Diplomarbeit

Validierung des Endlosschleifentetsts (EST) unter power sowie work-limit Vorgabe

Verfasserin

Verena Schön

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, im September 2011

Studienkennzahl:	298
Studienrichtung:	Psychologie
Betreuer:	Ao. Univ.- Prof. Dr. Georg Gittler

Danksagung

An dieser Stelle möchte ich meinem Betreuer Ao. Univ. Prof. Dr. Georg Gittler danken, der mir mit seiner tatkräftigen Unterstützung, auch in schwierigen Phasen dabei geholfen hat, die Arbeit in die richtigen Bahnen zu lenken, sowie für seine hilfreichen Ratschläge. Ebenso möchte ich mich bei Frau Mag. Eva Adlmann bedanken, die mir ebenfalls während der Diplomarbeit Hilfestellungen geleistet hat.

Mein Dank gilt vor allem meiner Mutter und meinem Vater, die mich während meines Studiums immer bestärkt haben und mir während Höhen und Tiefen immer zur Seite standen! Vielen Dank!! Ohne Euch wäre vieles nicht möglich gewesen!

Des Weiteren möchte ich einem sehr wichtigen Mensch im meinem Leben Danke, meiner Tante Ingrid Wagner, die mich immer unterstützt hat.

Auch meinem Freund Mathias möchte ich dafür danken, dass er immer ein offenes Ohr für mich hatte und mir mit Ratschlägen zur Seite stand.

Ganz herzlich möchte ich mich auch bei allen Versuchspersonen bedanken, die sich für die Testung Zeit genommen haben und ohne die diese Arbeit nicht möglich gewesen wäre.

Inhaltsverzeichnis

1 Einleitung

Die vorliegende Arbeit hat zum Ziel, den Endlosschleifentest (EST; Gittler & Arendasy, 2003), welcher das räumliche Vorstellungsvermögen erfasst, auf Rasch-Homogenität zu überprüfen. Dies sollte einerseits unter power (ohne Beschränkung der Bearbeitungszeit), andererseits unter work-limit Bedingung (enthält eine Zeitdruckkomponente), stattfinden. Außerdem wird das Testinventar einer Validierungsprüfung unterzogen.

Raumvorstellung stellt einen wichtigen Faktor der allgemeinen Intelligenz dar, und beschreibt eine Fähigkeit, die unsere Wahrnehmung und Vorstellung der Umwelt nachhaltig beeinflusst. Sie ermöglicht uns erst die reibungslose Interaktion mit unserem Umfeld. Räumliches Vorstellungsvermögen hat in vielen Bereichen des täglichen Lebens große Relevanz, auch wenn das oft nicht bewusst wahrgenommen wird, etwa beim Einschätzen der Entfernung beim Überholen eines Autos, beim Lesen eines Stadtplans oder bei vielen handwerklichen sowie sportlichen Aktivitäten. Auch im schulischen Bereich, besonders in den Unterrichtsfächern Chemie und Physik, ist ein gut ausgeprägtes Raumvorstellungsvermögen von Vorteil, um z.B. chemische Formeln oder molekulare Strukturen besser verstehen zu können (Maier, 1999).

Der Endlosschleifentest (EST; Gittler & Arendasy, 2003) ist auf Grundlage der probabilistischen Testtheroie aufgebaut und gilt unter power Vorgabe als Rasch-homogen (Gittler & Arendasy, 2003). In vielen Bereichen ist aber nicht nur die Leistung gefragt, sondern auch die Geschwindigkeit, mit der diese Leistung erzielt wurde. Das gleichzeitige Erfassen der Leistungs- und Geschwindigkeitskomponente in einer Testung verletzt aber mit hoher Wahrscheinlichkeit die Vorraussetzungen der Eindimensionalität eines Tests (Gittler & Arendasy, 2003). Deshalb ist es eine zentrale Fragestellung dieser Arbeit, ob der Endlosschleifentest auch unter der work-limit Instruktion Eindimensionalität im Sinne von Rasch aufweist.

Darüber hinaus soll diese Diplomarbeit erheben, ob aufgrund der work-limit Bedingung, bei der die Probanden aufgefordert werden „so rasch wie möglich“ zu arbeiten, im Gegensatz zur power Vorgabe, mit der Instruktion „ausreichend Zeit für die

Testaufgaben zur Verfügung zu haben“, zusätzliche Informationen im Bezug auf die Leistung der Probanden ermittelt werden können.

Außerdem wird der Frage nachgegangen, ob es unter Vorgabe der beiden Bedingungen (power oder work-limit) Unterschiede in der Höhe der Korrelationen mit Außenkriterien gibt. Hierfür werden die durchschnittlichen Schulnoten der Probanden in den Unterrichtsfächern Mathematik, Chemie, Physik, Englisch, Deutsch und Latein, sowie die Selbsteinschätzung der Begabung in Mathematik, Musik, Technik und Raumvorstellung herangezogen. Zusätzlich wurde ein Planzeichentest konstruiert und als weiteres Außenkriterium verwendet. Weiters wurde ein Fragebogen zur Persönlichkeitseinschätzung “Selbstbild“ (Gittler, 2003-2009) vorgegeben, mit dem das psychologische Geschlecht der Probanden erhoben und in Zusammenhang mit den beiden Testbedingungen ausgewertet wurde.

Die vorliegende Diplomarbeit gliedert sich in zwei Teile. Im theoretischen Teil wird zunächst ein Überblick über Modelle und Theorien zum Faktor Raumvorstellung gegeben. Des Weiteren werden die beiden Parameter Leistungsgeschwindigkeit und Leistungsgüte sowie das Speed-Power-Problem näher beschrieben (Nährer, 1986). Anschließend wird das Androgyniekonzept näher vorgestellt (Altstötter-Gleich, 1996; Alfermann, 2001). Abschließend wird auf das Gütekriterium Validierung näher eingegangen.

Im empirischen Teil erfolgt die Beschreibung der Stichprobe, des Testmaterials sowie des Versuchsplans. Die Ergebnisse der hypothesentestenden Prüfungen werden beschrieben, interpretiert und zusammengefasst.

THEORIETEIL

2 Raumvorstellung

2.1 Begriffsbestimmung

Linn und Petersen (1985) definieren Raumvorstellung wie folgt, *"*Spatial ability generally refers to skill in representing, transforming, generating, and recalling symbolic, nonlinguistic information*"* (S.1482).

Raumvorstellung stellt einen wichtigen Faktor der menschlichen Intelligenz dar (Rost, 1976). Wie viele Faktoren notwendig sind, um räumliches Vorstellungsvermögen adäquat zu beschreiben, ist jedoch bis heute nicht einheitlich geklärt (Gittler & Arendasy, 2003).

Nach Gardner (2002, S. 57) ist für den Beruf des Piloten, Chirurgen, Architekten oder Graphiker eine gut ausgeprägte räumliche Begabung von großer Bedeutung.

3 Modelle und Theorien zum Faktor Raumvorstellung

Auf den folgenden Seiten wird ein Überblick über die wichtigsten Theorien gegeben, welche den Faktor Raumvorstellung beinhalten.

3.1 Die Zwei-Faktoren-Theorie von Spearman

Spearman (1904, zitiert nach Holling, Preckel & Vock, 2004, S. 18) beobachtete, dass die Leistungen von verschiedenen Versuchspersonen in verschiedenen kognitiven Tests positiv miteinander korrelierten. Dies brachte ihn zu dem Schluss, dass allen kognitiven Fähigkeiten ein gemeinsamer Faktor zugrunde liegt. Dieser wurde von Spearman (1904, zitiert nach Holling et al., 2004, S. 18) als general factor (g-Faktor) beschrieben. Des Weiteren nahm er spezifische Faktoren (s- Faktor) an, welche die Varianz beschreiben, die durch g nicht erklärt werden konnte. Um seine Zwei-Faktoren-Theorie zu bestätigen führte Spearman (1904, zitiert nach Funke & Vaterrodt-Plünnecke, 1998, S.43) anfangs Korrelationsrechnungen und später Berechnungen mittels Faktorenanalyse durch.

El-Koussy (1935, zitiert nach Maier, 1999, S. 31) testete 162 männliche Versuchspersonen im Alter von 11 bis 13 Jahren anhand einer Testbatterie mit 28 Tests, von denen 17 zur Messung von räumlichen Vorstellungsvermögen dienten. Mit Hilfe eines modifizierten faktorenanalytischen Verfahrens von Spearman (1904) gelang El-Koussy (1935, zitiert nach Maier, 1999, S. 31) der Nachweis, dass einige Raumvorstellungstests eine klare Korrelation zum Faktor g zeigten. Die hohen Korrelationen zwischen den Variablen waren damit jedoch nur unzureichend erklärbar. El-Koussy (1935, zitiert nach Maier, 1999, S. 31) nahm einen Faktor k an, den er als notwendig sieht …. „to obtain and the facility to utilize visual spatial imagery“ (S.44). Im Jahre 1955 veränderte El-Koussy seine Sichtweise und unterschied zwischen einem 2- und einem 3-dimensionalen räumlichen Vorstellungsvermögen (zitiert nach Maier, 1999).

3.2 Thurstones Primärfaktoren

Louis Thurstone (1938, zitiert nach Holling et al., 2004, S. 19) vertrat die Ansicht, dass sich Denkleistung nicht ausreichend durch Spearmans Zwei-Faktoren-Theorie erklären lässt. Thurstone (1938, zitiert nach Amelang, Bartussek, Stemmler & Hagemann, 2011, S. 148) geht daher in seinem Modell davon aus, dass beim Lösen von kognitiven Aufgaben immer mehrere Gruppenfaktoren in wechselndem Gewichtsverhältnis beteiligt sind. Mit Hilfe der multiplen Faktorenanalyse ermittelte Thurstone 1938 (zitiert nach Amelang et al., 2011, S. 148 ff) zunächst neun, später sieben Primärfaktoren (primary mental abilities) der Intelligenz:

- verbal comprehension (Faktor v)
- word fluency (Faktor w)
- number (Faktor n)
- space (Faktor s)
- memory (Faktor m)
- perceptual speed (Faktor p)
- reasoning (Faktor i)

Der Faktor s (space) kann als räumliches Vorstellungsvermögen interpretiert werden und bezeichnet die Fähigkeit mit zwei- oder dreidimensionalen Objekten in der

Vorstellung zu operieren. Der Faktor space wird von Thurstone zuerst in zwei und 1950 in drei Faktoren unterteilt (Maier, 1999, S. 19ff.):

Faktor s1: spatial relations (Räumliche Beziehungen)

Nach Thurstone (1950, zitiert nach Maier, 1999, S. 38) kann spatial relation als die Fähigkeit, ein Objekt aus verschiedenen Blickwinkeln zu erkennen, beschrieben werden.

Nach Pawlik (1968, S. 336) handelt es sich um die Fähigkeit, räumliche Beziehungen zwischen unbewegten Gegenständen richtig zu erfassen. Es müssen verschiedene Ansichten eines Gegenstandes ohne anschauliche Hilfe vorgestellt werden (Pawlik, 1968, 336).

Faktor s2: visualization (Veranschaulichung)

Der Faktor s2 beschreibt die Fähigkeit zur gedanklichen Vorstellung von Drehung oder räumlicher Verschiebung ganzer Objekte oder einzelner Teile ohne anschauliche Hilfe (Pawlik, 1968, S. 336).

Für die richtige Bearbeitung von Aufgaben dieser Art sind komplizierte Denkvorgänge notwendig, die im Gegensatz zum Faktor spatial relations überwiegend dynamisch sind (Maier, 1994, S. 35).

Faktor s3: spatial orientation (Räumliche Orientierung)

Räumliche Orientierung bezeichnet die Fähigkeit „über räumliche Beziehungen zu denken, wenn dabei die Körperorientierung ein wesentlicher Teil des Problems ist“ (Thurstone, 1950, zitiert nach Rost, 1976, S. 124)

3.3 Die Einteilung der Raumvorstellung nach Michael, Guilford, Fruchter und Zimmerman

Michael, Guilford, Fruchter und Zimmerman (1957) gehen von drei Faktoren der Raumvorstellung aus:

Faktor Vz: visualization (Vz):
Nach Michael et al. (1957, S. 187) ist der Faktor visualization (Vz) gleichzusetzen mit dem Faktor s2 von Thurstone (1950).

Faktor (SR-O):spatial relations and orientations
Ausgehend von den drei Faktoren nach Thurstone (1950) entspricht der SR-O Faktor einer Kombination der Faktoren s1 und s2 (Michael et al., 1957, S. 187).

Faktor K: kinestetic imagery
Der Faktor k wird nach Michael et al. (1957) beschrieben als „This highly tentative factor represents merely a ***left-right discrimination*** with respect to the location of the human body…"(S.191).

3.4 Die Re-Analyse von Lohman

Lohman (1979, zitiert nach Eliot, 1987, S. 63 ff.) führte eine Re-Analyse der Daten aus 35 Studien durch und fand drei Hauptfaktoren und einige Nebenfaktoren des räumlichen Vorstellungsvermögens.

Zu den Hauptfaktoren zählen nach Lohman (1979, zitiert nach Eliot, 1987, S. 63ff):

- *spatial relations* wird durch Tests wie Thurstone`s Cards, Flags, oder Figuren erfasst. Es beschreibt die Fähigkeit zur raschen mentalen Rotation bzw. Spiegelung von Figuren und Objekten.
- *spatial orientation* kennzeichnet die Fähigkeit, ein Objekt aus verschiedenen Perspektiven zu betrachten. Es handelt sich dabei um einen schwer bestimmbaren Faktor t, da die Bearbeitung einer Aufgabe zur Messung dieser Fähigkeit auch durch Anwendung mentaler Rotation gelöst werden kann.

- *spatial visualization* ist die Fähigkeit, komplexe räumliche Aufgabenstellungen ohne Vorgabe von Zeitlimits zu bearbeiten.

Zu den Nebenfaktoren zählt Lohman (1979, zitiert nach Eliot, 1987, S. 64) closered speed, perceptual speed, visual memory und kinestetic.

3.5 Die Kategorien der Raumvorstellung nach Linn und Petersen

Linn und Petersen (1985, S. 1479ff.) beschreiben in ihrer viel zitierten Metaanalyse drei Kategorien der Raumvorstellung:

Spatial Perception (Räumliche Wahrnehmung):
Räumliche Wahrnehmung bezeichnet die Fähigkeit, räumliche Beziehungen zwischen Objekten zu erfassen und diese, trotz ablenkender Informationen, im räumlichen Verhältnis zum eigenen Körper zu erkennen.
Beispiele für Tests zur Messung der räumlichen Wahrnehmung sind der *„Rod and Frame Test“* (RFT) (Witkin, Dyk, & Faterson, 1962) oder der *„water level test“*(vgl Inheleder & Piaget, 1958) (Linn & Petersen, 1985).

Mental Rotation (Vorstellungsfähigkeit von Rotationen):
Mentale Rotation umfasst die Fähigkeit zwei- und dreidimensionale Objekte richtig und rasch in der Vorstellung zu rotieren.
Die Messung dieser Fähigkeit erfolgt zum Beispiel durch den Mental Rotation Test von Shepard und Metzler (1971), welcher von Vandeberg und Kuse (1978) und Peters (2005) modifiziert wurde (Linn & Petersen, 1985).

Spatial Visualization (Veranschaulichung oder Räumliche Visualisierung):
“Spatial visualization is the label commonly associated with those spatial ability tasks that involve complicated, multistep manipulations of spatially presented informations. The tasks may involve the processes required for spatial perception and mental rotations but are distinguished by the possibility of multiple solution strategies.” (Linn & Petersen, 1985, S.1484).

4 Leistungsgeschwindigkeit und Leistungsgüte

In der Psychologie stellt die Messbarkeit einer Person ein zentrales Thema dar. Besonders von Interesse ist die Messung der Leistung mit Hilfe von Leistungstests. Um über die Leistung einer Person etwas aussagen zu können, werden die beiden Parameter Leistungsgeschwindigkeit und Leistungsgüte verwendet (Nährer, 1986).

4.1 Definition von Leistungsgeschwindigkeit

In der Psychologie versteht man unter Geschwindigkeit die Zeit, die eine Person aufwendet, im Verhältnis zu einer anderen Person gesehen, um eine bestimmte Verhaltensweise auszuführen (Nährer, 1986, S. 1).

Nach Nährer (1986) bezeichnet Leistung „ die Menge und/oder Art von bewältigten Aufgaben“ (S.1).

4.2 Definition von Leistungsgüte

…„seit dem Durchbruch der probabilistischen Theorie, definiert man die Leistungsgüte entweder unmittelbar aufgrund der Lösungswahrscheinlichkeit in unbegrenzter Zeit oder aber als latente Eigenschaft (“latent trait“), die zusammen mit Aufgabenparametern die Lösungswahrscheinlichkeit nach einer von Modell zu Modell verschiedenen Funktion bestimmt“ (Iseler, 1970, S. 73)

4.3 Messen von Leistungsgeschwindigkeit und Leistungsgüte

Die beiden Parameter Leistungsgeschwindigkeit (speed) und Leistungsgüte (power) können entweder mittels einer einzigen Testung oder mit Hilfe von mindestens zwei separaten Tests erfasst werden (Nährer, 1986, S. 3).

4.3.1 Erfassung von Leistungsgeschwindigkeit und Leistungsgüte durch zwei getrennte Testungen

Um die beiden Parameter Leistungsgeschwindigkeit und Leistungsgüte durch zwei separierte Testungen zu erhalten werden so genannte “reine“ Geschwindigkeitstests (Speed-Test) oder “reine“ Fähigkeitstests (Power-Test) eingesetzt (Nährer, 1986, S. 4).

4.3.1.1 Geschwindigkeitstest (Speed-Test)

Nach Anastasi (1997) kann ein Speed-Test, wie folgt, definiert werden:

> A pure *speed test* is one in which individual differences depend entirely on speed of performance. Such a test is constructed from items of uniformly low difficulty, all of which are well within the ability level of the persons for whom the test is designed. The time limit is made so short that no one can finish all the items. Under these conditions, each person’s score reflects only the speed with which he or she worked (S.102).

Nach Nährer (1986, S. 8) weisen reine Geschwindigkeitstests eine Lösungswahrscheinlichkeit von P(+) → 1 auf. Der Konzentrations-Verlaufs-Test, KVT (ABELS, 1961), und der Aufmerksamkeits-Belastungs-Test d2 (BRICKENKAMP, 1966) sind Beispiele für Speed-Tests (Nährer, 1986, S.4).

4.3.1.2 Fähigkeitstest (Power-Test)

Reine Fähigkeits- oder Power-Tests werden als Maß für die Fähigkeit einer Person herangezogen). Die Items sind unterschiedlich schwer und haben eine Lösungswahrscheinlichkeit von p(+)< 1(Nährer, 1986, S. 4).

> A pure *power test,* [....] has a time limit long enough to permit everyone to attempt all items. The difficulty of the items is steeply graded, and the test includes some items too difficult for anyone to solve, so that no one can get a perfect score (Anastasi, 1997, S.102).

Der Dreidimensionale Würfeltest (3DW; Gittler, 1990) und der Wiener-Matrizen-Test (WMT; Formann & Piswanger, 1979) können als Beispiele für einen Power-Test angesehen werden.

4.3.2 Erfassung von Leistungsgeschwindigkeit und Leistungsgüte aus einer einzigen Testung

Um die beiden Komponenten speed und power aus einer einzigen Testung zu erhalten werden sogenannte Speeded-Tests eingesetzt.

4.3.2.1 Speeded-Test

Bei Speeded-Tests werden Items mit unterschiedlicher Schwierigkeit mit einer Lösungswahrscheinlichkeit von P(+) <1 vorgegeben, um den Fähigkeits- und den Geschwindigkeitsparameter aus einer einzigen Testung zu erhalten. Die Aufgaben werden unter Zeitdruck und/oder impliziten oder expliziten Zeitgrenzen vorgegeben (Nährer, 1986, S. 10).

Es werden nach Nährer (1986) folgende Personenkennwerte herangezogen:

- Die Zahl richtiger Antworten bis zur Erreichung einer Zeitgrenze (time-limit-Methode).
- Die für den Gesamttest benötigte Zeit, (meist) unter der Instruktion zur größten Sorgfalt in möglichst kurzer Zeit (work-limit Methode).
- Die Zahl richtiger Antworten für den Gesamttest ohne Zeitbegrenzung und der Instruktion, möglichst nur richtige Antworten zu liefern (S.10).

4.4 Problem bei der Erfassung von Leistungsgeschwindigkeit und Leistungsgüte

Im folgenden Abschnitt wird näher auf die Probleme eingegangen, die sich ergeben, wenn die beiden Parameter Leistungsgeschwindigkeit und Leistungsfähigkeit mittels einer einzigen Testung beziehungsweise mit Hilfe zweier getrennter Tests erhoben werden.

4.4.1 Kritik an der getrennten Erfassung von Leistungsgeschwindigkeit und Leistungsgüte

Nach Nährer (1986, S. 8) ist es schwierig die Leistungsgeschwindigkeit und Leistungsgüte getrennt zu erfassen.

Bei einem reinen Geschwindigkeitstest wird davon ausgegangen, dass das Testresultat nur von der Geschwindigkeit des Probanden abhängig ist. Jedoch ist anzunehmen, dass auch die Bearbeitung von sehr leichten Items, wie es bei Geschwindigkeitstests üblich ist, ein Mindestmaß an Fähigkeit von den Testpersonen erfordert. (Nährer,1986, S.8). Daher kann es nach Fischer (1970, S. 390) einen reinen Speed -Test gar nicht geben, da die Komponenten speed und power immer gleichzeitig wirksam sind.

Ein Power-Test kann lediglich die Leistung einer Versuchsperson erheben. Möchte man zusätzlich die Geschwindigkeit der erbrachten Leistung erfassen, stellt sich nach Nährer (1986) die Frage, *"*wie müssen die (sehr leichten) Items des zugehörigen Speed-Tests beschaffen sein, damit das Fähigkeits- und das Geschwindigkeitsmaß in Hinsicht auf den angestrebten Validitätsbereich zueinander in Beziehung gesetzt werden können?*"* (S.9). Es ergibt sich das Problem, dass sich die Items des Speed-Tests nicht unterscheiden (Lösungswahrscheinlichkeit = 1) und daher keine zufriedenstellende Dimensionsanalyse (außer nach dem Augenschein) durchgeführt werden kann (Nährer, 1986, S. 9).

4.4.2 Kritik an der Erfassung von Leistungsgeschwindigkeit und Leistungsgüte aus einer einzigen Testung

Bei Speeded-Tests ist die Anzahl der richtigen Antworten relevant, wobei richtige Antworten, die in kurzer Zeit gegeben werden, besser bewertet werden, als jene für die mehr Zeit beansprucht wird. Personen, welche die Aufgabe schnell beantworten, sind demnach die Fähigeren. Es wird also angenommen, dass Fähigkeit und Geschwindigkeit auf einer Dimension liegen, und ein Proband durch einen einzigen Parameter charakterisiert werden kann (Nährer, 1986, S. 11). Es gibt jedoch immer wieder Versuchspersonen, die einen Leistungstest in kurzer Zeit bearbeiten, aber mit wenig korrekt gelösten Items. Andere Testpersonen benötigen zur Bearbeitung viel Zeit, liefern dafür aber mehr richtige Items (Nährer, 1988, S. 211). Des Weiteren ist zu

bedenken, dass obwohl die Bearbeitungszeit für jede Aufgabe mit wachsender Schwierigkeit zunimmt (Nährer, 1986, S.221ff.) es vorkommen kann, dass eine Testperson x, leichte Items rascher bearbeitet als Versuchsperson y, dass aber bei der Bearbeitung von schwierigen Aufgaben die Person x langsamer ist (Nährer, S. 221-222). Aufgrund dieser Betrachtungen ist nach Nährer (1988, S. 211) davon auszugehen, dass die Anzahl richtig gelöster Antworten in Speeded-Tests, sowohl von der Fähigkeit des Probanden als auch von der Bearbeitungszeit abhängen. Aus der Anzahl richtig beantworteter Items ist kein Personenparameter bestimmbar, der eine allgemein gültige Auskunft über die Lösungswahrscheinlichkeit bei Aufgaben gibt (Nährer, 1988).

Nach Nährer (1988, S. 222) stellt es ein Problem dar, dass Versuchspersonen subjektive Hypothesen über die Verrechnung der Antworten im Leistungstest bilden, und diese das Testergebnis beeinflussen. Ein Proband, der denkt, dass nur die Anzahl der richtig gelösten Aufgaben von Bedeutung ist, wird sich anders in der Testsituation verhalten als eine Versuchsperson, die der Meinung ist, dass auch teilrichtige Antworten zählen (Nährer, 1988, S.222).

Weitere schwer kontrollierbare Faktoren sind nach Nährer (1988)…“individuelle Unterschiede in der Fähigkeit, Zeiten zu schätzen, sowie individuelle Differenzen im Anspruchsniveau bezüglich Genauigkeit bzw. Richtigkeit einer Lösung“ (S.222).

Außerdem hängt nach Nährer (1988) die Anzahl richtig beantworteter Aufgaben in Speeded–Tests auch von der Verteilung der Schwierigkeitsgrade der Items im Test und von der Höhe des Zeitdrucks ab.

Beim Speeded–Test wird eine explizite Zeitgrenze für die Bearbeitung der Items vorgegeben oder die Antworten werden im Nachhinein mit der tatsächlich benötigten Zeit gewichtet. Es gibt somit nur eine einzige Zeitdruckstärke, deren Validität jedoch fraglich ist, weil ihre Wahl keine messtheoretische Begründung aufweist. Des Weiteren ist zu bedenken, dass keine Aussage darüber gemacht werden kann, welche Ergebnisse der Proband bei einer anderen Zeitgrenze liefern würde oder über eine mögliche Kompensation geringer Fähigkeiten durch zusätzliche Zeit (Nährer, 1986, S. 49).

4.4.2.1 Die work-limit Instruktion

Die work-limt Instruktion stellt eine Möglichkeit dar, mit der die Nachteile die sich durch speeded–Tests ergeben, wegfallen (Nährer, 1988). Die Aufgaben werden ohne Zeitbegrenzung vorgegeben mit der Instruktion die Aufgaben „möglichst schnell und möglichst genau zu bearbeiten.“ Durch die Anweisung erhält man Auskunft über das Fähigkeitsniveau und die Bearbeitungsgeschwindigkeit des Probanden, wobei die Anzahl richtiger Antworten Informationen über das Fähigkeitsniveau (power-Komponente) liefert und die Bearbeitungszeit informiert über die Bearbeitungsgeschwindigkeit (speed-Komponente) (vgl Näher, 1988; Hummel, 2001).

Nährer (1986) konnte nachweisen, dass bei den Aufgaben des Figuren-Mengen-Tests die Rasch-Homogenität des Itemmaterials auch unter der work-limit Instruktion gegeben ist. Baar (1999) konnte zeigen, dass dieses auch für die dreidimensionalen Würfelaufgaben von Gittler (1999) anzunehmen ist (Hummel, 2001).

4.5 Speed-Power-Problem

Nach Nährer (1986) besteht das Speed-Power-Problem in der “fehlenden Separierbarkeit des Fähigkeitsaspektes von der Geschwindigkeitskomponente an einer beobachtbaren Reaktion“ (S.3). In der Testpraxis stellt sich die Frage, um wie viel eine richtige Antwort, die schnell gefunden wurde, besser bewertet werden soll, als eine korrekte Antwort, die später gegeben wurde (Fischer, 1970, S. 389).

Fischer (1970) formuliert das Speed-Power-Problem folgendermaßen: “Sagt die Geschwindigkeit (speed) einer Leistung auch etwas über die Leistungsgüte (power) der Vp aus, oder ist Schnelligkeit ein qualitativ anderer Aspekt des Leistungsverhaltens*?*“(S.389).

Fischer (1970, S. 390) formulierte zwei Hypothesen:

Die erste Hypothese geht davon aus, dass die beiden Parameter Leistungs-geschwindigkeit und Leistungsgüte Ausdruck einer gemeinsamen, zugrunde liegenden Leistungsdimension sind. Für die Messung der Leistungsfähigkeit ist ausschlaggebend,

wie viele Items die Versuchsperson korrekt gelöst hat, sowie die Länge der Bearbeitungszeit.

Die zweite Hypothese nimmt an, dass Leistungsgeschwindigkeit und Leistungsgüte zwei qualitativ verschiedene Dimensionen sind. Verschiedene Kombinationen der beiden Begabungsrichtungen sind möglich. Das drückt sich in den verschiedenen Reaktionsmustern der Versuchspersonen aus.

Die erste Hypothese wurde von Fischer (1970 S. 394) mit Hilfe des mehrkategoriellen Modell von Rasch (1960) überprüft. Als Stichprobe wählte Fischer (1970, S. 395) 200 männliche Testpersonen im Alter zwischen 18 und 20 Jahren. Er gab ihnen 15 ausgewählte Aufgaben aus dem Matrizentest von Raven, 16 Items aus Subtest 4 des Leistungsprüfsystems (LPS) von Horn (1962) und 16 Aufgaben aus dem Intelligenz-Struktur-Test (IST) von Amthauer (1970) vor.

Die erste Hypothese musste verworfen werden, Leistungsgeschwindigkeit und Leistungsgüte lassen sich nicht auf eine einzige Dimension reduzieren (Fischer, 1970, S. 397). Anschließend stellte sich Fischer (1970) die Frage, wie Leistungs-geschwindigkeit und Leistungsgüte als Determinanten des individuellen Reaktionsverhaltens einer Versuchsperson voneinander separiert werden können. Er nahm ein Zweistufenmodell für die Trennung der beiden Parameter speed und power an. Mit Hilfe des zweikategoriellen logistischen Modells konnte Fischer (1970, S. 398) eine Kovarianz zwischen Itemschwierigkeit und Zeitbedarf eines Items nachweisen. Diese lässt sich nach Fischer (1970, S. 397) damit erklären, dass schwierige Aufgaben allgemein mehr Zeit zur Bearbeitung benötigen, da die Schwierigkeit wie auch der Zeitbedarf eines Items von der Anzahl logischen oder anschaulichen Elementaroperationen im Lösungsprozess und von den nötigen Behaltensleistungen abhängen.

Auch Nährer (1986 S. 48) beschäftigte sich mit dem Zusammenhang zwischen den beiden Komponenten speed und power. Er geht von einer skalaren und vektoriellen Auffassung aus, die gleichzusetzen ist mit den oben formulierten Hypothesen von Fischer (1970).

Die skalare Auffassung nimmt eine starke Korrelation zwischen Schnelligkeit und Fähigkeit einer Person an und geht davon aus, dass ihr Problemlöseverhalten durch einen einzigen Parameter charakterisiert werden kann (Nährer, 1986, S. 48).

Die vektorielle Auffassung geht davon aus, dass die beiden Komponenten speed und power voneinander unabhängig sind und eine Versuchsperson daher nicht durch einen einzigen Parameter beschrieben werden kann. Demzufolge benötigt man mindestens zwei Parameter (Nährer, 1986, S. 48).

Nährer (1986, S. 54) formulierte zwei Lösungsstrategien, um die vektorielle Annahme prüfen zu können. Bei Strategie D (direkt) bleiben irrelevante und gleichsam redundante Informationen unberücksichtigt. Bei Strategie U (umständlich) werden auch redundante Informationen berücksichtigt. Nährer (1986, S. 54) geht davon aus, dass mit Hilfe der Strategie D die Bearbeitung leichter Aufgaben relativ wenig Zeit benötigt, jedoch bei schwierigen oder komplexeren Aufgaben Strategie U zur schnelleren Lösung der Aufgaben führt.

Zur Überprüfung der beiden Auffassungen wählte Nährer (1986, S. 59 ff) Denkaufgaben vom Typ des Figuren-Mengen-Test (SCHEIBLECHNER 1972). Die Items des Tests weisen unterschiedliche Schwierigkeitsgrade auf und sie können sowohl mit Strategie U als auch mit Strategie D gelöst werden. Die Items wurden ohne Zeitgrenze vorgegeben mit der Aufforderung die Aufgaben möglichst richtig in möglichst kurzer Zeit zu bearbeiten (maximal speed- maximal accuracy) (Nährer, S. 66).

Die Korrelation zwischen Fähigkeit und Bearbeitungszeit ist sehr niedrig und liegt bei r = -.21. Die skalare Auffassung in der strengen Formulierung, dass ohne Vorgabe einer Zeitgrenze, die fähigeren Probanden auch die schnelleren sind, kann daher nicht angenommen werden. Zumindest entspricht die Richtung des Zusammenhangs der Hypothese (Nährer, 1986, S. 75). Bei der Betrachtung der Itemschwierigkeit und der Bearbeitungszeit zeigt sich, dass die fähigeren Personen sowohl die leichten als auch die schwierigen Aufgaben schneller bearbeiten. Das unterstützt vorerst die skalare Auffassung (Nährer 1986, S. 75). Unterteilt man die Versuchspersonen jedoch nach der von ihnen angewanden Strategien, bearbeiten Versuchspersonen mit Strategie D leichte Items schneller, als jene Probanden die Strategie U anwenden. Schwierige Aufgaben lösen jedoch die Testpersonen mit Strategie U schneller. Dieses deutet auf eine

klassische Wechselwirkung zwischen Itemschwierigkeit und angewandter Strategie hin. Weiters lösen Testpersonen mit Strategie U sowohl bei leichten als auch bei schwierigen Aufgaben mehr Items richtig. Diese Ergebnisse sprechen eher für eine vektorielle Auffassung (Nährer, 1986, S. 76).

Ebenso für eine vektorielle Auffassung spricht die Analyse von Zeitbegrenzung und Bearbeitungsstrategie. Versuchspersonen mit Strategie U bearbeiten die Aufgaben bei geringem Zeitdruck besser. Mit zunehmendem Zeitdruck erzielen jedoch die Probanden mit Strategie D ein besseres Ergebnis (Nährer 1986, S. 86). Ausgehend von den Ergebnissen von Nährer (1986) kann man insgesamt gesehen, eher von einer vektoriellen Aufassung ausgehen.

Wedening (1991) beschäftigte sich ebenfalls mit der Frage, ob die beiden Parameter Leistungsgeschwindigkeit und Leistungsgüte auf eine gemeinsame Dimension zurück zuführen sind, oder ob mindestens zwei Parameter notwendig sind, um das Leistungsverhalten einer Person zu charakterisieren. Er bildete sechs Personengruppen, deren Testsituationen unterschiedlich waren. Die Versuchspersonen mussten dabei acht Subtests bearbeiten mit jeweils unterschiedlichen Zeitgrenzen je Item. Jeder Untertest bestand aus acht Matrizenitems, deren Schwierigkeitsgrad kontinuierlich anstieg (Wedening, 1991).Nach Wedening (1991) zeigte sich, dass der Zusammenhang zwischen Leistungsgüte und Leistungsgeschwindigkeit durch die Testdurchführung beeinflusst wird. Werden die Versuchspersonen von Beginn an unter Zeitdruck gesetzt erhält man Testergebnisse die eher für einen starken Zusammenhang zwischen den beiden Parametern sprechen. Diesem Ergebnis zufolge könnte eine Person anhand eines einzigen Parameters charakterisiert werden. Jedoch sollte nach Wedening (1991) aus Gründen der Objektivität eine Versuchsperson hinsichtlich ihrer Leistung durch mindestens zwei Parameter charakterisiert werden. Durch die Prüfung des Zusammenhangs zwischen intellektuellen und nichtintellektuellen Merkmalen einer Testperson wird dieses noch verstärkt.

Hornke (1997) gab einer Stichprobe von 110 Versuchspersonen einen computergestützten adaptiven Matrizentest vor. Nach Hornke (1970) werden richtig gelöste Matrizenaufgaben im Mittel rascher bearbeitet als falsch gelöste Items. Eine richtige Antwort wird um etwa ¼ schneller gegeben, als eine falsche. Nach Hornke

(1997) sprechen die Ergebnisse dafür, dass es keine substanzielle Korrelation zwischen der Leistung und der Bearbeitungszeit einer Versuchsperson gibt.

5 Geschlechtsstereotype, Geschlechtsrollen und Androgynie

Nachfolgend werden die Begriffe Geschlechtsstereotype, Geschlechtsrollen und Androgynie näher beschrieben.

5.1 Geschlechtsstereotype und Geschlechtsrollen

Stereotype sind allgemeine Annahmen über die relevanten Eigenschaften einer Personengruppe. Ausgangspunkt von Stereotypen ist ein Kategorisierungsprozess. Personen derselben Kategorie werden als ähnlich, Menschen unterschiedlicher Kategorien werden als unähnlich angesehen. (Alfermann, 2001 S. 30).

Eine Kategorie, die sehr bedeutend und in jeder Kultur vorhanden ist, ist das *biologische Geschlecht* (männlich/weiblich) (Alfermann, 2001 S. 31). Die Erfassung des biologischen Geschlechts erfolgt über die Wahrnehmung und Klassifikation primärer Geschlechtsmerkmale (Altstötter-Gleich, 1996, S. 95). Nach Alfermann (2001 S. 31) stellt die eigene biologische Geschlechtsidentität, das heißt die Erkenntnis über die Zugehörigkeit zur richtigen Geschlechterkategorie (entweder Junge/männlich oder Mädchen/weiblich) einen wichtigen Teil in der frühkindlichen Entwicklung dar.

Mit dem biologischen Geschlecht werden typische Eigenschaften und Handlungsweisen verknüpft. Diese sind aber weniger biologisch, sondern viel mehr sozial erworben und werden als Geschlechtsrollen bezeichnet. (Alfermann, 2001, S. 34). So gibt es Geschlechtesrollen-Stereotype, die nach Hampson (1986) wie folgt definiert werden “Sex-role stereotypes refer to beliefs about the appropriate activities for man and women“ (S.45). Maskulinität ist gekennzeichnet durch Aktivität, Kompetenz, Leistungsstreben und Durchsetzungsfähigkeit während Femininität Eigenschaften von Emotionalität (sanft, weinerlich), Soziabilität (einfühlsam, sozial, anpassungsfähig), Passivität und praktischer Intelligenz enthält (Alfermann, 2001, S. 32).

Die Geschlechtsrollen sind verbunden mit Geschlechtsrollenerwartungen. Werden diese von einem Menschen in das eigene Selbstbild übernommen, das sich die Person von sich selbst als Junge/Mann bzw. Mädchen/Frau macht, so spricht man von Geschlechtsrollenidentität. Handelt es sich um maskuline Inhalte wird von maskuliner Identität (Maskulinität) gesprochen, bei femininen Inhalten wird von femininer Identität (Femininität) gesprochen. Dieses *psychologische Geschlecht* kann bedeutend variabler ausfallen, als das biologische Geschlecht. Zum Beispiel kann ein feminines Selbstbild einhergehen mit maskulinen Eigenschaften. Maskuline Interessen wie zum Beispiel Fussbalspielen können einhergehen mit einer femininen Berufswahl (die fussballspielende Krankenschwester) (Alfermann, 2001, S. 35).

5.2 Androgynie

Nach C.G. Jung (1971, zitiert nach Hassler, 1990. S. 47) vereinigt das Wort Androgynie die griechischen Wörter für Mann und Frau. Es soll die Möglichkeit aufzeigen, dass männliche und weibliche Charakteristika in einer Person vereint sind.

Nach C.G. Jung (1971, zitiert nach Hassler, 1990, S. 47) sind Männlichkeit und Weiblichkeit in jedem Menschen vorhanden. Die Männlichkeit in der Frau (Animus) und die Weiblichkeit im Mann (Anima) streben danach wahrgenommen zu werden und in die Persönlichkeit integriert zu werden.

*"*Androgynie bedeutet auf der Persönlichkeitsebene die Kombination von positiven typisch weiblichen (femininen) und typisch männlichen (maskulinen) Eigenschaften" (Alfermann, 2001, S. 37).

Nach Alfermann (2001, S. 29) sind androgyne Menschen besser in der Lage ihre Fähigkeiten der Situation angemessen einzusetzen, da sie sich und andere Personen nicht nach der Geschlechtsrollenerwartung kategorisieren.

5.2.1 Das Androgyniekonzept

Lange Zeit wurde eine Dimension der psychologischen Geschlechtsrollenidentität angenommen, in der das gemeinsame Auftreten von Maskulinität und Femininität

ausgeschlossen war. Die beiden Kategorien wurden als zwei entgegengesetzte Pole eines Kontinuums dargestellt. Neuere Ansätze zur Einteilung der Geschlechtsrollenidentität gehen davon aus, dass die psychologische Geschlechtsrollenorientierung auf (mindestens) zwei Dimensionen, einer Maskulinitäts – und einer Femininitätsdimension, anzusiedeln ist. Unabhängig vom biologischen Geschlecht kann ein Mann oder eine Frau auf den beiden Dimensionen jeden beliebigen Punkt einnehmen (Alfermann, 2001, S. 36).

Nach Altstötter-Gleich (1996, S. 109) lassen sich aufgrund der mehrdimensionalen Betrachtung 4 Typen von Personen festlegen: maskuline, feminine, androgyne und undifferenzierte. Diese Personentypen ergeben sich aus der individuellen, vom biologischen Geschlecht unabhängigen Lokalisation auf den beiden Skalen:

	hohe Maskulinitätswerte	niedrige Maskulinitätswerte
hohe Feminitätswerte	Androgynität	Feminität
niedrige Feminitätswerte	Maskulinität	Undifferenziertheit

Abbildung 1: Geschlechtsidentität nach Altstötter-Gleich (1996, S. 109)

Ein Beispiel eines Mehrdimensionalen Verfahren ist das von Sandra L. Bem 1974 entwickelte Bem Sex-Role Inventory (BSRI). Das BSRI ist ein Fragebogen zur Selbsteinschätzung der Geschlechtsrolle und enthält eine Femininitätsskala eine Maskulinitätsskala sowie eine neutrale Skala. Jede dieser Skalen besteht aus 20 Persönlichkeitsmerkmalen. Die beiden Skalen “Femininität“ (F) und “Maskulinität“ (M) werden als von einander unabhängig gesehen (Bem, 1974). Personen, die im Selbsteinschätzungsfragebogen zur Geschlechteridentität hohe Werte auf der M-Skala haben, beschreiben sich als aggressiv, analytisch, ambitioniert, dominant, stark und unabhängig. Personen mit hohen F-Skalawerten beschreiben sich als gefühlsbetont, kindlich, sanft, loyal, schüchtern, und warmherzig (nach Amelang et al 2011, S. 551).

5.2.2 Geschlechtsrollenidentität und Raumvorstellung

Broverman, Broverman, Vogel, Palmer & Klaiber (1964, zit. nach Hassler, 1990, S. 22) untersuchte bei männlichen Jugendlichen und jungen Erwachsenen den Einfluss von Geschlechtshormonen auf kognitive Fähigkeiten. Die eingesetzten Tests prüften die verbalen, motorischen und räumlichen Fähigkeiten der Probanden. Sie kamen zu dem Ergebnis, dass sehr männliche Männer, eingeschätzt nach ihren sekundären Geschlechtsmerkmalen, schlechter bei räumlichen Aufgaben abschnitten und besser bei der Bearbeitung von sprachlichen und motorischen Items waren. Sehr männliche Männer hatten also eher ein weibliches kognitives Muster. Weniger männliche Männer hatten schlechtere motorische und verbale Fähigkeiten jedoch bessere räumliche. Nach Broverman et al. (1964 zit. .nach Hassler, 1990, S. 23) besitzen demnach physisch androgyne Personen höhere räumliche Begabung als geschlechtstypische Versuchspersonen.

In der Metaanalyse von Signorella und Jamison (1986) unter anderem über den Zusammenhang zwischen Geschlechtsrollenidentität und räumliches Vorstellungsvermögen zeigte sich,…“that persons who describe themselves as more masculine and/or less feminine do better than persons who describe themselves as more feminine and/or less masculine“ (S.218). Bei Aufgaben zur mentalen Rotation erzielten Personen mit hohen Maskulinitäts- und Femininitätswerten sowie Personen mit hohen Maskulintätswerten bessere Leistungen (Signorella & Jamison, 1986).

In einer Studie von Adlmann und Gittler (2010) wurde den Versuchspersonen ein Rasch-homogener Raumvorstellungstest vorgegeben, um mentale Rotation zu erheben. Weiters bearbeiteten die Testpersonen einen Fragebogen zur Erfassung der Geschlechtsrollenidentität, welcher aus 63 Adjektiven besteht, die aufgrund von Voruntersuchungen als männlich, weiblich oder neutral klassifiziert wurden. Es zeigte sich, dass die männlichen Probanden signifikant besser bei der Bearbeitung des Raumvorstellungstest, welcher mentale Rotation misst, abschnitten. Weiters ergab die Analyse der Daten, dass die Gruppe “androgyn“ signifikant höhere mentale Rotationsleistung als die Gruppe “undifferenziert“ erbrachte (Adlmann & Gittler, 2010).

6 Geschlechtsspezifische Unterschiede in der Raumvorstellung

In der Literatur wird immer wieder von Leistungsunterschieden zwischen den Geschlechtern berichtet, jedoch finden sich im kognitiven Bereich die größten Geschlechtsunterschiede in den räumlichen Fähigkeiten (Asendorf, 2007).

Maccoby und Jacklin (1974) stellten in ihrer Analyse von 1400 Untersuchungen zum Thema Geschlechtsunterschiede fest, dass Männer über bessere räumliche sowie mathematische Fähigkeiten verfügen, und Frauen bessere sprachliche Fähigkeiten besitzen (Maccoby & Jacklin, 1974).

Linn und Petersen (1985) kamen in ihrer Metaanalyse zu Geschlechtsunterschieden im Faktor Raumvorstellung zu dem Ergebnis, dass Geschlechtsunterschiede nicht in allen Komponenten der Raumvorstellung gleich hoch auftreten. Die größten Unterschiede sind im Bereich der mentalen Rotation zu finden. Geringere Unterschiede stellten sie bei Aufgaben zur räumlichen Wahrnehmung fest (Linn & Petersen, 1985).

Voyer, Voyer und Bryden, (1995) führten eine Metaanalyse aufgrund mehrerer Untersuchungsergebnisse zu geschlechtsspezifischen Differenzen bei Raumvorstellung durch und kamen dabei ebenfalls zu der Ansicht, dass der Geschlechtsunterschied bei mentaler Rotation am höchsten ist (Voyer, Voyer & Bryden, 1995).

In einigen Studien zeigte sich, dass sich der Geschlechtsunterschied zwischen männlichen und weiblichen Versuchspersonen in den verschiedenen Bereichen der Raumvorstellung verringert, wenn der Einfluss der Zeit minimiert wird (Maier, 1996).

In einer eigenen unveröffentlichen Studie von Gittler und Spiel (1985, zitiert nach Maier, 1999) wird darauf hingewiesen, dass Frauen bei der Bearbeitung von reinen power-Tests mehr Zeit zur Lösung der Aufgaben benötigen, aber nur geringfügig schlechter abschneiden, als Männer. Bei der Bearbeitung von reinen Speed-Tests jedoch weisen die weiblichen Probanden deutlich schlechtere Testergebnisse auf, als die männlichen.

Peters (2005) untersuchte in seiner Studie, ob der Faktor Zeit bei der Bearbeitung des Mental Rotation Tests (MRT) (Mental Rotation Test; Vandenberg & Kuse, 1978; Peters, Laeng, Latham, Jackson, Zaiyouna, & Richardson, 1995) Einfluss auf den Geschlechtsunterschied hat. Die erste Studie umfasste 1765 Probanden und wurde unter Standardbedingungen durchgeführt. Es zeigte sich, dass sich der Geschlechtsunterschied mit fortschreitender Testdauer erhöht und dass innerhalb eines vorgegebenen Zeitrahmens Frauen weniger Aufgaben lösen als Männer. Daher wurde in einer zweiten Studie die Bearbeitungszeit verdoppelt. Dabei zeigte sich, dass beide Geschlechter bessere Ergebnisse erbrachten, jedoch blieb der signifikante Geschlechtsunterschied zugunsten der Männer bestehen (Peters 2005).

Nach Kerkman, Wise und Harwood (2000, zitiert nach Lautenbacher, Güntürkün & Hausmann) schneiden weibliche Probanden bei Aufgaben zur mentalen Rotation in etwa gleich gut wie die männlichen Versuchspersonen ab, wenn sie nur zwei rotierte Figuren miteinander vergleichen müssen. Müssen mehrere Figuren mit dem Originalitem verglichen werden, erzielen Männer bessere Ergebnisse (Kerkman, Wise und Harwood (2000, zitiert nach Lautenbacher, Güntürkün & Hausmann).

Titze, Heil und Jansen (2008) gaben ihren Probanden den MRT (Mental Rotation Test; Vandenberg & Kuse, 1978; Peters, Laeng, Latham, Jackson, Zaiyouna, & Richardson, 1995) vor. Der MRT besteht aus zwei Aufgabensets mit jeweils12 Items. Das aus zusammengesetzten Würfeln vorgegebene Zielitem befindet sich auf der linken Seite und muss mit den vier Items auf der rechten Seite verglichen werden. Die Versuchspersonen müssen jene zwei Vergleichsitems identifizieren, die das Zielitem in rotierter Version zeigen. Ein Teil der Versuchspersonen bearbeitete den MRT in der Orginalversion, jedoch ohne Zeitvorgabe. Die Versuchspersonen mussten jene zwei Vergleichsaufgaben identifizieren, welche die Ausgangsaufgabe in rotierter Form zeigen. Die anderen Probanden bearbeiteten den MRT in einer weniger komplexen Form, sie mussten jeweils nur ein Item mit den Originalitem vergleichen, die übrigen drei Vergleichsitems waren unterdessen durch eine Schablone verdeckt und für den Probanden nicht sichtbar. So mussten die Versuchspersonen immer nur zwei Items miteinander vergleichen. Jedoch erzielten auch bei den weniger komplexen Aufgaben die männlichen Versuchspersonen durchschnittlich bessere Ergebnisse als die Frauen Titze, Heil und Jansen, 2008).

7 Die Gütekriterien

“Wenn man einen Test konstruieren will, muss man eine Vorstellung davon haben, was einen ’guten’ Test auszeichnet. Das ist die Frage nach den sogenannten *Gütekriterien* für Tests“ (Rost, 2004, S. 33).

Nach (Lienert & Raatz, 1998, S. 7) wird zwischen Haupt- und Nebengütekriterien unterschieden. Zu den Hauptgütekriterien zählen die Objektivität, Reliabilität und die Validität. Nach den Nebengütekriterien soll ein guter Test normiert, vergleichbar, ökonomisch und nützlich sein.

Es wird nachfolgend nur auf die Hauptgütekriterien näher eingegangen, nähere Ausführungen zu den Haupt- sowie Nebengütekriterien sind in den Büchern Lienert & Raatz, 1998 und Rost, 2004 sowie Kubinger, 2006 nachzulesen.

7.1 Die Hauptgütekriterien

7.1.1 Objektivität

“Unter Objektivität eines Tests verstehen wir den Grad, in dem die Ergebnisse eines Tests unabhängig vom Untersucher sind“ (Lienert & Raatz, 1998, S. 7). Demnach wäre ein Test vollkommen objektiv, wenn verschiedene Testleiter bei denselben Versuchspersonen zu gleichen Testergebnissen kommen (Lienert & Raatz, 1998, S. 7). Eine hohe Objektivität der Testdurchführung ist nach Rost (2004, S. 39) eine Voraussetzung für die Reliabilität und die Validität. Bei der Testvorgabe ist zu beachten, dass das Testergebnis unabhängig davon ist, wer den Test vorgibt (Durchführungsobjektivität), wer den Test auswertet (Auswertungsobjektivität) und wer den Test interpretiert (Interpretationsobjektivität) (Rost, 2004 S. 39).

7.1.2 Reliabilität

“Unter Reliabilität oder zu deutsch *Zuverlässigkeit* eines Tests bezeichnet die Präzision oder Genauigkeit, mit der ein Test eine Personeneigenschaft misst“ (Rost, 2004, S. 36).

Ein Test ist als reliabel anzusehen, wenn die mit Hilfe des Tests erzielten Testergebnisse die Versuchsperson genau, d.h fehlerfrei beschreiben (Lienert & Raatz ,1998, S. 9). Nach Lienert und Raatz (1998, S.9ff.) werden verschiedene Aspekte der Reliabilität eines Tests unterschieden:

Paralleltest- Reliabilität: Eine Stichprobe von Versuchspersonen bearbeitet zwei miteinander streng vergleichbare Tests (Paralleltests) und die Testergebnisse werden korreliert (Paralleltest-Methode).

Retest-Reliabilität: Den Versuchspersonen wird ein und derselben Test zweimal vorgegeben (Testwiederholung) und die Testergebnisse beider Zeitpunkte werden korreliert (Retest-Methode).

Innere Konsistenz:

- *Testhalbierung* (Halbierungsreliabilität oder –konsistenz): Die vorgegebenen Items eines Tests werden in zwei Hälften geteilt und das Testergebnis für jede Testhälfte getrennt ermittelt (split-half methode). Anschließend werden die Testergebnisse der beiden Testshälften korreliert. Der für den halbierten Test geltende Reliabilitätskoeffizient wird aufgewertet, so dass er für den ganzen Test Gültigkeit hat (Halbierungs-oder Split-Half-Konsistenzkoeffizient) (Lienert & Raatz, 1998, S. 10).

- *Konsistenzanalyse*: Ein Test wird in so viele Teile geteilt, wie er Items hat (Kubinger, 2006). Die Reliabilität wird indirekt über bestimmte Kennwerte der Testteile, wie zum Beispiel die Aufgabenschwierigkeits- und Trennschärfestatistiken, ermittelt (Lienert & Raatz, 1998, S. 10).

7.1.3 Validität

„Die *Validität* oder die Gültigkeit eines Tests gibt den Grad der Genauigkeit an, mit dem dieser Test dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das (die) er messen oder vorhersagen soll, tatsächlich misst oder vorhersagt“ (Lienert & Raatz, 1998, S. 10).

Ein Test ist vollkommen valide, wenn die Ergebnisse einen unmittelbaren und fehlerfreien Rückschluss auf den Ausprägungsgrad des zu erfassenden Persönlichkeits- oder Verhaltensmerkmals zulassen (Lienert & Raatz, 1998, S. 10). Nach Kubinger (2006, S. 50) stellt die Validität eines Tests das wichtigste Gütekriterium dar, ist aber am schwierigsten zu überprüfen.

Es werden verschiedene Konzepte der Validität unterschieden (Kubinger, 2006, S.50):

Nach Lienert und Raatz (1998, S. 10) ist von inhaltlicher Gültigkeit eines Tests zu sprechen, wenn dieser selbst das optimale Kriterium für das Persönlichkeitsmerkmal darstellt.

Von Konstruktvalidität spricht man, wenn ein Test gewisse theoretische respektive theoriegeleitete Vorstellungen in Bezug auf irgendein so genanntes „Konstrukt“ erfüllt (Kubinger, 2006, S. 53).

Bei der Kriteriumsvalidität wird eine bestimmte als relevant angesehene Variable (sog. „Außenkriterium“) mit dem interessierenden Test korreliert. wobei die Außenkriterien folgendermaßen eingeteilt werden können (Kubinger, 2006, S. 61):

1. Die „Vorhersagegültigkeit“ oder „prognostische Validität“ bestimmt sich aus der Korrelation des fraglichen Tests mit einem Außenkriterium, das in der fernen Zukunft liegt (z.B Prüfungserfolg); ein bestimmtes später beobachtbares Kriterium soll also vorhergesagt werden.

2. Die sog. „Übereinstimmungsvalidität“ bezieht sich auf sie Korrelation mit einem anderen Test, der (angeblich) dasselbe Konstrukt erfasst.

8 Korrelation von Raumvorstellung mit Schulleistungen

Im schulischen Bereich stellt ein gut ausgeprägtes räumliches Vorstellungsvermögen einen großen Vorteil dar. In einigen Studien konnte nachgewiesen werden, dass zwischen Raumvorstellung und den Leistungen in der Schule ein positiver Zusammenhang besteht. Wobei die stärksten Zusammenhänge zwischen Raumvorstellung und den Leistungen in den Fächern Mathematik, Chemie, Physik und Werken gefunden wurden (Maier, 1999).

Nach Smith (1964) gibt es einen positiven Zusammenhang zwischen höherer Mathematik und Raumvorstellung. Demnach haben Personen, die gut bei der Lösung schwieriger Mathematikaufgaben sind, ein besseres Raumvorstellungsvermögen, als Personen, die diese Mathematikaufgaben schlechter lösen. Nach Smith (1964) lässt sich jedoch zwischen einfachen Mathematikaufgaben, wie etwa Kopfrechnen, und den Fähigkeiten bei Raumvorstellungsaufgaben kein signifikanter Zusammenhang finden.

Guay und Mc Daniel (1977) führten eine Untersuchung zum Zusammenhang von einfachen sowie schwierigen Raumvorstellungsaufgaben und leichten Mathematikaufgaben bei Grundschulkindern durch. “Low–level spatial ability were defined as requiring the visualization of two-dimensional configurations, but no mental transformations of these visual images. High–level spatial ability were characterized as requiring the visualization of three-dimensional configurations, and the mental manipulation of these visual images” (Guay und Mc Daniel (1977, S. 211). Die Mathematikleistung der Schüler wurde mit Hilfe des “*Iowa Tests of Basic skills (ITBS)”* erhoben. Guay und Mc Daniel (1977) kamen in ihrer Studie zu dem Ergebnis, dass es einen positiven Zusammenhang zwischen einfachen Mathematikaufgaben und einfachen sowie auch schwierigen Raumvorstellungsaufgaben gibt.

Lehmann und Jüling (2003) gaben einer Stichprobe von 10 bis 11 Jährigen Schülerinnen und Schülern verschiedener Schulformen den Mental Rotation Test MRT (Peters et al., 1995) und den dreidimensionalen Würfeltest 3DW (Gittler, 1990) zur Erfassung der räumlichen Vorstellungsfähigkeit vor. Die mathematischen Leistungen wurden mit einem von Lehmann und Jüling (2003) selbst entwickelten Mathematiktest gemessen, der als Subskalen Zahlenreihen, Rechenaufgaben und algebraische Probleme beinhaltete. Weiters wurde ein kognitiver Fähigkeitstest (KFT 4-13+) von Heller, Gaedike und

Weinländer (1985) vorgegeben. Nach Lehmann und Jüling (2003) sind Raumvorstellungsfähigkeiten nicht nur für die Bearbeitung von Geometrieaufgaben hilfreich, sondern auch für algebraische – arithmetische Anforderungen. In ihrer Studie zeigte sich, dass mathematisch überdurchschnittlich leistungsfähige Schülergruppen über ein sehr gutes räumliches Vorstellungsvermögen verfügen. Besonders ausgeprägt war dabei die Fähigkeit zur mentalen Rotation. Nach Lehmann und Jüling (2003) ist der Zusammenhang zwischen Raumvorstellungs- und mathematischen Leistungen umso stärker, je weniger Algorithmen zur Lösung einer Aufgabe eingesetzt werden können.

Carter, LaRussa und Bodner (1987) gaben zwei Gruppen von Studenten Raumvorstellungstests vor. Die Ergebnisse lassen darauf schließen, dass es zwischen den Noten in Chemie und den Leistungen in Raumvorstellung einen signifikanten Zusammenhang gibt (Carter, LaRussa und Bodner, 1987).

Ausgehend davon, dass ausgeprägte Raumvorstellungsfähigkeiten eine wichtige Vorraussetzung für mindestens 84 Berufsbilder darstellen, haben Sorby, Charlesworth und Drummer (2006) eine Untersuchung vorgenommen, die diese Annahme für die Bereiche organische Chemie und fortgeschrittene Chemie überprüfen. Es wurden Studenten der Technischen Universität Michigan einem Raumvorstellungstest, dem Purdue Spatial Visualization Test (PSVT:R), unterzogen. Voraussetzung war, dass die Probanden eines von zwei verschieden fortgeschrittenen Chemieseminaren (Chemistry II bzw. Organic Chemistry II) belegten. 152 Teilnehmer an Chemistry II bzw. 57 Teilnehmer an Organic Chemistry II legten den PSVT:R ab. Die Ergebnisse des PSVT:R wurden in Folge mit dem Abschneiden in den genannten Chemie Seminaren in Relation gesetzt. Im Falle von Chemistry II konnte ein statistisch signifikanter Zusammenhang zwischen Raumvorstellungsfähigkeit und dem (positivem) Seminarabschneiden festgestellt werden. Im Falle von Organic Chemistry II gelang dies nicht. Hingegen wurde in beiden Fällen ein robuster Zusammenhang zwischen Vorerfahrung in geometrischem Zeichnen und dem (positiven) Seminarabschneiden belegt. Auch die bereits hinreichend untersuchte geschlechtsspezifische Tendenz, dass Männer über besseres Raumvorstellungsvermögen verfügen, wurde in dieser Studie bestätigt (Sorby, Charlesworth und Drummer, 2006).

Auch in den Fächern Werkerziehung und Technisches Werken konnten positive Zusammenhänge mit der Fähigkeit zur Raumvorstellung nachgewiesen werden (Stückrath, 1968).

EMPIRIETEIL

9 Zielsetzung und Fragestellung

Ziel der vorliegenden Studie ist, die Items des zwei dimensionalen Endlosschleifentests (EST), der unter power sowie work-limit Bedingung vorgegeben wurde, zu validieren.

9.1 Fragestellungen

9.1.1 Rasch-Homogenität des Testmaterials

Der Endlosschleifentest (Gittler & Arendasy, 2003) gilt unter power Vorgabe, das bedeutet ohne Beschränkung der Bearbeitungszeit, als Rasch-homogen.

Es wurde einerseits untersucht, ob die Items unter power Bedingung als auch unter work-limit Bedingung eindimensional im Sinne von Rasch messen, andererseits wurde der Gesamttest auf Rasch-Homogenität überprüft.

9.1.2 Auswirkungen der work- limit Instruktion auf die Leistung

Die work-limit Instruktion, die vor der Bearbeitung der zweiten Testhälfte vorgegeben wurde, enthält im Gegensatz zur reinen power Vorgabe, eine Zeitdruckkomponente. Es wurde der Frage nachgegangen, ob aufgrund der work-limit Vorgabe im Vergleich zur power Vorgabe zusätzlich diagnostisch relevante Informationen bezüglich der Leistung gewonnen werden kann.

9.1.3 Korrelationen der unterschiedlichen Instruktionen mit den Außenkriterien

Weiters wurde die Frage untersucht, ob unter Vorgabe der unterschiedlichen Instruktionen (power oder work-limit) Unterschiede in der Höhe der Korrelationen mit den Außenkriterien zu beobachten sind.

Als Außenkriterium wurden die durchschnittlichen Schulnoten der Testteilnehmer in den Unterrichtsfächern Mathematik, Chemie, Physik, Englisch, Deutsch und Latein herangezogen. Es konnten positive Zusammenhänge zwischen den Leistungen im Schulfach Chemie (Carter, LaRussa & Bodner, 1987), Physik (Maier, 1994) und

Mathematik (Pearson & Ferguson, 1989) und den Raumvorstellungsfähigkeiten festgestellt werden.

Des Weiteren wurde die Selbsteinschätzung in Mathematik, Technik, Musik und Raumvorstellung erhoben. Die Ergebnisse sprechen dafür, dass zwischen Leistung und Selbsteinschätzung der Leistung ein Zusammenhang besteht (Epstein, 1984, zitiert nach Laskowski, 2000).

Ein weiteres Außenkriterium stellt der Planzeichentest dar. Das Abrufen von kognitiven Karten aus dem Gedächtnis durch das Skizzieren von Karten erlaubt Rückschlüsse auf das Raumvorstellungsvermögen.[1]

9.1.4 Unterschiede in der Raumvorstellung in Abhängigkeit vom biologischen und psychologischen Geschlecht

Es wurde zudem der Frage nachgegangen, ob Geschlechtsunterschiede im Bezug auf die räumliche Vorstellungsfähigkeit angenommen werden können, da in zahlreichen Studien festgestellt werden konnte, dass Männer über ein besseres Raumvorstellungsvermögen, als Frauen verfügen (vgl. Linn & Petersen 1985; Voyer, Voyer & Bryden, 1995). Weiters wurde untersucht, ob die Vorgabe der unterschiedlichen Instruktionen einen Einfluss auf den Geschlechtsunterschied hat. In einer Studie von Gittler und Spiel (1985, zitiert nach Maier, 1999, S. 194) konnte gezeigt werden, dass sich der Geschlechtsunterschied bei der Vorgabe von Raumvorstellungsaufgaben ohne Zeitdruck verringert. Peters (2005) kam wiederum zu dem Ergebnis, dass die Vorgabe der Items ohne Zeitdruck keinen Effekt auf den Geschlechtsunterschied hat.

In Bezug auf die Geschlechtsrollenidentität konnte festgestellt werden, dass Personen, die sich als eher maskulin und wenig feminin einschätzen, besser bei der Bearbeitung von Raumvorstellungen abschneiden, als Personen, die sich als mehr feminin und weniger maskulin beschreiben. Des weiteren konnte gezeigt werden, dass Personen mit hoher Ausprägung auf der Femininitäts- (F) und Maskulinitätsskala (M), die somit der Gruppe „androgyn“ zugeordnet werden, besser bei Aufgaben zur mentalen

[1] URL: http://www.geodz.com/deu/d/kognitive_Karte, Zugriff am 16.08.2011

Rotationsleistung abschneiden, als jene, die auf beiden Skalen niedrige Ausprägungen aufweisen und der Gruppe „undifferenziert“ zugeordnet werden (vgl. Signorella & Jamison, 1986 und Adlmann & Gittler, 2010).

9.2 Erhebungsinstrumente

9.2.1 Der Endlosschleifentest (EST)

Der Endlosschleifentest wurde zur Messung des räumlichen Vorstellungsvermögens eingesetzt. Ausgangspunkt für die Entwicklung des rasch-homogenen Endlosschleifentests (Gittler & Arendasy, 2003) waren psychometrische Untersuchungen zum Schlauchfiguren-Test (Stumpf & Fay, 1983) im Sommersemester 1995 im „Forschungsseminar für Differentielle Psychologie“ unter der Leitung von Univ.-Prof. Dr. Gittler. Die Analyse nach dem dichotomen logistischen Modell nach Rasch ergab, dass für die Items des Schlaufigurentests (SFT) keine Rasch-Homogenität angenommen werden kann (Arendasy, 2000). Rasch-homogen bedeutet, dass die Items für alle Personen bzw. für alle Personengruppen einer Population dieselbe latente Fähigkeitsdimension messen (Gittler & Arendasy, 2003).

Beim Endlosschleifentest (Gittler & Arendasy, 2003) wird der Proband angewiesen, die abgebildeten Endlosschleifen mental zu rotieren. Der Versuchsperson wird die Endlosschleife in der Startansicht und Zielansicht vorgegeben und soll wählen, um wie viel Grad (90° oder 180°) und in welche Drehrichtung (nach oben, unten, links oder rechts) die Endlosschleife der Startansicht gedreht werden muss, damit Start - und Zielansicht ident sind. Zusätzlich gibt es die Antwortmöglichkeit „ich weiss nicht“ (siehe Abbildung 2).

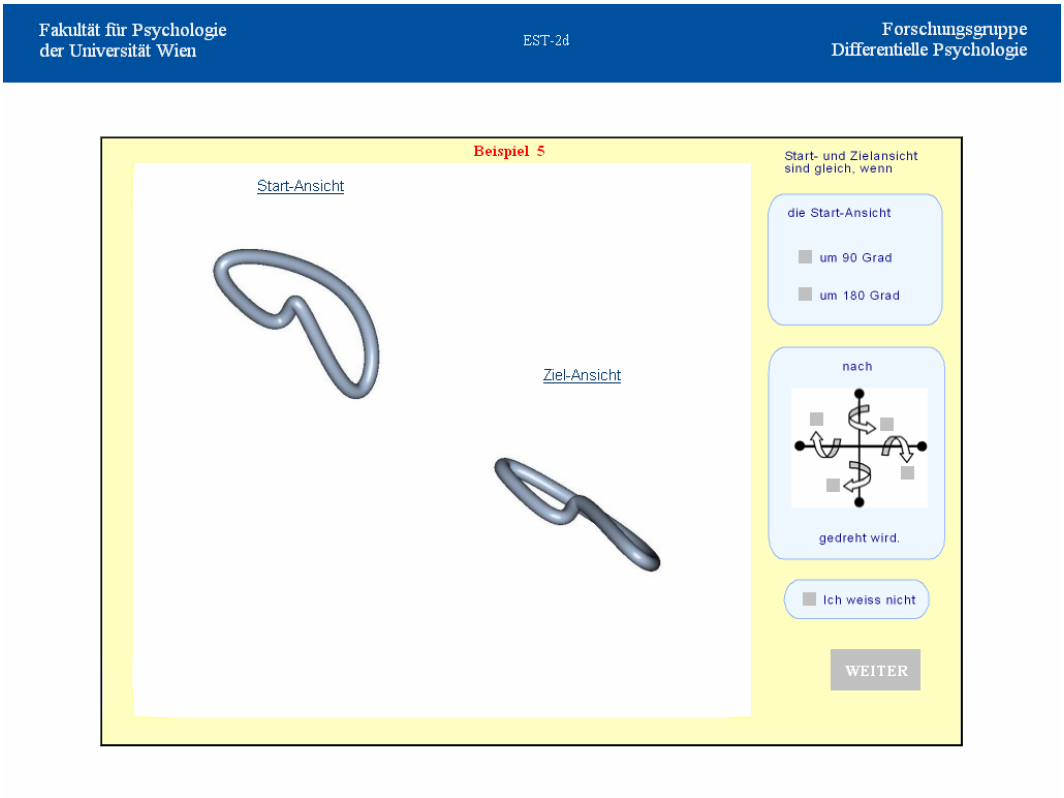


Abbildung 2: Beispielaufgabe 5 aus dem Endlosschleifentest (Gittler & Arendasy, 2003)

9.2.2 Fragebogen zur Persönlichkeitseinschätzung “Selbstbild“

Die Forschungsversion des Fragebogens zur Persönlichkeitseinschätzung „Selbstbild“ (unveröffentlichte Version; Gittler, 2003-2009) dient der Erfassung der Geschlechtsrollenidentität und besteht aus 63 Adjektiven, von denen 17 als typisch männlich (Bsp.: risikobereit, aggressiv), 24 als typisch weiblich (Bsp.: herzlich, romantisch, weichherzig) und 22 als neutral (Bsp.: wehleidig, nachtragend, schüchtern) klassifiziert werden. Die Einstufung der Adjektive basiert auf Voruntersuchungen (Gittler, 2003-2009). In der Untersuchung wurden die Testpersonen ersucht anzugeben, wie sie sich selbst „ganz privat“ einschätzen. Pro Bildschirmseite werden sieben Adjektive (Eigenschaften) vorgegeben, auf welcher der Teilnehmer für jede Eigenschaft auf einer sechsstufigen Antwortskala angeben kann, in welchem Ausmaß die Aussage auf den Probanden zutreffend ist (unveröffentlichte Version; Gittler, 2003-2009) (siehe Abbildung 3).

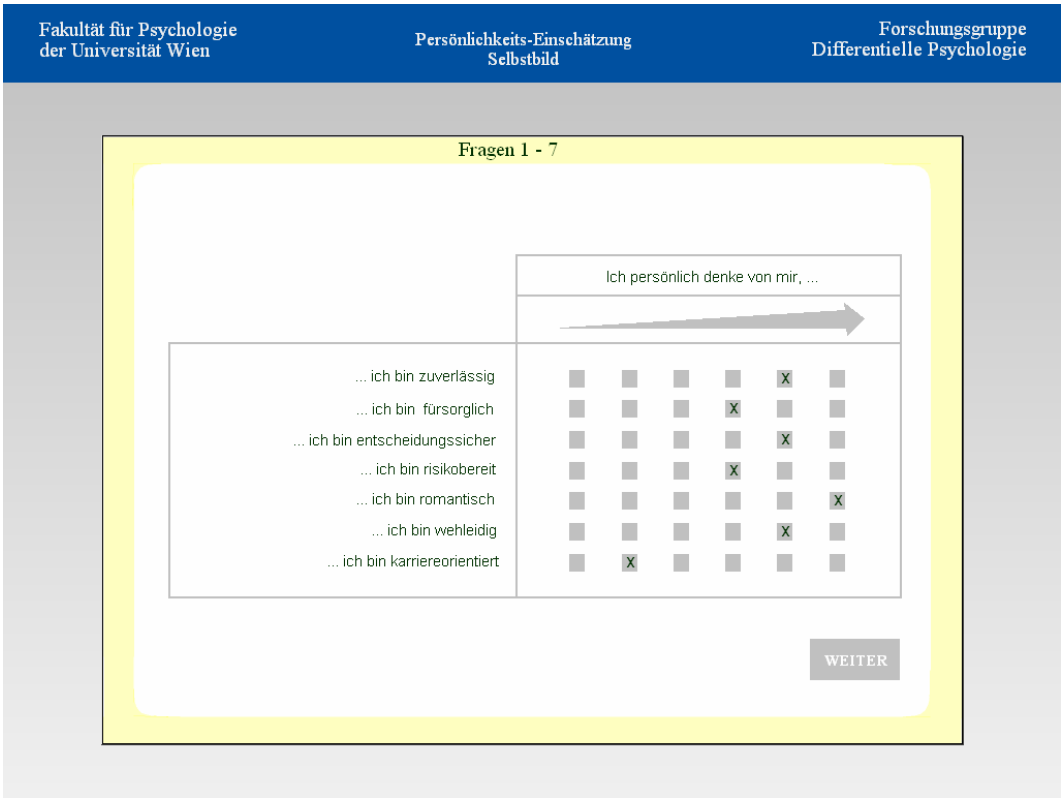


Abbildung 3: Beispielseite der verwendeten Adjektive sowie Antwortskala des Fragebogens Persönlichkeitseinschätzung "Selbstbild" (unveröffentliche Ausgabe, 2003-2009)

Die Auswertung basiert auf dem in Kapitel 4.2.1 beschriebenen zweidimensionalen Ansatz der psychologischen Geschlechtsorientierung. Es wird von einer Maskulinitäts- (M) und einer Femininitätsskala (F) ausgegangen, die unabhängig vom biologischen Geschlecht sind (vgl. Altstötter-Gleich, 1996; Alfermann, 2001). Personen mit hoher Ausprägung sowohl auf der Maskulinitäts- (M) als auch auf der Femininitätsskala (F) werden als „androgyn" (M+/F+) bezeichnet. Eine hohen Ausprägung auf einer Skala und eine niedrigen Ausprägung auf der anderen Skala wird als „maskulin" (M+/F-) bzw. „feminin" (M-/F+) bezeichnet. Personen mit niedrigen Werten auf der Maskulinitäts- (M) und Femininitätsskala (F) gelten als „undifferenziert" (M-/F-) (vgl. Altstötter-Gleich, 1996, S. 109).

9.2.3 Der Planzeichentest

Der Planzeichentest (siehe Anhang) ist ein von der Versuchsleiterin eigens konstruierter Papier-Bleistift-Test. Die Bearbeitung erfordert das Abrufen kognitiver Karten aus dem Gedächtnis. Die Versuchspersonen wurden gebeten, anzugeben, seit wie vielen Jahren sie in Wien leben, wobei Sie zwischen "einem Jahr", "ein bis fünf Jahre" und "mehr als

fünf Jahre“ wählen konnten. Weiters sollten die Versuchspersonen Ihre Ortskenntnisse in Wien anhand einer Schulnotenskala (“sehr gut“ bis “nicht genügend“) einschätzen. Es wurden Gebäude (Oper) Kirchen (Votivkirche) und Bahnhöfe (Westbahnhof) etc. angeführt, anhand deren die Probanden angeben mussten, ob diese Ihnen von der Lage her bekannt seien. Dabei konnte zwischen “ja“ bzw. “nein“ gewählt werden. Im Anschluss daran wurde dem Probanden ein Blatt vorgegeben, auf dem der “alte“ Südbahnhof, der Westbahnhof, das neue Allgemeines Krankenhaus (AKH), die Hauptuniversität und der Stadtpark zu sehen waren. Diese dienten als Orientierungspunkte, anhand derer die Versuchspersonen die folgenden Suchpunkte möglichst genau zu schätzen und einzuzeichnen hatten:

Suchpunkte:

- Stephansplatz
- Karlskirche
- Votivkirche
- Oper
- Urania
- Praterstern
- Stadthalle

In der vorliegenden Untersuchung wurde der Endlosschleifentest und der Fragebogen zur Persönlichkeitseinschätzung “Selbstbild“ am Computer vorgegeben. Der Planzeichentest wurde von den Probanden als Papier-Bleistift-Test bearbeitet.

10 Erhebungsdesign

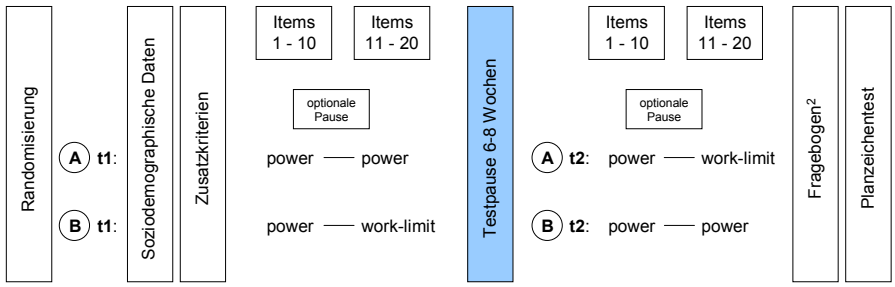


Abbildung 4: Schematischer Überblick über den Ablauf der Testvorgaben[2]

Bei der vorliegenden Untersuchung handelte es sich um eine Wiederholungstestung mit zwei Testzeitpunkten (t1, t2).

Die Versuchspersonen wurden randomisiert der Vorgabebedingung (Instruktion) und somit Gruppe A oder B zugewiesen. Zu beiden Testzeitpunkten bestand der Endlosschleifentest (EST) aus zwei Testhälften. Das erste Item wurde in der vorliegenden Untersuchung als “warm-up“ Item klassifiziert und ging somit in die nachfolgende Auswertung nicht mit ein. Die erste Testhälfte bestand aus Itemset 1 (Items 2-10) und die zweite Testhälfte setzte sich aus Itemset 2 (Items 11-20) zusammen. Die Items waren in beiden Gruppen identisch, es gab lediglich Unterschiede in der Instruktion, die vor der Bearbeitung der zweiten Testhälfte folgte (power oder work-limit Instruktion).

[2] Fragebogen zur Persönlichkeitseinschätzung “Selbstbild“

Gruppe A bearbeitete zum ersten Testzeitpunkt beide Testhälften unter der power Vorgabe (reine power Bedingung: po-po). Zum zweiten Testzeitpunkt wurde die erste Testhälfte unter der power Bedingung vorgegeben und die zweite Testhälfte erfolgte unter der work-limit Instruktion (po-wl).

Gruppe B bearbeitete zu t1 die erste Testhälfte unter der power Bedingung und die zweite Testhälfte unter der work-limit Vorgabe (po-wl). Zu t2 wurde die reine power Bedingung (po-po) vorgegeben. Jede Versuchsperson bearbeitete somit beide Vorgabebedingungen des EST (po-po und po-wl), aber zu unterschiedlichen Zeitpunkten (t1 oder t2).

10.1 Versuchsablauf

Zum ersten Testzeitpunkt wurde jeder Versuchsperson von der Versuchsleiterin ein Probandencode zugewiesen. Zunächst wurden in beiden Gruppen (A und B) die soziodemographischen Daten wie Alter, Geschlecht und höchst abgeschlossene Ausbildung erhoben. Weiters wurden Außenkriterien erfasst. Die Probanden wurden gebeten, ihre technische, mathematische und musikalische Begabung sowie Ihr räumliches Vorstellungsvermögen auf einer Analogskala von 0 (sehr schlecht) bis 100 (sehr gut) einzuschätzen.

Ebenfalls erhoben wurden die durchschnittlichen Leistungen der Probanden in den Fächern Mathematik, Physik, Chemie, Deutsch, Englisch und Latein. Die Probanden wurden gefragt, ob sie das jeweilige Unterrichtsfach in der Schule hatten. Bei Beantwortung der Frage mit “Ja“ wurden die Versuchspersonen ersucht, die Note welche die durchschnittliche Leistung in dem Schulfach am besten charakterisiert, auf einer Analogskala anzugeben. Der Proband konnte von 1 (sehr gut) bis 5 (nicht genügend) wählen. Bei Beantwortung der Frage mit “nein“ wurde die Frage übersprungen und das nächste Unterrichtsfach abgefragt.

Danach erfolgte der Endlosschleifentest (EST). Die Items der beiden Gruppen waren, wie bereits erwähnt, identisch und unterschieden sich lediglich in der Instruktion die vor der Bearbeitung der zweiten Testhälfte folgte (power oder work-limit Instruktion).

Zu Beginn der Erhebung erfolgten eine genaue Instruktion des EST und fünf Beispielaufgaben, um den Probanden die Möglichkeit zu geben, sich mit dem Testmaterial vertraut zu machen.
Anschließend begann die Testphase mit Vorgabe der ersten Testhälfte, bestehend aus Itemset 1 (Items 2-10). Die Instruktion lautete:

Für die folgenden Testaufgaben haben Sie ausreichend Zeit zur Verfügung. Sie sollen daher ruhig und konzentriert arbeiten.

Im Anschluss an die erste Testhälfte folgte eine Pause, in der je nach Bearbeiten der power oder work-limit Bedingung jeweils zwei unterschiedliche Arbeitsanweisungen folgten. In beiden Instruktionen wurde darauf hingewiesen, dass die Versuchsperson nun die erste Testhälfte geschafft hat und falls notwendig, sich etwas ausruhen könne. Bei der work-limit Instruktion erfolgte zusätzlich die Anweisung, in der zweiten Testhälfte nicht nur *so richtig wie möglich* zu arbeiten, sondern auch *so rasch wie möglich* zu arbeiten. Die Versuchspersonen der power Vorgabe erhielten diese Instruktion nicht, sondern bearbeiteten die Items weiter unter der power Instruktion.

Danach folgte die zweite Testhälfte mit Itemset 2 (Items 11-20). Nach der Bearbeitung der beiden Testhälften des EST war die Testung für den ersten Testzeitpunkt beendet, und es erfolgte eine sechs bis achtwöchige Testpause.

Zum zweiten Testzeitpunk (t2) erhielten die Probanden von der Versuchsleiterin denselben Code wie zu t1, um die Probanden zur richtigen Versuchsbedingung und somit Gruppe A oder B zuordnen zu können. Anschließend erfolgte der Endlosschleifentest, der, wie schon erwähnt, zu Testzeitpunkt 2 aus denselben zwei Itemsets wie zu Testzeitpunkt 1 bestand, und sich lediglich in der zweiten Testhälfte durch die Instruktion unterschied. Je nach Testgruppe (A oder B) wurden die Items des EST in der reinen power Bedingung (power-power) oder in der work-limit Bedingung (power-worklimit) vorgegeben.
Zu t2 erfolgte im Anschluss an die Bearbeitung des Endlosschleifentests der Fragebogen zur Persönlichkeitseinschätzung “Selbstbild“ sowie der Planzeichentest.

11 Datenerhebung und Stichprobe

Die Datenerhebung erfolgte im Zeitraum von Dezember 2009 bis Oktober 2010. Die Erhebungen wurden als Einzeltestung an einem ruhigen Ort durchgeführt.

11.1 Stichprobe

Die Stichprobe besteht aus insgesamt 100 Personen und weist ein Geschlechtsverhältnis von 53 Männern (durchschnittliches Alter *M* = 36.40, *SD* = 10.85) und 47 Frauen (*M* = 34.91, *SD* = 9.85) auf.
Der Altersbereich der Versuchspersonen erstreckt sich von 19 bis 61 Jahren (siehe Abbildung 5) mit einem Mittelwert von 35.70 (*SD* = 10. 37).

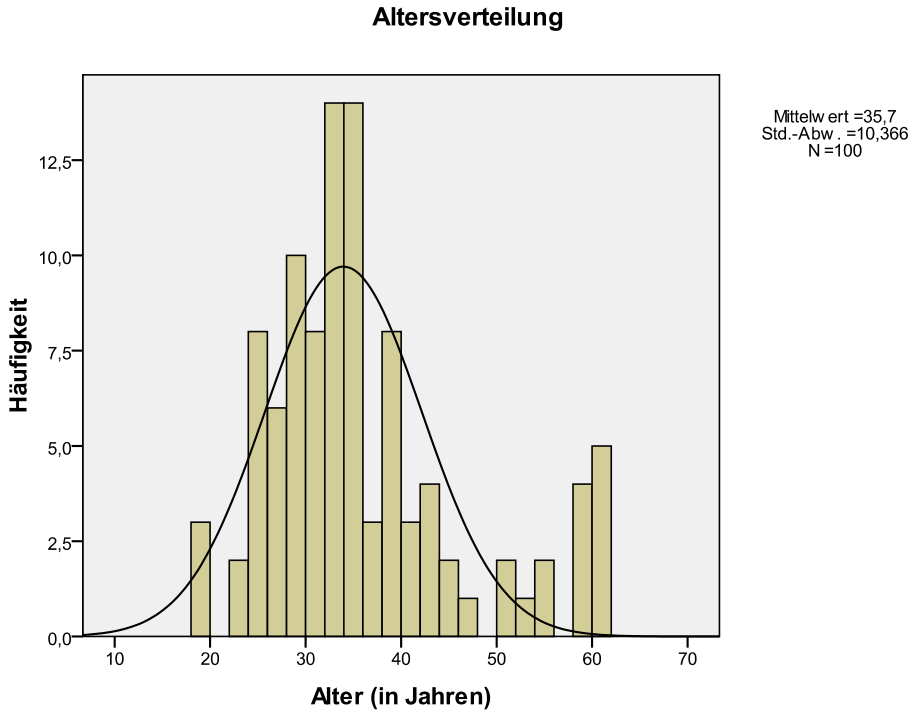


Abbildung 5: Altersverteilung der Stichprobe (N= 100)

Innerhalb der Gesamtstichprobe gaben 37 % der Versuchspersonen Matura bzw. Abitur als höchst abgeschlossene Ausbildung an. 29 % der Probanden wiesen einen

Universitätsabschluss auf. Nur 1% der Stichprobe gab an, die neunte Schulstufe als höchst abgeschlossene Ausbildung zu haben (siehe Abbildung 6).

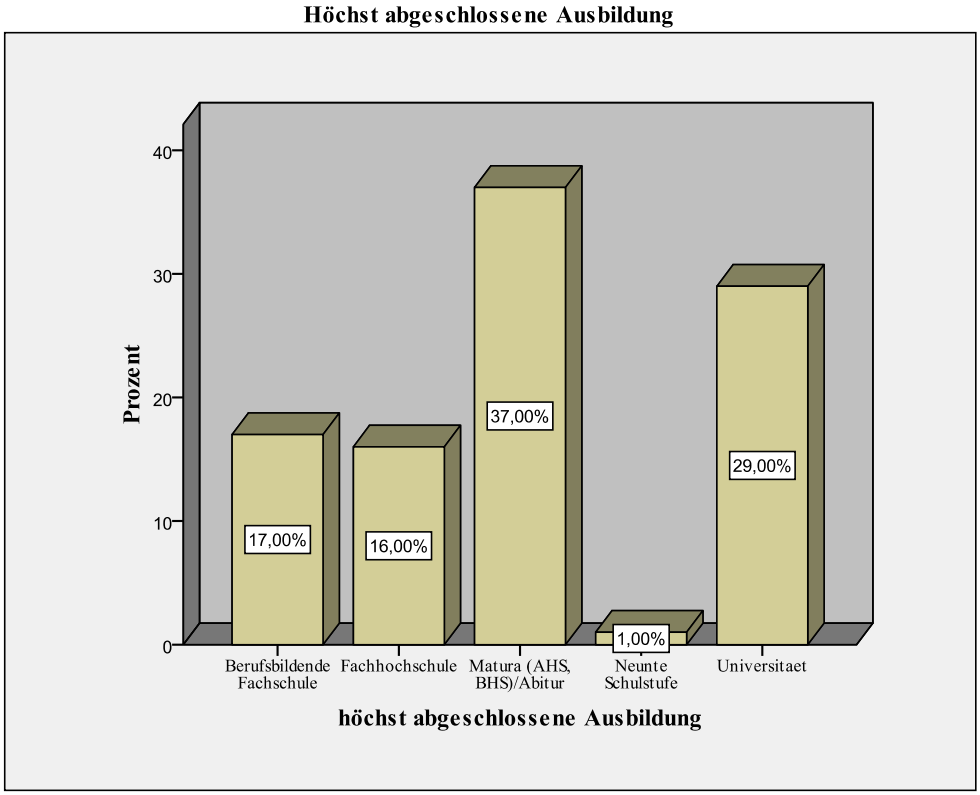


Abbildung 6: Höchst abgeschlossene Ausbildung

Von den 100 Personen waren 51 Personen der Gruppe A (t1: power - power, t2: power - work-limit) zugeordnet, davon waren 28 Männer und 23 Frauen. Gruppe B (t1: power - work-limit, t2: power - power) waren 49 Personen, von denen 25 Männer und 24 Frauen waren.

12 Modellprüfung des EST auf Rasch-Homogenität

In der vorliegenden Arbeit wurde überprüft, ob der EST unter der work-limit Bedingung sowie unter der reinen power Bedingung eindimensional im Sinne von Rasch ist und es wurde zudem der Gesamttest (EST) auf Rasch-Homogenität überprüft. Um die Geltung des Rasch Modells zu überprüfen, wurde mit der Software LpcM-WIn 1.0 (Fischer & Ponocny-Seliger, 1998) der Likelihood-Quotienten-Test (LQT) nach Andersen (1973) und Martin-Löf (1973) berechnet. Der Likelihood-Quotienten-Test nach Andersen (1973) ist der bekannteste Rasch-Modelltest zur Prüfung auf Personenhomogenität. Es wird angenommen, dass alle getesteten Personen den Test aufgrund derselben Fähigkeit bzw. Eigenschaft bearbeiten. Das Prinzip des Likelihood-Quotienten-Test besteht darin, die Itemparameter in verschiedenen Subgruppen der Personenstichprobe zu schätzen und zu prüfen, ob sie sich zwischen den Gruppen unterscheiden (Rost, 2004, S. 347).
Der Martin-Löf-Test ist ein Signifikanztest zur Prüfung auf Itemhomogenität. Hierfür werden die Items in Subgruppen geteilt (Rost, 2004, S. 351).

Das Signifikanzniveau für die Modellkontrolle wurde auf α = .01 festgesetzt. Die Gültigkeit des Rasch Modells kann angenommen werden, sobald das Ergebnis des Modelltests nicht signifikant ausfällt, d.h. der empirische χ^2 - Wert kleiner als der kritische χ^2 - Wert ist.

Als internes Teilungskriterium wurden der Mittelwert (mean) und der Median des EST-Rohscores (niedrige vs. hohe Leistung) herangezogen. Als externe Kriterien dienten das Geschlecht (männlich/weiblich), das Alter (≤ 33 / ≥34 Jahre), die Selbsteinschätzung der Raumvorstellungsbegabung (gut/schlecht) sowie die Testbedingung (Gruppe A: t1: po-po; t2: po-wl/Gruppe B:t1: po-wl; t2: po-po). Das Item 1 fließt, wie bereits erwähnt, in die Berechnung nicht mit ein, da es als “warm-up“ Item klassifiziert wurde.

Die Nachfolgenden Modelle werden als vollständige Rasch-Modelle behandelt.

12.1 Rasch-Homogenität unter der reinen power Bedingung

Zuerst erfolgt, mit Hilfe des LQT nach Andersen (1973), die Modellprüfung der Items 2-10 und der Items 11-20, die der Gruppe A zum ersten Testzeitpunkt und der Gruppe B zum zweiten Testzeitpunkt als reine power Bedingung vorgegeben wurde (siehe Tabelle 1).

Tabelle 1: Veranschaulichung der Modellprüfung der reinen power Bedingung, N = 100, k = 19

		Items 2-10	Items 11-20
t1	Gruppe B $n^{3}_{=49}$	power	power
t2	Gruppe A $_{n=51}$	power	power

Tabelle 2: Ergebnisse der Likelihood-Quaotienten-Tests nach Andersen (1973) der Daten unter der reinen power Bedingung, N[4] = 100, k[5] = 19, und *df*[6] =18, α = .01; angegeben werden der empirischen $\chi 2$ - Wert, der kritische $\chi 2$ - Wert, sowie die Signifikanz (*p*)

Teilungskriterien	Teilgruppe 1	n_1	Teilgruppe 2	n_2	$\chi^2_{empirisch}$	$\chi^2_{kritisch}$	Sign. (*p*)
Rohscore (mean)	niedrig	61	hoch	39	19.2347	34.8309	.378
Rohscore (median)	niedrig	53	hoch	47	22.8068		.198
Geschlecht	männlich	53	weiblich	47	23.6229		.168
Alter	≤ 33	51	≥ 34	49	26.4231		.090
Raumvorstellung	niedrig	52	hoch	48	22.2027		.233
Testbedingung	A	51	B	49	10.6117		.910

Keines der Teilungskriterien fiel signifikant aus, somit kann unter der reinen power Bedingung von Homogenität der Daten ausgegangen werden (siehe Tabelle 2).

12.2 Rasch-Homogenität unter der work-limit Bedingung

Es wurde mittels Likelihood-Quotien-Test nach Andersen (1973) untersucht, ob die Daten der 51 Personen der Gruppe A, die zum zweiten Testzeitpunkt in der zweiten Testhälfte die work-limit Bedingung bearbeiteten und die Daten der 49 Personen der

[3] n = Anzahl der Personen

[4] N = Gesamtstichprobe

[5] k = Anzahl der Items

[6] *df* = Freiheitsgrade

Gruppe B, die zum ersten Testzeitpunkt in der zweiten Testhälfte die work-limit Bedingung bearbeiteten, im Sinne von Rasch homogen sind.(siehe Tabelle 3)

Tabelle 3: Veranschaulichung der Modellprüfung unter der work-limit Bedingung. für die Daten der Gruppe B, die die work-limit Bedingung zu t1 bearbeitete und für die Gruppe A, die die work-limit Bedingung zu t2 bearbeitete. N=100, k=19

		Items 2-10	Items 11-20
t1	Gruppe B $_{n=49}$	power	work-limit
t2	Gruppe A $_{n=51}$	power	work-limit

Tabelle 4: Ergebnisse der Likelihood-Quaotienten-Tests nach Andersen (1973) der Daten unter der work-limit Bedingung, N = 100, k = 19, und *df* = 18, α = .01; angegeben werden der empirischen $\chi 2$ - Wert, der kritische $\chi 2$ - Wert, sowie die Signifikanz (*p*)

Teilungskriterien	Teilgruppe 1	n_1	Teilgruppe 2	n_2	$\chi^2_{empirisch}$	$\chi^2_{kritisch}$	Sign. (*p*)
Rohscore (mean)	niedrig	64	hoch	36	14.7079	34.8309	.682
Rohscore (median)	niedrig	50	hoch	50	16.3009		.572
Geschlecht	männlich	53	weiblich	47	11.5774		.868
Alter	≤ 33	51	≥ 34	49	20.7044		.295
Raumvorstellung	niedrig	52	hoch	48	24.3303		.145
Testbedingung	A	51	B	49	49.7572		.940

Wie Tabelle 4 zu entnehmen ist, liegen die empirischen χ^2 - Werte der internen Kriterien und die empirischen χ^2 - Werte der externen Kriterien unterhalb des kritischen χ^2 - Werts. Die nicht signifikanten Ergebnisse lassen auf Rasch-Homogenität des Modells unter der work-limit Bedingung schließen.

Ergänzend wurden das erste Itemset, das unter der power Bedingung (power Items) vorgegeben wurde, und das zweite Itemset, welches unter der work-limit Instruktion (work-limit Items) bearbeitet wurde, miteinander verglichen. Hierfür wurde mit Hilfe des Martin-Löf-Tests (1973) untersucht, ob die beiden Subskalen power und work-limit dieselbe Fähigkeitsdimension messen. Die Items wurden in zwei Subgruppen geteilt.

Die erste Gruppe setzt sich aus den neun power-Items 2 bis10 zusammen.
Die zweite Gruppe besteht aus den zehn work-limit-Items 11 bis 20.

Die Berechnung des Martin-Löf-Tests (1973) ergab ein nicht signifikantes Ergebnis, mit einem empirischen χ^2 - Wert = 69.3144, der kleiner als der kritische χ^2 - Wert (α =1%, *df*=89) = 122.9555 ausfällt.

Die Homogenität der beiden Subskalen power und work-limit kann als gegeben betrachtet werden.

12.3 Rasch-Homogenität des Gesamttests

12.3.1 Erste Modelprüfung des Gesamttests nach Andersen (1973)

Es wurde mit Hilfe des LQT nach Andersen (1973) die Geltung des Rasch-Modells für den Gesamttest überprüft. Um die Rasch-Homogenität des Gesamttests, der sowohl aus zwei Testzeitpunkten als auch zwei Gruppen – Gruppe A (t1: power - power; t2: power - worklimit) und Gruppe B (t1: power - worklimit; t2: power - power) – besteht, zu überprüfen, wurde das erste Itemset von beiden Testzeitpunkten für beide Gruppen zusammengefasst und erhielt die Itemnummerierung 2-10. Das zweite Itemset erhielt nach Zusammenfassung die Itemnummerierung 11-20 (siehe Tabelle 5).

Tabelle 5: Veranschaulichung des Modelltest, mit Itemset 1 (2-10) und Itemset 2 (11-20)

		Items 2-10	Items 11-20
t1	Gruppe A $_{n=51}$	power	power
	Gruppe B $_{n=49}$	power	work-limit
t2	Gruppe A $_{n=51}$	power	work-limit
	Gruppe B $_{n=49}$	power	power

Tabelle 6: Ergebnisse der Likelihood- Quptienten-Tests nach Andersen (1973), N = 200, k = 19, und *df* = 18, α = .01; angegeben werden der empirischen $\chi 2$ - Wert, der kritische $\chi 2$ - Wert sowie die Signifikanz (*p*)

Teilungskriterien	Teilgruppe 1	n_1	Teilgruppe 2	n_2	$\chi^2_{empirisch}$	$\chi^2_{kritisch}$	Sign. (*p*)
Rohscore (mean)	niedrig	125	hoch	75	19.2676	34.8309	.376
Rohscore (median)	niedrig	103	hoch	97	26.8725		.081
Geschlecht	männlich	106	weiblich	94	21.8631		.238
Alter	≤ 33	102	≥ 34	98	26.6759		.085
Raumvorstellung	niedrig	104	hoch	96	31.0397		.028
Testbedingung	A	102	B	98	10.0508		.930

Die empirischen χ^2 - Werte aus dieser ersten Modellprüfung des Gesamttests, fällt sowohl für die internen Kriterien (mean, median) als auch bei den externen Kriterien (Geschlecht, Alter, Raumvorstellungsbegabung und Testbedingung) kleiner als der kritische χ^2 - Wert aus. Aufgrund der nicht signifikanten Ergebnisse der Teilungskriterien kann von der Geltung des Rasch-Modells ausgegangen werden (siehe Tabelle 6).

Zusätzlich wurden mit Hilfe des Martin-Löf-Tests (1973) die beiden Itemsets einander gegenüber gestellt und untersucht, ob das erste Itemset, dieselbe Fähigkeitsdimension wie das zweite Itemset erfassen.
Hierfür wurden die Items in zwei Subgruppen geteilt.

Die erste Gruppe setzt sich aus den Items 2 bis10 zusammen.
Die zweite Gruppe besteht aus den Items 11 bis 20.

Die Berechnung des Martin-Löf-Tests (1973) ergab ein nicht signifikantes Ergebnis, mit einem empirischen χ^2 - Wert = 118.1642, der kleiner als der kritische χ^2 - Wert (α =1%, *df*=89) = 122.9555 ausfällt.

Daher ist anzunehmen, dass das das erste Itemset und das zweite Itemset dieselbe Fähigkeitsdimension messen.

12.3.2 Zweite Modellprüfung des Gesamttests nach Andersen (1973)

Um diese zweite Modelltestung durchführen zu können, erhielt das erste Itemset (Items 2-10) vom zweiten Testzeitpunkt die Itemnummerierung 22-30 und das zweiten Itemset (Items 11-20) vom zweiten Testzeitpunkt die Itemnummerierung 31-40 (siehe Tabelle 7).

Tabelle 7: Veranschaulichung der Modelltestung, die Daten des zweiten Testzeitpunktes der beiden Gruppen werden neben die Daten zu t1 gestellt. N = 100, k = 38

	t1		t2	
	Items 2-10	**Items 11-20**	**Items 22-30**	**Items 31-40**
Gruppe A $_{n=51}$	power	power	power	work-limit
Gruppe B $_{n=49}$	power	work-limit	power	power

Tabelle 8: Modelltestung, Ergebnisse der Likelihood- Quotienten-Tests nach Andersen (1973), N= 100, k= 38, *df* =37, und α = .01, angegeben werden der empirischen χ2 - Wert der kritische χ2 - Wert, sowie die Signifikanz (*p*)

Teilungskriterien	Teilgruppe 1	n_1	Teilgruppe 2	n_2	$\chi^2_{empirisch}$	$\chi^2_{kritisch}$	Sign. (*p*)
Rohscore (mean)	niedrig	63	hoch	37	40.4513	59.9122	.320
Rohscore (median)	niedrig	51	hoch	49	29.7902		.040
Geschlecht	männlich	53	weiblich	47	38.6107		.397
Alter	≤33	51	≥ 34	49	41.9144		.266
Raumvorstellung	niedrig	52	hoch	48	42.8857		.233
Testbedingung	A	51	B	49	28.1157		.853

Der Vergleich der internen Kriterien und der externen Kriterien mit dem kritischen χ^2-Wert von 59.9122 zeigt, dass bei dieser Modellprüfung, die empirischen χ^2 - Werte für die Teilungskriterien unterhalb des kritischen χ^2 - Werts liegen. Die Ergebnisse des internen Kriteriums und der externen Kriterien fallen nicht signifikant aus (siehe Tabelle 8).

12.4 Unvollständiges Rasch-Modell

Nachfolgend wurde ein unvollständiges Raschmodell berechnet. Das erste Itemset von Gruppe A und Gruppe B wurde zusammengefasst und erhielt die Itemnummerierung 2-10. Das zweite Itemset von Gruppe A und B, welches zum ersten Testzeitpunkt vorgegeben wurde, erhielt die Itemnummerierung 11-20. Das zweite Itemset von Gruppe A und B, das zum zweiten Testzeitpunkt vorgegeben wurde, erhielt die Itemnummerierung 21-30 (siehe Tabelle 9).

Tabelle 9: Veranschaulichung der Modellprüfung des unvollständigen Rasch Modells, N = 200, k = 29

		Items 2-10	Items 11-20	Items 21-30
t1	Gruppe A $_{n=51}$	power	power	
	Gruppe B $_{n=49}$	power		work-limit
t2	Gruppe A $_{n=51}$	power		work-limit
	Gruppe B $_{n=49}$	power	power	

Tabelle 10: Ergebnisse der Likelihood-Quaotienten-Tests nach Andersen (1973) der Daten des unvollständigen Rasch Modell, N = 200, k = 29, und *df* =28, α = .01; angegeben werden der empirischen $\chi 2$ - Wert, der kritische$\chi 2$ - Wert, sowie die Signifikanz (*p*)

Teilungskriterien	Teilgruppe 1	n_1	Teilgruppe 2	n_2	$\chi^2_{empirisch}$	$\chi^2_{kritisch}$	Sign. (*p*)
Rohscore (mean)	niedrig	117	hoch	88	26.0103	48.3002	.572
Rohscore (median)	niedrig	109	hoch	91	21.9516		.784
Geschlecht	männlich	106	weiblich	94	29.6293		.381
Alter	≤33	102	≥ 34	98	40.6885		.057
Raumvorstellung	niedrig	104	hoch	96	41.2706		.051
Testbedingung	A	102	B	98	14.0237		.987

Die Modellkontrolle des unvollständigen Rasch-Modells ergab für alle Teilungskriterien ein nicht signifikantes Ergebnis (siehe Tabelle 10), womit von der Homogenität der Daten ausgegangen werden kann.

Darüber hinaus wurden die Daten des unvollständigen Rasch-Modell (N= 200, k= 29 und *df*=28) und die Daten der ersten Modellprüfung des Gesamttests nach Andersen, welches als vollständiges Modell betrachtet werden kann (N= 200, k= 19 und *df*= 18), miteinander verglichen.

Es wurde die Total Log-Likelihood (Total Log-Likelihood $_{\text{vollständiges Rasch-Modell:}}$ -1599.239752; Total Log-Likelihood $_{\text{unvollständigen Rasch-Modell:}}$ -1592.120551) für den Vergleich der beiden Modelle herangezogen, um händisch einen Likelihood-Quotienten-Test zu berechnen. Der empirische χ^{2-} - Wert fällt kleiner als der kritische χ^{2-} - Wert ($\chi^2_{\text{empirisch}}$ = 14.24 < χ^2_{kritisch} = 23.21) aus, somit kann von einem nicht signifikanten Ergebnis ausgegangen werden.

Zusammenfassend kann festgehalten werden, dass keiner der Modelltests ein signifikantes Ergebnis liefert. Aufgrund der durchwegs nicht signifikanten Ergebnisse der Modellprüfungen ist es möglich, anschließende inferenzstatistische Auswertungen auf Rohwertniveau zu berechnen.

13 Hypothesenprüfung

Nachfolgende inferenzstatistische Berechnungen wurden mit Hilfe des Programms SPSS 17.0 durchgeführt. Das Signifikanzniveau wurde auf $\alpha = .05$ festgesetzt.

13.1 Auswirkungen der work-limit Instruktion auf die Leistung

Es soll nun der Frage nachgegangen werden, ob aufgrund der work-limit Instruktion, welche in der zweiten Testhälfte vorgegeben wurde, zusätzliche Informationen bezüglich der Leistung der Probanden gewonnen werden kann.

13.1.1 Leistungsunterschiede zwischen den Itemsets

Gruppe A und B bestehen, wie bereits erwähnt, zu beiden Zeitpunkten aus Itemset 1 (Items 2-10) und Itemset 2 (11-20) und unterscheiden sich jeweils nur in der Vorgabebedingung (power oder work-limit) des zweiten Itemsets (siehe Abbildung 11).

Tabelle 11: Versuchsbedingung der Gruppe A und B zu t1 und t2

	t1		t2	
	Items 2-10	**Items 11-20**	**Items 1-10**	**Items 11-20**
Gruppe A $_{n=51}$	power	power	power	work-limit
Gruppe B $_{n=49}$	power	work-limit	power	power

Aufgrund der unterschiedliche Anzahl an Items (k=9 und k=10) in den Itemsets werden die Rohwertleistungen in Prozentwerte transformiert, um so die Itemsets vergleichen zu können.

Aufgrund der signifikanten Ergebnisse des Kolmogrov–Smirnov-Tests (K-S; siehe Tabelle 12) kann die Normalverteilung für die Daten nicht angenommen werden.

Tabelle 12: Mittelwert *(M)*, Standardabweichung (*SD*) der in Prozentwerte transformierten Rohwerte der Itemsets. Ergebnisse des Kolmogorov – Smirnov – Tests *(p*- Werte) zur Prüfung der Normalverteilung der Itemsets, α = .05

	Gesamt	n	*M*	*SD*	K-S
t1	Itemset 1 (k = 9)	100	37.33	25.83	< .001
	Itemset 2 (k = 10)		32.80	21.47	< .001
	Gruppe A				
	Itemset 1 (power)	51	34.86	28.46	< .001
	Itemset 2 (power)		34.12	22.27	.013
	Gruppe B				
	Itemset 1 (power)	49	39.91	22.78	< .001
	Itemset 2 (work-limit)		31.43	20.72	< .001
t2	**Gesamt**				
	Itemset 1 (k = 9)	100	43.11	25.77	.001
	Itemset 2 (k = 10)		38.70	23.47	< .001
	Gruppe A				
	Itemset 1 (power)	51	44.66	26.71	.004
	Itemset 2 (work-limit)		39.02	23.85	.013
	Gruppe B				
	Itemset 1 (power)	49	41.50	24.93	.003
	Itemset 2 (power)		38.37	23.30	.033

Die Varianzanalyse kann jedoch als verhältnismäßig robust gegenüber Verletzungen der Voraussetzungen der Normalverteilung angesehen werden (Backhaus, Erichson, Plinke & Weiber, 2003), weshalb sie nachfolgend für die Berechnung herangezogen wurde. Um zu untersuchen, ob es Unterschiede in der Leistung der Probanden zwischen Itemset 1 und Itemset 2 gibt, wurde eine gemischte abhängige Varianzanalyse, getrennt für den ersten Testzeitpunkt und zweiten Testzeitpunkt gerechnet.

Für den ersten Testzeitpunkt fällt der Levene-Test auf Varianzhomogenität für Itemset 1 (F (1,98) = 3.873, p = .053) und Itemset 2 (F (1,98) = .437, p = .437) nicht signifikant aus, womit von Homogenität der Varianzen ausgegangen werden kann.

Tabelle 13: Tafel der gemischten abhängigen Varianzanalyse für den ersten Testzeitpunkt für die abhängige Variable Testleistung und der unabhängigen Variablen Itemset und Testbedingung, α = .05 (N = 100)

Quelle	Quadratsumme	Mittel der Quadrate	*F*(1,98)	Signifikanz (*p*)
Itemset	1062.518	1062.518	4.224	.043
Itemset * Testbedingung	748.543	748.543	2.976	.088
Fehler	24651.062	251.541		

Die Berechnung der gemischten abhängigen Varianzanalyse zeigt bei nicht signifikanter Wechselwirkung (p = .088) einen signifikanten Unterschied zwischen den beiden Itemsets (p = .043). Wie anhand der Mittelwerte der Tabelle 12 und der Abbildung 7 zu sehen ist, erzielten die Versuchspersonen in Itemset 1 bessere Ergebnisse als in Itemset 2: M = 37.33 (Itemset 1) vs. M = 32.80 (Itemset 2). Zwischen den beiden Versuchsgruppen kann kein signifikanter Unterschied (p = .779) angenommen werden.

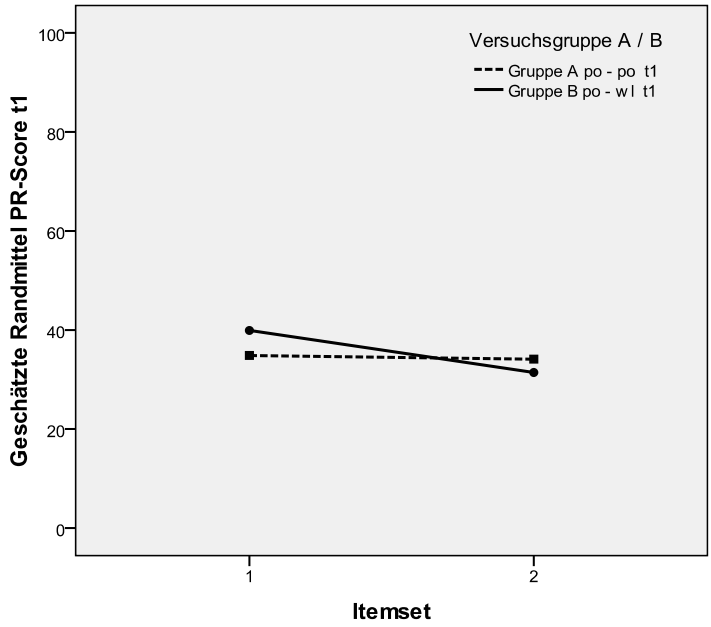


Abbildung 7: Leistungen der in Prozentwerte transformierte Rohscores der Itemsets beider Versuchsgruppen zu Zeitpunkt 1.

Zum zweiten Testzeitpunkt kann sowohl für Itemset 1 (F(1,98) = 0.522, p = .472,) als auch für Itemset 2 (F(1,98) = .010, p = .920) die Homogenität der Varianzen angenommen werden.

Tabelle 14: Tafel der gemischten abhängigen Varianzanalyse für den zweiten Testzeitpunkt für die abhängige Variable Testleistung und den unabhängigen Variablen Itemset und Testbedingung, α = .05 (N = 100)

Quelle	Quadratsumme	Mittel der Quadrate	*F* (1,98)	Signifikanz
Itemset	96.455	96.455	3.391	.069
Itemset * Testbedingung	78.936	78.936	0.278	.599
Fehler	27789.527	283.567		

Wie anhand der Ergebnisse der gemischten abhängigen Varianzanalyse (siehe Tabelle 14) zu erkennen ist, kann keine signifikante Wechselwirkung *(p* = .599) und kein signifikanter Leistungsunterschied zwischen Itemsets 1 und 2 (*p* = 599) angenommen werden. Zwischen den Leistungen der Versuchsgruppen kann ebenfalls kein signifikanter Niveaunterschied (*p* = .661) beobachtet werden.

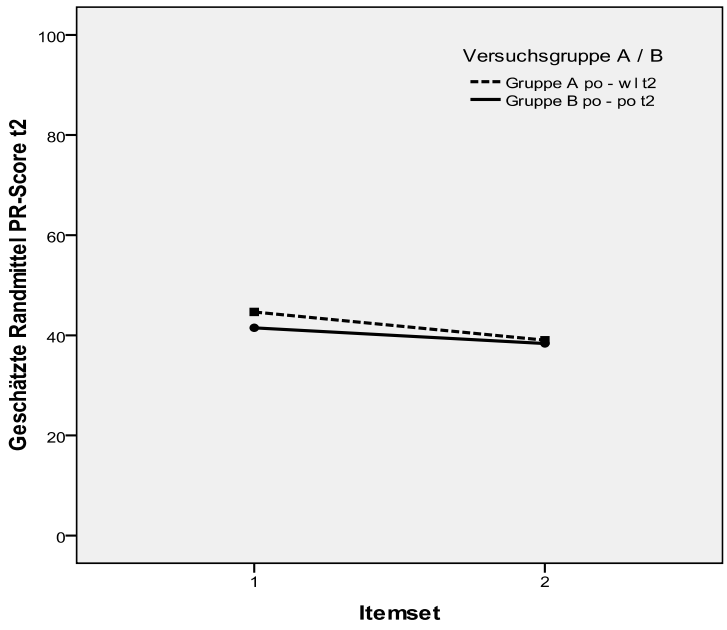


Abbildung 8: Leistungen der in PR-Werte transformierten Rohscores in den Itemsets beider Versuchsgruppen zu Zeitpunkt 2

Darüber hinaus wird mittels gemischter abhängiger Varianzanalyse geprüft, ob es einen Unterschied in der Testleistung zwischen den beiden Itemsets über die Zeit hinweg gibt.

Tabelle 15: Ergebnisse des Levene – Tests auf Varianzhomogenität, α = .05

	Itemset	*F* (1,98)	Signifikanz *(p)*
t1	Itemset 1	3.837	.053
	Itemset 2	0.609	.437
t2	Itemset 1	0.522	.472
	Itemset 2	0.010	.920

Der Levene-Test auf Varianzhomogenität fällt mit ≥ .053 jeweils nicht signifikant aus, womit die Varianzen als homogen angenommen werden können (siehe Tabelle 15).

Tabelle 16: Tafel der gemischten abhängigen Varianzanalyse für beide Testzeitpunke für die abhängige Variable Testleistung und der unabhängigen Variable Itemset, Testbedingung und Zeitpunkte, α = .05 (N = 100)

Quelle	Quadratsumme vom Typ III	Mittel der Quadrate	*F* (1,98)	Signifikanz*(p)*
Itemset	2022.71	2022.71	5.695	.019
Itemset * Testbedingung	170.661	170.661	.480	.490
Fehler(Itemset)	34809.818	355.202		
Zeitpunkt	3371.925	3371.925	11.955	.001
Zeitpunkt * Testbedingung	238.592	238.592	.846	.360
Fehler(Zeitpunkt)	27640.109	282.042		
Itemset * Zeitpunkt	1.262	1.262	.007	.933
Itemset * Zeitpunkt * Testbedingung	656.818	656.818	3.651	.059
Fehler (Itemset*Zeitpunkt)	17630.772	179.906		

Das Ergebnis der gemischten abhängigen Varianzanalyse (siehe Tabelle 16) zeigt bei jeweils nicht signifikanten Wechselwirkungen einen signifikanten Unterschied in den Leistungen der Probanden in Abhängigkeit vom Itemset (p = .019) sowie ein signifikantes Ergebnis in Abhängigkeit von den Zeitpunkte (p = .001).

Die Versuchspersonen schneiden in Itemset 1 und Itemset 2 zum zweiten Testzeitpunkt besser ab ($M_{\text{Itemset 1}}$= 43.11, $M_{\text{Itemset 2}}$ = 38.70) als zum ersten Testzeitpunkt ($M_{\text{Itemset 1}}$ = 37.33, $M_{\text{Itemset 2}}$ = 32.80).

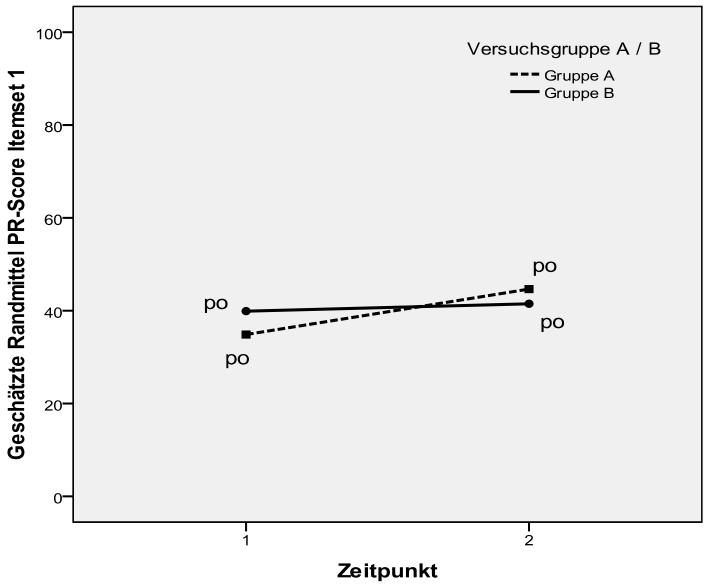


Abbildung 9: Leistungen der in Prozentwerte transformierten Rohscores beider Versuchsgruppen in Itemset 1 zu beiden Testzeitpunkten

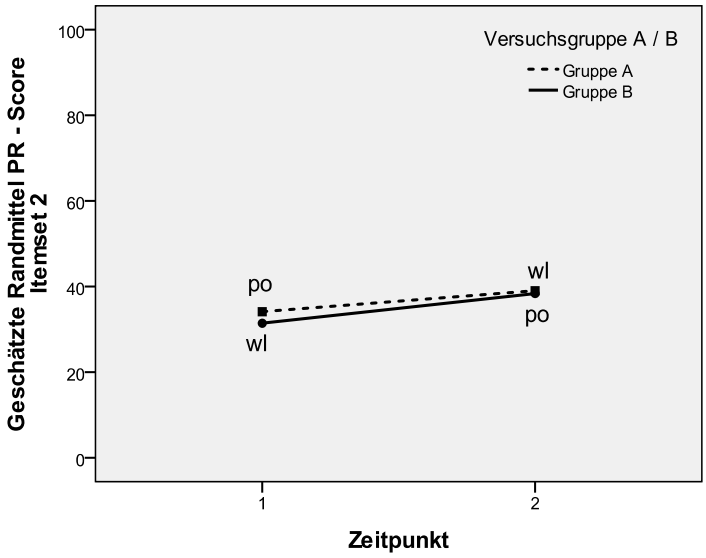


Abbildung 10: Leistungen der in Prozentwerte transformierten Rohscores beider Versuchsgruppen in Itemset 2 zu beiden Testzeitpunkten

Zur Prüfung der Frage, ob es Leistungsunterschiede innerhalb der Gruppe (A sowie B) und den beiden Itemsets in Abhängigkeit der beiden Messzeitpunkte gibt, wurden *t*-Tests für abhängige Stichproben berechnet.

Tabelle 17: Kolmogorov-Smirnov-Test (K-S) der Itemsetdifferenzen (*p*-Werte) zur Prüfung der Normalverteilung sowie Ergebnisse des t-Tests für abhängige Stichprobe mit t-Werten (t), Freiheitsgraden *(df)* und *p*-Werten (Signifikanz, 2-seitig).

	Gruppe A $_{n = 51}$	**K-S**	***t***	***df***	**Sign. (*p*)**
t1 vs t2	**Itemset 1 (po vs po)**	.267	-3.530	50	.001
	Itemset 2 (po vs wl)	.438	-1.557		.126
	Gruppe B $_{n = 49}$				
t1 vs t2	**Itemset 1 (po vs po)**	.159	-0.467	48	.643
	Itemset 2 (wl vs po)	.196	-2.478		.017

Die Normalverteilung der Messwertdifferenzen je Versuchsgruppe und je Itemset kann angenommen werden ($p \geq .159$). Die Ergebnisse des *t* -Tests für abhängige Stichprobe zeigen, dass in Gruppe A in Itemset 1 ein signifikanter Unterschied ($p = .001$) in der erzielten Leistung zwischen t1 und t2 anzunehmen ist. Gruppe A schneidet in Itemset 1 zum zweiten Testzeitpunkt besser ab ($M_{t2} = 44.66$) als im ersten Testzeitpunkt ($M_{t1} = 34.86$). In Itemset 2 der Gruppe A kann kein signifikanter Unterschied ($p = .126$) in den Testergebnissen der Probanden in Abhängigkeit der Testzeitpunkte beobachtet werden.

In Gruppe B gibt es keinen signifikanten Unterschied ($p = .643$) in der Leistung der Versuchspersonen in Itemset 1 zwischen t1 und t2. In Itemset 2 ist anzunehmen, dass es signifikante Unterschiede ($p = .017$) in der erzielten Leistung der Probanden zwischen den Testzeitpunkten gibt. Gruppe B erzielt in Itemset 2 zum zweiten Testzeitpunkt ($M_{t2} = 38.37$) höhere Leistungen als zum ersten Testzeitpunkt ($M_{t1} = 31.43$).

Zusätzlich wird geprüft, ob sich die Testleistung der Gruppe A von der Testleistung der Gruppe B im ersten Itemset zu t1 signifikant voneinander unterscheidet (siehe Tabelle 18), danach werden die Leistungsunterschiede zwischen Gruppen A und Gruppe B im ersten Itemset für t2 untersucht (siehe Tabelle 19). Da die Normalverteilung der Daten nicht angenommen werden kann (s.o), werden Unterschiede nachfolgend mittels nichtparametrischen U–Tests nach Mann und Whitney geprüft.

Tabelle 18: Vergleich der Leistungen in Itemset 1 der Gruppen A und B zu t1.

	t1	
	Itemset 1	Itemset 2
Gruppe A	power	power
Gruppe B	power	work-limit

Tabelle 19: Vergleich der Leistungen in Itemset 2 der Gruppen A und B zu t2.

	t2	
	Itemset 1	Itemset 2
Gruppe A	power	power
Gruppe B	power	work-limit

Die Berechnung ergibt, dass sich die Leistung der Gruppe A im ersten Itemset sowohl für t1 als auch für t2 nicht signifikant von der Leistung der Gruppe B unterscheidet (t1: z =.-1.567, p = .117; t2: z = -0.496, p = .620).

Tabelle 20: Vergleich der Leistungen in Itemset 2 der Gruppen A und B zu t1.

	t1	
	Itemset 1	Itemset 2
Gruppe A	power	power
Gruppe B	power	work-limit

Tabelle 21: Vergleich der Leistungen in Itemset 2 der Gruppen A und B zu t2.

	t2	
	Itemset 1	Itemset 2
Gruppe A	power	work-limit
Gruppe B	power	power

Darüber hinaus werden Leistungsunterschiede zwischen den Gruppen im zweiten Itemset (power vs. work-limit) für beide Zeitpunkte untersucht. (siehe Tabelle 20 und Tabelle 21). Zu t1 unterscheiden sich die Leistung der Gruppe A, welche das zweite Itemset unter der power Bedingung bearbeitete nicht signifikant (t1: z = -0.667, p = .505) von der Leistung der Gruppe B, die das zweite Itemset unter der work-limit Bedingung bearbeitete. Auch zu t2 kann kein signifikanter Unterschied (t2: z = -0.021, p = .983) zwischen den beiden Gruppen in Abhängigkeit der zu bearbeiteten Vorgabe im zweiten Itemset angenommen werden.

Tabelle 22: Vergleich der Leistung in Itemset 1 der Gruppe A mit der Leistung in Itemset 2 der Gruppe B zu t1

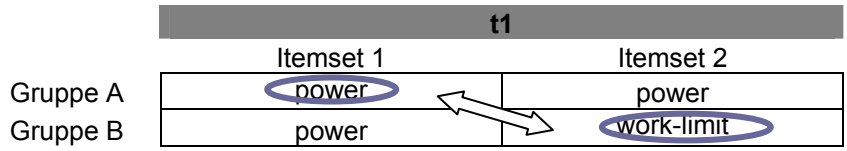
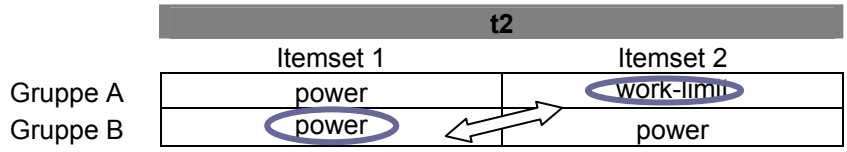


Tabelle 23: Vergleich der Leistung in Itemset 1 der Gruppe B mit der Leistung in Itemset 2 der Gruppe A zu t2



Außerdem wird die Leistung im ersten Itemset (power) der Gruppe A mit jener im zweiten Itemset (work-limit) der Gruppe B zu t1 verglichen (sieheTabelle 22). Im Anschluss wird das erste Itemset (power) der Gruppe B mit dem zweiten Itemset (work-limit) der Gruppe A zu t2 (siehe Tabelle 23) verglichen.

Weder zu t1 noch zu t2 können signifikante Leistungsunterschiede in Abhängigkeit der zu bearbeiteten Itemsets beobachtet werden (t1: z = -0.529, p = .596, t2: z = -1.038, p = .299).

Tabelle 24: Vergleich der Leistung in Itemset 1 der Gruppe B mit der Leistung in Itemset 2 der Gruppe A zu t1

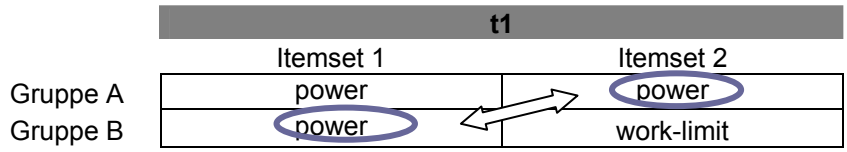
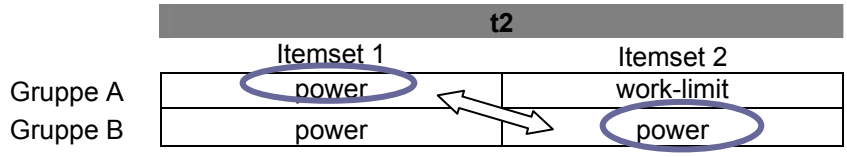


Tabelle 25: Vergleich der Leistung in Itemset 1 der Gruppe A mit der Leistung in Itemset 2 der Gruppe B zu t2



Um Leistungsunterschiede in Abhängigkeit der Testbedingung zu untersuchen, wird die Leistung unter der power Bedingung (Itemset 2) von Gruppe A zum ersten Testzeitpunkt mit der Leistung unter der power Bedingung (Itemset 1) von Gruppe B verglichen (siehe Tabelle 24). Anschliessend wird die Leistung unter der power Bedingung (Itemset 1) der Gruppe A mit der power Bedingung (Itemset 2)der Versuchsgruppe B zu t2 verglichen (siehe Tabelle 25).

Die Ergebnisse des U–Tests lassen darauf schliessen, dass es zu t1 keinen signifikanten Leistungsunterschied (z = -1.719, p = .086) zwischen Gruppe A und B bei den zu bearbeiteten Itemsets gibt und auch zu t2 weisen die Ergebnisse auf keinen signifikanten Unterschied hin (z = -1.372, p = .170).

Um zu prüfen, ob es Leistungsunterschiede zwischen der Vorgabeart power-power der Gruppe A zu t1 und der Vorgabeart power-power der Gruppe B zu t2 gibt (siehe Tabelle 26), wird aufgrund der schiefen Verteilung der Daten ein U-Test nach Mann & Whitney berechnet (siehe Tabelle 26).

Tabelle 26: Mittelwert (*M*) und Standardabweichung (*SD*) des in Prozentwerte transformierten Rohwerts. Ergebnisse des Kolmogorov-Smirnov-Tests zur Prüfung der Normalverteilung der verschiedenen Testbedingungen

		n	*M*	*SD*	*K-S*
	Gruppe A $_{n = 51}$				
t1	power-power (k =19)	51	34.47	22.26	.005
t2	power-worklimit (k =19)		41.69	21.90	.015
	Gruppe B $_{n = 49}$				
t1	power-worklimit (k =19)	49	35.45	19.27	<.001
t2	power- power (k=19)		39.85	21.35	.017

Tabelle 27: Vergleich der Leistung der reinen power- Bedingung über die Gruppen und Zeitpunkte

	t1		t2	
	Itemset 1	Itemset 2	Itemset 1	Itemset 2
Gruppe A	Power	power	power	work-limit
Gruppe B	Power	work-limit	power	power

Die Ergebnisse zeigen, dass sich die beiden Gruppen in der zu bearbeitenden Vorgabeart (power-power) nicht signifikant voneinander unterscheiden (z = -1.588, p = .112).

Tabelle 28: Vergleich der Leistung der power – work limit- Bedingung über die Gruppen und Zeitpunkte

	t1		t2	
	Itemset 1	Itemset 2	Itemset 1	Itemset 2
Gruppe A	Power	power	power	work-limit
Gruppe B	power	work-limit	power	power

Auch die Prüfung, ob sich die Vorgabeart power–work-limit über die Zeitpunkte und über die Gruppen hinweg unterscheidet (siehe Tabelle 28), zeigt, dass kein signifikanter

Unterschied (z = - 1.403, p = .161) zwischen den Leistungen in der Vorgabe power-work-limit über die Zeitpunkte und Gruppen hinweg beobachtet werden kann.

Zusammenfassend ist festzuhalten, dass sich die work-limit Instruktion nicht auf die Leistung der Probanden auswirkt.

13.1.2 Unterschiede in den Bearbeitungszeiten zwischen den Itemsets

Tabelle 29: Mittelwert (*M*) und Standardabweichung (*SD*) der Bearbeitungszeiten in Minuten für die einzelnen Itemsets des EST, k = 19 (n = Anzahl der Personen, k = Itemanzahl)

		n	*M*	*SD*
t1				
Gesamt		100	19.26	12.02
	Itemset 1		9.93	6.90
	Itemset 2		9.34	6.19
Gruppe A (po - po)		51	22.00	15.15
	Itemset 1		11.05	8.65
	Itemset 2		10.96	6.77
Gruppe B (po - wl)		49	16.41	8.57
	Itemset 1		8.76	4.17
	Itemset 2		7.65	5.07
t2				
Gesamt		100	19.56	12.78
	Itemset 1		10.19	7.50
	Itemset 2		9.38	6.02
Gruppe A (po - wl)		51	19.61	12.70
	Itemset 1		10.39	8.12
	Itemset 2		9.22	5.66
Gruppe B (po - po)		49	19.51	13.00
	Itemset 1		9.98	6.88
	Itemset 2		9.54	6.44

Die Zeit entspricht der reinen Bearbeitungszeit der Items (k= 19), ohne Instruktion und warm-up Item. Die durchschnittliche Bearbeitungszeit für den EST zu t1 lag bei 19.26 Minuten (SD = 12.02). Zu t2 betrug die durchschnittliche Bearbeitungszeit 19.56 Minuten (SD = 12.78). Mittels U-Test nach Mann & Whitney wurde untersucht, ob es Unterschiede in den Bearbeitungszeiten im zweiten Itemset zwischen den beiden Gruppen zu t1 wie auch zu t2 gibt. Die Ergebnisse lassen darauf schliessen, dass es zu t1 einen signifikante Unterschied (z = -2.665, p = .008) zwischen den beiden Gruppen in der Bearbeitungszeit des zweiten Itemsets gibt. Die Gruppe B bearbeitet das zweite Itemset unter der work-limit Bedingung signifikant schneller als die Gruppe A unter der

power Bedingung ($M_{\text{Gruppe A}}$: 10.96 vs. $M_{\text{Gruppe B}}$: 7.65). Zum zweiten Testzeitpunkt erwies sich das Ergebnis als nicht signifikant. Es kann kein signifikanter Einfluss der work-limit Instruktion auf die Bearbeitungszeit des zweiten Itemsets angenommen werden ($M_{\text{Gruppe A}}$: 9.22 vs $M_{\text{Gruppe B}}$: 9.54).

13.2 Zusammenhang der unterschiedlichen Instruktionsvorgaben mit den Außenkriterien

In der vorliegenden Arbeit ist ein weiteres Ziel die Klärung, ob Unterschiede in der Höhe der Korrelationen mit den Außenkriterien in Abhängigkeit von der Instruktionsbedingung (power oder work-limit) anzunehmen sind. Es stellt sich die Frage, ob die power- work-limit Vorgabe höhere Zusammenhänge zu den Außenkriterien aufweist als die reine power Bedingung.

Um die Korrelationen des EST mit den Schulleistungen, den selbsteingeschätzten Begabungen der Probanden sowie den Planzeichentest berechnen zu können, wurden die Items des Endlosschleifentests (EST), die unter der reinen power Bedingung der Gruppe A zum ersten Testzeitpunkt (t1) und der Gruppe B zum zweiten Testzeitpunkt (t2) vorgegeben wurden, zu “einer“ power-power Bedingung zusammengefasst (siehe Abbildung 11).

Die Items, welche die Gruppe A zum zweiten Testzeitpunkt (t2) und die Gruppe B zum ersten Testzeitpunkt /t1) unter der power-work-limit Instruktion bearbeitet hatte, wurden zu “einer“ power-work-limit Bedingung zusammengefasst (Abbildung 12). Wenn nachfolgend von der power-power oder power-work-limit Bedingung gesprochen wird, bezieht sich dies für jede der beiden Vorgabebedingungen jeweils immer auf beide Versuchsgruppen zusammen. Diese beiden Bedingungen power-power und power-work-limit werden nun anschließend mit den Außenkriterien korreliert, um zu untersuchen, ob Unterschiede in der Höhe der Korrelationen auftreten.

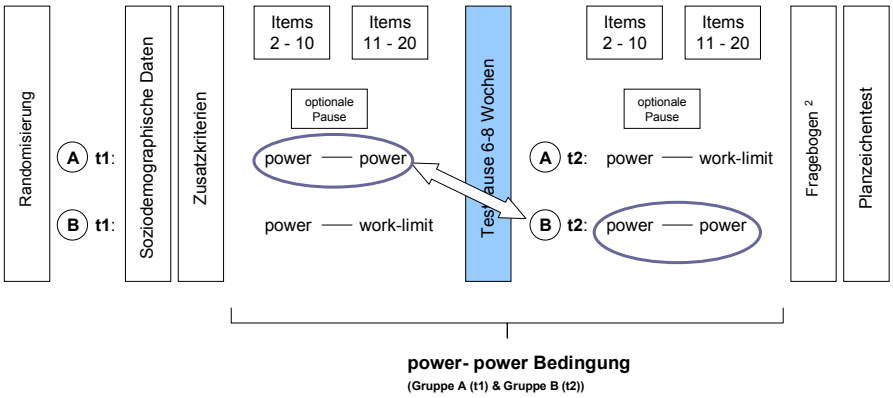


Abbildung 11: Veranschaulichung der Zusammenfassung der power-power Items beider Gruppen über die Zeitpunkte hinweg zu "einer" power- power Bedingung

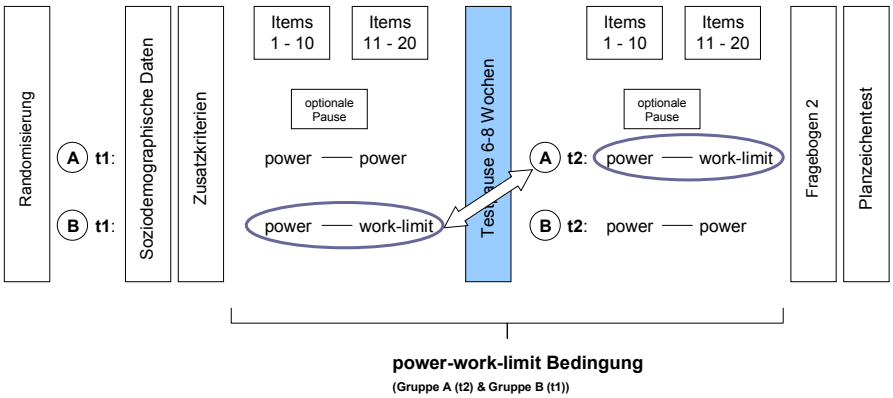


Abbildung 12: Veranschaulichung der Zusammenfassung der power-work-limit Items beider Gruppen über die Zeitpunkte hinweg zu "einer" power-work-limit Bedingung

13.2.1 Zusammenhang Raumvorstellung und Schulleistung

Die Versuchspersonen sollten, wie bereits in Kapitel 10.1 beschrieben, auf einer Analogskala ihre durchschnittliche Leistung in den Fächern Mathematik, Physik, Chemie, Deutsch und Englisch angeben. Die Versuchspersonen konnten von 1 (sehr gut) bis 5 (nicht genügend) wählen. Im Zuge der Auswertung wurden die Angaben der Probanden in 0 ("schlecht") und 100 ("sehr gut") umkodiert, und diese Werte für weitere Berechnungen herangezogen. In Tabelle 30 sind die Mittelwerte (*M*) und die Standardabweichungen (*SD*) der selbsteingeschätzten durchschnittlichen Schulleistung zu entnehmen.

Tabelle 30: Mittelwert (*M*) und Standardabweichung (*SD*) der selbsteingeschätzten durchschnittlichen Schulleistung in den einzelnen Fächern Mathematik (MA), Physik (PH), Chemie (CH), Deutsch (DE), Englisch (EN(und Latein (L) in Abhängigkeit der Zwischensubjektfaktoren Geschlecht und Testbedingung (n = Anzahl der Personen), (0 = "schlecht" und 100 = "sehr gut")

Mathematik	**n_{MA}**	***M*** **$_{MA}$**	***SD*** **$_{MA}$**
Gesamt	99	54.79	20.66
Männer	53	56.26	21.59
Frauen	46	53.09	19.62
Gruppe A	50	53.38	20.66
Gruppe B	49	56.22	20.76
Physik	**n_{PH}**	***M*** **$_{PH}$**	***SD*** **$_{PH}$**
Gesamt	98	56.80	20.40
Männer	52	56.46	18.86
Frauen	46	57.17	22.21
Gruppe A	51	58.33	18.97
Gruppe B	47	55.13	21.92
Chemie	**n_{CH}**	***M*** **$_{CH}$**	**SD $_{CH}$**
Gesamt	98	56.23	21.50
Männer	52	54.21	20.59
Frauen	46	58.52	22.47
Gruppe A	51	57.08	22.27
Gruppe B	47	55.32	20.82
Deutsch	**n_{DE}**	***M*** **$_{DE}$**	**SD $_{DE}$**
Gesamt	96	62.05	21.88
Männer	51	58.04	21.06
Frauen	45	66.60	22.13
Gruppe A	49	64.43	20.21
Gruppe B	47	59.57	23.46
Englisch	**n_{EN}**	***M*** **$_{EN}$**	**SD $_{EN}$**
Gesamt	97	61.01	22.36
Männer	52	60.37	22.82
Frauen	45	61.76	22.04

Gruppe A	50	62.36	21.6
Gruppe B	47	59.57	23.32
Latein	**n_L**	**M_L**	**SD_L**
Gesamt	31	42.84	29.39
Männer	16	34.88	27.51
Frauen	15	51.33	29.83
Gruppe A	15	44.27	32.82
Gruppe B	16	41.50	26.81

Um die Zusammenhänge von EST-Leistung und den von den Probanden angegebenen durchschnittlichen Leistungen in den Schulfächern zu ermitteln, werden die Korrelationskoeffizienten nach Pearson berechnet. Nach Bortz und Döring (2002) ist der Korrelationskoeffizient ein

> Quantitatives Maß für Enge und Richtung des Zusammenhangs zwei oder mehrerer Variablen. […] Je höher der Betrag eines Korrelationskoeffizienten, umso enger der Zusammenhang. Ob es sich um einen statistisch bedeutsame Korrelation handelt, zeigt jedoch erst der *Korrelationstest*. […] Der Richtung des Zusammenhangs nach unterscheidet man positive und negative Korrelationen. Eine positive Korrelation besagt, daß hohe Werte in der einen Variablen mit hohen Werten in der anderen Variable einhergehen[…]. Bei einer negativen Korrelation ist die Beziehung zwischen den Variablen gegensinnig: hohe Werte in der einen Variable gehen mit niedrigen Werten in der anderen Variable einher und umgekehrt (S.682).

Der Korrelationskoeffizient (r), kann Werte zwischen -1 und +1 annehmen, wobei ein Betrag nahe 1 einen starken und ein Betrag nahe bei 0 ein schwachen Zusammenhang bedeutet (Bühl, 2010, S. 386).

In Bezug auf die durchschnittlichen selbsteingeschätzten Schulnoten in den Fächern Mathematik, Physik, Chemie, Deutsch, Englisch und Latein steht ein positiver Korrelationskoeffizient für höhere Angaben (Analogskala: 0 =“schlecht“ und 100 = “sehr gut“) in den Schulfächern und höhere Testwerte im EST und ein negativer Korrelationskoeffizient für eine gegensinnige Beziehung der Variable Schulnote und Leistung im EST.

Tabelle 31: Pearson-Korrelation *r* zwischen der durchschnittlichen Leistung im Schulfach Mathematik und der Testleistung in der power-power und power-work-limit Vorgabe des EST (n = Anzahl der Personen, k = Itemanzahl)

Mathematik	n	k	Pearson-Korrelation	Signifikanz (*p*)
power-power Bedingung (Gruppe A (t1)& Gruppe B (t2))				
Gesamt $_{\text{power-power}}$	99	19	.165	.102
Frauen $_{\text{power-power}}$	46		.130	.388
Männer $_{\text{power-power}}$	53		.160	.251
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))				
Gesamt $_{\text{power-work-limit}}$	99	19	.121	.234
Frauen $_{\text{power-work-limit}}$	46		.152	.313
Männer $_{\text{power-work-limit}}$	53		.087	.538

Die durchschnittliche Leistung im Schulfach Mathematik korreliert weder signifikant mit der erzielten Leistung der Probanden unter der power-power Bedingung (r = .165, p = .102) noch mit der erbrachten Leistung unter der power- work-limit Bedingung (r = .121, p = .234). Auch sonst lassen sich keine statistisch nachweisbaren Zusammenhänge beobachten (siehe Tabelle 31).

Tabelle 32: Pearson-Korrelation *r* zwischen der durchschnittlichen Leistung im Schulfach Physik und der der Testleistung in der power-power und power-work-limit Vorgabe des EST (n = Anzahl der Personen, k = Itemanzahl)

Physik	n	k	Pearson-Korrelation	Signifikanz (*p*)
power-power Bedingung (Gruppe A (t1)& Gruppe B (t2))				
Gesamt $_{\text{power-power}}$	99	19	.272	**.007**
Frauen $_{\text{power-power}}$	46		.313	**.034**
Männer $_{\text{power-power}}$	53		.304	**.029**
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))				
Gesamt $_{\text{power-work-limit}}$	98	19	.177	.082
Frauen $_{\text{power-work-limit}}$	46		.230	.124
Männer $_{\text{power-work-limit}}$	52		.213	.129

Wie anhand der Ergebnisse der Tabelle 32 ersichtlich ist, sind mäßige, signifikant positive Zusammenhänge zwischen der durchschnittlichen Physiknote und der Leistung (r =.272, p = .007) unter der power-power Bedingung des EST sowie der Leistung der Frauen (r =.313, p = .034) wie auch der Männer (r = .304, p = .029) unter der reinen power Vorgabe anzunehmen.

Tabelle 33: Pearson-Korrelation *r* zwischen der durchschnittlichen Leistung im Schulfach Chemie der Testleistung in der power-power und power-work-limit Vorgabe des EST (n = Anzahl der Personen, k = Itemanzahl)

Chemie	n	k	Pearson-Korrelation	Signifikanz (*p*)
power-power Bedingung (Gruppe A (t1)& Gruppe B (t2))				
Gesamt $_{\text{power-power}}$	98	19	.140	.169
Frauen $_{\text{power-power}}$	46		.389	**.008**
Männer $_{\text{power-power}}$	53		.063	.657
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))				
Gesamt $_{\text{power-work-limit}}$	98	19	.035	.733
Frauen $_{\text{power-work-limit}}$	46		.294	**.048**
Männer $_{\text{power-work-limit}}$	52		-.005	.970

In Bezug auf die durchschnittliche Note im Fach Chemie lassen sich, wie aus Tabelle 33 ersichtlich, signifikant positive Korrelationen mit den Testleistungen von Frauen in der power-power Vorgabe (r = .389, p =.008), sowie für die power-work-limit Bedingung (r = .294, p =.048) nachweisen.

Tabelle 34: Pearson-Korrelation *r* zwischen der durchschnittlichen Leistung im Schulfach Deutsch und der der Testleistung in der power-power und power-work-limit Vorgabe des EST (n = Anzahl der Personen, k = Itemanzahl)

Deutsch	n	k	Pearson-Korrelation	Signifikanz (*p*)
t1				
power-power Bedingung (Gruppe A (t1)& Gruppe B (t2))				
Gesamt $_{\text{power-power}}$	96	19	-.127	.219
Frauen $_{\text{power-power}}$	45		.050	.747
Männer $_{\text{power-power}}$	51		-.140	.326
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))				
Gesamt $_{\text{power-work-limit}}$	96	19	-.125	.226
Frauen $_{\text{power-work-limit}}$	45		.092	.548
Männer $_{\text{power-work-limit}}$	51		-.121	.399

Die durchschnittliche Note im Unterrichtsfach Deutsch korreliert durchwegs sehr gering negativ mit der Testleistung der Probanden, es konnten keine signifikanten Korrelationen festgestellt werden (siehe Tabelle 34).

Tabelle 35: Pearson-Korrelation *r* zwischen der durchschnittlichen Leistung im Schulfach Englisch und der der Testleistung in der power-power und power-work-limit Vorgabe des EST (n = Anzahl der Personen, k = Itemanzahl)

Englisch	n	k	Pearson-Korrelation	Signifikanz (*p*)
power-power Bedingung (Gruppe A (t1)& Gruppe B (t2))				
Gesamt $_{\text{power-power}}$	97	19	-.115	.263
Frauen $_{\text{power-power}}$	45		-.144	.345
Männer $_{\text{power-power}}$	52		-.095	.505
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))				
Gesamt $_{\text{power-work-limit}}$	97	19	-.130	.206
Frauen $_{\text{power-work-limit}}$	45		.063	.679
Männer $_{\text{power-work-limit}}$	52		-.209	.137

Wie aus Tabelle 35 zu entnehmen ist, sind zwischen der durchschnittlichen Schulleistung im Schulfach Englisch und den Ergebnissen der Probanden im Raumvorstellungstest keine signifikanten Zusammenhänge zu beobachten.

Tabelle 36: Pearson-Korrelation *r* zwischen der durchschnittlichen Leistung im Schulfach der Testleistung in der power-power und power-work-limit Vorgabe des EST (n = Anzahl der Personen, k = Itemanzahl)

Latein	n	k	Pearson-Korrelation	Signifikanz (p)
power-power Bedingung (Gruppe A (t1)& Gruppe B (t2))				
Gesamt $_{\text{power-power}}$	31	20	.061	.744
Frauen $_{\text{power-power}}$	15		.258	.354
Männer $_{\text{power-power}}$	16		.096	.723
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))				
Gesamt $_{\text{power-work-limit}}$	31	20	.034	.858
Frauen $_{\text{power-work-limit}}$	15		.305	.269
Männer $_{\text{power-work-limit}}$	16		.041	.880

Zwischen der durchschnittlichen Leistung im Fach Latein und der Testleistung der Probanden kann durchwegs kein signifikanter Zusammenhang festgestellt werden (siehe Tabelle 36).

13.2.2 Zusammenhang Raumvorstellung und Begabung

Wie bereits in Kapitel 10.1 beschrieben, wurden die Versuchspersonen gebeten, sich hinsichtlich ihrer Begabungen in Mathematik, Musik, Technik und Chemie auf einer Analogskala von 0 ("schlecht") und 100 ("sehr gut") selbst einzuschätzen.

Tabelle 37: Mittelwert (*M*) und Standardabweichung (*SD*) der selbsteingeschätzten Begabungen in Mathematik, Musik, Technik und Raumvorstellung, (n = Anzahl der Personen), (0 = "schlecht", 100 = "sehr gut")

Mathematische Begabung	n_{MA_BG}	M_{MA_BG}	SD_{MA_BG}
Gesamt	100	51.21	22.76
Männer	53	57.85	22.01
Frauen	47	43.72	21.43
Gruppe A	51	52.61	22.38
Gruppe B	49	49.76	23.29
Musikalische Begabung	n_{MU_BG}	M_{MU_BG}	SD_{MU_BG}
Gesamt	100	45.30	28.39
Männer	53	39.98	28.38
Frauen	47	51.30	27.48
Gruppe A	51	46.02	27.25
Gruppe B	49	44.55	29.81
Technische Begabung	n_{TECH_BG}	M_{TECH_BG}	SD_{TECH_BG}
Gesamt	100	50.99	25.65
Männer	53	61.87	24.45
Frauen	47	38.72	21.21
Gruppe A	51	48.08	26.09
Gruppe B	49	54.02	25.10
Raumvorstellungs- Begabung	n_{RV_BG}	M_{RV_BG}	SD_{RV_BG}
Gesamt	100	56.46	24.17
Männer	53	63.85	21.22
Frauen	47	48.13	24.79
Gruppe A	51	57.41	25.98
Gruppe B	49	55.47	22.37

Um die Zusammenhänge zwischen der Leistung im EST und den von den Versuchspersonen angegebenen selbsteingeschätzten Begabungen zu ermitteln, werden Pearson-Korrelationen berechnet. Höhere Werte stehen aufgrund der vorgegebenen Analogskala (0 ="schlecht" und 100 ="sehr gut") für höhere Angaben in der selbsteingeschätzten Begabung.

Tabelle 38: Pearson-Korrelation *r* zwischen der selbsteingeschätzten mathematischen Begabung der Testleistung in der power-power und power-work-limit Vorgabe des EST (n = Anzahl der Personen, k = Itemanzahl)

Mathematische Begabung	n	k	Pearson-Korrelation	Signifikanz(*p*)
power-power Bedingung (Gruppe A (t1)& Gruppe B (t2))				
Gesamt $_{\text{power-power}}$	100	19	.321	**.001**
Frauen $_{\text{power-power}}$	47		.137	.360
Männer $_{\text{power-power}}$	53		.306	**.026**
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))				
Gesamt $_{\text{power-work-limit}}$	100	19	.424	**<.001**
Frauen $_{\text{power-work-limit}}$	47		.201	175
Männer $_{\text{power-work-limit}}$	53		.434	**.001**

Im Bezug auf die Selbsteinschätzung der mathematischen Begabung und der Gesamtleistung zeigt sich sowohl in der power-power Bedingung (r = .321, p = .001), als auch in der power-work-limit Vorgabe (r = .424, p = <.001), wie der Tabelle 38 zu entnehmen ist, ein signifikant positiver Zusammenhang, dieser zeigt sich auch bei den männlichen Personen für beide Vorgabebedingungen (power-power: r = .306, p = .026; power-work-limit: r = .434, p = .001)

Tabelle 39: Pearson-Korrelation *r* zwischen der selbsteingeschätzten musikalischen Begabung der Testleistung in der power-power und power-work-limit Vorgabe des EST (n = Anzahl der Personen, k = Itemanzahl)

Musikalische Begabung	n	k	Pearson-Korrelation	Signifikanz(*p*)
power-power Bedingung (Gruppe A (t1)& Gruppe B (t2))				
Gesamt $_{\text{power-power}}$	100	19	-.047	.645
Frauen $_{\text{power-power}}$	47		.089	.553
Männer $_{\text{power-power}}$	53		-.014	.919
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))				
Gesamt $_{\text{power-work-limit}}$	100	19	-.197	**.050**
Frauen $_{\text{power-work-limit}}$	47		-.048	.747
Männer $_{\text{power-work-limit}}$	53		-.177	.205

Tabelle 39 ist zu entnehmen, dass zwischen der selbsteingeschätzten musikalischen Begabung und der Testleistung in der power-work-limit Bedingung ein negativer signifikanter Zusammenhang anzunehmen ist.

Tabelle 40: Pearson-Korrelation *r* zwischen der selbsteingeschätzten technischen Begabung und der Testleistung in der power-power und power-work-limit Vorgabe des EST (n = Anzahl der Personen, k = Itemanzahl)

Technische Begabung	n	k	Pearson-Korrelation	Signifikanz(*p*)
power-power Bedingung (Gruppe A (t1)& Gruppe B (t2))				
Gesamt $_{\text{power-power}}$	100	19	.477	**<.001**
Frauen $_{\text{power-power r}}$	47		.430	**.003**
Männer $_{\text{power-power}}$	53		.366	**.007**
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))				
Gesamt $_{\text{power-work-limit}}$	100	19	.358	**<.001**
Frauen $_{\text{power-work-limit}}$	47		.158	.288
Männer $_{\text{power-work-limit}}$	53		.260	.060

Zwischen der selbsteingeschätzten technischen Begabung und der Gesamtleistung in der power-power (*r* = .477, *p* =. <.001) als auch in der power-work-limit Bedingung (*r* = .358, *p* =. <.001) kann ein mäßiger, signifikant positive Zusammenhang festgestellt werden. Auch können signifikant positive Korrelationen zwischen der selbsteingeschätzten technischen Begabung und der Testleistung der Frauen (*r* =.430, *p* = .003) wie auch der Männer (*r* =.366, *p* = .007) in der power-power Bedingung beobachtet werden (siehe Tabelle 40).

Tabelle 41: Pearson-Korrelation *r* zwischen der selbsteingeschätzten Raumvorstellung Begabung und der Testleistung in der power-power und power-work-limit Vorgabe des EST (n = Fälle, k = Items)

Raumvorstellung Begabung	n	k	Pearson-Korrelation	Signifikanz(*p*)
power-power Bedingung (Gruppe A (t1)& Gruppe B (t2))				
Gesamt $_{\text{power-power}}$	100	19	.325	**.001**
Frauen $_{\text{power-power}}$	47		.239	.106
Männer $_{\text{power-power}}$	53		.257	.064
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))				
Gesamt $_{\text{power-work-limit t}}$	100	19	.309	**.002**
Frauen $_{\text{power-work-limit}}$	47		.152	.307
Männer $_{\text{power-work-limit}}$	53		.270	.051

Wie anhand der Tabelle 41 ersichtlich, korreliert die Testleistung in der power-power Bedingung (*r* = .325, *p* = .001) als auch die Leistung im EST in der power-work-limit Vorgabe (*r* = .309, *p* = .001) signifikant positiv mit der selbsteingeschätzten Raumvorstellung.

13.2.3 Zusammenhang Raumvorstellung und Planzeichentest

Wie in Kapitel 9.2.3 beschrieben, wurden den Versuchspersonen auf einem A4-Blatt, Orientierungspunkte der Stadt Wien („alter“ Südbahnhof, Westbahnhof, neues allgemeines Krankenhaus, Hauptuniversität und Stadtpark) vorgegeben und die Versuchspersonen hatten die Aufgabe, die folgenden Suchpunkte möglichst genau mittels eines Punktes einzuzeichnen: Stephansplatz, Karlskirche, Votivkirche, Oper, Urania, Praterstern und Stadthalle. Die Auswertung des Planzeichentests erfolgte anhand einer Folie, die über das A4-Blatt gelegt wurde, auf der die richtigen Koordinaten der Suchpunkte eingezeichnet waren. Die Distanz zwischen dem eingezeichneten Suchpunkt des Probanden und der tatsächlichen Koordinate des Suchpunktes lieferte das Ergebnis für jeden einzelnen Suchpunkt. Im Anschluss wurden die einzelnen Distanzen für jede Versuchspersonen aufsummiert, somit wurde für jeden Testteilnehmer ein Gesamtwert gebildet. Je höher dieser Wert ausfällt, desto schlechter konnte die Versuchsperson die Suchpunkte einschätzen und einzeichnen.

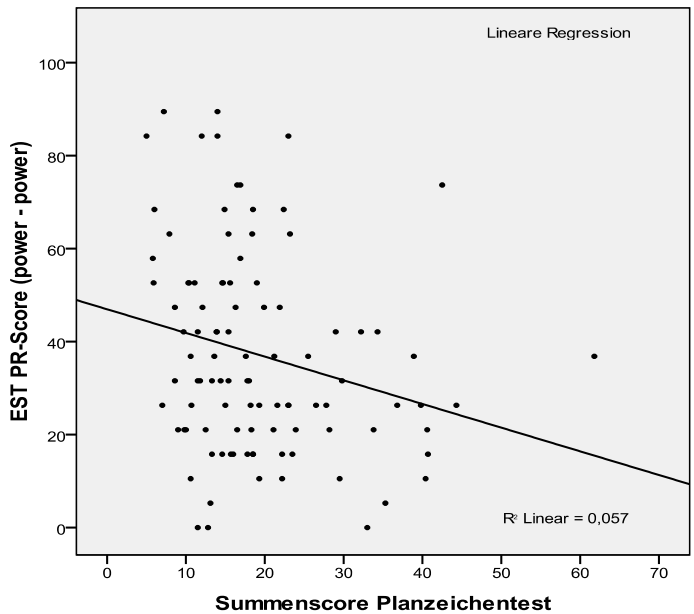
Um Zusammenhänge zwischen der Leistung im Raumvorstellungstest und den Leistungen im Planzeichentest zu ermitteln, werden Pearson-Korrelationen berechnet. Das negative Vorzeichen des Korrelationskoeffizienten gibt an, dass ein hoher Wert [7]im Planzeichentest mit niedrigen Testwerten im EST einhergeht.

[7] hohe Werte im Planzeichentest bedeuten schlechtere Leistung in diesem

Tabelle 42: Pearson-Korrelation *r* zwischen dem Planzeichentest und der Testleistung in der power-power und power-work-limit Vorgabe des EST (n = Anzahl der Personen, k = Itemanzahl)

Planzeichentest	n	*k*	Pearson-Korrelation	Signifikanz(*p*)
power-power Bedingung (Gruppe A (t1)& Gruppe B (t2))				
Gesamt $_{\text{power-power}}$	100	19	-.238	**.017**
Frauen $_{\text{power-power}}$	47		-.350	**.016**
Männer $_{\text{power-power}}$	53		-.174	.212
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))				
Gesamt $_{\text{power-work-limit}}$	100	19	-.167	.096
Frauen $_{\text{power-work-limit}}$	47		-.400	**.005**
Männer $_{\text{power-work-limit}}$	53		-.081	.564

Es können signifikant negative Zusammenhänge (r = -.238, p = .017) zwischen der Gesamttestleistung der power-power Bedingung und der Leistung im Planzeichentest angenommen werden. Zwischen den Leistungen im Planzeichentest der Testpersonen und der Testleistung im EST können signifikant negative Zusammenhänge für die Frauen sowohl unter der power-power (r = -.350, p = .016) wie auch unter der power-work-limit Vorgabe (r =.-.400, p = .005) festgestellt werden.



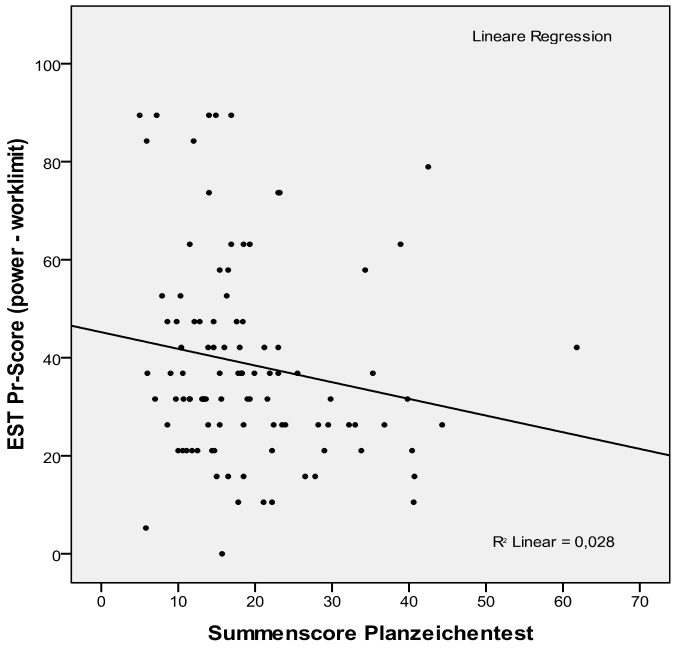


Abbildung 13: Bivariate Streudiagramme zum Zusammenhang zwischen Leistungen im EST und Leistungen im Planzeichentest unter der reinen power Bedingung und unter power-worklimit Bedingung

13.2.4 Vergleich der Korrelationskoeffizienten

Tabelle 43: Korrelationskoeffizienten *r* der Gesamtstichprobe sowie der Teilstichprobe Frauen und Männer in der power-power Bedingung und der power-work-limit Bedingung zu den Außenkriterien, Schulleistung in Mathematik (MA), Physik (PH), Chemie (CH), Deutsch (DE), Englisch (EN) und Latein (L) sowie den Begabungen in Mathematik (MA BG), Musik (MU BG), Technik (TECH BG), Raumvorstellung (RV BG) und dem Planzeichentest (PLAN); k = 19

			Korrelationskoeffizient *r*							
			Gesamt			Frauen			Männer	
		n	**po-po**[8]	**po-wl**[9]	**n**	**po-po**	**po-wl**	**n**	**po-po**	**po-wl**
Schul-leistung	**MA**	99	.165	.121	46	.130	.152	53	.160	.087
	PH	98	.272	.177	46	.313	.230	52	.304	.213
	CH		.140	.035		.389	.294		.063	-.005
	DE	96	-.127	-.125	45	.050	.092	51	-.140	-.121
	EN	97	-.115	-.130	45	-.144	.063	52	-.095	-.209
	L	31	.061	.034	15	.258	.305	16	.096	.041
Begab-ung	**MA BG**	100	.321	.424	100	.137	.201	100	.306	.434
	MU BG		-.047	-.197		.089	-.048		-.014	-.177
	TECH BG		.477	.358		.430	.158		.366	.260
	RV BG		.325	.309		.239	.152		.257	.270
	PLAN		-.238	-.167		-.350	-.400		-.174	-.081

[8] po-po = Abkürzung: power-power Bedingung

[9] po-wl = Abkürzung: power-work-limit Bedingung

Wie ein Vergleich der Korrelationskoeffizienten in Tabelle 43 zeigt, unterscheiden sich die beiden Vorgabearten durchwegs nur geringfügig in der Höher der Korrelationen mit den Außenkriterien. Die größten Unterschiede in der Höhe des Korrelationskoeffizienten, finden sich bei den Frauen in der selbsteingeschätzten technischen Begabung ($r_{\text{power-power}}$ = .430; $r_{\text{power-work-limit}}$ = .158). Frauen, die sich bezüglich ihrer technischer Begabung und Raumvorstellungsbegabung besser einschätzen, weisen unter Zeitdruck-komponente schlechtere Ergebnisse im EST auf als ohne Zeitdruckkomponente.

13.2.5 Retest-Reliabilität

> Um die Reliabilität nach der Retest-Methode zu bestimmen, wird ein und derselbe Test (unter der idealen Annahme, dass sich das zu messende Merkmal selbst nicht verändert hat) zu zwei verschiedenen Zeitpunkten vorgelegt. Die Reliabilität wird dann als Korrelation zwischen den beiden Testergebnissen ermittelt. Bei der Retest-Reliabilität ist zu beachten, dass die ermittelte Korrelation in Abhängigkeit vom Zeitintervall zwischen beiden Testungen variieren kann. Je nach Zeitabstand ist nämlich eine Vielzahl von Einflüssen auf die Messungen denkbar, die sich reliabilitätsverändernd auswirken können, insbesondere Übungs- und Erinnerungseffekte oder ein sich tatsächlich veränderndes Persönlichkeitsmerkmal (Moosbrugger & Kelava, 2007, S. 12).

Die Retest-Reliabilität des Gesamttests des EST beträgt r_{tt} = .689. Zwischen ersten Testzeitpunkt und zweiten Testzeitpunkt ist daher ein mittlerer Zusammenhang anzunehmen.

Getrennt nach den beiden Vorgabebedingungen betrachtet, kann bei der Gruppe A von einer höheren Korrelation zwischen den beiden Testzeitpunkten des EST ausgegangen werden, als für die Gruppe B (Gruppe A: r_{tt} = .754 vs Gruppe B: r_{tt} = .611).

13.2.6 Unterschied in der Raumvorstellung in Abhängigkeit vom biologischen und psychologischen Geschlecht

Es wird geprüft, ob auf Basis des psychologischen Geschlechts - erhoben mittels dem Fragebogen zur Persönlichkeitseinschätzung “Selbstbild“ (siehe Kapitel 9.2.2) - und des biologischen Geschlechts Unterschiede in der Raumvorstellung anzunehmen sind.

Tabelle 44: Selbsteinschätzung der Geschlechterrollenverteilung (N = 100).

	Häufigkeit	Prozent	Kumulierte Prozente
undifferenziert	7	7,0	7,0
feminin	37	37,0	44,0
maskulin	42	42,0	86,0
androgyn	14	14,0	100,0
Gesamt	100	100,0	

Aufgrund der ungleichen Verteilung der einzelnen Geschlechtsrollengruppen und der damit verbundenen Schwierigkeit, weitere Berechnungen durchzuführen, werden die vier Gruppen in Gruppe 1 mit maskulin & undifferenziert sowie Gruppe 2 mit feminin & androgyn zusammengefasst.

Tabelle 45: Kreuztabelle für biologisches Geschlecht und Geschlechtsrollenidentität

			Geschlechtsrollenidentität		Gesamt
			Männlich/ undifferenziert	weiblich/ androgyn	
Biologisches Geschlecht	m	Anzahl	40	13	53
		erwartet	26,0	27,0	53,0
		% innerhalb	75,5%	24,5%	100,0%
		Stand. Residuen	2,8	-2,7	
	w	Anzahl	9	38	47
		erwartet	23,0	24,0	47,0
		% innerhalb	19,1%	80,9%	100,0%
		Stand. Residuen	-2,9	2,9	
Gesamt		Anzahl	49	51	100
		erwartet	49,0	51,0	100,0
		% innerhalb	49,0%	51,0%	100,0%

Die Berechnung der entsprechenden Prüfgröße ergibt mit χ^2 (1) = 31.621, p < .001 ein signifikantes Ergebnis, es kann ein deutlicher Zusammenhang zwischen psychologischem Geschlecht in Abhängigkeit vom biologischen Geschlecht beobachtet werden (siehe Tabelle 45).

Tabelle 46: Mittelwert (*M*) und Standardabweichung (*SD*) der in Prozentwerten transformierten Rohwerte für das biologische und psychologische Geschlecht der power- power und power- worklimit Bedingung beider Gruppen

Test-bedingung	biologisches Geschlecht	psychologisches Geschlecht	n	*M*	*SD*
power-power Bedingung (Gruppe A (t1) & Gruppe B (t2))	Gesamt	Männlich/undifferenziert	49	42.96	22.85
		weiblich/androgyn	51	31.48	19.51
		Gesamt	100	37.11	21.88
	w	Männlich/undifferenziert	9	38.01	14.84
		weiblich/androgyn	38	27.15	16.03
		Gesamt	47	29.23	16.23
	m	Männlich/undifferenziert	40	44.08	24.30
		weiblich/androgyn	13	44.13	23.69
		Gesamt	53	44.09	23.92
power-work-limit Bedingung (Gruppe A (t2) & Gruppe B (t1))	Gesamt	Männlich/undifferenziert	49	44.58	23.14
		weiblich/androgyn	51	32.92	16.54
		Gesamt	100	38.63	20.79
	w	Männlich/undifferenziert	9	36.26	11.90
		weiblich/androgyn	38	28.81	10.65
		Gesamt	47	30.24	11.16
	m	Männlich/undifferenziert	40	46.45	24.71
		weiblich/androgyn	13	44.94	24.08
		Gesamt	53	46.08	24.34

Um mögliche Unterschiede in Abhängigkeit der Geschlechtsrollengruppen und dem Geschlecht sowie dem Innersubjektfaktor Testbedingung zu untersuchen, wird eine gemischte abhängige Varianzanalyse berechnet.

Der Levene-Test auf Varianzhomogenität fällt für die power-power Bedingung (F (3,96) = 3.447, p = .020) und für die power- work-limit Bedingung (F (3,96) = 9.127, p = < .001),

signifikant aus, womit von Heterogenität der Varianzen ausgegangen werden kann. Da, wie vorher schon erwähnt, die Varianzanalyse jedoch als verhältnismäßig robust gegenüber Verletzungen ihrer Vorraussetzungen angesehen werden kann (Backhaus, Erichson, Plinke & Weiber, 2003), wird sie nachfolgend für die Berechnung möglicher Unterschiede herangezogen.

Tabelle 47: Tafel der gemischten abhängigen Varianzanalyse mit Geschlecht und Geschlechtsrollengruppe als Zwischensubjektfaktoren und Testbedingung als Innersubjektfaktor.

Quelle	Quadratsumme vom Typ III	Mittel der Quadrate	*F*(1,96)	Sig.(*p*)
Testbedingung	19.892	19.892	0.124	.726
Testbedingung * biologisches Geschlecht	22.344	22.344	0.139	.710
Testbedingung * Geschlechtsrolle	7.209	7.209	0.045	.833
Testbedingung * biologisches Geschlecht *Geschlechtsrolle	51.707	51.707	0.322	.572
Fehler (Testbedingung)	15426.632	160.694		

Bei nicht signifikanter Wechselwirkung zwischen der Testbedingung und dem biologischen Geschlecht (p = .710) sowie zwischen Testbedingung und dem psychologischen Geschlecht (p = .833), lassen die Ergebnisse auf einen nicht signifikanten Unterschied (p = .726) in der erzielten Leistung in Abhängigkeit von der Testbedingung schliessen (siehe Tabelle 47).

Im Bezug auf das psychologische Geschlecht zeigen sich keine signifikanten Unterschiede (p = .263) in den erzielten Testergebnissen, es kann jedoch ein signifikanter Unterschied (p = .006) in den erbrachten Leistungen der Probanden in Abhängigkeit vom biologischen Geschlecht angenommen werden. Männliche Versuchspersonen schneiden sowohl in der power-power Bedingung ($M_{\text{Männer}}$ = 44.09 vs. M_{Frauen} = 29.23) als auch der power- work-limit Bedingung ($M_{\text{Männer}}$ = 46.08 vs. M_{Frauen} = 30.24) besser ab als weibliche Versuchspersonen (siehe Abbildung 14)

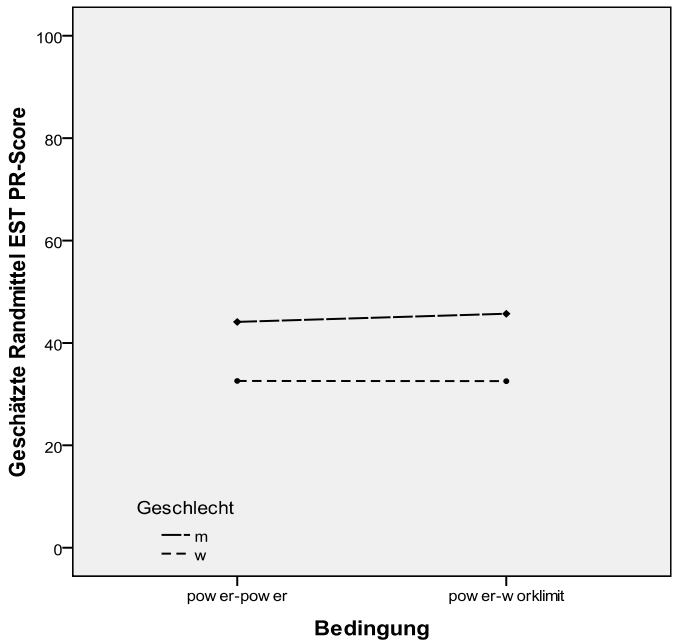


Abbildung 14: Testleistung der in Prozentwerte transformierten Rohscores der Männer und Frauen in der power Bedingung und in der work-limit Bedingung (Gruppe A & Gruppe B)

14 Zusammenfassung und Diskussion

Das Thema der vorliegenden Arbeit war, ob die work-limit Instruktion, die im Gegensatz zur power Vorgabe eine Zeitdruckkomponente besitzt, zusätzliche Informationen bezüglich der Leistung bei der Bearbeitung von Endlosschleifentestaufgaben (EST) liefert, sowie die Validierung des Testinventars.

Es wurden 100 Personen zu zwei Testzeitpunkten getestet. Zu beiden Messzeitpunkten wurde von beiden Gruppen (A und B) der Endlosschleifentest bearbeitet, der jeweils aus zwei Testhälften (Itemsets) bestand. Die Items waren zu beiden Messzeitpunkten für beide Versuchsgruppen identisch, einzig die Instruktion vor der Bearbeitung der zweiten Testhälfte war unterschiedlich. Zusätzlich wurden Zusatzkriterien erhoben.

Der Endlosschleifentest gilt unter der power-Vorgabe als Rasch-homogen (Gittler & Arendasy, 2003) und eine zentrale Frage war demnach, ob die Rasch-Homogenität auch unter der work-limit Bedingung gegeben ist.

Zunächst wurden die Daten unter der power Bedingung auf Rasch-Homogenität geprüft. Aufgrund der Modellgeltungsprüfung nach Andersen (1973) kann davon ausgegangen werden, dass die Daten unter der power Bedingung eindimensional im Sinne von Rasch sind.

Die anschließend durchgeführte Modellüberprüfung nach Andersen (1973) lässt darauf schließen, dass auch unter der work-limit Bedingung angenommen werden kann, dass die Daten in allen Untergruppen von Personen dieselbe latente Dimension Raumvorstellung erfassen. Ergänzend konnte mit dem Modelltest nach Martin-Löf (1973) gezeigt werden, dass die beiden Skalen power und work-limit dieselbe latente Dimension erheben. Die Zeitdruckkomponente beeinflusst die Dimensionalität des Testinventars demnach nicht wesentlich.

Nachdem die beiden vorangegangenen Modellgeltungskontrollen sowohl unter der power als auch unter der work-limit Vorgabe nicht signifikante Ergebnisse lieferten, wurde der Gesamttest auf Rasch - Homogenität überprüft. Auch hier lassen die Ergebnisse darauf schliessen, dass der Gesamttest dieselbe latente Dimension misst. Nachfolgend wurde eine Itemhomogenitätsprüfung nach Martin-Löf (1973) durchgeführt, welche belegt, dass die Itemhomogenität für das erste Itemset und das zweite Itemset gegeben ist.

Auch die zusätzlich durchgeführte Modellkontrolle nach Andersen (1973) des unvollständigen Raschmodells ergab, dass die Items in beiden Itemsets, welche der Gruppe A zu t1 als power-power Bedingung und zu t2 als power-worklimit Bedingung vorgegeben wurden, sowie der Gruppe B zu t1 als power-worklimit und zu t2 als power-power Bedingung vorgegeben wurde, dieselbe latente Dimension messen.

Zusammenfassend kann festgehalten werden, dass alle Modellgeltungskontrollen nicht signifikante Ergebnisse lieferten.

Anschließend wurden ergänzend weitere inferenzstatistische Analysen durchgeführt. Der Vergleich der beiden Itemsets zum ersten Testzeitpunkt ergab, dass die Versuchspersonen im ersten Itemset signifikant besser abschnitten als im zweiten Itemset. Dieses könnte auf mögliche Motivations- oder Ermüdungseffekte

zurückzuführen sein. Zum zweiten Testzeitpunkt konnte kein statistisch bedeutender Leistungsunterschied zwischen den Testhälften festgestellt werden.

Des Weiteren kann aufgrund der Ergebnisse der gemischten abhängigen Varianzanalyse angenommen werden, dass die Versuchspersonen zum zweiten Testzeitpunkt bessere Ergebnisse im EST erzielten als zum ersten Testzeitpunkt. Die Betrachtung der Leistungsunterschiede innerhalb der einzelnen Gruppen über die Zeitpunkte hinweg zeigt, dass Gruppe A im ersten Itemset zum zweiten Testzeitpunkt signifikant bessere Leistungen erzielt. In Itemset 2 lassen sich keine statistisch nachweisbaren Unterschiede in der Leistung zwischen den Messzeitpunkten feststellen. Gruppe B weist in Itemset 2 signifikant bessere Ergebnisse auf, in Itemset 1 sind keine statistisch bedeutenden Unterschiede zu erkennen. Als Ursache für die teilweisen besseren Ergebnisse zum zweiten Testzeitpunkt können mögliche Lerneffekte angenommen werden.

Ergänzend wurden mit nichtparametrischen U-Tests nach Mann und Whitney Itemsets der Gruppen miteinander verglichen. Aufgrund der durchwegs nicht signifikanten Ergebnisse kann angenommen werden, dass die Randomisierung gelungen ist, und es keinen Unterschied in der Leistung der beiden Gruppen gibt. Im Bezug auf die Bearbeitungszeit konnte ausschließlich für den ersten Testzeitpunkt ein signifikanter Einfluss der work-limit Instruktion nachgewiesen werden.

Eine weitere Fragestellung war, ob die power-work-limit Vorgabe höhere korrelative Bezüge bezüglich der Außenkriterien liefert, als die reine power Bedingung. Als Außenkriterien wurde einerseits die durchschnittliche Schulleistung in den Fächern Mathematik, Physik, Chemie, Deutsch, Englisch und Latein herangezogen, andererseits die selbsteingeschätzten Begabungen in Mathematik, Musik, Technik und Raumvorstellung. Ein weiteres Außenkriterium stellte der Planzeichentest dar.

Entgegen den Erwartungen konnten keine signifikanten Zusammenhänge zwischen den durchschnittlichen Leistungen im Unterrichtsfach Mathematik und Raumvorstellung gefunden werden. Bezüglich der Gesamtstichprobe lassen die Ergebnisse darauf schliessen, dass Personen, die ihre durchschnittliche Schulleistung in Physik höher

angeben, auch signifikant bessere Leistungen in der reinen power Vorgabe des EST erzielten, dieses Ergebnis gilt auch, getrennt betrachtet, für Frauen und Männer.

Des Weiteren konnte festgestellt werden, dass Frauen, die bessere durchschnittliche Leistungen in Chemie aufweisen, unter beiden Vorgabebedingungen signifikant höhere Ergebnisse erbrachten. Zwischen den Leistungen in Deutsch, Englisch und Latein konnten erwartungsgemäß keine statistisch bedeutsamen Korrelationen nachgewiesen werden.

Sowohl für die Leistungen unter der reinen power Bedingung als auch für die Ergebnisse unter der power-work-limit Vorgabe konnten korrelative Bezüge zur selbsteingeschätzten Begabung in Mathematik festgestellt werden. Des Weiteren ergaben sich bei den Männern signifikante Zusammenhänge zwischen den Leistungen unter der power-power Vorgabe, wie auch unter der power-work-limit Bedingung des EST und der mathematischen Begabung.

Personen, die sich besser im Bezug auf ihre technische Begabung einschätzen, erzielten sowohl unter der reinen power als auch unter der power-work-limit Bedingung signifikant bessere Ergebnisse. Getrennt nach dem Geschlecht betrachtet, erreichen Frauen und Männer, die ihre technische Begabung höher einschätzten, signifikant höhere Testwerte im EST unter der reinen power-Vorgabe. Bezüglich der selbsteingeschätzte Raumvorstellungsbegabung konnten signifikante Korrelationen für die Leistung der Probanden für die reine power Vorgabe gefunden werden.

Im Bezug auf die Gesamtstichprobe, erzielten Personen die schlechtere Ergebnisse im Planzeichentest lieferten, auch signifikant schlechtere Ergebnisse in der power-power Vorgabe. Auch konnte beobachtet werden, dass Frauen, die im Planzeichentest weniger gut abschnitten, signifikant schlechtere Leistungen in beiden Vorgabebedingungen des EST erbrachten.

Die Retest-Reliabilität fällt in der Gruppe A, mit der Vorgabebedingung power-power zum ersten Testzeitpunkt und power-work-limit zum zweiten Messzeitpunkt, mit r_{tt} = .754 höher aus als in der Gruppe B mit r_{tt} = .611, mit der Vorgabebedingung power-work-limit zum ersten Erhebungszeitpunkt und power-power zum zweiten Testzeitpunkt.

Zudem wurde der Frage nachgegangen, ob Unterschiede in der Raumvorstellung in Abhängigkeit vom biologischen und psychologischen Geschlecht angenommen werden können.

Im Bezug auf das biologische Geschlecht konnte der aus der Literatur bekannte Geschlechtsunterschied (vgl. Voyer, Voyer & Bryden, 1995; Linn & Petersen, 1985) sowohl für die Leistungen der Versuchspersonen unter der reinen power als auch für die power-work-limit Bedingung bestätigt werden. Männer schneiden in beiden Vorgaben besser ab.

Zwischen dem psychologischen Geschlecht und den beiden Versuchsbedingungen konnte kein signifikanter Zusammenhang gefunden werden. Es konnte nicht bestätigt werden, dass Personen, die aufgrund des Fragebogen zur Persönlichkeitseinschätzung "Selbstbild" als androgyn bezeichnet werden können, besser bei Aufgaben zur Raumvorstellungsfähigkeit abschneiden (vgl. Signorella & Jamison, 1986; Gittler & Adlmann, 2010).

.

Zusammenfassend können folgende Hauptergebnisse aus der vorliegenden Untersuchung gewonnen werden:

- Die work-limit Vorgabe des EST ist mit den strengen Kriterien des Rasch-Modells vereinbar. Auch unter der work-limit Instruktion wird dieselbe latente Dimension erfasst. Die work-limit Bedingung stellt somit eine Vorgabeart dar, dessen Einsatz als legitim angesehen werden kann.
- Durch work-limit Vorgabe, die eine Zeitdruckkomponente enthält, kann im Vergleich zur power Bedingung, die ohne Zeitdruck vorgegeben wird, keine zusätzliche diagnostische relevante Information bezüglich der Leistung der Versuchspersonen gewonnen werden.
- Aus dem Vergleich der Korrelationskoeffizienten ist ersichtlich, dass sich die beiden Vorgabearten nur geringfügig in der Höher der Korrelationen mit den Außenkriterien unterscheiden. Der bedeutendste Unterschied in der Höhe des Korrelationskoeffizienten, findet sich bei den Frauen in der selbsteingeschätzten technischen Begabung. Frauen, die sich bezüglich ihrer technischen Begabung besser einschätzten, weisen ohne Zeitdruckkomponente signifikant bessere Ergebnisse im EST auf, als mit Zeitdruckkomponente. Für Frauen stellt die reine

power Bedingung den besseren Prädiktor zur Vorhersagbarkeit ihrer technischen Begabung dar.

- Männer schneiden in beiden Vorgabebedingungen besser ab als Frauen.
- Zwischen dem psychologischen Geschlecht und den beiden Vorgabebedingungen konnten keine signifikanten Zusammenhänge beobachtet werden.

15 Literaturverzeichnis

Primärliteraturverzeichnis

Abels D. (1961). *K-V-T (Konzentrations- Verlaufs-Test)* (2. Auflage). Göttingen: Hofgrefe.

Adlmann E., & Gittler G. (2010). *Geschlechtsunterschiede in der Raumvorstellung: Zum Einfluss der Geschlechtsrollenidentität.* In F. Petermann & U. Koglin (Hrsg.), 47. Kongress der Deutschen Gesellschaft für Psychologie, 26.-30. September 2010 (S. 242-247). Lengerich: Papst Science Publishers.

Alfermann, D. (2001). *Männlich-Weiblich-Menschlich: Androgynie und die Folgen*. *In Passero, U., & Braun, F. (Hrsg.). Konstruktion von Geschlecht.*(S. 29-50). Herbholzheim: Cantaurus.

Altstötter- Gleich, C. (1996). *Theoriegeleitete Itemkonstruktion und –auswahl. Eine Modifikation des Einsatzes der Repetory-Grid-Technik, dargestellt am Beispiel der Erfassung der Geschlechteridentität* (Psychologie Band.13) .Landau: Empirische Pädagogik.

Amelang, M., Bartussek, D., Stemmler, G., & Hagemann, D. (2011). *Differentielle Psychologie und Persönlichkeitsforschung* (7. vollständig überarbeitete Auflage). Stuttgart: Kohlhammer.

Amthauer, R. (1970). *Intelligenzstrukturtest (I-S-T 70).* Göttingen: Hogrefe.

Anastasi, A., & Urbina,S.(1997). *Psychological Testing*. Upper Saddle River, NJ: Prentice-Hall.

Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika, 38* (1), 123-140.

Arendasy, M. (2000). *Psychometrischer Vergleich computergestützter Vorgabeformen bei Raumvorstellungsaufgaben: Stereoskopisch – dreidimensionale und*

herkömmlich – zweidimensionale Darbietung. Unveröffentliche Dissertation, Universität, Wien.

Asendorf, J.B.(2007). *Psychologie der Persönlichkeit* (4. überarbeitete Auflage). Berlin: Springer.

Baar, A. (2000). *Score und Bearbeitungszeit bei 3DW- Würfelaufgaben unter power sowie work-limit-Bedingung*. Unveröff. Diplomarbeit. Universität Wien. Fakultät für Psychologie.

Backhaus K., Erichson B., Plinke W.& Weiber R. (2003). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung.* (10. neu bearbeitete und erweiterte Auflage). Berlin Heidelberg: Springer.

Bem, S.L. (1974). The measurement of psychological androgyny. *Journal of consulting and Clinical Psychology, 45,* 196-205.

Bortz, J., & Döring, N. (2002). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (3. Auflage). Berlin Heidelberg: Springer-Verlag.

Brickenkamp, R. (1966). *Test d2. Aufmerksamkeits-Belastungs-Test* (2. Aufl.). Göttingen: Hogrefe.

Bühl, A. (2010). *PASW 18 Einführung in die moderne Datenanalyse* (12. aktualisierte Aufl.). München: Pearson Studium.

Carter, C. S., LaRussa M. A., & Bodner, G. M. (1987). A study of two measures of spatial ability as predictors of success in diifferent levels of general chemistry. *Journal of Research in Science Teaching, 24*, 645-657.

Eliot, J. (1987). *Models of psychological space.* New York; Berlin; Heidelberg; London; Paris; Tokyo: Springer Verlag.

Fischer, G.H. (1970). *Ein Beitrag zum Speed-Power-Problem*. In G. Reinert (Hrsg.), Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie, 1973 (S. 389-404). Göttingen: Hogrefe.

Fischer, G.H., & Ponocny-Seliger, E. (1998). *Structural Rasch modeling. Handbook of the usage of LpcM-WiN 1.0.* Groningen: PROGAMMA.

Formann, A.,.& Piswanger, K. (1979). *WMT (Wiener-Matrizen-Test)*. Weinheim: Beltz.

Funke, J., & Vaterrodt-Plünnecke, B. (1998). *Was ist Intelligenz?* München: C. H. Beck.

Gardner, H. (2002). *Intelligenzen. Die Vielfalt des menschlichen Geistes*. (U.Spengler, Übers.). Stuttgart: Klett-Cotta. (Original erschienen 1999: IntelligencesReframed. Multiple Intelligences for the 21st Century)

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

Gittler, G. (2003-2009). *Fragebogen zur Persönlichkeitseinschätzung "Selbstbild"*. Universität Wien: Unveröffentlichte Forschungsversion.

Gittler, G. & Arendasy, M. (2003). Endlosschleifen: Psychometrische Grundlagen des Aufgabentyps E^p. *Diagnostica*, *49(4)*, 164-175.

Guay, R. B. (1977). *Purdue spatial visualization test: Rotations*. West Lafayette; IN,ud8u898fdfdfoiod Purdue Research Foundation.

Guay, R.B. & McDaniel, E. (1977). The relationship between mathematics achivement and spatial abilities among elementary school children; In: *Journal of Research in Mathematics Education (Vol. 8)* pp.211-215.

Hampson, S.(1986). *Sex roles and personality. In. D. J Hargreaves & A. M. Colley (hg.). The psychology of sex-roles* (pp.45-59). London: Harper & Row.

Hassler, M. (1990). *Androgynie.* Göttingen: Verlag für Psychologie, Hogrefe.

Heller, K.A., Gaedike, A., & Weinläder, H. (1985). *Kognitiver Fähigkeitstest KFT 4 - 13 +.* Weinheim: Beltz.

Holling, H., Preckel, F., & Vock, M. (2004). *Intelligenzdiagnostik. Kompendien. Psychologische Diagnostik. Band 6*.Göttingen; Bern; Toronto; Seattle; Oxford; Prag: Hogrefe.

Horn, W. (1963). *Leistungsprüfsystem (LPS)*. Göttingen: Hogrefe.

Hornke, L.F. (1997). Untersuchung von Itembearbeitungszeiten beim computergestützten adaptiven Testen. *Diagnostica, 43* (1), 27-39.

Hummel, M. (2001). *Vergleich zweier Instruktionsformen.* Unveröffentliche Diplomarbeit, Universität, Wien, Institut für Psychologie.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic.

Iseler, A. (1970). *Leistungsgeschwindigkeit und Leistungsgüte. Theoretische Analysen unter besonderer Berücksichtigung von Intelligenzanalysen*. Weinheim [u.a.]: Beltz.

Kubinger, K.D. (2006). *Psychologische Diagnostik. Theorie und Praxis psychologischen Diagnostizierens*. Göttingen: Hogrefe.

Laskowski, A. (2000). *Was den Menschen antreibt: Entstehung und Beeinflussung des Selbstkonzepts.* Frankfurt: Campus.

Lautenbacher S., Güntürkün O., & Hausmann M. (2007). *Gehirn und Geschlecht. Neurowissenschaft des kleinen Unterschieds zwischen Frau und Mann.* Heidelberg: Springer.

Lehman W., & Jüling I. (2002). Raumvorstellungsfähigkeit und mathematische Fähigkeiten -unabhängige Konstrukte oder zwei Seiten der Medaille? *Psychologie in Erziehung und Unterricht, 49*, 31-43.

Lienert, G.A. & Ratz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Psychologie-Verl.-Union.

Linn, M.C., & Petersen, A.C. (1985). Emergence and characterisation of gender differences in spatial abilities: A meta-analysis. *Child Developement*, *56*,1479-1498.

Maccoby, E. E., & Jacklin, C. N. (1974). *The Psychology of Sex Differences*. Stanford University Press.

Maier, P.H. (1994). *Räumliches Vorstellungsvermögen*. Europäische Hochschulschriften, Reihe 6, Psychologie, Band.493.

Maier, P.H. (1999*). Räumliches Vorstellungsvermögen : Ein theoretischer Abriß des Phänomens räumliches Vorstellungsvermögen* (1.Auflage). Donauwörth: Auer Verlag.

Maier, P. H. (1996). Geschlechtsspezifische Differenzen im räumlichen Vorstellungsvermögen. *Psychologie in Erziehung und Unterricht, 43,* 245-265.

Martin-Löf, P. (1973). *Statistika Modeller (Statistical Models): Anteckningar fran seminarier läsåret 1969–1970 (Notes from seminars in the academic year 1969–1970), with the assistance of Rolf Sundberg.* Stockholm University.

Michael, W.B., Zimmerman, W.S., & Guilford, J.P. (1957). The description of spatial-visualization abilities. In *Educational and Psycholocical Measurement,17 ,*185-199

Moosbrugger, H., & Kelava A. (2007). Theorie und Fragebogenkonstruktion. Heidelberg: Springer.

Nährer, W. (1986). *Schnelligkeit und Güte als Dimensionen kognitiver Leistung*. Berlin Heidelberg: Springer.

Nährer (1988). *„Schnelligkeitsangepaßtes Testen“: Testökonomie unter Berücksichtigung der Testzeit.* In K.D. Kubinger (Hrsg.), Moderne Testtheorie. Ein Abriß samt neuesten Beiträgen, 219-236.

Pawlik, K. (1976). *Dimensionen des Verhaltens. Eine Einführung in Methodik und Ergebnisse faktorenanalytischer psychologischer Forschung.* Bern; Stuttgart: Hans Huber.

Peters, M. (2005). Sex differences and the factor of time in solving Vandenberg and Kuse mental rotation problems. *Brain and Cognition*,*57 ,*176-184.

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.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Raven, J. C. (1947). *Advanced progressive matrices (APM)*. London: Lewis & Co.

Rost, D. (1976). *Der Begabungsfaktor „Raumvorstellung“-Theorie und Training.* Dissertation, Universität Hamburg.

SCHEIBLECHNER H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. *Zeitschrift für experimentelle und angewandte Psychologie, 19*, 18-38.

Shepard, R.N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701-703.

Signorella, M. L., & Jamison, W. (1986). Masculinity, femininity, androgyny, and cognitive performance: A meta-analysis. *Psychological Bulletin, 100*, 207-228.

Smith, I. (1964). *Spatial ability*. San Diego: Knapp.

Sorby, S. A, Charlesworth, P.; & Drummer, D. (2006). *Spatial skills and their relationship to performance in chemistry courses*. 12 TH international conference on geometry and graphics, August 6-10, Salvador, Brazil.

Stückrath, H. (1968). *Kind und Raum – Psychologische Voraussetzungen der Raumlehre in der Volksschule* (3. Auflage). München: Kösler-Verlag.

Stumpf, H. & Fay, E. (1983). Schlauchfiguren - *Ein Test zur Beurteilung des räumlichen Vortsellungsvermögens.* Göttingen: Hogrefe.

Titze C., Heil M., & Jansen P. (2008). Gender Differences in the Mental Rotation Test (MRT) Are Not Due to Task Complexity. *Journal of Individual Differences, 29(3)*, 130-133.

Vandenberg, S.G., & Kuse, A.R. (1978). Mental rotations. A group tests of three-dimensional spatial visualization: *Perceptual and Motor Skills*, *47*, 599-604.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities:a meta-analyses and consideration of critical variables. *Psychological Bulletin*, *117*, 250-270.

Wedening, G. (1991). *Differentielle Leistungsveränderung durch Zeitlimitierung*. Unveröffentliche Dissertation, Universität, Graz.

Witkin, H.A., Dyk, R.B., & Faterson, H.F. (1962). *Psychological differentiation*. New York: Wiley.

Internetquellen

http://www.geodz.com/deu/d/kognitive_Karte
(Zugriff am 16.08.2011)

Sekundärliteraturverzeichnis

Broverman, D.M., Broverman, I.K., Vogel, W., Palmer, R.D., & Klaiber, E.L. (1964). The automatization cognitive style and physical development. *Child Development, 35*, 1342-1359.

Epstein, S. (1984). „Entwurf einer integrativen Persönlichkeitstheorie.“ In: S. H. Filipp (Hrs): Selbstkonzept Forschung (2. Auflage). Stuttgart: Klett-Cotta.

El-Koussy, A. A. H. (1935). Visual perception of space. *British Journal of psychology: Monograph Supplement*, *20*, 1-30.

Gittler, G., & Spiel, C. (1985). Vorträge auf der Multiplikatorentagung für Darstellende-Geometrie-Lehrer, März 1985, (nicht veröffentlicht): Wien.

Kerkman D.D, Wise J.C, Harwood E.A (2000). Impossible “mental rotation” problems. A mismeasure of women’s spatial abilities? *Learning and Individual Differererences* , *12,* 253–269.

Jung, C.G. (1971). *Die Beziehung zwischen dem Ich und dem Unbewussten*. Studienausgabe, Olten: Walter.

Lohman, D.F. (1979) *Spatial Ability: Review and Re-analysis of the Correlational Literature*. Stanford University Technical Report 8.

Spearman, C. (1904). “General intelligence”, objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

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

Thurstone, L. L. (1950). Some primary mental abilities in visual thinking. *Psychometric Laboratory Report, 59*, Chicago: University of Chicago.

16 Abbildungsverzeichnis

Ich habe mich bemüht, sämtliche Inhaber der Bildrechte ausfindig zu machen und ihre Zustimmung zur Verwendung der Bilder in dieser Arbeit eingeholt. Sollte dennoch eine Urheberrechtsverletzung bekannt werden, ersuche ich um Meldung bei mir.

17 Tabellenverzeichnis

18 Anhang

18.1 Instruktion zum Planzeichentest

Bitte geben Sie Ihren Probandencode an:…….

Seit wie vielen Jahren leben Sie in Wien?

bis 1 Jahr
1 bis 5 Jahre
mehr als 5 Jahre

Wie schätzen Sie Ihre Ortskenntnisse in Wien auf der folgenden Schulnoten-Skala ein?

sehr gut gut befriedigend genügend nicht genügen

Welche der angeführten Gebäude, Plätze, Kirchen, Bahnhöfe etc. in Wien sind Ihnen von Ihrer Lage her bekannt?

(Neues) Allgemeines Krankenhaus	ja	nein
Westbahnhof	ja	nein
(alter) Südbahnhof	ja	nein
Stadtpark	ja	nein
Universität Wien(Haupt-Uni)	ja	nein
Stephansplatz	ja	nein
Karlskirche	ja	nein
Votivkirche	ja	nein
Haus des Meeres (Flakturm)	ja	nein
Oper	ja	nein
Urania	ja	nein
Praterstern	ja	nein
Stadthalle	ja	nein

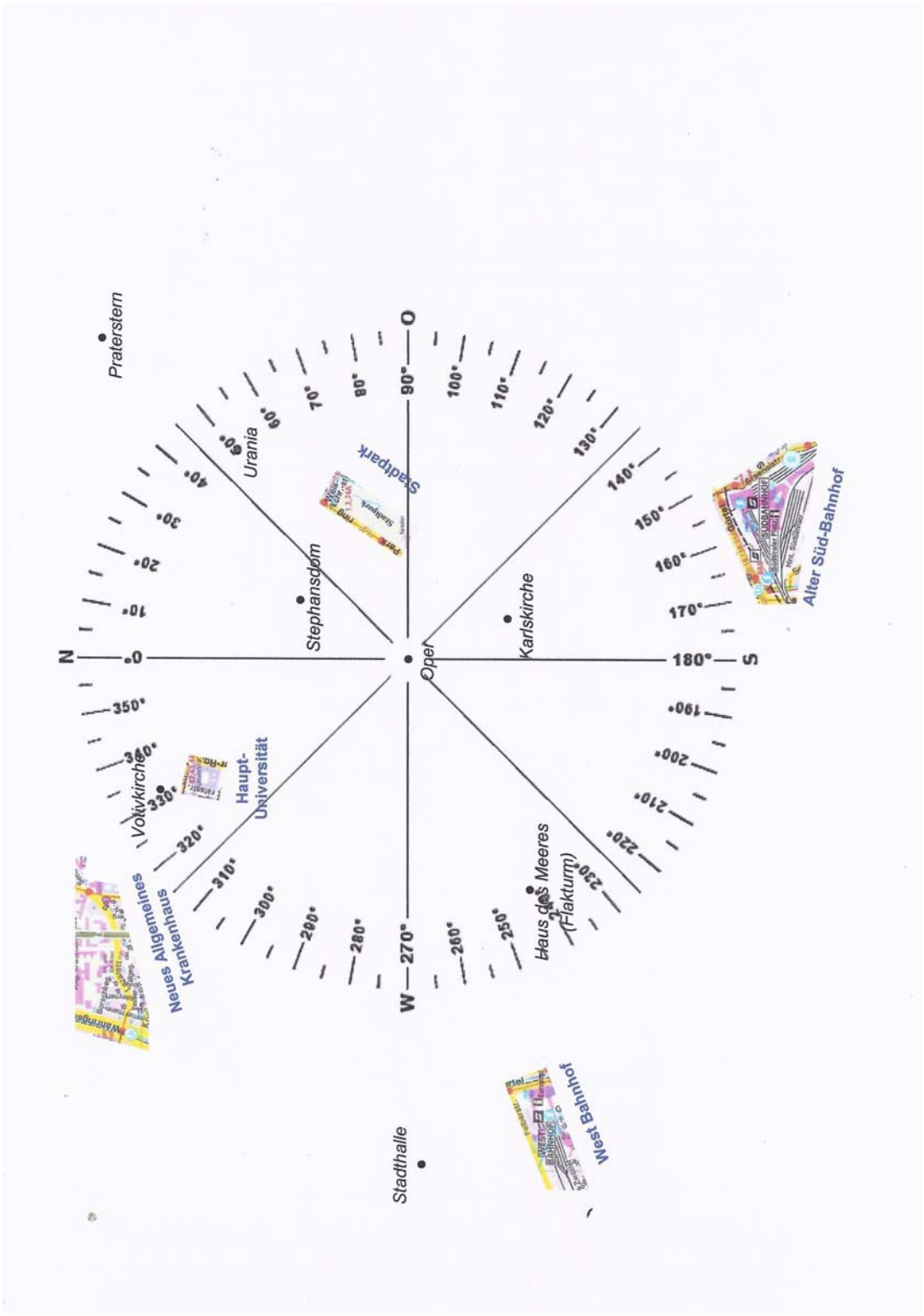
Sie sehen auf dem Blatt, Plätze, Gebäude, Kirchen etc., die Ihnen als Orientierungspunkte dienen. Bitte schätzen Sie nun ein, wo sich folgende Such-Punkte Ihrer Meinung nach befinden und zeichnen Sie diese möglichst genau, mittels eines Punktes ein. Als Beispiel sehen Sie das Haus des Meeres (Flakturm)

Such-Punkte:

- Stephansplatz
- Karlskirche
- Votivkirche
- Oper
- Urania
- Praterstern
- Stadthalle

18.2 Planzeichentest

18.3 Auswertung Planzeichentest



18.4 Abstract - Deutsch

Ziel der vorliegenden Diplomarbeit ist die Validierung des Endlosschleifentests (EST) unter der power sowie work-limit Instruktion. Eine zentrale Fragestellung ist, ob der Endlosschleifentest (EST), welcher das räumliche Vorstellungsvermögen erfasst, und unter der power Bedingung als Rasch-Homogen gilt, auch unter der work-limit Bedingung Eindimensionalität im Sinne von Rasch aufweist. Außerdem wird der Frage nachgegangen, ob die work-limit Vorgabe, die im Gegensatz zur power Vorgabe eine Zeitdruckkomponente enthält, zusätzliche Informationen, bezüglich der Leistung der Versuchspersonen liefert, sowie ob es aufgrund der work-limit Bedingung höhere korrelative Bezüge zu bestimmten Außenkriterien gibt. Die work-limit Bedingung des EST ist mit den strengen Kriterien des Rasch-Modells vereinbar, jedoch konnte durch die work-limit Bedingung im Bezug auf die Leistungen der Versuchspersonen keine zusätzlich diagnostisch relevanten Informationen gewonnen werden. Bezüglich der Außenkriterien, lassen die Ergebnisse darauf schliessen, dass sich die beiden Vorgabebedingungen nur geringfügig in der Höhe der Korrelationen unterscheiden.

Schlüsselwörter: Raumvorstellung, power Instruktion, work-limit Instruktion, Rasch-Homogenität, Kriteriumsvalidität, Geschlechtsrollenidentität

18.5 Abstract

The objective of this thesis is to validate the Endless-loops-test (ger.: Endlosschleifentest – EST; Gittler & Arendasy, 2003) under the power and the work-limit condition. The EST determines the spatial ability of the test person. It has been examined before that under the power condition the EST can be described as Rasch homogeneous. The key question to be answered is whether or not the same holds true under the work-limit condition. Furthermore it was tested if the EST results under the work-limit condition reveal additional information about the performance of the test person. In addition it was examined if higher correlative scores can be found in respect of certain external criteria. The work-limit condition of the EST indeed is Rasch homogeneous – even under the strict criteria of the Rasch model. However the predictive capability of the performance of the test persons under the work-limit condition of the EST did not improve in a meaningful way. Regarding the external criteria the results conclude that the correlation between the power and the work-limit condition differ only in a minor way.

Keywords: spatial ability, "power" Instruction, "work-limit" Instruction, Rasch homogeneity, criteria-oriented validation, gender-role identity

19 Lebenslauf

Persönliche Angaben

Name:	Verena Schön
Geburtsdatum:	05.Jänner 1978 in Wien
Geburtsort:	Wien
Eltern:	Margareta und Manfred Schön

Schulbildung und Ausbildung:

1984-1988	Volksschule Friesgasse
1988-1997	Wirtschaftskundliches Realgymnasium
	Diefenbachgasse, 1150 Wien
seit 1998	Psychologiestudium an der Universität Wien

Praktikum

01.08.2006-17.10.2006	Pflichtpraktikum: Sozialmedizinisches Zentrum Baumgartner Höhe
	Otto Wagner Spital mit Pflegezentrum
	Gerontopsychiatrische Tagesklinik 19/3 III.
	Psychiatrische Abteilung