



universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Modifikation des Itemdesigns bei einem Raumvorstellungstest –
Auswirkungen auf Rasch-Homogenität und Arbeitsmotivation

verfasst von / submitted by

Robert Sahlender, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on
the student record sheet:

UA 066 840

Studienrichtung lt. Studienblatt /
degree programme as it appears on
the student record sheet:

Masterstudium Psychologie

Betreut von / Supervisor:

Ao. Univ.-Prof. i.R. Dr. Georg Gittler

Danksagung

Die letzten Zeilen, die ich im Zuge meiner Masterarbeit schreibe, sollen zugleich am Beginn dieser stehen. Sie richten sich an alle Menschen, die mir im Rahmen dieser Abschlussarbeit und im gesamten Studium zur Seite gestanden sind. Ohne sie wären die folgenden Seiten nie entstanden.

Ein großer Dank gilt Herrn Ao. Univ.-Prof. i.R. Dr. Georg Gittler. Für die Unterstützung bei der Themenauswahl, bei der Umsetzung der Testung, der Bereitstellung der notwendigen Materialien sowie der inhaltlichen und fachlichen Betreuung dieser Masterarbeit. Danke für Ihren wertvollen Input und die angenehme und verständnisvolle Betreuung. Nicht zuletzt auch für die Geduld und das Vertrauen, dass diese Arbeit doch noch ein Ende findet.

Ebenso möchte ich mich bei Tobias Alfes, MSc bedanken, welcher ebenfalls maßgeblich an der Erstellung der Testung beteiligt war und immer mit Rat und Tat zur Seite stand.

Weiters danken möchte ich meinen Eltern, die es mir ermöglicht haben, meinen Interessen zu folgen und meinen eigenen Weg zu gehen. Danke für eure Unterstützung in allen Lebenslagen.

Meine Partnerin Lisa hat mich in allen aber besonders in den schwierigen Phasen unterstützt und ermutigt. Danke für deinen Input, deine Hilfe und die vielen Momente, in denen du den Unterschied ausgemacht hast. Ohne dich wäre ich nicht fertig geworden und das Leben ein großes Stück grauer.

Ein weiterer Dank gilt allen Personen, die mich im Rahmen meines Studiums begleitet haben, die mit mir gelernt und gefeiert haben. Ihr habt diese Reise unvergesslich gemacht.

Zu guter Letzt möchte ich mich bei den vielen Testpersonen bedanken, ohne die diese Arbeit nicht möglich gewesen wäre. Eure investierte Zeit war der Grundbaustein für diese Abschlussarbeit.

Ein Studium ist kein Sprint, sondern ein Marathon. Man bereitet sich darauf vor und dennoch sind die letzten Kilometer hart. Die Momente auf der Zielgerade waren aber alle Anstrengung wert.

Vielen Dank.

Inhaltsverzeichnis

Einleitung	6
Theoretischer Teil	9
1. Raumvorstellung	9
1.1 Definition von Raumvorstellung	9
1.2 Anwendungsgebiete und Relevanz	10
1.3 Raumvorstellung als Bestandteil klassischer Intelligenztheorien	11
1.3.1 Zwei-Faktoren-Theorie	11
1.3.2 Primärfaktorenmodell.....	12
1.3.3 Gruppenfaktormodell	12
1.3.4 Modell der fluiden und kristallinen Intelligenz	13
1.3.5 Intelligenzstrukturmodell	13
1.3.6 Berliner Intelligenzstrukturmodell	13
1.3.7 Drei-Schichten-Modell.....	14
1.3.8 Theorie der multiplen Intelligenzen	14
1.4 Subfaktoren der Raumvorstellung.....	15
1.5 Geschlechtsunterschiede	16
1.5.1 Ursachen für Geschlechtsunterschiede.....	17
1.6 Raumvorstellungsaufgaben und Bearbeitungsstrategien.....	20
2. Der Dreidimensionale Würfeltest	21
2.1 Entwicklung des 3DW	22
2.2 Testgütekriterien.....	23
2.3 Anwendung in der Praxis	24
2.4 Beschreibung des Tests	24
2.5 Auswertung	25
2.6 Bearbeitungszeit und Testverkürzung	26
2.7 Adaptionen	28
2.8 Neue Entwicklungen	29
2.9 Dreidimensionaler Würfeltest- π (3DW- π)	29
Empirischer Teil.....	32
3. Ziele und Fragestellungen	32
3.1 Rasch-Homogenität	33
3.2 Designeffekt	33
3.3 Arbeitsmotivation.....	34
3.4 Nebenfragestellungen.....	34

4. Methode.....	35
4.1 Stichprobe.....	35
4.2 Untersuchungsablauf.....	36
4.3 Testdesign.....	37
4.4 Messinstrumente.....	38
4.4.1 Der Dreidimensionale Würfeltest (3DW)	38
4.4.2 Der Dreidimensionale Würfeltest- π (3DW- π).....	39
4.4.3 Arbeitsmotivationsfragebogen	39
4.5 Statistische Analyse.....	40
4.5.1 Das dichotom logistische Modell von Rasch	40
4.5.2 Modellkontrolle	42
4.5.3 Das linear-logistische Testmodell (LLTM).....	43
4.6 Datenauswertung.....	43
4.6.1 Rasch-Homogenität	44
4.6.2 Designeffekt	45
4.6.3 Arbeitsmotivation.....	45
4.6.4 Nebenfragestellungen.....	46
5. Ergebnisse	46
5.1 Rasch-Homogenität	46
5.2 Designeffekt	48
5.3 Arbeitsmotivation.....	49
5.4 Nebenfragestellungen.....	51
6. Diskussion	52
6.1 Interpretation	53
6.1.1 Rasch-Homogenität.....	53
6.1.2 Designeffekt	53
6.1.3 Arbeitsmotivation.....	54
6.1.4 Nebenfragestellungen.....	54
6.2 Ausblick	56
Literaturverzeichnis.....	58
Abbildungsverzeichnis.....	68
Tabellenverzeichnis.....	68
Abkürzungsverzeichnis.....	68
Anhang	70
Anhang A	70
Anhang B.....	72

Anhang C.....	74
Anhang D	75
Anhang E.....	77

Einleitung

Das räumliche Vorstellungsvermögen, kurz Raumvorstellung (RV), ist eine der faszinierendsten Fähigkeiten von Menschen und anderen Lebewesen. Sie ermöglicht räumliche Strukturen und Relationen in Gedanken abzubilden. Sie ist so grundlegend für den Alltag, dass sie Menschen oft gar nicht bewusst ist. Sie wird benötigt, wenn wir nachts im Dunkeln ins Bad gehen, ohne uns dabei die Füße zu stoßen. Wir brauchen Raumvorstellung, wenn wir durch Menschenansammlungen auf Einkaufsstraßen gehen, versuchen uns dabei zu orientieren und wenn wir die besorgten Einkäufe schließlich im Kofferraum eines Autos stapeln wollen. Ebenso hat sie Feliks Zemdeg gebraucht, als er 2010 im Rahmen des Melbourne Cube Days den Weltrekord im Lösen eines 3x3x3 Zauberwürfels aufgestellt hat. Ohne sein räumliches Vorstellungsvermögen hätte er das kaum in unglaublichen 7.03 Sekunden geschafft.

Raumvorstellung ist für uns Menschen somit eine grundlegende Fähigkeit und dennoch oft im Hintergrund unserer Wahrnehmung. Ein Teil unserer Intelligenz, der unterschiedlich stark ausgeprägt sein kann. Als solcher steht Raumvorstellung daher bereits seit Jahrzehnten im Interesse der Forschung. Zum Beispiel als Bestandteil in den frühen Modellen der Intelligenz (Spearman, 1904; Thurstone, 1938) und in neueren Theorien (Carroll, 1993; Gardner, 2011). Die wichtigsten Theorien und Modelle der Intelligenz und der Platz der Raumvorstellung in diesen werden in Kapitel 1 näher vorgestellt und erläutert. In diesem wird auch das Konstrukt Raumvorstellung noch einmal näher erläutert und diskutiert. Denn obwohl die Raumvorstellung bereits lange erforscht wird, wurde bis heute keine einheitliche Definition gefunden, und immer wieder werden auch ihre Komponenten diskutiert (Linn & Petersen, 1985).

Gemessen wird Raumvorstellung mittels eigens konstruierter Tests wie dem Dreidimensionalen Würfeltest (3DW) von Gittler (1990). Er gibt nicht nur Aufschluss darüber, wie gut eine Person in dieser Fähigkeit abschneidet, sondern dient durch Normvergleiche auch als Instrument zur Eignungs- und Auswahl Diagnostik. Aufgrund dieser Relevanz und seiner guten testtheoretischen Fundierung, findet der Test besonders im deutschsprachigen Raum vielfach Anwendung. Er misst eindimensional im Sinne des dichotomen logistischen Modells von Rasch (1960) und verfügt über einen robusten Itempool, der auch Anpassungen gegenüber stabil zu sein scheint. So wurde in der Vergangenheit beispielsweise die Umstellung von Paper-Pencil- auf eine PC-Version durchgeführt (Arrer,

1992) oder das Hinzufügen einer farblichen Komponente getestet (Gittler & Arendasy, 2003). Eine weitere Veränderung des 3DW ist auch Gegenstand dieser Arbeit.

Grundsätzlich wird zwischen einer intermedialen Änderung (wie zum Beispiel von Paper-Pencil- auf PC-Version) oder einer intramedialen Änderung (Änderung innerhalb eines Mediums) unterschieden (Gittler & Arendasy, 2003). In der vorliegenden Arbeit handelt es sich um eine intramediale Änderung, die untersucht wird. Konkret geht es um eine Anpassung des Itemdesigns im 3DW und um einen neuen Interaktionsmechanismus bei der Bearbeitung der Aufgaben. Die Überlegungen, die dieser Änderung vorausgehen, sind vielfältig. Angefangen vom Sicherheitsaspekt des Tests über bessere Informationsgewinnung bis hin zur Attraktivität des Designs für Testpersonen lassen sich viele Gründe für eine Neugestaltung finden. Diese Aspekte sollen mittels des neuen Designs und Interaktionsmechanismus aufgegriffen und verbessert werden. Die Gründe für die Neugestaltung sowie die genaue Beschreibung des 3DW und des neuen Testdesigns finden sich in Kapitel 2 dieser Arbeit, inklusive einer ausführlichen Beschreibung des 3DW.

Bei Adaptionen und Anpassungen bestehender Tests muss die Messäquivalenz überprüft werden. Dabei wird der Frage nachgegangen, ob zwei Testversionen gleichwertig messen oder Unterschiede vorliegen. Mögliche Unterschiede könnten vorliegen, wenn das Testdesign einen Einfluss auf die Schwierigkeit hat, die Testpersonen verwirrt sind oder neue Strategien anwenden. Es geht in dieser Arbeit daher um die Überprüfung der Messäquivalenz der verwendeten Testversionen. Nicht alle Aspekte werden behandelt, zentral ist die Überprüfung der Rasch-Homogenität der Testitems sowie der mögliche Einfluss des Testdesigns. Darüber hinaus sollen mögliche Unterschiede in der Motivation der Testpersonen aufgedeckt werden. Wenn die Items des verwendeten Tests homogen nach Rasch sind, kann die Prüfung der Messäquivalenz unabhängig von der verwendeten Stichprobe erfolgen, indem nur die Itemparameter verglichen werden. Die Geltung des Rasch Modells ist wichtig für die Einsetzbarkeit des neuen Testdesigns in der Praxis, das Wissen um einen möglichen Designeffekt relevant für die Anpassung der Normwerttabellen. Die Motivation der Testpersonen gibt Aufschlüsse darüber, wie das neue Design in der Praxis ankommt. In Nebenfragestellungen werden der Gesamtscore und die Testdauer genauer betrachtet, um etwaige weitere Unterschiede in den Testdesigns aufzudecken.

Die zentrale Fragestellung dieser Arbeit ist, ob die Änderung des Itemdesigns bei einem Raumvorstellungstest, dem 3DW, Auswirkungen auf die Rasch-Homogenität und Arbeitsmotivation hat. Ziel ist es, erste Ergebnisse über das neue Design zu liefern, Hinweise

auf die Motivation der Testpersonen zu finden und einen Ausblick auf weitere Untersuchungen zu geben. Die einzelnen Fragestellungen werden in Kapitel 3 aufgelistet und beschrieben.

In Kapitel 4 folgt die Methode der Untersuchung. Die Stichprobe wird beschrieben, das Testdesign und die Messinstrumente vorgestellt sowie ein Einblick in den Ablauf der Untersuchung und der statistischen Analyse gegeben. Anschließend folgt in Kapitel 5 die Darstellung der Ergebnisse der einzelnen Forschungsfragen. Kapitel 6 beginnt mit einer kurzen Zusammenfassung und diskutiert die Ergebnisse. Die Arbeit schließt mit einem Ausblick auf zukünftige Forschungen.

Theoretischer Teil

1. Raumvorstellung

1.1 Definition von Raumvorstellung

Obwohl sich die Fachwelt darin einig ist, wie wichtig und zentral die Fähigkeit zur Raumvorstellung für Menschen ist, so herrscht große Uneinigkeit beim Finden einer gemeinsamen Definition (Eliot & Smith, 1983). Das liegt auch daran, dass der Begriff Raumvorstellung ein recht großes Konstrukt erfasst, welches genaugenommen aus mehreren Aspekten besteht (Hegarty & Waller, 2005). Dennoch finden sich in der Literatur einige Definitionen, die gut beschreiben, was Raumvorstellung bedeutet. Einige der gängigsten davon sind in Folge beispielhaft angeführt.

„Das räumliche Vorstellungsvermögen ist ein bedeutsamer Intelligenzfaktor und eine zentrale Fähigkeit, die unsere Wahrnehmung... von unserer Umwelt und damit die Art und Weise der Interaktion mit ihr nachhaltig beeinflusst.“ (Maier, 1994, S.415)

„The ability to generate, retain and manipulate abstract visual images. At the most basic level, spatial thinking requires the ability to encode, remember, transform and match spatial stimuli.“ (Lohman, 1979, S.126)

„Some scholars describe spatial ability broadly in terms of individual differences in the processing of non-linguistic information, while others describe it narrowly in terms of individual differences in performance on spatial tests.“ (Eliot & Smith, 1983, S.1)

„Spatial visualization refers to skill in representing, transforming, generating and recalling symbolic, nonlinguistic information.“ (Linn & Petersen, 1985, S.1482)

„Spatial and other visual perceptual abilities have to do with individuals abilities in searching the visual field, apprehending forms, shapes and positions of objects as visually perceived, forming mental representations of those forms, shapes and positions and manipulating such representations.“ (Carroll, 1993, S.304)

Gemeinsam ist diesen Definitionen, dass Raumvorstellung unabdingbar für Menschen ist und die geistige Manipulation von räumlichen Sachverhalten beinhaltet. Im Folgenden wird besonders darauf eingegangen, warum sie für uns Menschen so wichtig ist und in welchen Bereichen sie Anwendung findet.

1.2 Anwendungsgebiete und Relevanz

Raumvorstellung erfüllt eine zentrale Funktion im Leben von Menschen. Sie ist ein wesentlicher Bestandteil der menschlichen Intelligenz und beeinflusst maßgeblich die Art und Weise, wie Menschen mit ihrer Umwelt interagieren (Maier, 1994). Sie hilft ihnen, sich in ihrer Umgebung zurechtzufinden, Hindernisse zu überwinden, räumliche Verbindungen herzustellen und sich physisch wie gedanklich mit räumlichen Anordnungen auseinanderzusetzen (Gittler, 1994).

Eine besonders hohe Ausprägung in der Fähigkeit Raumvorstellung ist auch eine Voraussetzung in vielen Berufen. Zum Beispiel müssen Pilot*innen, Architekt*innen oder Sportler*innen außerordentliche Leistungen in dieser Dimension erbringen, um ihren Job möglichst erfolgreich auszuüben (Baenninger & Newcombe, 1989). Es ist schwierig eine genaue oder auch nur ungefähre Anzahl von Berufen zu nennen, die Raumvorstellung voraussetzen oder stark ausgeprägt benötigen. Die aktuelle Webseite des österreichischen Arbeitsmarktservices (AMS, 2023) sieht Raumvorstellung in über 30 Berufen als nötige Voraussetzung. Diese Darstellung ist jedoch stark verkürzt. Grob geschätzt müssen es hunderte oder tausende von Berufen sein, die Raumvorstellung als wichtigen Faktor beinhalten. Eine realistischere Zahl nennt Müller (1986) als er schlussfolgert, dass Raumvorstellung in über 3000 Berufen benötigt wird und daher bereits in der Ausbildung eine wichtige Rolle spielen sollte.

Besonders gut erforscht ist das im Bereich der so genannten MINT-Fächer. Gemeint sind damit die Bereiche Mathematik, Informatik, Naturwissenschaft und Technik. Im englischen wird die Bezeichnung STEM verwendet, die für Science, Technology, Engineering und Mathematics steht. Diesen Bereichen wird eine besonders hohe Bedeutung am Arbeitsmarkt und in der Wirtschaft nachgesagt. Ein Zusammenhang mit Raumvorstellung besteht darin, dass festgestellt werden konnte, dass Studierende der MINT-Fächer bessere räumliche Fähigkeiten haben als Studierende anderer Fachrichtungen (Moè et al., 2018). Außerdem korrelieren die räumlichen Fähigkeiten mit dem Studienerfolg in diversen Fächern. Zum Beispiel bei Mathematik Studierenden (Wei et al., 2012), Medizin Studierenden (Guillot et al., 2007) oder Studierenden der Ingenieurwissenschaften (Berkowitz & Stern, 2018). Lubinski (2010) ist der Ansicht, dass Raumvorstellung nicht nur ein guter Prädiktor für Erfolg in diesen Bereichen ist, sondern deswegen auch ein wesentlicher Faktor in der Talentfindung.

Auch das gezielte Training der Raumvorstellung steht in Zusammenhang mit Erfolg im beruflichen Kontext. Provo und Kolleg*innen fanden (2002) einen positiven Effekt bei

Veterinärmediziner*innen, nachdem diese ihre mentale Rotation trainiert haben. Kalun und Kolleg*innen (2020) untersuchten diesen Effekt im Bereich der Medizin bei Chirurg*innen und fanden ebenfalls einen positiven Zusammenhang. Letzterer Eindruck wurde auch bereits 2005 von Brandt und Wright angedeutet, als sie räumliche Fähigkeiten als Prädiktor für die Berufswahl und Ausbildung von Mediziner*innen untersuchten. Auch in der Ausbildung von Pilot*innen finden sich positive Effekte im Zusammenhang mit einem Training der Raumvorstellung (Maroco & Bártoło-Ribeiro, 2013). Es lohnt sich auch ein Blick in den Sport, wo festgestellt werden konnte, dass Raumvorstellung nicht nur ein hilfreicher Faktor bei der Ausübung ist, sondern auch durch den Sport selbst trainiert wird (Ozel et al., 2002).

Das Konstrukt Raumvorstellung steht somit nach wie vor im Interesse der Forschung, wird oft diskutiert und bringt jedes Jahr eine große Anzahl an Publikationen hervor. Nicht zuletzt, da das Konstrukt vor allem in der Praxis Relevanz besitzt (Gittler & Fischer, 2011). Bereits seit langem wird diskutiert, ob und wie Raumvorstellung Teil der menschlichen Intelligenz ist. Im Verlauf der Jahrzehnte findet sie sich daher in zahlreichen Modellen und Theorien wieder, die versuchen, die menschliche Intelligenz zu beschreiben. Außerdem wurden zahlreiche Verfahren und Tests entwickelt, um sie zu messen (Eliot & Smith, 1983).

1.3 Raumvorstellung als Bestandteil klassischer Intelligenztheorien

1.3.1 Zwei-Faktoren-Theorie

Charles Spearman (1904, 1927) geht in seiner Theorie davon aus, dass die menschliche Intelligenz aus einem *Zentralfaktor* (*g*) sowie aus mehreren Subfaktoren besteht, den *spezifischen Faktoren* (*s*). Die Grundlage seiner Idee basiert auf einer Faktorenanalyse. Eine Leistung in einer speziellen Aufgabe hängt immer von der Kombination dieser beiden Faktoren ab, nämlich vom allgemeinen Faktor für Intelligenz sowie einem oder mehreren spezifischen Faktoren. Einer dieser spezifischen Faktoren ist Raumvorstellung (Maltby et al., 2011).

Nach dieser Theorie ergibt sich die Korrelation diverser Leistungstests daher, dass sie alle einen gewissen Anteil des *g*-Faktors messen. Der Rest der Varianz wird durch die spezifischen Faktoren erklärt, die miteinander nicht korrelieren sollten. Obwohl die Zwei-Faktoren-Theorie später mehrfach kritisiert wurde, war sie Vorbild für viele Verfahren, deren Ziel es ist, einen allgemeinen Faktor für Intelligenz zu extrahieren. Viele der heutigen Tests zur Messung der Intelligenz basieren darauf. Zudem bildete es die Grundlage für weitere Forschungen. Eine zentrale Intelligenz, wie der *g*-Faktor, ist dabei fast immer ein Teil davon (Asendorpf, 2004).

1.3.2 Primärfaktorenmodell

Thurstone (1938) postulierte in seinem Primärfaktorenmodell, dass es mehrere unterschiedliche Intelligenzfaktoren gibt. Beim Lösen einer Aufgabe kommen sie jeweils in einem unterschiedlichen Verhältnis zu tragen. Die Grundlage für seine Erkenntnisse war eine Faktorenanalyse. Mittels Einfachstrukturkriterium fand er in seiner Faktorenanalyse neun Faktoren, die er als Primärfaktoren bezeichnete. In späteren Untersuchungen wurden zwei davon gekürzt, sodass heute nur noch von sieben Primärfaktoren gesprochen wird (Thurstone & Thurstone, 1941).

Laut ihm war der von Spearman gefundene g-Faktor das Ergebnis der zugrunde liegenden Primärfaktoren, die wie unterschiedliche Intelligenzen zu interpretieren sind. Einer davon ist räumliches Vorstellungsvermögen. Die sieben Faktoren, die Thurstone vorschlug, lauten (Maltby et al., 2011):

- *Assoziatives Gedächtnis*
- *Rechenfähigkeit*
- *Wahrnehmungs-/Auffassungsgeschwindigkeit*
- *Schlussfolgerndes Denken*
- *Räumliches Vorstellungsvermögen*
- *Sprachbeherrschung*
- *Wortflüssigkeit*

1.3.3 Gruppenfaktormodell

Vernon (1950) wählte eine hierarchische Form für sein Modell. Er nahm an, dass es verschiedene Klassen von Intelligenz gibt, denen ein gemeinsamer Faktor zugrunde liegt. Die verschiedenen Faktoren korrelieren aufgrund ihrer Zugehörigkeit zu diesem Generalfaktor auch miteinander. Das Modell umfasst vier Ebenen. Unter dem Generalfaktor finden sich zwei Gruppenfaktoren, denen jeweils weitere untergeordnete Gruppenfaktoren zugeordnet sind. In der untersten Ebene gibt es spezifische Intelligenzfaktoren. Diese sind aus den darüberliegenden Gruppenfaktoren und den untergeordneten Gruppenfaktoren abgeleitet. Räumliches Vorstellungsvermögen wird nach dieser Theorie vom Gruppenfaktor für räumliche und mechanische Fähigkeiten und einem darunterliegenden Gruppenfaktor für räumliche Fähigkeiten abgeleitet (Maltby et al., 2011).

1.3.4 Modell der fluiden und kristallinen Intelligenz

Cattell (1963) stellte eine Vereinigung von Spearman's und Thurstone's Modellen vor. Laut ihm gibt es einen zentralen g-Faktor, der aus zwei Komponenten besteht: der *fluiden Intelligenz* und der *kristallinen Intelligenz*. Diesen beiden verdankt die Theorie auch ihren Namen. Fluide Intelligenz beschreibt die Fähigkeit der Problemlösung, ohne auf Erfahrung zurückgreifen zu können. Sie ist eine angeborene Fähigkeit, die zur Leistung befähigt. Die kristalline Intelligenz hingegen umfasst alle Fähigkeiten, die durch Lernen erworben wurden. Daher sollte die kristalline Intelligenz im Alter ansteigen, während die fluide Intelligenz abnimmt (Maltby et al., 2011). Raumvorstellung wäre in dieser Theorie als Teil der fluiden Intelligenz zu sehen.

1.3.5 Intelligenzstrukturmodell

Guilford (1967) war ein Kritiker der klassischen Modelle von Spearman und Thurstone und ging nicht von einem zentralen Faktor für Intelligenz aus. Er nannte 120 verschiedene Faktoren für menschliche Intelligenz. In seinem Intelligenzstrukturmodell beschreibt er verschiedene Leistungsprozesse über drei Variablen: *Input-* und *Outputvariablen* sowie *Operationsvariablen*. Insgesamt gibt es vier Inhalte, fünf Operationen und sechs Produkte. Daraus ergibt sich eine Anzahl von 120 verschiedenen Faktoren der menschlichen Intelligenz. Einer der Inhalte war *figural*, mit dessen Beteiligung sich auch Aspekte der Raumvorstellung beschreiben lassen. Guilfords Modell beschreibt das Zusammenwirken vieler verschiedener Aspekte von Intelligenz, ist insgesamt aber sehr komplex und daher als Theorie eher ungeeignet (Maltby et al., 2011).

1.3.6 Berliner Intelligenzstrukturmodell

Das Berliner Intelligenzstrukturmodell stammt von Adolf Otto Jäger (1984). Es versucht die Modelle von Spearman, Thurstone und Guilford zu integrieren und mithilfe neuerer Erkenntnisse zu einem neuen Modell zusammenzufassen (Jäger et al., 1997). Das Berliner Intelligenzstrukturmodell ist hierarchisch und bimodal. Auch hier gibt es einen Generalfaktor, der die allgemeine Intelligenz beschreibt, der sich aus Inhalten und Operationen zusammensetzt. Die vier Operationsklassen sind: *Bearbeitungsgeschwindigkeit*, *Gedächtnis*, *Einfallsreichtum* und *Verarbeitungskapazität* und die drei Inhaltsklassen: *figural-bildhaft*, *verbal* und *numerisch* (Maltby et al., 2011). Raumvorstellung wird in diesem Modell durch die Inhaltsklasse *figural-bildhaft* im Zusammenspiel mit den Operationsklassen beschrieben.

1.3.7 Drei-Schichten-Modell

Das Drei-Schichten-Modell der Intelligenz wurde von Carroll (1993) mithilfe einer Faktorenanalyse entwickelt. In einem hierarchischen Modell aus drei Ebenen geht er von einem allgemeinen Intelligenzfaktor aus. Auf einer zweiten Ebene befinden sich acht breitere Intelligenzfaktoren: *fluide Intelligenz*, *kristalline Intelligenz*, *allgemeine Gedächtnisfähigkeit*, *visuelle Wahrnehmung*, *auditive Wahrnehmung*, *Abruffähigkeit*, *kognitive Geschwindigkeit* und *Verarbeitungsgeschwindigkeit*. Auf der dritten Ebene finden sich 69 spezifische Fähigkeiten. Sein Modell kombiniert die Zweifaktorentheorie von Spearman, das Primärfaktorenmodell von Thurstone, das Modell der fluiden und kristallinen Intelligenz von Cattell sowie das hierarchische Modell von Vernon (Maltby et al., 2011).

1.3.8 Theorie der multiplen Intelligenzen

Eine der moderneren Theorien ist das Modell der multiplen Intelligenzen von Howard Gardner (2011). Entgegen seinen Vorgängern ist er der Ansicht, dass voneinander unabhängige eigenständige Intelligenzen definiert werden können. Diese können unterschiedlich stark ausgeprägt sein und sind sogar trainierbar (Funke & Vaterrodt-Plünnecke, 2004).

Gardner formulierte ursprünglich sieben Arten von Intelligenz, ehe er die Liste später noch um zwei weitere Arten ergänzte. Die insgesamt neun Intelligenzen lauten (Maltby et al., 2011):

- *Sprachlich-linguistische Intelligenz*
- *Logisch-mathematische Intelligenz*
- *Bildlich-räumliche Intelligenz*
- *Körperlich-kinästhetische Intelligenz*
- *Interpersonale Intelligenz*
- *Intrapersonale Intelligenz*
- *Naturalistische Intelligenz*
- *Existenzialistische Intelligenz*

Howard Gardner geht in seiner Theorie davon aus, dass die verschiedenen Intelligenzen unabhängig voneinander existieren und arbeiten sowie kein zentraler Intelligenzfaktor dahinter liegt. Eine Messung gestaltet sich schwierig und die Theorie wird bis heute als nicht ausreichend evaluiert angesehen (Maltby et al., 2011). Das räumliche

Vorstellungsvermögen ist in Howard Gardners Theorie im Einflussbereich der bildlich-räumlichen Intelligenz zu finden.

1.4 Subfaktoren der Raumvorstellung

Wie bei der Definition gibt es auch über die Zusammensetzung der Fähigkeit Raumvorstellung viele Meinungen und verschiedene Überlegungen. Die meisten Forscher*innen versuchten die Frage mittels faktorenanalytischer Untersuchungen zu beantworten. Obwohl sich fast alle einig sind, dass Raumvorstellung aus mehreren Unterfaktoren besteht, so herrscht kein klarer Konsens darüber, welche das sind und wie sie benannt werden sollten. In einer zusammenfassenden Arbeit schreibt Lohmann (1979) von drei Subfaktoren der Raumvorstellung, die immer wieder vorkommen und damit eine breite Mehrheit finden:

- *Spatial Relations*: kommt bei Aufgaben vor, die mentale Rotation erfordern.
- *Spatial Orientation*: die Fähigkeit, die Perspektive wechseln zu können.
- *Visualization*: wird benötigt, um komplexere Aufgaben zu lösen, die meist mehr Zeit erfordern.

Carroll (1993) nennt in einer weiteren Metaanalyse faktorenanalytischer Untersuchungen mehr Faktoren, die seiner Meinung nach zu unterschiedlichen Leistungen in Raumvorstellungstests führen. Die am besten belegten sind:

- *Spatial Relations*: wie bei Lohmann geht es hier um mentale Rotation.
- *Visualization*: ähnlich wie bei Lohmann geht es darum, visuellen Input geistig zu manipulieren, ohne Zeitdruck zu haben.
- *Closure Speed*: die Geschwindigkeit, mit der neue Muster erkannt werden.
- *Flexibility of Closure*: die Geschwindigkeit, mit der bekannte Muster erkannt werden.
- *Perceptual Speed*: die Geschwindigkeit, mit der visuelle Muster erkannt oder verglichen werden.

Die Faktorenanalyse ist eine wirksame Methode, um die Binnenstruktur von Raumvorstellung zu untersuchen und hat in der Vergangenheit zu einigen wichtigen Erkenntnissen geführt. Dennoch lässt sie einen wichtigen Gesichtspunkt außer Acht. Nämlich, dass möglicherweise nicht nur Unterschiede in den Fähigkeiten einer Person zu unterschiedlichen Ergebnissen führen, sondern auch verschiedene Strategien. So überlappen sich die Geltungsbereiche der Fähigkeiten *spatial visualization* und *spatial orientation* beispielsweise in diversen Tests, je nachdem, welche Strategie zur Bearbeitung gewählt wird.

Daher ist die strenge Trennung der beiden nicht möglich (Just & Carpenter, 1985). Unterstützung zu dieser These findet sich auch schon früher von French (1965), der darauf aufmerksam machte, dass verschiedene Strategien sich auf die Faktorenstruktur auswirken. Schultz (1991) ist sogar der Ansicht, dass die Ergebnisse eines Tests mehr die verwendeten Strategien widerspiegeln als die zugrunde liegenden Fähigkeiten.

Eine weitere Metaanalyse von Linn und Petersen (1985) fokussierte sich daher eher auf die Prozesse, die beim Bearbeiten von Aufgaben ablaufen. Sie fanden drei grundlegende Subfaktoren bei Raumvorstellung. Diese sind aufgrund der guten statistischen Fundierung weitgehend anerkannt und somit ein möglicher Konsens in der Diskussion um die Struktur der Raumvorstellung (Leopold et al., 2001).

- *Spatial perception*: beschreibt die räumliche Wahrnehmung. Zum Beispiel die Anordnung von Dingen in einem Raum, den Zusammenhang verschiedener Anordnungen und die Berücksichtigung der eigenen Position bei der Orientierung.
- *Mental rotation*: die Fähigkeit, Dinge im Kopf zu drehen, ohne Hilfsmittel zu verwenden.
- *Spatial visualization*: erfordert die Fähigkeit, aus komplexen räumlichen Informationen Sinn zu machen, indem sie mental bearbeitet werden. Die Vorstellung von Bewegungen und die mentale Umwandlung von Zwei- in Dreidimensionalität fallen unter diese Kategorie. Besonders gefragt ist dieser Faktor vor allem in technischen Berufen.

1.5 Geschlechtsunterschiede

Beim Faktor Raumvorstellung gibt es zahlreiche Belege für die Existenz von Geschlechtsunterschieden. Das Thema gehört zu den am meisten erforschten Aspekten der Raumvorstellung. Ein Großteil der Studien findet einen Geschlechtsunterschied in der Raumvorstellung. Dieser zeigt einen Vorteil für Männer. Die Frage, warum das so ist, beschäftigt die Forschung daher schon lange und ist bis heute nicht eindeutig geklärt. Die Suche danach reicht von evolutionären über biologische bis zu sozialen Ursachen.

Eine der ersten, die Geschlechtsunterschiede bei Raumvorstellung untersuchten, waren Maccoby und Jacklin (1974). Sie schlussfolgerten, dass Männer bessere Leistungen in mathematischen und räumlichen Testaufgaben erbringen als Frauen. Eine vielzitierte Metaanalyse von Voyer und Kolleg*innen (1995) fand die stärksten Unterschiede im oben beschriebenen Subfaktor *mental rotation*. Auch im Bereich *spatial perception* zeigte sich ein

signifikanter Unterschied. Beide fielen zugunsten der Männer aus. Eine weitere Metaanalyse von Maeda und Yoon (2013) fand ebenfalls einen Effekt zugunsten der Männer. Zudem erhöhte sich der Effekt signifikant, wenn Zeitlimits hinzukamen.

Einige Studien fanden bereits Unterschiede im Kindesalter (Moore & Johnson, 2008; Moè, 2018), während andere das Einsetzen der Unterschiede erst nach der Pubertät vermuten (Roberts & Bell, 2000). Moore und Johnson (2020) untersuchten die Grundlage der Unterschiede bei besonders jungen Kindern unter einem Jahr. Sie fanden zwar nicht in allen Fällen Geschlechtsunterschiede, wenn sie welche fanden, jedoch immer zugunsten der Männer. Geiser und Kolleg*innen (2008) fanden einen stabilen Vorteil von Männern im Alter von neun bis 23 Jahren, der sich sogar mit der Zeit vergrößerte. In diesem Zeitraum scheint die größte Entwicklung stattzufinden.

Doch diese Geschlechtsunterschiede sind keineswegs unveränderbar. Zum Beispiel fanden Casey und Kolleg*innen (2008), dass bereits im Vorschulalter durch das Darbieten geometrischer Inhalte eine Steigerung des Raumvorstellungsvermögens bei Mädchen erreicht werden kann. Allerdings scheint dies in gleichem Maße auch für Buben zu gelten, denn Gittler (1994) zeigte, dass entsprechendes Training für beide Geschlechter zu gesteigerten Leistungen in gleichem Ausmaß führte. Moè (2021) startete ein Programm unter Studierenden, indem versucht wurde, die Ergebnisse in mentaler Rotation durch ein besonders motivationsförderndes Programm zu steigern, anstatt durch reines Training. Dieses zeigte gute Ergebnisse und konnte helfen, Unterschiede zu verringern.

1.5.1 Ursachen für Geschlechtsunterschiede

Die Frage, woher die Unterschiede kommen, ist bis heute nicht restlos geklärt und brachte im Laufe der Zeit unterschiedliche Theorien und Erklärungsansätze hervor. Eine bekannte Theorie sucht die Lösung des Rätsels in den Wurzeln der Menschheit. Eals und Silverman (1994) berufen sich in ihrer *Hunter-Gatherer Theory* auf die ursprüngliche Rollenverteilung in der Menschheitsgeschichte. Männer benötigten damals als Jäger eine bessere Raumvorstellung als Frauen, die eher mit dem Sammeln beauftragt waren. Dies würde laut dieser Theorie auch eine verbesserte Wiedererkennungsfähigkeit bei Frauen erklären. Eine Theorie, die zwar nicht besonders plausibel ist, aber zumindest kulturübergreifend Bestand hat. Die Vermutungen wurden auch international bestätigt (Silverman et al., 2007).

Eine wesentlich plausiblere Erklärung sind Unterschiede in den biologischen Voraussetzungen bei Männern und Frauen. Voyer und Kolleg*innen (1995) fanden heraus, dass Mädchen und Buben vor der Pubertät noch keinerlei Unterschiede in der Raumvorstellungsleistung aufweisen. Danach sind diese jedoch deutlich nachweisbar. Das spricht für eine biologische, wahrscheinlich hormonelle, Ursache des Unterschieds. Shirazi und Kolleg*innen (2021) untersuchten den Einfluss der Hormone Progesteron und Estrogen auf die kognitive Entwicklung und den Zusammenhang mit Performance in mentaler Rotation. Sie fanden einen leichten Einfluss von Progesteron, dessen Level die Vorhersage der Performance in einem *mental rotation task* voraussagen konnte. Griksiene und Kolleg*innen (2018) fanden einen Zusammenhang mit hormoneller Empfängnisverhütung. Frauen, die ein östrogenhaltige Pille einnahmen, schnitten schlechter in Raumvorstellung ab als die Vergleichsstichproben. Sie vermuteten allerdings die Verwendung unterschiedlicher Strategien zur Lösung der Aufgaben als Hauptgrund. Peragine und Kolleg*innen (2020) sprachen sich für eine dynamische Betrachtungsweise aus. Da der Hormonhaushalt besonders bei Frauen stark variiert, hat er einen großen Einfluss auf die Performance in mentaler Rotation. Bei Berücksichtigung der verschiedenen zyklischen Phasen, schwindet der Unterschied in mentaler Leistung. Der Hormonstatus im Allgemeinen hat somit mehr Einfluss auf die Performance als das Geschlecht selbst.

Interessanterweise weist beim Bearbeiten von Raumvorstellungsaufgaben die rechte Hemisphäre des Gehirns bei Männern eine vermehrte Aktivität auf (Gur et al., 2000), während Frauen einen höheren Anteil an bilateraler, ausgeglichener, Aktivität zeigen (Howard et al., 1992). Andere Forscher*innen argumentierten den Geschlechtsunterschied mit einer geringeren kognitiven Anstrengung bei Männern. Campbell und Kolleg*innen (2018) wollten dies mittels Pupillenmessung belegen. In einem *mental rotation task* schnitten Männer bei Würfelaufgaben besser ab und zeigten eine geringere kognitive Anstrengung als Frauen. Dieser Effekt verschwand allerdings, wenn soziale Aspekte in den Aufgaben auftauchten. Frauen schnitten besser ab, wenn die Würfel zum Beispiel durch Hände ersetzt wurden (Tarampi et al., 2016). Voyer und Kolleg*innen (2017) untersuchten in einer Meta-Analyse den Einfluss des visuell-räumlichen Arbeitsgedächtnisses auf die räumlichen Fähigkeiten von Männern und Frauen. Sie fanden leichte Vorteile für Männer, die als Erklärung dienen können. Kaufman (2007) wiederum fand Unterschiede im Arbeitsgedächtnis, die zu Unterschieden in den räumlichen Fähigkeiten beitragen. Trotz der großen Anzahl an Befunden können biologische Ursachen allein jedoch nicht den stabilen Effekt des Geschlechtsunterschieds erklären.

Ein weiterer großer Teil der Forschung konzentriert sich bei der Suche nach den Gründen der Unterschiede im räumlichen Vorstellungsvermögen auf den sozialen Aspekt. Eine bereits ältere Studie von Newcombe und Kolleg*innen (1983) nannte die unterschiedliche Erfahrungswelt von Männern und Frauen als Grund dafür. Die Interaktion mit der Umwelt, der Familie und Bezugspersonen als auch Freizeitaktivitäten, sind laut den Autor*innen sehr unterschiedlich und in Folge auch für die unterschiedliche Ausprägung von Fähigkeiten verantwortlich. Auch Etaugh und Riley (1983) fanden diese These plausibel und argumentierten zum Beispiel mit dem unterschiedlichen Spielmaterial von Mädchen und Buben. Auch hier wurde die unterschiedliche Freizeitgestaltung thematisiert, welche die Raumvorstellung bei Buben eher fördert und fordert. Besonders das Spielzeug, welches eher Buben verwenden, fördert die Raumvorstellung besser als Spielzeug, das eher Mädchen verwenden. Dieser Effekt bleibt bis ins Erwachsenenalter bestehen (Cherney & London, 2006). Moè und Kolleg*innen (2018) untersuchten Studierende aus dem Fachbereich STEM. Frauen in diesem Fachbereich hatten in ihrer Kindheit eher mit Spielzeug gespielt, welches die mentale Rotation förderte oder Sportarten ausgeübt, die diesen Faktor trainierten. Dies ist ein weiterer Beleg für den frühen Einfluss der Umgebung auf das spätere Leben und sogar die Berufswahl. Obwohl solche Effekte in der modernen Gesellschaft kleiner werden, gibt es immer noch klassisch männliche und weibliche Kategorien. Diese werden in Spielzeuggeschäften besonders oft deutlich. Umso wichtiger ist es, in der Erfahrungswelt von Kindern auf möglichst unterschiedliche Reize zu setzen, um eine breite Palette an Fähigkeiten zu fördern (Cherney et al., 2003).

Guizzo und Kolleg*innen (2019) legten in ihrer Untersuchung einen besonderen Fokus auf Stereotype. Sie fanden zwar deutliche Effekte zugunsten der Männer in Aufgaben zur mentalen Rotation, konnten diese jedoch verringern, indem sie in einer Bedingung in der Anweisung bewusst darauf hindeuteten, dass keine Geschlechtsunterschiede in Raumvorstellungsaufgaben existieren. Stereotype sind demnach ein wichtiger Faktor in der Entstehung von Geschlechtsunterschieden.

Eine plausible Erklärung für den Geschlechtsunterschied bei Raumvorstellungsaufgaben ist das Zusammenspiel mehrerer Ursachen. Zum Beispiel führt das Üben einer Aufgabe zu einem so genannten *Lateralitätsshift*, einem Wechsel der hemisphärischen Verarbeitung von Inhalten. Der verschiedene Übungsgrad von Aufgaben ist demnach verantwortlich für die andersartige Verarbeitung von Raumvorstellungsaufgaben im Gehirn bei Männern und Frauen. Männer bearbeiten diese eher links- und Frauen rechtshemisphärisch (Voyer et al., 1995). Unterstützt wird diese These von Goldberg und

Costa (1981), die herausfanden, dass neue Aufgaben rechts verarbeitet werden und bereits bekannte links. Mit einer gewissen Übung können sie also die Hemisphäre wechseln. Dasselbe passiert eventuell bei Raumvorstellung. Die dafür benötigte Übung kommt vom sozialen Aspekt des Sachverhalts. Die *Bent Twig* Theorie besagt, dass Buben bereits früh dazu tendieren, Aufgaben zu wählen, die ein gewisses Maß an Raumvorstellung benötigen. Das machen sie aufgrund ihrer angeborenen Affinität dazu (Sherman, 1967). Die Übung wiederum verstärkt den Effekt bis hin zur späteren Berufswahl. Eine Studie von Casey (1996) bestätigte auch die Vermutung über das Zusammenspiel von biologischen und Umweltfaktoren. Frauen mit einer gewissen Prädisposition wurden bei entsprechender Förderung durch die Umwelt deutlich besser in Raumvorstellung. Übung macht demnach den Meister, aber nur mit der richtigen Umgebung.

1.6 Raumvorstellungsaufgaben und Bearbeitungsstrategien

Aufgaben zur Messung des räumlichen Vorstellungsvermögens können auf verschiedene Art und Weise gelöst werden. Barratt (1953) war einer der ersten, der zu diesem Thema etwas publizierte. Er befragte dazu Menschen retrospektiv nach ihren Bearbeitungsstrategien, die bereits Raumvorstellungsaufgaben bearbeitet haben. Eine anschließende Faktorenanalyse ergab einen verbalen Faktor sowie drei weitere Faktoren. Diese entsprachen weitgehend jenen drei Faktoren, die später auch durch die Metaanalyse von Linn und Petersen (1985) bestätigt wurden. Zudem fand Barratt (1953) heraus, dass sich die Strategien in einen *whole approach* und einen *part-approach* teilen lassen. Beim *whole-approach* betrachten Versuchspersonen das Problem als Ganzes, während beim *part-approach* einzelne Teile schrittweise begutachtet werden. Zu einem ähnlichen Schluss kamen auch Kail und Kolleg*innen (1979), die die beiden unterschiedlichen Herangehensweisen als eine *holistische* und eine *iterative* Strategie beschrieben. Bei schwierigeren Aufgaben scheinen Menschen eher iterative, analytische Strategien zu verwenden, während bei leichteren Aufgaben holistische Strategien den Vorzug bekommen (Lohmann, 1979).

Ebenfalls scheint es einen Unterschied zwischen grundsätzlich begabten und unbegabten Menschen zu geben. Begabte überlegen weit mehr und raten weniger als Unbegabte, die oft nach dem Trial und Error Prinzip vorgehen. Lohmann (1979) kritisierte daher den retrospektiven Ansatz von Barratt (1953) in seiner Forschung. Besser wäre es laut ihm Versuchspersonen vorher, währenddessen und nachher zu fragen. Konsens zwischen allen Parteien besteht in der Tatsache, dass die Lösung einer Aufgabe sowohl vom Schwierigkeitsgrad, der Fähigkeit einer Person als auch von den verwendeten

Bearbeitungsstrategien abhängt. Besonders der Einsatz letzterer hat maßgeblichen Einfluss auf die Testleistung (Schultz, 1991).

Putz-Osterloh (1977) analysierte Items der Würfelaufgaben aus dem Intelligenz-Struktur-Test (IST) (Amthauer, 1973). Sie fand dabei heraus, dass es drei unterschiedliche Arten von Würfelaufgaben gibt: *Flächenwürfel*, *Flächenwürfel plus* und *Raumwürfel*. Um einen Flächenwürfel zu lösen, müssen die drei sichtbaren Flächen des Antwortwürfels mit den Flächen der Alternativen überprüft werden. Bei einem Flächenwürfel plus sollte zusätzlich auch die Relation der Flächen zueinander geprüft werden. Bei einem Raumwürfel ist die Fähigkeit zur mentalen Rotation erforderlich, indem sich die Testperson auch das Zeigen und Verschwinden von Würfelflächen gedanklich vorstellen muss (Putz-Osterloh, 1977). In einer Folgestudie wurde auch das Blickverhalten von Proband*innen untersucht. Dabei wurden Raum- und Flächenstrategien gefunden. Raumstrategen wiesen höhere Fixationszeiten auf, Flächenstrategen waren hingegen schneller (Putz-Osterloh & Lüer, 1979). Köller und Kolleg*innen (1994) beschrieben ebenfalls unterschiedliche Strategien zur Lösung einer Würfelaufgabe, geteilt in *analytische* versus *holistische*. Die Wahrscheinlichkeit, eine Würfelaufgabe zu lösen, variierte je nach Art und verwendeter Strategie. Eine andere Methode zur Lösung schlugen Just und Carpenter (1985) vor, nämlich die Verwendung des kognitiven Koordinatensystems. Eine gedankliche Repräsentation, die beim Lösen der Aufgaben hilft.

Aufgrund der hohen Relevanz der Raumvorstellung schafften es zahlreiche Tests auf den Markt, die diese Fähigkeit messen können. Eliot und Smith (1983) präsentierten in ihrer Studie 392 Beispiele aus diversen Tests und unterteilen diese in zwei Kategorien: *recognition division* und *manipulative division*. Erstere erfordern das Erkennen von Vorgaben, zweitere erfordern eine mentale Bearbeitung der Aufgaben. Zwei im deutschsprachigen Raum weit verbreitete Tests sind der Endlos-Schleifen-Test (Stumpf & Fay, 1983) und der Dreidimensionale Würfeltest (Gittler, 1990). Letzterer ist Gegenstand dieser Arbeit und wird daher im Folgenden ausführlicher beschrieben.

2. Der Dreidimensionale Würfeltest

Der Dreidimensionale Würfeltest (3DW) wurde 1978 von Georg Gittler am Institut für Psychologie in der Abteilung für Methodenlehre an der Universität Wien entwickelt und als fertiger Test 1990 veröffentlicht. Er ist im deutschsprachigen Raum einer der am meisten

verwendeten und besterforschten Tests zur Messung des räumlichen Vorstellungsvermögens. Aufgrund seiner zentralen Stellung in dieser Arbeit wird der 3DW in den folgenden Unterkapiteln näher beleuchtet und erklärt.

Der 3DW ist ein Niveau-Power-Test zur Messung des räumlichen Vorstellungsvermögens. Er wird ab einem Alter von 13 Jahren eingesetzt und es existiert keine Altersobergrenze. Er kann für Einzel- und Gruppentestungen eingesetzt werden. Eine zentrale Besonderheit des Tests ist die Gültigkeit des dichotom logistischen Modells von Rasch (1960). Aufgrund seiner guten Messeigenschaften eignet er sich auch für das adaptive Testen.

2.1 Entwicklung des 3DW

Die Grundlage für den 3DW lieferte eine lernpsychologische Arbeit von Gittler (1978). Er benötigte Testaufgaben, die der Rasch-Homogenität entsprachen. Der Intelligenz-Struktur-Test (IST) von Amthauer (1973), genauer gesagt die Würfelaufgaben des Subtest 8, sollten dieser Forderung standhalten. Eine erste Analyse bei 188 Personen stieß auf zwei Probleme: die Speed-Bedingung sowie der Einsatz von zwei unterschiedlichen Strategien zur Lösung. Sowohl die Verifikations- als auch Falsifikationsstrategie wurden zur Lösung der Aufgaben eingesetzt. In Folge entsprachen die Aufgaben nicht den Anforderungen des Rasch Modells. Der 3DW sollte dieses Problem mit ein paar Änderungen beheben und wurde sozusagen als Nachfolgetest geplant. Die Itemanalyse zur Neukonstruktion der Aufgaben war theoriegeleitet, um zu klären, warum die Aufgaben heterogen sind. Nach einer Rasch-Analyse wurden immer engere Hypothesen formuliert, so lange, bis diese erfolgreich war. Die Änderungen sind im Folgenden kurz beschrieben.

Die Anzahl der Antwortalternativen wurde von sechs auf sieben erhöht. Neben sechs unterschiedlichen Antwortwürfeln wurde die Alternative „Kein Würfel richtig“ hinzugefügt. Die Aufstockung der Alternativen auf sieben mögliche richtige Antworten reduzierte die Ratewahrscheinlichkeit auf ein Siebtel (14.29%). Die Alternative „Kein Würfel richtig“ sollte die Falsifikationsstrategie unterbinden und den Testpersonen nahelegen, den richtigen Antwortwürfel zu suchen. Dies war im Subtest 8 des IST noch zu umgehen, indem nach dem Ausschlussprinzip vorgegangen wurde, da einer der richtige Würfel sein musste. Die Alternative „Ich weiß die Lösung nicht“ wurde zusätzlich als Option hinzugefügt. Diese sollte zusätzlich dazu führen, Rateeffekte zu minimieren, indem die Versuchspersonen ermutigt wurden, nicht zu raten, wenn die Antwort nicht gefunden werden kann (Gittler, 1990).

Die größte Schwäche des Subtest 8 des IST war jedoch die zeitliche Komponente. Diese gefährdete die Eindimensionalität stark, weswegen er den Rasch-Analysen nicht standhielt. Die ursprünglich 30 entwickelten Items des neuen 3DW kamen daher als reine Power Testung zur Anwendung. Dennoch scheiterte eine erste Überprüfung auf Rasch-Homogenität. Nachdem zehn Items entfernt wurden, scheiterte eine weitere Analyse. Daher waren weitere Untersuchungen notwendig (Gittler, 1990).

In einer genaueren Analyse wurden die Items schließlich nach drei unterschiedlichen Aufgabentypen klassifiziert. Das Kriterium hierfür war die Anzahl der notwendigen Lösungsschritte, um eine Aufgabe richtig zu lösen. Der erste Itemtyp erforderte dabei nur eine Drehung, der zweite zwei und der dritte drei Drehungen. Die Annahme liegt nahe, zu denken, dass die Aufgaben mit steigender Anzahl an Lösungsschritten komplizierter und damit schwieriger wurden. Jedoch war es in diesem Fall anders und die Aufgaben des zweiten Itemtyps waren die am einfachsten zu lösenden. Sie bestanden aus zwei Teilschritten und waren mittels einer analytischen Strategie einfach zu lösen. Durch die Analyse der Muster an den Seiten konnten sie ohne Zuhilfenahme der mentalen Rotation gelöst werden, weswegen sie auch Flächenwürfel genannt wurden. Diese Items wurden daraufhin aus dem Test entfernt und es gelang zum ersten Mal die Testung auf Rasch-Homogenität. Übrig blieben 17 Testaufgaben plus ein Warm-Up Item, welches der Gewöhnung an die Testaufgaben dient und nicht in die Berechnung miteinfließt. Ein Likelihood-Quotienten-Test, Kreuzvalidierungen sowie der Modelltest nach Martin-Löf (1973) dienten zur Überprüfung der Eindimensionalität (Gittler, 1990). Der 3DW war somit einsatzbereit und wurde in weiteren Analysen auf seine Testgütekriterien geprüft.

2.2 Testgütekriterien

Die Interpretations- sowie Auswertungsobjektivität ist gegeben, da die Summe der richtigen Antworten als Testscore fungiert. Etwas schwieriger zu bestimmen ist die Durchführungsobjektivität, da es durch die Anwesenheit einer*s Testleiter*in zu unerwünschten Verhaltensweisen kommen kann. Bei der PC-Version spielt das keine Rolle. Die Reliabilität hängt grundsätzlich immer auch von den verwendeten Aufgaben und der Heterogenität der Stichprobe ab. In einer repräsentativen Stichprobe von Schüler*innen und Student*innen, für die die meisten Daten vorliegen, lag die Retest-Reliabilität bei $r = .61$. Für die Überprüfung der Validität wurden spezifische Hypothesen aufgestellt, mit welchen anderen Tests der 3DW hoch oder niedrig korrelieren sollte. Die Würfelaufgaben des IST von Amthauer (1973) sowie der Form-Lege-Test von Lienert (1958) korrelierten niedriger mit

dem 3DW als die Schlauchfiguren von Stumpf und Fay (1983) und der Mechanical Comprehension Test von Bennett (1969). Letztere messen die dreidimensionale Raumvorstellung, während erstere nur die zweidimensionale Vorstellungskraft messen. Die Übereinstimmungsvalidität des 3DW und Noten in Darstellender Geometrie beträgt $r = -.37$. Die Vorhersagevalidität in derselben Stichprobe im Abstand von zwei Jahren beträgt $r = -.32$. Ebenso korrelierte die Selbsteinschätzung von Student*innen mit dem 3DW zu $r = .44$ (Gittler, 1990).

Die Ökonomie des Tests ergibt sich aus dessen variabler Gestaltung mit einer Testlänge von 8-18 Aufgaben, in der mehrere Versionen möglich und machbar sind und aus der flexiblen Bearbeitungszeit. Je nach Bedarf können verkürzte Testversionen vorgegeben werden, die trotzdem mit der Originalversion vergleichbar sind. Die praktische Bedeutung ist vor allem durch die Anwendbarkeit in der Diagnostik gegeben (Rost, 1977). Um das Leistungsniveau von Personen zu vergleichen, wurden Normwerttabellen erstellt. Die ursprüngliche Normierung des 3DW beruht auf einer Untersuchung von $N = 5673$ Österreicher*innen, die meisten davon Schüler*innen (4064). Für die letzte Gruppe wurden auch repräsentative Quotenstichproben gezogen (Gittler, 1990).

2.3 Anwendung in der Praxis

Der 3DW lässt sich aufgrund seiner Eigenschaften und Messgenauigkeit sehr vielseitig einsetzen. In der Berufsberatung kann er zum Einsatz kommen, um die Eignung für diverse Berufe zu ermitteln, die ein hohes Maß an Raumvorstellung erfordern. Zum Beispiel bei Pilot*innen. In der Studienberatung kann es interessant sein, angehende Architekt*innen mit dem Test zu konfrontieren. In einem Auswahlverfahren filtert der 3DW in der Personalauswahl diejenigen Kandidat*innen heraus, die die besten Leistungen im Bereich Raumvorstellung liefern. In einer Studie wurden 81 Student*innen (davon 36 Techniker*innen) die Aufgaben des 3DW sowie die Würfelaufgaben aus dem IST vorgegeben. Der 3DW erwies sich als das bessere Instrument zur Diskriminierung zwischen Techniker*innen und Nicht-Techniker*innen (Gittler, 1990). Auch Mappingleistungen (Orientierung in der Stadt) konnten mittels des Tests gefiltert werden (Auer & Pircher, 1988). Ebenso wird der 3DW gerne zu Forschungszwecken eingesetzt, da er testtheoretisch sehr gut fundiert ist.

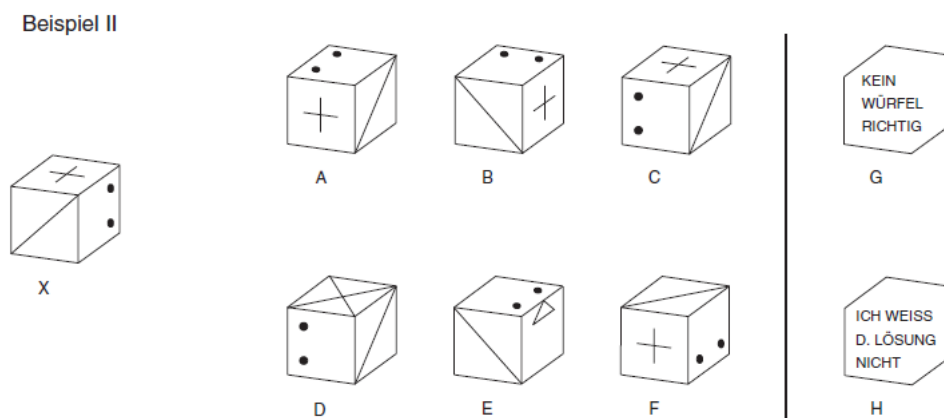
2.4 Beschreibung des Tests

Die Originalversion des 3DW wird als Paper-Pencil Test vorgegeben und besteht aus einem Testheft und einem Antwortbogen. Im Testheft wird auf der linken Seite ein Zielwürfel

X gezeigt, rechts davon befinden sich sechs verschiedene Würfel, die mit den Buchstaben von A bis F nummeriert sind. Alle Würfel haben sechs verschiedene Muster an den Würfelflächen, von denen jeweils nur drei sichtbar sind. Das Ziel der Aufgabe ist, aus den sechs möglichen Antwortwürfeln A bis F denjenigen zu finden, der mit dem Zielwürfel X ident sein kann. Der Antwortwürfel stellt den Zielwürfel in veränderter Lage dar. Er kann gedreht und/oder gekippt worden sein und im Zuge dessen auch ein neues Muster zeigen, welches vorher nicht sichtbar war. Die Aufgabe soll mittels mentaler Rotation gelöst werden, durch gedankliches Drehen und Kippen der Würfel. Der Lösungsbuchstabe muss anschließend am Antwortblatt markiert werden. Zusätzlich besteht die Möglichkeit, die Alternativen „Kein Würfel richtig“ und „Ich weiß die Lösung nicht“ anzukreuzen. Die Alternative „Kein Würfel richtig“ dient dem Ausschluss der Falsifikationsstrategie und die Alternative „Ich weiß die Lösung nicht“ soll Rateeffekte verringern. Pro Aufgabe gibt es immer nur eine richtige Lösung (Gittler, 1990). Ein Beispielitem des 3DW in seiner originalen Paper-Pencil Version findet sich in Abbildung 1 (Gittler, 1999).

Abbildung 1

Beispielitem aus dem 3DW (Paper-Pencil Version)



(Gittler, 1999)

2.5 Auswertung

Die Auswertung des 3DW erfolgt zweikategoriell. Entweder die Antwort ist richtig oder falsch. Der Rohwert (R) eines Tests wird durch die Anzahl der richtig gelösten Aufgaben gebildet. Bei der klassischen Version des 3DW kann dieser maximal 17 betragen. Von den insgesamt 18 vorgegebenen Items fließt das erste als „Aufwärmitem“ nicht in die Bewertung mit ein. Der Rohwert bei diversen Kurzversionen ist daher immer $k-1$, wobei k die Anzahl der vorgegebenen Aufgaben beschreibt. Etwaige Aufwärmitems werden auch bei den

Kurzversionen nicht bewertet. Die Rohwerte der Kurzversionen können mittels Umwandlung mit den Werten des Originaltests verglichen und mittels Tabellen sogar in Personenparameter, t-Werte und Prozentränge umgerechnet werden. Je größer die Personenparameter sind, desto besser ist die Person. Schwierigkeiten gibt es bei extremen Scores wie 0 und 17. Hier ist der Test für die Personen entweder zu schwierig oder zu leicht, sodass keine Aussage über den Personenparameter getroffen werden kann. Er ist entweder kleiner oder größer als der messbare Parameter und kann nicht genau bestimmt werden (Gittler, 1990). Abhilfe schafft hier adaptives Testen.

2.6 Bearbeitungszeit und Testverkürzung

Die Bearbeitungszeit hängt vom individuellen Tempo der Testpersonen ab. Es gibt keine Zeitbeschränkung, da eine reine Powerbedingung eine Bedingung für die Gültigkeit des Rasch Modells ist. Das bedeutet, dass ausschließlich die Leistung gemessen werden soll und die Testdauer dabei keine Rolle spielt. Deswegen ist die Bearbeitungszeit generell schwierig abzuschätzen und kann am besten mit dem Begriff „open end“ beschrieben werden. Die mittlere Bearbeitungszeit liegt bei $M = 22.44$ Minuten mit einer Standardabweichung von $SD = 7.92$ für die Bearbeitung des ganzen Tests mit 18 Aufgaben inklusive dem Warm-Up Item. Dieser Aspekt des Tests ist je nach Anwendungsfall sehr unökonomisch, zum Beispiel im Rahmen einer Batterietestung, in der es nur um eine grobe Abschätzung der Fähigkeiten im Bereich Raumvorstellung geht oder im Rahmen einer Gruppentestung. Die Powerbedingung ist allerdings für die Gültigkeit des Rasch Modells und damit für die zahlreichen Vorteile des Tests unerlässlich (Gittler, 1990).

Jedoch gibt es durch die Rasch-Skalierung der Items eine Möglichkeit, einen Kompromiss zwischen der Ökonomie und der Messgenauigkeit zu finden. Im Rahmen der probabilistischen Testtheorie wurden mehrere Kurzformen des 3DW entwickelt, in denen die Powertestung beibehalten wird und die Leistungen der Testpersonen immer noch auf der latenten Fähigkeitsskala abschätzbar sind. Der 3DW kann auf minimal acht bewertete Aufgaben gekürzt werden, jede Variante von 8 bis 17 Wertungsaufgaben ist dabei möglich, ausgenommen Aufwärmitems. Die Fähigkeits- und Personenparameter wurden im Normierungsdatensatz auch für alle möglichen Kurztestversionen zwischen 8 und 17 Wertungsaufgaben geschätzt und ermöglichen damit eine gemeinsame Verrechnung auf der Fähigkeitsdimension. Bearbeitet eine Person zum Beispiel nur acht Aufgaben, kann eine Rohwertschätzung der Langversion vorgenommen werden. Die Ergebnisse bleiben somit vergleichbar. Natürlich geht eine Testverkürzung mit einer etwas geringeren Messgenauigkeit

einher. Das Minimum an acht Aufgaben wurde allerdings so gewählt, sodass die Korrelation der Rohwertschätzungen der Kurzversion mit den Rohwerten der Langversion bei $r > .9$ liegt (Gittler, 1990).

Die Verkürzung des Tests auf weniger als 18 Aufgaben kann auf drei verschiedene Arten geschehen. Erstens durch Reduktion der Anzahl der Aufgaben um eine bestimmte Menge (*Testverkürzung*) im Vorhinein. Alle Testpersonen bearbeiten dieselbe (kürzere) Version des 3DW. Die Bearbeitungszeit wird insgesamt zwar kürzer sein, ist aber immer noch an das *open end-Prinzip* gebunden. Die zweite Möglichkeit, ist die Testung zu dem Zeitpunkt zu stoppen, an dem die langsamste Person zumindest acht Aufgaben bearbeitet hat (*indirekter Testabbruch*). Ab diesen Zeitpunkt können alle Personen verrechnet werden und ein Vergleich ist möglich, auch wenn die Testpersonen höchstwahrscheinlich eine unterschiedliche Anzahl an Aufgaben bearbeitet haben. Die letzte Variante zur Testverkürzung, ist die Arbeitszeit zu limitieren (*direkter Testabbruch*). Der Test wird hierbei nach einer bestimmten Zeit beendet. Der Vorteil dieser Variante liegt darin, dass die Bearbeitungszeit hierbei sehr genau planbar ist. Allerdings kann es vorkommen, dass nicht alle Personen in die Auswertung miteinbezogen werden können, wenn sie bis dahin nicht mindestens acht Aufgaben bearbeitet haben. Der zeitliche Abbruch muss daher genau überlegt werden. Je nach Anwendungsfall kann die Testleitung selbst entscheiden, welche Variante am besten zur Situation passt (Gittler, 1990).

Eine besondere Form der Testverkürzung stellt das adaptive Testen dar. Üblicherweise sind in einem Test nur wenige der Testaufgaben aussagekräftig und der Rest entweder zu leicht oder zu schwierig. Bei einer adaptiven Version lässt sich die Anzahl der Aufgaben auf ein Minimum beschränken, da jede bereits bearbeitete Aufgabe als Information in die weitere Bearbeitung miteinfließt. Die Schwierigkeiten der vorgegebenen Aufgaben passen sich dem Fähigkeitsparameter der Testperson an. Die Software sucht nach dem nächsten noch aussagekräftigeren Item. Damit nähert sich der geschätzte Parameter dem wahren Leistungsparameter einer Person immer weiter an (Kubinger, 2019). Vor allem in extremen Leistungsbereichen kann die adaptive Version wesentlich besser differenzieren. Aufgrund der guten testtheoretischen Fundierung des 3DW und der Geltung des Rasch Modells, ist eine solche adaptive Version möglich. Die adaptive Version des 3DW (A3DW) wird in der Testbatterie „Aviation Psychology“ von Schuhfried (2005) eingesetzt. Diese testet die Dimension *visualization* bei angehenden Pilot*innen und trägt damit wesentlich zu deren Auswahl bei und ist ein gutes Beispiel für die Verwendung des 3DW in der Praxis.

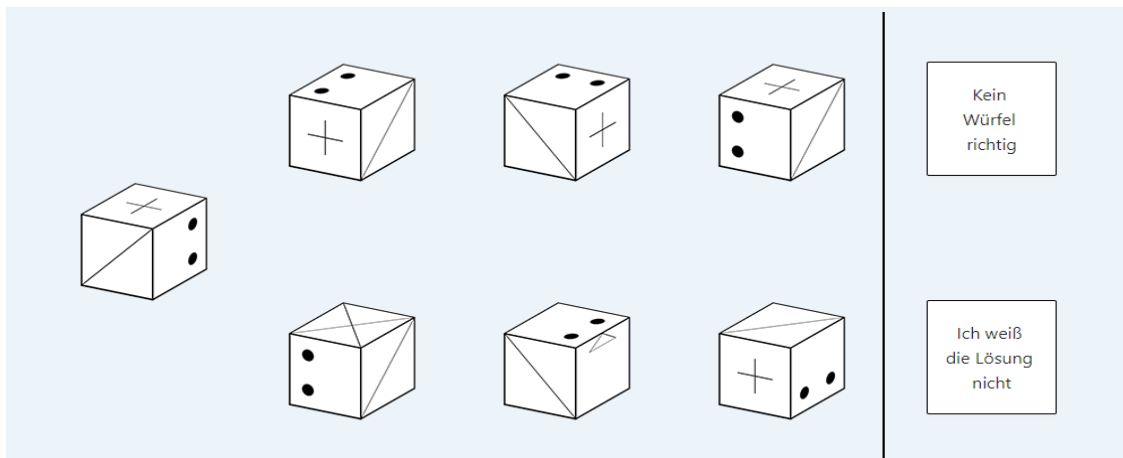
2.7 Adaptionen

1988 erschien der 3DW auch als PC-Version. Von nun an konnte der Test auch am Bildschirm bearbeitet werden. Der Vorteil der PC-Version liegt darin, dass er einfach adaptierbar ist und die Daten automatisch gesammelt werden. Ein weiterer Vorteil liegt in der höheren Durchführungsobjektivität, da die Instruktion durch eine Testleitung wegfällt. Etwaige Bevorzugung oder Benachteiligung von Testpersonen, ob gewollt oder nicht, sind damit ausgeschlossen. Auch die Auswertung erfolgt automatisch mittels Software und ist damit weniger fehleranfällig als das händische Auswerten des Tests. Die automatische Zeiterfassung ist ebenfalls positiv und wesentlich genauer als eine manuelle Zeiterfassung bei der Paper-Pencil-Version. Sie kann in der PC-Version auch einfach und automatisch für alle Einzelpersonen einer Gruppentestung vorgenommen werden. Da der Test einfacher zu verteilen ist und auch bei Gruppentestungen nicht alle Personen in einem Raum sein müssen, kommt es mitunter auch zu weniger sozialem Druck und in Folge zu objektiveren Ergebnissen. In Folge von weniger sozialem Einfluss, der Interaktion mit dem PC und einer in der Regel auch etwas kürzeren Testzeit, wird die Testung mittels PC auch mit einer erhöhten Testmotivation in Verbindung gebracht (Kisser & Wenninger, 1983). Arrer (1992) hat in ihrer Diplomarbeit die Gültigkeit der Normen in der PC-Version des 3DW untersucht, welche für die Version im Paper-Pencil Format gegolten haben. Sie erwiesen sich als gültig.

Die PC-Version des 3DW wurde auch in der vorliegenden Arbeit als Messinstrument verwendet und unterscheidet sich im Design nur minimal von der Paper-Pencil Variante. Die sechs Antwortwürfel sind auch hier rechts vom Zielwürfel angeordnet. Rechts daneben befinden sich die beiden zusätzlichen Möglichkeiten „Kein Würfel richtig“ und „Ich weiß die Lösung nicht“. Das Design der PC-Version ist also eine digitale Abbildung der originalen Paper-Pencil-Version des 3DW. Der einzige Unterschied besteht in der Markierung der korrekten Antwort. Diese kann in der PC-Version mittels eines Klicks auf den gewünschten Antwortwürfel geschehen. Dieser ist dann eingeloggt. Mittels eines Klicks auf die Schaltfläche „Weiter“ schließt die Testperson eine Aufgabe ab und gelangt zur nächsten. Ein Beispielitem der PC-Version des 3DW, so wie es auch in dieser Testung vorkam, findet sich in Abbildung 2 (Gittler & Alfes, 2021a).

Abbildung 2

Beispielitem aus dem 3DW (PC-Version)



(Gittler & Alferts, 2021a)

2.8 Neue Entwicklungen

Die Umstellung von der Paper-Pencil auf die PC-Version war nicht die letzte Veränderung, der der 3DW unterzogen wurde. Unter anderem wurde von Gittler und Arendasy (2003) untersucht, ob etwaige Farbelemente die Eigenschaften des 3DW maßgeblich beeinflussen. Durch Farbelemente an den Würfelflächen wurde vermutet, die Enkodierung zu verbessern. Ziel der Studie war eine Äquivalenzüberprüfung des Tests mit verändertem Design. Es galt herauszufinden, ob der veränderte 3DW dieselben messtheoretischen Eigenschaften besitzt wie das Original, ob das Rasch Modell immer noch gilt und ob es einen möglichen „Farbeffekt“ gibt. Das wäre der Fall, wenn die Farbe insgesamt einen Einfluss auf die Ergebnisse gehabt hätte. Eine Testung an 446 Personen mit verschiedenen Testversionen sowie einer gemischten *Link-Version* wurde anschließend mittels probabilistischer Methoden untersucht. Weder quantitative noch qualitative Unterschiede konnten festgestellt werden. Die Ergebnisse sprechen für die Robustheit der Aufgaben des 3DW sowie die Notwendigkeit solcher Überprüfungen beim Verändern vorhandener Testmaterialien (Gittler & Arendasy, 2003).

Die neueste Weiterentwicklung des 3DW ist Thema dieser Arbeit und betrifft das Design des Tests. Auch ein möglicher Interaktionsmechanismus wurde hinzugefügt. Der neu adaptierte Test wird im Folgenden näher beschrieben.

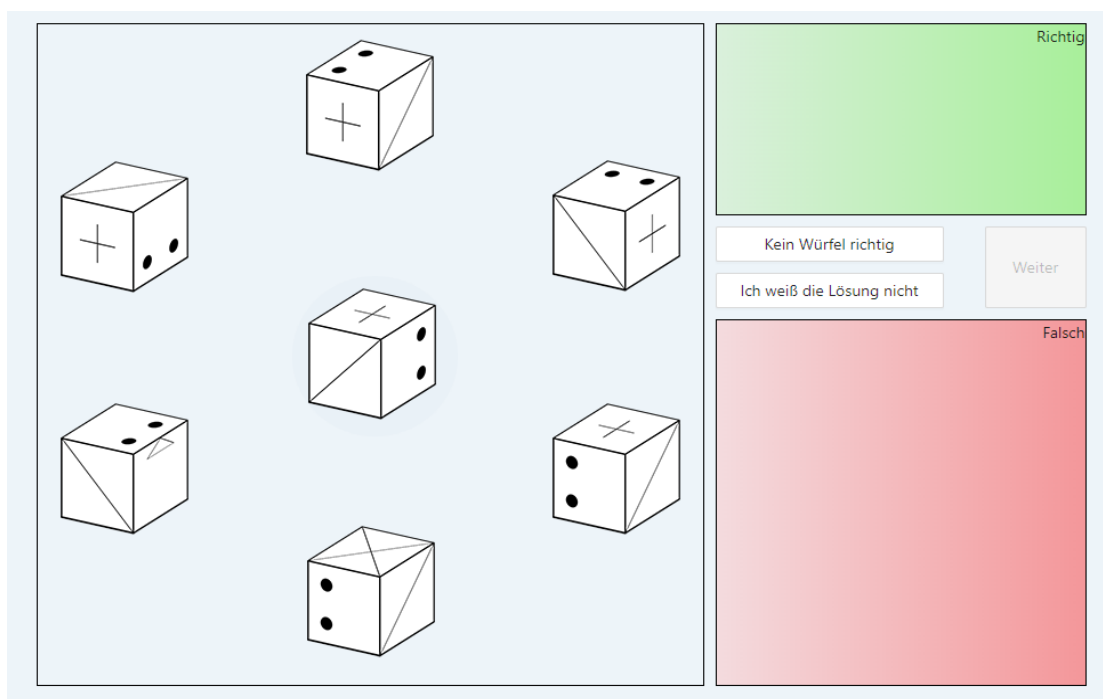
2.9 Dreidimensionaler Würfeltest- π (3DW- π)

Neben der PC-Version des 3DW wurde in dieser Arbeit auch eine neue Version desselben Testverfahrens verwendet, der Dreidimensionale Würfeltest- π (3DW- π). Seinen

Namen verdankt er der besonderen Anordnung der Testitems. Die Antwortwürfel sind im Kreis um den Zielwürfel angeordnet, während in dem weiter oben dargestellten 3DW in der Paper-Pencil-Version und der dazugehörigen PC-Version die Antwortwürfel rechts neben dem Zielwürfel angeordnet sind. Eine weitere Änderung betrifft die mögliche Interaktion mit den Würfeln. Den Testpersonen ist es hier möglich, die verschiedenen Antwortwürfel mittels Mauszeiger und -tasten über den Bildschirm zu ziehen und in dafür vorgesehene Felder abzulegen. Sie können die Würfel entweder in das Feld „Falsch“ oder „Richtig“ ziehen. Um einen Antwortwürfel als richtig zu markieren, muss dieser in das dafür vorgesehene grün hinterlegte Feld „Richtig“ verschoben werden. Erst dann kann durch einen Klick auf die Schaltfläche „Weiter“ mit der nächsten Aufgabe fortgefahren werden. Das rot hinterlegte Feld „Falsch“ dient als Hilfestellung und kann, muss aber nicht genutzt werden. Damit ist es möglich, alle falschen Alternativen der Reihe nach in das rote Feld zu schieben und nur die noch in Frage kommenden Würfel miteinander zu vergleichen, bis die richtige Alternative gefunden wurde. Die beiden Antwortalternativen „Kein Würfel richtig“ und „Ich weiß die Lösung nicht“ sind wie auch im originalen Design vorhanden und können ebenfalls gewählt werden. Ein Beispielitem des 3DW- π , so wie es in dieser Testung vorkam, findet sich in Abbildung 3 (Gittler & Alferts, 2021a).

Abbildung 3

Beispielitem aus dem 3DW- π (adaptierte Version)



(Gittler & Alferts, 2021a)

Durch die Änderung des Designs und den Interaktionsmechanismus sollen mehrere Aspekte des Tests verbessert werden. Die kreisrunde und im weiteren Verlauf geplante zufällige Anordnung der Antwortwürfel soll unter anderem die Sicherheit des Tests verbessern. Dadurch ist es bei mehrmaliger Bearbeitung des Tests nicht mehr möglich, sich die korrekten Antworten anhand der Anordnung der Würfel zu merken, da sie bei jeder Bearbeitung anders ausfallen würde. Diese Gefahr besteht bei einer weiten Verbreitung und häufiger Anwendung in Auswahlverfahren durchaus. Auch der systematischen Vorgangsweise von Würfel A bis F soll vorgebeugt werden, da die Reihenfolge in der kreisrunden Anordnung nicht mehr eindeutig vorgegeben ist (Gittler & Alferts, 2021a).

Ein weiterer Vorteil liegt in der Interaktivität des 3DW- π . Durch die Möglichkeit, die Würfel zu bewegen und die gleichzeitige Aufzeichnung sämtlicher Schritte der Testpersonen im zugehörigen log-file, können Rückschlüsse auf die eingesetzten Strategien der Testpersonen getroffen werden. Auch die Bearbeitungszeit wird automatisch erfasst und gewährt Einblicke in die Details einer Bearbeitung. Durch die mögliche systematische Bearbeitung mittels drag and drop-Prinzip können auch besonders attraktive Distraktoren erkannt werden. So wäre es möglich, dass manche Antwortwürfel sehr schnell als falsch erkannt werden, während andere nur sehr schwierig ausgeschlossen werden können. Das kann Hinweise zur weiteren Verbesserung der Items geben. Dies sind alles Aspekte, die für die Diagnostik als auch die Weiterentwicklung des Tests besonders wertvoll sind (Gittler & Alferts, 2021a).

Außerdem erwartet werden Effekte in Hinblick auf die Attraktivität und die Arbeitsmotivation. Zuletzt wurde vermehrt beobachtet, dass Testpersonen den 3DW als sperrig erleben. Auch subjektive Beobachtungen lassen darauf schließen, dass die Arbeitsmotivation aufgrund des eher klassischen Designs nachlässt. Immerhin handelt es sich bei der derzeit verwendeten PC-Variante immer noch um das gleiche Design, wie einst in der Paper-Pencil-Version. Die neue Version macht den Test interaktiver und damit zugänglicher und ist näher an der modernen Realität als die Originalversion des 3DW. Schließlich interagieren Menschen heutzutage täglich mit dem Content auf ihrem Smartphone und auf anderen technischen Endgeräten. In weiterer Folge geht es darum, die Zielsetzung dieser Arbeit genau zu beleuchten und die weitere Vorgehensweise zu erläutern (Gittler & Alferts, 2021a).

Empirischer Teil

3. Ziele und Fragestellungen

Das Ziel der vorliegenden Untersuchung ist herauszufinden, inwiefern sich der 3DW in der Originalversion (3DW) und die neue Version mit verändertem, interaktivem Design (3DW- π) voneinander unterscheiden. Die Fragestellung dieser Arbeit beschäftigt sich demnach mit der Messäquivalenz der zwei verschiedenen Testversionen. Die Arbeit soll einen ersten Einblick in die Unterschiede der zwei Testversionen geben. Obwohl das neue Design zahlreiche Vorteile bietet, wie die leichtere Informationsgewinnung, Sicherheitsaspekte und Einblicke in die Lösungsstrategien der Teilnehmer*innen, konzentriert sich diese Arbeit vorwiegend auf drei zentrale Aspekte. Die Rasch-Homogenität der Daten, einen möglichen Designeffekt sowie Hinweise auf die Arbeitsmotivation der Testpersonen. Die zentrale Fragestellung dieser Arbeit lautet daher:

Wie unterscheiden sich der 3DW und der 3DW- π hinsichtlich:

- *Rasch-Homogenität*
- *Designeffekt*
- *Arbeitsmotivation*

Zunächst wird die Frage gestellt, ob die Items dem Rasch Modell entsprechen und die Eindimensionalität gegeben ist. Eine zweite Frage soll klären, ob es einen spezifischen Effekt des Designs gibt (*Designeffekt*). Das wäre der Fall, wenn die neue Version beispielsweise für alle Items einfacher zu bearbeiten ist. Die dritte Frage beschäftigt sich mit der Arbeitsmotivation der Testpersonen. Sie soll erste Hinweise darauf geben, wie das neue Testdesign bei ihnen ankommt. Dabei geht es um die Aspekte Arbeitsmotivation, Erfolgserwartung, Testangst und Gefallen.

Das zu erreichende Ziel mit der adaptierten Version ist die praktische Einsetzbarkeit des neuen Designs. Um dieses in der Praxis einzusetzen, muss es denselben Qualitätsstandards entsprechen wie das originale Design. Die Geltung des Rasch Modells muss gegeben sein. Das gilt sowohl für interne Kriterien als auch als Voraussetzung für das adaptive Testen. Sollte sich ein Designeffekt herauskristallisieren, müssten auch die Normwerte entsprechend angepasst werden. Nicht zuletzt spielt auch das subjektive Gefallen von Testpersonen eine Rolle. Das Ziel der Weiterentwicklung ist letzten Endes nicht nur eine sicherere Version des Tests zu bekommen, die vielfältige Möglichkeiten zur

Informationsgewinnung bietet, sondern auch eine moderne und interaktive Version zu haben, die Testpersonen zusagt und motiviert, ihr Bestes zu geben. Im Folgenden finden sich die zentralen Hypothesen dieser Arbeit sowie die Vermutungen, auf denen sie beruhen.

3.1 Rasch-Homogenität

Der 3DW wurde mit dem Ziel der Rasch-Homogenität konstruiert. Die Rasch-Homogenität der Originalversion wurde bereits in vielen Stichproben bestätigt und gilt als sehr robust. Auch die Einführung und Testung neuerer Testversionen wie zum Beispiel die Übertragung des 3DW von einer Paper-Pencil-Version zu einer PC-Version, hielt dem Rasch Modell stand (Arrer, 1992). Ebenso hat eine Einführung eines Farbelements in den Aufgaben für keine Veränderungen in der Rasch-Homogenität gesorgt (Gittler & Arendasy, 2003). Aufgrund dieser Ergebnisse ist davon auszugehen, dass das Rasch Modell auch für die verwendeten Items in der vorliegenden Stichprobe gilt. Liegt eine Verletzung des Rasch Modells vor, kann dies aber auch ein Hinweis auf eine schlechte Verteilung der Stichprobe sein. Für den 3DW- π ist das Erreichen der Rasch-Homogenität von besonderer Bedeutung, da er in Zukunft auch als Testversion in der Praxis einsetzbar sein soll. Eine Verletzung dieser würde die Entwicklung der neuen Version bremsen sowie umfangreichere Anpassungen und Untersuchungen notwendig machen.

Hypothese 1: Die Verteilung der Daten in der vorliegenden Stichprobe lassen sich mit dem Rasch Modell beschreiben.

3.2 Designeffekt

Die Interaktivität des 3DW- π entspricht eher unserer modernen Umwelt als die originale Version des 3DW. Im Alltag interagieren Menschen oft mit Objekten, auch das Hin- und Herschieben von digitalen Inhalten auf mobilen Endgeräten findet häufig statt. Gao (2003) untersuchte die Lernerfolge und Motivation von Testpersonen nach einer Auseinandersetzung mit Lernmaterialien, die einen unterschiedlich hohen Interaktionsgrad aufwiesen. Studierende, die einen höheren Interaktionsgrad vorfanden, waren motivierter als die Vergleichsgruppen. Cantoni und Kolleg*innen (2004) beschäftigten sich vor allem mit der visuellen Komponente von Lernmaterialien im E-Learning Bereich. Auch sie kamen zum Schluss, dass sich die interaktive und reaktive Gestaltung von Materialien positiv auf das Lernen auswirken könnte, da es natürliches Lernen fördert. Aufgrund dieser Erkenntnisse und den oben beschriebenen Überlegungen wird ein Effekt des Designs als wahrscheinlich angenommen.

Hypothese 2: Es gibt einen Designeffekt. Ein Modell mit einem freien Designparameter beschreibt die Daten genauso gut wie ein vollständiges Rasch Modell.

3.3 Arbeitsmotivation

Wie oben bereits erwähnt wird vermutet, dass sich eine interaktive Version des Tests positiv auf die Motivation von Testpersonen auswirkt. Um diese Frage aufzugreifen, kommt ein Arbeitsmotivationsfragebogen zum Einsatz. Dieser fragt nach unterschiedlichen Aspekten der Arbeitsmotivation und soll damit erste Einblicke in das Thema liefern.

Hypothese 3: Es gibt Unterschiede in der Arbeitsmotivation von Testpersonen, die unterschiedliche Versionen des 3DW bearbeitet haben. Personen, die den 3DW- π bearbeitet haben, waren motivierter als Personen, die den 3DW bearbeitet haben.

3.4 Nebenfragestellungen

Im Falle einer erfolgreichen Rasch Analyse können auch Gesamtscores interpretiert werden. Damit soll ein weiterer Einblick in die Daten gegeben und mögliche Unterschiede sichtbar gemacht werden. Allen voran wird nach Unterschieden in den verschiedenen Untersuchungsgruppen gesucht. Aber auch Unterschiede zwischen Männern und Frauen sowie verschiedenen Altersgruppen sollen gefunden werden. Auch die Frage, welcher Wert am häufigsten erreicht wurde, soll geklärt werden. Die Verteilung der Scores kann gute Hinweise auf die Qualität der Stichprobe liefern. Die ersten Nebenfragestellungen lautet daher:

Unterscheidet sich der Gesamtscore in folgenden Untergruppen?

- *Untersuchungsgruppen*
- *Geschlecht*
- *Alter*

Welcher Wert wurde am häufigsten erreicht?

Einen weiteren spannenden Einblick in die Daten bietet die Auswertung der Bearbeitungszeiten. Durch den Interaktionsmechanismus könnte die Bearbeitungszeit des 3DW- π höher liegen als im 3DW. Genauso gut könnte es sein, dass durch die Möglichkeit, falsche Würfel auch optisch auszuschließen, schneller eine Antwort gefunden wird als im Originaldesign. Möglicherweise haben auch die beiden Gruppen, die unterschiedliche Versionen bearbeitet haben, höhere Werte. Interessant ist daher auch folgende Fragestellung:

Unterscheidet sich die Bearbeitungszeit

- zwischen dem 3DW und dem 3DW- π ?

- in den vier Untersuchungsgruppen?

4. Methode

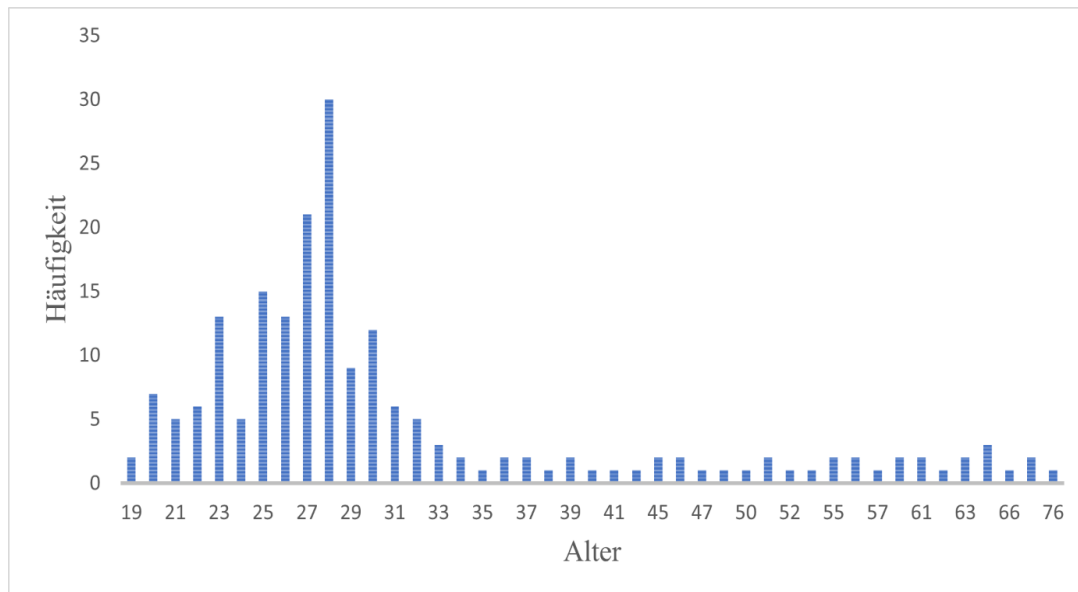
4.1 Stichprobe

Die Datenerhebung fand im Jahr 2021 zwischen April und Juli statt. Die Testpersonen wurden vor allem über den Bekannten- und Freundeskreis rekrutiert (60%). Ein weiterer Teil wurde mittels Social-Media gefunden sowie durch verschiedene Universitätsgruppen und Verteiler (31.3%) auf die Studie aufmerksam. 6.7% fanden über diverse andere Kanäle ihren Weg zur Untersuchung oder konnten im Nachhinein keiner der oben genannten Gruppen zugeordnet werden.

Insgesamt umfasste die Stichprobe $N = 195$ Testpersonen. In die endgültige Stichprobe wurden nur Testpersonen mit vollständigen Datensätzen aufgenommen. Ein Mindestalter von 18 Jahren war Voraussetzung für die Teilnahme. Auf ein Höchstalter wurde verzichtet. Die Altersspanne umfasste 19 bis 76 Jahre. Der Mittelwert lag bei $M = 31.78$ Jahren bei einer Standardabweichung von $SD = 11.83$. Der Median lag bei $Md = 28$ Jahren. Zudem war 28 der Wert, der am häufigsten vorkam ($Mo = 28$). Mehr als die Hälfte der Testpersonen war unter 30 Jahre alt. Die übrige Verteilung war gleichmäßig auf alle Altersgruppen verteilt, bis hin zur ältesten Testperson mit 76 Jahren. Die Abbildung 4 zeigt die Altersverteilung über die gesamte Stichprobe.

Abbildung 4

Altersverteilung der Stichprobe



Die Geschlechtsverteilung innerhalb der Stichprobe war ausgeglichen. 111 Testpersonen waren weiblich (56.9%) und 84 (43.1%) männlich. 81% der Testpersonen kamen aus Österreich, weitere 12.8% aus Deutschland. Die übrigen 6.2% verteilten sich auf diverse andere Nationen. Der Bildungsgrad der Stichprobe war unausgeglichen. 59% der Testpersonen hatten einen akademischen Abschluss, 51.8% von einer Universität und 7.2% von einer Fachhochschule. Weitere 28.7% hatten als höchsten Abschluss eine AHS-Matura, 10.3% eine abgeschlossene berufsbildende mittlere Schule und nur 2.1% eine abgeschlossene Pflichtschulausbildung.

4.2 Untersuchungsablauf

Die Testung wurde mit einer Online-Testbatterie (Gittler & Alferts, 2021a)¹ durchgeführt. Diese wurde mithilfe eines Links an die Testpersonen verschickt und konnte am privaten PC bearbeitet werden. Die Testung fand somit in privaten Räumlichkeiten statt und unterlag keinerlei Aufsicht.

Im ersten Schritt wurden die Testpersonen begrüßt und ihre demografischen Daten aufgenommen. Abgefragt wurde das Geschlecht, das Alter, die Herkunft sowie die höchste abgeschlossene Ausbildung. Im Anschluss begann die Testung. Diese startete bei allen Testpersonen mit einer allgemeinen Instruktion zum 3DW samt einem Beispielitem. Im ersten

¹ Der Autor bedankt sich an dieser Stelle bei Herrn Ao. Univ.-Prof. i.R. Dr. Georg Gittler für die Bereitstellung des Testmaterials und bei Tobias Alferts, MSc für die Programmierung der Testbatterie.

Testteil (T1) wurden zehn Items aus dem Itempool des A3DW vorgegeben und entweder in der Version 3DW oder 3DW- π , bearbeitet. Das erste dieser zehn Items war immer ein Aufwärmitem, das nicht in die Bewertung miteinfluss. Insgesamt wurden in T1 also neun Wertungsitems vorgegeben. Anschließend folgte die Vorgabe des Arbeitsmotivationsfragebogens mit insgesamt 17 Items. Nach einer kurzen Pause, die durch einen Text eingeleitet wurde, folgte der zweite Testteil (T2). Dieser bestand aus weiteren zehn Wertungsitems aus dem Itempool des A3DW, die ebenfalls in der Version 3DW oder 3DW- π bearbeitet wurden sowie der erneuten Beantwortung der 17 Fragen zur Arbeitsmotivation. Die Länge der Pause zwischen den Teilen konnte individuell gewählt werden.

Es gab kein Zeitlimit und es mussten alle Items bearbeitet werden, um in der Testbatterie voranzukommen. Lediglich ein vorzeitiger Testabbruch war möglich. Unvollständige Daten wurden jedoch nicht in der Untersuchung berücksichtigt und ausgeschlossen. Nach Abschluss der gesamten Testung wurde auf das Ende verwiesen und eine Danksagung erschien. Die Testpersonen konnten das Fenster schließen und die Testung war damit beendet. Als Dankeschön für die Teilnahme erhielten die Testpersonen im Anschluss eine Auswertung über ihr Abschneiden im Raumvorstellungstest per E-Mail zugesandt.

4.3 Testdesign

Es gab vier Untersuchungsgruppen². Die 195 Testpersonen wurden diesen randomisiert und automatisch zugewiesen. Die Untersuchungsgruppe 1 (UG1) beinhaltete 45 Testpersonen, die Untersuchungsgruppe 2 (UG2) 51 Testpersonen, die Untersuchungsgruppe 3 (UG3) 47 Testpersonen und die Untersuchungsgruppe 4 (UG4) 52 Testpersonen. Die Geschlechtsverteilung innerhalb der einzelnen Untersuchungsgruppen war ebenfalls ausgeglichen. Die Verteilung der Untersuchungsgruppen findet sich in Tabelle 1.

² Der Autor bedankt sich an dieser Stelle bei Herrn Ao. Univ.-Prof. i.R. Dr. Georg Gittler und Herrn Tobias Alfes, MSc für die Unterstützung bei der Planung des Versuchsdesigns.

Tabelle 1*Untersuchungsgruppen*

Gruppe	T1	T2	N	M	W
UG 1	3DW	3DW- π	45 (23.1%)	51.1%	48.9%
UG 2	3DW- π	3DW	51 (26.2%)	41.2%	58.8%
UG 3	3DW	3DW	47 (24.1%)	66%	34%
UG 4	3DW- π	3DW- π	52 (26.7%)	53.8%	46.2%

Die vier Untersuchungsgruppen unterschieden sich in dem unterschiedlichen Design der Items, die sie im Laufe der Testung bearbeiteten. Die Reihenfolge der bearbeiteten Items war immer gleich. In Testteil 1 (T1) wurden die Items 1-10 bearbeitet und in Testteil 2 (T2) die Items 11-20, unabhängig vom vorgegebenen Design. UG1 und UG2 arbeiteten mit beiden Versionen und hatten somit jeweils einmal mit dem 3DW und einmal mit dem 3DW- π zu tun. Die Gruppen UG3 und UG4 bearbeiteten jeweils nur eine Version, entweder 20 Items des 3DW oder 20 Items des 3DW- π . Die Untersuchungsgruppen sind in Tabelle 1 dargestellt. Darin finden sich auch die bearbeiteten Testversionen in T1 und T2. Nach jedem Testteil wurde derselbe Arbeitsmotivationsfragebogen mit 17 Items vorgegeben. Er bezog sich retrospektiv auf die zuvor bearbeiteten Testaufgaben.

4.4 Messinstrumente

4.4.1 Der Dreidimensionale Würfeltest (3DW)

In der vorliegenden Arbeit wurde zur Messung der Raumvorstellungsfähigkeit der Dreidimensionale Würfeltest (3DW) von Gittler (1990) verwendet und am PC vorgegeben. Dieser ist nicht nur das verwendete Testinstrument, sondern auch Ausgangspunkt dieser Arbeit und wurde im Theorieteil bereits ausführlich beschrieben. Details zur Entwicklung, der Fundierung des Verfahrens, Einsatzgebieten sowie eine ausführliche Beschreibung befinden sich im theoretischen Teil in Kapitel 2.

Die Aufgabe im 3DW besteht darin, einen Zielwürfel aus sechs verschiedenen Antwortwürfeln zu erkennen. Gelöst werden soll sie durch den Einsatz von mentaler Rotation, durch geistiges Drehen oder Kippen der Würfel. Es besteht zudem die Möglichkeit, die Alternativen „Kein Würfel richtig“ oder „Ich weiß die Lösung nicht“ zu wählen. Der 3DW

wird ohne Zeitvorgabe vorgegeben, es handelt sich somit um eine reine Power Testung. Ein Beispielitem des 3DW in der PC-Version findet sich in Kapitel 2.7 in Abbildung 2.

4.4.2 Der Dreidimensionale Würfeltest- π (3DW- π)

Neben dem Originaldesign des 3DW wurde in dieser Arbeit auch eine neue Version desselben Testverfahrens verwendet, der 3DW- π . Nähere Details sowie alle Aspekte und Gründe der Neugestaltung finden sich im theoretischen Teil in Kapitel 2.4.

Seinen Namen verdankt der 3DW- π der besonderen Anordnung der Testitems. Die Antwortwürfel sind im Kreis um den Zielwürfel angeordnet. Eine weitere Änderung betrifft die mögliche Interaktion mit den Würfeln. Den Testpersonen ist es hier möglich, die verschiedenen Antwortwürfel mittels Mauszeiger und -tasten über den Bildschirm zu ziehen und in dafür vorgesehene Felder abzulegen. Erst wenn der korrekte Antwortwürfel im richtigen Feld liegt, kann die Aufgabe abgeschlossen werden. Ein Beispielitem des 3DW- π findet sich in Kapitel 2.9 in Abbildung 3.

4.4.3 Arbeitsmotivationsfragebogen

Der in dieser Arbeit verwendete Fragebogen zur Arbeitsmotivation basiert auf einer Ausarbeitung von Studierenden des Seminars "Forschungsfelder der psychologischen Diagnostik und Persönlichkeitspsychologie" unter der Leitung von Prof. Dr. Steffi Pohl an der Freien Universität Berlin von 2009 bis 2010. Die in dieser Studie verwendete Version des Fragebogens wurde von Georg Gittler und Tobias Alferts im Wortlaut und Kontext modifiziert sowie von ursprünglich insgesamt 30 auf 17 Fragen gekürzt (Gittler & Alferts, 2021b)³.

Die 17 verwendeten Fragen bezogen sich im Allgemeinen auf die Arbeitsmotivation der Proband*innen, indem diese möglichst variantenreich abgefragt wurde. Es wurden Fragen zur allgemeinen Motivation, dem Testgefallen, der Erfolgserwartung sowie zur Testangst gestellt. Beispielfragen aus dem Fragebogen lauteten:

„Ich war bemüht, die Testaufgaben so gut wie möglich zu bearbeiten.“

„Ich war während der Bearbeitung des Tests angespannt.“

„Es hat mir Freude bereitet, diesen Test zu bearbeiten.“

³ Der Autor bedankt sich an dieser Stelle bei Herrn Ao. Univ.-Prof. i.R. Dr. Georg Gittler und Herrn Tobias Alferts, MSc für die Zusammen- und Bereitstellung des Arbeitsmotivationsfragebogens.

Zur besseren Beurteilung der dahinterliegenden Konstrukte des neu zusammengestellten Fragebogens sowie deren weiteren Verwendung in der Analyse dieser Arbeit wurde mit den verwendeten Items eine Faktorenanalyse durchgeführt. Die vollständige Liste der verwendeten Fragen sowie eine Zuordnung zu den gefundenen Faktoren befindet sich in Anhang B. Die genaue Beschreibung zum Vorgehen folgt in Kapitel 4.6.2.

4.5 Statistische Analyse

4.5.1 Das dichotom logistische Modell von Rasch

Das in dieser Arbeit angewendete dichotom logistische Rasch Modell wurde von Georg Rasch (1960) eingeführt und begründete die probabilistische Testtheorie oder auch Item-Response Theorie. Sie stellt im Vergleich zur klassischen Testtheorie eine modernere Variante der Testkonstruktion dar und bietet gegenüber seinem Konterpart weitreichende Vorteile. Sie umfasst eine ganze Reihe von Modellen, die das Antwortverhalten von Personen modellieren (Rost, 2004). In der klassischen Testtheorie sind die Reliabilitäten und Validitäten immer von der Stichprobe abhängig. Die probabilistische Testtheorie geht von einer latenten Fähigkeitsdimension aus, die nicht direkt gemessen werden kann. Lediglich durch Beobachtungen manifester Merkmale kann auf sie geschlossen werden. Es wirken zum Beispiel die Items eines Tests als solche manifesten Merkmale, um auf die dahinterliegende Kompetenz, zum Beispiel die Raumvorstellung, schließen zu können. Die Antwort auf ein raschhomogenes Item hängt dabei immer nur von einem Parameter ab, der sich aus der Fähigkeit einer Person und der Schwierigkeit eines Items zusammensetzt. In anderen Modellen innerhalb der IRT können diese Parameter variieren (Rost, 2004).

Das Rasch Modell hat als Item-Charakteristik eine logistische Funktion gewählt. Diese hat zum Vorteil, dass der Wertebereich der Funktion immer zwischen 0 und 1 liegt und ausschließlich von einem einzigen Parameter abhängt, nämlich von dem zu messenden latenten Merkmal. Dieser Parameter findet sich in der Formel des Rasch Modells wieder, die im Folgenden genauer aufgeschlüsselt wird (Rasch, 1960, zitiert nach Fischer, 1974). Eine Person v löst das Item i mit folgender Wahrscheinlichkeit:

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

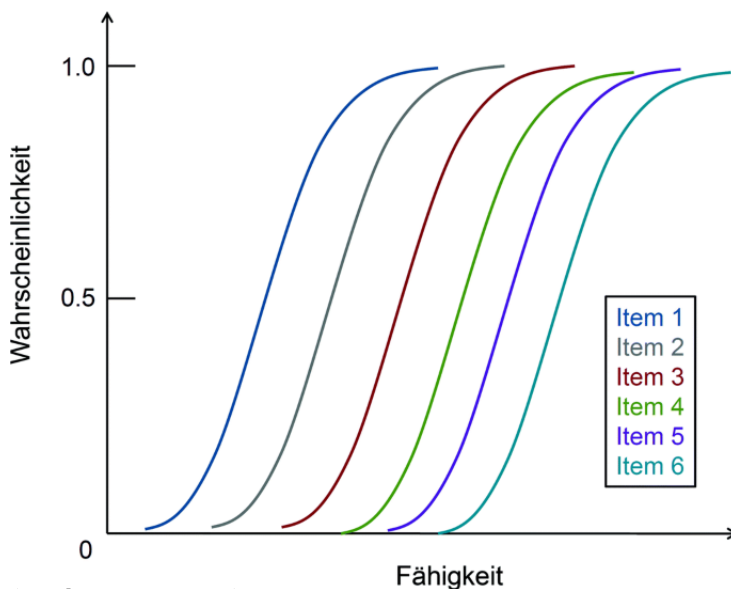
ξ_v beschreibt den Fähigkeitsparameter und σ_i den Schwierigkeitsparameter. Aus dieser Formel ergibt sich die Lösungswahrscheinlichkeit. Ist die Personenfähigkeit gleich der Itemschwierigkeit, wird das Item mit einer Wahrscheinlichkeit von 50% gelöst. Ist die Personenfähigkeit darunter, liegt die Lösungswahrscheinlichkeit bei unter 50%, ist sie

darüber, liegt die Lösungswahrscheinlichkeit bei über 50%. Konkret bedeutet das, dass je höher die Fähigkeit und je leichter das Item, desto höher ist die Wahrscheinlichkeit, dass das Item auch gelöst wird. Sind zwei Items gleich schwierig, wird die fähigere Person ein Item eher lösen. Ein leichteres Item wird von gleich fähigen Personen eher gelöst als ein schwieriges Item (Fischer & Molenaar, 1995).

Die Beziehung der Lösungswahrscheinlichkeit und der latenten Dimension kann über die so genannte Item-Charakteristik-Kurve (ICC) gezeigt werden. Je nach Ausprägung der Fähigkeit gilt eine höhere Lösungswahrscheinlichkeit. Besonders aussagekräftige Items diskriminieren sehr gut zwischen fähigen und weniger fähigen Personen. Ist die Person gleich fähig wie das Item schwierig, liegt die Lösungswahrscheinlichkeit bei 50%. An diesem Punkt liefert ein Item die meiste Information (Fischer, 1974; Rost, 2004). Zur Veranschaulichung findet sich in Abbildung 5 eine beispielhafte ICC-Kurve für sechs Items (Wittke, 2019, S.142).

Abbildung 5

ICC-Kurve von sechs Items



Ein Test, der mehrere Items enthält, darf keine sich überschneidenden ICC-Kurven aufweisen, sie sollten auf der Skala lediglich anhand der Itemschwierigkeiten verschoben sein. Dies ist eine der zentralen Grundannahmen von Rasch. Seine konkreten Forderungen an das Modell sind im Folgenden dargestellt (Rasch, 1960).

1. Das Verhältnis der Schwierigkeiten von zwei zu vergleichenden Items sollte unabhängig von der Stichprobe sein.
2. Das Verhältnis der Fähigkeiten zweier Personen sollte unabhängig davon sein, welche Aufgaben bearbeitet wurden.
3. Die Anzahl der bearbeiteten Aufgaben sollte die gesamte Information über die Fähigkeit einer Person beinhalten.
4. Die Anzahl der Personen, die ein Item lösen konnten, sollte die gesamte Information über die Schwierigkeit eines Items beinhalten.

Diese Forderungen beinhalten zwei Grundannahmen der probabilistischen Testtheorie und die zwei zentralen Thesen des Rasch Modells. Nämlich die *spezifische Objektivität* und die *erschöpfende Statistik*. Die erschöpfende Statistik besagt, dass die reine Anzahl der gelösten Aufgaben Rückschlüsse über die Fähigkeit der Person ermöglicht. Es ist dabei nicht wichtig, welche Aufgaben die Person gelöst hat, sondern nur wie viele. Der Rohwert ist somit bereits ein faires Maß der Leistung. Die spezifische Objektivität besagt, dass ein Vergleich von zwei Personen möglich ist, unabhängig von den bearbeiteten Aufgaben. Ebenso ist ein Vergleich zweier Items möglich, unabhängig von den Personen, die sie gelöst haben. Daraus folgt auch die Tatsache, dass sich die ICC-Kurven, wie oben beschrieben, nicht schneiden dürfen. Sie sollten immer dieselbe Steigung haben (wie in Abbildung 5) (Rost, 2004).

4.5.2 Modellkontrolle

Die Geltung des Rasch Modells kann mittels spezieller Geltungstests überprüft werden. Zum Beispiel mittels des Likelihood-Quotienten-Tests (LQT) von Andersen (1973), welcher auch in dieser Arbeit Anwendung fand. Er vergleicht Untergruppen hinsichtlich ihrer Ergebnisse. Zum Beispiel werden Männer und Frauen oder verschiedene Altersgruppen verglichen. Ziel der Testung ist herauszufinden, ob die Nullhypothese behalten werden kann und das Rasch Modell somit für alle Teilgruppen gleichermaßen gilt (Anderson, 1973).

Die Geltung des Rasch Modells bringt Vorteile. Zum Beispiel lässt sich die Eindimensionalität der Daten beweisen. Alle Testaufgaben messen dasselbe Konstrukt, wenn das Rasch Modell gilt. Genau das sollte im Idealfall auch in dieser Untersuchung der Fall sein. Weitere Anwendungen und Fallbeispiele des Rasch Modells finden sich unter anderem bei Gittler (1990), Fischer und Molenaar (1995), Gittler und Arendasy (2003) und Bortz und Döring (2006).

Ausgehend vom Rasch Modell wurden zahlreiche weitere Modelle auf dessen Grundlage entwickelt. Besonders interessant für diese Arbeit ist das linear-logistische Testmodell (LLTM) von Fischer (1973), das im Folgenden ebenfalls beschrieben wird.

4.5.3 Das linear-logistische Testmodell (LLTM)

Das LLTM geht auf Scheiblechner (1972) und Fischer (1973) zurück und stellt ein restriktiveres Modell als das dichotom logistische Modell von Rasch dar. Die Gültigkeit des Rasch Modells ist daher eine Voraussetzung für die Geltung des LLTM (Moosbrugger, 2012). Die ursprüngliche Idee war, die Schwierigkeit eines dem Modell von Rasch entsprechenden Items, auf die Schwierigkeit jener kognitiven Fertigkeiten zurückzuführen, die aufgrund theoretischer Überlegungen im Vorfeld der Lösung des Items zugrunde liegen. Zur Kontrolle der Gültigkeit des LLTM werden die im Modell geschätzten Parameter mit den Parametern aus dem dichotom logistischen Modell von Rasch mithilfe von Modellkontrollen verglichen. Raschhomogene Testaufgaben, die unter unterschiedlichen Bedingungen vorgegeben wurden, können als strukturell unterschiedliche Aufgaben betrachtet werden. Die Itemparameter des dazugehörigen Rasch Modells werden einer linearen Nebenbedingung unterworfen, was in der so genannten Q-Matrix abgebildet wird. Dies soll den vermuteten Effekt beschreiben. Mittels LQT kann in Folge geprüft werden, ob das neu formulierte LLTM die Daten ebenso gut beschreibt wie ein flexibleres Modell. Ist das der Fall, können die Parameter reduziert werden, solange bis das passende Modell gefunden wurde (Gittler & Arendasy, 2003).

Das LLTM eignet sich auch zur Untersuchung von experimentellen Versuchsplänen. Dies ist in der vorliegenden Studie der Fall. Weitere Anwendungsmöglichkeiten sind die Veränderungsmessung (Gittler, 1994) oder ein Vergleich von Gruppen. Weitere Anwendungsbeispiele finden sich zum Beispiel bei Fischer und Formann (1982) und bei Gittler (1990). Der bekannteste Test, der auf dem LLTM basiert, ist der Wiener Matrizen Test von Formann und Piswanger (1979).

4.6 Datenauswertung

Die statistischen Auswertungen in dieser Arbeit wurden mittels der Statistikprogramme SPSS (Version 27) und R (Version 4.1.1) durchgeführt. Das Signifikanzniveau wurde mit $\alpha = .05$ festgelegt. Für Deskriptiv- und Inferenzstatistik sowie für faktorenanalytische Auswertungen wurde SPSS verwendet, zur Messung der Eindimensionalität und für die Modellvergleiche wurden Analysen in R durchgeführt. In die Auswertung der Daten flossen immer 19 Wertungsaufgaben mit ein. Das erste vorgegebene Item, das Aufwärmitem, wurde lediglich in die Analyse der Testdauer aufgenommen.

Die Auswertung der in dieser Studie erhobenen Daten erfolgte in mehreren Schritten. Zuerst wurde für Hypothese 1 die Geltung des Rasch Modells und die Eindimensionalität der Daten überprüft. Dies war eine notwendige Voraussetzung für weitere Modellvergleiche mit LLTM-Modellen. In einem nächsten Schritt wurde für Hypothese 2 überprüft, ob ein Designeffekt vorliegt. Für die Beantwortung von Hypothese 3 wurde eine Faktorenanalyse durchgeführt und nach Mittelwertsunterschieden gesucht. Im Anschluss wurden die Nebenfragestellungen beantwortet. Hierfür wurden Gesamtscores gebildet und diese in den verschiedenen Gruppen verglichen. Außerdem wurde die Gesamtdauer des Tests berechnet und Gruppenunterschiede geprüft. Die genaue Beschreibung der Datenauswertung folgt im nächsten Abschnitt.

4.6.1 Rasch-Homogenität

Zunächst wurden die Daten im Hinblick auf die Eindimensionalität und Geltung des dichotom logistischen Rasch Modells untersucht. Die Geltung eines eindimensionalen Rasch Modells diente zur Beantwortung der ersten Hypothese und galt als Voraussetzung für die Durchführung weiterer Analysen mit LLTM-Modellen. Die Gültigkeit eines eindimensionalen Rasch Modells würde belegen, dass alle Items durch dieselbe latente Fähigkeit gelöst werden, in diesem Fall durch die Ausprägung in der Fähigkeit Raumvorstellung. Die Gültigkeit des Rasch Modells wurde mittels Modellvergleichen überprüft. Die Vorgehensweise bei den Modellvergleichen war dabei immer gleich. Ein Modell mit mehr freien Parametern wurde mit einem sparsameren, mit weniger freien Parametern, verglichen. Bei einem nicht-signifikanten Ergebnis, beschreibt das sparsamere Modell mit weniger freien Parametern die Daten genauso gut, wie das flexiblere Modell. Ziel war es, ein möglichst sparsames Modell für die Daten zu finden. Verglichen wurde mittels Chi-Quadrat-Test (χ^2 -Test) von Pearson (1900). In der vorliegenden Untersuchung wurde dafür die ANOVA Funktion des Statistik Programms R verwendet.

In einem ersten Schritt wurden ein mehrdimensionales Rasch Modell (mdRM) und ein eindimensionales Rasch Modell (uniRM) Modell auf seine Gültigkeit hin überprüft. Zur weiteren Überprüfung wurde das eindimensionale Rasch Modell auch mit einem Birnbaum Modell (2PL) verglichen. Sollte das eindimensionale Rasch Modell die Daten ebenso gut beschreiben wie ein mehrdimensionales Rasch Modell und ein 2PL Modell, kann von Eindimensionalität ausgegangen werden und weitere Modellvergleiche mit LLTM-Modellen dürfen durchgeführt werden. Zudem wäre die erste Hypothese bestätigt.

In weiterer Folge wurden ein Rasch Modell für unvollständige Daten (uRM) sowie ein vollständiges Rasch Modell (vRM) modelliert und auf ihre Gültigkeit überprüft. Die Gültigkeit der Modelle wurde mittels Likelihood-Ratio-Tests geprüft. Die Teilungskriterien waren der Median, der Mittelwert, das Geschlecht und das Alter. Bei einem signifikanten Ergebnis der Tests würde das Rasch Modell in der Untergruppe dieser Personen nicht gelten. Bei nicht signifikanten Ergebnissen kann von der Geltung des Modells in allen Untergruppen ausgegangen werden.

4.6.2 Designeffekt

Im weiteren Verlauf wurden das unvollständige Rasch Modell und das vollständige Rasch Modell mit einem LLTM mit einem Designparameter (LLTMdp), der neben den Itemparametern modelliert wurde, verglichen. Sollte das Design einen Einfluss auf die Testergebnisse haben, würde ein Vergleich des LLTMdp mit einem vollständigen Rasch Modell signifikant werden. Das würde bedeuten, dass das sparsamere vollständige Rasch Modell die Daten nicht genauso gut beschreiben kann wie das LLTM mit einem Designparameter. Umgekehrt kann bei einem nicht signifikanten Ergebnis von der Gültigkeit des vollständigen Rasch Modells ausgegangen werden. Dieses beschreibt die Daten dann ebenso gut wie das weniger restriktive Modell des LLTM mit einem Designparameter. Hypothese 2 wäre also mit einem signifikanten Modellvergleich bestätigt.

4.6.3 Arbeitsmotivation

Für die Frage, ob sich Testpersonen in ihrer Motivation hinsichtlich des bearbeiteten Testdesigns unterscheiden, wurde ein Arbeitsmotivationsfragebogen vorgegeben. Nähere Details dazu finden sich in Kapitel 4.4.3, die verwendeten Fragen sowie dazugehörige Faktoren befinden sich in Anhang B. Aufgrund eines technischen Fehlers lagen die Daten für den Arbeitsmotivationsfragebogen nur für den ersten Testteil (T1) vor, weshalb die gefundenen Ergebnisse nur für T1 aussagekräftig sind. Da im vorliegenden Testdesign jeweils die Hälfte der Personen im ersten Teil den 3DW bearbeitet hat und die andere Hälfte den 3DW- π , war ein direkter Vergleich dennoch möglich und sinnvoll interpretierbar.

Da der Fragebogen ein neu zusammengestelltes Testinstrument war (Gittler & Alferts, 2021b), konnten über die zugrundeliegenden Faktoren vor Beginn der Arbeit keine Aussagen gemacht werden. Eine explorative Faktorenanalyse wurde durchgeführt, um mehr Einblicke zu gewinnen und die Daten damit besser beschreiben zu können. Mittels Hauptkomponentenanalyse wurden die wichtigsten, unabhängigen Faktoren extrahiert. Die Ermittlung der Anzahl der Faktoren wurde mittels Kaiser-Guttman Kriterium (Guttman,

1954) durchgeführt sowie nach einer Varimax Rotation inhaltlich benannt. Danach wurde ein Vergleich der Summenscores berechnet, um mögliche Unterschiede in den Designs aufzudecken. Gesucht wurde nach Mittelwertsunterschieden auf Ebene der Einzelitems sowie auf Faktorenebene. Diese wurden mittels t-Tests überprüft.

4.6.4 Nebenfragestellungen

Für die Beantwortung der Nebenfragestellungen wurden zunächst Gesamtscores gebildet. Diese wurden mittels ANOVA auf Unterschiede zwischen den vier Untersuchungsgruppen geprüft. Ebenso wurden Mittelwertsunterschiede zwischen den Geschlechtern und zwei Altersgruppen, getrennt nach Median, mittels t-Tests geprüft. Zuletzt wurden speziell die UG3 und UG4 miteinander verglichen, da diese jeweils nur die Version 3DW oder die Version 3DW- π bearbeiteten. Das Ziel war, signifikante Unterschiede zwischen den Teilgruppen aufzudecken.

Zur Berechnung der Gesamtdauer der Testung wurden die zwei Datenpunkte zusammengezählt, nämlich die Dauer der 3DW Testung (entweder in der Version 3DW oder 3DW- π) in T1 und in T2. Aus beiden Zeiten wurde ein Gesamtscore berechnet. Dies war notwendig, da die Bearbeitung nach zehn Aufgaben durch die Bearbeitung des Arbeitsmotivationsfragebogens sowie einer kurzen Pause unterbrochen wurde. Für die Berechnung der Gesamtdauer des Tests wurden fünf Versuchspersonen ausgeschlossen, da diese unrealistische Zeiten aufwiesen. Das Ausschlusskriterium wurde mit einer Dauer von über 120 Minuten gewählt. Dies betraf fünf Personen, sodass mit einem Gesamtdatensatz von $N = 190$ Personen gerechnet wurde. Ermittelt wurden zunächst deskriptive Statistiken für die Gesamtstichprobe sowie für die vier Untersuchungsgruppen. Anschließend wurde mittels ANOVA ein Vergleich der Mittelwerte zwischen den Gruppen sowie ein t-Test zum Vergleich der Mittelwerte zwischen UG3 und UG4 durchgeführt. Das Ziel bestand darin, herauszufinden, ob sich die Bearbeitungszeit in den einzelnen Untersuchungsgruppen, besonders in UG3 und UG4, unterschied.

5. Ergebnisse

5.1 Rasch-Homogenität

Für die Prüfung auf Eindimensionalität und die Gültigkeit des dichotom logistischen Modells von Rasch wurde ein mehrdimensionales Rasch Modell mit einem eindimensionalen Rasch Modell verglichen. Ein ANOVA Vergleich zeigte kein signifikantes Ergebnis

(mdRm/uniRM; $\chi^2 = 28.03$, $p = .139$). Ein mehrdimensionales Rasch Modell bietet als flexibleres Modell keinen signifikanten Mehrwert gegenüber einem eindimensionalen Rasch Modell. Das eindimensionale Rasch Modell beschreibt die Daten ebenso gut und kann in weiterer Folge angenommen werden. Als nächstes wurde zur Überprüfung ein Vergleich zwischen dem eindimensionalen Rasch Modell und dem 2PL Modell durchgeführt. Auch dieses Ergebnis war nicht signifikant (unidRm/2PL; $\chi^2 = 25.865$, $p = .103$). Das eindimensionale Rasch Modell beschreibt auch hier die Daten gleich gut wie das flexiblere 2PL Modell. Die Eindimensionalität der Daten wurde somit bewiesen und die erste Hypothese dieser Arbeit bestätigt. Außerdem war damit die Voraussetzung für weiterführende Analysen mit LLTM-Modellen gegeben. Die Ergebnisse der Modellvergleiche zur Ermittlung der Eindimensionalität und der Gültigkeit des eindimensionalen dichotom logistischen Rasch Modells sind in Tabelle 2 abgebildet. Zur besseren Verständlichkeit dieser wird auf das Abkürzungsverzeichnis sowie das Kapitel 4.6.1 verwiesen.

Tabelle 2

Modellvergleiche zur Prüfung der Eindimensionalität

Modellvergleich	CL	Deviance	npär	χ^2	df	Sig.
<i>mdRM</i>	-1893.10	3786.10	41			
<i>uniRM</i>	-1907.10	3814.10	20	28.03	21	.14
<i>2PL</i>	-1894.10	3788.26	38			
<i>uniRM</i>	-1907.10	3814.10	20	25.7	19	.10

Anmerkung: Conditional Likelihood (CL), Deviance (Abweichung), Anzahl der Parameter (npär), Freiheitsgrade (df)

Infolgedessen wurde der eindimensionale Ansatz fortgeführt und zwei verschiedene Rasch Modelle auf die Daten mittels Conditional Likelihood (CL) Methode angepasst. Zunächst wurde ein Rasch Modell für unvollständige Daten berechnet. Mittels eines Likelihood Ratio Tests wurde es auf seine Gültigkeit hin überprüft. Die Teilungskriterien waren: Median, Mittelwert, Geschlecht und Alter. Keine der vier Subgruppen zeigten ein signifikantes Ergebnis (Median: $p = .11$, Mittelwert: $p = .23$, Geschlecht: $p = .7$, Alter: $p = .41$). Zudem wurde ein vollständiges Rasch Modell auf seine Gültigkeit hin überprüft. Die Likelihood Ratio Tests nach denselben Teilungskriterien wurden ebenso nicht signifikant (Median: $p = .05$, Mittelwert: $p = .19$, Geschlecht: $p = .68$, Alter: $p = .79$). Es kann also von

der Gültigkeit des vollständigen Rasch Modells in den genannten Untergruppen ausgegangen werden. Dies gilt ebenfalls als Voraussetzung für weitere Modellgleichungstests mittels LLTM-Analysen.

5.2 Designeffekt

Tabelle 3 zeigt die Modellvergleiche zwischen dem unvollständigen Rasch Modell, einem LLTM-Modell mit einem Designparameter und einem vollständigen Rasch Modell.

Tabelle 3

Modellvergleiche zur Ermittlung eines Designeffekts

Modellvergleich	CL	Deviance	np _{par}	LR	df	Sig.
<i>uRM</i>	-1314.80	2629.60	37			
<i>LLTMdp</i>	-1326.10	2652.20	20	11.33	17	.84
<i>vRM</i>	-1327.10	2654.20	19	12.30	18	.83
<i>LLTMdp</i>	-1326.10	2652.20	20			
<i>vRM</i>	-1327.10	2654.20	19	.97	1	.33

Anmerkung: Conditional Likelihood (CL), Deviance (Abweichung), Anzahl der Parameter (np_{par}), Likelihood Ratio (LR) Freiheitsgrade (df)

Das flexiblere Modell des unvollständigen Rasch Modells zeigte keine signifikant bessere Passung, sodass das LLTM mit einem Designparameter die Daten genauso gut beschreibt (*uRM/LLTMdp*; LR = 11.33, $p = .84$). Das vollständige Rasch Modell beschreibt die Daten ebenso gut wie das flexiblere LLTM mit einem Designparameter (*vRM/LLTMdp*; LR = .97, $p = .33$). Auch ein direkter Vergleich des unvollständigen Rasch Modells mit dem vollständigen Rasch Modell wurde nicht signifikant (*uRM/vRM*; LR = 12.30, $p = .83$).

Das sparsamste Modell, welches damit bevorzugt wird, ist ein vollständiges Rasch Modell mit einer total log-likelihood von -1327.10. Dieses unterscheidet sich nicht signifikant von flexibleren Modellen wie einem LLTM mit einem Designparameter oder einem unvollständigen Rasch Modell. Die Itemparameter der 19 Wertungsitems des bevorzugten vollständigen Rasch Modells sind in Tabelle 4 aufgelistet.

Tabelle 4*Itemparameter des vollständigen Rasch Modells*

Item	Itemparameter	Std. Error	Item	Itemparameter	Std. Error
1	.77	.18	10	2.04	.23
2	-.54	.08	11	-.96	.10
3	-.63	.09	12	.32	.10
4	-.89	.09	13	-.26	.09
5	-.63	.09	14	1.49	.20
6	-1.47	.13	15	-.70	.09
7	.09	.10	16	.25	.09
8	.09	.10	17	.91	.16
9	.32	.10	18	-.86	.09
			19	.66	.13

Um zu ermitteln, welches Design tendenziell schwieriger ist, wurden zwei separate Rasch Modelle für beide Test-Versionen aufgestellt (3DW und 3DW- π). Die vollständigen Itemparameter für beide Versionen finden sich in Anhang C. Ein Vergleich der Itemparameter beider Designs brachte kein signifikantes Ergebnis ($t(36) = -.57$, $p = .57$). Die nicht signifikante mittlere Differenz zwischen den Itemparametern der beiden Versionen betrug $-.176$ zugunsten des 3DW. Die Korrelation zwischen den Itemparametern beider Versionen erreichte einen hohen Wert ($r = .87$). Aufgrund der oben beschriebenen Ergebnisse konnte kein Designeffekt festgestellt werden, Hypothese 2 ist somit nicht bestätigt worden.

5.3 Arbeitsmotivation

Für die Frage, ob sich Testpersonen in ihrer Motivation hinsichtlich des bearbeiteten Testdesigns unterscheiden, wurde ein Arbeitsmotivationsfragebogen vorgegeben. Nähere Details dazu finden sich in Kapitel 4.4.3, die vollständigen Fragen finden sich in Anhang B. Aufgrund eines technischen Fehlers lagen die Daten für diesen Fragebogen nur für den ersten Testteil (T1) vor. Daher sind die folgenden Ergebnisse nur für T1 aussagekräftig.

Als erstes wurden alle Items des Fragebogens hinsichtlich eines Gruppenunterschieds geprüft. Dafür wurde ein t-Test für unabhängige Stichproben über alle Items hinweg in T1 berechnet. Das Teilungskriterium war die bearbeitete Testversion (3DW oder 3DW- π). Da im vorliegendem Testdesign jeweils die Hälfte der Personen im ersten Teil die Items 1-10 des

3DW bearbeitet hat und die andere Hälfte dieselben Items 1-10 des 3DW- π , war ein direkter Vergleich der Einzelitems möglich. Die Daten von 195 Testpersonen ergaben dabei keine signifikanten Unterschiede. Die einzelnen Ergebnisse aller Items finden sich in Anhang D.

Der Arbeitsmotivationsfragebogen wurde im Anschluss einer Faktorenanalyse unterzogen, um zugrunde liegende Faktoren zu identifizieren. Mittels Hauptkomponentenanalyse wurden die Daten analysiert. Die Überprüfung des Kaiser-Guttman Kriterium (Guttman, 1954) rechtfertigte die Extraktion von vier Faktoren, jeweils mit Eigenwerten größer 1. Sie erklärten eine Gesamtvarianz von 62.25%. Eine Varimax rotierte Vierfaktor-Lösung lieferte die beste Beschreibung der Daten, um sie bestmöglich zu interpretieren. Die vollständig rotierte Komponentenmatrix befindet sich in Anhang E. Die 17 Items wurden anschließend aufgrund ihrer Faktorladungen den vier Faktoren zugeordnet und entsprechend ihrer inhaltlichen Passung benannt. Die vier neu gewonnenen Faktoren waren Testangst (tAngst), Testmotivation (tMot), Erfolgserwartung(tErw) und Gefallen (tGef). Die vollständige Zuordnung findet sich gemeinsam mit den Items im Anhang B.

Anhand der vier gefundenen Faktoren wurden die Ergebnisse noch ein weiteres Mal beleuchtet, um etwaige Unterschiede zwischen den Testversionen zu finden. Ein Vergleich der Summenscores für die vier Faktoren (Tabelle 5) zeigte jedoch keine Unterschiede in den Faktoren Testangst (tAngst: $t(193) = -.12$, $p = .91$), Testmotivation (tmot: $t(193) = .76$, $p = .45$), Erfolgserwartung (tErw: $t(193) = -.02$, $p = .99$) und Gefallen (tGef: $t(193) = -1.169$, $p = .24$). Hypothese 3 wurde somit nicht bestätigt.

Tabelle 5

Vergleich der Summenscores des Arbeitsmotivationsfragebogens

Faktor	t	df	Mittl. Diff.	Sig.
<i>Testangst</i>	-.12	193	-.08	.91
<i>Testmotivation</i>	.76	193	.42	.45
<i>Erfolgserwartung</i>	-.02	193	.01	.99
<i>Gefallen</i>	-1.17	193	-.38	.24

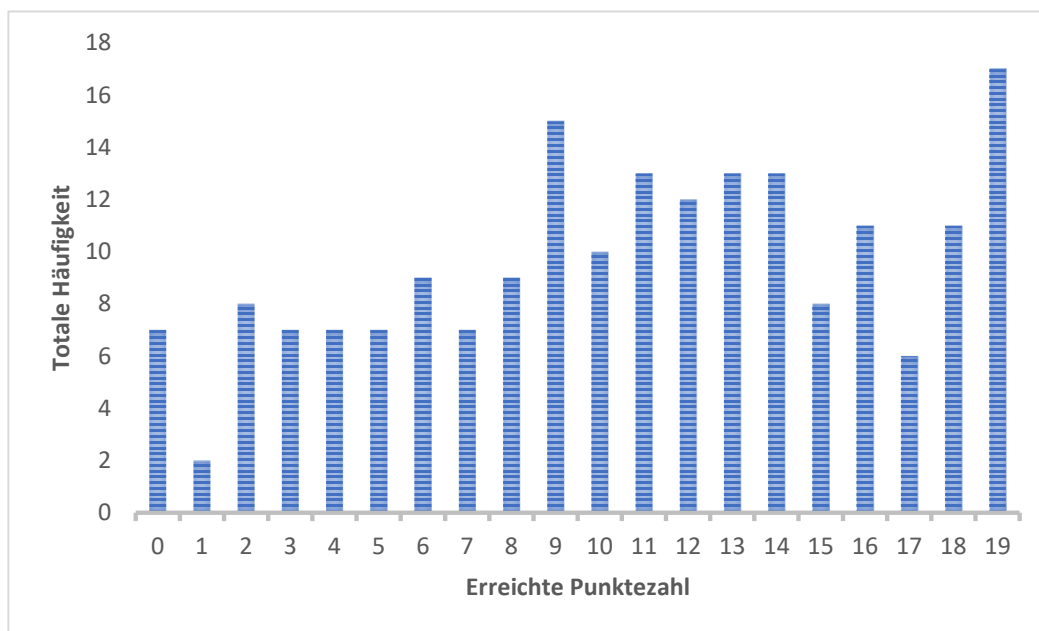
Anmerkung: t-Wert (t), Freiheitsgrade (df)

5.4 Nebenfragestellungen

Nach erfolgreicher Rasch-Analyse wurden die Ergebnisse auch mittels Deskriptivstatistik betrachtet. Die maximal erreichbare Punktezahl lag bei 19 Punkten. In Abbildung 6 wird die Verteilung des Gesamtscores über alle Gruppen hinweg dargestellt. Auffallend hierbei sind die beiden Werte am Anfang und Ende der Skala bei null richtigen Items (3.65%) und 19 richtigen Items (8.85%). Der Rohwert 19 wurde 17-mal erreicht, was gleichzeitig dem Modalwert der Untersuchung entspricht. Der Mittelwert lag bei $M = 10.81$ mit einer Standardabweichung von $SD = 5.46$.

Abbildung 6

Häufigkeiten der erreichten Punktescores im Test



Ein Vergleich der Rohwerte der vier Untersuchungsgruppen ergab keine statistisch signifikanten Unterschiede in den Ergebnissen ($F(3,19) = 116.8$, $p = .27$). Ein Vergleich zwischen Männern ($M = 12.8$, $SD = 5.28$) und Frauen ($M = 9.31$, $SD = 5.12$) ergab ein signifikantes Ergebnis zugunsten der Männer ($t(193) = 4.65$, $p < .001$.) von 3.49 Punkten. Ein Vergleich von zwei Gruppen getrennt nach Alter (unter 28 Jahre und über 28 Jahre) zeigte kein signifikantes Ergebnis ($t(193) = .04$, $p = .97$). Es gibt somit keine Unterschiede im Gesamtscore zwischen den Testversionen, aber einen Unterschied zwischen den Geschlechtern zugunsten der Männer und keinen Unterschied in Bezug auf das Alter.

Die Auswertung der Gesamtdauer der Testung mit dem 3DW (in der Originalversion 3DW oder der adaptierten Version des 3DW- π) ergab eine mittlere Bearbeitungszeit von $M =$

31.6 Minuten mit einer Standardabweichung von $SD = 18.45$. Die schnellste Person brauchte nur sechs Minuten, während die langsamste Person 113 Minuten benötigte. Der häufigste Wert war 22 Minuten. Die Unterschiede der Bearbeitungszeit in den vier Untersuchungsgruppen sind in Tabelle 6 abgebildet.

Tabelle 6

Testdauer in den vier Untersuchungsgruppen

Gruppe	N	Mittelwert	Std. Error
UG 1	45	31.04	18.98
UG 2	49	30.80	15.05
UG 3	47	28.68	18.02
UG 4	49	35.60	21.13

Die kürzeste Zeit wurde mit einem Mittelwert von $M = 28.68$ ($SD = 18.02$) in UG3 erzielt. Der höchste Wert ist mit einem Mittelwert von $M = 35.6$ ($SD = 21.13$) in UG4 vorgekommen. Ein Vergleich der Mittelwerte mittels ANOVA zeigte keine signifikanten Unterschiede zwischen den Gruppen ($F(3) = 1.26$, $p = .29$). Ein direkter Vergleich zwischen UG3 und UG4 mittels t-Test zeigte einen Mittelwertsunterschied von 7.05 Minuten zugunsten der UG3. Dieser Unterschied war jedoch nicht signifikant ($t(94) = -1.76$, $p = .08$). Es konnte somit kein signifikanter Unterschied in der Testdauer zwischen den Gruppen festgestellt werden. In UG3 und UG4 wurden jeweils nur Items im Design des 3DW oder 3DW- π bearbeitet, weswegen dieser Vergleich gesondert betrachtet wurde. Die Gruppen UG3 und UG4 unterschieden sich nicht signifikant in ihrer Testdauer.

6. Diskussion

In dieser Arbeit wurde die Frage behandelt, ob sich ein adaptiertes Design des Dreidimensionalen Würfeltests von Gittler (1990) als äquivalent im Sinne der Rasch-Homogenität zur aktuellen Version verhält. Außerdem wurde die Frage geklärt, ob ein Designparameter die Ergebnisse beeinflusst. Außerdem sollten auch erste Rückschlüsse auf die Motivation der Testpersonen gefunden werden. In Nebenfragestellungen wurde zudem ein Blick auf den Gesamtscore der Testpersonen sowie deren Bearbeitungszeiten geworfen, um weitere Hinweise auf Unterschiede zu finden.

Dazu wurden zwei Versionen des 3DW vorgegeben, die PC-Version im originalen Design (3DW) sowie eine adaptierte und interaktivere Version (3DW- π). Außerdem wurde ein speziell adaptierter Arbeitsmotivationsfragebogen verwendet, um auf Unterschiede in der Motivation, der Testangst, des Gefallens sowie der Erfolgserwartung schließen zu können. Das Versuchsdesign wurde so gewählt, dass eine Verrechnung beider Versionen möglich war. Zwei Gruppen bearbeiteten jeweils die adaptierte und originale Version, während zwei weitere Gruppen entweder den adaptierten oder den originalen Test bearbeiteten. Die Datenerhebung fand im Sommer 2021 statt und wurde vollständig online durchgeführt. Die Versuchspersonen konnten die Testbatterie am eigenen PC zu Hause bearbeiten. Ausgewertet wurden die Ergebnisse mittels der Statistikprogramme R und SPSS. Mithilfe des dichotom logistischen Modells von Rasch und des LLTM im Rahmen der Item-Response Theorie sowie unter Zuhilfenahme von Deskriptiv- und Inferenzstatistik wurden die Daten analysiert, um die Fragestellungen dieser Arbeit zu beantworten. Die oben dargestellten Ergebnisse werden in Folge noch einmal zusammengefasst und diskutiert.

6.1 Interpretation

6.1.1 Rasch-Homogenität

Die Hauptfragestellung der Arbeit war, ob das dichotom logistische Rasch Modell die vorliegenden Daten beschreibt. Das wäre im Sinne der Neukonstruktion des Tests ein erfreuliches Ergebnis, da der Anwendung des neuen Designs in der Praxis nichts im Wege stünde. Es konnte gezeigt werden, dass dies der Fall ist und ein vollständiges Rasch Modell die Daten ebenso gut beschreibt wie alle flexibleren Modelle, die geprüft wurden. Dieses Ergebnis wurde erwartet und ist ein Beleg für die Robustheit der verwendeten Items. Diese ist bereits in der Vergangenheit belegt worden, als beispielsweise die Paper-Pencil Version auf eine PC-Version umgestellt (Arrer, 1992) oder die Einführung eines Farbparameters untersucht wurde (Gittler & Arendasy, 2003). Das nicht signifikante Ergebnis des Vergleichs der Itemparameter zwischen den beiden Testversionen sowie die hohe Korrelation dieser bestätigt die Ergebnisse der Modellvergleiche und zeigt ein homogenes Bild der Daten.

6.1.2 Designeffekt

In einer weiteren Fragestellung wurde untersucht, ob es einen möglichen Designeffekt gibt, der das Ergebnis über alle Items hinweg beeinflusst. Die Ergebnisse zeigen, dass es keinen Designeffekt gibt. Modellvergleiche mit einem LLTM mit einem Designparameter führten zu keinem signifikanten Ergebnis. Die Einführung eines Designparameters brachte demnach keinen signifikanten Vorteil gegenüber einem strengeren vollständigen Rasch

Modell. Konkret bedeutet das, dass das bearbeitete Testdesign (3DW oder 3DW- π) keinen Einfluss auf die Testleistung einer Person hatte. Dies hätte sich besonders auf die Erstellung von Normwerttabellen ausgewirkt, die sich von den bisherigen Versionen unterscheiden hätten. Das Ergebnis ist überraschend, da vermutet wurde, dass ein interaktives Design zu höheren Personenparametern führt (Cantoni et al., 2004), eventuell auch motivationsbedingt (Gao, 2003). Die hohe Korrelation der Itemparameter, der nicht signifikante Mittelwertsunterschied zwischen den zwei Test-Versionen sowie die nicht signifikanten Modellvergleiche sind Belege, dass kein Designeffekt vorliegt. Ein möglicher Grund dafür sind die wie erwartet sehr robusten Items des 3DW. Die Schwierigkeit der Items wird offensichtlich auch durch die Änderung des Designs nicht maßgeblich beeinflusst.

6.1.3 Arbeitsmotivation

Hinsichtlich der Arbeitsmotivation wurden keine signifikanten Unterschiede zwischen den beiden Testversionen festgestellt. Dies wurde sowohl auf Itemebene als auch auf Faktorebene geprüft. Das Ergebnis ist mit Vorsicht zu betrachten, da die Daten nur lückenhaft vorlagen. Dass nur Daten für den ersten Testteil vorhanden waren, schmälert daher die Aussagekraft der Ergebnisse. Die verwendete Stichprobe war für einen ersten Vergleich allerdings ausreichend groß. Demnach bearbeiteten in T1 92 Personen den 3DW und 103 Personen den 3DW- π . Die Ergebnisse lieferten einen ersten Hinweis darauf, dass möglicherweise keine bis kaum Unterschiede in der Motivation von Testpersonen vorhanden waren, wenn sie die beiden unterschiedlichen Versionen des 3DW bearbeiteten. Diese Schlussfolgerung ist eher überraschend, da durchaus Unterschiede zu erwarten waren. Dennoch steht das Ergebnis der weiteren Verwendung des 3DW- π nicht im Wege, da dieser noch zahlreiche weitere Vorteile bietet, wie eine verbesserte Testsicherheit und eine optimierte Datengewinnung.

6.1.4 Nebenfragestellungen

Die Auswertung der Gesamtscores nach erfolgreicher Rasch-Analyse brachte keinerlei überraschende Ergebnisse. Der Gesamtscore der Stichprobe war gleichmäßig verteilt mit einem Hang zu guten und sehr guten Ergebnissen. Zum Beispiel lag der Modalwert beim höchsten zu erreichenden Rohwert von 19 Punkten. Eine Erklärung dafür könnte der hohe Bildungsgrad der Stichprobe mit 59% Akademiker*innen und kaum Versuchspersonen ohne Matura (12.4%) liefern. Generell waren die beiden Scores am Anfang und Ende der Skala (0 und 19) überrepräsentiert, da auch der Wert 0 von 3.6% der Versuchspersonen erreicht wurde. Das könnte auf wenig motivierte Testpersonen hinweisen, die die Items nur durchgeklickt

haben. Bei einer zufälligen Auswahl von Antworten würde man allerdings im Schnitt 2.70 Punkte erzielen (berechnet aus 19 Items und 7 möglichen Antworten). Viel wahrscheinlicher ist, dass sich manche Versuchspersonen zwar Mühe geben, den Test gut zu bearbeiten, aufgrund mangelnder Fähigkeiten aber einfach nicht gut abschneiden und falsche Entscheidungen treffen. Die guten Ergebnisse, speziell die vielen Testpersonen mit der höchsten Punktezahl (19), lassen auf eine motivierte Stichprobe schließen, die sich bei der Bearbeitung Mühe gegeben hat.

Ein Vergleich nach Geschlechtern fiel signifikant aus und zeigte einen Mittelwertsunterschied von 3.49 Punkten zugunsten der Männer. Dieses Ergebnis war zu erwarten und ist im Einklang mit bisherigen Ergebnissen. Für genauere Analysen müssten jedoch vor allem die Personenparameter genauer analysiert werden, weshalb dieses Ergebnis in dieser Arbeit nur eine Randnotiz bleibt. Ein Vergleich der Gruppen nach Altersgruppen wurde nach einem Mediansplit durchgeführt. Er führte zu keinem signifikanten Ergebnis. Auf einen detaillierten Vergleich der Altersgruppen wurde aufgrund der jungen Stichprobe verzichtet. Für weitere Analysen, wie zum Beispiel Clusteranalysen, hätte es eine anders verteilte Stichprobe benötigt. Wären zum Beispiel ähnlich große Gruppen von Personen in den Zwanzigern, Dreißigern und so weiter vorhanden, könnte man mittels Clusteranalysen mögliche Alterseffekte feststellen.

Ein Vergleich der Mittelwerte des Gesamtscores in Abhängigkeit der bearbeiteten Testversion (3DW oder 3DW- π) zeigte keine signifikanten Unterschiede. Auch die Einbeziehung der gemischten Untersuchungsgruppen 3 und 4 führte zu keinem signifikanten Unterschied. Dies ist ein weiterer Hinweis auf fehlende Unterschiede in der Schwierigkeit der verwendeten Designs. Die Ergebnisse sind in Einklang mit den Hauptfragestellungen dieser Arbeit, die keine nennenswerten Unterschiede in den Versionen feststellen konnten.

Die Auswertung der Gesamtbearbeitungszeit wurde nur für 190 Testpersonen vorgenommen, da fünf Personen aufgrund unrealistischer Werte ausgeschlossen wurden. Dies war der Fall, weil diese eine besonders hohe Bearbeitungszeit von über 120 Minuten aufwiesen, eine Testperson erreichte sogar einen Wert von über 1000 Minuten. Ein Grund dafür war die nicht kontrollierte Testumgebung, die in dieser Untersuchung vorlag. Es war zum Beispiel möglich, zwischen den Aufgaben eine Pause zu machen und später fortzufahren. Dies und mögliche Messfehler könnten eine Erklärung für die fünf extremen Werte sein, die ausgeschlossen wurden, um das Ergebnis sinnvoll interpretieren zu können. Die mittlere Bearbeitungszeit lag bei 31.6 Minuten. Das ist deutlich höher als der Mittelwert von $M =$

22.44, der im Manual des 3DW angegeben ist (Gittler, 1990). Dies könnte mehrere Gründe haben. Zum einen wurden in dieser Testung mit 20 Items eine Aufgabe mehr bearbeitet als in der Originalversion des 3DW. Außerdem kann es, wie oben beschrieben, durch die unkontrollierte Testumgebung zu Verzögerungen in der Testumgebung der Testpersonen gekommen sein. Zuletzt sei erwähnt, dass jeweils unterschiedliche Items des 3DW verwendet wurden. Die verwendeten Items in dieser Untersuchung stammen aus dem Pool des A3DW und sind damit nicht ident mit jenen im originalen 3DW. Da die einzelnen Items unterschiedliche mittlere Bearbeitungszeiten haben, wirken sie sich auch auf die Gesamtdauer einer Testung aus. Ein Vergleich wäre also nur mittels mittlerer Bearbeitungszeiten der einzelnen Items sinnvoll. Ein weiterer Aspekt für zukünftige Untersuchungen, den es sich lohnen würde genauer zu betrachten. Die adaptierte Version des 3DW- π bot zudem neue Möglichkeiten der Interaktion, die ebenfalls zu einer höheren Bearbeitungszeit führen könnten. Ein Vergleich der vier Untersuchungsgruppen ergab jedoch keine signifikanten Mittelwertsunterschiede. Ein Vergleich der beiden Untersuchungsgruppen, die jeweils nur eine Version über die gesamte Testung hinweg bearbeiteten (UG3 und UG4), zeigte einen Unterschied von 7.05 Minuten zugunsten des 3DW. Personen, die den 3DW- π bearbeiteten, brauchten im Schnitt um knapp 7 Minuten länger. Allerdings war dieser Unterschied nicht signifikant und womöglich nur ein Zufall. Weitere Untersuchungen sollten sich dieser Fragestellung noch ausführlicher widmen. Wie oben bereits erwähnt könnte auf Ebene der einzelnen Items ebenfalls spannende Erkenntnisse zu erwarten sein.

6.2 Ausblick

Aufgrund der Ergebnisse in dieser Arbeit ergeben sich eine Reihe von weiteren Forschungsmöglichkeiten im Hinblick auf den 3DW- π . In zukünftigen Untersuchungen sollte vor allem eine möglichst große und repräsentative Stichprobe angestrebt werden, um etwaige Unterschiede zwischen den verschiedenen Personengruppen zu finden. Gerade im Bereich der Altersunterschiede könnten interessante Ergebnisse zu erwarten sein. Es wäre spannend zu untersuchen, ob die Arbeitsmotivation und das Gefallen bei jüngeren Personen höher ausgeprägt sind als bei älteren Personen. Aufgrund der Interaktivität des Designs, das dem modernen Alltag näherkommt, kann das durchaus zu erwarten sein. Hierfür würde sich die Analyse von verschiedenen Altersgruppen eignen. Die einzelnen Personengruppen müssten allerdings eine ausreichend große Anzahl an Testpersonen beinhalten, um sinnvolle Ergebnisse zu erhalten. Besonders auf die Stichprobenverteilung müsste daher schon in der Datenerhebung geachtet werden.

Da der 3DW- π in dieser Untersuchung zum ersten Mal zum Einsatz kam, wird jede weitere Untersuchung mehr Daten und damit mehr Einblicke bringen. Sinn ergeben würde zum Beispiel auch eine qualitative Befragung von Testpersonen vor, während und nach der Testung. Dadurch könnte ein tieferer Einblick in die Motivation und die Befindlichkeit der Testpersonen gewonnen werden. Gemeinsam mit den gespeicherten Daten im log-file wären hier auch spannende Erkenntnisse über die verwendeten Strategien zu finden. Besonders schlechte und gute Distraktoren könnten auf diese Weise identifiziert und gegebenenfalls ersetzt werden. Auf Ebene der Einzelitems könnten sich ebenfalls Effekte zeigen. Es wäre beispielsweise möglich, dass das neue Design nur bei einzelnen Items einen Unterschied macht und andere robuster sind. Die Bearbeitungszeiten und die vorhandenen Daten zur Interaktion halten viel Information für weitere Untersuchungen bereit.

Auch eine genauere Betrachtung der Geschlechtsunterschiede, anhand von Personenparametern, wäre in einer Folgearbeit denkbar und sinnvoll. Die Bearbeitungszeit von einzelnen Items und die Gesamtdauer könnten ebenfalls ein möglicher Fokus sein. Auf diese und weitere Analysen wurde in vorliegender Untersuchung aufgrund des Umfangs verzichtet. Dass die testtheoretischen Voraussetzungen für die Verwendung des 3DW- π gegeben sind, konnte in dieser Arbeit gezeigt werden. Das ist ein Ergebnis, das für die Robustheit der Testitems des 3DW spricht. Dadurch sind nicht nur weitere Forschungen gerechtfertigt, sondern auch die Anwendung des 3DW- π in der Praxis ist ein Stück näher gerückt. Damit steht einer spannenden Zukunft des 3DW nichts im Wege.

Literaturverzeichnis

- AMS - Berufsinformationssystem (2023, März 13). *Räumliches Vorstellungsvermögen*.
https://bis.ams.or.at/bis/kompetenz/4665-R%C3%A4umliches+Vorstellungsverm%C3%B6gen?stammdaten_noteid=132
- Amthauer, R. (1973). *I-S-T 70: Intelligenz-Struktur-Test* (4., überarb. Aufl.). Hogrefe.
- Andersen, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, 38(1), 123–140. <https://doi.org/10.1007/BF02291180>
- Arrer, S. (1992). *Unterschiede zwischen Computer- und Papier-Bleistift-Vorgabe des 3DW*. Diplomarbeit. Universität Wien.
- Asendorpf, J. B. (2004). *Persönlichkeit im Kulturvergleich*. Springer.
https://doi.org/10.1007/978-3-662-09570-6_8
- Auer, J., & Pircher, P. (1988). *Räumliche Orientierungsleistung in einer Großstadt und Raumvorstellungsfähigkeit*. Planungspraktikumsbericht. Universität Wien.
- Baenninger, M., & Newcombe, N. (1989). The role of experience in spatial test performance: A meta-analysis. *Sex Roles: A Journal of Research*, 20(5-6), 327–344.
<https://doi.org/10.1007/BF00287729>
- Barratt, E. S. (1953). An Analysis of Verbal Reports of Solving Spatial Problems as an Aid in Defining Spatial Factors. *The Journal of Psychology*, 36(1), 17–25.
<https://doi.org/10.1080/00223980.1953.9712874>
- Bennett, G. K. (1969). *Bennett Mechanical Comprehension Test (BMCT)*. Psychological Corporation.
- Berkowitz, M., & Stern, E. (2018). Which cognitive abilities make the difference? Predicting academic achievements in advanced stem studies. *Journal of Intelligence*, 6(4), 1–24.
<https://doi.org/10.3390/jintelligence6040048>
- Bortz, J., & Döring, N. (2006). Quantitative Methoden der Datenerhebung. In J. Bortz & N. Döring (Hrsg.), *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (S. 137-293). Springer. https://doi.org/10.1007/978-3-540-33306-7_4
- Brandt, M. G., & Wright, E. D. (2005, Juni). Medical student career choice and mental rotations ability. *Clinical and investigative medicine*, 28(3), 112-117.

- Campbell, M. J., Toth, A. J., & Brady, N. (2018). Illuminating sex differences in mental rotation using pupillometry. *Biological Psychology*, 138, 19–26.
<https://doi.org/10.1016/j.biopsycho.2018.08.003>
- Cantoni, V., Cellario, M., & Porta, M. (2004). Perspectives and challenges in e-learning: towards natural interaction paradigms. *Journal of Visual Languages & Computing*, 15(5), 333–345. <https://doi.org/10.1016/j.jvlc.2003.10.002>
- Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies* (Illustrated). Cambridge University Press.
<https://doi.org/10.1017/CBO9780511571312>
- Casey, M. (1996). Understanding Individual Differences in Spatial Ability within Females: A Nature/Nurture Interactionist Framework. *Developmental Review*, 16(3), 241–260.
<https://doi.org/10.1006/drev.1996.0009>
- Casey, B. M., Andrews, N., Schindler, H., Kersh, J. E., Samper, A., & Copley, J. (2008). The Development of Spatial Skills Through Interventions Involving Block Building Activities. *Cognition and Instruction*, 26(3), 269–309.
<https://doi.org/10.1080/07370000802177177>
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1), 1–22. <https://doi.org/10.1037/h0046743>
- Cherney, I. D., Kelly-Vance, L., Gill Glover, K., Ruane, A., & Oliver Ryalls, B. (2003). The Effects of Stereotyped Toys and Gender on Play Assessment in Children Aged 18-47 Months. *Educational Psychology*, 23(1), 95–106.
<https://doi.org/10.1080/01443410303222>
- Cherney, I. D., & London, K. (2006). Gender-linked Differences in the Toys, Television Shows, Computer Games, and Outdoor Activities of 5- to 13-year-old Children. *Sex Roles*, 54(9–10), 717–726. <https://doi.org/10.1007/s11199-006-9037-8>
- Eals, M., & Silverman, I. (1994). The Hunter-Gatherer theory of spatial sex differences: Proximate factors mediating the female advantage in recall of object arrays. *Ethology and Sociobiology*, 15(2), 95–105. [https://doi.org/10.1016/0162-3095\(94\)90020-5](https://doi.org/10.1016/0162-3095(94)90020-5)
- Eliot, J., & Smith I. M. (1983). *An international directory of spatial tests*. Nfer-Nelson.

- Etaugh, C., & Riley, S. (1983). Evaluating competence of women and men: Effects of marital and parental status and occupational sex-typing. *Sex Roles*, 9(9), 943–952.
<https://doi.org/10.1007/bf00290055>
- Fischer, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6), 359–374. [https://doi.org/10.1016/0001-6918\(73\)90003-6](https://doi.org/10.1016/0001-6918(73)90003-6)
- Fischer, G. (1974). Einführung in die Theorie psychologischer Tests: Grundlagen und Anwendungen. Huber. <https://doi.org/10.30965/25890581-05304010>
- Fischer, G. H., & Formann, A. K. (1982). Veränderungsmessung mittels linear-logistischer Modelle. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 3(2), 75–99.
<https://psycnet.apa.org/record/1983-22500-001>
- Fischer, G. H., & Molenaar, I.W. (1995). Estimation of Item Parameters. In G. H. Fischer & I. W. Molenaar (Hrsg.), *Rasch Models: Foundations, Recent Developments, and Applications* (S.39-51). https://doi.org/10.1007/978-1-4612-4230-7_3
- Formann, A. K., & Piswanger, K. (1979). *WMT. Wiener Matrizen-Test. Ein Raschskalierter sprachfreier Intelligenztest. Manual*. Beltz-Test.
- French, J. W. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement*, 25(1), 9-28.
<https://doi.org/10.1177/001316446502500102>
- Funke, J., & Vaterrodt-Plünnecke, B. (2004). *Was ist Intelligenz?* (2., überarb. Aufl.). Beck.
- Gao, T. (2003). The effects of different levels of interaction on the achievement and motivational perceptions of college students in a web-based learning environment. *Journal of Interactive Learning Research*, 14(4), 367-386.
- Gardner, H. E. (2011). *Frames of minds. The theory of multiple intelligence*. Basic Books.
- Geiser, C., Lehmann, W., & Eid, M. (2008). A note on sex differences in mental rotation in different age groups. *Intelligence*, 36(6), 556-563.
<https://doi.org/10.1016/j.intell.2007.12.003>
- Gittler, G. (1978). *Analyse erfolgskontingenten Lernens unter verschiedenen Rückmeldungsbedingungen*. Unveröffentlichte Dissertation. Universität Wien.

- Gittler, G. (1990). *Dreidimensionaler Würfeltest (3DW). Ein Rasch-skaliertes Test zur Messung des räumlichen Vorstellungsvermögens, Theoretische Grundlagen und Manual*. Beltz Test.
- Gittler, G. (1994). Intelligenzförderung durch Schulunterricht: Darstellende Geometrie und räumliches Vorstellungsvermögen. In G. Gittler, M. Jirasko, U. Kastner-Koller, C. Korunka & A. Al-Roubaie (Hrsg.), *Die Seele ist ein weites Land. Aktuelle Forschung am Wiener Institut für Psychologie* (S.105-122). WUV.
- Gittler, G. (1999). Sind Raumvorstellung und Reasoning separierbare Fähigkeitsdimensionen? Dimensionalitätsanalysen zweier Rasch-skalierten Tests: 3DW und WMT. *Diagnostica* 45(2), 69–81. <https://doi.org/10.1026//0012-1924.45.2.69>
- Gittler, G., & Alferts, T. (2021a). *3DW- π 2021*. Unveröffentlichtes Manuskript und Online-Test-Batterie.
- Gittler, G., & Alferts, T. (2021b). *Motivationsfragebogen 2021*. Unveröffentlichtes Manuskript.
- Gittler, G., & Arendasy, M. (2003). Psychometrische Effekte von Farbgebung. *Zeitschrift Für Differentielle Und Diagnostische Psychologie*, 24(1), 23–31. <https://doi.org/10.1024//0170-1789.24.1.23>
- Gittler, G., & Fischer, G. (2011). IRT-Based Measurement of Short-Term Changes of Ability, With an Application to Assessing the "Mozart Effect". *Journal of Educational and Behavioral Statistics*, 36(1), 33–75. <https://doi.org/10.3102/1076998610366260>
- Goldberg, E., & Costa, L. D. (1981). Hemisphere differences in the acquisition and use of descriptive systems. *Brain and Language*, 14(1), 144–173. [https://doi.org/10.1016/0093-934x\(81\)90072-9](https://doi.org/10.1016/0093-934x(81)90072-9)
- Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.
- Guillot, A., Champely, S., Batier, C., Thiriet, P., & Collet, C. (2007). Relationship between spatial abilities, mental rotation and functional anatomy learning. *Advances in Health Sciences Education*, 12(4), 491–507. <https://doi.org/10.1007/s10459-006-9021-7>
- Guizzo, F., Moè, A., Cadinu, M., & Bertolli, C. (2019). The role of implicit gender spatial stereotyping in mental rotation performance. *Acta Psychologica*, 194, 63–68. <https://doi.org/10.1016/j.actpsy.2019.01.013>

- Gur, R. C., Alsop, D., Glahn, D., Petty, R., Swanson, C. L., Maldjian, J. A., Turetsky, B. I., Detre, J. A., Gee, J., & Gur, R. E. (2000). An fMRI Study of Sex Differences in Regional Activation to a Verbal and a Spatial Task. *Brain and Language*, 74(2), 157–170. <https://doi.org/10.1006/brln.2000.2325>
- Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, 19(2), 149–161. <https://doi.org/10.1007/bf02289162>
- Griksiene, R., Monciunskaitė, R., Arnatkeviciute, A., & Ruksenas, O. (2018). Does the use of hormonal contraceptives affect the mental rotation performance? *Hormones and Behavior*, 100, 29–38. <https://doi.org/10.1016/j.yhbeh.2018.03.004>
- Hegarty, M., & Waller, D. (2005). Individual Differences in Spatial Abilities. In P. Shah & A. Miyake (Hrsg.), *The Cambridge Handbook of Visuospatial Thinking* (S.121-169). <https://doi.org/10.1017/CBO9780511610448.005>
- Howard, R., Fenwick, P., Brown, D., & Norton, R. (1992). Relationship between CNV asymmetries and individual differences in cognitive performance, personality and gender. *International Journal of Psychophysiology*, 13(3), 191–197. [https://doi.org/10.1016/0167-8760\(92\)90069-n](https://doi.org/10.1016/0167-8760(92)90069-n)
- Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven [Structural research on intelligence: Competing models, new developments, perspectives]. *Psychologische Rundschau*, 35(1), 21–35.
- Jäger, A. O., Süß, H. M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test: BIS-Test*. Hogrefe.
- Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, 92(2), 137–172. <https://doi.org/10.1037/0033-295X.92.2.137>
- Kail, R., Carter, P., & Pellegrino, J. (1979). The locus of sex differences in spatial ability. *Perception & Psychophysics*, 26(3), 182–186. <https://doi.org/10.3758/bf03199867>
- Kalun, P., Dunn, K., Wagner, N., Pulakunta, T., & Sonnadara, R. (2020). Recent evidence on visual-spatial ability in surgical education: A scoping review. *Canadian medical education journal*, 11(6), e111–e127. <https://doi.org/10.36834/cmej.69051>

- Kaufman, S. B. (2007). Sex differences in mental rotation and spatial visualization ability: Can they be accounted for by differences in working memory capacity? *Intelligence*, 35(3), 211–223. <https://doi.org/10.1016/j.intell.2006.07.009>
- Kisser, R., & Wenninger, U. (1983). Computergestütztes Testen im Rahmen der Fahreignungsdiagnostik (Act & React Testsystem ART-90). *Kleine Fachbuchreihe des Kuratoriums für Verkehrssicherheit 20*. Kuratorium für Verkehrssicherheit.
- Köller, O., Rost, J., & Köller, M. (1994): Individuelle Unterschiede beim Lösen von Raumvorstellungsaufgaben aus dem IST—bzw. IST-70-Untertest “Würfelaufgaben”. *Zeitschrift für Psychologie*, 202(1), 65–85.
- Kubinger, K. D. (2019). *Psychologische Diagnostik. Theorie und Praxis psychologischen Diagnostizierens*. Hogrefe. <https://doi.org/10.1026/02779-000>
- Leopold, C., Gorska, R. A., & Sorby, S. A. (2001). International experiences in developing the spatial visualization abilities of engineering students. *Journal for Geometry and Graphics*, 5(1), 81-91.
- Lienert, G. A. (1958). *F-L-T: Form-Lege-Test*. Hogrefe.
- Linn, M. C., & Petersen, A. C. (1985). Emergence and Characterization of Sex Differences in Spatial Ability: A Meta-Analysis. *Child Development*, 56(6), 1479–1498. <https://doi.org/10.2307/1130467>
- Lohman, D. F. (1979). *Spatial Ability: A Review and Reanalysis of the correlational literature*. Stanford University School of Education.
- Lubinski, D. (2010). Spatial ability and STEM: A sleeping giant for talent identification and development. *Personality and Individual Differences*, 49(4), 344–351. <https://doi.org/10.1016/j.paid.2010.03.022>
- Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford University Press.
- Maeda, Y., & Yoon, S. Y. (2013). A Meta-Analysis on Gender Differences in Mental Rotation Ability Measured by the Purdue Spatial Visualization Tests: Visualization of Rotations (PSVT:R). *Educational Psychology Review*, 25(1), 69–94. Springer Science and Business Media, LLC. <https://doi.org/10.1007/s10648-012-9215-x>

- Maier, P. H. (1994). *Räumliches Vorstellungsvermögen: Komponenten, geschlechtsspezifische Differenzen, Relevanz, Entwicklung und Realisierung in der Realschule*. Internationaler Verlag Der Wissenschaften.
<https://doi.org/10.1007/BF03338824>
- Maltby, J., Day, L. & Macaskill, A. (2011). *Differentielle Psychologie, Persönlichkeit und Intelligenz*. Pearson Deutschland GmbH.
- Maroco, J., & Bártoło-Ribeiro, R. (2013). Selection of air force pilot candidates: A case study on the predictive accuracy of discriminant analysis, logistic regression, and four neural network types. *The International Journal of Aviation Psychology*, 23(2), 130-152. <https://doi.org/10.1080/10508414.2013.772837>
- Martin-Löf, P. (1973). *Statistika Modeller: Anteckningar fran seminarier Lasaret 1969-1970, utarbetade av Rolf Sunberg. Obetydligt ändrat nytryck, oktober 1973*. Institutet för s kringsmatematisk och matematisk vid Stockholms universitet.
- Mo , A. (2018). Mental rotation and mathematics: Gender-stereotyped beliefs and relationships in primary school children. *Learning and Individual Differences*, 61, 172–180. <https://doi.org/10.1016/j.lindif.2017.12.002>
- Mo , A., Jansen, P., & Pietsch, S. (2018). Childhood preference for spatial toys. Gender differences and relationships with mental rotation in STEM and non-STEM students. *Learning and Individual Differences*, 68, 108–115.
<https://doi.org/10.1016/j.lindif.2018.10.003>
- Mo , A. (2021). Doubling mental rotation scores in high school students: Effects of motivational and strategic trainings. *Learning and Instruction*, 74, 101-461.
<https://doi.org/10.1016/j.learninstruc.2021.101461>
- Moore, D. S., & Johnson, S. P. (2008). Mental Rotation in Human Infants: A Sex Difference. *Psychological Science*, 19(11), 1063–1066. <https://doi-org./10.1111/j.1467-9280.2008.02200.x>
- Moore, D. S., & Johnson, S. P. (2020). The development of mental rotation ability across the first year after birth. *Advances in Child Development and Behavior*, 58, 1–33.
<https://doi.org/10.1016/bs.acdb.2020.01.001>

- Moosbrugger, H. (2012). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S.227-274). Springer.
https://doi.org/10.1007/978-3-642-20072-4_10
- Müller, Kurt P. (1986). Raumvorstellung. Was ist das, u. warum ist sie wichtig? In *Pädagogische Welt*, 40 (S.23-26 u. 31) Pädagogische Stiftung Cassianeum.
- Newcombe, N., Bandura, M. M., & Taylor, D. G. (1983). Sex differences in spatial ability and spatial activities. *Sex Roles*, 9(3), 377–386. <https://doi.org/10.1007/bf00289672>
- Ozel, S., Larue, J., & Molinaro, C. (2002). Relation between sport activity and mental rotation: Comparison of three groups of subjects. *Perceptual and Motor Skills*, 95(3), 1141-1154. <https://doi.org/10.2466/pms.2002.95.3f.1141>
- Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157–175.
<https://doi.org/10.1080/14786440009463897>
- Peragine, D., Simeon-Spezzaferro, C., Brown, A., Gervais, N. J., Hampson, E., & Einstein, G. (2020). Sex difference or hormonal difference in mental rotation? The influence of ovarian milieu. *Psychoneuroendocrinology*, 115, 104-488.
<https://doi.org/10.1016/j.psyneuen.2019.104488>
- Provo, J., Lamar, C., & Newby, T. (2002). Using a cross section to train veterinary students to visualize anatomical structures in three dimensions. *Journal of Research in Science Teaching*, 39(1), 10–34. <https://doi.org/10.1002/tea.10007>
- Putz-Osterloh, W. (1977). Über Problemlöseprozesse bei dem Test Würfelaufgaben aus dem Intelligenztest IST und IST-70 von Amthauer. *Diagnostika*, 23, 252-265.
- Putz-Osterloh, W., & Lüer, G. (1979). Do subjects produce spatial images while taking spatial tests? *Zeitschrift für Experimentelle und Angewandte Psychologie*, 26(1), 138–156.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. The Danish Institute for Educational Research.

- Roberts, J. E., & Bell, M. A. (2000). Sex differences on a computerized mental rotation task disappear with computer familiarization. *Perceptual and Motor Skills*, 91(3, Pt 2), 1027–1034. <https://doi.org/10.2466/PMS.91.7.1027-1034>
- Rost, D. H. (1977). *Raumvorstellung: Psychologische und pädagogische Aspekte*. Beltz.
- Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2., überarb. Aufl.). Huber.
- Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. *Zeitschrift für experimentelle und angewandte Psychologie*, 19, 476–506.
- Schuhfried, G. (2005). *Aviation Psychology. Test battery based on JAR-FCL3*. Dr. G. Schuhfried Ges.m.b.H.
- Schultz, K. (1991). The contribution of solution strategy to spatial performance. *Canadian Journal of Psychology / Revue canadienne de psychologie*, 45(4), 474–491. <https://doi.org/10.1037/h0084301>
- Sherman, J. A. (1967). Problem of sex differences in space perception and aspects of intellectual functioning. *Psychological Review*, 74(4), 290–299. <https://doi.org/10.1037/h0024723>
- Shirazi, T. N., Levenberg, K., Cunningham, H., Self, H., Dawood, K., Cárdenas, R., Ortiz, T. L., Carré, J. M., Breedlove, S. M., & Puts, D. A. (2021). Relationships between ovarian hormone concentrations and mental rotations performance in naturally-cycling women. *Hormones and Behavior*, 127, 104-886. <https://doi.org/10.1016/j.yhbeh.2020.104886>
- Silverman, I., Choi, J., & Peters, M. (2007). The Hunter-Gatherer Theory of Sex Differences in Spatial Abilities: Data from 40 Countries. *Archives of Sexual Behavior*, 36(2), 261–268. <https://doi.org/10.1007/s10508-006-9168-6>
- Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *The American Journal of Psychology*, 15(2), 201–292. <https://doi.org/10.2307/1412107>
- Spearman, C. (1927). The Abilities of Man their Nature and Measurement. *Nature*, 120(3014), 181–183. <https://doi.org/10.1038/120181a0>
- Stumpf, H., & Fay, E. (1983). *Schlauchfiguren. Ein Test zur Beurteilung des räumlichen Vorstellungsvermögens*. Handanweisung. Hogrefe.

- Tarampi, M. R., Heydari, N., & Hegarty, M. (2016). A tale of two types of perspective taking: Sex differences in spatial ability. *Psychological Science*, 27(11), 1507-1516.
<https://doi.org/10.1177/0956797616667459>
- Thurstone, L.L. (1938). *Primary mental abilities*. University of Chicago Press.
- Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs*, 2, 94.
- Vernon, P. E. (1950). *The structure of human abilities*. Methuen.
- Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, 117(2), 250–270. <https://doi.org/10.1037/0033-2909.117.2.250>
- Voyer, D., Voyer, S. D., & Saint-Aubin, J. (2017). Sex differences in visual-spatial working memory: A meta-analysis. *Psychonomic Bulletin and Review*, 24(2), 307–334.
<https://doi.org/10.3758/s13423-016-1085-7>
- Wei, W., Yuan, H., Chen, C., & Zhou, X. (2012). Cognitive correlates of performance in advanced mathematics. *The British journal of educational psychology*, 82(1), 157–181. <https://doi.org/10.1111/j.2044-8279.2011.02049.x>
- Witte, K. (2019). Grundlagen der Testtheorie und Testkonstruktion. In K. Witte (Hrsg.), *Angewandte Statistik in der Bewegungswissenschaft* (Band 3) (S.139-158). Springer.
https://doi-org.uaccess.univie.ac.at/10.1007/978-3-662-58360-9_10

Abbildungsverzeichnis

Abbildung 1	Beispielitem aus dem 3DW (Paper-Pencil Version)	25
Abbildung 2	Beispielitem aus dem 3DW (Originaldesign)	29
Abbildung 3	Beispielitem aus dem 3DW- π (adaptiertes Design)	30
Abbildung 4	Altersverteilung der Stichprobe	36
Abbildung 5	ICC-Kurve von drei Items	41
Abbildung 6	Häufigkeiten des erreichten Punktescores im Test	51

Tabellenverzeichnis

Tabelle 1	Untersuchungsgruppen	38
Tabelle 2	Modellvergleiche zur Überprüfung der Eindimensionalität	47
Tabelle 3	Modellvergleiche zur Ermittlung eines Designeffekts	48
Tabelle 4	Itemparameter des vollständigen Rasch Modells	49
Tabelle 5	Vergleich der Summenscores des Arbeitsmotivationsfragebogens	50
Tabelle 6	Testdauer in den vier Untersuchungsgruppen	52

Abkürzungsverzeichnis

RV	Raumvorstellung
R	Rohwert
3DW	Dreidimensionaler Würfeltest
3DW- π	Dreidimensionaler Würfeltest- π (adaptierte Version)
UG (1-4)	Untersuchungsgruppe 1-4
ICC	Item-Charakteristik-Kurve
LLTM	Linear-logistisches Testmodell
IST	Intelligenz-Struktur-Test
LQT	Likelihood-Quotienten-Test
T1	Testteil 1
T2	Testteil 2
uRm	unvollständiges Rasch Modell
vRM	vollständiges Rasch Modell
mdRm	mehrdimensionales Rasch Modell

uniRm	eindimensionales Rasch Modell
LLTMdp	Linear-logistisches Testmodell mit Designparameter
MINT	Mathematik, Informatik, Naturwissenschaften, Technik
STEM	Science, Technology, Engineering, Mathematics
g	Zentralfaktor für Intelligenz
s	spezifischer Intelligenzfaktor
A3DW	adaptive Version des 3DW
2PL	Birnbaum Modell
CL	Conditional Likelihood
LR	Likelihood Ratio
Npar	Anzahl der Parameter
df	Freiheitsgrade
t	t-Wert aus dem t-Test zum Mittelwertsvergleich
χ^2 -Test	Chi-Quadrat Test
tAngst	Faktor: Testangst
tMot	Faktor: Testmotivation
tErw	Faktor: Erfolgserwartung
tGef	Faktor: Gefallen

Anhang

Anhang A

Zusammenfassung

Die Relevanz der Raumvorstellung für das tägliche Leben ist unumstritten. Daher stößt das Konstrukt nicht nur in der Forschung auf großes Interesse, sondern ist auch in der Berufsauswahl ein wichtiger Faktor. Einer der führenden Tests zur Messung des räumlichen Vorstellungsvermögens ist der Dreidimensionale Würfeltest von Gittler. Seine Items messen eindimensional im Sinne von Rasch und erweisen sich als überaus robust gegenüber verschiedenen Anpassungen. Eine solche ist auch Gegenstand dieser Arbeit. Der Dreidimensionale Würfeltest wurde in dieser Untersuchung 195 Versuchspersonen in zwei verschiedenen Versionen vorgelegt. In der Originalversion sowie einer adaptierten neuen Version, dem 3DW- π . Das neue, kreisrunde Design sowie die Möglichkeit zur Interaktion soll der nächste Evolutionsschritt im Rahmen der Weiterentwicklung des Tests sein. Diese Arbeit bildet die Grundlage für dessen Einführung, indem geprüft wird, inwiefern sich die Änderungen auf die Eindimensionalität im Sinne der Rasch-Homogenität auswirken. Zudem wurde untersucht, ob es einen möglichen Designeffekt gibt. Außerdem wurde nach ersten Hinweisen gesucht, ob sich die Motivation der Testpersonen zwischen den beiden Versionen unterscheidet. Die Analyse bestätigte, dass der Test auch mit Änderung des Designs eindimensional im Sinne von Rasch misst. Es konnte zudem kein Designeffekt festgestellt werden und auch hinsichtlich der Motivation zeigten sich vorerst keine Unterschiede. Weitere Forschungsansätze wurden diskutiert.

Abstract

The relevance of the concept of spatial ability for our daily lives is undisputed. Thus, the construct is not only of great interest in research, but also a key factor in career selection. One of the leading tests for measuring spatial abilities is Gittler's three-dimensional dice test. Its items measure one-dimensionally in the sense of Rasch and prove to be extremely robust to various adaptations. One such adaptation is also the subject of this paper. The three-dimensional dice test was presented to 195 test persons in two different versions in this study. The original version and an adapted new version, the 3DW- π . The new circular design as well as the possibility of interaction should be the next evolutionary step in the further development of the test. This work is intended to provide the basis for its introduction by examining the extent to which the change affects unidimensionality in terms of Rasch homogeneity. In addition, it was investigated whether there is a possible design effect.

Furthermore, indications were sought as to whether the subjects' motivation differs between the two versions. The analysis confirmed that even with a change in design, the test can be measured in terms of unidimensionality and therefore used. No design effect could be detected either. Regarding motivation, no differences were found for the time being. Further research approaches were discussed.

Anhang B

Items und Faktoren des Arbeitsmotivationsfragebogens

Item	Faktor	Ladung	Faktorname	Notiz
<i>Ich war bemüht, die Testaufgaben so gut wie möglich zu bearbeiten.</i>	2	-.44	Testmotivation	
<i>Ich hätte mich bei der Bearbeitung des Tests mehr bemüht, wenn für mich Konsequenzen mit dem Testergebnis verbunden wären.</i>	2	.71	Testmotivation	umgepolt
<i>Als ich die Aufgabenstellung des Tests gesehen habe, habe ich Angst bekommen, im Test schlecht abzuschneiden.</i>	1	.79	Testangst	
<i>Ich war während der Bearbeitung des Tests angespannt.</i>	1	.83	Testangst	
<i>Ich war während der Bearbeitung des Tests so nervös, dass ich nicht mehr klar denken konnte.</i>	1	.69	Testangst	
<i>Ich fühlte mich unter Druck, bei diesem Test gut abzuschneiden.</i>	1	.85	Testangst	
<i>Ich habe bei der Bearbeitung des Tests mein Bestes gegeben.</i>	2	-.76	Testmotivation	
<i>Ich habe in diesem Test gut abgeschnitten.</i>	3	.78	Erfolgserwartung	
<i>Der Test hat zu lange gedauert.</i>	4	.69	Gefallen	umgepolt
<i>Ich habe im Vergleich zu anderen Studierenden in diesem Test überdurchschnittlich gut abgeschnitten.</i>	3	.80	Erfolgserwartung	

<i>Es hat mir Freude bereitet, diesen Test zu bearbeiten.</i>	4	-.55	Gefallen	
<i>Der Test gibt Aufschluss über das räumliche Vorstellungsvermögen.</i>	4	-.47	Gefallen	ausgeschlossen
<i>Ich halte diesen Test zur Beurteilung der Eignung in technischen Berufen relevant.</i>	4	-.62	Gefallen	
<i>Ich hätte mir bei der Bearbeitung des Tests mehr Mühe geben können.</i>	2	.77	Testmotivation	umgepolt
<i>Ich empfand die Bearbeitung des Tests als Zeitverschwendung.</i>	4	.70	Gefallen	umgepolt
<i>Ich musste bei der Beantwortung von manchen Testaufgaben raten, weil ich die Lösung nicht finden konnte.</i>	3	-.56	Erfolgserwartung	umgepolt
<i>Ich habe bei der Beantwortung von manchen Testaufgaben geraten, weil ich nicht motiviert war sie ernsthaft zu beantworten.</i>	2	.63	Testmotivation	umgepolt

Anhang C

Itemparameter getrennt nach Design

Item	Itemparameter 3DW	Std. Error 3DW	Itemparameter 3DW- π	Std. Error 3DW- π
1	-1.30	.29	-.26	.26
2	.35	.26	.74	.25
3	.61	.26	.68	.25
4	.68	.26	1.11	.26
5	.09	.26	1.17	.26
6	1.25	.28	1.70	.27
7	-.04	.26	-.13	.26
8	-.30	.26	.13	.26
9	-.30	.26	-.33	.22
10	-2.49	.40	-1.71	.31
11	1.25	.26	.69	.26
12	-.44	.27	-.20	.26
13	.35	.26	.18	.26
14	-1.27	.30	-1.71	.31
15	.93	.26	.50	.26
16	-.30	.26	-.20	.26
17	-1.03	.29	-.81	.27
18	.86	.26	.89	.26
19	-.58	.27	-.74	.27

Anhang D

Mittelwertsvergleich des Arbeitsmotivationsfragebogens auf Einzelitemebene

Item	t	df	Sig.	Mittl.Diff.	Std.Error
<i>Ich war bemüht, die Testaufgaben so gut wie möglich zu bearbeiten.</i>	-.58	193	.56	-.12	.20
<i>Ich hätte mich bei der Bearbeitung des Tests mehr bemüht, wenn für mich Konsequenzen mit dem Testergebnis verbunden wären.</i>	1.28	193	.20	.33	.26
<i>Als ich die Aufgabenstellung des Tests gesehen habe, habe ich Angst bekommen, im Test schlecht abzuschneiden.</i>	-.35	193	.72	-.08	.24
<i>Ich war während der Bearbeitung des Tests angespannt.</i>	-1.04	193	.30	-.22	.21
<i>Ich war während der Bearbeitung des Tests so nervös, dass ich nicht mehr klar denken konnte.</i>	.58	193	.56	.08	.14
<i>Ich fühlte mich unter Druck, bei diesem Test gut abzuschneiden.</i>	.64	193	.52	.14	.22
<i>Ich habe bei der Bearbeitung des Tests mein Bestes gegeben.</i>	-.65	193	.51	-.10	.15
<i>Ich habe in diesem Test gut abgeschnitten.</i>	-.09	193	.93	-.02	.18
<i>Der Test hat zu lange gedauert.</i>	-.73	193	.46	-.15	.21
<i>Ich habe im Vergleich zu anderen Studierenden in diesem Test überdurchschnittlich gut abgeschnitten.</i>	.22	193	.83	.04	.18
<i>Es hat mir Freude bereitet, diesen Test zu bearbeiten.</i>	-.99	193	.33	-.21	.21
<i>Der Test gibt Aufschluss über das räumliche Vorstellungsvermögen.</i>	-.90	193	.37	-.13	.14
<i>Ich halte diesen Test zur Beurteilung der Eignung in technischen Berufen relevant.</i>	-.40	193	.69	-.07	.17
<i>Ich hätte mir bei der Bearbeitung des Tests mehr Mühe geben können.</i>	.70	193	.49	.15	.22

<i>Ich empfand die Bearbeitung des Tests als Zeitverschwendung.</i>	.34	193	.73	.05	.15
<i>Ich musste bei der Beantwortung von manchen Testaufgaben raten, weil ich die Lösung nicht finden konnte.</i>	-.14	193	.89	-.03	.20
<i>Ich habe bei der Beantwortung von manchen Testaufgaben geraten, weil ich nicht motiviert war sie ernsthaft zu beantworten.</i>	-.73	193	.47	-.13	.18

Anhang E

Rotierte Komponentenmatrix des Arbeitsmotivationsfragebogens

Item	Komponente			
	1	2	3	4
<i>Ich war bemüht, die Testaufgaben so gut wie möglich zu bearbeiten.</i>	.10	-.44	.30	.02
<i>Ich hätte mich bei der Bearbeitung des Tests mehr bemüht, wenn für mich Konsequenzen mit dem Testergebnis verbunden wären.</i>	-.09	.71	.10	.11
<i>Als ich die Aufgabenstellung des Tests gesehen habe, habe ich Angst bekommen, im Test schlecht abzuschneiden.</i>	.73	.14	-.33	-.03
<i>Ich war während der Bearbeitung des Tests angespannt.</i>	.83	-.19	-.11	-.01
<i>Ich war während der Bearbeitung des Tests so nervös, dass ich nicht mehr klar denken konnte.</i>	.69	.07	-.29	.10
<i>Ich fühlte mich unter Druck, bei diesem Test gut abzuschneiden.</i>	.85	-.03	-.01	.03
<i>Ich habe bei der Bearbeitung des Tests mein Bestes gegeben.</i>	-.01	-.76	.37	-.17
<i>Ich habe in diesem Test gut abgeschnitten.</i>	-.36	-.24	.78	-.06
<i>Der Test hat zu lange gedauert.</i>	.21	.21	.28	.69
<i>Ich habe im Vergleich zu anderen Studierenden in diesem Test überdurchschnittlich gut abgeschnitten.</i>	-.21	.00	.80	-.06
<i>Es hat mir Freude bereitet, diesen Test zu bearbeiten.</i>	-.17	-.28	.51	-.55
<i>Der Test gibt Aufschluss über das räumliche Vorstellungsvermögen.</i>	.40	-.13	.35	-.47
<i>Ich halte diesen Test zur Beurteilung der Eignung in technischen Berufen für relevant.</i>	.20	.28	.29	-.62

<i>Ich hätte mir bei der Bearbeitung des Tests mehr Mühe geben können.</i>	.01	.77	-.04	.05
<i>Ich empfand die Bearbeitung des Tests als Zeitverschwendung.</i>	.05	.38	-.10	.70
<i>Ich musste bei der Beantwortung von manchen Testaufgaben raten, weil ich die Lösung nicht finden konnte.</i>	.21	.39	-.56	.17
<i>Ich habe bei der Beantwortung von manchen Testaufgaben geraten, weil ich nicht motiviert war sie ernsthaft zu beantworten.</i>	.11	.63	-.30	.35

Extraktionsmethode: Hauptkomponentenanalyse.

Rotationsmethode: Varimax mit Kaiser-Normalisierung.^a

^a. Die Rotation ist in 16 Iterationen konvergiert.