah & A. Miyake (Hrsg.), *The Cambridge handbook of visuospatial thinking* (S. 121–169). Cambridge: Cambridge University Press.

Hirnstein, M., Ocklenburg, S., Schneider, D. & Hausmann, M. (2009). Sex differences in left-right confusion depend on hemispheric asymmetry. *Cortex*, *45*, 891–899. doi:10.1016/j.cortex.2008.11.009

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L. & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quarterly*, *50*, 391–402. Verfügbar unter `http://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/3-2008/06_Hohensinn.pdf`

Horan, P. F. & Rosser, R. A. (1984). A multivariable analysis of spatial abilities by sex. *Developmenal Review*, *4*, 387–411. doi:10.1016/0273-2297(84)90023-6

Hänsgen, K. D. & Spicher, B. (2011). *EMS Eignungstest für das Medizinstudium 2011* (Technical Report No. 18). Verfügbar unter Universität Freiburg/Schweiz, Zentrum für Testentwicklung und Diagnostik (ZTD) website: `http://www.unifr.ch/ztd/ems/berichte/Bericht18.pdf`

Johnson, E. S. & Meade, A. C. (1987). Developmental patterns of spatial ability: An early sex difference. *Child Development*, *58*, 725–740. doi:10.1111/j.1467-8624.1987.tb01413.x

Johnson, W. & Bouchard Jr., T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*, 393–416. doi:10.1016/j.intell.2004.12.002

Jonkisz, E., Moosbrugger, H. & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebogen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 27–74). Berlin: Springer.

Jordan, K., Wüstenberg, T., Jaspers-Feyer, F., Fellbrich, A. & Peters, M. (2006). Sex differences in left/right confusion. *Cortex*, *42*, 69–78. doi:10.1016/S0010-9452(08)70323-X

Kalichman, S. C. (1986). Horizontality as a function of sex and academic major. *Perceptual and Motor Skills*, *63*, 903–906. doi:10.2466/pms.1986.63.2.903

Kaufman, S. B. (2007). Sex differences in mental rotation and spatial visualization ability: Can they be accounted for by differences in working memory capacity? *Intelligence*, *35*, 211–223. doi:10.1016/j.intell.2006.07.009

Kelley, T. L. (1928). *Crossroads in the mind of man: A study of differentiable mental abilities*. Stanford: Stanford University Press.

Kerkman, D. D., Wise, J. C. & Harwood, E. A. (2000). Impossible "mental rotation" problems: A mismeasure of women's spatial abilities? *Learning and Individual Differences*, *12*, 253–269. doi:10.1016/S1041-6080(01)00039-5

Kozhevnikov, M. & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, *29*, 745–756. doi:10.3758/BF03200477

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model - Some critical suggestions on traditional approaches. *International Journal of Testing*, *5*, 377–394.

Kubinger, K. D. (2009). *Psychologische Diagnostik: Theorie und Praxis psychologsichen Diagnostizierens* (2. Aufl.). Göttingen: Hogrefe.

Kubinger, K. D., Rasch, D. & Yanagida, T. (2011). *Statistik in der Psychologie: Vom Einführungskurs bis zur Dissertation*. Göttingen: Hogrefe.

Larson, P., Rizzo, A. A., Buckwalter, J. G., van Rooyen, A., Kratz, K., Neumann, U. et al. (1999). Gender issues in the use of virtual environments. *CyberPsychology & Behavior*, *2*, 113–123. doi:10.1089/cpb.1999.2.113

Lassnigg, L. & Vogtenhuber, S. (2009). Geschlechterverteilung der Schüler/innen in öffentlichen und privaten Schulen nach Schultyp und Fachrichtung. In W. Specht (Hrsg.), *Nationaler Bildungsbericht Österreich 2009, Band 1: Das Schulsystem im Spiegel von Daten und Indikatoren* (S. 38–39). Graz: Leykam.

Leopold, C., Górska, R. & Sorby, S. A. (2001). International experiences in developing the spatial visualization abilities of engineering students. *Journal of Geometry and Graphics*, *5*, 81–91. Verfügbar unter `http://www.heldermann-verlag.de/jgg/jgg01_05/jgg0509.pdf`

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:10.1111/j.1467-8624.1985.tb00213.x

Lohman, D. F. (1979). *Spatial ability: A review and reanalysis of the correlational literature* (Technical Report No. 8). Verfügbar unter Stanford University, School of Education website: `http://dlib.stanford.edu:6520/text1/dd-ill/spatial-ability.pdf`

Lohman, D. F. (1988). Spatial abilities as traits, processes, and knowledge. In R. J. Sternberg (Hrsg.), *Advances in the psychology of human intelligence* (S. 181–248). Hillsdale: Lawrence Erlbaum Associates.

Macfarlane Smith, I. (1964). *Spatial ability: Its educational and social significane*. London: University of London Press LTD.

Mair, P., Hatzinger, R. & Maier, M. (2011). eRm: Extended Rasch modeling (Version 0.14-0) [Software-Handbuch]. Verfügbar unter `http://cran.r-project.org/package=erm`

Masters, M. S. & Sanders, B. (1993). Is the gender difference in mental rotation disappearing? *Behavior Genetics*, *23*, 337–341. doi:10.1007/BF01067434

McFarlane, M. (1925). A study of practical ability [Supplement]. *The British Journal of Psychology. Monograph Supplements*, *8*.

McGee, M. G. (1979). *Human spatial abilities: Sources of sex differences.* New York: Praeger.

McWilliams, W., Hamilton, C. J. & Muncer, S. J. (1997). On mental rotation in three dimensions. *Perceptual and Motor Skills*, *85*, 297–298.

Michael, W. B., Guilford, J. P., Fruchter, B. & Zimmerman, W. S. (1957). The description of spatial-visualization abilities. *Educational and Psychological Measurement*, *17*, 185–199. doi:10.1177/001316445701700202

Moffat, S. D., Hampson, E. & Hatzipantelis, M. (1998). Navigation in a "virtual" maze: Sex differences and correlation with psychometric measures of spatial ability in humans. *Evolution and Human Behaviour*, *19*, 73–87. doi:10.1016/S1090-5138(97)00104-9

Moosbrugger, H. (2012). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 227–274). Berlin: Springer.

Moosbrugger, H. & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Berlin: Springer.

Neubauer, A. C., Bergner, S. & Schatz, M. (2010). Two- vs. three-dimensional presentation of mental rotation tasks: Sex differences and effects of training on performance and brain activation. *Intelligence*, *38*, 529–539. doi:10.1016/j.intell.2010.06.001

Nordvik, H. & Amponsah, B. (1998). Gender differences in spatial abilities and spatial activity among university students in an egalitarian educational system. *Sex Roles*, *63*, 1009–1023. doi:10.1023/A:1018878610405

Ocklenburg, S., Hirnstein, M., Ohmann, H. A. & Hausmann, M. (2011). Mental rotation does not account for sex differences in left–right confusion. *Brain and Cognition*, *76*, 166–171. doi:10.1016/j.bandc.2011.01.010

Ofte, S. H. (2002). Right-left discrimination: Effects of handedness and educational background. *Scandinavian Journal of Psychology*, *43*, 213–219. doi:10.1111/1467-9450.00289

Ofte, S. H. & Hugdahl, K. (2002). Rightleft discrimination in male and female, young and old subjects. *Journal of Clinical and Experimental Neuropsychology*, *24*, 82–92. doi:10.1076/jcen.24.1.82.966

Olson, D. M. & Eliot, J. (1986). Relationships between experiences, processing style, and sex-related differences in performance on spatial tests. *Perceptual and Motor Skills*, *62*, 447–460. doi:10.2466/pms.1986.62.2.447

Parsons, T. D., Larson, P., Kratz, K., Thiebaux, M., Bluestein, B., Buckwalter, J. G. et al. (2004). Sex differences in mental rotation and spatial rotation in a virtual environment. *Neuropsychologia*, *42*, 555–562. doi:10.1016/j.neuropsychologia.2003.08.014

Peters, M. (2005). Sex differences and the factor of time in solving Vandenberg and Kuse mental rotation problems. *Brain and Cognition*, *57*, 176–184. doi:10.1016/j.bandc.2004.08.052

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R. & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test: Different versions and factors that affect performance. *Brain and Cognition*, *28*, 39–58. doi:10.1006/brcg.1995.1032

Peters, M., Lehmann, W., Takahira, S., Takeuchi, Y. & Jordan, K. (2006). Mental Rotation Test performance in four cross-cultural samples (n = 3367): Overall sex differences and the role of academic program in performance. *Cortex*, *42*, 1005–1014. doi:10.1016/S0010-9452(08)70206-5

Piaget, J. & Inhelder, B. (1967). *The child's conception of space.* New York: Norton.

Quaiser-Pohl, C. & Lehmann, W. (2002). Girls' spatial abilities: Charting the contributions of experiences and attitudes in different academic groups. *British Journal of Educational Psychology*, *72*, 245–260. doi:10.1348/000709902158874

R Development Core Team. (2011). R: A language and environment for statistical computing [Software-Handbuch]. Wien, Österreich. Verfügbar unter `http://www.r-project.org/`

Rasch, D., Pilz, J., Verdooren, R. & Gebhardt, A. (2011). OPDOE (Version 1.0-3) [Software-Handbuch]. Verfügbar unter `http://wwwu.uni-klu.ac.at/agebhard/OPDOE`

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.*

Kopenhagen: Danish Institute for Educational Research.

Rilea, S. L. (2008). A lateralization of function approach to sex differences in spatial ability: A reexamination. *Brain and Cognition*, *67*, 168–182. doi:10.1016/j.bandc.2008.01.001

Robert, M. & Chevrier, E. (2003). Does men's advantage in mental rotation persist when real three-dimensional objects are either felt or seen? *Memory & Cognition*, *31*, 1136–1145. doi:10.3758/BF03196134

Roberts, J. E. & Bell, M. A. (2003). Two- and three-dimensional mental rotation tasks lead to different parietal laterality for men and women. *International Journal of Psychophysiology*, *50*, 235–246. doi:10.1016/S0167-8760(03)00195-8

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2. Aufl.). Bern: Hans Huber.

Shah, P. & Miyake, A. (2005). *The Cambridge handbook of visuospatial thinking*. Cambridge: Cambridge University Press.

Souvignier, E. (2001). Training räumlicher Fähigkeiten. In K. J. Klauer (Hrsg.), *Handbuch Kognitives Training* (2. Aufl., S. 293–319). Göttingen: Hogrefe.

Stumpf, H. & Eliot, J. (1995). Gender-related differences in spatial ability and the k factor of general spatial ability in a population of academically talented students. *Personal and Individual Differences*, *19*, 33–45. doi:10.1016/0191-8869(95)00029-6

Stumpf, H. & Fay, E. (1983). Schlauchfiguren: Ein Test zur Beurteilung des räumlichen Vorstellungsvermögens [Manual]. Göttingen: Hogrefe.

Stumpf, H. & Klieme, E. (1989). Sex-related differences in spatial ability: More evidence for convergence. *Perceptual and Motor Skills*, *69*, 915–921. doi:10.2466/pms.1989.69.3.915

Teng, E. L. & Lee, A. L. (1982). Right-left discrimination: No sex differences among normals on the hand test and the route test. *Perceptual and Motor Skills*, *55*, 299–302. doi:10.2466/pms.1982.55.1.299

Thorndike, E. L. (1921). On the organization of intellect. *Psychological Review*, *28*, 141–151. doi:10.1037/h0070821

Thurstone, L. L. (1949). *Mechanical aptitude III: Analysis of group tests* (Psychometric

Laboratory No. 55). Verfügbar unter Defense Technical Information Center website: `http://handle.dtic.mil/100.2/AD490040`

Thurstone, L. L. (1950). Some primary abilities in visual thinking. *Proceedings of the American Philosophical Society*, *94*, 517–521. Verfügbar unter `http://www.jstor.org/stable/3143593`

Thurstone, L. L. (1969). *Primary mental abilities*. Chicago: University of Chicago Press. (Originalausgabe 1938)

Uecker, A. & Obrzut, J. E. (1993). Hemisphere and gender differences in mental rotation. *Brain and Cognition*, *22*, 42–50. doi:10.1006/brcg.1993.1023

Vandenberg, S. G. & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*, 599–604. doi:10.2466/pms.1978.47.2.599

Voyer, D. (1997). Scoring procedure, performance factors, and magnitude of sex differences in spatial performance. *The American Journal of Psychology*, *110*, 259–276. doi:10.2307/1423717

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: a meta-analysis. *Psychonomic Bulletin & Review*, *18*, 267–277. doi:10.3758/s13423-010-0042-0

Voyer, D., Voyer, S. & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250–270. doi:10.1037/0033-2909.117.2.250

Walter, O. & Rost, J. (2011). Psychometrische Grundlagen von Large Scale Assessments. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Methoden der psychologischen Diagnostik* (S. 87–149). Göttingen: Hogrefe.

Weiss, E. M., Kemmler, G., Deisenhammer, E. A., Fleischhacker, W. W. & Delazer, M. (2003). Sex differences in cognitive functions. *Personality and Individual Differences*, *35*, 863–875. doi:10.1016/S0191-8869(02)00288-X

Weitensfelder, L. (2011, September). Conception of a facet-oriented spatial ability test (TARV – Test of Applied Relations and Visuo-spatial abilities). In N. Hirschmann (Chair), *A cross section of new instruments in psychological assessment*. 11th European Conference on Psychological Assessment, Riga. Verfügbar unter `http://www.ecpa11.lu.lv/files/Lisbeth%20Weitensfelder.pdf`

Weitensfelder, L. (2012). Test zur Angewandten Raumvorstellung. In K. D. Kubinger, M. Frebort, M. Khorramdel & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“) (Hrsg.), *Self-Assessment: Theorie und Konzepte.* Lengerich: Pabst.

Weitensfelder, L., Grubesic, A., Kubinger, K. D. & Gittler, G. (2010). *Zur Notwendigkeit einer facettenorientierten Raumvorstellungsmessung in der Eignungsbeurteilung aus Gründen der Genderfairness.* Unveröffentlichtes Manuskript, Test und Beratungsstelle des Arbeitsbereichs Psychologische Diagnostik, Fakultät für Psychologie, Universität Wien, Österreich.

Witkin, H. A., Lewis, H. B., Hertzman, M., Machover, K., Bretnall Meissner, P. & Wapner, S. (1972). *Personality through perception: An experimental an clinical study.* Westport, Conn.: Greenwood. (Originalausgabe 1954)

Zentrum für Testentwicklung und Diagnostik (ZTD). (1995). Eignungstest für das Medizinstudium in der Schweiz [Manual]. Freiburg, Schweiz: Universität.

Zimmerman, W. S. (1953). A revised orthogonal rotational solution for Thurstone's original primary mental abilities test battery. *Psychometrika*, *18*, 77–93. doi:10.1007/BF02289029

Tabellenverzeichnis

Abbildungsverzeichnis

A. Termini verschiedener Raumvorstellungsfaktoren

Tabelle A
Termini von Raumvorstellungsfaktoren in Studien ab 1947

	Benennung	Definition
Guilford und Lacey (1947)	Spatial Relations (S_1, SR)	„ability to perceive visual-spatial arrangements (...) to organize movements in spatially-determined order (...) to relate specific spatial locus or arrangement within the stimulus pattern with specific locus or arrangement within the response pattern“ (S. 479)
	S_2	„right-hand-left-hand discrimination“ (S. 838), „”project”himself into the test objects or into a more favorable posture from which to judge the positions of the objects“ (S. 499)
	Visualization (Vz)	„visual-thinking activity in which objects must be moved or transformed in order to solve a problem“ (S. 857), „mentally to move, turn, twist, or rotate an object or objects and to recognize a new appearance or position after the prescribed manipulation is performed“ (S. 292)
	Length Estimation (LE)	„ability to estimate linear and nonlinear extent“ (S. 448), „comparison of lines or simple distances between points“ (S. 823)
Thurstone	S_1	„ability to recognize the identity of an object when it is seen from different angles (...) the subject imagines how the object looks from different directions“ (1950, S. 518), „ability to visualize a rigid configuration as it is moved from one position to another“ (1949, S. 13)
	S_2	„imagine the movement or internal displacement among the parts of a configuration that one is thinking about“ (1950, S. 518), „ability to visualize a configuration in which there is movement or displacement among the parts of the configuration“ (1949, S. 13)

weiter auf der nächsten Seite

Tabelle A
Termini von Raumvorstellungsfaktoren in Studien ab 1947 (Fortsetzung)

	Benennung	Definition
	S_3	„ability to think about those spatial relations in which the body orientation of the observer is an essential part of the problem“ (1950, S. 519)
	K	„represents kinesthetic imagery“ (1949, S. 16)
French (zitiert nach Michael et al., 1957)	Space	„ability to perceive spatial patterns accurately and to compare them with each other“ (S. 308)
	Spatial Orientation	„ability to remain unconfused by the varying orientations in which a spatial pattern may be presented (...) dimensionality is not so important to the factor as the rotational position of presentation“ (S. 188)
	Visualization	„ability to comprehend imaginary movements in a 3-dimensional space or the ability to manipulate objects in imagination“ (S. 308)
	Length Estimation	„ability to compare the length of lines or distances on a sheet of paper“ (S. 308)
Zimmerman (1953)	Spatial Relations	„When only a slight degree of turning or rotating is required for an individual to orient himself with an external object, he is more likely either to move himself or to feel himself adjust empathically, perhaps kinesthetically, to the stimulus situation.“ (S. 91)
	Visualization	„If a greater degree of adjustment is required (...) in order to bring himself and the object into alignment, he would have to form a mental image of the object and then manipulate it into position.“ (S. 91)

weiter auf der nächsten Seite

Tabelle A
Termini von Raumvorstellungsfaktoren in Studien ab 1947 (Fortsetzung)

	Benennung	Definition
Michael et al. (1957)	Spatial Relations and Orientation	„ability to comprehend the nature of the arrangement of elements within a visual stimulus pattern primarily with respect to the examinee's body as the frame of reference.“ (S. 189), „movement of the entire configuration, or a major component thereof, into a different position (little or no movement among the parts)“ (S. 193)
	Visualization	„mental manipulation of visual objects involving a specified sequence of movements (...) to rotate, turn, twist, or invert one or more objects, or parts, of a configuration“ (S. 191)
	Kinesthetic Imagery	„left-right discrimination with respect to the location of the human body“ (S. 191)
Ekstrom et al. (1976)	Spatial Orientation	„ability to perceive spatial patterns or to maintain orientation with respect to objects in space“ (S. 149), „the whole figure is manipulated in spatial orientation“ (S. 173)
	Visualization	„ability to manipulate or transform the image of spatial patterns into other arrangements (...) requires that the figure be mentally resturctured into components for manipulation“ (S. 173)
McGee (1979)	Spatial Orientation	„comprehension of the arrangement of elements within a visual stimulus pattern, the aptitude to remain unconfused by the changing orientations in which a spatial configuration may be presented, and the ability to determine spatial relations in which the body orientation of the observer is an essential part of the problem“ (S. 54)

weiter auf der nächsten Seite

Tabelle A
Termini von Raumvorstellungsfaktoren in Studien ab 1947 (Fortsetzung)

	Benennung	Definition
	Spatial Visualization	„ability to mentally rotate, manipulate, and twist two- and three-dimensional stimulus objects“ (S. 49)
Lohman (1979)	Spatial Relations	„Although mental rotation is the common element, the factor probably does not represent speed of mental rotation. Rather, it represents the ability to solve such problems quickly, by whatever means.“ (S. 127)
	Spatial Orientation	„(...) the ability to imagine how a stimulus array will appear from another perspective. (...) The subject most imagine he is reoriented in space, and then make some judgment about the situation. There is often a left-right discrimination component in these tasks, but this discrimination must be made from the imagined perspective.“ (S. 127)
	Visualization	„In addition to their spatial-figural content, the tests that load on this factor share two important features: (a) all are administered under relatively unspeeded conditions, and (b) most are much more complex than corresonding tests that load on the more peripheral factors.“ (S. 127)
	K	„speed of making left-right discriminations“ (S. 129)
Linn und Petersen (1985)	Mental Rotation	„ability to rotate a two or three dimensional figure rapidly and accurately“ (S. 1483)
	Spatial Perception	„to determine spatial relationships with respect to the orientation of their own bodies, in spite of distracting information“ (S. 1482)

weiter auf der nächsten Seite

Tabelle A
Termini von Raumvorstellungsfaktoren in Studien ab 1947 (Fortsetzung)

	Benennung	Definition
	Spatial Visualization	„(...) spatial ability tasks that involve complicated, multistep manipulations of spatially presented information. These tasks may involve the processes required for spatial perception and mental rotations but are distinguished by the possibility of multiple solution strategies.“ (S. 1484)
Lohman (1988)	Speeded Rotation	„(...) to determine whether a given stimulus is a rotated version of the target or is a rotated and reflected version of the target. Many subjects solve such problems by mentally rotating and reflecting the stimuli (...). The factor appears to represent the ability to solve simple rotation problems quickly, by whatever means.“ (S. 187)
	Spatial Orientation	„(...) require the examinee to determine how an object or scene will appear when viewed from a new perspective. Some tests require a left-right dscrimination from the imagined perspective.“ (S. 185)
	General Visualization	„One important characteristic of tests (...) is their complexity. Some require the rotation, reflection, or folding of complex figures, others require that figures be combined, some require multiple transformations, others require no transformations.“ (S. 185)
	Kinesthetic factor	„(...) factor represents the ability to make rapid left-right discriminations. (...) It may also be involved in discriminating a standard from a reflected version of a pattern.“ (S. 188f.)
Carroll (1993)	Spatial Relations	„Speed in manipulating relatively simple visual patterns, by whatever means (mental rotation, transformation, or otherwise).“ (S. 363)

weiter auf der nächsten Seite

Tabelle A
Termini von Raumvorstellungsfaktoren in Studien ab 1947 (Fortsetzung)

	Benennung	Definition
	Visualization	„Ability in manipulating visual patterns, as indicated by level of difficulty and complexity in visual stimulus material that can be handled successfully, without regard to the speed of task solution.“ (S. 362)
	Length Estimation	„Ability to make accurate estimates or comparisons of visual lengths or distances (without using measuring instruments).“ (S. 363)
Colom et al. (2001)	General Visualization	„generation, retention, retrieval and transformation of visual images“ (S. 904)
Hegarty und Waller (2004)	Spatial Orientation	„perspective-taking ability“ (S. 183), „ability to make egocentric spatial transformations (i.e., to imagine the results of changing one's egocentric frame of reference with respect to the environment)“ (S. 187f.)
	Spatial Visualization	„mental rotation ability“ (S. 183), „ability to make object-based transformations (i.e., to imagine the results of changing the positions of objects in the environment, while maintaining one's current orientation in the environment)“ (S. 188)
W. Johnson und Bouchard Jr. (2005)	Spatial factor	„comprehension of the arrangement of elements within a visual pattern and the ability to retain spatial orientation with respect to one's body, even in changing conditions“ (S. 413)
	Image Rotation factor	„ability to mentally rotate, manipulate, and twist two- and three-dimensional objects“ (S. 413)

B. Elternbrief

Stefan Haberstroh
[redacted]
e-mail: [redacted]
mobil: [redacted]

Liebe Eltern,

im Rahmen meiner Diplomarbeit an der Universität Wien (Institut für Entwicklungspsychologie und Psychologische Diagnostik) führe ich derzeit eine Studie an der Schule Ihrer Tochter / Ihres Sohnes durch. In meiner Diplomarbeit geht es um Raumvorstellung, also zum Beispiel um die Fähigkeit, Objekte in Gedanken rotieren zu können. Ziel der Arbeit ist es, einen neuen Raumvorstellungstest zu entwickeln. Dieser versucht, verschiedene Teilbereiche der Raumvorstellung zu erfassen und somit ein differenziertes Bild dieser Fähigkeit zu erhalten.

Dieser Ansatz bietet zwei Vorteile, die den Test auszeichnen sollen und weswegen ich um Ihre Unterstützung bitte. Zum einen soll überprüft werden, ob sich SchülerInnen (z. B. verschiedenen Geschlechts oder verschiedener Schultypen) je nach Teilbereich anders unterscheiden. Zum anderen soll das Erfassen verschiedener Teilbereiche der Raumvorstellung es auch ermöglichen, ganz konkrete Fördermaßnahmen anzusetzen, um damit die Leistungen der jeweiligen Person zu verbessern.

Die Entwicklung des Tests erfolgt des Weiteren im Rahmen der Wiener Self-Assessments für die Studiengänge Architektur und Maschinenbau (Technische Universität Wien). Studieninteressierten wird dadurch die Möglichkeit geboten, ihre Eignung vor Beginn eines Studiums unverbindlich, also nicht in Form eines Auswahlverfahrens, zu testen. Zusätzliche Informationen hierzu finden Sie auf:

- http://studienwahl.tuwien.ac.at
- http://derstandard.at/1289609229334/TU-Wien-bietet-Eignungstest-fuer-Studieninteressierte-an

bitte wenden

Ihre Tochter / Ihr Sohn würde gemeinsam mit anderen SchülerInnen den Raumvorstellungstest bearbeiten. Die Testdauer beläuft sich auf eine Schulstunde und würde während der Unterrichtszeit stattfinden. Sie / Er erhält zu Beginn des Tests einen anonymen Testcode, den nur Ihre Tochter / Ihr Sohn kennt. Daher ist es für andere Personen nicht möglich, die Testleistungen einzelnen SchülerInnen zuzuordnen. Mit diesem Testcode ist es nach Auswertung aller Tests möglich, dass eigene Ergebnis online abzurufen und ebenso mit mir für Rückfragen in Kontakt zu treten.
Die gesammelten Daten werden selbstverständlich streng vertraulich behandelt und nur für wissenschaftliche Zwecke genutzt.

Ich würde es sehr begrüßen, in die Erhebung auch Ihr Kind miteinbeziehen zu können. Daher wende ich mich mit der Bitte an Sie, Ihr Kind an der Studie teilnehmen zu lassen. Bei Fragen stehe ich Ihnen unter der oben genannten E-Mail Adresse sowie telefonisch gerne zur Verfügung.
Bitte geben Sie die ausgefüllte Einverständniserklärung Ihrer Tochter / Ihrem Sohn wieder mit.

Mit freundlichen Grüßen und der Bitte um Ihre Unterstützung,
Stefan Haberstroh

✂----------

Ich bin mit der Teilnahme meiner Tochter / meines Sohnes _____ ,
Name des Kindes

Klasse _____ , an der oben beschriebenen Testung und der Verwendung der Daten zu rein

wissenschaftlichen Zwecken einverstanden.

_____ _____
Ort, Datum Unterschrift des / der Erziehungsberechtigten

C. Q-Matrix: weitere Möglichkeiten

Tabelle C

Q-Matrix: weitere Möglichkeiten Q2 bis Q4

Item	Basisparameter, Q2		Basisparameter, Q3		Basisparameter, Q4			
	Relationen	Rotation & Orientierung	Relationen & Orientierung	Rotation	Relationen	Rotation	Orientierung	Anzahl markierter Objekte
MC1	1	0	1	0	1	0	0	3
MC2	0	1	0	1	0	1	0	4
MC3	0	1	1	0	0	0	1	3
MC4	1	1	1	1	1	1	0	3
MC5	1	1	1	0	1	0	1	4
MC6	0	1	1	1	0	1	1	3
MC7	1	1	1	1	1	1	1	2
MC8	1	0	1	0	1	0	0	3
MC9	0	1	0	1	0	1	0	4
MC10	0	1	1	0	0	0	1	3
MC11	1	1	1	1	1	1	0	3
MC12	1	1	1	0	1	0	1	3
MC13	0	1	1	1	0	1	1	3
MC14	1	1	1	1	1	1	1	3

D. MC1 bis SQ14: Grafische-Modell-Tests

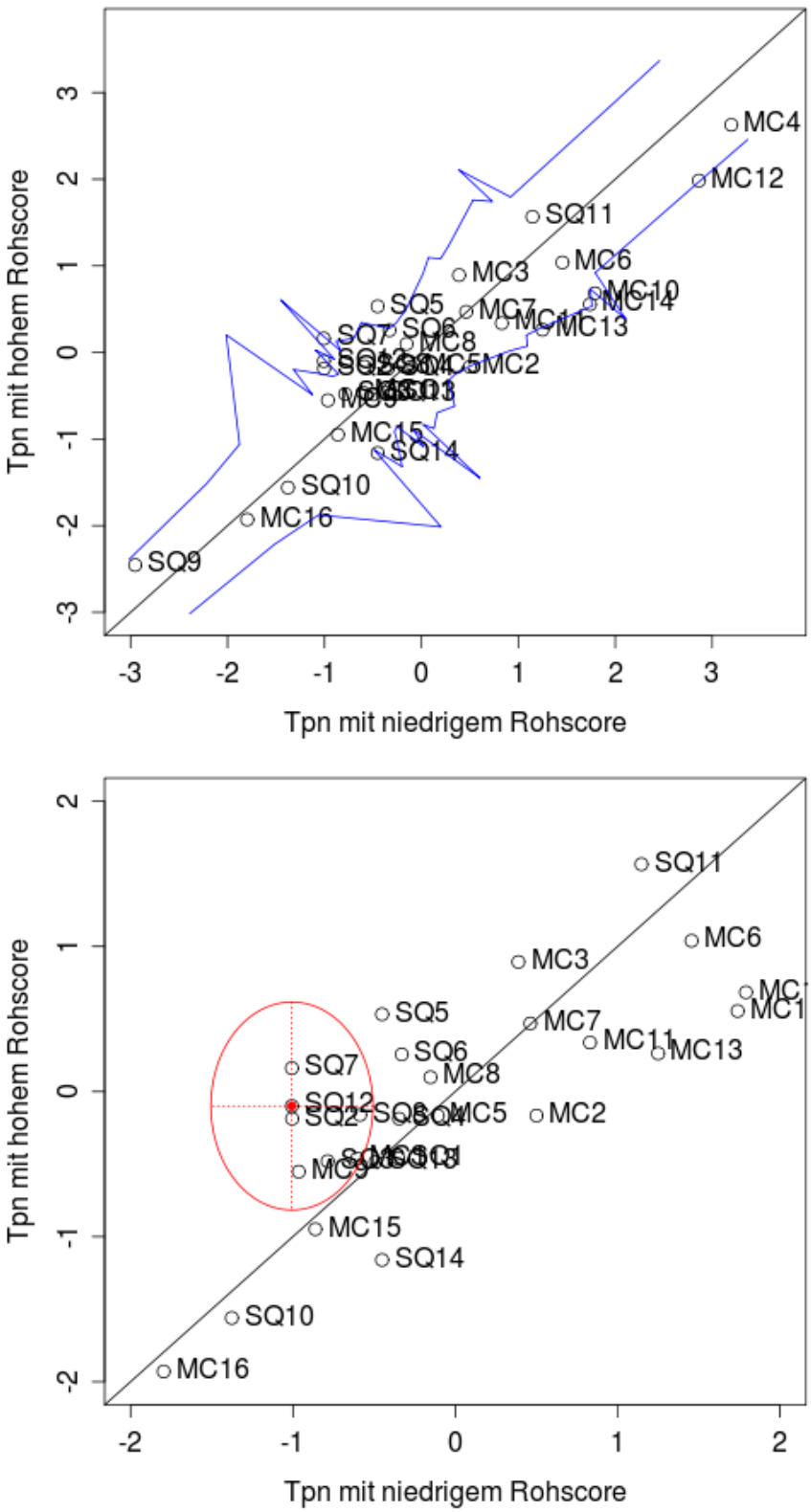


Abbildung 39. MC1 bis SQ14: Grafischer-Modell-Test, Teilungskriterium Rohscore.
Untere Abbildung mit der als Ellipse eingezeichneten Standardabweichung von Item SQ12.

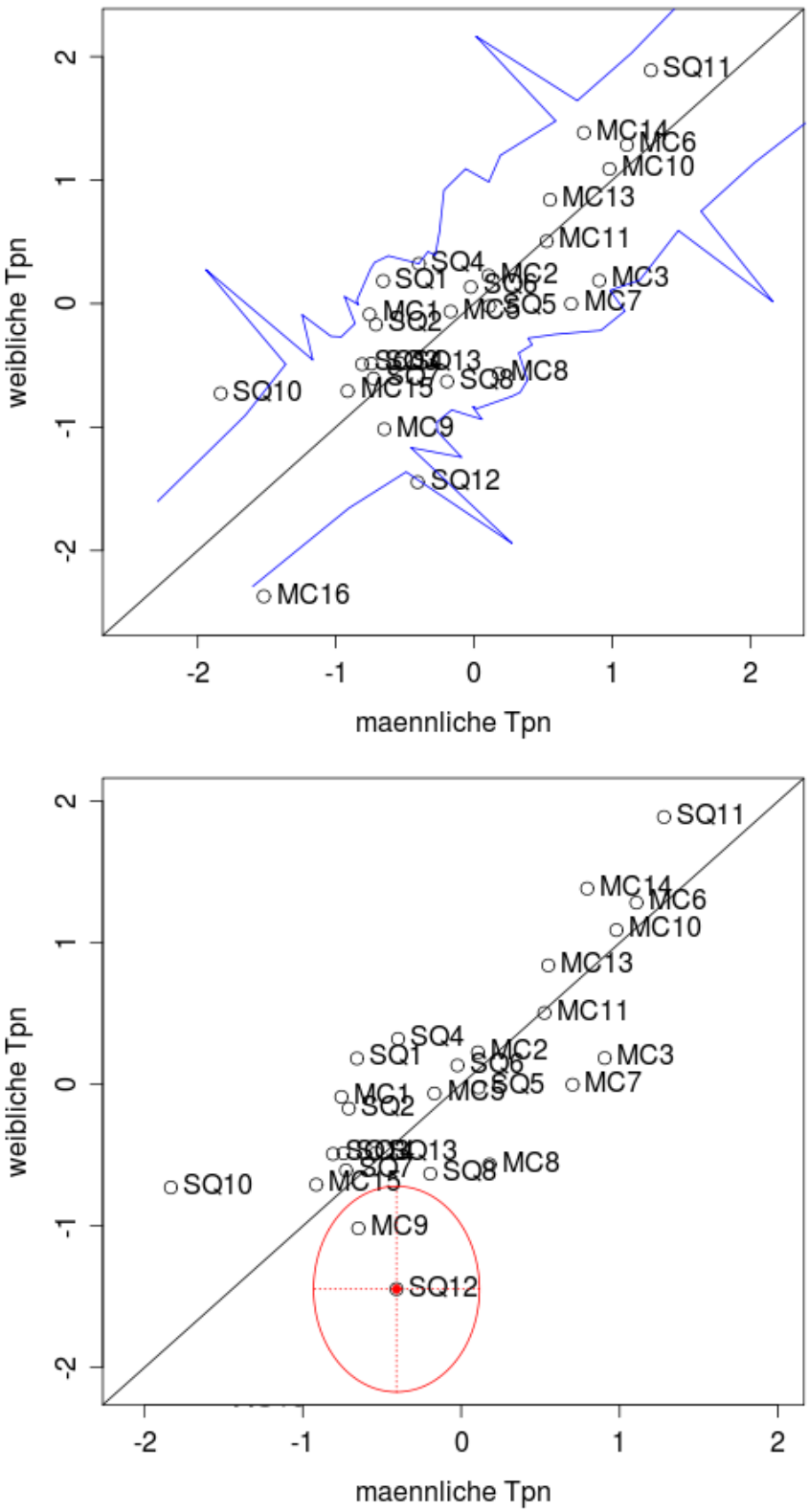


Abbildung 40. MC1 bis SQ14: Grafischer-Modell-Test, Teilungskriterium Geschlecht. Untere Abbildung mit der als Ellipse eingezeichneten Standardabweichung von Item SQ12.

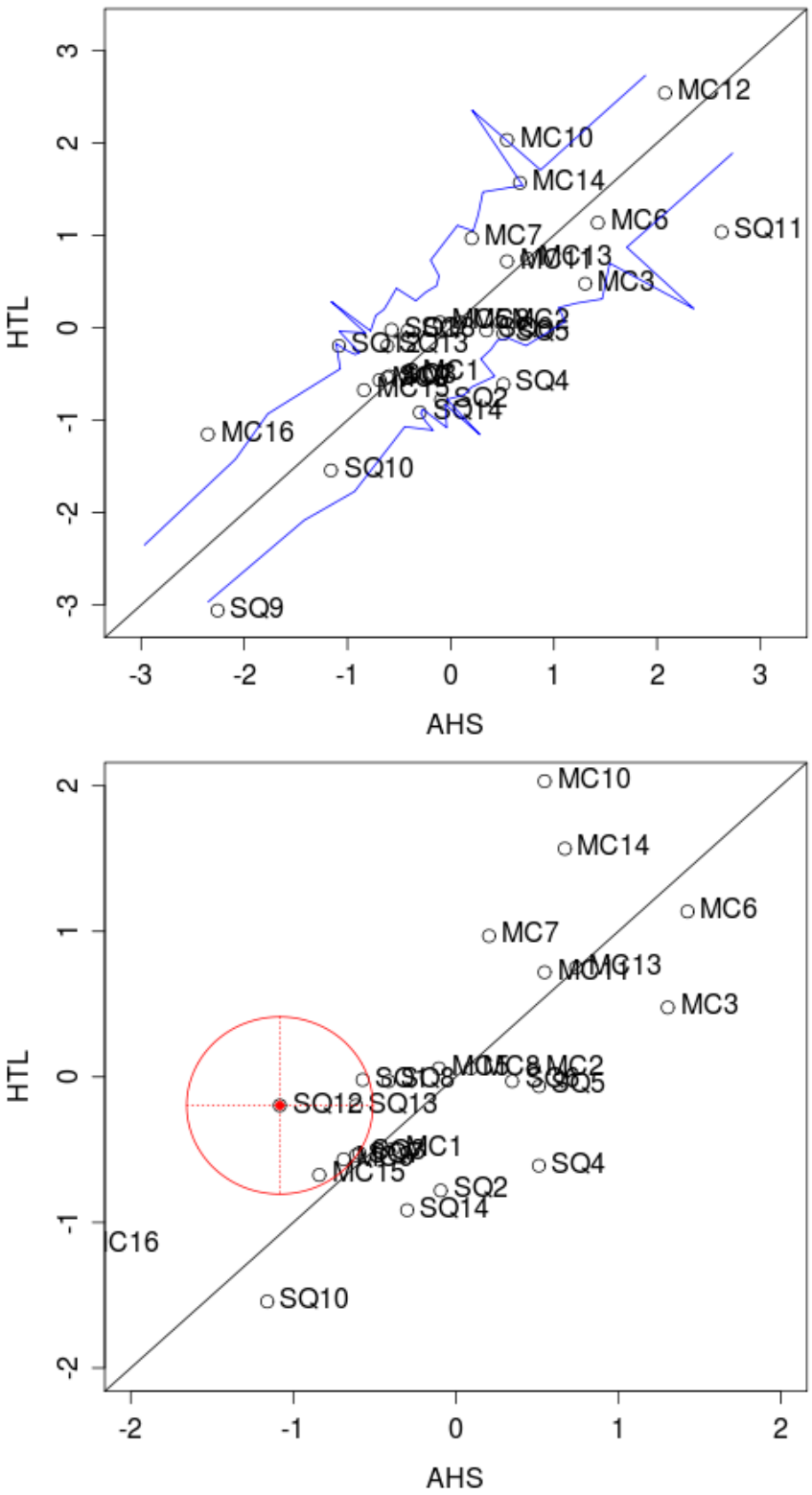


Abbildung 41. MC1 bis SQ14: Grafischer-Modell-Test, Teilungskriterium Ausbildung. Untere Abbildung mit der als Ellipse eingezeichneten Standardabweichung von Item SQ12.

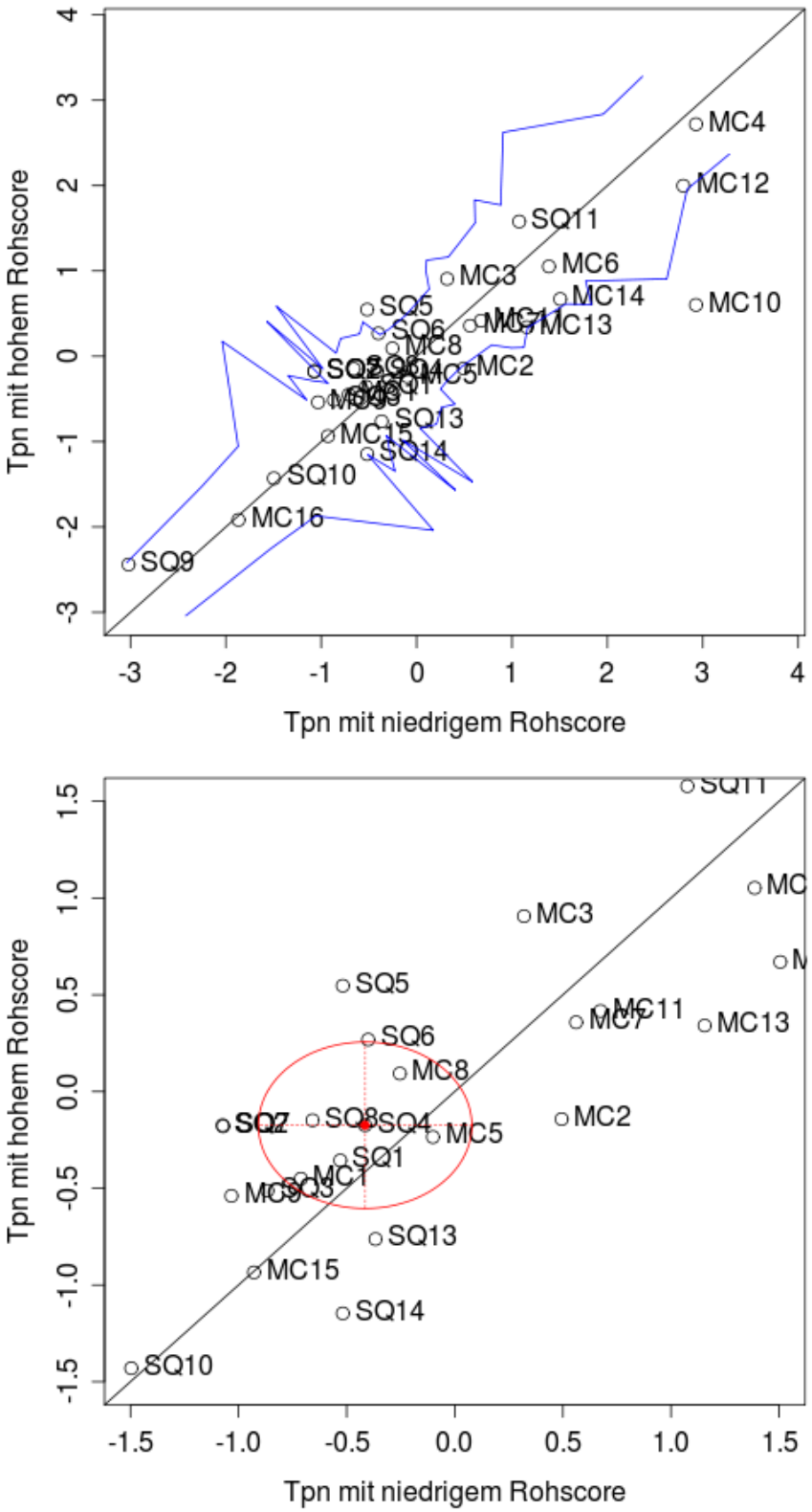


Abbildung 42. MC1 bis SQ14 ohne SQ12: Grafischer-Modell-Test, Teilungskriterium Rohscore. Untere Abbildung mit der als Ellipse eingezeichneten Standardabweichung von Item SQ4.

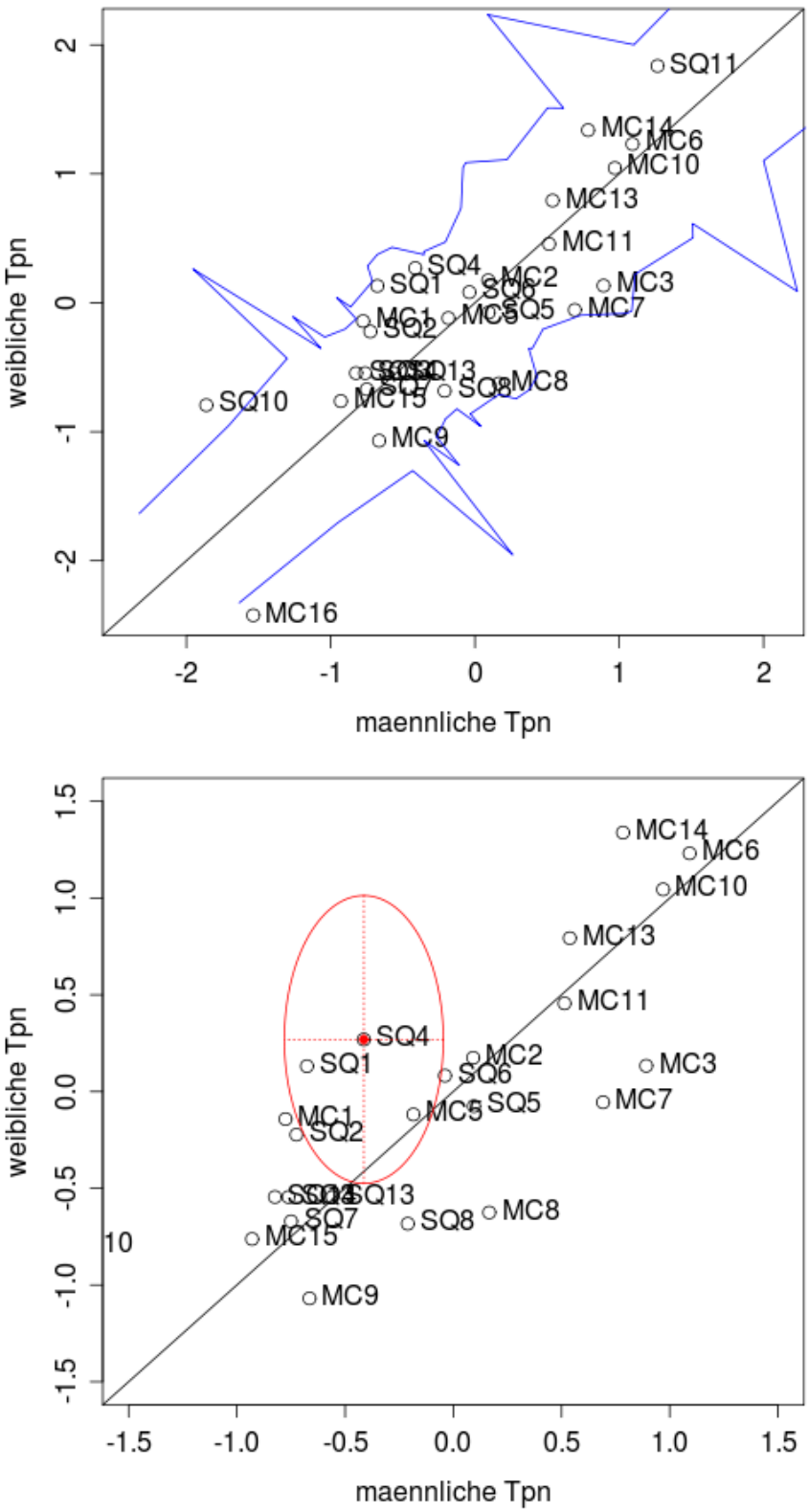


Abbildung 43. MC1 bis SQ14 ohne SQ12: Grafischer-Modell-Test, Teilungskriterium Geschlecht.
Untere Abbildung mit der als Ellipse eingezeichneten Standardabweichung von Item SQ4.

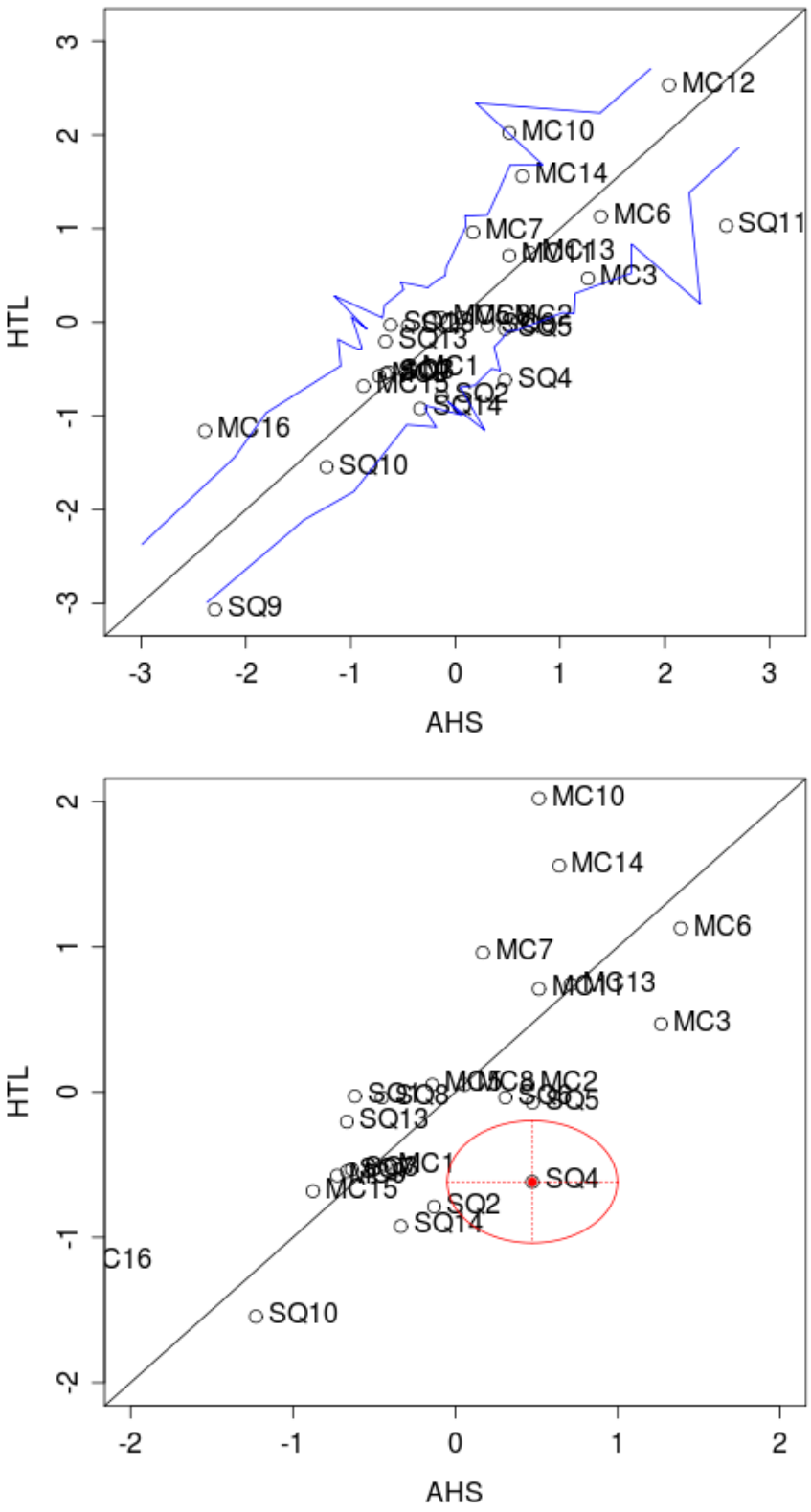


Abbildung 44. MC1 bis SQ14 ohne SQ12: Grafischer-Modell-Test, Teilungskriterium Ausbildung.
Untere Abbildung mit der als Ellipse eingezeichneten Standardabweichung von Item SQ4.

Curriculum Vitae

Persönliche Daten

01.03.1986	Geboren in Pfaffenhofen an der Ilm (D)

Ausbildung

2006-2012	Diplom-Studium Psychologie, Universität Wien
06.2005	Abitur, Hallertau Gymnasium, Wolnzach (D)

Praktika

05.07.-27.08.2010	Institut für gerichtliche Psychologie und Psychiatrie, Homburg (D)
27.07.-04.09.2009	Begutachtungsabteilung, Justizanstalt Wien Mittersteig