KONSTRUIERE DIE LÖSUNG SELBST

Auswirkungen des Einsatzes von und des Verzichts auf Distraktoren bei Intelligenztestaufgaben zur Erfassung räumlichen Denkens und fluider Intelligenz

DISSERTATION

zur Erlangung des akademischen Grades

eines Doktors der Philosophie

der Fakultät HW

Bereich Empirische Humanwissenschaften

der Universität des Saarlandes

vorgelegt von

Alica Thissen

aus Saarbrücken

Saarbrücken, 2018

Dekan:

Prof. Dr. Cornelius König, Universität des Saarlandes

BerichterstatterInnen:

Prof. Dr. Frank M. Spinath, Universität des Saarlandes

Prof. Dr. Gisa Aschersleben, Universität des Saarlandes

Tag der Disputation: 24.04.2018

Danksagungen

Ich danke Herrn Prof. Dr. Frank M. Spinath für die Möglichkeit, an seinem Lehrstuhl zu promovieren und für seine Unterstützung in dieser Zeit.

Ich danke außerdem Frau Prof. Dr. Gisa Aschersleben dafür, dass sie sich bereit erklärt hat, meine Dissertation als zweite Berichterstatterin zu begutachten.

Des Weiteren danke ich Marco Koch, Nisa Yazici und Thiemo Schmitt für die Hilfe bei der Vorbereitung und der Durchführung meiner Datenerhebungen.

Ich danke meinen lieben Kolleginnen und Kollegen für die schöne Zeit, die wir zusammen erlebt haben, und für die vielen Momente, in denen wir zusammen gelacht und uns ausgetauscht haben, sodass eine angenehme und freundschaftliche Arbeitsatmosphäre entstand.

Ein besonderer Dank gilt Dr. Nicolas Becker, der immer ein offenes Ohr für mich hatte und mir vom ersten Tag an motiviert, freundlich und engagiert zur Seite stand, sodass ich mich nie verloren fühlte und viel lernen konnte. Ich danke ihm sehr für seine Unterstützung und die gute Zeit, die wir zusammen hatten.

Danken möchte ich außerdem meiner Mutter, Claudia Haas, für ihre Liebe, ihre Stärke, ihren Rückhalt und ihre Unterstützung in dieser Zeit und meinem ganzen Leben.

Inhaltsverzeichnis

Abbildungsverzeichnis ..... **VI**

Tabellenverzeichnis ..... **VII**

Überblick über die relevanten Studien ..... **VIII**

1 Einführung **1**

2 Theoretischer Hintergrund **3**

2.1 Intelligenz 3

2.2 Räumliches Denken 7

2.2.1 Definition und Einordnung 7

2.2.2 Relevanz räumlichen Denkens 9

2.2.3 Würfelrotationsaufgaben 10

2.3 Probleme eines Antwortformats mit Distraktoren 11

2.3.1 Ratewahrscheinlichkeit 11

2.3.2 Zwei Lösungsstrategien 12

2.3.3 Mangel an Strategien zur Manipulation der Itemschwierigkeiten 14

3 Studien **16**

3.1 Studie 1 16

3.1.1 Abstract 16

3.1.2 Hinführung 17

3.1.3 Ziele der ersten Studie 19

3.1.4 Materialen und Methodik 20

3.1.5 Ergebnisse 21

3.1.6 Diskussion 24

3.1.7 Ausblick 25

3.1.8 Zusammenfassung 26

3.2 Studie 2 27

3.2.1 Abstract 27

3.2.2 Hinführung 27
3.2.3 Ziele und Hypothesen der zweiten Studie 29
3.2.4 Materialien und Methodik 31
3.2.5 Ergebnisse 35
3.2.6 Diskussion 40
3.2.7 Zusammenfassung 44
3.3 Studie 3 45
3.3.1 Abstract 45
3.3.2 Hinführung 45
3.3.3 Ziele der Studie 3 51
3.3.4 Materialien und Methodik 52
3.3.5 Ergebnisse 57
3.3.6 Diskussion 58
3.3.7 Zusammenfassung 60
4 Generelle Diskussion 61
4.1 Zusammenfassung der Ergebnisse 61
4.2 Einordnung der Ergebnisse und Ausblick 64
4.2.1 Vorteile der Würfelkonstruktionsaufgabe und damit verbundene Forschungsmöglichkeiten 64
4.2.2 Einschränkungen und offene Forschungsfragen 70
4.3 Fazit 75
5 Literaturverzeichnis 77
6 Anhang 89
6.1 Anhang 1: Training aus Studie 3 89

Abbildungsverzeichnis

Abbildung 1: Hierarchisches Intelligenzmodell, modifiziert nach Carroll (1996) 4
Abbildung 2: Aufgabenformat einer gängigen Würfelrotationsaufgabe (Lösung: D) 11
Abbildung 3: Die Würfelkonstruktionsaufgabe 17
Abbildung 4: Würfelkonstruktionsaufgabe mit Lösung 18
Abbildung 5: Würfelkonstruktionsaufgabe mit zwei vorgegebenen Lösungssymbolen 28
Abbildung 6: Würfelkonstruktionsaufgabe im distraktorbasierten Format (korrekte Lösung hervorgehoben) 29
Abbildung 7: Strukturmodell für die distraktorfreie Gruppe 39
Abbildung 8: Strukturmodell für die distraktorgestützte Gruppe 40
Abbildung 9: Klassisches figurales Matrizen-Item 47
Abbildung 10: Rekonstruierte Antwortoptionen des WMT (Formann & Piswanger, 1979) 49
Abbildung 11: Antwortformat des Post-BOMATs 54
Abbildung 12: Antwortausschlusstest-Item 55

Tabellenverzeichnis

Tabelle 1: Fünf räumliche Subfaktoren nach Lohman (1979) und Carroll (1993) 8
Tabelle 2: Ergebnisse der statistischen Analysen aus Studie 1..... 23
Tabelle 3: Itemschwierigkeiten und part-whole-korrigierte Trennschärfen für beide Testversionen aus Studie 2 37
Tabelle 4: Korrelation zwischen Leistung in Würfelaufgabe und Intelligenz in Studie 2 38
Tabelle 5: Ablauf der Testung in Studie 3 56

Überblick über die relevanten Studien

Diese Dissertation umfasst insgesamt drei Studien. Studie I wurde in einer international anerkannten wissenschaftlichen Fachzeitschrift (mit Peer-Review) veröffentlicht, während Studie II bereits zur Veröffentlichung eingereicht wurde und sich noch im Begutachtungsprozess befindet. Das Manuskript zu Studie III befindet sich in Vorbereitung zur zeitnahen Einreichung. Die vollständigen Schriften können in ihrer veröffentlichten Form in den jeweiligen Fachzeitschriften eingesehen werden.

Studie I: **Thissen, A.**, Koch, M., Becker, N. & Spinath, F. M. (2016). Construct Your Own Response. *European Journal of Psychological Assessment*. doi: 10.1027/1015-5759/a000342

Studie II: **Thissen, A.**, Spinath, F. M. & Becker, N. (eingereicht). Manipulate me: The cube construction task allows for a better manipulation of item difficulties than current cube rotation tasks. *European Journal of Psychological Assessment.*

Studie III: Becker, N., **Thissen, A.** & Spinath, F. M. (in Vorbereitung). The knowledge of response elimination strategies that ignore the item stem doesn't impair the construct validity of figural matrices.

Wenn in der folgenden Arbeit der Begriff ‚wir' benutzt wird, beziehe ich mich dabei auf mich und meine Co-Autoren, die mit mir zusammen an den jeweiligen Studien beteiligt waren. Zur besseren Lesbarkeit der Dissertation und um Redundanz zu vermeiden, wurden einzelne Textpassagen der Studien im Vergleich zu den publizierten/eingereichten Manuskripten leicht abgewandelt.

1 Einführung

Intelligenz ist ein bedeutender Prädiktor für verschiedene Kriterien des menschlichen Lebens (Jensen, 1998) und es ist deshalb von Relevanz, sie möglichst valide zu erfassen. Die Validität eines kognitiven Fähigkeitstests kann unter anderem durch die Beschaffenheit des Antwortformats der Aufgabe beeinträchtigt werden (Vigneau, Caissie & Bors, 2006). Mit den Auswirkungen, die die Vorgabe einer Auswahl an falschen Antwortoptionen, sog. Distraktoren, auf die Güte eines Tests haben kann, beschäftigt sich diese Arbeit. Forschung hierzu kommt aus dem Bereich figuraler Matrizenaufgaben (Arendasy & Sommer, 2013; Becker et al., 2016). Neben der Möglichkeit zu raten erlaubt ein solches Antwortformat den Testpersonen die Anwendung unterschiedlicher Lösungsstrategien, von denen anzunehmen ist, dass sie kognitiv unterschiedlich anspruchsvoll sind. Dies führt dazu, dass einzelne Testpersonen sich besser darstellen können, als es ihrer tatsächlichen Fähigkeit entspricht, und die Rangreihe der Leistungen somit verfälscht wird. Infolgedessen verringert sich die konvergente Validität zu anderen Tests, die dasselbe Konstrukt erfassen sollen. Zusätzlich bringt der Einsatz von Distraktoren einen zu geringen Spielraum bei der Schwierigkeitsabstufung der Items mit sich, was sich einerseits ebenfalls negativ auf die Validität auswirkt und andererseits ein Problem bei der Erfassung von Hochbegabung darstellt, da es vor allem an Items sehr hoher Schwierigkeit mangelt (Becker, Preckel, Karbach, Raffel & Spinath, 2015; Gittler, 1990; Liepmann, Beauducel, Brocke & Amthauer, 2007).

Im Bereich figuraler Matrizenaufgaben, die der Erfassung fluider Intelligenz dienen und ein sehr guter Prädiktor für die allgemeine Intelligenz *g* sind (Marshalek, Lohman & Snow, 1983), existiert bereits ein distraktorfreies Aufgabenformat als Alternative zu bisherigen Aufgaben (Becker & Spinath, 2014). Diese Arbeit konzentriert sich hauptsächlich auf die Erfassung räumlichen Denkvermögens, welchem besonders bei der Vorhersage für Berufserfolg in

naturwissenschaftlichen Berufen sowie der Luftfahrt und der Architektur eine wesentliche Bedeutung zukommt (Lohman, 1996; Wai, Lubinski & Benbow, 2009). Hier gibt es ebenfalls einige Aufgaben mit einer hohen *g*-Sättigung (Lohman, 1988, 1996), auf die sich die beschriebenen Probleme mit hoher Wahrscheinlichkeit übertragen lassen. Der Schwerpunkt dieser Arbeit liegt auf der Erprobung eines distraktorfreien Aufgabenformats zur Erfassung räumlichen Denkvermögens.

Nach einer theoretischen Einführung wird in Studie I eine distraktorfreie Würfelrotationsaufgabe, die *Würfelkonstruktionsaufgabe*, vorgestellt und auf relevante psychometrische Eigenschaften hin überprüft. Studie II beschäftigt sich mit der Variation der Itemschwierigkeiten der Würfelkonstruktionsaufgabe und ihrer Validierung. Zudem wird die Würfelkonstruktionsaufgabe im Hinblick auf die Schwierigkeitsverteilung der Items und die Validität direkt mit einer distraktorgestützten Version verglichen. Studie III befasst sich mit der Auswirkung des Einsatzes von Distraktoren im Kontext figuraler Matrizenaufgaben. Hier soll eine in der Fachliteratur zwar erwähnte, aber bisher noch nicht empirisch überprüfte alternative Lösungsstrategie untersucht werden, bei der die Lösung der Aufgabe durch alleinige Betrachtung der Antwortoptionen gefunden werden kann, ohne den Itemstamm zu berücksichtigen (Mittring & Rost, 2008; White & Zammarelli, 1981). Die Ergebnisse der Studien werden abschließend noch einmal zusammengefasst und im Rahmen eines breiteren Kontextes diskutiert.

2 Theoretischer Hintergrund

2.1 Intelligenz

Dank einer mehr als 100-jährigen Forschungstradition ist das Konstrukt Intelligenz das am besten erforschte Konstrukt der Psychologie (Rost, 2013). Es dient der „Beschreibung interindividueller Unterschiede in der geistigen Leistungsfähigkeit bei der Bewältigung neuer geistiger Anforderungen“ (Neubauer & Stern, 2007, S. 22). Auch wenn es keine einheitlich festgelegte Intelligenzdefinition gibt, herrscht in der Wissenschaft doch große Übereinstimmung bei der Definition des Intelligenzbegriffs. Gottfredson (1997) liefert folgende Definition:

> Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings, “catching on”, “making sense” of things, or “figuring out” what to do. (S. 13)

Durch die faktoranalytische Untersuchung kognitiver Leistungen gelten hierarchische Intelligenzmodelle mit einem allgemeinen Intelligenzfaktor *g* an der Spitze und untergeordneten Gruppenfaktoren in der Forschung als Goldstandard (Carroll, 1996; McGrew, 2009; Rost, 2013). Ein prominentes Modell stammt von John B. Carroll (Carroll, 1996) und soll hier exemplarisch erklärt werden. Eine vereinfachte Darstellung des Modells findet sich in Abbildung 1.

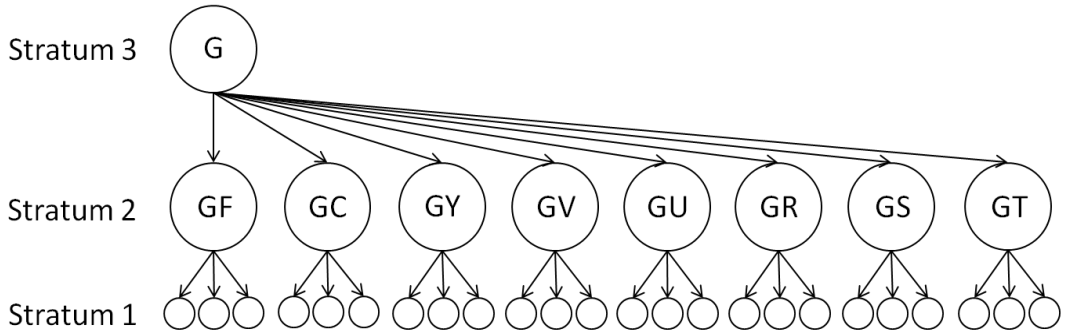


Abbildung 1: Hierarchisches Intelligenzmodell, modifiziert nach Carroll (1996)

Hierarchische Intelligenzmodelle resultieren aus der Beobachtung, dass alle kognitiven Leistungen miteinander positiv korreliert sind. Es korrelieren jedoch nicht alle Testleistungen im selben Maße miteinander, manche korrelieren höher als andere. So lassen sich Gruppen von Testleistungen zusammenstellen, in denen die Testleistungen besonders hoch miteinander korrelieren, und es lässt sich jeweils ein latenter Faktor extrahieren, der den Zusammenhang zwischen den gruppierten Testleistungen stiftet. Diese latenten Faktoren befinden sich in Stratum 1. Es handelt sich hier um sehr spezifische mentale Fähigkeiten. Werden die Testleistungen etwas gröber zusammengefasst, können immer noch Gruppen zusammengestellt werden, in denen die Testleistungen zwar weniger hoch miteinander korrelieren als in den spezifischen Gruppen vorher, in denen die Interkorrelationen der Testleistungen einer Gruppe sich aber weiterhin von anderen Gruppen abgrenzen lassen. Die Interkorrelationen werden wiederum jeweils durch einen übergeordneten Faktor zweiter Ordnung gestiftet (Stratum 2). Hierbei handelt es sich um breiter gefasste mentale Fähigkeiten. Die Gruppenfaktoren aus Stratum 2 bei Carroll sind logisches Schlussfolgern (fluide Intelligenz, *GF*), deklaratives Wissen (kristalline Intelligenz, *GC*), allgemeine Gedächtnisfähigkeit (*GY*), breite visuelle Wahrnehmung (*GV*), breite auditive Wahrnehmung (*GU*), breite Abruffähigkeit (*GR*), breite kognitive Schnelligkeit (*GS*) und Verarbeitungsgeschwindigkeit (*GT*). Auch diese Gruppenfaktoren korrelieren alle noch miteinander und es lässt sich ein robuster latenter Faktor extrahieren, der

die Zusammenhänge zwischen allen Testleistungen stiftet. Dieser kann 50–80 % der Varianz in Leistungsunterschieden zwischen Personen aufklären (Deary, 2001). Er wird als allgemeine Intelligenz *g* bezeichnet und wurde von Pionierforscher Charles Spearman als geistige Energie (Spearman, 1927) und später allgemein als Fähigkeit zum Lernen komplexer Prozesse (Snow, 1989) verstanden. *G* schlägt sich also in jeder kognitiven Leistung nieder. Um eine Aufgabe zu lösen, braucht es demnach immer die allgemeine Intelligenz *g*, eine spezifische mentale Fähigkeit über *g* hinaus (Gruppenfaktoren zweiter Ordnung), eine noch weiter spezifizierte Fähigkeit (Gruppenfaktoren erster Ordnung) und eine aufgabenspezifische Fähigkeit, die nur der jeweiligen spezifischen Aufgabe eigen ist und die mit keiner anderen Testleistung korreliert (Rost, 2013). Je größer das Ausmaß, in dem zum Lösen einer Aufgabe eine aufgabenspezifische Fähigkeit nötig ist, desto geringer sind die Interkorrelationen zwischen verschiedenen Aufgaben, da die Zusammenhänge nur durch *g* und die systematischen Gruppenfaktoren gestiftet werden. Da in allen Gruppenfaktoren *g* enthalten ist, korrelieren sie alle positiv miteinander, sie unterscheiden sich jedoch in der Höhe ihrer *g*-Sättigung.

Aufgaben, die typischerweise eine hohe *g*-Sättigung zeigen, sind Aufgaben, die fluide Intelligenz erfassen, sog. figurale Matrizenaufgaben (Marshalek et al., 1983). Sie gelten deshalb als bester eigenständiger Prädiktor für die allgemeine Intelligenz *g*. Durch ihr figurales Aufgabenformat bieten sie den Vorteil, dass sie sprachfrei und möglichst kulturfair das logische Schlussfolgern erfassen. Ebenso zeigen bestimmte Aufgaben zum räumlichen Denken eine hohe *g*-Sättigung (Lohman, 1988, 1996).

Auf dem Markt existieren eine Reihe von etablierten Intelligenztests, von denen die meisten durch den Einsatz von mehreren heterogenen Aufgaben die Berechnung eines Intelligenzquotienten erlauben, der mit *g* gleichgesetzt werden kann (Kubinger, 2009; Liepmann et al., 2007; Petermann & Petermann, 2007; Wechsler, 2012; Weiß & Osterland, 2012). Manche Tests erfassen hingegen auch nur eine spezielle mentale Fähigkeit (Gittler, 1990; Stumpf & Fay, 1983).

Die Diagnostik von Intelligenz spielt eine entscheidende Rolle bei der Vorhersage von schulischem und beruflichem Erfolg. Die Korrelationen zwischen Intelligenz und Schulerfolg sind mit die höchsten, die es in der empirischen Psychologie gibt und darüber hinaus existieren nur wenige inkrementelle Prädiktoren (Jensen, 1998; Rost, 2013). Ebenso ermöglicht Intelligenz überdurchschnittliche Vorhersagen für Berufserfolg, wenn auch weniger genaue als für Schulerfolg, da in selegierten Berufsgruppen wie Akademikern die Varianz der Intelligenz eingeschränkt ist und hier andere Faktoren wie Motivation, Kreativität und Persönlichkeit mehr zum Tragen kommen (Neubauer & Stern, 2007). Dennoch ist sie auch hier der beste zur Verfügung stehende Prädiktor. Intelligenz ermöglicht weiterhin Vorhersagen über Delinquenz, Unfälle, Krankheit, Alkoholismus, ungewollte Schwangerschaften, Lebenserwartung und Lebenszufriedenheit (Ceci, 1996; Neubauer & Stern, 2007; Süß, 2001). Des Weiteren hat die Diagnostik von Hochbegabung bzw. kognitiver Behinderung besondere Relevanz, da diese Personengruppen nicht selten von einer besonderen Förderung und Beratung profitieren oder auf diese angewiesen sind, um ihr Potential zu entfalten bzw. um individuelle Verhaltens- und Leistungsprobleme sowie soziale Konflikte zu vermeiden (Heller, 1987; Heller & Perleth, 2007).

Die Forschung zu Intelligenz befasste sich seit ihren Anfängen vor über 100 Jahren nicht nur mit der Definition von Intelligenz, sondern auch mit ihrer Diagnostik (Stemmler, Hagemann, Amelang & Spinath, 2016). Die Überlegung, wie ein Test beschaffen sein muss, um Intelligenz bestmöglich zu erfassen, war von Anfang an zentral. Dabei kann die Beschaffenheit der Aufgabe selbst (d. h. der Items) Auswirkungen auf die Güte der Messung haben, aber auch die Beschaffenheit des Antwortformats. Mit den Auswirkungen, die die Beschaffenheit des Antwortformats einer Aufgabe auf die Güte eines kognitiven Fähigkeitstests haben kann, beschäftigt sich diese Arbeit. Die Folgen der Vorgabe einer Auswahl an falschen Antwortoptionen, sog. Distraktoren, stehen dabei im Fokus. Hierzu gibt es bereits einige Studien im Bereich figuraler Matrizenaufgaben (Arendasy & Sommer, 2013; Bethell-Fox, Lohman & Snow, 1984; Vigneau et al., 2006), die, wie oben

bereits erwähnt, eine hohe *g*-Sättigung aufweisen. Ebenfalls eine hohe *g*-Sättigung zeigen viele Aufgaben zum räumlichen Denken, bei denen die Auswirkungen des Einsatzes von Distraktoren noch nicht näher empirisch untersucht wurden. Der Schwerpunkt dieser Arbeit liegt deshalb auf der Erfassung räumlichen Denkens, auf das im folgenden Abschnitt genauer eingegangen werden soll.

2.2 Räumliches Denken

2.2.1 Definition und Einordnung

Räumliches Denken ist als Faktor nahezu in jedem gängigen Intelligenzmodell vertreten (Carroll, 1996; Marshalek et al., 1983; McGrew, 2009; Vernon, 1961). Im prominenten Modell von Carroll (1996) findet sich räumliches Denken beispielsweise als Gruppenfaktor unter *g* wieder. Im Radex-Modell von Marshalek et al. (1983) ist räumliches Denken einer der zentralen Indikatoren von *g*. Infolge dessen gibt es Aufgaben zum räumlichen Denken als Teil von Intelligenztestbatterien (Jäger, Süß & Beauducel, 1997; Liepmann et al., 2007), es existieren aber auch eine ganze Reihe von eigenständigen Tests, die einzelne Aspekte des räumlichen Denkens erfassen (Eliot & Macfarlane Smith, 1983; Gittler, 1990; Stumpf & Fay, 1983).

Räumliches Denken wird in der Fachliteratur nicht einheitlich definiert. Es existieren mehrere Definitionen, die inhaltlich zwar Unterschiede, aber auch große Überschneidungen aufweisen. Lohman (1979) definiert räumliches Denken als „[...] 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“ (S. 126). Räumliches Denken als mentale Fähigkeit wurde bereits 1880 von Galton (1883) thematisiert und ab den 1920er Jahren erstmals als eigenständiger Faktor identifiziert, der von der allgemeinen Intelligenz *g* zu trennen ist (Kelley, 1928; Thurstone, 1938). In den darauffolgenden Jahrzehnten konzentrierten sich die Bemühungen darauf, zu definieren, was räumliches Denken

genau ausmacht. Aufgrund verschiedener faktoranalytischer Techniken und einer Fülle von entwickelten Tests zum räumlichen Denken entstand eine uneinheitliche Taxonomie. Es wurde ein ganzes Bündel von trennbaren, aber hoch korrelierten Subfaktoren extrahiert, die teilweise widersprüchlich benannt wurden und über deren Anzahl keine Einigkeit herrschte (Cattell, 1971; Guilford & Lacy, 1947; Thurstone, 1944).

Die Ergebnisse zweier Meta-Analysen werden häufig zitiert und liefern eine Re-Analyse einer großen Anzahl von Datensets und somit einen guten Überblick. Lohman (1979) fand in seiner Meta-Analyse zehn Subfaktoren räumlichen Denkens, Carroll (1993) fünf, wobei diese fünf sich auch bei Lohman (1979) fanden. Diese fünf Faktoren, die in beiden Meta-Analysen auftauchten, sind in Tabelle 1 aufgeführt.

Tabelle 1: Fünf räumliche Subfaktoren nach Lohman (1979) und Carroll (1993)

Faktor	**Beschreibung der geforderten Fähigkeit**
Spatial Relations	Simple mentale Rotationen von Objekten
Spatial Visualization	Komplexe mentale Rotationen / Bewegung und Verschiebung einzelner Teile eines Objektes
Closure-Speed	Erkennen eines unvollständigen Bildes durch Abgleich mit einer Langzeitgedächtnis-Repräsentation
Flexibility of Closure	Erkennen eines versteckten, eingebetteten Musters und Herauslösen aus einem größeren, komplexen Muster
Perceptual Speed	Schneller Vergleich visueller Stimuli / schnelles Finden eines Items in einer Gruppe von Items

Anhand der einzelnen Faktorbeschreibungen wird deutlich, dass die Faktoren hinsichtlich ihrer benötigten mentalen Prozesse Überschneidungen aufweisen und nicht immer eindeutig voneinander zu trennen sind. Einige sind typischerweise mehr, andere weniger *g*-gesättigt, möglicherweise in Abhängigkeit von der Komplexität der ablaufenden mentalen Prozesse (Lohman, 1988; Marshalek et al.,

1983). Spatial-Visualization-Aufgaben zeigen typischerweise eine besonders hohe *g*-Sättigung (Lohman, 1988, 1996). Neuere Untersuchungen extrahieren einen dynamischen Faktor, der die Fähigkeit, bewegte Objekte zu erfassen, beschreibt (Pellegrino & Hunt, 1991). Die Debatte über die Anzahl, Definition und Trennbarkeit zu extrahierender Subfaktoren und die von vielen Forschern geäußerte Frustration darüber hält bis heute an (Harle & Towns, 2011).

2.2.2 Relevanz räumlichen Denkens

Wie oben bereits beschrieben, ist räumliches Denken ein Faktor in jedem Intelligenzmodell und deshalb als Teil einer Testbatterie zur Erfassung der allgemeinen Intelligenz geeignet. Besonders komplexe Aufgaben können ein guter Indikator für *g* sein (Marshalek et al., 1983). Zudem hat räumliches Denken besondere Relevanz in sog. MINT-Fächern (Mathematik, Informatik, Naturwissenschaften, Technik; engl. STEM), aber auch in Bereichen wie der Luftfahrt, Architektur, Zahnmedizin, Chirurgie, des Designs etc. (Lohman, 1996). Personen mit einem hohen räumlichen Vorstellungsvermögen sind häufiger interessiert an solchen Fächern und zeigen dort mehr Persistenz und eine bessere Leistung als in diesem Bereich weniger fähige Kandidaten (Kell & Lubinski, 2013; Lohman, 1988; Smith, 1964; Wai et al., 2009; Webb, Lubinski & Benbow, 2007). Um für diese Berufe geeignete Kandidaten zu identifizieren, sind Tests zum räumlichen Vorstellungsvermögen also von großer Relevanz.

Abseits von speziellen Berufsgruppen wird räumlichem Denken als Intelligenzfacette oft zu Unrecht eine untergeordnete Rolle zugeschrieben (Lohman, 1996). Forschung aus den USA zeigt, dass räumliches Vorstellungsvermögen dort als Faktor in Programmen zur allgemeinen Talentsuche meist vernachlässigt wird (Shea, Lubinski & Benbow, 2001). Es konnte gezeigt werden, dass durch die Konzentration auf verbale und mathematische Fähigkeiten in Programmen zur Talentsuche mehr als die Hälfte der besten 1 % im Bereich räumliches Vorstellungsvermögen nicht identifiziert werden (Lohman & Korb, 2006;

Shea et al., 2001). Es werden somit eine ganze Reihe von überdurchschnittlich begabten Personen nicht erkannt und infolge dessen auch nicht optimal gefördert, obwohl sie in MINT-Bereichen unter Umständen zu besonderer Leistung fähig wären. Zusätzlich könnte es sein, dass generell in Unterricht und Ausbildung zu wenig Wert auf die Schulung von räumlichem Vorstellungsvermögen gelegt wird (Harle & Towns, 2011; Wai et al., 2009; Webb et al., 2007) und SchülerInnen, deren Begabung vor allem im räumlichen Denken liegt, ihr Potential im normalen Schulalltag nicht adäquat zum Ausdruck bringen können bzw. SchülerInnen, denen es umgekehrt an Erfahrung mit räumlichem Material mangelt, in MINT-Fächern Schwierigkeiten entwickeln.

2.2.3 Würfelrotationsaufgaben

Ein gängiges Format zur Erfassung räumlichen Denkens sind Würfelrotationsaufgaben (Gittler, 1990; Jäger et al., 1997; Liepmann et al., 2007). Würfelrotationsaufgaben sind in der Regel folgendermaßen aufgebaut: Ein Referenzwürfel wird zusammen mit einer Reihe von anderen Würfeln, die die Antwortoptionen darstellen, dargeboten. Die Aufgabe besteht darin, den Referenzwürfel mit den Antwortoptionen zu vergleichen und zu entscheiden, welche Antwortoption eine rotierte Version des Referenzwürfels sein könnte. Dabei gibt es nur eine richtige Antwort, während die anderen falschen Antworten als sog. Distraktoren fungieren. Manchmal gibt es außerdem die Antwortoptionen ‚Kein Würfel richtig' und ‚Ich weiß die Lösung nicht', um einem unerwünschten Lösungsverhalten entgegenzuwirken. Abbildung 2 zeigt eine typische Würfelrotationsaufgabe. Auch bei anderen Aufgaben zum räumlichen Denken kommen überwiegend Distraktoren zum Einsatz (Eliot & Macfarlane Smith, 1983).

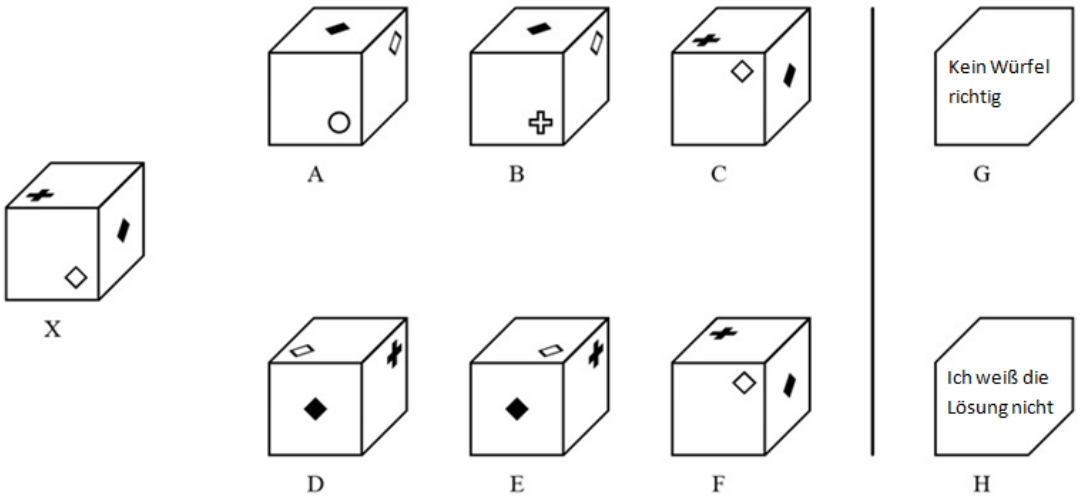


Abbildung 2: Aufgabenformat einer gängigen Würfelrotationsaufgabe (Lösung: D)

2.3 Probleme eines Antwortformats mit Distraktoren

Das oben beschriebene Aufgabenformat, bei dem die richtige Lösung aus einer Reihe von Antwortalternativen ausgewählt werden muss, bringt verschiedene Probleme mit sich, die im Folgenden genauer erläutert werden sollen. Die beiden entscheidenden Folgen dieser Probleme sind eine eingeschränkte Konstruktvalidität des Tests und die fehlende Möglichkeit, Hochbegabung durch Items ausreichender Schwierigkeit adäquat zu erfassen.

2.3.1 Ratewahrscheinlichkeit

Zunächst einmal kann die richtige Lösung mit einer Ratewahrscheinlichkeit von $1/k$ (k = Anzahl an Antwortoptionen) erraten werden. Bei einer ausreichend großen Anzahl an Distraktoren ist die Ratewahrscheinlichkeit zwar vergleichsweise gering, trotzdem ist Raten nicht erwünscht und kann nicht komplett ausgeschlossen werden. Durch die Möglichkeit, die Lösung zu erraten, verringert sich die Schwierigkeit der Aufgabe. Zusätzlich wird die Konstruktvalidität beeinträchtigt, da

Raten nicht die Fähigkeit widerspiegelt, die mit der Aufgabe eigentlich erfasst werden soll. Die Antwortoption ‚Ich weiß die Lösung nicht' (Gittler, 1990) kann hier nicht als hinreichende Strategie betrachtet werden, um Raten auszuschließen.

2.3.2 Zwei Lösungsstrategien

Gravierender ist der Umstand, dass der Einsatz von Distraktoren den Testpersonen ermöglicht, zwei verschiedene Lösungsstrategien anzuwenden (Arendasy & Sommer, 2013; Bethell-Fox et al., 1984; Gittler, 1990; Vigneau et al., 2006). Ergebnisse hierzu stammen aus der Forschung zu figuralen Matrizenaufgaben. Das sogenannte *Constructive Matching* ist durch den Versuch definiert, aktiv eine potentielle Lösung zu generieren und diese anschließend mit den Antwortoptionen zu vergleichen, um die richtige Lösung auszuwählen. Im Kontext von Würfelrotationsaufgaben würde dies bedeuten, ein mentales Modell des Referenzwürfels zu generieren und den Würfel anschließend mental zu drehen, um ihn mit dem jeweiligen Antwortwürfel zu vergleichen und zu entscheiden, ob es sich um denselben Würfel handeln könnte. Der gesamte Würfel und damit die Beziehungen aller sichtbaren Würfelseiten zueinander müssen in diesem Fall mental repräsentiert und die korrekte Antwort muss im Sinne eines *Top-down*-Prozesses verifiziert werden. Constructive Matching stellt die Lösungsstrategie dar, die angewendet werden soll. Im Gegensatz dazu steht eine zweite mögliche Lösungsstrategie, die sogenannte *Response-Elimination*. Hier werden einzelne Elemente der beiden Würfel verglichen, um falsche Antworten ausschließen zu können, und zwar ohne dass der Referenzwürfel oder die Antwortwürfel im Ganzen mental repräsentiert werden müssen. Durch den paarweisen Vergleich zweier Symbole auf dem Referenzwürfel und der Antwortoption können falsche Antworten systematisch ausgeschlossen werden. Auch wenn der Referenzwürfel mental repräsentiert und rotiert wird, genügt eine einzige Diskrepanz, um eine Antwortoption als falsch zu identifizieren. Es erfolgt keine Prüfung der vollständigen Widerspruchsfreiheit, ein Prozess, für den mehr kognitive Schritte

nötig wären. Im Sinne eines *Bottom-up*-Prozesses kann die Lösung hier durch Falsifikation gefunden werden, indem falsche Antworten eliminiert werden, bis nur noch eine Antwortoption übrig bleibt oder die Testperson aus einer verminderten Anzahl von Antwortoptionen raten kann. Es ist anzunehmen, dass Response-Elimination kognitiv weniger aufwändig ist als Constructive Matching und die Ansprüche an räumliches Denken hier deutlich geringer sind.

Aus der Forschung zu figuralen Matrizenaufgaben ist bekannt, dass die Auswahl einer Lösungsstrategie durch individuelle Unterschiede in der allgemeinen Intelligenz und Arbeitsgedächtniskapazität beeinflusst wird (Jarosz & Wiley, 2012; Vigneau et al., 2006). Testpersonen mit einer niedrigeren allgemeinen Intelligenz und Arbeitsgedächtniskapazität zeigen eine größere Tendenz Response-Elimination anzuwenden, während Testpersonen mit einer hohen allgemeinen Intelligenz und Arbeitsgedächtniskapazität wahrscheinlicher Constructive Matching anwenden. Response-Elimination wird deshalb als eine Ausweichstrategie angesehen, auf die zurückgegriffen wird, wenn die erforderliche Verarbeitung die Kapazitätsgrenzen der Testperson überschreitet (Arendasy & Sommer, 2013; Bethell-Fox et al., 1984).

Weniger fähige Kandidaten im Bereich räumlichen Denkens können die Aufgabe also durch eine kognitiv weniger aufwändige, nicht erwünschte Strategie lösen und den eigentlichen Lösungsprozess umgehen. Sie stellen sich dann in ihrer Leistung dar wie fähigere Kandidaten oder sogar besser. Dies wirkt sich negativ auf die Konstruktvalidität des Tests aus, denn es kann nicht mehr sichergestellt werden, dass er bei allen Testpersonen misst, was er zu messen beabsichtigt (Arendasy & Sommer, 2013; Becker et al., 2016). Die Korrelation zu einem allgemeinen Intelligenztest (konvergente Validität) wird dadurch vermindert, dass die Rangreihen der Leistungen im Vergleich nicht übereinstimmen, da Kandidaten im einen Fall eine bessere Leistung ‚vortäuschen' können, als es ihrer tatsächlichen Fähigkeit entspricht. Des Weiteren wird durch die Möglichkeit, Response-Elimination anzuwenden, automatisch auch die Erstellung von Items hoher Schwierigkeit erschwert. Es findet sich also eine Varianzeinschränkung in der Verteilung der Testwerte, da im niedrigen Fähigkeitsbereich zu wenige Personen zu

finden sind und im hohen zu viele. Die differenzierte Erfassung von hoher Begabung kann so nicht erfolgen. Es ist fraglich, ob die Antwortoption ‚Kein Würfel richtig', wie sie bei Gittler (1990) zu finden ist, einem Falsifikationsvorgehen entgegenwirken kann. Selbst wenn dem so wäre, könnte am Ende zwischen einer übrig gebliebenen Antwortoption und der Option ‚Kein Würfel richtig' mit einer ungünstigen Ratewahrscheinlichkeit von 50 % geraten werden.

2.3.3 Mangel an Strategien zur Manipulation der Itemschwierigkeiten

Ein weiteres Problem, das der Einsatz von Distraktoren mit sich bringt, ist eine zu geringe Manipulierbarkeit der Itemschwierigkeiten. Die Itemschwierigkeiten von geläufigen Würfelrotationsaufgaben ohne Zeitlimit sind überwiegend moderat und außerdem homogen (Gittler, 1990). Abgesehen von der möglichen Anwendung von Response-Elimination-Strategien, kann dies auch an einem Mangel an Strategien zur Variation der Itemschwierigkeiten liegen. Im Vergleich zu anderen figuralen Aufgabenformen wie figuralen Matrizenaufgaben, bei denen die Items durch die Anwendung verschiedener Regeln in ihrer Schwierigkeit variiert werden können, sind die Möglichkeiten bei Würfelaufgaben begrenzter. Eine systematische Variation bestimmter Symbole auf dem Referenzwürfel, um die Schwierigkeit zu beeinflussen, findet sich bis jetzt nicht. Bei Gittler (1990) soll die Itemschwierigkeit manipuliert werden, indem der Referenzwürfel unterschiedlich oft gekippt bzw. gedreht werden muss, um zur Lösung zu gelangen. Eine systematische Abstufung der Itemschwierigkeiten oder ein ausreichend heterogener Itempool wird dadurch jedoch nicht erzeugt. Genauso sind die Möglichkeiten der Distraktorgestaltung begrenzt. Wenn ein Referenzwürfel erstellt wird, wird die richtige Antwort erzeugt, indem der Würfel auf eine bestimmte Art rotiert wird und evtl. Seiten zu sehen sind, die vorher nicht zu sehen waren. Distraktoren werden dann erzeugt, indem der Referenzwürfel rotiert wird und die Seiten, die darauf sichtbar waren, verfälscht werden (Gittler, 1990). Die Erstellung von Items variierender

Schwierigkeit, also von leichter, mittlerer und hoher Schwierigkeit, wird somit erschwert. Die Differenzierung in Gruppen von sehr fähigen und sehr wenig fähigen Kandidaten ist damit nur begrenzt möglich und der Test ist nicht für eine solche Diagnostik geeignet (Mittring & Rost, 2008). Des Weiteren vermindert sich auch die konvergente Validität, d. h. die Korrelation zu anderen ähnlichen Fähigkeitstests, die im unteren und oberen Fähigkeitsspektrum genauer differenzieren können. Der einzige Ansatz, die Itemschwierigkeiten tatsächlich zu variieren, liegt derzeit darin, ein Zeitlimit für die gesamte Testung einzuführen (Jäger et al., 1997; Liepmann et al., 2007). Das bedeutet jedoch, dass die Items am Ende der Testung nur deshalb schwieriger werden, weil einige Testpersonen die Items aus Zeitgründen nicht mehr in Angriff nehmen können, nicht aber, weil das Item selbst schwieriger wird. Dies ist jedoch problematisch, da die Ergebnisse einer zeitlimitierten Testung mit einer mentalen Geschwindigkeitskomponente konfundiert sind (Wilhelm & Schulze, 2002).

3 Studien

3.1 Studie 1

3.1.1 Abstract

Räumliches Denkvermögen wird häufig mit Würfelrotationsaufgaben gemessen, bei denen die Testperson die richtige Antwort aus verschiedenen Antwortoptionen auswählen muss. Ein solches Antwortformat bringt verschiedene Probleme mit sich. In der folgenden Studie wird die Würfelkonstruktionsaufgabe vorgestellt. Die Würfelkonstruktionsaufgabe stellt ein neuartiges, alternatives Format zur Erfassung räumlichen Denkvermögens dar. Anstatt die richtige Antwort aus verschiedenen Antwortoptionen auszuwählen, muss die Testperson hier die Lösung in einer computerisierten Testumgebung selbst konstruieren. Das Format hat verschiedene Vorteile: Es ist nicht länger möglich, zu raten oder die Antwortwürfel mit dem Referenzwürfel zu vergleichen, wodurch anzunehmen ist, dass die Anforderungen an räumliches Denken steigen. Zudem ist es möglich, Items sehr hoher Schwierigkeit zu erstellen, die in geläufigen Würfelrotationsaufgaben fehlen, die allerdings für die Erfassung von Hochbegabung unerlässlich sind. In der folgenden Studie wurden 28 Items entwickelt und einer Stichprobe von 128 Studierenden vorgelegt. Die Ergebnisse zeigen, dass die Items eine sehr hohe statistische Schwierigkeit aufweisen [*M(pi)* = .20, *SD* = .07] und die Skala eine sehr hohe interne Konsistenz (α = .93) besitzt. Die Ergebnisse einer exploratorischen Faktoranalyse zeigen, dass sowohl eine einfaktorielle Lösung (räumliches Denken) als auch eine zweifaktorielle Lösung (Spatial Relations und Spatial Visualization) plausibel ist und sinnvoll interpretiert werden kann. Perspektiven zukünftiger Forschung und praktische Anwendungsmöglichkeiten der Würfelkonstruktionsaufgabe werden diskutiert.

3.1.2 Hinführung

Es stellt sich die Frage, wie den unter Punkt 2.3 beschriebenen Problemen begegnet werden kann. Um die negativen Auswirkungen des Einsatzes von Distraktoren zu umgehen, haben wir eine Würfelrotationsaufgabe konstruiert, die ohne Distraktoren auskommt. Diese Aufgabe haben wir *Würfelkonstruktionsaufgabe* genannt. Eine solche distraktorfreie Würfelrotationsaufgabe existiert in der Fachliteratur nach unserem Wissen bis jetzt noch nicht. Das Aufgabenformat ist in Abbildung 3 zu sehen.

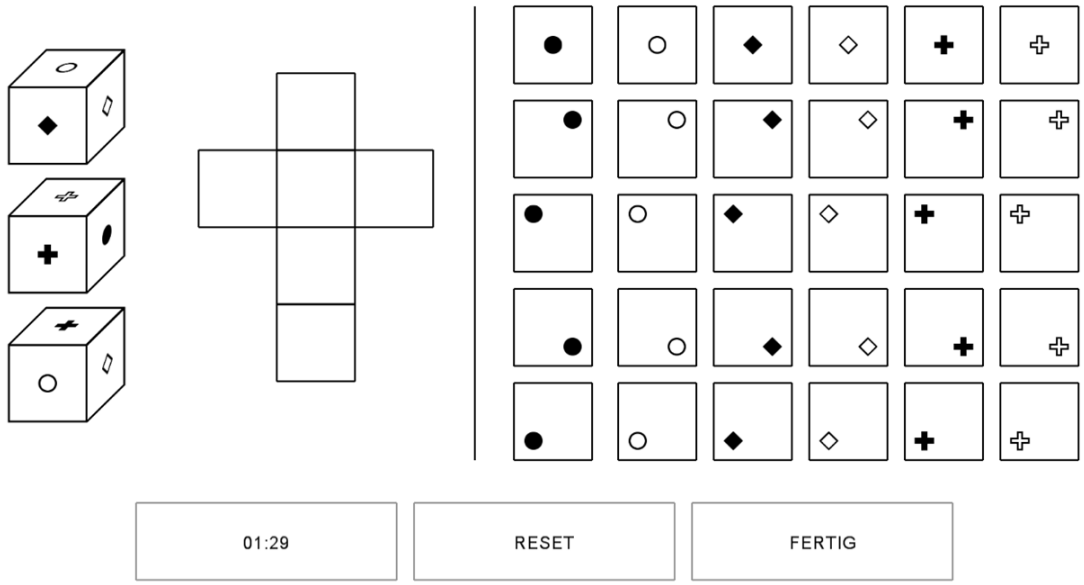


Abbildung 3: Die Würfelkonstruktionsaufgabe

Das Item ist auf der linken Seite abgebildet. Es besteht aus drei Würfeln. Hierbei handelt es sich um ein und denselben Würfel, der aus drei unterschiedlichen Perspektiven gezeigt wird. Die Items sind alle nach derselben Regel aufgebaut: Ein Würfelpaar überschneidet sich in zwei der gezeigten Seiten, ein Würfelpaar überschneidet sich in einer gezeigten Seite und beim letzten Würfelpaar gibt es keine Überschneidungen. Diese Darstellungsform beinhaltet die minimale Anzahl an Überschneidungen, die nötig ist, um jede Seite des Würfels und

somit jedes vorhandene Symbol auf dem Würfel sichtbar zu machen. Die Aufgabe besteht darin, den Würfel mental zu entfalten und anzugeben, wie der Würfel im aufgeklappten Zustand aussehen würde. Dazu befindet sich rechts neben dem Item eine aufgeklappte leere Würfelfläche, die mit den entsprechenden Symbolen gefüllt werden soll. Dazu stehen der Testperson auf der rechten Seite alle Symbole, die auf den Items verwendet wurden, zur Verfügung. Die Testperson kann aus diesem Symbol-Pool die Symbole auswählen, die auf dem jeweiligen Item vorkommen und sie per Klick an die gewünschte Stelle auf der leeren Würfelfläche setzen. Die Lösung der Aufgabe ist in Abbildung 4 zu sehen.

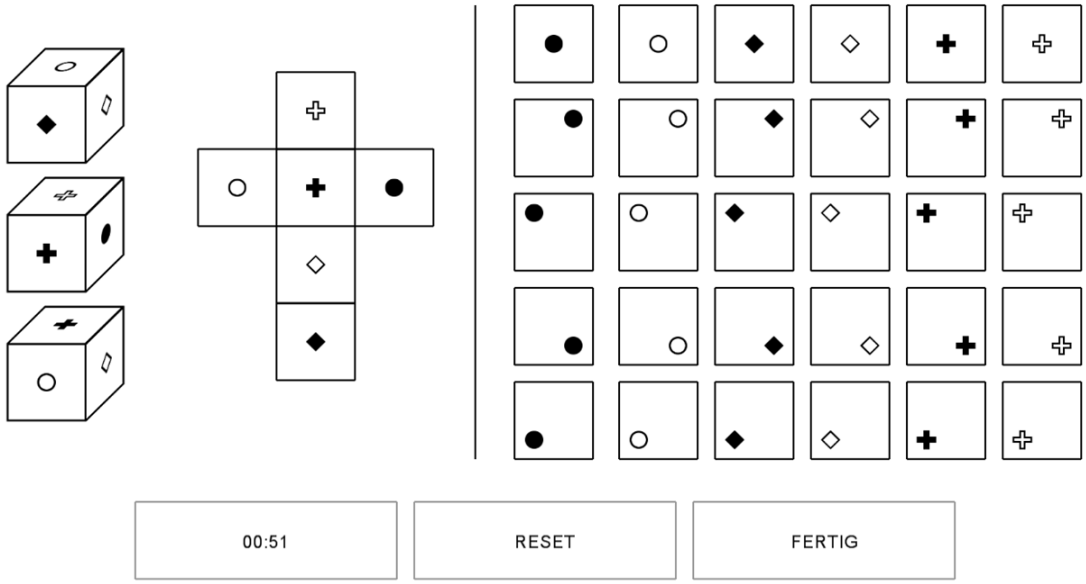


Abbildung 4: Würfelkonstruktionsaufgabe mit Lösung

Als Anmerkung soll hier noch erwähnt werden, dass es möglich ist, den Würfel an unterschiedlichen Stellen aufzuklappen und es somit verschiedene Varianten der Lösung gibt, die bei der Auswertung der Aufgabe berücksichtigt werden. Sie alle stellen Rotationsmöglichkeiten eines richtigen Würfels dar, den es nur in einer Form als eindeutig korrekte Lösung gibt. Die Lösung muss selbst konstruiert werden und der Lösungsprozess kann nicht umgangen werden, da nicht die Möglichkeit besteht, aus einer Reihe von Distraktoren auszuwählen und die

Response-Elimination-Strategie anzuwenden. Die Würfelkonstruktionsaufgabe verspricht daher, den Faktor räumliches Vorstellungsvermögen in einer reineren Form zu messen als Aufgabenformate mit Distraktoren, und es ist zu erwarten, dass sich Items hoher Schwierigkeit erstellen lassen. Es kann davon ausgegangen werden, dass die Ratewahrscheinlichkeit dank einer großen Anzahl an Symbolen und Platzierungsmöglichkeiten sehr gering ist.

Wie bereits erwähnt, werden in der Fachliteratur wiederholt die beiden Faktoren Spatial Relations und Spatial Visualization beschrieben (Carroll, 1993; Lohman, 1979). Es stellt sich die Frage nach der Faktorstruktur der Würfelkonstruktionsaufgabe. Die Aufgabe könnte eindimensional den Faktor räumliches Denken erfassen oder zwei trennbare, aber verbundene Faktoren. Der Faktor Spatial Relations könnte durch die Aufgabe gemessen werden, da eine mentale Rotation durchgeführt werden soll. Ebenso könnte die Aufgabe Spatial Visualization erfassen, da drei Würfel zueinander in Beziehung gesetzt werden müssen und die Würfeloberfläche auseinander gefaltet werden muss, was die Aufgabe im Vergleich zu einer simplen mentalen Rotationsaufgabe komplexer macht. Sowohl eine eindimensionale als auch eine zweidimensionale Lösung mit zwei interkorrelierten Faktoren wären plausibel. Da die Aufgabe räumliches Vorstellungsvermögen in möglichst reiner Form messen soll, sollte sie nicht zwei unkorrelierte Faktoren erfassen.

3.1.3 Ziele der ersten Studie

Ziel der ersten Studie war es, die Würfelkonstruktionsaufgabe auf ihre psychometrischen Eigenschaften zu prüfen. Es sollte gezeigt werden, dass es möglich ist, ohne Distraktoren eine reliable Skala zu entwickeln, deren Items eine hohe Schwierigkeit aufweisen und somit auch im hohen Fähigkeitsbereich differenzieren können. Des Weiteren sollte die Faktorstruktur der Aufgabe geklärt werden und Geschlechtsunterschiede sollten untersucht werden, da bei

räumlichen Aufgaben männliche Testpersonen häufig besser abschneiden (Halpern & Collaer, 2005; Linn & Petersen, 1985).

3.1.4 Materialen und Methodik

3.1.4.1 Stichprobe

Die Stichprobe umfasste 130 Studierende eines Psychologie-Einführungskurses (*M(Alter)* = 21.54; *SD(Alter)* = 3.09; 72.30 % weiblich) der Universität des Saarlandes. Für die Teilnahme am Experiment wurden zusätzliche Credit-Points vergeben. Die Testung fand in einem laboratorischen Setting am Computer in Kleingruppen von höchstens 20 Studierenden statt. Vor der Testung füllten die Studierenden einen demografischen Fragebogen aus.

3.1.4.2 Erstellung der Würfelkonstruktionsaufgabe

Der Test bestand aus 28 Items. Während des Erstellungsprozesses wurden die Symbole zufällig auf die Referenzwürfel verteilt. Die durchschnittlich benötigte Bearbeitungszeit wurde in einer Pilotstudie ermittelt. Sechs Testpersonen bearbeiteten eine vorläufige Version des Tests mit 40 Items mit einer Zeitbeschränkung von 4 Min. pro Item. Die Testpersonen brauchten im Durchschnitt 1.43 Min. (*SD* = .37 Min.) für die Bearbeitung eines Items. Um die Ökonomie des Tests zu erhöhen, wurde die Anzahl der Items auf 28 Items reduziert. Eine Analyse der Itemschwierigkeiten und -trennschärfen ergab keine bedeutsamen Unterschiede, weshalb die 28 Items zufällig aus den 40 Items ausgesucht wurden und auch deren Reihenfolge zufällig festgelegt wurde. Für jedes Item wurde die Bearbeitungszeit auf 2.17 Min. festgelegt. Dies entspricht der in der Pilotstudie ermittelten durchschnittlichen Bearbeitungszeit plus zwei Standardabweichungen und stellt eine liberale Zeitbegrenzung dar. Zwischen den Items gab es jeweils eine Pause von 5 Sek., die auch übersprungen werden konnte. Insgesamt dauerte die Testung maximal 62.91 Min.

3.1.4.3 Statistische Methoden

Um die Reliabilität der Aufgabe festzustellen, wurde Cronbachs Alpha (α) berechnet. Des Weiteren wurde die Trennschärfe der Items berechnet, indem jeweils die Korrelation eines Items mit dem Rest der Skala (part-whole-Korrektur) gebildet wurde. Die Schwierigkeit eines Items wird durch seine Lösungswahrscheinlichkeit *p* ausgedrückt, also durch den Anteil der Testpersonen, die das Item lösen konnten. Eine niedrige Lösungswahrscheinlichkeit beschreibt also eine hohe Itemschwierigkeit. Die Dimensionalität der Aufgabe wurde durch eine exploratorische Faktoranalyse (EFA) mittels Geomin Rotation der polychorischen Korrelationsmatrix in Mplus (Version 6; Muthén & Muthén, 2007) ermittelt. Um die Schiefe der Itemschwierigkeitsverteilung zu normalisieren, wurde der kategorische Algorithmus WLSMV (Beauducel & Herzberg, 2006) benutzt. Ebenfalls mit Mplus wurde die Passung der Daten auf die aus der explorativen Faktoranalyse stammenden Modelle konfirmatorisch geprüft. Um die Geschlechtsunterschiede zu ermitteln, wurde ein t-Test für unabhängige Stichproben durchgeführt.

3.1.5 Ergebnisse

3.1.5.1 Reliabilität

Mit Cronbachs α = .93 und McDonalds ω = .97 kann die Skala als sehr reliabel angesehen werden. Die durchschnittliche Itemtrennschärfe lag bei .54. Nur fünf der 28 Items zeigten einen Wert unter .40 (siehe Tabelle 2).

3.1.5.2 Schwierigkeiten

Eine Analyse der Anzahl korrekt gelöster Items ergab eine durchschnittliche Lösungshäufigkeit von 5.59 Items (*SD* = 6.48). Die minimale Anzahl gelöster Items betrug 0, die maximale 27. Die Verteilung der Summenwerte zeigte eine Schiefe von 1.27 und einen Exzess von .7. Die Lösungswahrscheinlichkeiten der Items

(Anzahl der Personen, die das Item lösen können/*n*) variierten zwischen .08 und .34, die mittlere Lösungswahrscheinlichkeit betrug .20 (*SD* = .07). Die Verteilung der Lösungswahrscheinlichkeiten zeigte eine Schiefe von .28 und einen Exzess von .14 (siehe Tabelle 2).

3.1.5.3 Faktorielle Validität

Die EFA ergab die folgenden Eigenwerte für die ersten fünf Faktoren: 15.48, 2.48, 2.01, 1.22, 1.14. Die Ein-Faktor-Lösung [$\chi^2(350) = 413.58$; $p = .01$; CFI = .98; RMSEA = .04] sowie die Zwei-Faktor-Lösung [$\chi^2(323) = 351.53$; $p = .13$; CFI = .99; RMSEA = .03] zeigten einen guten Modellfit. Zwischen den beiden Faktoren der Zwei-Faktor-Lösung bestand eine Korrelation von $r = .52$. Der χ^2-Test auf Unterschiedlichkeit zeigte einen signifikant besseren Modellfit des zweifaktoriellen Modells, $\chi^2(27) = 62.05$; $p < .01$. Die Faktorladungen der einfaktoriellen und der zweifaktoriellen Lösung sind in Tabelle 2 aufgelistet. In der einfaktoriellen Lösung ergaben sich substantielle Ladungen für jedes Item auf dem ersten Faktor. Genauso konnten substantielle Ladungen für alle Items auf beiden Faktoren der Zwei-Faktor-Lösung gefunden werden.

3.1.5.4 Geschlechtsunterschiede

Bezüglich der Anzahl der gelösten Items konnten keine signifikanten Geschlechtsunterschiede gefunden werden [*M(weiblich)* = 5.91, *SD (weiblich)* = 6.84; *M(männlich)* = 4.78, *SD(männlich)* = 5.43; $t(128) = .89$, $p = .38$].

3.1.5.5 Fehlende Werte

Die Anzahl der fehlenden Werte war vernachlässigbar und wurde deshalb nicht weiter untersucht (siehe Tabelle 2).

Tabelle 2: Ergebnisse der statistischen Analysen aus Studie 1

Item	*p*	r_{it}	$\lambda_{EFA1;1}$	$\lambda_{EFA2;1}$	$\lambda_{EFA2;2}$	Missings
1	.14	.34	.58	.73		0 %
2	.24	.34	.48	.49	.42	0 %
3	.21	.53	.72	.75	.61	0 %
4	.22	.59	.77	.70	.71	0 %
5	.08	.34	.58	.45	.55	0 %
6	.15	.38	.58	.75	.44	0 %
7	.14	.40	.61	.68	.50	0 %
8	.17	.53	.73	.57	.70	.8 %
9	.24	.63	.82	.63	.79	1.5 %
10	.08	.29	.54	.78		.8 %
11	.15	.56	.78	.54	.77	.8 %
12	.22	.61	.79	.63	.75	.8 %
13	.21	.67	.88	.75	.82	.8 %
14	.12	.47	.76	.71	.68	1.5 %
15	.11	.51	.76	.70	.69	.8 %
16	.17	.65	.84	.68	.80	.8 %
17	.22	.47	.66	.65	.58	.8 %
18	.24	.68	.88	.48	.89	1.5 %
19	.24	.53	.76		.81	.8 %
20	.32	.58	.76	.47	.76	.8 %
21	.22	.63	.83	.44	.84	.8 %
22	.19	.68	.88	.49	.89	1.5 %
23	.24	.64	.86		.89	.8 %
24	.19	.57	.77	.48	.77	.8 %
25	.34	.59	.79	.45	.80	1.5 %
26	.18	.66	.88	.40	.90	1.5 %
27	.23	.67	.89		.92	2.3 %
28	.34	.58	.80		.83	2.3 %

Anmerkungen: p = Itemschwierigkeit; r_{it} = Trennschärfe; $\lambda_{EFA1;1}$ = Ladungen auf dem ersten Faktor der einfaktoriellen EFA; $\lambda_{EFA2;1}$ = Ladungen auf dem ersten Faktor der zweifaktoriellen EFA; $\lambda_{EFA2;2}$ = Ladungen auf dem zweiten Faktor der zweifaktoriellen EFA; Missings = fehlende Werte in Prozent; Faktorladungen < .40 sind nicht abgebildet.

3.1.6 Diskussion

In der ersten Studie sollte die Würfelkonstruktionsaufgabe als neues Format zur Erfassung räumlichen Vorstellungsvermögens vorgestellt und auf ihre basalen psychometrischen Kriterien hin überprüft werden. Es konnte gezeigt werden, dass die Items der Würfelkonstruktionsaufgabe eine reliable Skala bilden und eine sehr hohe Schwierigkeit aufweisen, womit sie im Gegensatz zu aktuellen Würfelrotationsaufgaben die Möglichkeit bieten, im hohen Fähigkeitsspektrum zu differenzieren. Die Ergebnisse der EFA zeigten einen guten Modellfit für die einfaktorielle sowie die zweifaktorielle Lösung, während die Ergebnisse des χ^2-Tests die zweifaktorielle Lösung unterstützten. Der Eigenwert des ersten Faktors war deutlich höher als die der restlichen Faktoren und die meisten Items zeigten Ladungen auf beiden Faktoren der zweidimensionalen Lösung, die zudem moderat interkorrelierten. Es ist also plausibel einen latenten Faktor (räumliches Denken) anzunehmen. Ebenso ist es plausibel, dass die Aufgabe zwei korrelierte Faktoren erfasst, z. B. die beiden oben genannten Subfaktoren Spatial Relations und Spatial Visualization, deren gemeinsame Varianz in einem hierarchischen Modell durch den übergeordneten Faktor ‚räumliches Denken' erklärt wird. Ob die beiden Faktoren tatsächlich durch die Aufgabe erfasst werden, könnte eine Studie klären, in der beide Faktoren in reiner Form gemessen und die Ergebnisse zu den Ergebnissen der Würfelkonstruktionsaufgabe in Beziehung gesetzt werden.

Es ergaben sich keine signifikanten Geschlechtsunterschiede in dieser Studie. Damit widersprechen die Ergebnisse unserer Studie einigen anderen Studien, in denen Frauen in räumlichen Aufgaben schlechter abschneiden als Männer (Halpern & Collaer, 2005; Linn & Petersen, 1985). Dieser Befund kann als Hinweis darauf gedeutet werden, dass unsere Aufgabe bezüglich des Geschlechts fairer ist. Es ist aber auch möglich, dass sich weniger Geschlechtsunterschiede zeigen, da unsere Aufgabe eine höhere Reasoning-Fähigkeit erfordert als andere räumliche Aufgaben. Geschlechtsunterschiede treten bei mentalen Rotationsaufgaben stärker auf als bei Visualisierungsaufgaben, die durch einen

höheren Reasoning-Anteil gekennzeichnet sind und eine Kombination aus visuellen und nicht-visuellen Strategien erlauben (Linn & Petersen, 1985; Voyer, Voyer & Bryden, 1995). Unsere Aufgabe ist möglicherweise zum größten Teil eine Spatial-Visualization-Aufgabe. Generell zeigen Reasoning-Aufgaben kleinere Geschlechtsunterschiede als räumliche Aufgaben (Irwing & Lynn, 2005). Die Interpretierbarkeit der Ergebnisse zu den Geschlechtsunterschieden ist allerdings stark eingeschränkt, da es sich um eine kleine Stichprobe handelte und besonders der Anteil der männlichen Studierenden (*n* = 36) gering war.

3.1.7 Ausblick

Die Würfelkonstruktionsaufgabe könnte eine wesentliche Rolle bei der Auswahl von besonders fähigen Kandidaten in MINT-Berufen spielen. Sehr fähige Kandidaten sind interessierter an solchen Berufen und erfolgreicher (hoher *Ability-Preference-Fit*; Webb et al., 2007). Durch die hohen Schwierigkeiten der Items kann die Aufgabe im Vergleich zu aktuellen Würfelrotationsaufgaben besser im hohen Fähigkeitsbereich differenzieren und somit mit hoher Wahrscheinlichkeit besser die Leistung in MINT-Domänen vorhersagen. In zukünftigen Studien könnte die inkrementelle Validität der Aufgabe gegenüber herkömmlichen Würfelrotationsaufgaben und allgemeinen Intelligenztests, die nach wie vor der beste Prädiktor für Berufserfolg sind (Jensen, 1998; Schmidt & Hunter, 1998), ermittelt werden.

Des Weiteren ist die Konstruktvalidität der Würfelkonstruktionsaufgabe genauer zu untersuchen. Die Rolle, die das Fehlen von Distraktoren in diesem Zusammenhang spielt, ist hier von besonderer Bedeutung. Die Würfelkonstruktionsaufgabe sollte jeweils in einer Version mit und ohne Distraktoren mit einer Reihe von anderen kognitiven Tests (klassische Würfelrotationsaufgabe, Tests zu logischem Schlussfolgern, verbalen und numerischen Fähigkeiten und Arbeitsgedächtnis) verglichen werden. Zudem könnte eine Studie zum lauten Denken mehr Aufschluss darüber geben, welche Strategien

beim Lösen der Aufgabe tatsächlich angewendet werden und welche kognitiven Prozesse ablaufen.

Eine Studie, in der die Würfelkonstruktionsaufgabe in einer Version mit Distraktoren mit einer Version ohne Distraktoren verglichen wird, könnte außerdem die Rolle von Distraktoren bei der Entstehung von Geschlechtsunterschieden klären. Ein Lautes-Denken-Paradigma könnte Aufschluss darüber geben, ob Frauen und Männer unterschiedliche Strategien anwenden, die unterschiedlich zielführend beim Lösen der Aufgabe sind bzw. ob unterschiedliche Strategien unterschiedlich stark Geschlechtsunterschiede fördern.

Um die Einsatzmöglichkeiten der Würfelkonstruktionsaufgabe zu vergrößern, könnte außerdem der Versuch unternommen werden, Items abgestufter Schwierigkeit zu erstellen, die im gesamten Fähigkeitsspektrum differenzieren können. Mit der Abstufung der Itemschwierigkeiten und der Untersuchung der Konstruktvalidität der Würfelkonstruktionsaufgabe befasst sich Studie 2.

3.1.8 Zusammenfassung

In Studie 1 wurde aufgrund bekannter Mängel distraktorgestützer Verfahren der Versuch unternommen, einen Test zum räumlichen Denken zu entwickeln, der ohne Distraktoren auskommt, und diesen auf seine psychometrischen Eigenschaften hin zu überprüfen. Die Würfelkonstruktionsaufgabe zeigte eine sehr gute Reliabilität und eine schlüssige Faktorstruktur. Vor allem weisen die Items eine hohe Schwierigkeit auf und die Aufgabe bietet damit einen Vorteil gegenüber gängigen Würfelrotationsaufgaben. Studie 1 stellt eine gute Grundlage für weitere Forschung zur Weiterentwicklung und Validierung der Würfelkonstruktionsaufgabe dar.

3.2 Studie 2

3.2.1 Abstract

Die Würfelkonstruktionsaufgabe stellt ein alternatives, distraktorfreies Aufgabenformat dar, bei dem die Lösung selbst konstruiert anstatt aus einer Anzahl von Antwortoptionen ausgewählt werden muss. In einer vorhergehenden Studie wurde gezeigt, dass die Items durchweg eine sehr hohe Schwierigkeit besitzen. In gängigen Würfelrotationsaufgaben fehlen Items einer solchen Schwierigkeit. Mit der folgenden Studie sollte eine Strategie, die Itemschwierigkeiten der Würfelkonstruktionsaufgabe zu variieren, getestet werden. Des Weiteren sollte die Würfelkonstruktionsaufgabe in einer Version mit Distraktoren mit einer Version ohne Distraktoren mit identischem Itemstamm verglichen und die konvergente Validität zu einem allgemeinen Intelligenztest überprüft werden. Die Stichprobe bestand aus 146 GymnasialschülerInnen (62 in der distraktorfreien Gruppe, 84 in der distraktorgestützten). Die Items der distraktorfreien Würfeltestversion konnten gut in ihrer Schwierigkeit abgestuft werden und zeigten einen breiteren Schwierigkeitsrange und damit eine genauere Differenzierungsfähigkeit ($.02 \leq p_i \leq .95$) als die Items der Version mit Distraktoren, die äquivalent zu gängigen Würfelrotationsaufgaben in ihrer Schwierigkeit homogener waren ($.37 \leq p_i \leq .63$). Des Weiteren zeigte sich für die distraktorfreie Würfeltestversion eine signifikant höhere konvergente Validität zum allgemeinen Intelligenztest als in der distraktorgestützten Version, sowohl auf manifester ($r = .57$ vs. $r = .17$) als auch auf latenter Ebene ($r = .72$ vs. $r = .31$). Die Implikationen der Ergebnisse werden diskutiert.

3.2.2 Hinführung

Wie sich in Studie 1 zeigte, besitzt die Würfelkonstruktionsaufgabe gute psychometrische Eigenschaften und Items sehr hoher Schwierigkeit. Diese Items sind dazu geeignet, zwischen sehr fähigen Kandidaten zu differenzieren. In einer

normalverteilten Stichprobe gelingt mit solchen Items jedoch keine ausreichende Differenzierung, hierfür braucht es Items jeder Schwierigkeit. In der zweiten Studie wurde deshalb der Versuch unternommen, die Items der Würfelkonstruktionsaufgabe in ihrer Schwierigkeit abzustufen, damit eine valide Messung im gesamten Fähigkeitsspektrum möglich wird. Der Ansatz zur Abstufung der Itemschwierigkeiten bestand darin, eine variierende Anzahl an Symbolen im Lösungsfeld bereits vorzugeben. Je mehr vorgegebene Symbole sich im Lösungsfeld befanden, desto leichter sollte die Aufgabe werden, da nur ein Teil des Würfels berücksichtigt werden musste. Abbildung 5 zeigt eine Beispielaufgabe mit zwei vorgegebenen Symbolen.

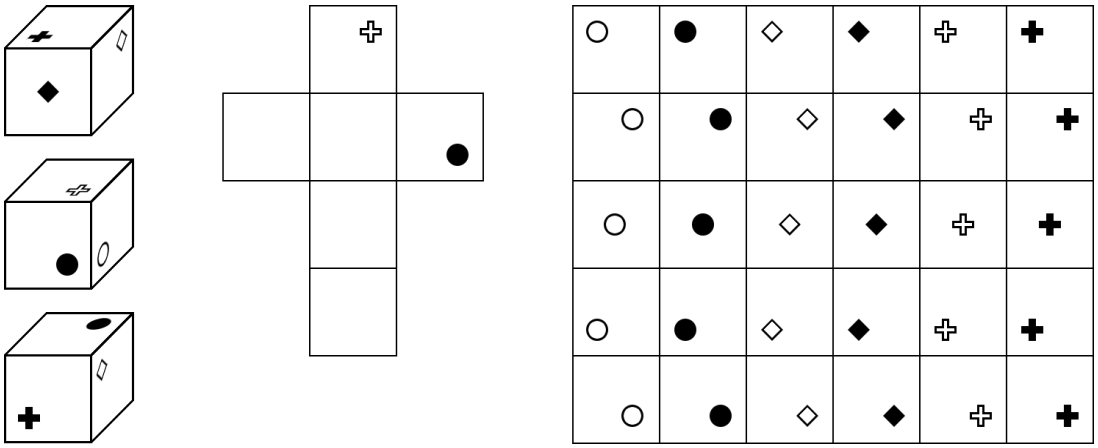


Abbildung 5: Würfelkonstruktionsaufgabe mit zwei vorgegebenen Lösungssymbolen

Ziel war zudem, eine bessere Abstufung der Itemschwierigkeiten als in gängigen Würfelrotationsaufgaben zu ermöglichen. Somit würde die Würfelkonstruktionsaufgabe allen der unter Punkt 2.3 angesprochenen Probleme gängiger Aufgabenformate begegnen. Zum einen soll die Ratewahrscheinlichkeit minimiert und das Anwenden von Response-Elimination-Strategien durch das Weglassen von Distraktoren ausgeschlossen werden. Zum anderen soll durch das neue Aufgabenformat eine Möglichkeit gefunden werden, Items variierender

Schwierigkeit zu entwickeln. Alle diese Faktoren sollten sich positiv auf die Konstruktvalidität der Aufgabe auswirken und sie für einen breiten Praxiseinsatz, bei dem im gesamten Fähigkeitsspektrum ausreichend differenziert werden kann, geeignet machen.

Um auszuschließen, dass die Itemschwierigkeiten und die konvergente Validität mit der Beschaffenheit des Itemstamms der Würfelkonstruktionsaufgabe zusammenhängen, wurde eine Version der Aufgabe mit Distraktoren erstellt, die mit gängigen Antwortformaten von Würfelrotationsaufgaben vergleichbar sein sollte. Der Itemstamm der jeweiligen Aufgabe wurde nicht verändert, d. h., die Itemstämme waren in beiden Testversionen mit und ohne Distraktoren identisch. Verändert wurde lediglich das Antwortformat. Abbildung 6 zeigt das distraktorbasierte Aufgabenformat.

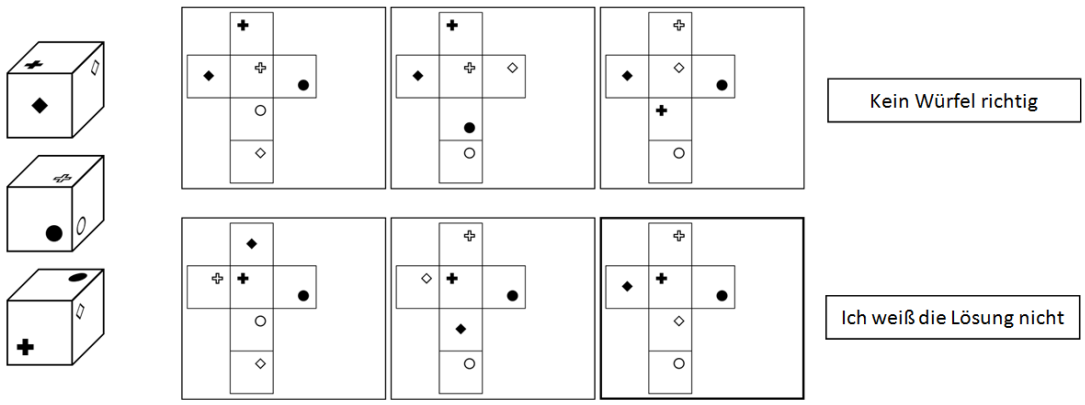


Abbildung 6: Würfelkonstruktionsaufgabe im distraktorbasierten Format (korrekte Lösung hervorgehoben)

3.2.3 Ziele und Hypothesen der zweiten Studie

Ziel der zweiten Studie war es zum einen, zu überprüfen, ob die Abstufung der Itemschwierigkeiten in der distraktorfreien Würfelkonstruktionsaufgabe möglich ist. Des Weiteren sollte das Aufgabenformat in einer Version mit Distraktoren mit einer Version ohne Distraktoren verglichen werden, um zu sehen, inwiefern sich

die Itemschwierigkeitsverteilungen und die konvergente Validität zu dem allgemeinen Intelligenztest Intelligenz-Struktur-Test-Screening (IST-Screening; Liepmann, Beauducel, Brocke & Nettelnstroth, 2012) unterscheiden. Ebenso wurden Geschlechtsunterschiede in den beiden Testversionen untersucht. Außerdem sollten die Faktorstruktur beider Testversionen sowie die latenten Korrelationen zum allgemeinen Intelligenztest überprüft werden. Dabei bestanden die folgenden Hypothesen:

Manipulation der Itemschwierigkeiten

Wir erwarteten einen positiven Zusammenhang zwischen der Anzahl vorgegebener Symbole im Lösungsfeld und der Lösungswahrscheinlichkeiten der Items.

Unterschiede in den Itemschwierigkeiten der beiden Testversionen

Da bei klassischen Würfelrotationsaufgaben abgesehen von der Einführung eines Zeitlimits für den gesamten Test kein Ansatz gefunden werden kann, die Itemschwierigkeiten zu manipulieren, sollte in der Version mit Distraktoren eine homogene Verteilung der Itemschwierigkeiten vorliegen. Unter der Annahme, dass der Ansatz zur Abstufung der Itemschwierigkeiten im distraktorfreien Format gelingt, sollte die Verteilung der Itemschwierigkeiten hier heterogener sein, also einen breiteren Range zeigen. Bezüglich der Anzahl richtig gelöster Items in den beiden Testversionen erwarteten wir keinen signifikanten Unterschied, da sich die Effekte der zusätzlichen Items sehr niedriger und sehr hoher Schwierigkeit in der distraktorfreien Version gegenseitig aufheben sollten.

Unterschiede in der konvergenten Validität der beiden Testversionen

Für die distraktorfreie Version wurde eine höhere Korrelation zu einem allgemeinen Intelligenztest erwartet, da sich das Verhindern von Raten, Response-Elimination-Strategien und die bessere Differenzierung im hohen und niedrigen Fähigkeitsspektrum positiv auf die Konstruktvalidität des Tests auswirken sollten. Indem die zu messende Fähigkeit differenzierter und valider erfasst wird, sollten

die Rangreihen der Leistungen in der Würfelkonstruktionsaufgabe und dem allgemeinen Intelligenztest mehr übereinstimmen und sich somit höhere Korrelationen ergeben. Diese unterschiedlichen Korrelationen sollten sich auch auf latenter Ebene finden lassen.

Geschlechtsdifferenzen

Für die Geschlechtsunterschiede bestand keine gerichtete Hypothese. In Studie 1 zeigten sich keine Geschlechtsdifferenzen. Da sich in Studien zum räumlichen Denken, speziell bei mentaler Rotation, jedoch oft ein Geschlechtsunterschied zugunsten der Männer zeigt (Voyer et al., 1995), sollte überprüft werden, ob sich hier Unterschiede finden lassen und ob diese je nach Testversion variieren.

Latente Modellierung

Es wurde erwartet, dass in beiden Testversionen sowohl die Items des Würfeltests als auch des ISTs auf einen latenten Faktor reduziert werden können. Des Weiteren sollte die latente Korrelation zwischen Würfelfaktor (räumliches Denken) und IST-Faktor (allgemeine Intelligenz) in der distraktorfreien Version ebenfalls höher ausfallen als in der distraktorbasierten Version.

3.2.4 Materialien und Methodik

3.2.4.1 Stichprobe und Ablauf der Testung

Die Stichprobe bestand aus 146 SchülerInnen [*M(Alter)* = 16.96, *SD(Alter)* = 1.77, 61.8 % weiblich] eines Gymnasiums in Nordrhein-Westfalen. Die Teilnahme an der Studie war freiwillig. Die Testung wurde während der regulären Unterrichtszeiten durchgeführt. Sie fand in einem laboratorischen Setting in Kleingruppen mit je nach Klassengröße bis zu 30 Testpersonen statt. Die Testpersonen füllten zunächst einen demografischen Fragebogen aus und führten anschließend gemeinsam einen ‚Paper-and-Pencil'-Gruppenintelligenztest durch (IST-Screening, siehe Punkt

3.2.4.2). Nach einer kurzen Pause starteten die Testpersonen mit dem Würfeltest am Computer. Sie wurden zufällig entweder der distraktorfreien Testversion (62 Testpersonen; *M(Alter)* = 16.8, *SD(Alter)* = 1.53, 66.7 % weiblich) oder der distraktorbasierten Testversion (84 Testpersonen; *M(Alter)* = 17.08, *SD(Alter)* = 1.93, 58.2 % weiblich) zugeteilt. Aufgrund technischer Probleme konnten die Ergebnisse des ISTs und der Würfelaufgabe von 10 Testpersonen (5 aus der distraktorfreien Gruppe und 5 aus der distraktorbasierten Gruppe) einander nicht mehr zugeordnet werden. Diese Testpersonen wurden aus der Analyse zur Konstruktvalidität ausgeschlossen.

3.2.4.2 Materialien

Allgemeiner Intelligenztest

Als Indikator für *g* wurde das IST-Screening (Liepmann et al., 2012) eingesetzt. Das IST-Screening ist ein ökonomischer Kurz-Intelligenztest, der auf der Basis des gut etablierten Intelligenz-Struktur-Tests 2000 R (IST 2000 R; Liepmann et al., 2007) erstellt wurde. Er besteht aus den drei Subtests Analogien, Zahlenreihen und Matrizen, die als Indikatoren für verbale, numerische und figurale Reasoning-Fähigkeiten angesehen werden und jeweils aus 20 Items bestehen. Aus den drei Skalen ist ein Gesamtwert ‚Schlussfolgerndes Denken' als Indikator für *g* bestimmbar. Das IST-Screening zeigt eine gute Reliabilität ($.72 \leq \alpha \leq .90$), auf konfirmatorischer Faktoranalyse basierende faktorielle Validität, konvergente Validität zu anderen Intelligenztests und Kriteriumsvalidität bezüglich Schulnoten. Der Test beinhaltet eine Zeitlimitierung für jeden Subtest und dauert insgesamt 26 Min.

Itemstämme für die Würfelaufgaben

Die Itemstämme der konstruktionsbasierten und distraktorbasierten Würfelaufgaben-Testversion waren identisch. Es wurden die in Studie 1 entwickelten Itemstämme verwendet, die Anzahl der Items wurde jedoch auf 23

Items reduziert, um die Testung ökonomischer zu gestalten. Die Auswahl der Items fand zufällig statt, da sich zwischen den Items keine bedeutsamen Unterschiede in Schwierigkeit und Trennschärfe ergaben. Ebenso wurde die liberale Zeitbegrenzung von 130 Sek. pro Item übernommen, mit einer überspringbaren Pause von 5 Sek. Somit war die maximale Bearbeitungszeit für beide Versionen der Würfelaufgabe 51.67 Min.

Distraktorfreie Würfeltestversion

Die distraktorfreien Items waren die zuvor erwähnten 23 ausgewählten Items aus Studie 1. Diesmal wurde jedoch eine variierende Anzahl an Symbolen im Lösungsfeld bereits vorgegeben (Default Symbols), um die Itemschwierigkeiten zu beeinflussen (siehe Abbildung 5). Die Variation sah wie folgt aus: Items 1–4: fünf Default Symbols, Items 5–9: vier Default Symbols, Items 10–14: drei Default Symbols, Items 15–19: zwei Default Symbols, Items 20–23: ein Default Symbol. Die Default Symbols wurden zufällig ausgewählt.

Distraktorbasierte Würfeltestversion

In der Version mit Distraktoren wurden den Testpersonen eine Reihe von Antwortoptionen präsentiert, von denen nur eine korrekt war, während es sich bei den anderen um Distraktoren handelte (siehe Abbildung 6). Der geläufigen Strategie folgend wurden die Distraktoren konstruiert, indem Symbole der Lösung durch andere Symbole ersetzt wurden. Bei zwei Distraktoren wurden zwei Symbole ersetzt, bei drei Distraktoren wurden vier Symbole ersetzt. Des Weiteren gab es die Antwortoption ‚Kein Würfel richtig' und ‚Ich weiß die Lösung nicht', um Raten und die Anwendung von Response-Elimination-Strategien zu minimieren und die Vergleichbarkeit mit gängigen Aufgabenformaten herzustellen (Gittler, 1990).

3.2.4.3 Statistische Methoden

Um die Vergleichbarkeit der allgemeinen Intelligenz der Testpersonen in den beiden Testbedingungen sicherzustellen, wurden die IST-Ergebnisse der beiden Gruppen mithilfe eines t-Tests für unabhängige Stichproben verglichen. Um sicherzustellen, dass Unterschiede zwischen den beiden Testversionen nicht auf eine unterschiedliche Reliabilität zurückzuführen sind, wurden Cronbachs Alpha (α) und die part-whole-korrigierten Trennschärfen (r_{it}) für beide Versionen ermittelt. Um zu überprüfen, ob die Vorgabe der Default Symbols den erwünschten Effekt bezüglich der Itemschwierigkeiten (p_i) hatte, wurde die Korrelation zwischen der Anzahl der Default Symbols und den Itemschwierigkeiten der distraktorfreien Version ermittelt. Der Unterschied in den Summenwerten der beiden Versionen wurde mittels t-Test für unabhängige Stichproben auf Signifikanz überprüft. Der Unterschied der Varianzen der Summenwerteverteilungen wurde mittels Levene-Test (Levene, 1960) auf Signifikanz überprüft. Die Signifikanz des Unterschiedes zwischen den mittleren Itemschwierigkeiten der beiden Versionen wurde mit einem t-Test für abhängige Stichproben ermittelt. Des Weiteren wurde mittels Levene-Test überprüft, ob sich die Varianz der Schwierigkeitsverteilung der Testversion mit Distraktoren von der Varianz der Schwierigkeitsverteilung der Testversion ohne Distraktoren unterschied. Die Konstruktvaliditäten der beiden Testversionen wurden analysiert, indem jeweils Pearson-Korrelationen gebildet wurden zwischen dem Summenwert der Würfelaufgaben und dem Aggregatwert des ISTs bzw. der IST-Subtests. Unterschiede in der Validität der beiden Testversionen wurden mit dem Signifikanztest von Millsap, Zalkind & Xenos (1990) überprüft. Die Faktorstruktur der beiden Testversionen und die latenten Korrelationen wurden mittels konfirmatorischer Faktoranalyse (CFA) in R (R Core Team, 2015) mit dem Paket lavaan (Rosseel, 2012) berechnet. Aufgrund der Stichprobengröße wurden hierzu aus den Items der drei IST-Subtests drei Parcels gebildet und äquivalent dazu aus den Items der Würfelaufgabe ebenfalls drei Parcels, die in ihrer Schwierigkeit jeweils vergleichbar heterogene Items enthielten.

Die Geschlechtsunterschiede in den Summenwerten der Würfelaufgabe wurden mittels t-Test für unabhängige Stichproben untersucht.

3.2.5 Ergebnisse

Vergleich der allgemeinen Intelligenz

Die Testpersonen der distraktorfreien Gruppe zeigten einen mittleren Summenwert von *M(sum)* = 37.89 [*SD(sum)* = 8.56] im IST, die Testpersonen aus der distraktorbasierten Gruppe einen mittleren Summenwert von *M(sum)* = 39.96 [*SD(sum)* = 6.54]. Ein t-Test für unabhängige Stichproben offenbarte keinen signifikanten Unterschied zwischen den beiden Gruppen [$t(134) = -1.59$, $p = .11$].

Interne Konsistenzen

Mit Cronbachs $\alpha = .81$ für die distraktorbasierte Würfelaufgabe und Cronbachs $\alpha = .83$ für die distraktorfreie Würfelaufgabe zeigten beide Skalen eine vergleichbar gute Reliabilität. Die part-whole-korrigierten Trennschärfen der Items beider Skalen finden sich in Tabelle 3. Die mittleren Trennschärfen von $M(r_{it}) = .40$ [$SD(r_{it}) = .15$; $09. \leq r_{it} \leq .49$] für die distraktorbasierte Version und von $M(r_{it}) = .36$ [$SD(r_{it}) = .12$; $-01. \leq r_{it} \leq .67$] für die distraktorfreie Version waren ebenfalls gut und vergleichbar.

Itemschwierigkeiten

Die Itemschwierigkeiten sind zusammen mit der Anzahl der Default Symbols in der distraktorfreien Version in Tabelle 3 aufgelistet. Die Korrelation der Anzahl an vorgegebenen Default Symbols in der distraktorfreien Version mit den dazugehörigen Itemschwierigkeiten war $r = .71$ ($p < .01$), was darauf hinweist, dass die Strategie zur Manipulation der Itemschwierigkeiten erfolgreich war. Die Betrachtung der mittleren Itemschwierigkeiten zeigte, dass die Items der distraktorfreien Version im Mittel schwerer zu lösen waren [$M(p_i) = .27$] als die der distraktorbasierten Version [$M(p_i) = .46$]. Ein t-Test für abhängige Stichproben zeigte, dass der Unterschied zwischen den beiden Mittelwerten signifikant war

[$t(22) = 3.07$, $p < .01$]. Der Range der Itemschwierigkeiten der distraktorfreien Version ($.02 \leq p_i \leq .95$) war substantiell größer als in der distraktorbasierten Version ($.37 \leq p_i \leq .63$). Des Weiteren war die Standardabweichung der Verteilung der Itemschwierigkeiten in der distraktorfreien Version [$SD(p_i) = .26$] größer als in der distraktorbasierten [$SD(p_i) = .07$]. Die Ergebnisse eines Levene-Tests zeigten, dass sich die Varianzen der Itemschwierigkeitsverteilungen der beiden Testversionen signifikant voneinander unterschieden [$F(1,44) = 17.14$, $p < .01$].

Summenwerte

Der mittlere Summenwert in der distraktorbasierten Testversion [$M(sum) = 10.63$, $SD(sum) = 5.03$] war höher als in der distraktorfreien Version [$M(sum) = 6.29$, $SD(sum) = 3.94$]. Der Unterschied in den Summenwerten war signifikant [$t(44) = -5.84$, $p < .01$], genauso wie der Unterschied in den Varianzen der Summenwerteverteilungen [$F(1,44) = 5.27$, $p < .05$].

Geschlechtsunterschiede

In der distraktorfreien Version ergab sich ein signifikanter Unterschied [$t(55) = 2.49$, $p < .05$] in den Summenwerten zugunsten der Männer gegenüber den Frauen [$M(sum) = 8.21$ vs. $M(sum) = 5.50$]. Dieser Unterschied konnte in der distraktorbasierten Testversion nicht gefunden werden [$M(sum) = 10.82$ vs. $M(sum) = 10.57$; $t(77) = .22$, $p = .83$].

Tabelle 3: Itemschwierigkeiten und part-whole-korrigierte Trennschärfen für beide Testversionen aus Studie 2

	distraktorfrei			*distraktorbasiert*	
Item	*Default Symbols*	p_i	r_{it}	p_i	r_{it}
1	5	.44	.39	.40	.10
2	5	.11	.29	.37	.09
3	5	.84	.37	.40	.26
4	5	.95	.31	.45	.32
5	4	.55	.38	.38	.42
6	4	.11	.67	.55	.16
7	4	.26	.33	.38	.46
8	4	.42	.38	.52	.25
9	4	.60	.32	.40	.48
10	3	.03	−.01	.63	.40
11	3	.24	.43	.43	.42
12	3	.34	.55	.50	.40
13	3	.31	.49	.45	.49
14	3	.32	.35	.37	.29
15	2	.06	.43	.50	.41
16	2	.23	.50	.49	.47
17	2	.08	.55	.51	.38
18	2	.02	.39	.43	.44
19	2	.18	.58	.51	.45
20	1	.05	.42	.55	.35
21	1	.08	.45	.45	.41
22	1	.02	.09	.50	.46
23	1	.06	.45	.44	.29
M	3	.27	.36	.46	.40
SD	1.38	.26	.12	.07	.15

Anmerkungen: p_i = Itemschwierigkeit*;* r_{it} = part-whole-korrigierte Trennschärfe; *M* = Mittelwert; *SD* = Standardabweichung.

Konvergente Validitäten

Die Korrelationen der Summenwerte der beiden Testversionen mit den Werten der IST-Subtests und dem IST-Aggregatwert sind in Tabelle 4 dargestellt. Der Summenwert der distraktorfreien Würfelaufgabe zeigte substantielle und signifikante Korrelationen zu allen Werten des ISTs. Für die distraktorbasierte Würfelaufgabe waren die Korrelationen substantiell geringer und nur in einem Fall (d. h. zum figuralen Subtest des ISTs) signifikant. Signifikanztests offenbarten, dass mit Ausnahme des verbalen Subtests des ISTs alle Korrelationen für die distraktorfreie Version zum IST höher waren als für die distraktorbasierte.

Tabelle 4: Korrelation zwischen Leistung in Würfelaufgabe und Intelligenz in Studie 2

	DF		**DB**		**DF vs. DB**	
	r	***p***	***r***	***p***	***z***	***p***
IST_V	.36	.01	.11	.36	1.50	.07
IST_N	.47	< .001	.02	.88	1.73	.04
IST_F	.54	< .001	.28	.01	1.78	.04
IST_A	.57	< .001	.17	.13	2.67	< .001

Anmerkungen: DF = distraktorfreie Version; DB = distraktorbasierte Version; IST_V = verbaler Subtest des IST; IST_N = numerischer Subtest des IST; IST_F = figuraler Subtest des IST; IST_A = Aggregatwert des IST; *r* = Pearson-Korrelation; *z* = *z*-Wert für die Differenz zwischen Korrelationen; *p* = Signifikanz.

Latente Modellierung

Das Strukturmodell zeigte in der Gruppe der distraktorfreien Würfeltestversion einen guten Modellfit: $\chi^2(8) = 6.58$; $p = .58$; CFI = 1; RMSEA = 0. Die Ladungen der drei Parcels für den Würfelfaktor lagen hier zwischen .78 und .82, für den IST-Faktor zwischen .60 und .79. Die Interkorrelation der beiden latenten Faktoren betrug $r = .72$ ($p < .01$). Das Strukturmodell für die distraktorfreie Version ist in Abbildung 7 dargestellt.

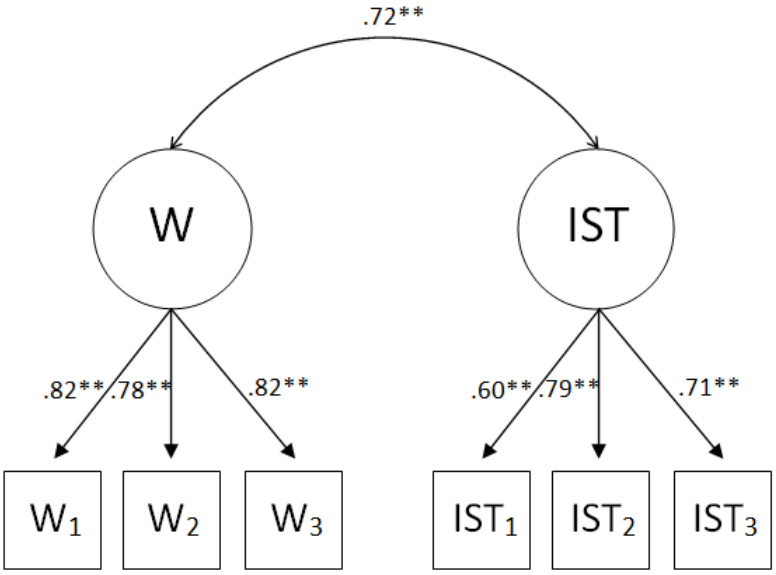


Abbildung 7: Strukturmodell für die distraktorfreie Gruppe

Für die distraktorbasierte Version zeigte sich ebenfalls ein guter Modellfit für das Strukturmodell: $\chi^2(8) = 11.512$; $p = .19$; CFI = .96; RMSEA = .07. Die Ladungen der Parcels für den Würfelfaktor lagen zwischen .66 und .80, für den IST-Faktor zwischen .50 und .60. Die Interkorrelation der beiden latenten Faktoren betrug $r = .31$ ($p = .09$). Das Strukturmodell der distraktorgestützten Version findet sich in Abbildung 8.

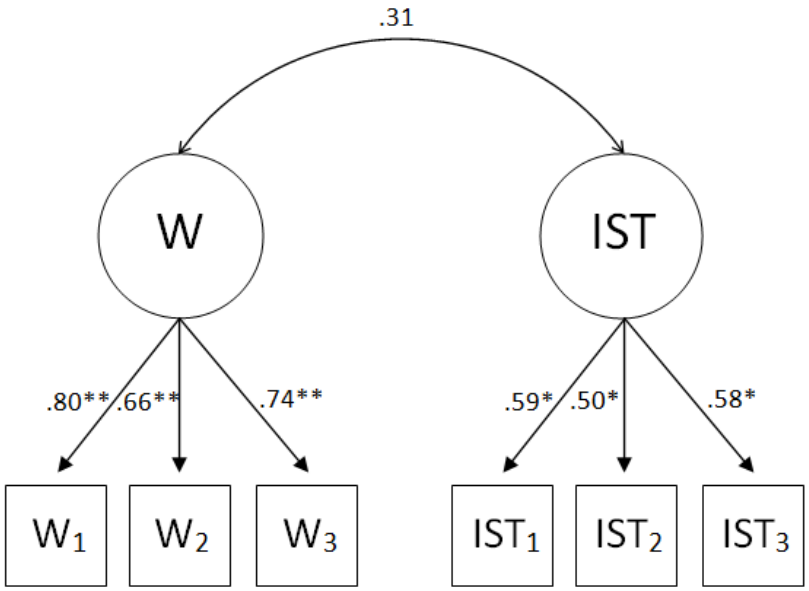


Abbildung 8: Strukturmodell für die distraktorgestützte Gruppe

3.2.6 Diskussion

Ziel der Studie war es, zu überprüfen, ob der Ansatz zur Manipulation der Itemschwierigkeiten der Würfelkonstruktionsaufgabe funktioniert. Weiterhin sollten die Summenwerte, die Schwierigkeiten und die Konstruktvaliditäten der distraktorfreien und der distraktorbasierten Würfeltestversion mit gleichen Itemstämmen verglichen werden. Des Weiteren sollte die Korrelation zwischen IST und Würfelaufgabe latent modelliert werden.

Die Ergebnisse der Studie sind geradlinig. Eine positive Korrelation zwischen der Anzahl vorgegebener Lösungssymbole und den Itemschwierigkeiten in der distraktorfreien Würfeltestversion wies darauf hin, dass der Ansatz zur Manipulation der Itemschwierigkeiten erfolgreich war. Da die beiden Gruppen sich hinsichtlich ihrer allgemeinen Intelligenz nicht unterschieden und die Reliabilitäten beider Würfeltestversionen vergleichbar waren, können Unterschiede zwischen den beiden Würfeltestversionen nicht auf diese Variablen zurückgeführt werden. Die Items der distraktorfreien Würfelaufgabe waren signifikant schwerer zu lösen

als die Items der distraktorbasierten Würfelaufgabe. Da einige Items der distraktorfreien Version Schwierigkeiten aufwiesen, die niedriger waren als die mittlere Itemschwierigkeit der distraktorbasierten Version, kann trotzdem nicht geschlussfolgert werden, dass die Items in der distraktorfreien Würfelaufgabe generell schwerer sind als in der distraktorbasierten Version. Stattdessen kann angenommen werden, dass ein Würfelkonstruktionstest mit Items sehr niedriger Schwierigkeit generiert werden kann, indem in allen Items viele Lösungssymbole vorgegeben werden. Der Range der Itemschwierigkeiten der distraktorfreien Würfelaufgabe deckte ein breites Spektrum ab und zeigte, dass es mit der Würfelkonstruktionsaufgabe möglich ist, Items fast jeder Schwierigkeit zu generieren. Im Gegensatz dazu war der Schwierigkeitsrange der distraktorbasierten klassischen Würfelaufgabe enger gefasst und lag im mittleren Bereich des Fähigkeitsspektrums. Dies spiegelt sich auch darin wider, dass die Varianz der Itemschwierigkeitsverteilung der distraktorfreien Würfelaufgabe signifikant größer war als die der distraktorbasierten Version.

Entgegen unserer Hypothese unterschieden sich die Summenwerte der beiden Würfeltestversionen signifikant. In der distraktorbasierten Version wurden mehr Items gelöst als in der distraktorfreien Version und die Verteilung der Summenwerte zeigte eine größere Varianz. Dieser Befund könnte sich ändern, wenn mehr leichtere Items in die Würfelkonstruktionsaufgabe aufgenommen würden, sodass der Test besser dem Fähigkeitsniveau einer vergleichbaren Stichprobe entspräche.

Es kann festgehalten werden, dass die Würfelkonstruktionsaufgabe mit Default Symbols eine bessere Differenzierbarkeit entlang des gesamten Fähigkeitsspektrums erlaubt als gängige distraktorbasierte Würfelrotationsaufgaben. Es ist eventuell möglich, auch die Schwierigkeiten von distraktorbasierten Würfelaufgaben mehr zu variieren, es finden sich in gängigen Aufgaben jedoch keine Strategien hierzu und dies war auch nicht Ziel der Studie. Stattdessen sollte gezeigt werden, dass die Würfelkonstruktionsaufgabe eine gute

Alternative zu gängigen Aufgaben darstellt, da sie eine Manipulation der Itemschwierigkeiten ermöglicht.

Dass es in der distraktorfreien Version im Gegensatz zur distraktorbasierten einen signifikanten Geschlechtsunterschied zugunsten der männlichen Testpersonen gab, spricht dafür, dass in dieser Version räumliches Denkvermögen valide gemessen wird, denn dies ist ein typischer replizierter Unterschied. Warum sich der Unterschied in der distraktorbasierten Version nicht zeigt, sollte genauer untersucht werden. Möglicherweise zeigen sich hier durch die starke Anwendung von Response-Elimination-Strategien keine Geschlechtsunterschiede. Ebenso kann die Stichprobengröße und vor allem deren Zusammensetzung für den Befund verantwortlich sein. Möglicherweise handelt es sich um einen kleinen Effekt, der bei einer Stichprobengröße von 84 Personen nicht signifikant werden kann. Zudem ist die Stichprobe auch insofern varianzeingeschränkt, als der überwiegende Teil der Testpersonen zwischen 12 und 16 Jahre alt war und es sich nur um GymnasialschülerInnen handelte. Bei Heranwachsenden sind je nach Geschlecht unterschiedliche Entwicklungssprünge möglich, die die Vergleichbarkeit erschweren (Voyer et al., 1995). Zur genaueren Untersuchung der Geschlechtsunterschiede ist die Erstellung einer Stichprobe von Erwachsenen oder von SchülerInnen, bei denen ausreichend große homogene Altersgruppen verglichen werden können, sinnvoll. Zudem sollten andere räumliche Aufgaben zum Vergleich erfasst werden.

Schließlich waren die konvergenten Validitäten der distraktorfreien Würfeltestversion substantiell und in fast allen Fällen signifikant höher als bei der distraktorbasierten Version. Dies zeigte sich auch auf latenter Ebene. Die konvergente Validität der distraktorfreien Würfeltestversion ist mit einer Höhe von r = .57 des Weiteren höher als die des 3DW (Gittler, 1990), der den einzigen zum Vergleich heranziehbaren eigenständigen Würfelrotationstest im deutschsprachigen Raum darstellt. Hier finden sich Korrelationen von .48 zum Wiener Matrizen-Test (Forman & Piswanger, 1979) und von .35 zum IST 70 (Amthauer, 1970). Dieser Befund kann durch die bessere Differenzierbarkeit der

Würfelkonstruktionsaufgabe im hohen und niedrigen Fähigkeitsspektrum erklärt werden. Eine weitere Erklärung ist die Verhinderung von Response-Elimination-Strategien. Inwiefern die höhere konvergente Validität des Konstruktionsformats auf die Verhinderung von Response-Elimination-Strategien und/oder auf den breiteren Schwierigkeitsrange zurückgeht, muss in einer weiteren Studie geklärt werden. Um die beiden Effekte voneinander zu trennen, braucht es eine Version mit Distraktoren, die den gleichen Schwierigkeitsrange in den Items aufweist, sodass Unterschiede in den Korrelationen auf die Verhinderung von Response-Elimination zurückgeführt werden können. Für den Fall, dass sich dann keine Unterschiede mehr ergeben, lässt das den Rückschluss zu, dass die Unterschiede in den Korrelationen in unserer Studie auf die unterschiedlichen Itemschwierigkeiten zurückgehen. Wenn sich ein kleinerer Unterschied als in unserer Studie ergibt, deutet das darauf hin, dass sich sowohl die Itemschwierigkeiten als auch die Möglichkeit Response-Elimination anzuwenden auf die Konstruktvalidität auswirken. Um die Lösungsstrategien, die tatsächlich angewendet werden, noch genauer zu beleuchten, kann eine Studie, in der lautes Denken zum Einsatz kommt, aufschlussreich sein. Erst mit einem solchen Paradigma werden die ablaufenden kognitiven Prozesse beim Lösen der Aufgabe deutlich.

Auch hier kann die Frage gestellt werden, warum die Korrelation der distraktorbasierten Würfeltestversion zum allgemeinen Intelligenztest so niedrig und nicht signifikant war. Bei anderen distraktorgestützten Verfahren zeigen sich keine viel größeren Korrelationen, aber zumindest etwas höhere, z. B. bei Gittler (1990) r = .35 zum IST 70 (Amthauer, 1970). Auch hier kann die Stichprobengröße und -zusammensetzung verantwortlich sein. Insofern sollte die Studie mit einer größeren und weniger varianzeingeschränkten Stichprobe repliziert werden, um den Effekt zu überprüfen. Sollte sich dann derselbe Effekt zeigen, kann es auch sein, dass unser Aufgabenmaterial besonders gut Response-Elimination erlaubt und sich deshalb eine kleine Korrelation einstellt. In dem Fall weisen die Ergebnisse aber weiterhin darauf hin, dass die Beschaffenheit des Antwortformats einen entscheidenden Einfluss auf die Validität hat. Eine Korrelation von r = .17 ist zudem

nicht bedeutungslos. Der Fakt, dass die konstruktionsbasierte Version eine deutlich bessere konvergente Validität zeigt als ein distraktorbasiertes Format wie z. B. der 3DW von Gittler (1990), bleibt bestehen.

3.2.7 Zusammenfassung

Studie 2 hat gezeigt, dass die Würfelkonstruktionsaufgabe mit vorgegebenen Lösungssymbolen einen vielversprechenden Ansatz darstellt, einen Test mit einer hohen Konstruktvalidität zu generieren, mit dem im gesamten Fähigkeitsspektrum räumlichen Denkens zwischen Testpersonen differenziert werden kann. Zudem zeigte sich für das konstruktionsbasierte Format eine höhere konvergente Validität und ein breiterer Schwierigkeitsrange der Items als für ein mit üblichen Aufgaben in diesem Bereich vergleichbares distraktorgestütztes Format. Die Ergebnisse könnten ForscherInnen und PraktikerInnen ermutigen, distraktorfreie Würfelkonstruktionsaufgaben zur Erfassung räumlichen Denkens einzusetzen.

3.3 Studie 3

3.3.1 Abstract

Es ist nach theoretischen Überlegungen möglich, figurale Matrizenaufgaben durch eine Methode des Auszählens und der Schnittmengenbildung zu lösen, ohne den Itemstamm überhaupt zu betrachten. Eine solche Methode könnte eine kognitiv weniger aufwändige Umgehungsstrategie des eigentlichen Lösungsweges darstellen. In der folgenden Studie sollte überprüft werden, inwiefern die Kenntnis einer solchen Antwortausschlussstrategie sich auf die Leistung der Testpersonen und die konvergente Validität eines figuralen Matrizentests zu einem allgemeinen Intelligenztest auswirkt. Dazu durchliefen 64 Studierende ein Training zur Antwortausschlussstrategie und bearbeiteten davor und danach einen figuralen Matrizentest. Es konnte gezeigt werden, dass sich die Leistung der Testpersonen nach Instruktion zur Strategie (M = 5.56, SD = 1.80) im Vergleich zu vorher nicht verbesserte [M = 5.94, SD = 1.6; $t(63)$ = 1.30, p = .12, d = .26]. Des Weiteren verschlechterte sich die Konstruktvalidität des figuralen Matrizentests nicht signifikant ($r(prä)$ = .26 vs. $r(post)$ = .42). Zudem zeigte sich eine Korrelation von r = .50 (p < .01) zwischen dem Antwortausschlusstest und dem allgemeinen Intelligenztest, was darauf hindeutet, dass die Fähigkeit, die Antwortausschlussstrategie anzuwenden, Ausdruck der allgemeinen Intelligenz und somit keine Gefahr für die Konstruktvalidität des Tests ist. Die Ergebnisse werden eingeordnet und ihre Implikationen diskutiert.

3.3.2 Hinführung

Studie 3 befasst sich weiterhin mit dem Einfluss, den die Vorgabe von Distraktoren auf die Testvalidität haben kann, diesmal jedoch im Bereich figuraler Matrizenaufgaben. Die Forschung zu den Auswirkungen des Antwortformats auf die Validität kognitiver Fähigkeitstests konzentrierte sich anfangs, wie weiter oben bereits erwähnt, auf den Kontext figuraler Matrizenaufgaben. Hier existieren

bereits eine Reihe von Studien zur Anwendung von Constructive-Matching- / Response-Elimination-Strategien bei der Testbearbeitung und deren Auswirkungen auf die Aufgabenschwierigkeit und die Konstruktvalidität (Arendasy & Sommer, 2013; Becker et al., 2016; Bethell-Fox et al., 1984; Jarosz & Wiley, 2012). Über diese beiden Strategien hinaus wurde in diesem Kontext in der Fachliteratur eine neue, zuvor nicht angesprochene und wenig erforschte dritte mögliche Lösungsstrategie genannt, nämlich die, die Aufgabe nur durch Inspektion der Distraktoren bei völliger Außerachtlassung des Itemstamms zu lösen (Mittring & Rost, 2008; White & Zammarelli, 1981). Diese Strategie sollte in Studie 3 genauer untersucht werden. Zum besseren Verständnis soll hier zunächst auf den Aufbau von Matrizenaufgaben näher eingegangen werden.

3.3.2.1 Figurale Matrizenaufgaben

Figurale Matrizenaufgaben sind ein bedeutsames Aufgabenformat im Bereich der Intelligenzdiagnostik (Jensen, 1998). Sie zeigen besonders hohe Korrelationen zur allgemeinen Intelligenz *g* (Marshalek et al., 1983) und der Lösungsprozess, der beim Bearbeiten der Aufgaben stattfindet, ist gut erforscht (Carpenter, Just & Shell, 1990). Matrizenaufgaben erfassen die sog. fluide Intelligenz. Rost (2009) definiert fluide Intelligenz (bzw. *Gf*, *Reasoning*) als „grundlegende Prozesse schlussfolgernden Denkens (Induktion und Deduktion, Klassifikation und Begriffsbildung), welche nur minimal von Lernerfahrungen und Akkulturation abhängen“ (S. 60). Fluide Intelligenz besitzt in verschiedenen Intelligenzmodellen einen zentralen Stellenwert (Carroll, 1996; Horn & Cattell, 1966; McGrew, 2009) und ist ein bedeutender Prädiktor für viele Bereiche des menschlichen Lebens (Brand, 1987; Gottfredson, 1997, 2004; Neisser et al., 1996). Abbildung 9 zeigt ein Beispiel einer klassischen figuralen Matrizenaufgabe.

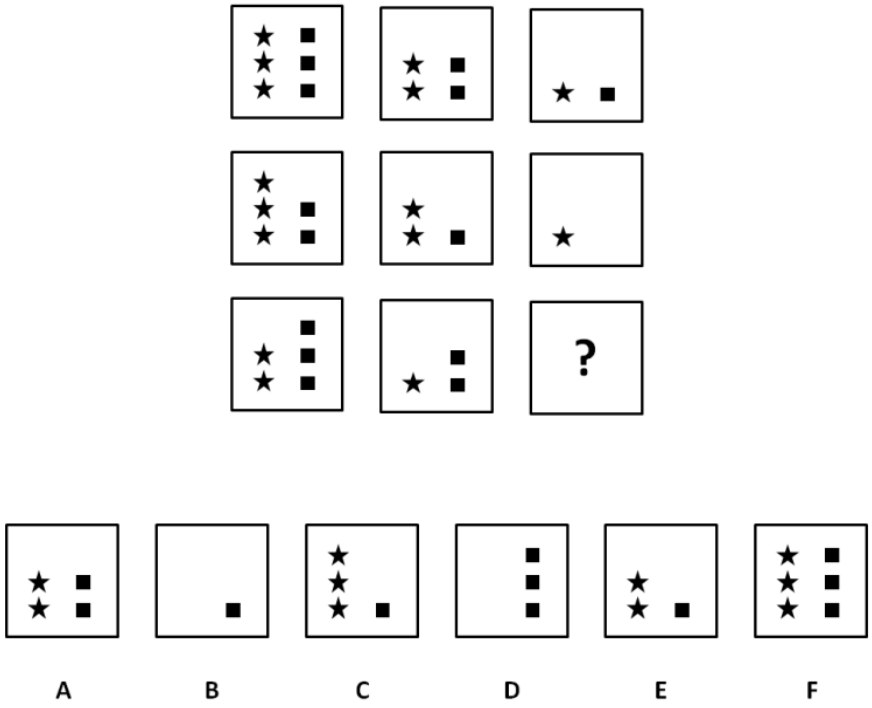


Abbildung 9: Klassisches figurales Matrizen-Item

Der Itemstamm befindet sich im oberen Teil der Abbildung. Er besteht aus einer 3-x-3-Matrix von Zellen mit figuralen Elementen (d. h. geometrischen Symbolen). Diese Elemente folgen bestimmten Gestaltungsregeln, die es zu erkennen gilt. In diesem Beispiel besteht eine Zelle aus zwei Spalten, in der sich links Sterne und rechts Quadrate befinden. Von links nach rechts gesehen verschwinden in einer Reihe sukzessive jeweils ein Stern und ein Quadrat pro Zelle, hier wird also eine Subtraktionsregel realisiert. Die letzte Zelle (Lösungszelle) ist leer.

Im unteren Teil der Abbildung befindet sich das Antwortformat. Es besteht u. a. aus der korrekten Lösung (Attraktor), die die Matrix nach den Regeln im Itemstamm logisch vervollständigt. In diesem Fall ist die korrekte Lösung Antwortoption B. Die anderen Antwortoptionen sind falsche Antworten (Distraktoren) und vervollständigen die Matrix nicht logisch. Die Aufgabe besteht darin, die richtige Antwortoption auszuwählen. Die Allgemeine Intelligenz *g*

befähigt Personen, die in den Matrizen vorkommenden Regeln zu erfassen. Die Arbeitsgedächtniskapazität erlaubt es Personen, den Lösungsprozess zu überwachen.

3.3.2.2 Lösungsstrategien

Betrachtung des Itemstamms und des Antwortformats

Wie bereits angesprochen, ergeben sich für das distraktorgestützte Aufgabenformat verschiedene Probleme. Neben der Ratewahrscheinlichkeit können auch bei figuralen Matrizenaufgaben zwei Lösungsstrategien angewendet werden (Arendasy & Sommer, 2013; Becker et al., 2015; Mitchum & Kelley, 2010; Vigneau et al., 2006). Die erwünschte Lösungsstrategie stellt das schon genannte Constructive Matching dar, bei dem lediglich der Itemstamm betrachtet wird und die Lösung dann selbst kognitiv generiert wird. Anschließend wird die richtige Lösungsoption ausgewählt. Die alternative Lösungsstrategie Response-Elimination besteht darin, die Antwortoptionen mit dem Itemstamm zu vergleichen, um falsche Antworten sukzessive auszuschließen und eine der übrig bleibenden Optionen auszuwählen. Hier wird also nicht nur der Itemstamm, sondern auch das Antwortformat bereits in den Lösungsprozess mit einbezogen.

Ausschließliche Betrachtung des Antwortformats

Eine dritte, bisher noch nicht ausführlich betrachtete Lösungsstrategie besteht darin, dass bei Außerachtlassung des Itemstamms lediglich das Antwortformat betrachtet wird, um zur Lösung zu gelangen. Diese Lösungsstrategie wurde theoretisch beschrieben, ihre tatsächliche Anwendung bis jetzt aber noch nicht empirisch überprüft. Die Lösung kann durch eine Kombination von einfachen Auszählprinzipien gefunden werden, wobei es nicht nötig ist, die Regeln der Matrizen zu erkennen. Laut Mittring und Rost (2008) kann eine beträchtliche Anzahl von gängigen Matrizenaufgaben auf diese Art und Weise gelöst werden. Die

Autoren führten theoretische Analysen verschiedener figuraler Matrizenaufgaben durch und konnten zeigen, dass es möglich ist, in drei Schritten den erwünschten Lösungsprozess zu umgehen. Die dort beschriebene Antwortausschlussstrategie soll in Abbildung 10 am Beispiel des Antwortformats eines Items des Wiener Matrizen-Tests (WMT; Formann, 1979) erklärt werden. Wie auch bei anderen Matrizenaufgaben bestehen die Items des WMT aus einer 3-x-3-Matrix, wobei die letzte Zelle rechts unten fehlt. Es gibt acht Antwortoptionen: sieben Distraktoren und eine richtige Lösung (Attraktor). Die Testperson muss die richtige Lösung aus den Antwortoptionen auswählen.

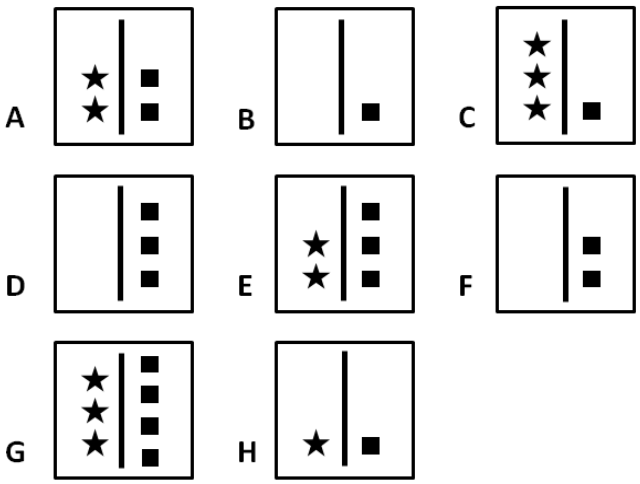


Abbildung 10: Rekonstruierte Antwortoptionen des WMT (Formann & Piswanger, 1979)

Die Analyse erfolgt in drei Schritten:

Schritt 1: Identifikation der Matrix-Komponenten

Eine Matrix kann mehrere Komponenten haben, zum Beispiel Kreise, Quadrate, Punkte, Linien, Positionen, Farben, etc. Diese Komponenten müssen identifiziert werden. Im Beispiel bestehen alle Antwortoptionen aus einer linken und einer rechten Hälfte, die durch eine Linie getrennt werden. Die linke Hälfte ist jeweils mit

0–3 Sternen (Komponente 1) und die rechte Hälfte jeweils mit 1–4 Quadraten gefüllt (Komponente 2).

Schritt 2: Auszählen der Komponenten

In der linken Hälfte befinden sich dreimal kein Stern (B, D, F), einmal ein Stern (H), zweimal zwei Sterne (A, E) und zweimal drei Sterne (C, G). Die Alternative mit keinem Stern ist am häufigsten vertreten. Folgt man der simplen Annahme, dass die Matrix-Komponente, die am häufigsten vorkommt, wahrscheinlich die richtige ist, sollte also B, D oder F die Lösung sein. In der rechten Hälfte befinden sich dreimal ein Quadrat (B, C, H), zweimal zwei Quadrate (A, F), zweimal drei Quadrate (D, E) und einmal vier Quadrate (G). Die Alternative mit einem Quadrat kommt am häufigsten vor und sollte somit Teil der Lösung sein. Es bleiben hier also die Antwortoptionen B, C und H übrig.

Schritt 3: Bildung einer Schnittmenge

Als Letztes werden die übrig gebliebenen Antwortoptionen verglichen: Auf der linken Seite bleiben die Optionen B, D und F übrig, auf der rechten Seite die Optionen B, C und H. Nur B ist in beiden Mengen enthalten, B sollte also die richtige Lösung sein. Tatsächlich ist Antwortoption B der Attraktor.

Laut Mittring und Rost (2008) können 75 % der Items des WMT mithilfe dieser kognitiv wenig fordernden Strategie des Auszählens und Schnittmengenbildens gelöst werden. Ebenso 35 % der Items des BOMAT (Hossiep, Turck & Hasella, 1999), 50 % der Items der Raven's Advanced Progressive Matrices (1976) und 70 % der Items von Naglieris (2003) non-verbalem Intelligenztest. Es muss die Frage gestellt werden, ob die Konstruktvalidität eines solchen Tests beeinträchtigt ist, wenn es möglich ist, die Aufgaben auf eine weniger fordernde Weise, als bei der Testkonstruktion beabsichtigt, zu lösen.

White und Zammarelli (1981) beschreiben ebenfalls mehrere Prinzipien, nach denen es möglich ist, figurale Matrizenaufgaben ohne Inspektion des Itemstamms zu lösen. Sie sollen hier nicht im Detail erklärt werden, da sie eine große Ähnlichkeit

mit den Prinzipien von Mittring und Rost (2008) aufweisen. Erwähnenswert ist jedoch, dass sie eine Studie durchführten, um ihre Annahme zu überprüfen. Einer Stichprobe von 35 SchülerInnen wurden nur die Antwortoptionen von fünf Items des Culture Fair Test (Cattell & Cattell, 1973) vorgelegt. Die richtige Lösung war immer angegeben und die SchülerInnen sollten sich eine Strategie überlegen, auf die richtige Lösung zu kommen. Sie durchliefen also ein Training. Danach sollten sie andere Items des Tests nur durch Inspektion der Antwortoptionen lösen. Es gab keine Instruktion, wie die Items zu lösen waren. Die Ergebnisse zeigten, dass die tatsächliche Häufigkeitsverteilung der richtigen Antworten signifikant verschieden von der erwarteten zufälligen Häufigkeitsverteilung war. Die durchschnittliche Anzahl richtiger Antworten war höher als bei zufälligem Lösen zu erwarten und es gab mehr Extremwerte. Die SchülerInnen hatten offensichtlich Strategien angewendet, um die Aufgaben ohne Inspektion des Itemstamms zu lösen. Die Korrelation zu einem allgemeinen Intelligenztest (AH4 Group Test of General Intelligence; Heim, 1968, 1970) war $r = .31$. Nach Meinung der Autoren ist dies eine so niedrige Korrelation, dass nicht davon ausgegangen werden kann, dass beim Bearbeiten der beiden Tests dieselben kognitiven Prozesse zum Tragen kommen.

3.3.3 Ziele der Studie 3

Studie 3 hatte mehrere Ziele: Erstens sollte untersucht werden, ob die Kenntnis der theoretischen Prinzipien zum Ausschließen falscher Antworten bei ausschließlicher Inspektion der Antwortoptionen wie sie oben beschrieben wurden (Antwortausschlussstrategie, kurz AAS) zu einer verringerten konvergenten Validität des Tests führt. Zu diesem Zweck wurde ein Prä-Post-Design entwickelt, in dem die Testpersonen figurale Matrizenaufgaben zunächst ohne Kenntnis der Antwortausschlussstrategien bearbeiteten und noch einmal nach einem Training, bei dem die Prinzipien erlernt wurden. Die Ergebnisse wurden dann mit den Ergebnissen eines allgemeinen Intelligenztests korreliert. Zweitens sollte überprüft werden, ob die Kenntnis der Antwortausschlussstrategien zu einer Verbesserung

der Testleistung führt. Um zu überprüfen, ob die AAS überhaupt angewendet wird, wurde ihre Anwendungshäufigkeit erfragt. Ebenfalls untersucht werden sollte, ob die Häufigkeit, mit der eine Testperson die AAS im Post-Matrizentest anwendet, mit der allgemeinen Intelligenz oder der Leistung im Post-Matrizentest korreliert. Abschließend sollte untersucht werden, ob die Fähigkeit, Antwortausschlussstrategien anzuwenden, Ausdruck von allgemeiner Intelligenz ist. Hierzu wurde ein Antwortausschlusstest (AAT) konzipiert, bei dem die Testpersonen instruiert wurden, die Items durch die Antwortausschlussstrategien zu lösen, die sie im Training zuvor gelernt hatten. In diesem AAT wurden ausschließlich Antwortoptionen einer Matrizenaufgabe vorgegeben und die Lösung konnte durch die korrekte Anwendung der zuvor gelernten Strategie des Auszählens und Schnittmengenbildens gefunden werden. Die konvergente Validität des Antwortausschlusstests wurde dann ebenfalls bestimmt, indem die Ergebnisse mit den Ergebnissen des allgemeinen Intelligenztests korreliert wurden.

3.3.4 Materialien und Methodik

3.3.4.1 Stichprobe

Die Stichprobe umfasste 64 Testpersonen [*M(Alter)* = 22.28; *SD(Alter)* = 3.43; 75 % weiblich], von denen 92 % Studierende der Universität des Saarlandes waren.

3.3.4.2 Materialien

Leistungsprüfsystem-2K

Das LPS-2K (Kreuzpointner, 2013) ist eine Kurzversion des LPS-2 (Leistungsprüfsystem 2; Kreuzpointner, Lukesch & Horn, 2013), das die allgemeine Intelligenz *g* auf Basis von Carrolls (1993) Three-Stratum-Modell der Intelligenz misst. Das LPS-2 erfasst elf Faktoren erster Ordnung (elf Subtests), die sich auf vier Faktoren zweiter Ordnung reduzieren lassen [Kristalline Intelligenz (*Gc*), fluide

Intelligenz (*Gf*), visuelle Wahrnehmung (*Gv*), kognitive Schnelligkeit (*Gs*)]. Die elf Subtests lassen sich zu einem IQ-Wert aggregieren. Das LPS-2 besitzt eine hohe interne Konsistenz ($.86 \leq \alpha \leq .94$), durch konfirmatorische Faktoranalyse überprüfte faktorielle Validität und konvergente Validität hinsichtlich anderer Testverfahren. Das LPS-2K beinhaltet die vier Subtests Allgemeinwissen (*Gc*), Zahlenfolgen (*Gf*), Mentale Rotation (*Gv*) und Addieren (*Gs*). Die Subtests erlauben ebenfalls die Berechnung eines Aggregatwerts, der indikativ für *g* ist.

BOMAT-advanced

Der Bochumer Matrizentest BOMAT-advanced (Hossiep et al., 1999) misst allgemeine Intelligenz im hohen Fähigkeitsspektrum, d. h. komplexes logisches Schlussfolgern. Die Items bestehen aus einer 5-x-3-Matrix mit einem leeren Quadrat. Das leere Quadrat muss gefüllt werden, indem die logischen Regeln, die in der Matrix angewendet werden, verstanden werden. Die richtige Antwort muss aus sechs Antwortalternativen ausgewählt werden. Es gibt zehn Übungsitems, in denen die im Test verwendeten logischen Regeln erklärt werden, um zu gewährleisten, dass alle Testpersonen dieselbe Testvoraussetzung haben. Der BOMAT-advanced besitzt eine gute interne Konsistenz ($\alpha = .90$) und zeigt eine gute konvergente Validität zum Zahlenverbindungstest (ZVT; Oswald & Roth, 1978) sowie Kriteriumsvalidität zu Schulnoten. In dieser Studie wurden 20 Items aus dem BOMAT-advanced ausgewählt. Für das Prä-Post-Design wurden zwei Itemstämme mit je zehn Items konstruiert, die hinsichtlich ihrer Itemschwierigkeiten und Trennschärfen untereinander vergleichbar waren und für den Gesamttest möglichst repräsentativ sein sollten. Die Testpersonen wurden den beiden Gruppen zufällig zugeteilt und die Reihenfolge der beiden Itemsets gekreuzt. Die Items des Prä-Tests von Gruppe A waren die Items des Post-Tests von Gruppe B und umgekehrt, um Reihenfolgeeffekte auszuschließen. Bei den Items des Post-BOMATs sollten die Testpersonen außerdem jeweils angeben, ob sie die AAS bei diesem Item angewendet haben oder nicht (siehe Abbildung 11).

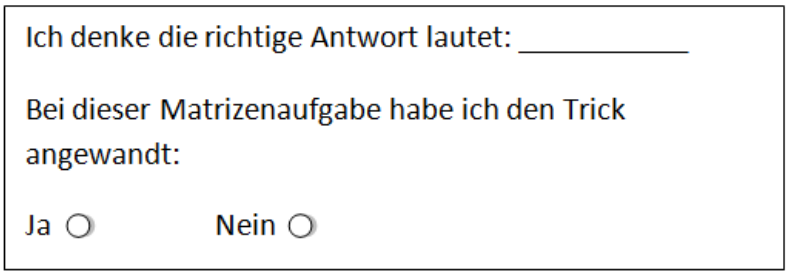


Abbildung 11: Antwortformat des Post-BOMATs

3.3.4.3 Training

Die von Mittring und Rost (2008) beschriebene Antwortausschlussstrategie wurde den Testpersonen in einer detaillierten schriftlichen Instruktion erklärt (siehe Anhang 6.1). Auf der Basis von zwei Übungsitems, bei denen die sechs Antwortoptionen ohne Itemstamm abgebildet waren, wurde Schritt für Schritt erklärt, wie das Item mit dem ‚Trick' der AAS gelöst werden konnte. Die Testpersonen konnten während des Trainings Fragen stellen. Die TestleiterInnen hatten sich zuvor intensiv mit der AAS auseinander gesetzt.

3.3.4.4 Antwortausschlusstest (AAT)

Der Antwortausschlusstest (AAT) bestand aus zehn figuralen Matrizenaufgaben, bei denen jeweils nur das Antwortformat, also die sechs Antwortoptionen, vorgegeben wurde (siehe Abbildung 12). Die Testpersonen sollten die richtige Antwort mithilfe der AAS herausfinden.

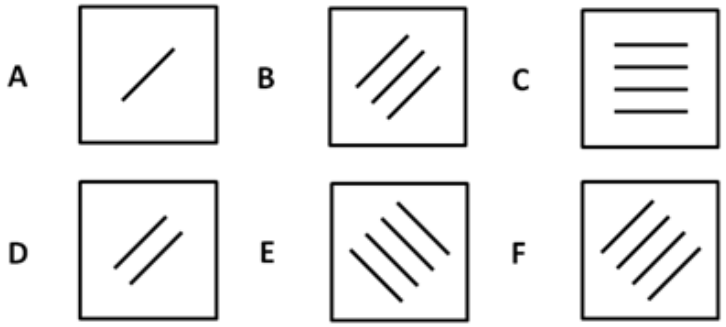


Abbildung 12: Antwortausschlusstest-Item

3.3.4.5 Testablauf

Die Testung fand in Kleingruppen von bis zu fünf Testpersonen an der Universität des Saarlandes statt. Vor der Testung füllten die Testpersonen einen demografischen Fragebogen aus. Danach starteten sie mit dem LPS-2K, der ca. 30 Min dauerte. Anschließend bearbeiteten sie zehn Items des BOMAT-advanced mit einer Zeitbegrenzung von 20 Min. Nach einer Pause von fünf Min. durchliefen die Testpersonen das Training, in dem ihnen die Antwortausschlussstrategie detailliert erklärt wurde. Danach bearbeiteten sie mit einer Zeitbegrenzung von 20 Min. den Antwortausschlusstest. Abschließend bearbeiteten sie wieder zehn Items des BOMAT-advanced mit einer Zeitbegrenzung von 20 Min. Die Testpersonen waren in zwei Gruppen eingeteilt, bei denen die Reihenfolge der BOMAT-Itemsets umgekehrt war. Die Testung dauerte insgesamt 120–150 Min. Der Testablauf ist in Tabelle 5 dargestellt.

Tabelle 5: Ablauf der Testung in Studie 3

Gruppe 1	Gruppe 2
LPS-2K	LPS-2K
BOMAT-Set A	BOMAT-Set B
Training	Training
BOMAT-Set B	BOMAT-Set A

3.3.4.6 Statistische Methodik

Da die beiden Prä- und Post-BOMAT-Sets in den beiden Gruppen hinsichtlich ihrer Itemschwierigkeiten und internen Konsistenzen vergleichbar waren, wurden die Ergebnisse von Gruppe 1 und 2 zum Zwecke der statistischen Auswertung aggregiert. Um zu überprüfen, ob sich die Verfahren hinsichtlich ihrer internen Konsistenz unterschieden, wurde Cronbachs α für den LPS-2K und die beiden BOMAT-Versionen berechnet. Ein t-Test für abhängige Stichproben sollte Aufschluss darüber geben, ob es eine Verbesserung in der BOMAT-Testleistung vor und nach dem Training gab. Um zu untersuchen, ob sich die konvergente Validität des BOMATs nach dem Training im Vergleich zu vorher geändert hat, wurde die Korrelation zwischen dem jeweiligen BOMAT-Summenwert und dem LPS-2K-Summenwert berechnet. Unterschiede in den Korrelationen wurden mit dem Signifikanztest von Millsap et al. (1990) untersucht. Die konvergente Validität des AATs wurde berechnet, indem die Korrelation zwischen dem AAT-Summenwert und dem LPS-2K-Summenwert gebildet wurde. Um zu untersuchen, ob die Häufigkeit, mit der die AAS im Post-BOMAT angewendet wird, mit der allgemeinen Intelligenz korreliert, wurde die Korrelation zwischen der Häufigkeit, mit der die Testpersonen die AAS angewendet hatten, und den Ergebnissen des LPS-2K ermittelt. Um zu prüfen, ob die Häufigkeit, mit der die AAS angewendet wird, mit der Leistung im BOMAT korreliert, wurde die Korrelation zwischen den Post-

BOMAT-Ergebnissen und der angegebenen Häufigkeit der AAS-Anwendung berrechnet.

3.3.5 Ergebnisse

3.3.5.1 Interne Konsistenzen

Der LPS-2K zeigte eine hohe interne Konsistenz (α = .96). Der Prä-BOMAT zeigte eine interne Konsistenz von α = .29, der Post-BOMAT eine interne Konsistenz von α = .44. Die interne Konsistenz des AATs betrug α = .73.

3.3.5.2 Leistungsunterschiede in den BOMAT-Versionen

Die durchschnittliche Anzahl korrekt gelöster Items in der Prä-BOMAT-Version betrug M = 5.94 (SD = 1.6), in der Post-BOMAT-Version M = 5.56 (SD = 1.80). Der t-Test für abhängige Stichproben zeigte keinen signifikanten Unterschied zwischen den beiden Mittelwerten [$t(63) = 1.30$, $p = .12$, $d = .26$].

3.3.5.3 Konvergente Validitäten

Die Korrelation zwischen dem Prä-BOMAT und dem LPS-2K betrug $r = .26$ ($p < .05$), die Korrelation zwischen Post-BOMAT und LPS-2K $r = .42$ ($p < .01$). Der Signifikanztest ergab keinen signifikanten Unterschied zwischen den beiden Korrelationen ($z = 1$, $p = .16$). Die Korrelation zwischen LPS-2K und dem AAT betrug $r = .50$ ($p < .01$).

3.3.5.4 Korrelationen der AAS-Anwendungshäufigkeit mit den Leistungen im BOMAT und LPS-2K

Die AAS wurde im Schnitt bei der Hälfte der Items des Post-BOMATs angewandt (M = 4.83, SD = 1.88). Zwischen der Häufigkeit der AAS-Anwendung und der Leistung im BOMAT ergab sich ein negativer Zusammenhang von r = −.30 (p < .05). Zwischen der Häufigkeit, mit der die AAS im Post-BOMAT angewendet wurde, und den Ergebnissen des LPS-2K ergab sich keine signifikante Korrelation (r = −.06, p = .63).

3.3.6 Diskussion

Ziel von Studie 3 war es, zu untersuchen, ob sich die konvergente Validität eines figuralen Matrizentests verringert, wenn Testpersonen Kenntnis von der von Mittring und Rost (2008) beschriebenen Antwortausschlussstrategie haben, die den Itemstamm außer Acht lässt. Des Weiteren sollte überprüft werden, ob sich die Testleistung mit Kenntnis der AAS verbessert und inwiefern die Häufigkeit, mit der die AAS angewendet wird, mit der BOMAT-Leistung und der allgemeinen Intelligenz korreliert. Ebenso sollte untersucht werden, inwiefern die Fähigkeit, diese AAS anzuwenden, ein Ausdruck von allgemeiner Intelligenz ist.

Die internen Konsistenzen des Prä- und Post-BOMATs waren klein und nicht zufriedenstellend. Dies ist vermutlich dem Umstand geschuldet, dass die Items in ihrer Schwierigkeit bewusst heterogen ausgewählt wurden, um den gesamten BOMAT möglichst gut zu repräsentieren. In diesem Fall ist die interne Konsistenz eventuell nicht das adäquate Maß für die Reliabilität des Tests. Eine Testverlängerung zur Steigerung der Reliabilität ist im Hinblick auf die Testökonomie hier nicht sinnvoll. Zukünftige Studien könnten die Retestreliabilität der Tests überprüfen. Die interne Konsistenz des AATs war auf einem guten Niveau.

Die Veränderung in der Leistung im BOMAT nach dem Training der AAS war im Gegensatz zur Studie von White und Zammarelli (1981) nicht signifikant. Wir

schlussfolgern deshalb, dass die Kenntnis der von Mittring und Rost (2008) beschriebenen AAS die Leistung in einem figuralen Matrizentest nicht verbessert.

Die Korrelationen zwischen den beiden BOMATs und dem LPS-2K waren gering Dies kann zum Teil an den niedrigen internen Konsistenzen der BOMATs liegen. Werden die minderungskorrigierten Korrelationen berechnet, ergibt sich für den Prä-BOMAT eine Korrelation von r = .48 zum LPS-2K und für den Post-BOMAT eine Korrelation von r = .63. Diese sind im Vergleich zu den ursprünglichen Werten deutlich höher und bestätigen die Vermutung, dass die internen Konsistenzen (mit-)verantwortlich für die niedrigen konvergenten Validitäten der beiden BOMATs sind. Der Befund, dass die Korrelationen sich nicht signifikant voneinander unterscheiden, bleibt bestehen (z = 1.21, p = .11).

Die konvergente Validität des Post-BOMATs unterschied sich nicht signifikant von der des Prä-BOMATs. Wir nehmen deshalb an, dass die Anwendung von AAS die Konstruktvalidität eines figuralen Matrizentests nicht negativ beeinflusst. Laut Mittring und Rost (2008) sollte sich die konvergente Validität nach dem Training der AAS verschlechtern. In unserer Studie ist das gegensätzliche Muster zu beobachten, der numerische Wert der Korrelation des BOMATs nach dem Training mit dem LPS-2K ist höher als vor dem Training.

Zwischen dem AAT und dem LPS-2K ergab sich eine substantielle Korrelation. Es stellt sich also heraus, dass der AAT eine Messung von allgemeiner Intelligenz darstellt und die Möglichkeit, Antwortausschlussstrategien zu verwenden, die Konstruktvalidität nicht einschränkt. Stattdessen ist die Fähigkeit, eine solche AAS anzuwenden, eng verwandt mit der Fähigkeit, die notwendig ist, um einen allgemeinen Intelligenztest zu lösen. Da es in unserer Stichprobe vermutlich eine Varianzeinschränkung gibt, könnte die Korrelation in der Gesamtpopulation sogar noch höher ausfallen.

Die AAS wurde ungefähr bei der Hälfte der Items angewendet. Ein Zusammenhang zwischen der Häufigkeit der Anwendung und der allgemeinen Intelligenz konnte nicht gefunden werden. Es zeigte sich jedoch, dass mit

vermehrter Anwendung der AAS die Testleistung abnimmt. Das bedeutet, je häufiger die AAS angewandt wird, desto schlechter schneidet die Testperson im BOMAT ab. Die AAS kann also nicht als Fallback-Strategie angesehen werden, mit der weniger begabte Testpersonen ihre Leistung verbessern können, sodass die Konstruktvalidität gemindert wird.

3.3.7 Zusammenfassung

Die Ergebnisse der Studie 3 legen nahe, dass die mögliche Anwendung den Itemstamm außer Acht lassender Antwortausschlussstrategien, wie sie von Mittring und Rost (2008) beschrieben werden, die Konstruktvalidität figuraler Matrizenaufgaben nicht einschränkt. Die Testleistung verbessert sich mit Kenntnis und Häufigkeit der Anwendung der AAS nicht und die Fähigkeit, eine AAS anzuwenden, kann als Ausdruck allgemeiner Intelligenz angesehen werden. Die Ergebnisse weisen darauf hin, dass der Einsatz von Distraktoren in dieser Hinsicht der Konstruktvalidität des Tests nicht abträglich ist.

Abschließend soll hier noch angemerkt werden, dass die Wahrscheinlichkeit, mit der eine AAS, die den Itemstamm völlig außer Acht lässt, angewandt wird, als gering einzustufen ist. Die Strategie ist, wenn auch nicht als kompliziert, so doch vom Umfang her als aufwändig zu betrachten und ohne ein entsprechendes Training werden die wenigsten Testpersonen auf die Idee kommen, eine solche AAS anzuwenden. Die empirischen Ergebnisse unserer Studie weisen zudem darauf hin, dass im Falle der Anwendung der AAS die von Mittring und Rost (2008) sowie White und Zammarelli (1981) vorgebrachten kritischen Überlegungen vernachlässigbar sind.

4 Generelle Diskussion

4.1 Zusammenfassung der Ergebnisse

In der vorliegenden Arbeit sollten die Auswirkungen des Einsatzes bzw. des Verzichts von Distraktoren im Antwortformat von figuralen Intelligenztestaufgaben zur Erfassung von räumlichem Denken und fluider Intelligenz untersucht werden. Verschiedene Studien aus der Forschung zu figuralen Matrizenaufgaben (Arendasy & Sommer, 2013; Bethell-Fox et al., 1984; Vigneau et al., 2006) weisen darauf hin, dass der Einsatz von Distraktoren verschiedene Probleme mit sich bringt. Zusätzlich zur Ratemöglichkeit kann eine Testperson zwei verschiedene Strategien anwenden, um eine Aufgabe zu lösen: zum einen das gewünschte Constructive Matching, bei dem die Lösung selbst konstruiert bzw. vollständig überprüft wird, zum anderen aber auch eine Response-Elimination-Strategie, bei der einzelne Teile der Antwortoptionen mit dem Itemstamm verglichen werden, um falsche Antworten auszuschließen. Diese Strategie ermöglicht es Testpersonen, sich besser darzustellen als es ihrer eigentlichen Fähigkeit entspricht. Zudem zeigt sich, dass es einen Mangel an Strategien zur Schwierigkeitsmanipulation der Items gibt. Die Validität des Tests sowie die Möglichkeit zur Diagnostik von Hochbegabung im entsprechenden Fähigkeitsbereich werden dadurch eingeschränkt. Es ist anzunehmen, dass sich eben diese Probleme auch bei Aufgaben zum räumlichen Denken zeigen. Schwerpunkt dieser Arbeit war deshalb die Konstruktion und Validierung eines distraktorfreien Aufgabenformats zur Erfassung räumlichen Denkens (Studie 1 und 2). Mittring und Rost (2008) sowie White und Zammarelli (1981) weisen zudem auf eine dritte mögliche Strategie zur Lösung figuraler Matrizenaufgaben hin, die noch nicht nähergehend empirisch erforscht wurde: Bei reiner Inspektion des Antwortformats könne mithilfe einer Antwortausschlussstrategie durch Auszählen und Schnittmengenbildung die

richtige Lösung gefunden werden, ohne dass der Itemstamm überhaupt betrachtet werden muss. Die Relevanz dieser Problematik sollte in Studie 3 empirisch untersucht werden.

Aufgrund der bekannten Probleme, die der Einsatz von Distraktoren mit sich bringt, wurde in Studie 1 eine Würfelrotationsaufgabe ohne Distraktoren entwickelt, die Würfelkonstruktionsaufgabe. Ziel der ersten Studie war es, dieses neue Aufgabenformat auf seine psychometrischen Eigenschaften hin zu untersuchen. Es zeigte sich, dass die Würfelkonstruktionsaufgabe als Skala sehr reliabel ist, eine eindimensionale Faktorstruktur aufweist und in der Anfangsversion Items sehr hoher Schwierigkeit besitzt. Dies ist wünschenswert, da es in gängigen Würfelrotationstests an schweren Items mangelt, wodurch die Validität gängiger Würfelrotationsaufgaben beeinträchtigt und die Diagnostik von Hochbegabung deutlich erschwert wird.

Für die differenzierte Erfassung des gesamten Fähigkeitsspektrums ist jedoch ein breiterer Schwierigkeitsrange der Items nötig. Zu diesem Zweck wurde in Studie 2 der Versuch unternommen, die Items der Würfelkonstruktionsaufgabe in ihrer Schwierigkeit durch die Vorgabe von Teillösungen abzustufen. Zudem sollte die konvergente Validität der Aufgabe überprüft werden. Um auszuschließen, dass Unterschiede in der Validität und Schwierigkeit im Vergleich zu anderen existierenden Würfelrotationsaufgaben auf die Beschaffenheit der Items der Würfelkonstruktionsaufgabe zurückzuführen sind, wurde eine Version der Würfelkonstruktionsaufgabe mit Distraktoren entwickelt, die mit gängigen auf dem Markt existierenden Formaten vergleichbar sein sollte, deren Itemstämme jedoch identisch mit denen der distraktorfreien Würfelkonstruktionsaufgabe waren. Es zeigte sich, dass es im Format ohne Distraktoren durch die Vorgabe von Teillösungen möglich ist, Items jeder Schwierigkeit zu erstellen, während die Itemschwierigkeit im Format mit Distraktoren überwiegend moderat war. Des Weiteren konnte gezeigt werden, dass die konvergente Validität des Aufgabenformats ohne Distraktoren zu einem allgemeinen Intelligenztest signifikant höher war als die des Formats mit Distraktoren, bei dem sich keine

signifikante Korrelation zum allgemeinen Intelligenztest ergab. Dieses Ergebnis erklärt sich unserer Auffassung nach zum einen dadurch, dass es ohne Distraktoren besser möglich ist, Items variierender Schwierigkeit zu konstruieren und somit ein breiterer Differenzierungsgrad erreicht werden kann, der mit dem des allgemeinen Intelligenztests mehr übereinstimmt als im Format ohne Distraktoren. Zum anderen erklärt es sich aber auch dadurch, dass hier Response-Elimination-Strategien verhindert werden, sodass die Items auch dadurch schwieriger werden und eine verzerrte Leistungsdarstellung nicht mehr möglich ist, der Test also valider wird.

Die dritte Studie befasste sich mit der Problematik, dass figurale Matrizenaufgaben gelöst werden können, indem eine Antwortausschlussstrategie angewendet wird, die keine Inspektion des Itemstamms erfordert. Da erste Hinweise auf diese Problematik im Zusammenhang mit figuralen Matrizenaufgaben in der Fachliteratur bereits gegeben wurden, eingehende empirische Studien jedoch fehlen, knüpfte Studie 3 hier an und befasste sich mit figuralen Matrizenaufgaben. Es konnte gezeigt werden, dass die mögliche Anwendung von Antwortausschlussstrategien bei reiner Inspektion des Antwortformats keine Verringerung der konvergenten Validität des Tests zur Folge hat und sich die Leistung der Testpersonen auch nicht verbessert. Des Weiteren konnte gezeigt werden, dass die Fähigkeit, eine solche Antwortausschlussstrategie anzuwenden, mit allgemeiner Intelligenz korreliert ist. Die Möglichkeit der Anwendung einer solchen Antwortausschlussstrategie ist also nach unseren Ergebnissen als wenig problematisch einzustufen.

4.2 Einordnung der Ergebnisse und Ausblick

4.2.1 Vorteile der Würfelkonstruktionsaufgabe und damit verbundene Forschungsmöglichkeiten

Wird die Würfelkonstruktionsaufgabe mit bestehenden Aufgaben im Feld verglichen, lässt sich sagen, dass sie einige Vorteile gegenüber existierenden Formaten mitbringt. Wie bereits erwähnt, lässt sich *1) die Schwierigkeit der Items sukzessive abstufen*, was eine Differenzierung in verschiedenen Fähigkeitsbereichen, u.a. auch dem Hochbegabungsbereich, ermöglicht. Dabei ist die Abstufung der Itemschwierigkeiten nicht mit einem Geschwindigkeitsfaktor konfundiert. Bei bestehenden Würfeltests zeigt sich das folgende Bild: Die Itemschwierigkeiten werden entweder über eine Begrenzung der Bearbeitungszeit für den gesamten Test abgestuft, indem die letzten Items dadurch schwieriger werden, dass mit abnehmender Bearbeitungszeit auch die Lösungswahrscheinlichkeit bzw. die Wahrscheinlichkeit der Inangriffnahme sinkt, oder aber die Itemschwierigkeiten sind unter einer Power-Bedingung relativ homogen und moderat. Im ersten Fall ist anzunehmen, dass die zu messende Fähigkeit mit einer kognitiven Speed-Komponente konfundiert ist (Wilhelm & Schulze, 2002) und die Validität somit beeinträchtigt wird. Im zweiten Fall ist anzunehmen, dass das Konstrukt in reinerer Form gemessen wird, es ergibt sich so jedoch eine Varianzeinschränkung bei den Itemschwierigkeiten, die sich ebenfalls negativ auf die Validität des Tests auswirkt, da an den Rändern der Verteilung nicht differenziert genug gemessen wird. Durch die Vorgabe der Würfelkonstruktionsaufgabe kann nun in einer Power-Bedingung mit sehr liberalem Zeitlimit pro Item eine größere Varianz der Itemschwierigkeiten hergestellt und somit ein Ausweg aus dem Speed-Power-Dilemma gefunden werden. Die Konfundierung mit einer kognitiven Speed-Komponente ist eine weitere denkbare Erklärung, warum manche Tests zum räumlichen Denken eine weniger hohe *g*-Sättigung aufweisen als andere, besonders wenn der zum

Vergleich herangezogene Intelligenztest ein Power-Test ist. In einer Folgestudie könnten beide Versionen der Würfelkonstruktionsaufgabe in einer Power- und einer Speed-Bedingung durchgeführt und mit einem Mental Speed-Test und einem allgemeinen Intelligenztest korreliert werden, um den Effekt, den die Zeitbegrenzung auf die Validität des Tests hat, genauer zu beleuchten.

Des Weiteren ergibt sich *2) eine hohe konvergente Validität* zum allgemeinen Intelligenztest. Inwiefern dies mit der Anwendung verschiedener Lösungsstrategien und/oder der fehlenden Schwierigkeitsabstufung der Items des distraktorbasierten Formats zusammenhängt, bleibt genauer zu untersuchen. Hierbei bietet *3) die computerbasierte Testform* mehrere Vorteile gegenüber anderen, nicht computerisierten Verfahren. Ein Vorteil der computerbasierten Testung ist ihre *ökonomische Durchführ- und Auswertbarkeit*. Es ist möglich, die Aufgaben ortsungebunden online durchzuführen, was die weitere Forschung zum Beispiel im Hinblick auf die Rekrutierung von StudienteilnehmerInnen erleichtert und auch in der Praxis, z. B. bei der Auswahl geeigneter KandidatInnen für relevante Studiengänge und Berufszweige, einen komfortablen Einsatz des Tests erlaubt. Bei der Untersuchung der möglicherweise angewandten verschiedenen Lösungsstrategien bietet sie den Vorteil, dass sie eine *Blickbewegungsanalyse* (Hayes, Petrov & Sederberg, 2011; Vigneau et al., 2006) der Testpersonen erlaubt, die Unterschiede zwischen Testpersonen bei der Bearbeitung der Aufgaben zeigen kann. Ebenso ist in der computerbasierten Testung der Vorteil zu sehen, dass *Reaktionszeiten* erfasst werden können, aus denen Rückschlüsse über die angewandten Strategien gezogen werden können. In dieser Arbeit war der Fokus der Fragestellung ein anderer, die Untersuchung von Reaktionszeiten ist jedoch Gegenstand zahlreicher Studien, die sich mit kognitiven Leistungstests unter anderem auch zum räumlichen Denken befassen (Goldhammer & Klein Entink, 2011; Klein Entink, Fox & van der Linden, 2009; Lohman, 1986; Scherer, Greiff & Hautamäki, 2015). Es gibt divergierende Befunde hinsichtlich der Fragestellung, ob eine längere Reaktionszeit mit einer größeren Wahrscheinlichkeit einhergeht, die Aufgabe richtig zu lösen, oder ob umgekehrt eine kurze Reaktionszeit mit einer

hohen Lösungswahrscheinlichkeit korreliert ist. Die Beziehung zwischen Reaktionszeit und Lösungswahrscheinlichkeit wird von verschiedenen Parametern wie Itemcharakteristiken (leicht vs. schwer), Personencharakteristiken (fähig vs. unfähig), dem zu messenden Konstrukt und den involvierten Prozessen beeinflusst (Goldhammer, Naumann & Greiff, 2015; Goldhammer et al., 2014). Die Würfelkonstruktionsaufgabe erlaubt die weitere Untersuchung solcher Fragestellungen im Kontext räumlicher Denkaufgaben. Sinnvoll wäre die Kombination von Daten aus Studien zum lauten Denken, Blickbewegungsstudien, Reaktionszeiten und Fehlermustern, da diese in Kombination am besten Rückschlüsse auf die angewendeten Lösungsstrategien und, bei zusätzlicher Erfassung anderer konstruktnaher Tests, Rückschlüsse auf den Einfluss der Lösungsstrategien auf die Konstruktvalidität zulassen (Glück & Fitting, 2003; Lohman & Kyllonen, 1983).

An dieser Stelle sei noch auf einen weiteren Vorteil der Würfelkonstruktionsaufgabe hingewiesen, nämlich, *4)* dass ein *Vergleichsprozess nicht mehr zwingend Teil des Lösungsprozesses* ist. Bereits existierende Würfelrotations- oder Faltaufgaben sind immer so aufgebaut, dass die Distraktoren zwingend Teil der Aufgabenstellung sind, d. h., dass ein Vergleichsprozess immer Teil des Lösungsprozesses ist, selbst wenn die Anwendung von Response-Elimination-Strategien als Kompensation mangelnder Fähigkeit nicht intendiert ist. Die Aufgaben bestehen nicht wie bei Matrizenaufgaben aus einem Itemstamm, der die Aufgabenstellung enthält. Bei Matrizenaufgaben ist es möglich, die Aufgabe ausschließlich durch Betrachtung des Itemstamms zu lösen und die gefundene Lösung dann mit den Antwortoptionen abzugleichen. Bei Würfelaufgaben gibt es keine eigenständige Lösung, wenn nur der Itemstamm betrachtet wird. Erst durch Betrachten der Antwortoptionen fängt die Testperson an, den Würfel zu rotieren. Es wird nicht die fertige Rotation mental auf das Vergleichsbild gelegt, sondern die Rotation erfolgt immer im Abgleich mit der Antwortoption. Es kann die Frage gestellt werden, ob dieser Prozess nicht immer zwingend rückwärtsgerichtet und weniger kognitiv aufwändig ist bzw. stärker zu paarweisen Vergleichen der Seiten

einlädt, als es zum Beispiel bei Matrizenaufgaben der Fall ist. Insofern stellt sich das Problem der Distraktorvorgabe hier vielleicht sogar verschärft dar. Einzelne Aufgaben, die ein ähnliches Antwortformat wie das der Würfelkonstruktionsaufgabe verfolgten, waren zur Zeit ihrer Erstellung noch nicht computerbasiert und deshalb möglicherweise nicht reliabel auswertbar. Andere Aufgaben zeigten nur einen Referenzwürfel, sodass es für die Testperson nicht möglich war, zu wissen, welche Symbole sich auf den verdeckten Würfelseiten befanden (Eliot & Mcfarlane Smith, 1983). Somit war die Konstruktion von Aufgaben ohne Distraktoren schwer möglich. Durch die Vorgabe von drei Referenzwürfeln kann die Testperson jede Seite des Würfels sehen und somit den kompletten Würfel rekonstruieren. Ein Vergleichsprozess ist nicht mehr zwingend Teil des Lösungsprozesses und die Konstruktion der Lösung verlangt mehr räumliche Fähigkeit in dem Sinne, dass hier ‚pures' Constructive Matching angewendet werden muss. Ein Vergleichsprozess kann jedoch bei der Vorgabe von Teillösungen auch nicht komplett ausgeschlossen werden. Dies spricht dafür, die Items noch auf anderem Wege in ihrer Schwierigkeit zu manipulieren als durch die Vorgabe von Teillösungen. Auf diesen Gedanken wird unter Punkt 4.2.2 noch einmal genauer eingegangen.

Dass bei der Würfelkonstruktionsaufgabe die Lösung generiert, statt durch Vergleich ausgesucht werden muss, hat *5)* noch einen anderen Vorteil, nämlich den der Möglichkeit eines *effektiveren Trainings*. Mehr als bei anderen kognitiven Fähigkeiten steht bei räumlichem Denkvermögen die Frage im Fokus, ob dieses trainierbar ist oder nicht. Eine aktuelle Meta-Analyse, die die Ergebnisse aus 217 Studien integriert, kommt zu dem Schluss: „Spatial skills are highly malleable and training in spatial thinking is effective, durable, and transferable" (Uttal et al., 2013, S. 365). Welche Art von Training den nachhaltigsten Effekt hat, muss noch genauer untersucht werden, ein falsches Training gibt es nach den Ergebnissen von Uttal et al. (2013) jedoch nicht. Es zeigte sich bei verschiedenen Arten von Training ein Trainingseffekt, der über den durchschnittlichen Retesteffekt hinausging und der sich eventuell noch festigen würde, wenn das Training entsprechend lange

andauern würde (die meisten Studien setzen nur ein Training von wenigen Stunden bis Wochen ein). Bei Studien mit entsprechend ausführlichem Training zeigte sich auch ein Transfer, was besonders bemerkenswert und unabdingbar ist, wenn von effektivem Training gesprochen werden soll. Uttal et al. (2013) weisen darauf hin, dass grundsätzlich jede Person unabhängig von ihrem Geschlecht oder ihrem Fähigkeitsniveau von einem Training profitieren kann. Die Förderung solcher Fähigkeiten ist in der Kindheit am effektivsten, weshalb es bereits verschiedene Überlegungen gibt, wie räumliches Denken möglichst grundlegend und langfristig im Unterricht gefördert werden kann (Newcombe, 2010).

Es gibt Hinweise darauf, dass ein Training zum räumlichen Denken, das eine aktive Konstruktion der Lösung erfordert, hinsichtlich seiner Effektivität einem Training, bei dem lediglich die richtige Lösung ausgesucht werden muss, überlegen ist (Sorby, 2009). Auch andere Studien zeigen, dass der Akt des Generierens im Gegensatz zum Akt des Aussuchens oder zu passiver Instruktion wünschenswerte Schwierigkeiten mit sich bringt, die den Lernerfolg in der jeweiligen Fähigkeit steigern können (Bjork, 1999; Kornell, Hays & Bjork, 2009; Linn, Chang, Chiu, Zhang & McElhaney, 2010). Somit kann die Würfelkonstruktionsaufgabe die Möglichkeit bieten, räumliches Denken auf besonders effektive Weise zu trainieren.

Das effektive Training räumlichen Vorstellungsvermögens ist möglicherweise für Frauen von besonderer Bedeutung, da diese in Aufgaben zum räumlichen Denken häufig schlechter abschneiden als Männer (Linn & Petersen, 1985; Voyer et al., 1995). Während die Unterschiede zwischen Männern und Frauen für bestimmte Aufgabentypen in den letzten 30 Jahren zurückgegangen sind oder sich sogar aufgehoben haben (Voyer et al., 1995), bestehen sie in bestimmten Bereichen wie der mentalen Rotation weiterhin, hier ist der Effekt am größten (Peters et al., 2007). Für die beobachteten Geschlechtsunterschiede gibt es eine Vielzahl von Erklärungsansätzen. Sie reichen von biologischen/evolutionspsychologischen bis zu kulturellen/sozialen Ansätzen (Mohler, 2008). Ein schlechtes Selbstkonzept, ‚Stereotype-Threat' und Testangst werden als Gründe für die schlechtere Testleistung von Frauen diskutiert (Shih, Pittinsky & Ambady, 1999). Ebenso gibt es

Hinweise darauf, dass die Art des Testmaterials einen Einfluss auf die Leistung hat und sich Geschlechtsdifferenzen vor allem bei ‚technischem' Material wie Polygonen zeigen (Jansen-Osmann & Heil, 2007). Von einer sozialisationsbedingten Perspektive aus betrachtet, könnte es sein, dass Frauen durch ihre Erziehung weniger häufig mit räumlichem Material in Berührung kommen und somit weniger Erfahrung mit räumlichem Denken haben und geringeres Interesse an Aktivitäten, die räumliches Denken erfordern (Baenninger & Newcombe, 1995). Es mangelt ihnen möglicherweise an der Entwicklung bzw. Ausbildung ihres Potentials zum räumlichen Denken. In diesem Fall könnten sie von einem Training besonders profitieren.

Die Frage, wie räumliches Denken möglichst effektiv trainiert werden kann, hat für Frauen unter anderem deshalb besondere Relevanz, da diese in MINT-Fächern traditionell unterrepräsentiert sind, ein Umstand, den zu ändern nicht zuletzt Gegenstand verschiedener politischer Bemühungen ist (Quaiser-Pohl, 2012). Für den niedrigen Frauenanteil in MINT-Fächern können verschiedene Faktoren wie z. B. die Vereinbarkeit von Familie und Beruf oder sozialisationsbedingte Rollenstereotype (Ceci, Williams & Barnett, 2009), aber eben auch die schlechtere räumliche Denkleistung von Frauen im Vergleich zu Männern verantwortlich sein. Falls Letzteres zutrifft, könnte eine durch gezieltes, langfristiges Training entstandene Verbesserung der räumlichen Denkleistung die Affinität von Frauen für solche Fächer und ihren Erfolg in diesen steigern.

Darüber hinaus ist die Abbrecherquote von Studierenden der MINT-Fächer im Vergleich zu anderen Fächern besonders hoch und liegt derzeit in Deutschland ähnlich wie in den USA bei 40 % (Heublein, Richter, Schmelzer & Sommer, 2014; Price, 2010). Besonders in der Anfangsphase solcher Studiengänge sind räumliche Fähigkeiten besonders von Bedeutung, während später semantische Fähigkeiten in den Vordergrund rücken (Hambrick et al., 2011). Im Fach Chemie ist es beispielsweise von Bedeutung, sich Moleküle gut räumlich vorstellen zu können (Harle & Towns, 2011). Somit könnte räumliches Denkvermögen als „gatekeeper" (Uttal et al., 2013, S. 369) für den weiteren Studienerfolg fungieren. Ein Training

räumlichen Vorstellungsvermögens könnte der hohen Abbrecherquote in MINT-Fächern entgegenwirken (Sorby, 2009).

Im Hinblick auf die genannten Punkte sollten die Trainingsmöglichkeiten räumlichen Denkvermögens und die Effizienz verschiedener Antwortformate in diesem Kontext genauer untersucht werden. In einer zukünftigen Studie könnte die Effektivität eines Trainings mit dem distraktorfreien Format der Würfelkonstruktionsaufgabe mit der Effektivität eines Trainings mit dem distraktorgestützten Format verglichen werden.

4.2.2 Einschränkungen und offene Forschungsfragen

Aus Studie 2 ergibt sich die Einschränkung, dass zunächst nur festgestellt werden konnte, dass das Antwortformat einen Einfluss auf die Validität des Tests hat, jedoch nicht geklärt werden konnte, warum sich der Unterschied in den Daten bezüglich der distraktorfreien und der distraktorbasierten Version der Würfelkonstruktionsaufgabe ergibt. Hier könnten sich die unterschiedlichen Lösungsstrategien genauso auswirken wie die unterschiedlichen Schwierigkeitsverteilungen. Die angewendeten Strategien lassen sich wie bereits erwähnt mit einem Paradigma zum lauten Denken, Reaktionszeitanalysen und Blickbewegungsanalysen untersuchen. Inwiefern die höhere konvergente Validität der distraktorfreien Testform mit der fehlenden Schwierigkeitsabstufung der Items des distraktorbasierten Formats zusammenhängt, lässt sich nur untersuchen, indem eine distraktorbasierte Form mit entsprechender Schwierigkeitsverteilung der Items erstellt wird, wobei sich die Frage stellt, auf welchem Wege dies zu erreichen ist. Neben der Manipulation der Distraktoren, wäre eine Möglichkeit, die Schwierigkeiten der Items noch auf anderem Wege zu manipulieren, ohne dass Teillösungen vorgegeben werden. Eventuell ist es möglich, durch eine systematische Variation der Symbole auf dem Referenzwürfel eine Schwierigkeitsabstufung herzustellen. Es könnte zum Beispiel einen Unterschied machen, ob sich ein Symbol zentral in der Mitte des Würfels befindet oder in einer

der Ecken. Auch könnte die Art des Symbols sich auf die Schwierigkeit auswirken. Bei unserer zufälligen Zusammenstellung der Symbole ergab sich kein systematischer Effekt auf die Schwierigkeit, bei genauerer systematischer Untersuchung des Einflusses der Symbolplatzierung könnte sich jedoch ein solcher finden lassen. Damit wäre die Schwierigkeitsmanipulation der Items unabhängig vom Antwortformat möglich und auch bei der Version mit Distraktoren einsetzbar. Eine andere Möglichkeit ist, einzelne Seiten auf dem Würfel leer zu lassen und somit die Anzahl der zueinander in Beziehung zu setzenden Seiten zu variieren (und damit vermutlich auch die Schwierigkeit der Items) oder die Seiten zusätzlich zu den Symbolen mit verschiedenen Farben zu versehen. Ebenso könnte die Anzahl der Überschneidungen der sichtbaren Seiten auf dem Referenzwürfel und auf dem Lösungsfeld variiert werden oder die Anzahl erforderlicher Rotationen, um auf die Lösung zu kommen.

Dies sind alles denkbare Ansätze, um eine Abstufung der Schwierigkeiten im distraktorbasierten Format zu ermöglichen und somit den Einfluss der Schwierigkeitsabstufung auf die konvergente Validität zu überprüfen. Zusätzlich ergibt sich in diesem Fall der Vorteil, dass im konstruktionsbasierten Format die Lösung ohne jegliche Hilfestellung selbst konstruiert werden muss und die Messung somit noch eine ‚reinere' Constructive-Matching-Messung wäre. Es müsste dann jedoch sichergestellt werden, dass die Aufgaben nicht wieder zu schwer werden (siehe Studie 1). Mit der Vorgabe von Teillösungen besteht eine gute Möglichkeit, die Schwierigkeiten abzustufen, weitere Ansätze sind trotzdem denkbar und vor allem für die Variation im distraktorbasierten Format sinnvoll.

Die Ergebnisse bezüglich der Leistungsunterschiede zwischen den Geschlechtern in der distraktorfreien Würfelkonstruktionsaufgabe in Studie 1 und 2 sind widersprüchlich. In Studie 2 konnten Unterschiede zugunsten der Männer festgestellt werden, in Studie 1 zeigte sich das umgekehrte Muster, die Frauen waren hier leistungsstärker als die Männer. Die Gruppe der Männer war in diesem Fall jedoch sehr klein. Da die Studien nicht primär darauf ausgelegt waren, Geschlechtsunterschiede zu untersuchen, wurden die Stichproben wie unter Punkt

3.2.6 bereits erwähnt dementsprechend nicht optimal ausgewählt. Die Ergebnisse bezüglich der Geschlechtsunterschiede in dieser Arbeit sind deshalb nicht gut interpretierbar. Sie sollten aber explorativ miterfasst werden, da Geschlechtsunterschiede einen zentralen Befund bei Studien zum räumlichen Denken darstellen, und können einen Ausgangspunkt für weitere Studien darstellen, in denen sie unter besseren Bedingungen untersucht werden können.

Für zukünftige Forschung relevant ist eine weitere Validierung der Würfelkonstruktionsaufgabe im Hinblick auf weitere Maße für ihre Konstruktvalidität und vor allem für ihre Kriteriumsvalidität. Räumliches Denken spielt vor allem in MINT-Fächern eine bedeutsame Rolle, die, wie z. B. die Informatik, in der heutigen digitalisierten Welt besondere Relevanz haben. Wenn die Würfelkonstruktionsaufgabe einen besseren Prädiktor für Studiums- oder Berufserfolg in solchen Fächern darstellt, dann ist dies ein weiteres Argument für ihre praktische Anwendung und ein weiterer bedeutsamer Hinweis darauf, dass das Antwortformat einen entscheidenden Einfluss auf die Validität eines kognitiven Fähigkeitstests hat. Es ist sinnvoll, zu untersuchen, für welche Tätigkeiten genau die Würfelkonstruktionsaufgabe eine relevante Vorhersagekraft hat. Ebenso kann untersucht werden, ob die Würfelkonstruktionsaufgabe im Rahmen einer Diskriminanzanalyse besser dazu in der Lage ist, zwischen Personen, die beispielsweise einer Fachrichtung angehören, in der die Fähigkeit zum räumlichen Denken besonders gefordert ist, und Personen, die einer solchen Fachrichtung nicht angehören, zu unterscheiden.

Es kann die Frage gestellt werden, ob eine Antwortausschlussstrategie, wie sie in Studie 3 untersucht wurde, auch bei Würfelrotationsaufgaben angewendet werden kann. Da Würfelrotationsaufgaben anders aufgebaut sind und nicht direkt mit Matrizenaufgaben vergleichbar sind, ist diese Frage nicht einfach zu beantworten. Die Ergebnisse aus Studie 3 weisen zudem darauf hin, dass die negativen Auswirkungen einer solchen Antwortausschlussstrategie auf die Konstruktvalidität des Tests eher vernachlässigbar sind. Dennoch könnte sich zukünftige Forschung der Frage widmen, ob es auch bei Würfelrotationsaufgaben

eine vergleichbare Antwortausschlussstrategie gibt, die bis jetzt noch nicht untersucht wurde, und inwiefern diese die psychometrischen Eigenschaften der Aufgabe einschränkt.

Studie 3 zeigt, dass die mögliche Anwendung verschiedener Lösungsstrategien nicht immer zwingend ein Problem für die Validität des Tests darstellen muss. Auch bei Aufgaben zum räumlichen Denken gibt es abgesehen von Response-Elimination-Strategien, die als kompensatorische Ausweichstrategien angesehen werden, verschiedene Strategien, mit denen Testpersonen die Aufgaben im Sinne eines Constructive-Matching-Prozesses lösen können (Halpern, 2012; Lohman & Kyllonen, 1983). Es herrscht jedoch keine Einigung darüber, wie problematisch dies für die Validität eines Tests ist, was mitunter auch daran liegt, dass kognitive Strategien und deren Effekte schwer zu messen sind. Da der Fokus der Studien zum räumlichen Denken auf der Unterscheidung zwischen Response-Elimination und Constructive Matching lag, wurde auf unterschiedliche Constructive-Matching-Strategien nicht näher eingegangen. In der Fachliteratur sind jedoch verschiedene Lösungswege für räumliche Aufgaben bekannt. So können die Aufgaben mit einer holistischen Strategie gelöst werden, bei der das Item als Ganzes mental rotiert wird (Shepard & Metzler, 1971), oder mit einer analytischen Strategie, bei der die Informationen des Items listenartig gespeichert werden und die Lösung so ‚errechnet' wird (Bethell-Fox & Shepard, 1988). In der Mitte dieses Kontinuums liegt eine piecemealartige Verarbeitung, bei der einzelne Elemente des Items Stück für Stück rotiert werden (Mumaw, Pellegrino, Kail & Carter, 1984). Insgesamt hängt die Art der angewendeten Strategie von Item- aber auch von Personencharakteristiken ab. Testpersonen können eine Neigung zu einer bestimmten Strategie haben, aber auch die Strategie während des Tests adaptiv wechseln. Ein Repertoire an Strategien, aus dem flexibel ausgewählt werden kann, führt aber wohl zum größten Erfolg in räumlichen Aufgaben (Lohman & Kyllonen, 1983).

Sofern verschiedene Strategien als Ausdruck der gleichen zugrundeliegenden, zu messen beabsichtigten Fähigkeit angesehen werden können und zu gleicher

Leistung führen, stellt deren Anwendung kein Problem dar. Ist der Test jedoch darauf ausgelegt, mit nur einer bestimmten Strategie gelöst zu werden, sodass eine Strategie effektiver ist als eine andere, ist der Test in seiner Konstruktvalidität eingeschränkt, denn es kann nicht mehr sicher davon ausgegangen werden, dass er bei allen Personen akkurat die relevante Fähigkeit misst. Die Wahl einer bestimmten Strategie muss derweil nicht zwingend etwas mit einer geringeren Fähigkeit zu tun haben. So gibt es beispielsweise Hinweise darauf, dass Frauen eher zu analytischen Strategien neigen aus Gründen der Präferenz oder Übung, nicht aber wegen einer geringeren Verarbeitungskapazität, und dass sie deshalb in bestimmten räumlichen Aufgaben schlechter abschneiden (Glück & Fitting, 2003). Wenn eine Personengruppe unabhängig von ihrem Fähigkeitsniveau systematisch zu einer bestimmten Strategie neigt und diese eine Strategie in einem Test weniger effektiv ist als eine andere, ist der Test nicht nur nicht valide, sondern es muss auch gefragt werden, ob ein solcher Test fair ist, besonders wenn er das einzige herangezogene Maß für die übergeordnete Fähigkeit darstellt. Es sollte deshalb für die Würfelkonstruktionsaufgabe noch weiter untersucht werden, ob hier verschiedene Constructive-Matching-Strategien angewendet werden und welche Auswirkungen dies auf die Validität und die Fairness des Tests hat.

Es wäre sinnvoll, ein ähnliches Studiendesign wie in Studie 3 auch auf die Würfelkonstruktionsaufgabe anzuwenden, um zu sehen, inwiefern die in Studie 2 gefundene schlechtere Leistung der Frauen im distraktorfreien Format mit der Anwendung einer bestimmten Constructive-Matching-Strategie zusammenhängt und inwiefern der Test dadurch unfair wird. In Studie 2 ergaben sich für die distraktorfreie Würfeltestversion nicht nur Leistungsdifferenzen zugunsten der Männer, eine Post-hoc-Analyse der konvergenten Validitäten zeigte zusätzlich, dass die Korrelation der Testergebnisse des Würfeltests mit den Testergebnissen des IST bei den Männern deutlich höher war als bei den Frauen (r = .66 vs. r = .46). Der Unterschied war zwar nicht signifikant (z = .98, p = .16), es ist aber anzunehmen, dass dies nur an der zu kleinen Stichprobe für solche Analysen lag (n = 19 vs. n = 38). Die Annahme sollte an einer größeren Stichprobe überprüft werden. Wenn

den Testpersonen die gewünschte Lösungsstrategie erklärt würde und sie dazu aufgefordert würden, diese Strategie anzuwenden, könnten sie womöglich bessere Leistungen erbringen, denn es gibt bereits Studien, die einen erfolgreichen trainingsbedingten Strategiewechsel und Leistungsanstieg zeigen konnten (Gittler & Glück, 1998; Glück, Machat, Jirasko & Rollett, 2002). Möglicherweise würde sich die Leistung der Frauen im Vergleich zu den Männern in einem größeren Ausmaß verbessern und die konvergente Validität in einem höheren Maß ansteigen. Wenn dem so wäre, dann würde dies bedeuten, dass der Test zuvor unfair war, da er die Fähigkeit der Frauen zum räumlichen Denken weniger adäquat widergespiegelt hat als bei den Männern. Würde sie sich nicht verändern, dann ließe sich die schlechtere Testleistung der Frauen nicht durch die Anwendung weniger effektiver Strategien erklären und der Test wäre in dieser Hinsicht geschlechtsfair.

Ebenso könnte ein Studiendesign wie in Studie 3 die Bedeutung von Response-Elimination-Strategien genauer klären und untersuchen, welche Fähigkeit der Anwendung von Response-Elimination-Strategien zugrunde liegt. Hierzu könnte überprüft werden, inwiefern sich die konvergente Validität zu einem allgemeinen Intelligenztest oder einem anderen Test zum räumlichen Denken verändert, wenn die Testpersonen darauf hin trainiert werden, Response-Elimination-Strategien anzuwenden.

4.3 Fazit

Die vorliegende Arbeit zeigt deutlich, dass geläufige Würfelrotationsaufgaben zur Erfassung räumlichen Denkens mit Distraktoren dem Format ohne Distraktoren hinsichtlich der psychometrischen Kriterien unterlegen sind. Damit reihen sich die Ergebnisse in Ergebnisse aus der Matrizenforschung ein und zeigen, dass sich die Distraktorproblematik keinesfalls auf figurale Matrizenaufgaben beschränkt. Die Ergebnisse aus Studie 3 relativieren die Distraktorproblematik etwas. Es gibt dort keinen Hinweis darauf, dass eine Antwortausschlussstrategie, die den Itemstamm ignoriert, die Validität des Tests beeinträchtigt. Die Anwendung von Response-

Elimination-Strategien, die den Aufgabenstamm miteinbeziehen, bleibt als Möglichkeit jedoch bestehen und stellt das wahrscheinlichere Verhalten und somit das größere Problem für die Konstruktvalidität dar. Des Weiteren bleibt das Problem der eingeschränkten Variationsmöglichkeit der Schwierigkeiten beim Einsatz von Distraktoren bestehen. Die Ergebnisse von Studie 1 und 2 unterstützen diese Annahme. Im Vergleich zu existierenden distraktorgestützten Aufgabenformaten kann die Konstruktvalidität durch den Verzicht auf Distraktoren deutlich erhöht und die Itemschwierigkeit besser variiert werden. Das Konstruktionsformat stellt somit in diesem Fall, wie auch schon der distraktorfreie figurale Matrizentest DESIGMA (Becker & Spinath, 2014) im Feld figuraler Matrizentests, ein vielversprechendes alternatives Aufgabenformat zur Erfassung räumlichen Denkens dar und schließt eine Lücke hinsichtlich der bis jetzt fehlenden Aufgaben sehr hoher Schwierigkeit.

Obwohl die Auswirkungen, die das Antwortformat auf die Konstruktvalidität eines Tests haben kann, jenseits figuraler Aufgaben in der Fachliteratur meist vernachlässigt wird, gibt es bereits Studien, die darauf hinweisen, dass die problematischen Auswirkungen von Distraktoren keinesfalls auf figurale Aufgaben beschränkt sind, sondern z. B. auch bei Leseverständnisaufgaben zu finden sind (Rost & Sparfeldt, 2007). Es ist daher wünschenswert, in der Forschung einen neuen Fokus auf diese Problematik zu richten und weitere Aufgabentypen im Konstruktionsformat zu erstellen, sodass es möglicherweise eine ganze Intelligenztestbatterie geben wird, die ohne Distraktoren auskommt.

5 Literaturverzeichnis

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

Arendasy, M. E. & Sommer, M. (2013). Reducing response elimination strategies enhances the construct validity of figural matrices. *Intelligence*, *41*(4), 234–243. doi: 10.1016/j.intell.2013.03.006

Baenninger, M. & Newcombe, N. (1995). Environmental input to the development of sex-related differences in spatial and mathematical ability. *Learning and Individual Differences*, *7*(4), 363–379. doi: 10.1016/1041-6080(95)90007-1

Beauducel, A. & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186–203. doi: 10.1207/s15328007sem1302_2

Becker, N., Preckel, F., Karbach, J., Raffel, N. & Spinath, F. M. (2015). Die Matrizenkonstruktionsaufgabe. *Diagnostica*, *61*(1), 22–33. doi: 10.1026/0012-1924/a000111

Becker, N., Schmitz, F., Falk, A., Feldbrügge, J., Recktenwald, D., Wilhelm, O., … Spinath, F. (2016). Preventing response elimination strategies improves the convergent validity of figural matrices. *Journal of Intelligence*, *4*(1), 2. doi: 10.3390/jintelligence4010002

Becker, N. & Spinath, F. M. (2014). *DESIGMA. Design a matrix – advanced.* Göttingen: Hogrefe.

Bethell-Fox, C. E., Lohman, D. & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*(3), 205–238. doi: 10.1016/0160-2896(84)90009-6

Bethell-Fox, C. E. & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 12–23. doi: 10.1037/0096-1523.14.1.12

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Hrsg.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (S. 435–459). Cambridge, MA: MIT Press.

Brand, C. (1987). The importance of general intelligence. In S. Modgil & C. Modgil (Hrsg.), *Arthur Jensen: Consensus and controversy* (S. 251–265). New York, NY: Falmer.

Carpenter, P. A., Just, M. A. & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431. doi: 10.1037/0033-295X.97.3.404

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

Carroll, J. B. (1996). A three-stratum theory of intelligence: Spearman's contribution. In I. Dennis & P. Tapsfield (Hrsg.), *Human abilities: Their nature and measurement* (S. 1–17). Mahwah, NJ: Lawrence Erlbaum.

Cattell, R. B. (1971). *Abilities: Their structure and growth.* Boston, MA: Houghton Mifflin.

Cattell, R. B. & Cattell, A. K. (1973). *Handbook for the individual or group Culture Fair Intelligence Test (a measure of "g"): Scale 2, Form A and B.* Champaign, IL: Institute for Personality and Ability Testing.

Ceci, S. J. (1996). General intelligence and life success: An introduction to the special theme. *Psychology, Public Policy, and Law*, *2*(3–4), 403–417. doi: 10.1037/1076-8971.2.3-4.403

Ceci, S. J., Williams, W. M. & Barnett, S. M. (2009). Women's Underrepresentation in Science: Sociocultural and Biological Considerations. *Psychological Bulletin*, *135*(2), 218–261. doi: 10.1037/a0014412

Deary, I. J. (2001). Human intelligence differences: A recent history. *Trends in Cognitive Science*, *5*(3), 127–130. doi: 10.1016/S1364-6613(00)01621-1

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

Formann, A. K. (1979). *Wiener Matrizen-Test WMT. Ein Rasch-skalierter sprachfreier Intelligenztest. Manual.* Weinheim: Beltz Test.

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

Galton, F. (1883). *Inquiries into human faculty and its development.* London, England: Macmillan.

Gittler, G. (1990). *Dreidimensionaler Würfeltest 3DW*. Weinheim: Beltz Test.

Gittler, G. & Glück, J. (1998). Differential transfer of learning: Effects of instruction in descriptive geometry on spatial test performance. *Journal of Geometry and Graphics*, *2*(1), 71–84.

Glück, J. & Fitting, S. (2003). Spatial strategy selection: Interesting incremental information. *International Journal of Testing*, *3*(3), 293–308. doi: 10.1207/S15327574IJT0303_7

Glück, J., Machat, R., Jirasko, M. & Rollett, B. (2002). Training-related changes in solution strategy in a spatial test: An application of item response models. *Learning and Individual Differences*, *13*(1), 1–22. doi: 10.1016/S1041-6080(01)00042-5

Goldhammer, F. & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*(2), 108–119. doi: 10.1016/j.intell.2011.02.001

Goldhammer, F., Naumann, J. & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's Matrices. *Journal of Intelligence*, *3*(1), 21–40. doi: 10.3390/jintelligence3010021

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H. & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626. doi: 10.1037/a0034716

Gottfredson, L. S. (1997). Mainstream Science on Intelligence: An Editorial With 52 Signatories, History, and Bibliography. *Intelligence*, *24*(1), 13–23. doi: 10.1016/S0160-2896(97)90011-8

Gottfredson, L. S. (2004). Intelligence: Is it the epidemiologists' elusive "fundamental cause" of social class inequalities in health? *Journal of Personality and Social Psychology*, *86*(1), 174–199. doi: 10.1037/0022-3514.86.1.174

Guilford, J. P. & Lacy, J. I. (1947). *Printed classification tests, AAF, Aviation Psychological Research Report, #5*. Washington, DC: U.S. Government Printing Office.

Halpern, D. F. (2012). *Sex differences in cognitive abilities*. New York, NY: Taylor and Francis.

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

Hambrick, D. Z., Libarkin, J. C., Petcovic, H. L., Baker, K. M., Elkins, J., Callahan, C. N., ... LaDue, N. D. (2011). A test of the circumvention-of-limits hypothesis in scientific problem solving: The case of geological bedrock mapping. *Journal of Experimental Psychology: General*, *141*(3), 397–403. doi: 10.1037/a0025927

Harle, M. & Towns, M. (2011). A review of spatial ability literature, its connection to chemistry, and implications for instruction. *Journal of Chemical Education*, *88*(3), 351–360. doi: 10.1021/ed900003n

Hayes, T. R., Petrov, A. A. & Sederberg, P. B. (2011). A novel method for analyzing sequential eye movements reveals strategic influence on Raven's Advanced Progressive Matrices. *Journal of Vision*, *11*(10), 1–11. doi: 10.1167/11.10.10

Heim, A. W. (1968). *AH4 Group Test of General Intelligence*. Slough, England: NFER Publishing Company.

Heim, A. W. (1970). *AH4 Group Test of General Intelligence. Manual.* Windsor, England: NFER Publishing Company.

Heller, K. A. (1987). Perspektiven einer Hochbegabungsdiagnostik. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *8*(3), 159–172.

Heller, K. A. & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen,*

Revision. Göttingen: Hogrefe.

Heller, K. A. & Perleth, C. (2007). Talentförderung und Hochbegabtenförderung in Deutschland. In K. A. Heller & A. Ziegler (Hrsg.), *Begabt sein in Deutschland* (S. 139–170). Münster: LIT Verlag.

Heublein, U., Richter, J., Schmelzer, R. & Sommer, D. (2014). Die Entwicklung der Studienabbruchquoten an deutschen Hochschulen. *HIS: Forum Hochschule*, *4*, 1–20.

Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253–270. doi: 10.1037/h0023816

Hossiep, R., Turck, D. & Hasella, M. (1999). *BOMAT-advanced. Bochumer Matrizentest advanced. Handanweisung.* Göttingen: Hogrefe.

Irwing, P. & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology*, *96*(4), 505–524. doi: 10.1348/000712605X53542

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

Jansen-Osmann, P. & Heil, M. (2007). Suitable stimuli to obtain (no) gender differences in the speed of cognitive processes involved in mental rotation. *Brain and Cognition*, *64*(3), 217–227. doi: 10.1016/j.bandc.2007.03.002

Jarosz, A. F. & Wiley, J. (2012). Why does working memory capacity predict RAPM performance? A possible role of distraction. *Intelligence*, *40*(5), 427–438. doi: 10.1016/j.intell.2012.06.001

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport: Praeger.

Kell, H. J. & Lubinski, D. (2013). Spatial ability: A neglected talent in educational and occupational settings. *Roeper Review*, *35*(4), 219–230. doi: 10.1080/02783193.2013.829896

Kelley, T. L. (1928). The boundaries of mental life and a technique for their investigation. In T. L. Kelley (Hrsg.), *Crossroads in the mind of man* (S. 1–23).

Palo Alto, CA: Stanford University Press.

Klein Entink, R. H., Fox, J. P. & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21–48. doi: 10.1007/S11336-008-9075-Y

Kornell, N., Hays, M. J. & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. doi: 10.1037/a0015729

Kreuzpointner, L. (2013). *Konstruktionsrational der Kurzversion des LPS 2 (LPS-2K)*. Abgerufen von https://epub.uni-regensburg.de/35693/

Kreuzpointner, L., Lukesch, H. & Horn, W. (2013). *Leistungsprüfsystem 2 (LPS-2). Manual.* Göttingen: Hogrefe.

Kubinger, K. D. (2009). *Adaptives Intelligenzdiagnostikum 2 (Version 2.2).* Göttingen: Beltz Test.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow & H. B. Mann (Hrsg.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (S. 278–292). Stanford, CA: Stanford University Press.

Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007). *I-S-T 2000 R. Intelligenz-Struktur-Test 2000 Revision* (2. Aufl.). Göttingen: Hogrefe.

Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W. (2012). *Intelligenz-Struktur-Test Screening (IST Screening)*. Göttingen: Hogrefe.

Linn, M. C., Chang, H.-Y., Chiu, J. L., Zhang, H. & McElhaney, K. (2010). Can desirable difficulties overcome deceptive clarity in scientific visualizations? In A. Benjamin (Hrsg.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (S. 239–262). New York, NY: Routledge.

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. doi: 10.2307/1130467

Lohman, D. F. (1979). *Spatial ability: A review and re-analysis of the correlational literature (Technical Report No. 8).* Stanford, CA: Stanford University, Aptitude

Research Project, School of Education.

Lohman, D. F. (1986). The effect of speed-accuracy tradeoff on sex differences in mental rotation. *Attention, Perception, & Psychophysics*, *39*(6), 427–436. doi: 10.3758/BF03207071

Lohman, D. F. (1988). Spatial ability as traits, processes and knowledge. In R. J. Sternberg (Hrsg.), *Advances in the psychology of human intelligence, Bd. 4* (S. 181–248). Hillsdale, NJ: Erlbaum.

Lohman, D. F. (1996). Spatial ability and *g*. In I. Dennis & P. Tapsfield (Hrsg.), *Human abilities: Their nature and measurement* (S. 97–116). Mahwah, NJ: Erlbaum.

Lohman, D. F. & Korb, K. A. (2006). Gifted today but not tomorrow ? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted*, *29*(4), 451–484. doi: 10.4219/jeg-2006-245

Lohman, D. F. & Kyllonen, P. C. (1983). Individual differences in solution strategy on spatial tasks. In R. F. Dillon & R. R. Schmeck (Hrsg.), *Individual Differences in Cognition* (S. 105–135). New York, NY: Academic Press.

Marshalek, B., Lohman, D. F. & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*(2), 107–127. doi: 10.1016/0160-2896(83)90023-5

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10. doi: 10.1016/j.intell.2008.08.004

Millsap, R. E., Zalkind, S. S. & Xenos, T. (1990). Quick-reference tables to determine the significance of the difference between two correlation coefficients from two independent samples. *Educational and Psychological Measurement*, *50*(2), 297–307. doi: 10.1177/0013164490502008

Mitchum, A. L. & Kelley, C. M. (2010). Solve the problem first: Constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 699–710. doi: 10.1037/a0019182

Mittring, G. & Rost, D. H. (2008). Die verflixten Distraktoren: Über den Nutzen einer

theoretischen Distraktorenanalyse bei Matrizentests (für besser Begabte und Hochbegabte). *Diagnostica*, *54*(4), 193–201. doi: 10.1026/0012-1924.54.4.193

Mohler, J. L. (2008). A review of spatial ability research. *Engineering Design Graphics Journal*, *72*(2), 19–30.

Mumaw, R. J., Pellegrino, J. W., Kail, R. V. & Carter, P. (1984). Different slopes for different folks: Process analysis of spatial aptitude. *Memory & Cognition*, *12*(5), 515–521. doi: 10.3758/BF03198314

Muthén, B. O. & Muthén, L. K. (2007). *MPlus Version 6.* Los Angeles, CA: Muthén & Muthén.

Naglieri, J. A. (2003). *NNAT Naglieri Nonverbal Ability Test. Individual Administration. Stimulus book.* San Antonio, TX: The Psychological Corporation.

Neisser, U., Boodoo, G., Bouchard Jr., T. J., Boykin, A. W., Brody, N., Ceci, S. J., ... Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*(2), 77–101. doi:10.1037/0003-066X.51.2.77

Neubauer, A. & Stern, E. (2007). *Lernen macht intelligent: Warum Begabung gefördert werden muss.* München: DVA.

Newcombe, N. S. (2010). Picture this: Increasing math and science learning by improving spatial thinking. *American Educator*, *34*(2), 29–35, 43. doi: 10.1037/A0016127

Oswald, W. D. & Roth, E. (1978). *Der Zahlen-Verbindungs-Test ZVT. Ein sprachfreier Intelligenz-Schnell-Test.* Göttingen: Hogrefe.

Pellegrino, J. W. & Hunt, E. B. (1991). Cognitive models for understanding and assessing spatial abilities. In H. A. H. Rowe (Hrsg.), *Intelligence: Reconceptualization and measurement* (S. 203–225). Hillsdale, NJ: Lawrence Erlbaum.

Petermann, F. & Petermann, U. (2007). *HAWIK-IV. Hamburg-Wechsler Intelligenztest für Kinder – IV (Übersetzung und Adaption der WISC-IV von D. Wechsler).* Bern: Huber.

Price, J. (2010). The effect of instructor race and gender on student persistence in

STEM fields. *Economics of Education Review*, *29*(6), 901–910. doi: 10.1016/j.econedurev.2010.07.009

Quaiser-Pohl, C. (2012). Mädchen und Frauen in MINT: Ein Überblick. In H. Stöger, A. Ziegler & M. Heilemann (Hrsg.), *Mädchen und Frauen in MINT. Bedingungen von Geschlechtsunterschieden und Interventionsmöglichkeiten* (S. 13–40). Berlin: LIT Verlag.

R Core Team (2015). *R. A language and environment for statistical computing.* Wien, Österreich: R Foundation for Statistical Computing.

Raven, J. C. (1976). *Advanced Progressive Matrices, Set II.* Oxford, England: Oxford University Press.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi: 10.18637/jss.v048.i02

Rost, D. H. (2009). *Intelligenz*. Weinheim: Beltz.

Rost, D. H. (2013). *Handbuch Intelligenz.* Weinheim: Beltz.

Rost, D. H. & Sparfeldt, J. R. (2007). Leseverständnis ohne Lesen? Zur Konstruktvalidität von multiple-choice-Leseverständnistestaufgaben. *Zeitschrift Für Pädagogische Psychologie*, *21*(3–4), 305–314. doi: 10.1024/1010-0652.21.3.305

Scherer, R., Greiff, S. & Hautamäki, J. (2015). Exploring the relation between speed and ability in complex problem solving. *Intelligence*, *48*, 37–50. doi: 10.1016/j.intell.2014.10.003

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274. doi: 10.1037/0033-2909.124.2.262

Shea, D. L., Lubinski, D. & Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology*, *93*(3), 604–614. doi: 10.1037/0022-0663.93.3.604

Shepard, R. N. & Metzler, J. (1971). Mental rotation of three-dimensional

objects. *Science*, *171*(3972), 701–703. doi: 10.1126/science.171.3972.701

Shih, M., Pittinsky, T. L. & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, *10*(1), 80–83. doi: 10.1111/1467-9280.00111

Smith, I. M. (1964). *Spatial ability: Its educational and social significance.* London, England: University of London Press.

Snow, R. E. (1989). Aptitude-treatment interaction as a framework for research on individual differences in learning. In P. L. Ackerman, R. J. Sternberg & R. Glaser (Hrsg.), *Learning and individual differences: Advances in theory and research* (S. 13–59). New York, NY: W. H. Freeman & Company.

Sorby, S. A. (2001). A course in spatial visualization and its impact on the retention of female engineering students. *Journal of Women and Minorities in Science and Engineering*, *7*(2), 50–70. doi: 10.1615/JWomenMinorScienEng.v7.i2.50

Sorby, S. A. (2009). Educational research in developing 3-D spatial skills for engineering students. *International Journal of Science Education*, *31*(3), 459–480. doi: 10.1080/09500690802595839

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* London, England: Macmillan.

Stemmler, G., Hagemann, D., Amelang, M. & Spinath, F. (2016). *Differentielle Psychologie und Persönlichkeitsforschung.* Stuttgart: Kohlhammer.

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

Süß, H.-M. (2001). Prädiktive Validität der Intelligenz im schulischen und außerschulischen Bereich. In E. Stern & J. Guthke (Hrsg.), *Perspektiven der Intelligenzforschung* (S. 109–135). Lengerich: Pabst.

Terlecki, M. S., Newcombe, N. S. & Little, M. (2008). Durable and generalized effects of spatial experience on mental rotation: Gender differences in growth patterns. *Applied Cognitive Psychology*, *22*(7), 996–1013. doi: 10.1002/acp.1420

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

Press.

Thurstone, L. L. (1944). *A factorial study of perception.* Chicago, IL: University of Chicago Press.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C. & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. doi: 10.1037/a0028446

Vernon, P. E. (1961). *The structure of human ablilites* (2. Aufl.). London, England: Methuen.

Vigneau, F., Caissie, A. F. & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, *34*(3), 261–272. doi: 10.1016/j.intell.2005.11.003

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. doi: 10.1037/0033-2909.117.2.250

Wai, J., Lubinski, D. & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101*(4), 817–835. doi: 10.1037/a0016127

Webb, R. M., Lubinski, D. & Benbow, C. P. (2007). Spatial ability: A neglected dimension in talent searches for intellectually precocious youth. *Journal of Educational Psychology, 99*(2), 397–420. doi: 10.1037/0022-0663.99.2.397

Wechsler, D. (2012). *WAIS-IV. Wechsler Adult Intelligence Scale – fourth edition (Deutsche Adaption).* Frankfurt/Main: Pearson.

Weiß, R. H. & Osterland, J. (2012). *CFT 1-R. Grundintelligenztest Skala 1 Revision.* Göttingen: Hogrefe.

White, A. P. & Zammarelli, J. E. (1981). Convergence principles: Information in the answer sets of some multiple-choice intelligence tests. *Applied Psychological Measurement*, *5*(1), 2–27. doi: 10.1177/014662168100500103

Wilhelm, O. & Schulze, R. (2002). The relation of speeded and unspeeded reasoning

with mental speed. *Intelligence*, *30*(6), 537–554. doi: 10.1016/S0160-2896(02)00086-7

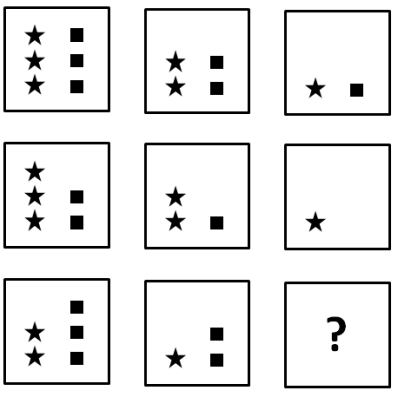
6 Anhang

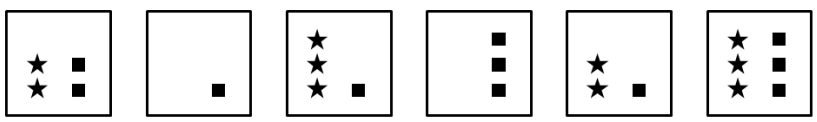
6.1 Anhang 1: Training aus Studie 3

Training

Wie Du bereits weißt, geht es in dieser Studie um das Bearbeiten von Matrizenaufgaben. Zur Erinnerung: Matrizenaufgaben bestehen aus einem Aufgabenstamm mit mehreren Quadraten, von denen ein Quadrat leer ist. Aus vorgegebenen Antwortmöglichkeiten soll die richtige Lösung für das leere Quadrat ausgewählt werden. Die Lösungsalternativen ähneln sich sehr, damit die richtige Lösung nicht zu leicht gefunden werden kann.

Das Ganze wird jetzt anhand eines Beispiels genauer erklärt. Im Folgenden siehst Du eine Matrizenaufgabe:



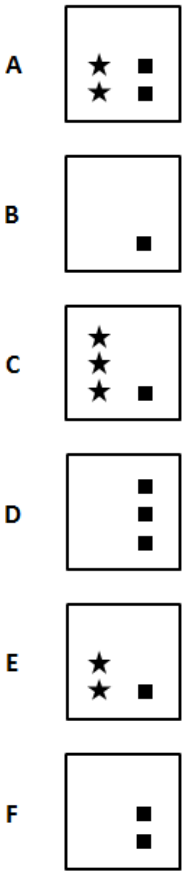


Im Aufgabenstamm erkennst Du, dass in jeder Zeile von links nach rechts immer ein Stern und ein Quadrat weggenommen werden. In der letzten Zeile gibt es im ersten Feld zwei Sterne und drei Quadrate und im zweiten Feld einen Stern und zwei Quadrate. Nimmt man noch einen Stern und ein Quadrat weg, bleibt nur noch ein Quadrat übrig. Die richtige Lösung müsste daher B sein.

Ich erkläre Dir jetzt einen Trick, mit dem Du Hinweise auf die richtige Lösung finden kannst: Du siehst Dir lediglich die vorgegebenen Lösungsalternativen genauer an. Der Trick besteht nun darin, zu schauen, welche verschiedenen Merkmale es gibt und welche Ausprägungen davon am häufigsten auftreten. Die Lösungsalternative, in der die am häufigsten auftretenden Merkmale alle enthalten sind, ist mit einer hohen Wahrscheinlichkeit die richtige.

Der Trick wird nun anhand des Beispiels von oben erklärt.
Zuerst musst Du erkennen, welche verschiedenen Merkmale es gibt, die unabhängig voneinander verändert werden. In diesem Fall sind dies die Merkmale **Sterne** und **Quadrate**.

Jetzt musst Du auszählen, welche Ausprägungen der Merkmale wie oft auftreten. Dies siehst Du in der folgenden Abbildung:



Jetzt schaust Du, wie oft die verschiedenen Ausprägungen der Merkmale auftreten. Für das Beispiel ergeben sich die folgenden Häufigkeiten:

Sterne:

3 Sterne kommen 1-mal vor,

2 Sterne kommen 2-mal vor,

0 Sterne kommen 3-mal vor.

Für die Quadrate ergeben sich die folgenden Häufigkeiten:

3 Quadrate kommen 1-mal vor,

2 Quadrate kommen 2-mal vor,

1 Quadrat kommt 3-mal vor.

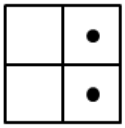
Du siehst, dass die Ausprägung 0 Sterne und 1 Quadrat am häufigsten vorkommt. Du siehst, dass dies auf Lösungsalternative B zutrifft, die die richtige ist. Du kannst also mit diesem Auszähltrick auch auf die richtige Lösung kommen.

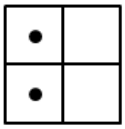
Ich fasse die **Schritte des Tricks** noch einmal zusammen:

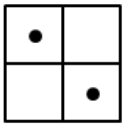
- Herausfinden, welche unterschiedlichen Merkmale es gibt
- Auszählen, wie oft verschiedene Merkmalsausprägungen vorkommen
- Die Alternative wählen, bei der die häufigsten Merkmalsausprägungen zusammen vorkommen

In dem ersten Beispiel ging es nur um die Anzahl der Merkmale. Es kann aber auch um die Position der Merkmale gehen. Schau Dir hierzu das folgende Beispiel an:

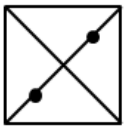
Du kannst hier generell zwei Merkmale unterscheiden: die Kreise und die Linien. Die Kreise können an vier Positionen auftreten: links oben (lo), links unten (lu), rechts oben (ro) und rechts unten (ru). Die Linien kommen in zwei Varianten vor: horizontal, vertikal (h, v) und diagonal (d).

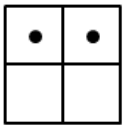
A 

B 

C 

D 

E 

F 

Wir zählen zuerst die Ausprägungen der Linien aus:

h, v kommt 4-mal vor,

d kommt 2-mal vor.

Für die Kreise zählen wir Folgendes aus:

ro, ru kommt 1-mal vor,

lo, lu kommt 1-mal vor,

lo, ru kommt 2-mal vor,

lu, ro kommt 1-mal vor,

lo, ro kommt 1-mal vor.

Du müsstest nun die Lösungsalternative C auswählen, da bei ihr die Linien in der Ausprägung h, v und die Kreise in der Ausprägung lo, ru vorkommen.

Hast Du noch Fragen?