universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„WIT-2 Kritik und Anpassung des Moduls M5 Merkfähigkeit“

verfasst von / submitted by

Roman Göttner

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Naturwissenschaften (Mag. rer. nat.)

Wien, 2016 / Vienna, 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 298
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Diplomstudium Psychologie
Betreut von / Supervisor:	Ao. Univ.-Prof. Dr. Georg Gittler

Danksagung

Sehr großer Dank geht an meinen Professor und Betreuer Prof. Dr. Georg Gittler, welcher mit besonderer Geduld und viel Verständnis, diese Diplomarbeit ermöglichte. Herr Prof. Dr. Georg Gittler gab mir die Möglichkeit und den thematischen Rahmen mein Studium erfolgreich abzuschließen und trägt auf diesem Wege auf besondere Art und Weise zu meiner Zukunft bei.

Der Studienassistentin Frau Magdalena Siegel danke ich aufrichtig für ihren kommunikativen und beratenden Beistand bezüglich der Fertigstellung dieser Arbeit. Ohne ihre Mühe und Zeit wäre diese Diplomarbeit nicht möglich gewesen.

Außerdem möchte ich mich bei Johann Münscher bedanken, in erster Linie für den fachlichen und inhaltlichen Diskurs, der zur Erfassung und Formulierung dieser Diplomarbeit positiv beigetragen hat, sowie für die langjährige und spezielle Freundschaft.

Tiefer herzlicher Dank geht an meine Freundin Marietta Lieb, welche mit besonderer Geduld und Hilfe, mir organisatorisch unter die Arme griff und mir Verständnis und Ruhe in schweren Zeiten entgegenbrachte. Hiermit danke ich ihr für die schönen Erlebnisse und die gemeinsame Zeit während dieses Studiums.

Unendliche Dankbarkeit und Zuneigung gilt meiner Familie. Sie haben mich in meiner ganzen Entwicklung und während meines Lebens mit unerschöpflicher Liebe und Zuneigung in allen Lebenslagen ständig begleitet. Sie waren und sind mein sicherer Hafen, der Wind in meinen Segeln und das Licht am Ende des Horizontes und die Kraft in meinem Herzen. Ihnen sei aus dem tiefsten meiner Seele gedankt und diese Arbeit gewidmet.

Aufrichtiger Dank geht an die Teilnehmer dieser Studie für ihre Zeit und ihren Entschluss mit zu machen.

INHALTSVERZEICHNIS

1 EINLEITUNG

Fragen nach bestimmten Eigenschaften und Fähigkeiten von Personen sind heutzutage essentiell und im Rahmen der psychologischen Arbeit sowie in anderen Bereichen des Lebens wie Einstellungsverfahren und dem empirischen wissenschaftlichen Arbeiten nicht mehr weg zu denken. Zur Beantwortung dieser Fragen braucht es verantwortungsvoll konzipierte und qualitativ hochwertige Standards, welche möglichst genau differenzieren und diese Eigenschaften messbar machen. Diesbezüglich ist es von entscheidender Wichtigkeit die Messinstrumente, wie Fragebögen oder Tests, auf eine Art und Weise zu konstruieren und zu überprüfen, dass diese als genaue und zuverlässige Werkzeuge dienen können.

Genau wie beim Eichen einer Waage geht dies mit einer deutlichen Verantwortung einher. Wenn die verwendeten Messwerkzeuge unpräzise sind, werden die Messungen Fehler beinhalten, welche dann wieder fehlerhafte und unpräzise Annahmen und Aussagen erzeugen. Im Rahmen der Psychologie stellt sich hier eine besondere Herausforderung an die Entwicklung dieser Werkzeuge, da psychologische Konstrukte, wie zum Beispiel Intelligenz, nicht mit einer einfach konzipierten Waage zu erfassen sind. Diese Konstrukte sind in erster Linie oft schwierig zu operationalisieren, nur bedingt greifbar und nicht einfach messbar zumachen. Solche latenten Konstrukte sind nicht direkt messbar, aber auf einer Theorie oder einem Modell begründet. Aus diesen Theorien lassen sich manifeste Indikatoren ableiten, welche zu einem gewissen Grad latente Konstrukte sinnvoll erklären und Rückschlüsse auf jene zulassen.

Die Konstruktion dieser Messwerkzeuge sollte somit einer besonderen Vorsicht unterliegen. Eine zu grob oder zu ungenaue angelegte Erfassung manifester Indikatoren führt zwar auch zu Messergebnissen, die jedoch möglicherweise über wenig Aussagekraft,

hinsichtlich der wirklichen Fähigkeiten und Kapazitäten einer Person verfügen. Werden diese falschen oder unpräzisen Ergebnisse dann im Rahmen einer spezifischen Situation entweder übermäßig und inkorrekt interpretiert oder auf einem für Laien missverständlichem Wege weitergeleitet, kann dies sogar der getesteten Person schaden. Aus diesen Gründen ist die Einhaltung und Evaluierung verschiedener Gütekriterien, wie zum Beispiel Reliabilität, Validität und Objektivität bei jedem dieser Diagnostika von entscheidender Bedeutung.

Die Konstruktion von Tests, die Erfassung von dem Konstrukt der Intelligenz sowie die Theoriemodellbildung jener, blicken auf eine hundertjährige Tradition zurück und stellen somit einen der meist beforschten und häufig diskutierten Themenbereiche der Psychologie dar (Blum et. al, 1998; Eysenck, Brengelmann, & Fulker, 1980; Petermann, 2006). Jedoch ist die Diskussion um Intelligenz weiterhin aktuell und bis heute ist jene Messung mit den Problemen und Herausforderungen der Testkonstruktion konfrontiert. Um die Problematik die mit dem Intelligenzbegriff und seiner Messung einhergeht zu verdeutlichen, soll hier das viel genannte Zitat von Edwin Boring (1923) Erwähnung finden: *„Intelligenz ist, was Intelligenztests testen.“* In dieser Aussage spiegelt sich die weiterhin aktuelle Diskussion über die gleichzeitige Ungenauigkeit des Intelligenzbegriffs und der Schwierigkeit des Erfassens wieder. Das Konstrukt der Intelligenz definiert sich nach Boring dementsprechend durch die Diagnostika, also durch den Intelligenztest. In diesem Sinne sind die Autoren dieser Werkzeuge in erster Linie damit beauftragt eine gute und sinnvolle Brücke zu schlagen, zwischen den Fähigkeiten der Testpersonen, also den Persönlichkeitseigenschaften, die Leistungen ermöglichen, und den erbrachten Leistungen, welche die Ergebnisse von Handlungen wiederspiegeln und mit gut und schlecht bewertbar sind.

Im Rahmen dieser Arbeit soll der Wilde-Intelligenz-Test 2 (WIT-2) von M. Kersting, K. Althoff und A. O. Jäger (2008) hinsichtlich dieser Gütekriterien näher betrachtet und beleuchtet werden, insbesondere das neu erstellte Modul *Merkfähigkeit*. In diesem Modul sind hinsichtlich dem Aspekt der Testkonstruktion und spezifischer der Kennwerte der Items Auffälligkeiten zu verzeichnen. Konkret erzeugen die erfassten Werte bezüglich Reliabilität und Item-Trennschärfe Unklarheiten. So sollen die Fragen beantwortet werden, ob der Untertest *Merkfähigkeit* den oben genannten Kriterien entspricht, sowie durch entsprechende Modifikationen eine mögliche Verbesserung des Moduls zu erreichen ist und eine weitere Vorlage dieses Moduls als empfehlenswert erscheint.

In den folgenden Abschnitten soll der WIT-2 kurz vorgestellt werden, sowie Literatur die Orientierungspunkte für Qualitätskriterien innerhalb der Testkonstruktion darstellen.

2 WILDE-INTELLIGENZ-TEST (WIT2)

Der Wilde-Intelligenz-Test 2 (WIT-2) geht als Intelligenztestbatterie basierend auf dem Wilde-Intelligenz-Test (WIT) auf die Arbeiten von Prof. Dr. Kurt Wilde sowie von dem Psychologenteam der Deutschen Gesellschaft für Personalwesen e. V. (DGP) zurück. Die grundlegende theoretische Basis dieser beider Verfahren lässt sich im Modell der Primary Mental Abilities (PMA) von Thurstone (1938) verwurzeln. Leistungen des Intellekts sind nach Thurstone auf sieben Primärfähigkeiten reduzierbar. Thurstone beschreibt *Reasoning*, *Space*, *Number*, *Verbal Comprehension*, *Word Fluency*, und *Perceptual Speed* sowie *Memory* als intellektuelle Fähigkeiten.

Diese Fähigkeiten werden auf Basis einer *Paper-Pencil* Version über sechs Module des WIT-2 erfasst, wobei zwei von allen freikombinierbaren Modulen berufsbezogenes Wissen überprüfen. Präsentiert werden acht Dimensionen innerhalb des Verfahrens (*Abkürzung-Name/Itemanzahl/Subtestanzahl): M1-Sprachliches Denken/40/2, M2-Rechnerisches Denken/40/2, M3-Räumliches Denken/40/2, Schlussfolgerndes Denken/60/3, M5-Merkfähigkeit/21/1, M6-Arbeitseffizienz/42/1, M7-Wissen Wirtschaft/20/1,* sowie *M8-Wissen Informationstechnologie/20/1*. Die Testdauer beläuft sich im Maximalfall mit Pausen auf etwa 2 Stunden und 30min, im Minimalfall auf etwa 20min und es liegen quasi-parallele Testformen A u. B vor. Der Test unterliegt damit einer gewissen Speed-Komponente.

Nach Angaben, der Autoren (Kersting, Althoff, & Jäger, 2008) orientiert sich der WIT-2 an einer Erweiterung des Modells von Thurstone, dem „Modifizierten Modell der Primary Mental Abilities“ (MMPMA). Dies soll das Modell Thurstone´s um die Annahmen des Facettenansatzes, der Hierarchie-Annahme und des „Cognitive Corelate Approachs“ erweitern (siehe. Kersting et. al., 2008). Das MMPMA differenziert, in Anlehnung an Carrol’s Three Stratum Theory of Cognitive Abilities (1993), drei Hierachie-Schichten (Strata) hinsichtlich ihres Generalisierungsgrades von gewissen Teilbereichen des Konstrukts der Intelligenz. So beinhaltet das Stratum III die generellen Faktoren der Fluiden Intelligenz, also dem Arbeitsgedächtnis, auf der einen Seite und die kristallisierte Intelligenz auf der anderen. In diesem Modell werden die Primärfaktoren Thurstone´s auf dem Stratum II, sowie die in den Modulen *M1, M2, M3, M4* und *M5* des WIT-2 erfassten Fähigkeiten lokalisiert. Carrol formuliert, dass ein Stratum die Breite und Generalität einer kognitiven Fähigkeit wiederspiegelt. Dementsprechend, vermutet Carrol auf dem Stratum I eine Anzahl von 68 Teilkonstrukten, welche nach zusätzlicher faktorieller Analyse verlangen (Carrol, 1993). Das MMPMA formuliert dort die Module *M6*, *M7* und *M8*.

Größtenteils wurden im WIT-2 sieben Untertests des WIT übernommen, wobei vier neuentwickelte Subtests inkludiert worden sind. Die methodische Grundlage dieses Verfahrens baut auf der klassischen Testtheorie auf. Als Einsatzbereiche des Testverfahrens werden in erster Linie die allgemeine Intelligenzdiagnostik und Eignungsdiagnostik im Rahmen der Personalauswahl und Personalentwicklung angeführt. Des Weiteren empfehlen die Autoren den WIT-2 als Instrument für die Intelligenzstrukturforschung (Kersting et. al., 2008).

Mit einer Normierung anhand einer Personenmenge von etwa 42.000 für eine Alterspanne von 28 Jahren deckt dieses Verfahren das Alter von 14 bis 42 Jahren ab. Hinsichtlich der Reliabilität für die Module des WIT-2, wurden diese über Retest-Reliabilität und interne Konsistenzen bestimmt, welche in einen Bereich von $\alpha = .77$ bis $\alpha = .98$ fallen. Bezüglich der Gültigkeit und damit der Kriteriums- und Konstruktvalidität, sind nach Sicht der Autoren mehrere Nachweise erbracht worden. Durch metaanalytisches Vorgehen soll die Kriteriumsvalidität anhand von mehreren Studien erbracht worden sein. Zusätzlich soll durch Zusammenhangsbestimmung mit weiteren Testverfahren eine Überprüfung der Konstruktvalidität (Kersting et. al., 2008) gewährleistet sein.

3 QUALITÄTSKRITERIEN DER TESTKONSTRUKTION

Eine der wichtigsten Aufgaben der psychologischen Diagnostik liegt in der Erfassung von Fähigkeitsausprägungen. Eid und Petermann (2006) unterstreichen diese Aussage als Notwendigkeit um überhaupt diagnostische Fragen beantworten und diesbezüglich Entscheidungen treffen zu können. Da auch die Intelligenzdiagnostik und ihre Ergebnisse einen gravierenden Einfluss auf den weiteren Werdegang der Testpersonen haben können, ist es von besonderer Bedeutung Verfahren zu wählen, welche Gütekriterien

gut erfüllen. (Kersting, 2008 ; Kersting, Häcker & Hornke, 2011). Gleichzeitig fordert Kersting (2007), dass auch sich in der Praxis befindende Verfahren, fortlaufend bezüglich ihrer Güte und Qualität überprüft werden sollten. Was macht also einen guten Test aus? Schermelleh-Engel, Kelava und Moosbrugger (2006) postulieren in ihrer Arbeit zehn Gütekriterien, welche bei der Konstruktion von psychologischen Tests von entscheidender Bedeutung sind: *1. Objektivität 2. Validität, 3. Reliabilität, 4.Skalierung, 5. Normierung, 6. Testökonomie, 7. Nützlichkeit, 8. Zumutbarkeit, 9. Unverfälschbarkeit* und *10. Fairness*.

Hauptgütekriterien sind die Objektivität, Validität und Reliabilität eines Testverfahrens (Eid & Schmidt, 2014). *Objektivität* wird als das Gütekriterium behandelt, welches die Unabhängigkeit der Testergebnisse zum/zur Untersuchungsleiter/-in einfordert. Inwiefern ein Instrument das erfasst, was es erfassen soll, und aus den Ergebnissen zutreffende und gültige Aussagen möglich sind, stellt die Frage nach dem Gütekriterium der *Validität* dar (Messick, 1989). Die *Reliabilität* wird von Eid und Schmidt (2014) als Gütekriterium beschrieben, welches anzeigt, inwieweit Messfehler Einflüsse auf das Messergebnis haben. Ist der Einfluss gering, spricht man von einem reliablen Messinstrument.

Im Allgemeinen wird die Objektivität in drei Aspekte unterteilt. *Durchführungsobjektivität* beschreibt, inwiefern Testergebnisse unabhängig von systematischen oder zufälligen Verhaltensvariationen der Testleitung sind. Um die Unabhängigkeit der Testergebnisse von der Testleitung, und somit die *Auswertungsobjektivität*, zu gewährleisten, ist die Wahl des Antwortformats von wichtiger Bedeutung sowie die Durchführung und Auswertung des jeweiligen Verfahrens. So ist beispielsweise ein geschlossenes Antwortformat vorteilhaft gegenüber einem offenen Format, da der Testwert hier keiner Deutung der Testleitung unterliegt (Kubinger, 2009).

Die sog. *Interpretationsobjektivität* ist dann gegeben, wenn durch verschiedene Untersucher, dieselben Auswertungsergebnisse von verschiedenen Testpersonen gleiche Schlüsse zulassen. Diese Form der Objektivität scheint vollkommen, wenn in der Auswertung numerische Werte festgehalten und/oder geliefert werden (Eid & Schmidt, 2014). Die Objektivität des WIT-2, hinsichtlich der Durchführungs-, Auswertungs- und Interpretationsobjektivität, werden von Gittler (2009) im Großen und Ganzen unter Betrachtung der oben genannten Aspekte attestiert.

Aspekte der *Validität* unterscheiden sich grundsätzlich in drei Bereichen. So ist die *Inhaltliche Validität* gekennzeichnet durch die Fragestellung, inwieweit das Testverfahren oder seine Teilelemente die zu erfassenden Fähigkeiten oder Merkmale repräsentieren (Cronbach, 1970). Hauptziel bei der Konstruktion von Testverfahren ist die Frage, ob die Testwerte auf Basis des angeführten Konstrukts interpretierbar sind, also ob *Konstruktvalidität* vorliegt (Messick, 1995). *Kriteriumsvalidität* ermöglicht im Gegensatz zu den oben genannten Aspekten durch die Korrelierung von einem sog. unabhängigen Außenkriterium mit den Testergebnissen einer Stichprobe eine Maßzahl, welche indirekt Aussagen über das gemessene Merkmal zulässt (Lienert & Ratz, 1998).

Hinsichtlich der Inhaltsvalidität erheben die Autoren des WIT-2 keinen Anspruch und berufen sich auf eine ganzheitliche Validitätsauffassung („*unifying concept of validity"*), bei der unterschiedliche Validierungsstrategien immer nur übergeordnete Validitätserkenntnisse erzeugen. (Kersting et. al., 2008). Wie oben erwähnt werden zwar im Falle des WIT-2 Validitätsstudien und Zusammenhangsstudien zur Erfassung der Konstruktvalidität herangezogen, jedoch fallen Ungereimtheiten bei näherer Betrachtung im Bereich der Strukturanalysen über konfirmatorische Faktorenanalysen sowie im Bereich der Einordnung im nomologischem Netz auf. Im Folgenden werden hier die prägnantesten

Auffälligkeiten aufgeführt. Eine Auffälligkeit spiegelt sich in der Aufnahme des Moduls *M5-Merkfähigkeit* wider bei der konfirmatorischen Faktorenanalyse der Module *M1* bis *M3* und *M5*, obwohl von den Autoren empfohlen wird, dass *M5* nicht als Indikator für Fluide Intelligenz heranzuziehen ist. Die konfirmatorische Faktorenanalyse aller Module, ausschließlich des Moduls *M4*, widerspricht in gleichem Maße durch die Aufnahme der speziellen Wissenstest (*M7* u. *M8*) der Empfehlung, dass diese nicht geeignet sind für die Abschätzung der Kristallisierten Intelligenz (siehe Kersting et. al., 2008). Bei der Einordnung in das nomologische Netz durch Korrelationen mit anderen Skalen anderer Verfahren fallen Besonderheiten auf. So korreliert das Modul *M5* mit der Skala *Verbal* (r = .49) des Berliner Intelligenzstruktur-Tests DGP (BIS-r-DGP) höher als mit der Skala *Merkfähigkeit* (r =.41). Bei der Korrelation der drei inhaltshomogenen Itembündel mit jeweils sieben Items (verbal α = .66; numerisch α = .58, figural α = .50) des Moduls *M5* mit den Analysegruppen des Lern- und Gedächtnistest (LGT-3) ergab sich die Besonderheit von einem einseitigen Korrelationswert nach Pearson von r = .12 zwischen dem Itembündel *M5-Merkfähigkeit,figural* und der Skala *Figural* des LGT-3 (siehe Kersting et. al., 2008). Dieser Wert ist nach allgemeingültiger Interpretation des Korrelationskoeffizienten Pearsons als sehr schwache Korrelation anzusehen. Die weiteren Korrelationen in diesem Fall bewegen sich zwischen r = .37 bis r = .63.

Als weiterer Punkt sei nach der Arbeit von Lienert (1969) angemerkt, dass Validität nicht nur auf der Ebene von Subtests und Testmodulen hinterfragt, sondern eine Bewertung auf Ebene der einzelnen Items vorgenommen werden sollte, also auch auf der Ebene der Trennschärfe (r_{it}) der jeweiligen Items. Die Trennschärfe bezeichnet im Kern die Fähigkeit einer Aufgabe zum Gesamtergebnis beizutragen, also wie gut das Testergebnis aufgrund der Beantwortung eines Items vorhersagbar ist (Bortz & Döring, 1995). Folgt man dieser Argumentation und betrachtet man die Trennschärfe des Moduls *M5-Merkfähigkeit* sind

folgende Angaben zu beachten. Das figurale Itembündel, bestehend aus sieben Items, liefert insgesamt eine mittlere Trennschärfe von $r_{it;figural} = .23$. Entfernt man das Item mit der höchsten Trennschärfe (Item 125, Form A, S.36; $r_{it} = .54$) beläuft sich der Wert für die mittlere Trennschärfe auf $r_{it,figural} = .18$. Das Item mit dem geringsten Wert ($r_{it} = .08$) ist Item 122 in der Parallelform A (S.36). Die anderen Itembündel weisen diesbezüglich höhere Werte auf ($r_{it;nummerisch} = .36$ und $r_{it;verbal} = .41$), das Gesamtmodul einen Wert von $r_{it;Gesamt} = .35$. Bezüglich der Einordnung dieser Werte schreibt Richardson (1936), dass bei Trennschärfen ein Wert von $r_{it} \sim .30$ als anstrebsam gelten sollte und führt weiter auf, dass Items mit einer Trennschärfe von $r_{it} = .00$ nutzlos erscheinen, da sie zwischen den Testpersonen nicht unterscheiden.

Den Grad der Genauigkeit eines Tests, mit der eine bestimmte Fähigkeit gemessen wird, beschreibt die Zuverlässigkeit oder *Reliabilität* (Lienert & Raatz, 1998). Über den Reliabilitätskoeffizienten wird erfasst, inwieweit Messwerte, sofern unter gleichen Bedingungen erhoben, bei ein und derselben Testperson übereinstimmen. Somit wird ein Maß geschaffen, welches angibt, zu welcher Intensität Testergebnisse reproduzierbar sind. Des Weiteren formulieren K. Schermelleh-Engel, A. Kelava, und H. Moosbrugger (2006), dass für psychologische Tests eine Reliabilität (Rel) von mindestens Rel = .70 notwendig ist. Diesbezüglich führt Weise (1975) auf, dass eine „geringe“ Reliabilität vorliegt, wenn $Rel \leq .80$. Zufriedenstellende Werte würden sich im Rahmen $.80 < Rel < .90$ bewegen. Erst bei $Rel \geq .90$ spricht Weise (1975) erst von einer „hohen“ Reliabilität.

Übertragen auf den WIT-2 finden sich mehrere Subtests, welche diese Bedingungen der Reliabilität nicht erfüllen *(M1a-Analogien; M1b-Gleiche Wortbedeutungen*; *M2b-Grundrechnen; M5-Merkfähigkeit).* Betrachtet man nun die Modulreliabilitäten, sog. Interne Konsistenzen (Cronbach´s α) so unterschreitet nur das Modul *M5-Merkfähigkeit*

mit α = .78, den zufriedenstellenden Reliabilitätsgrenzwert von α = .80. Alle anderen Modulreliabilitäten liegen in einem Bereich von einem Minimum α = .81 (*M8*) bis zu einem Maximum von α = .95 (*M6*). Für das Modul *M5* argumentieren die Autoren gegen die Anwendung von Maßen der internen Konsistenz bei heterogenen Iteminhalten, führen sie aber dennoch an (siehe Kersting et. al., 2008). Bei zeitlimitierter Testung führt Gittler (2009) an, dass Retestreliabilitäten (r_{tt}) strenger zu interpretieren sind und führt als zufriedenstellende Grenze $r_{tt} \geq .65$ an. In diesem Aspekt zeigt das Modul *M5-Merkfähigkeit* Auffälligkeiten. Verglichen mit den anderen Modulen/Subtests, welche in einem Retestintervall von sechs Wochen wiedervorgelegt wurden (*M7, M8, M3b-Spiegelbilder*) mit min. $r_{tt} = .73$ und max. $r_{tt} = .93$, erreicht das Modul *M5* den geringsten Wert von $r_{tt} = .67$.

Aus diesen Beobachtungen bezüglich dieser Auffälligkeiten der Test- und Itemkennwerte scheint eine nähere Betrachtung des Untertests *M5* als sinnvoll. Dieser Argumentation folgend stellen sich mehrere Fragen hinsichtlich der Ebenen des Moduldesigns und Testmaterials, welche fortfolgend adressiert werden sollen.

4 MERKFÄHIGKEIT MODUL *M5*

Das Modul *M5-Merkfähigkeit* soll die „Fähigkeit, sich kurz zuvor eingeprägte Information und Assoziationen zu merken und wiederzuerkennen“ erfassen. Dieses Modul beinhaltet 21 Items, und besteht aus insgesamt vier Phasen, die bei Testdurchführung durchlaufen werden. Instruktionsphase, Lernphase, Störphase und Reproduktionsphase. In der Instruktionsphase werden die Testpersonen auf das nachfolgende Material vorbereitet und in die Aufgabe des Einprägens eingeführt. Für diese Phase ist nach Angaben der Testautoren eine halbe Minute ausreichend. In der folgenden Lernphase, mit einer Laufzeit

von vier Minuten, sollen die Testpersonen sich eine Geschichte einprägen, welche mehrere Informationstypen (numerisch, verbal, figural) beinhaltet. Hinsichtlich der Störphase, also zwischen Lern- und Reproduktionsphase, soll nach Kersting et. al. (2008) eine „Störung" von etwa 17 Minuten liegen. So wird die Durchführung anderer Subtests empfohlen, in erster Linie der Subtest *M2b-Eingekleidete Rechenaufgaben*, da dieser eine Gesamtlaufzeit von 17 Minuten vorsieht. Die Reproduktionsphase (S.28-29) setzt sich aus einer Instruktionszeit von einer Minute und einer Laufzeit für das Abrufen der heterogen kodierten Inhalte von dreieinhalb Minuten zusammen. Die richtige Lösung ist aus sechs Antwortalternativen auszuwählen.

4.1 AUFFÄLLIGKEITEN *M5*

Was bei erster Betrachtung des Testmaterials in Lern- und Reproduktionsphase, auffällt, ist besonders der Umstand, in welcher Form die figuralen Inhalte (Piktogramme) in der Lernphase präsentiert werden sowie die Art und Weise den gelernten Inhalt letztendlich abzufragen. Zusätzlich werden alle figuralen Inhalte, bis auf das „*Logo der Personalberatung*" (Abbildung 2. des Testmaterials), in den Abbildungen 1. „*Anfahrtsskizze*" und 3. „*Lageplan des Hotels (Erdgeschoss)*", in räumlichen Bezug gesetzt. Dies geschieht erstens auf der Ebene der räumlichen Nähe, zweitens der farblichen Gruppierung (Engel & Singer, 1997; Lorenz, 1959) und drittens auf schriftlicher Ebene. So werden acht der neun Piktogramme in weißen Kästen innerhalb der Lernphase präsentiert, während die weiteren Informationen auf orangenem Hintergrund vorliegen. Die Wortwahl hinsichtlich der Instruktion zu den Abb.1 „*Anfahrtskizze*" und Abb.3 „*Lageplan des Hotel (Erdgeschoss)*" sowie die Satzteile „*Wichtige Orientierungspunkte*", „*Raumplan des Erdgeschosses*" und „*Zur besseren Orientierung*" instruiert die Testpersonen mit der Gewichtung nach einer gewissen räumlichen Relevanz. Jedoch folgt in der

Reproduktionphase (siehe Items *122, 125, 128, 131, 134, 137* u. *140*) eine graphische Detailabfrage der Piktogramme. Gittler (2009) beschreibt dies als Möglichkeit der Missinterpretation der Aufgabenstellung. Diese Dissonanz beinhaltet Problematiken hinsichtlich der Frage nach der Genauigkeit mit der *M5* die in diesem Fall zu messende Eigenschaft Merkfähigkeit misst. Die sich hieraus ergebende Frage ist demnach, ob es hier, um die Messung der Fähigkeitsdimension der *räumlichen Merkfähigkeit* (Wo befindet sich ein Piktogramm?) anhand der sieben figuralen Items, oder um die Erfassung einer *figuralen Merkfähigkeit* (Wie sieht ein Objekt/Piktogramm im Detail aus?)?

Zusätzlich finden sich weitere Aspekte in diesem Modul und seinem Testinhalt, welche Unklarheiten aufweisen. Fortfolgend sollen die betroffenen Items in Reihenfolge ihres Erscheinens im Testheft Form A berichtet werden. Item *121* erfragt die Uhrzeit des Eintreffens der sog. „*Teilnehmer*“, jedoch wird die abverlangte Information in der Lernphase mit „*Mitarbeiterinnen und Mitarbeiter*“ präsentiert. Hier kann es zu Missverständnissen kommen, welche Testpersonen in ihrem Denk- und Bearbeitungsprozesses behindern können (Sind Mitarbeiterinnen und Mitarbeiter dasselbe wie Teilnehmer?), und somit liegt prinzipiell keine korrekte Lösung für das Item *121* vor. Das Item *122* weist drei Kritikpunkte auf. So wird erstens nach dem Symbol des Restaurants im „*Lageplan des Hotels*“ gefragt, welcher aber innerhalb des Textes mit „*Raumplan des Hotels (Erdgeschoss)*“ instruiert wurde. Zweitens ist bezüglich dieser gewünschten Information, während der Lernphase keine eindeutige Zuordnung zu dem Begriff „*Restaurant*“ möglich, da auch das Piktogramm „*Küche*“, dargestellt als ein Koch, als Restaurant missverstanden werden kann und gleichzeitig keine eindeutige Beschriftung der Symbole vorliegt. Als dritter Aspekt soll auf die unterschiedlichen Darstellungsgrößen und Formen in Lern- und Reproduktionsphase hingewiesen werden. In der Lernphase ist das entsprechende Piktogramm leicht elliptisch geformt und kleiner im Vergleich zu der

auszuwählenden Antwortmöglichkeit in der Reproduktionsphase. Item *128* sowie die Items *131, 134, 137, 140* unterliegen den gleichen Aspekten wie das Item *122*. Bei Item *124* stellt sich die Frage zur Auswahl der Antwortmöglichkeiten. Antwortmöglichkeit A (1,2%) ist rein optisch sehr ähnlich mit der korrekten Antwort F (12%), was bei schnellerer Bearbeitung zu einer Fehlerquelle werden könnte. Hinsichtlich des Items *128* ist eine korrekte Zuordnung des Piktogrammes zu dem Begriff *„Bar Luna*“ erst dann möglich, wenn die Testperson weiß, dass „*Luna*“ „Mond“ bedeutet. Im Rahmen von Item *132* soll die Adresse des Hotels erfragt werden (korrekte Antwort A:„*Kirschblütenweg*“), jedoch könnte durch die Formulierung: „*Wie heißt die Straße, in der sich das Hotel befindet?*“ die falsche Antwortmöglichkeit D („*Kirschblütenstraße*“) nahe gelegt und möglicherweise unnötige Denkprozesse erzeugt werden (vgl. Item *121*). Innerhalb des Items *138* wäre der relativ weitgefasste Begriff „*Morgen*“ hinsichtlich seiner Definition zu bemängeln. Streng genommen könnte hier die Lösung „*Begrüßung und Vorstellung*“, welcher der erste Programmpunkt am Morgen ist, von einigen Testpersonen vermisst werden (vgl. Item *121*). Die korrekte Antwortmöglichkeit A („*Videokamera*“) in Item *141* liegt bei genauer Betrachtung nicht in dem Material der Lernphase vor. So ist dort nur von einer „*Kamera*“ die Rede (vgl. Item *121*).

So lässt sich für die Items *121, 122, 128, 131, 132, 134, 137, 138, 140* und *141* annehmen, dass nach Holland und Wainer (1993) sog. *Differential Item Functioning* (DIF) vorliegt. DIF beschreibt demnach, inwieweit spezielle Untergruppen innerhalb der Gesamtmenge der Testpersonen gewisse Items besser oder schlechter lösen als andere, basierend auf den Charakteristika der Untergruppe (Holland & Wainer, 1993). Beispiele hierfür wären das Geschlecht, Vorwissen oder Vorahnung über den Testablauf, Testerfahrung sowie sog. „*Test Wiseness*“ (Evans, 1984). Dies spiegelt sich auch in den Geschlechterunterschieden der jeweiligen Dimensionen des WIT-2 wieder. So berichten

Kersting et. al. (2008), dass hinsichtlich der Dimension Merkfähigkeit Frauen bessere Werte aufweisen als Männer (Cohen´s d = .47; M_{Frauen} = 13.4; $M_{Männer}$ = 11.4; SD_{Frauen} = 3.6; $SD_{Männer}$ = 4.5) ;

5 FORSCHUNGSFRAGEN

Vor diesem Hintergrund wird angenommen, dass die Auffälligkeiten und Mängel in den Testkennwerten und dem Testmaterial des Moduls M5-Merkfähigkeit des WIT-2 die Folgen von fehlerhafter Testkonstruktion darstellen. Dementsprechend und mit dem Verweis auf die Verantwortung innerhalb der psychologischen Diagnostik ist eine weitere Vorlage dieses Moduls als diagnostisches Werkzeug zu überprüfen. Insbesondere das missverständliche Zusammenspiel zwischen der Modulinstruktion, dem Material der Lernphase sowie der Form der Itemkonstruktion in der Reproduktionsphase legt die Vermutung nahe, dass die mit diesem Modul erhobenen Werte fehlerbehaftet sind und Testpersonen demnach schlechter abschneiden. In Anlehnung an die genannten Arbeiten (Eid & Petermann, 2006; Eid & Schmidt, 2014; Gittler, 2009; Kersting, 2007; Kersting, 2008; Lienert & Raatz, 1998) folgt diese Arbeit der zentralen Annahme, dass das Modul *M5-Merkfähigkeit* durch Modifikationen des Testmaterials hinsichtlich der oben genannten Kritikpunkte verbesserungsfähig ist. Basierend darauf lassen sich folgende Hypothesen ableiten:

- H1: Testwerte, die mit der modifizierten Version (Form A) des Moduls *M5-Merkfähigkeit* erhoben werden, fallen besser aus als die Testwerte des Originals (Form B).

- H2: Die Trennschärfen innerhalb der Form A fallen höher aus als die Trennschärfen der Form B.
    - H2.1: Trennschärfen des figuralen Itemsbündels in Form A weisen höhere Werte auf als die in Form B.
    - H2.2: Trennschärfen des numerischen Itemsbündels in Form A weisen höhere Werte auf als die in Form B.
    - H2.3: Trennschärfen des verbalen Itemsbündels in Form A weisen höhere Werte auf als die in Form B.
- H3. Die Reliabilität von Form A fällt höher aus als die des Originals (Form B).

6 METHODEN

6.1 ERHEBUNGSINSTRUMENTE

Als Grundlage für die Instrumente der Erhebung wurden das Modul *M5-Merkfähigkeit* und der Untertest *M2a-Eingekleidete Rechenaufgaben* des WIT-2 verwendet. Der Beschreibung der Testautoren folgend misst *M2a* die Dimension des Rechnerischen Denkens über verbal eingekleidete Rechenaufgaben. Bei der Bearbeitung sind Ziffern als Lösungen auf dem Antwortbogen anzukreuzen. Dieser Subtest wird, wie oben erwähnt, nur zur Zeitüberbrückung genutzt. Die durch *M2a* erhobenen Werte sind von peripherer Bedeutung hinsichtlich der Leithypothese, und werden nicht für die Beantwortung dieser genutzt.

6.2 UNTERSUCHUNGSPLAN

Innerhalb dieser Studie wurde ein Zwei-Gruppen-Design gewählt, in der die erste Gruppe die klassische Bedingung des Originals des Moduls *M5- Merkfähigkeit* bearbeiten sollte, wobei die zweite Gruppe eine modifizierte Version (siehe Anhang B) erhalten sollte. Als Störphase sollte der Untertest *M2a- Eingekleidete Rechenaufgaben* dienen. Damit belief sich die geplante Gesamtbearbeitungsdauer auf 32-42 Minuten.

Untersuchungsablauf		***Dauer***
1. Soziodemographische Daten	Alter, Geschlecht, Ausbildungsgrad	1 Min
2. Modul M5	Instruktion & Lernphase	4 ½ Min
3. Modul M2a	Eingekleidete Rechenaufgaben	17 Min
4. Modul M5	Instruktion & Reproduktionsphase	4 ½ Min
5. Individuelle Anmerkungen	„Haben Sie Anmerkungen zum Test?“	5 -15 Min
Gesamt		**32 – 42 Min**

Tabelle 1: Überblick über Untersuchungsablauf und den zeitlichen Rahmen

Aufgrund der nur geringen Veränderungen am Testmaterial war auch nur mit relativ geringen Mittelwertsunterschieden zwischen den jeweiligen Bedingungen zu rechnen, woraufhin die erwartete Effektgröße auf $d = .50$ festgelegt wurde. Aus der grundlegenden Leithypothese (H_1) heraus, mit einer Alpha-Fehlerwahrscheinlichkeit von $\alpha = .05$ sowie einem Orientierungswert der Power = .80 (Cohen, 1977), ergab sich in Bezug auf den Stichprobenumfang bei einem Gruppenverhältnis von 1/1 ein Wert von $N_{Gesamt} = 102$, und damit eine Gruppengröße für die jeweilige Bedingung von $N_{Form\ A/B} = 51$. Um einen Bezug zu der Normierung und des Alters hinsichtlich der empfohlenen Differenzierungsfähigkeit des WIT-2 (Kersting et. al, 2008) herzustellen, wurde das mögliche Alter zur Teilnahme auf den Bereich von 14 bis 42 Jahren beschränkt. Zusätzlich sollten nur Personen

zugelassen werden, welche sich nach eigenen Angaben vor Beginn der Testung, als physisch „*gesund*“ bezeichneten. In diesem Sinne sollten alle von den Testpersonen benötigten Sehhilfen genutzt werden. Angestrebt wurde eine Randomisierung der Zuordnung zu den jeweiligen Versuchsanordnungen mittels Urnenziehung sowie ein Bearbeitungsrahmen von 8:00 – 16:00Uhr um mögliche Einschränkungen der individuelle Leistungsfähigkeit aufgrund von Tagesformen weitestgehend zu eliminieren (Fort & Mills, 1976).

6.3 GEPLANTE UNTERSUCHUNGSDURCHFÜHRUNG

Der in beiden Formen identische Ablauf, anhand der im Manual vorgelegten Laufzeiten belief sich auf 26 Minuten und sollte strikt eingehalten werden. Zuerst sollten die Teilnehmerinnen und Teilnehmer die jeweilige Instruktion (Instruktionsphase; 30 Sekunden) durchlesen und verstehen. Dann folgte die Lernphase des Testmaterials in jeweiliger Form (4 Minuten), in der keine Notizen erlaubt waren. Der Lernphase nachgeschaltet wurde die sog. Störphase in Form der Instruktion (1 Minute) und der Bearbeitung (16 Minuten) des unmodifizierten Untertests *M2a-Eingekleidete Rechenaufgaben*. Darauf folgte die Instruktion zur Reproduktionsphase (1 Minute) der gelernten Inhalte und letztendlich die Reproduktionsphase (3 ½ Minuten). Abschließend wurden die Testpersonen hinsichtlich individueller Anmerkungen zum Testablaufes oder Testmaterials auf qualitativer Ebene befragt (verbal mit: „*Haben Sie Anmerkungen zum Test?*“). Dies diente der zusätzlichen Exploration des Testmaterials. Somit sollten Aspekte identifiziert werden, welche der Versuchsleitung möglicherweise entgangen sein könnte.

6.4 STICHPROBE

6.4.1 REKRUTIERUNG DER STICHPROBE

Die Erhebung der Stichprobe geschah im Zeitraum vom 1.3. – 3.8.2015 im Wohnraum Hamburg. Personen wurden von der Studienleitung persönlich hinsichtlich einer Studienteilnahme angesprochen. Dies beinhaltete das weite Umfeld mit Verwandten, Freunden, Bekannten und Arbeitskollegen sowie die jeweiligen Partner und Familien.

6.4.2 BESCHREIBUNG DER STICHPROBE

Die ursprüngliche Teilnehmeranzahl von 113 Personen reduzierte sich aufgrund von fehlender Durchführungsmotivation bei elf Personen, womit letztendlich ein Stichprobenumfang von N =102 erreicht wurde. Innerhalb der Bedingung mit der modifizierten Version (Form A, n = 51) von *M5* ergab sich ein Geschlechterverhältnis von 27 Frauen (52.9%) zu 24 Männern (47.1%).

Die Alterspanne reichte hier von 15 bis 42 Jahren, mit einem Mittelwert von M = 29.47; SD = 8,28. Bei der erfassten Hochschulreife zeigte sich eine Verteilung von 28 Personen ohne Hochschulreife (54.9%) zu 23 Personen mit Hochschulreife (45.1%). Die Bedingung der Originalversion von *M5* (Form B, n = 51) beinhaltete 25 Frauen (49%) zu 26 Männern (51%). Die Spanne des Alters betrug innerhalb dieser Bedingung 14 bis 42 Jahre (M = 26.43, SD = 6.89). Hinsichtlich der Hochschulreife zeigte sich eine Verteilung mit 27 Personen ohne Hochschulreife (52.9%) und 24 Personen mit Hochschulreife (47.1%).

6.5 STATISTISCHE AUSWERTUNGSVERFAHREN

Die Datenauswertung und Anwendung der statistischen Auswertungsverfahren für die vorliegende Diplomarbeit wurden mit Hilfe des Computerprogrammes SPSS 20 durchgeführt. Die formulierte Leithypothese H1 wurde anhand des Welch-Tests überprüft, da auf der Arbeit von D. Rasch, K. D. Kubinger und K. Moder (2009) von jeglicher Pre-Testung hinsichtlich der Überprüfung von Normalverteilung sowie Varianzhomogenität abgeraten wird. Des Weiteren wird der Welch-Test von den Autoren (Rasch et. al, 2009) als robusteres Verfahren angegeben im Vergleich zu den sonst üblicherweise angewandten Verfahren (Two-Sample *t* test, Wilcoxon-U test). Im Rahmen der Erfassung der Trennschärfen der jeweiligen Itembündel wurde eine Reliabilitätsanalyse durchgeführt. Dies sollte zur Beantwortung der Hypothesen 2 und 3 dienen. Als Orientierungswerte wurden folgende Werte nach einschlägiger Literatur festgelegt (Weise, 1976; Schermelleh-Engel et. al., 2006):

- ✓ Itemtrennschärfe mit $r_{it} \sim .30$
- ✓ Cronbach´s α mit $Rel \leq .80$

Zusätzlich wurden die im Testmanual angegebenen Werte mitangeführt. Auf eine direkte Gegenüberstellung der Werte des Manuals ($N_{Manual} = 2234$) mit den erhobenen Werten innerhalb dieser Studie ($N_{Form\ A} = 51$), verzichtete die Versuchsleitung da die Untersuchungsgruppen deutliche Größenunterschiede aufweist.

7 ERGEBNISSE

Im Folgenden wird eine ausführliche Darstellung der Ergebnisse der Studie angeführt, sowie die gleichzeitige Beantwortung der Leithypothese und den weiteren formulierten Hypothesen.

7.1 TESTERGEBNISSE DES MODULS *M5-MERKFÄHIGKEIT*

Bei dem Vergleich zwischen den jeweiligen Gesamtergebnissen der unveränderten Version (Form B; M = 1.69, SD = 3.108) und der modifizierten Version (Form A; M = 12.35, SD = 3.393) des Untertests *M5*, zeigte der durchgeführte Welch-Test (t = 6.691, p = .011, df = 99.240), dass die Testergebnisse in Form A signifikant höher ausfielen als die erreichten Ergebnisse in der Originalversion.

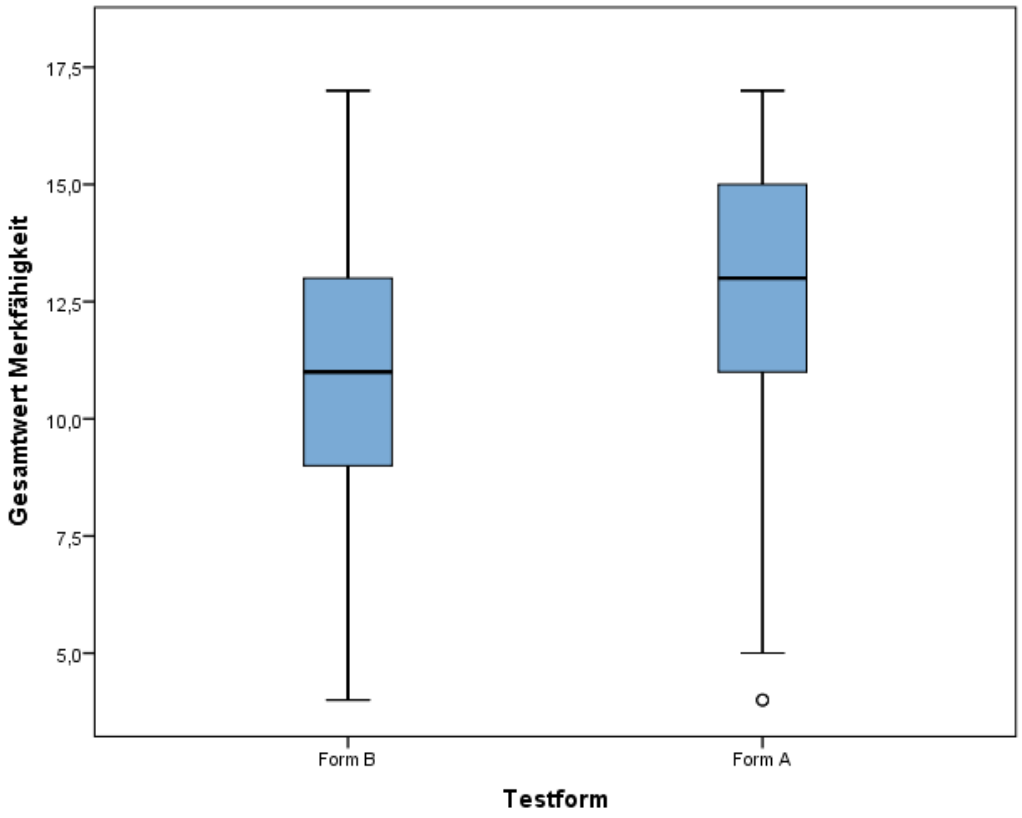

Abbildung 1: Box-Whisker-Plot. Ausreißer als Kreise dargestellt.

Der Unterschied der Mittelwerte war signifikant (p = .05) und erreichte nicht das durch die festgelegte Effektstärke von d = .50 inhaltlich bedeutsame Niveau und verfehlte dies mit einem Wert von SD = .03. Diese Abweichung von der festgelegten Grenze war in ihrem Ausmaß gering und somit sah die Versuchsleitung die Annahme der Leithypothese bestätigt.

Diesen Ergebnissen ergänzend gegenübergestellt ließ sich kein Unterschied zwischen den jeweiligen Bedingungen im Aufgabenbereich der *eingekleideten Rechenaufgaben* feststellen. Beide Gruppen erhielten die jeweilig unveränderte entsprechende Form für diesen Abschnitt. Für Form B (M = 8.18, SD = 3.167) und Form A (M = 9.49, SD = 3.711) ergab die Durchführung des Welch-Tests (t = 3.698, p = .057, df = 97.582) demnach keinen signifikanten Unterschied (p = .05) hinsichtlich der erreichten Werte im Subtest *Eingekleidete Rechenaufgaben*. Dieser Umstand wird im Folgekapitel 8.2 zur Ergebnisinterpretation genauer eruiert.

7.2 TRENNSCHÄRFENANALYSE

Die Durchführung der Trennschärfenanalyse sollte eine genauere Einsicht in die Art und Weise der Verbesserung liefern, welche durch die Modifikation des Testmaterials und insbesondere der Modulinstruktion erreicht wurde. Wie in der H2 angenommen, war eine Verbesserung der gesamten mittleren Trennschärfen zu verzeichnen. So fand sich eine höhere gesamte mittlere Trennschärfe bei Form A von r_A = .256 im Vergleich zur Form B mit r_B = .107. Jedoch lag dieser Wert nur knapp unter der als zufriedenstellenden Grenze von r_{Gesamt} = .30.

	Figural	Nummerisch	Verbal	Gesamte mittlere Trennschärfe
FORM B	0,054	0,196	0,071	0,107
FORM A	**0,305**	**0,246**	**0,216**	**0,256**
MANUAL	0,234	0,364	0,407	0,35

Tabelle 2: Trennschärfen der jeweiligen Formen sowie dem Manual. Fett markiert sind erhöhte Werte gegen über der Originalversion.

Die nähere Betrachtung der Itembündel und der Trennschärfen auf dem Niveau der spezifischen Items erbrachte ein diffenzierteres Bild über die genauen Probleme der Itemkonstruktion sowie der Itemauswahl. In erster Linie ließ sich eine Verbesserung der Itemtrennschärfen in Form A in 15 von insgesamt 21 Fällen feststellen.

Innerhalb des Bündels der figuralen Items konnten in sechs von sieben Fällen höhere Trennschärfen für die Items *122, 125, 128, 131, 134* und *137* aus Form A berichtet werden. Die Items *122, 125, 128,* und *137* befanden sich im zufriedenstellenden Bereich (r_{min} = .38, r_{max} = .46). Die Items *125, 131* und *133* erreichten in Form B leicht negative Trennschärfen.

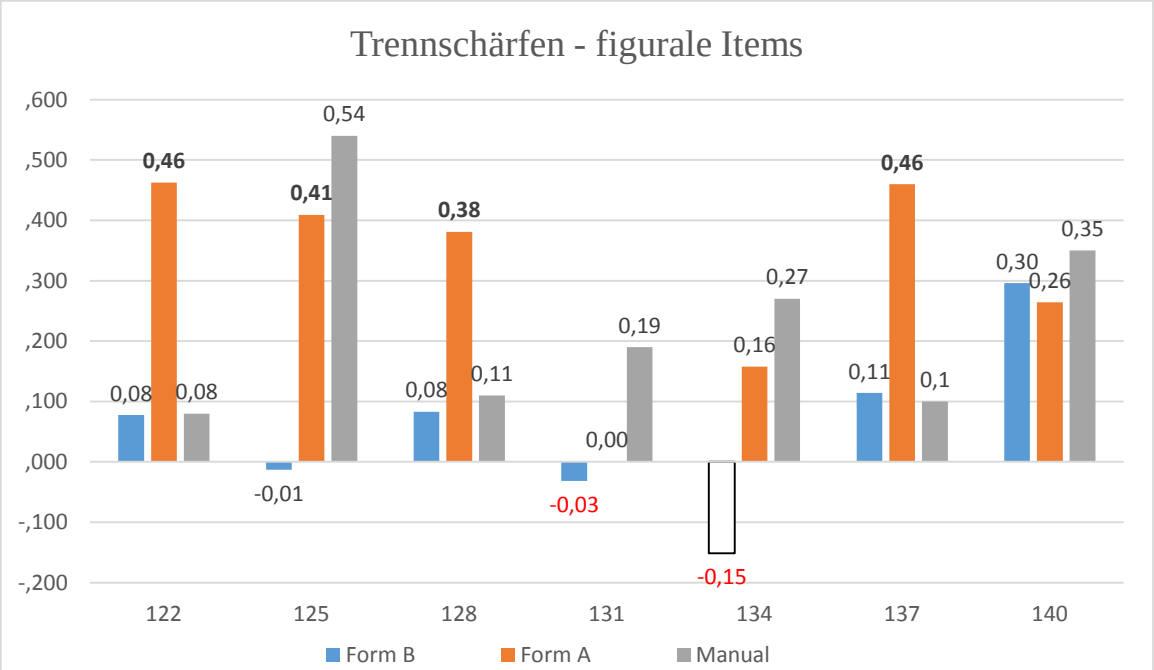

Abbildung 2: Trennschärfen des figuralen Itembündels. Fett markiert sind Trennschärfen im zufriedenstellenden Bereich. Rot markiert sind negative Trennschärfen.

Die Trennschärfen innerhalb des numerischen Itembündels wiesen bei vier von sieben Items der Form A höhere Werte auf, wovon die drei Items *124*, *136* und *139* (r_{min} = .33, r_{max} = .51) in den erwünschten Bereich fielen. Die Items *127* und *130* kennzeichneten sich in Form A mit deutlich schwächeren Trennschärfen gegenüber der Form B.

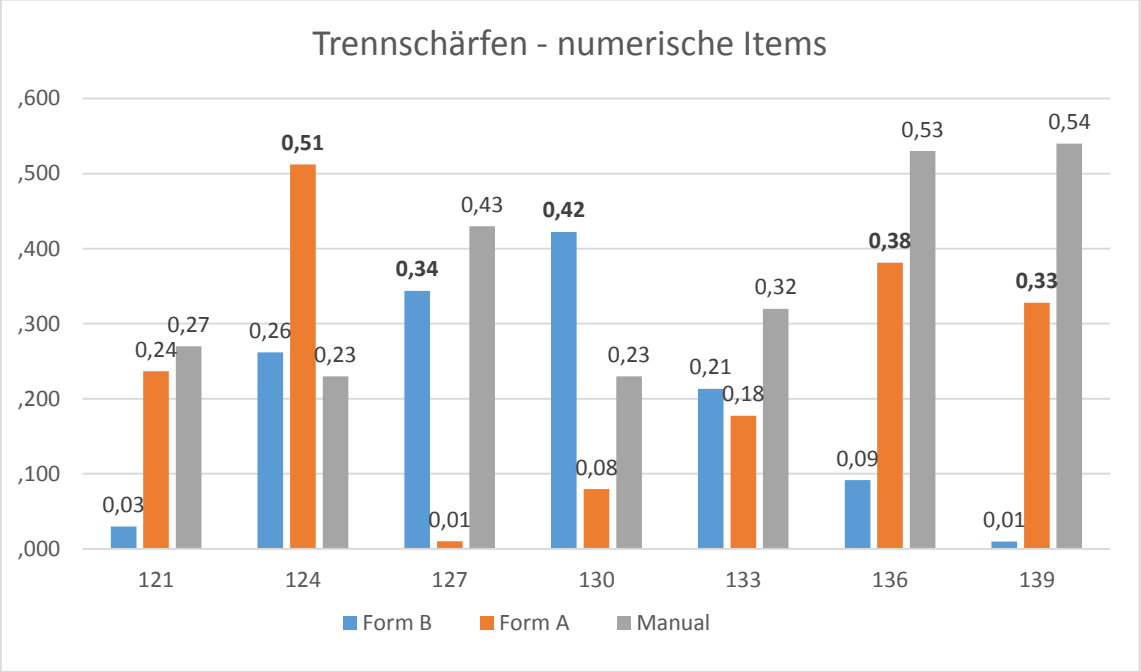

Abbildung 3: Trennschärfen des numerischen Itembündels. Fett markiert sind Trennschärfen im zufriedenstellenden Bereich. Rot markiert sind negative Trennschärfen

Bei den Items *123*, *129*, *135*, *138* und *141* aus dem verbalen Itembündel für Form A konnten, in fünf von sieben Fällen, höhere Werte der Trennschärfe erfasst werden. Der anstrebsame Bereich für Trennschärfen erreichten die Items *123*, *129* und *138* (r_{min} = .32, r_{max} = .44), nur das Item *132* aus Form B erreichte ebenfalls diesen Bereich. Innerhalb Form B fielen die Trennschärfen für die Items *123* und *135* leicht negativ und für das Item *141* deutlich negativ aus. Für Form A erreichte Item *132* eine leicht negative Trennschärfe.

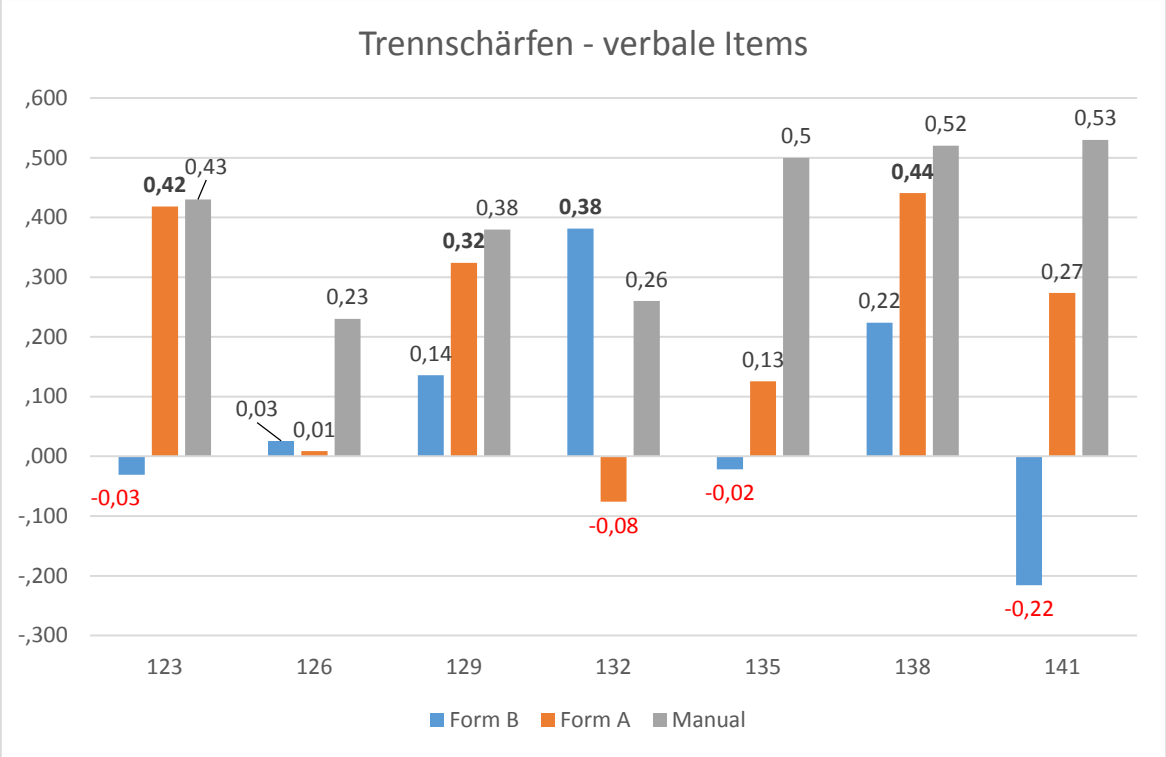

Abbildung 4: Trennschärfen des verbalen Itembündels. Fett markiert sind Trennschärfen im zufriedenstellenden Bereich. Rot markiert sind negative Trennschärfen.

7.3 RELIABILITÄTSANALYSE

Über die Durchführung einer Reliabilitätsanalyse wurde die interne Konsistenz des Moduls *M5-Merfähigkeit* geschätzt, hier angegeben als Cronbach`s α. Dieses Vorgehen orientiert sich an der Vorgehensweise der Testautoren um einen Vergleich mit den im Manual zu findenden Kennwerten zu schaffen. Aufgrund der Ergebnisse bezüglich der Trennschärfen der jeweiligen Items, wurden die Reliabilität der Itembündel sowie die des Gesamtmoduls errechnet. Des Weiteren wurden die Reliabilitätswerte für die jeweilige Skala (figural, numerisch, verbal) erfasst, welche durch die mögliche Eliminierung der jeweiligen Items erreicht werden können.

Form A wies eine deutlich höhere Reliabilität bei den figuralen und verbalen Itembündeln, und nur leicht erhöhte Werte im Bündel der numerischen Items auf. Auch

ließ sich ein leicht erhöhter Wert bei Form A mit Rel_A = .62 gegenüber einem internen Konsistenzwert der Form B von Rel_B = .57 verzeichnen.

	Figural	Verbal	Numerisch	Gesamt
Form B	.113	.170	.412	.57
Form A	**.552**	**.446**	**.492**	**.62**
Manual	k.A.	k.A.	k.A.	.78

Tabelle 3: Gesamtreliabilitäten der Itembündel. Fett markiert sind höhere Reliabilitäten gegenüber der Originalform. Manualangaben sind unvollständig für die jeweiligen Itembündel,

Bei Betrachtung der Reliabilitätswerte auf Itemniveau ergab sich folgendes Bild. Alle einzelne Items überblickend waren nur sehr geringe Verbesserungen der Reliabilität samtlicher Itembündel bei denkbarer Auslassung der spezifischen Items zu berichten. Auffällig im figuralen Itembündel für Form B waren die Items *125*, *131* und *134*. Durch Entfernen der einzelnen Items würde eine leichte Verbesserung der Gesamtreliabilität (Rel_B = .113) auf minimal Rel_B = .138 (Item *125*) oder maximal Rel_B = .269 (Item *134*) erreicht. Für Form A (Rel_A = .552) lieferte das mögliche Entfernen der Items *131* und *134* ebenfalls eine Verbesserung mit einem Erreichen des Reliabilitätswert von minimal Rel_A = .561 oder maximal Rel_A = .624.

Figural	122	125	128	131	134	137	140	Gesamt
Form B	.068	**.138**	.065	**.165**	**.269**	.032	-.110	.113
Form A	.436	.512	.468	**.624**	**.561**	.431	.519	.552

Tabelle 4: Cronbach's α für figurales Itembündel bei Entfernen des Items. Fettmarkiert sind mögliche Erhöhungen der Reliabilität

Für die Items *123*, *126*, *135* und *141* aus dem verbalen Itembündel für Form B (Rel_B = .170) ließ sich auch eine potenzielle Verbesserung mit einem minimal erlangten Wert von Rel_B = .178 (Item *126*) oder einem Maximum von Rel_B = .227 (Item *123*) festhalten. Bei Form A betraf dies die Items *126*, *132* und *135*, mit einem Minimum von

Rel_A = .447 oder einer maximal zu erreichenden Reliabilität durch Entfernen des Items *132* von Rel_A = .535.

Verbal	123	126	129	132	135	138	141	Gesamt
Form B	**.227**	**.178**	.088	-.145	**.215**	.020	**.323**	.170
Form A	.296	**.500**	.348	**.535**	**.447**	.286	.375	.446

Tabelle 5: Cronbach's α für verbales Itembündel bei Entfernen des Items. Fettmarkiert sind mögliche Erhöhungen der Reliabilität.

Ähnlich verhielten sich die Items *121*, *136* und *139* bei Form B aus dem numerischen Itembündel, mit einer geringen Erhöhung der Gesamtreliabilität (Rel_B = .412) auf minimal Rel_B = .425 (Item *136*) oder einem maximal realisierbarem Wert von Rel_B = .466 (Item 139). Für Form A (Rel_A = .492) fanden sich mögliche Verbesserungen der Reliabilität durch Auslassung der Items *127 (*Rel_A = .547) und *130* (Rel_A = .522).

Numerisch	121	124	127	130	133	136	139	Gesamt
Form B	**.456**	.340	.300	.229	.359	**.425**	**.466**	.412
Form A	.452	.332	**.547**	**.522**	.480	.389	.409	.492

Tabelle 6: Cronbach's α für numerisches Itembündel bei Entfernen des jeweiligen Items. Fettmarkiert sind mögliche Erhöhungen der Reliabilität.

7.4 Anmerkungen der Testpersonen

Abschließend nach Durchführung der jeweiligen Formen wurden die Probanden befragt, ob diese noch persönliche Anmerkungen hätten, hinsichtlich des Ablaufes, des vorgelegten Materials oder spezieller Items. Eine engere Auswahl dieser Anmerkungen soll hier exemplarische Erwähnung finden. Nennenswerte Äußerungen bezüglich des Testmaterials waren folgende:

- Form B
    - „Hätte ich gewusst, dass ich mir diese Bildchen im Detail merken muss dann hätte ich mir die genauer angeschaut.“
    - „Dieses Logo war am einfachsten zu merken, das stand ja auch alleine.“ (bezogen auf Item *125*).
    - „Was war das für ein rundes Ding in der Mitte des Raumplanes? Das hat mich furchtbar verwirrt.“
    - „Ich hätte es besser gefunden, wenn ich einzeichnen hätte müssen wo die Orte sind als wie sie genau aussehen. Wer merkt sich denn sowas?“
    - „Die Rechenaufgaben waren teilweise schon recht schwierig.“
    - „[…] es war doch eine Kamera, keine Videokamera oder? Ich wusste zuerst nicht was ich ankreuzen sollte, denn ganz genau genommen war die richtige Antwort ja nicht da.“ (bezogen auf Item *141*).
- Form A
    - „Namen kann ich mir einfach nicht merken. Egal ob ich sie höre oder geschrieben sehe.“ (bezogen auf Item *123*).
    - „Das Logo war im Vergleich zu den anderen Bildern zu einfach. Ich habe es mir nur einmal angeguckt“ (bezogen auf Item *125*).
    - „Gewisse Symbole haben eine klare gesellschaftliche Bedeutung, da ist es egal wie sie genau im Detail aussehen […]. Da schaut man schnell drüber und denkt nicht weiter drüber nach.“ (bezogen auf Item *134*).

Die hier enthaltenen Kommentare unterstreichen die Annahmen bezüglich der figuralen Items und des Weiteren stützen sie die weiter oben erwähnten Ergebnisse für die einzelnen Items.

8 Diskussion und Interpretation der Ergebnisse

Ziel dieser Diplomarbeit war es die Frage zu beantworten, ob eine Verbesserung der Einsatzfähigkeit des Moduls *M5-Merkfähigkeit* durch die spezifische Modifikation des Testmaterials möglich ist. Es wurde also vermutet, dass dieses Modul in seiner derzeitigen Form Unzulänglichkeiten und Auffälligkeiten aufweist, welche von einem Einsatz in der Praxis abraten würden. Dies wurde überprüft und entsprechende Anpassungen wurden in folgender Form vorgenommen: Hinsichtlich der Testkennwerte wurde besonderes Augenmerk auf die Veränderung der Instruktion des Moduls gelegt. Zusätzlich wurde die Abfrage und die Inhalte der Items betrachtet sowie eine geringfügige Anpassung und Ergänzung der Items untersucht. Im Folgenden werden die formulierten Hypothesen und Annahmen beantwortet und es wird, eine Auswertung der Ergebnisse und ihre Einordnung in den theoretischen Rahmen vorgenommen.

8.1 Verbesserung durch Modifikation

Die Leithypothese H1, dass eine modifizierte Instruktion bessere Testergebnisse liefere gegenüber des Originals, wurde anhand der vorliegenden Daten dieser Studie bestätigt. Personen in der veränderten Bedingung (M = 12.35, SD = 3.393) erreichten durchschnittlich einen besseren Score als Personen der klassischen Bedingung (M = 10.69, SD = 3.108).

Die Durchführung dieser Studie unterlag spezifischen Kriterien, um einen Vergleich zwischen Versuchsgruppe und Kontrollgruppe zu ermöglichen. Folgende Faktoren können die erreichten mittleren Scores für die jeweiligen Bedingungen beeinflussen:

Nach Reischies & Lindenberger (2010) spielt das Alter beim Lösen von Merkfähigkeitsaufgaben eine Rolle. Ältere Personen schneiden durchschnittlich schlechter ab als jüngere Personen. Bezüglich dieses Faktors lässt sich innerhalb der vorliegenden Stichprobe für beide Bedingungen jeweils eine ähnliche Verteilung beobachten (siehe Stichprobenbeschreibung). Der Umstand, dass innerhalb von Form B ältere Personen vorliegen, und somit eine schlechtere mittlere Leistung für diese Form erzeugen, liegt demnach nicht vor. Auch der gegensätzliche Umstand mit einer deutlich jüngeren Population in Form A ist nicht zu verzeichnen. Der Faktor des Alters ist, dieser Argumentation zufolge, kontrolliert und moduliert somit nicht die gefundenen Ergebnisse.

Entsprechend ist der Faktor des Geschlechts von entscheidender Rolle. So formulieren auf der einen Seite Lepach, Reimers, Pauls, Petermann & Daseking (2015) und auf der anderen Seite die Testautoren (Kersting et. al., 2008), dass Frauen durchschnittlich leicht höhere Leistungen in Merkfähigkeitsaufgaben erbringen als Männer. Innerhalb der beobachteten Stichprobe ist dieser Effekt nicht zu beobachten. Weder gibt es eine ungleiche Verteilung in den jeweiligen Formen, noch sichtbare Leistungsunterschiede in der Merkfähigkeit welche auf den Faktor des Geschlechts zurückzuführen sind. In den jeweiligen Bedingungen lösen Frauen und Männer die Aufgaben im Mittel gleich gut, für Form A werden jedoch insgesamt bessere Scores erreicht.

Da die Testautoren den Faktor der Hochschulreife als Differenzierungsmerkmal in die Normierung des WIT-2 einfließen ließen, wurde hier demzufolge auch dieser Aspekt

als möglicher Faktor betrachtet, welcher Merkfähigkeitsleistungen potentiell beeinflusst. Nach den Ergebnissen sowie dem vorliegenden Verhältnis innerhalb der betrachteten Stichprobe liegt auch hier eine durchwegs gleichartige Verteilung von diesem Faktor vor. Weder Form A noch Form B weisen eine ungleiche Mengen- sowie Leistungsverteilung hinsichtlich des Faktors Hochschulreife auf.

Weitere Aspekte, welche das beobachtete Ergebnis der Verbesserung in den erreichten Scores hinsichtlich der Merkfähigkeit modulieren oder beeinflussen könnten, wären unteranderem: Zeit der Erhebung (Tageszeit), Reihenfolge der Items, Erfahrung mit Intelligenztests (sog. „Testwiseness“). Hinsichtlich der Tageszeit und die damit verbundene Leistungs- und Konzentrationsfähigkeit der Probanden wurde während der Durchführung das zeitliche Fenster von 8:00 bis 16:00 eingehalten, jedoch wurde dies nicht für jede Testdurchführung festgehalten beziehungsweise nicht statistisch überprüft. Bezüglich des in der Literatur wiederkehrenden Themas der Itemreihenfolge und des Einflusses auf die jeweilige Lösung der Items (Liernert, & Raatz, 1998; Weise, 1975,), wurden die vorliegenden Quasi-parallel-Formen der Testautoren verwendet, um mögliche Effekte zu eliminieren. Nach Kersting et. al. (2008) sind diese Parallelformen als größtenteils identisch zu betrachten.

Testwiseness (Evans, 1984) als weiterer einflussnehmender Faktor ist aus Sicht der Versuchsleitung im Großen und Ganzen zu vernachlässigen, weil es sich in erster Linie bei den Versuchspersonen nicht um Psychologen oder Psychologiestudenten handelte, sondern um Personen mit geringerer Testexpertise und somit geringerer bis nicht vorhandener Testwiseness. Eine Leistungsverbesserung auf Grund individueller Testwiseness sollte in erster Linie durch die Stichprobengröße sowie durch die randomisierte Zuordnung zu den jeweiligen Bedingungen eliminiert worden sein.

Ein weiteres wichtiges Argument, welches dafür spricht, dass die Modifikation der Instruktion eine Verbesserung erbrachte, sind die Ergebnisse aus dem Aufgabenbereich der *eingekleideten Rechenaufgaben.* Hier gibt es keine statistisch signifikanten Unterschiede zwischen Form A und Form B. Weder Personen aus der klassischen Bedingung noch Personen aus der modifizierten Bedingung erbringen höhere mittlere Leistungen gegenüber Personen der jeweiligen anderen Gruppe. Betrachtet man nun diese Daten als eine Quasikontrollgruppe oder genauer als Indikator für ein allgemeines kognitives Leistungsniveau für die jeweiligen Gruppen, so lässt sich annehmen, dass beide Personengruppen sich hinsichtlich dieses allgemeineren kognitiven Leistungsniveaus nicht unterscheiden. Liegt also in beiden Gruppen ein ähnliches kognitives Leistungsniveau vor, so erbringt Form A eine präzisere Darstellung des Fähigkeitsparameters *Merkfähigkeit* und würde somit eine höhere Reliabilität in sich tragen.

Es kann also angenommen werden, dass eine mögliche Missinterpretation der Instruktion und das damit verbundene schlechtere Abschneiden in diesem Modul vorliegt (siehe Gittler, 2009). Demnach ist der Hauptgrund, warum Personen bessere Ergebnisse in Form A erzielen auf das genauere Verständnis der zu bearbeitenden Aufgabe und die gleichzeitige bessere Vorbereitung zurückzuführen. Dieses genauere Verständnis wurde durch eine spezifischere und präzisere Instruktion vermittelt.

8.2 ITEMIDENTIFIKATION

Die Annahme einer Verbesserung der Itemtrennschärfen konnte ebenfalls bestätigt werden. Es konnten höhere Trennschärfen für die modifizierte Form A gegenüber der Originalform B verzeichnet werden. Die Ergebnisse zeigten eine deutliche Verbesserung auf Niveau der jeweiligen Itembündel (figural, numerisch, verbal) sowie der gesamten mittleren Trennschärfe des Moduls. Im Vergleich zu den Angaben der Testautoren aus dem

Manual des WIT-2 zeichnete sich nur die mittlere Trennschärfe des figuralen Itembündels höher ab und lag damit am unteren Ende des zufriedenstellenden Bereichs. Alle anderen mittleren Trennschärfen lagen diesem Vergleich folgend unter den im Manual angegebenen Werten. Demnach werden die H2 und ihre Unterhypothesen als bestätigt angesehen.

Bei genauerer Betrachtung ergab sich jedoch ein differenziertes Bild. Wie gut ein Item eine Skala widerspiegelt, welche aus den weiteren oder restlichen Items gebildet wird, ist als Trennschärfe bezeichnet (Liernert & Raatz, 1998). In diesem Sinne sind hohe Werte wünschenswert. Die erfassten Trennschärfen für beide Formen lagen größtenteils nicht über dem zufriedenstellenden Grenzwert von $r_{it} > .30$. Um ungeeignete Items zu identifizieren wird das Kriterium angelegt, dass in Form A und in Form B kein Erreichen der genannten Grenze festzustellen ist.

Die geringsten Werte erreichten die figuralen Items der Form B. Drei der sieben Items zeigten negative Trennschärfen auf (Items *125*, *131, 134*), womit im engeren Sinne die Inhaltsvalidität verletzt ist. Eine nähere Betrachtung der Inter-Item-Korrelations-Matrix des figuralen Itembündels für Form B unterstreicht diese Auffälligkeit, da dort hauptsächlich negative oder sehr geringe Korrelationen aufzufinden sind (in 18 von 21 Fällen). Alle weiteren Itemtrennschärfen (Items *122, 128, 137*) des figuralen Itembündels in Form B liegen unter dem empfohlenen Grenzwert. Das Item *140* verfehlte mit $r_{B140} = .296$ diesen Bereich ebenfalls knapp. Jedoch muss diese vergleichsweise hohe Trennschärfe kritisch betrachtet werden da sich Item *140* relativ am Ende der Bearbeitungsreihenfolge befindet und dies dazu führen kann, dass langsamere oder weniger leistungsstarke Personen es nicht schaffen dieses Item zu bearbeiten. Dadurch erhöht sich indirekt die Trennschärfe für dieses Item. Für die figuralen Items der Form A ergab die Trennschärfenanalyse zwar gegenüber der Form B in sechs der sieben Fällen höhere Trennschärfen (Item *140*, $r_{A140} =$

.264), jedoch erreichten die Items *131* und *134* auch hier nicht die untere Grenze. Die Items *122, 125, 128* und *137* zeichneten sich durch zufriedenstellende Trennschärfen aus.

Für das numerische Itembündel aus Form B liegen die Trennschärfen der Items *121, 124, 133, 136* und *139* unter dem gewünschten Grenzwert. Unzureichende Trennschärfen liegen in Form A für die Items *121, 127, 130* und *133* vor.

Negative Trennschärfen lagen erneut für Form B im Bereich des verbalen Itembündels für die Items *123, 135* und *141* vor. Nur das Item *132* erreichte den zufriedenstellenden Bereich. Für Form A zeigte sich für dieses Item nun wiederum eine leicht negative und damit unzureichende Trennschärfe. Ungenügende Trennschärfe fanden sich zusätzlich für die Items *126* und *135*, sowie ein knappes unterschreiten der festgelegten Grenze bei Item *141* mit $r_{A141} = .273$.

Auf Basis dieser Kennwerte in beiden Formen wird eine Entfernung oder eine Rekonstruktion der Items *131* und *134* (figural), *121* und *133* (numerisch), *126* und *135* (verbal) aus Sicht der Versuchsleitung als notwendig betrachtet. Besonders im figuralen sowie im verbalen Teil liegen deutliche Mängel hinsichtlich der Itemtrennschärfe vor. Da bei diesen Items weder in der Originalform noch in der hier vorliegenden modifizierten Version eine Verbesserung der Kennwerte festzustellen war, sind es diese Items, welche einer genaueren weiteren Betrachtung hinsichtlich ihrer Konstruktion und somit ihres Inhaltes und ihrer Formulierung benötigen. Auf Grund der hier erhobenen Ergebnisse muss vor einer mögliche Fehlerquelle und Ungenauigkeit durch das Verwenden dieses Moduls in der Praxis gewarnt werden.

8.3 RELIABILITÄT

Die Ergebnisse der Analyse der internen Konsistenz folgten den oben erwähnten Funden. Auch die H3 wurde bestätigt. Es zeigten sich für die modifizierte Form in jedwedem Itembündel sowie für das gesamte Modul höhere Reliabilitäten im Vergleich zu den erfassten Kennwerten der Originalform. Jedoch blieb der Wert der internen Konsistenz auch hier für alle Itembündel unter den Manualangaben. Alle erfassten Kennwerte blieben unter der zufriedenstellenden Grenze von Rel ≤ .80. Da die interne Konsistenz in direkter Abhängigkeit zu der Trennschärfe der individuellen Items steht, lässt sich dies durch die aufgezeigten Mängel in diesem Bereich erklären. Des Weiteren steht die Frage nach der Itemhomogenität erneut im Raum. Durch das strukturelle Design des untersuchten Moduls mit den drei unterschiedlichen Itembündel, ist mit einer niedrigeren Modulreliabilität zu rechnen (Bühner, 2006). Gleichzeitig werden die Angaben der Testautoren bezüglich der internen Konsistenz als fraglich angesehen, da auch die Erfassung dieser Daten nur schwierig nachzuvollziehen ist. Aufgrund dieser Begebenheit wird die Analyse der Reliabilitäten auf Ebene der jeweiligen Items als Indikator und vielmehr als Wegweiser verstanden und interpretiert. In diesem Sinne soll deutliches Augenmerk auf die Items *131* und *134* aus dem figuralen Itembündel sowie auf die Items *126* und *135* auf dem verbalen Itembündel gelegt werden. Würde nun eine jeweilige Eliminierung beziehungsweise ein jeweiliges Entfernen dieser Items stattfinden, so käme es in Form A sowie in Form B zu einer Verbesserung der internen Konsistenz für die entsprechenden Itembündel. Dies unterstreicht die gefundenen Ergebnisse auf der Ebene der Trennschärfen. Erneut wird empfohlen besonders diese Items zu bearbeiten, da eine reine Anpassung der Instruktion für das Modul *M5-Merkfähigkeit* nicht ausreicht, um diese Items „brauchbarer“ für den Einsatz in der Praxis zu gestalten.

Es wird also abgeraten das Modul *M5-Merkfähigkeit* in seiner jetzigen Form vorzulegen. Die Instruktion dieses Modul sollten die hier angewandten Modifikationen übernehmen und integrieren, um als diagnostisches Werkzeug präziser die gemessene Dimension der Merkfähigkeit abzubilden.

9 LIMITATIONEN UND AUSBLICK

9.1 LIMITATIONEN DER STUDIE

Eine der Schwierigkeiten während der Durchführung dieser Studie war in erster Linie die Rekrutierung von Testpersonen, aufgrund des Umstandes, dass scheinbar mehrere Personen eine negative Einstellung gegenüber Rechenaufgaben aufweisen. Das reine Erwähnen des notwendigen Bearbeitens von eingekleideten Rechenaufgaben ließ eine Vielzahl an Personen nicht teilnehmen, da auch gleichzeitig keine monetäre Vergütung aus Kostengründen stattfinden konnte. Des Weiteren bleibt für die Durchführung anzumerken, dass es weitere Leistungsschwankungen geben kann, wenn Testpersonen mehrere Module und Subtests des WIT-2 in einer Sitzung bearbeiten. Dadurch spiegelt diese Untersuchung nur die alleinige Vorlage der Module *M5-Merkfähigkeit* und *M2a-eingekleidete Rechenaufgaben* wider.

Bezüglich der Teilnahme wurde versucht eine möglichst breite Auswahl an Personen zu rekrutieren. Alle Teilnehmer wurden persönlich vom Versuchsleiter angesprochen, ob sie teilnehmen möchten. Darunter litt zu einem gewissen Grad die Repräsentativität der untersuchten Population, da die Selektion in diesem Sinne das soziale Geflecht des Versuchsleiters graduell abbildet. Dieser Effekt konnte vermutlich nicht gänzlich umgangen werden, da fremde Personen nur selten Motivation zur Bearbeitung und somit zur Teilnahme mitbrachten.

Hinsichtlich des WIT-2, insbesondere der Konstruktion der verwendeten Items in Modul *M5-Merkfähigkeit*, blieben mehrere Fragen offen. So waren dem Manual zwar teilweise Testkennwerte für diese spezifischen Items zu entnehmen, jedoch war der genaue Prozess der Itemkonstruktion nicht im Manual beschrieben und dadurch schwer nachzuvollziehen. Da diese Diplomarbeit sich im Kern mit einer näheren Betrachtung und der damit verbundenen Überarbeitung des Testmaterials befasst, wäre genau der Prozess der Konstruktion des Moduls und seiner Items von besonderem Interesse. Aus Sichtweise der Versuchsleitung sind bei der Testkonstruktion Fehler geschehen, welche von Anfang an hätten vermieden werden können. Diese Arbeit stützt sich hauptsächlich auf die Ergebnisse aus dem Vergleich zwischen den vorgelegten Formen A und B. Ein Vergleich mit den im Manual beschriebenen Daten ist nur eingeschränkt möglich und somit kann auf die Daten der deutlich größeren Population der Modulkonstruktion nur im beschränkten Umfang zurückgegriffen werden. Dementsprechend konnten für die Instruktion sowie die jeweiligen Items nur geringfügige Anpassungen vorgenommen werden, um gleichzeitig eine Vergleichbarkeit mit der Originalform zu gewährleisten und aufrechtzuerhalten.

9.2 AUSBLICK

Basierend auf den Ergebnissen dieser Studie lässt sich unterstreichen, dass eine sorgfältige Testkonstruktion von entscheidender Wichtigkeit für die Werkzeuge der Diagnostik ist. Weiterhin wäre eine stärkere Veränderung des Moduls bis hin zu einer kompletten Neukonstruktion empfehlenswert, um zusätzliche Erkenntnisse zur Modulkonstruktion zu gewinnen. Die Inhalte der Items des Moduls *Merkfähigkeit* bleiben aus Sicht der Versuchsleitung kritisch zu betrachten. Inwieweit zum Beispiel eine Verwendung von Piktogrammen, welche eine gesellschaftlich-kulturelle Bedeutung oder Symbolik in sich tragen, als sinnvoll zu erachten ist, bleibt offen. Besonders beim Erfassen

von figuraler Merkfähigkeit sollten möglichst unbekannte und abstrakte Symbole verwendet werden. Eine räumliche Zugehörigkeit bei figuralen Inhalten sollte für zukünftige Abwandlungen dieses Moduls vermieden werden, da sonst unklar ist, ob die Fähigkeit des „räumlichen Merkens“ oder die Fähigkeit des „figuralen Merkens“ erfasst wird. Eine weitere aussichtsreiche Modulation wäre eine Item- und Modulkonstruktion, welche stärker dem Rasch-Modell und somit der probabilistischen Testtheorie folgt anstatt der klassischen Testtheorie.

10. LITERATURVERZEICHNIS

Blum, F., Didi, H. J., Fay, E., Maichle, U., Trost, G., Wahlen, J. H., & Gittler, G. (1998). *Intelligenz Struktur Analyse: Ein Test zur Messung der Intelligenz.* Frankfurt: Swets Test Services.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion.* München: Pearson Studium.

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

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences.* New York: Acadamic Press.

Cronbach, L. J. (1970). *Essentials of Psychological testing* (3 Ausg.). New York: Harper International Edition.

Eid, M., & Petermann, F. (2006). Aufgaben, Zielsetzungen und Strategien der psychologischen Diagnostik. In F. Petermann, & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 15-25). Göttingen: Hogrefe.

Eid, M., & Schmidt, K. (2014). *Testtheorie und Testkonstruktion.* Göttingen: Hogrefe.

Engel, A. K., & Singer, W. (1997). Neuronale Grundlagen der Gestaltwahrnehmung. *Spektrum der Wissenschaft*(Dossier: Kopf oder Computer), S. 66-73.

Evans, W. (1984). Test wiseness: An examination of cue using strategies. *Journal of Experimental Education, 52*(3), S. 141-144.

Eysenck, H. J., Brengelmann, J. C., & Fulker, D. W. (1980). *Intelligenz: Struktur und Messung.* Springer.

Fort, A., & Mills, J. N. (1976). Der Einfluss der Tageszeit und des vorhergehenden Schlaf-Wach-Musters auf die Leistungsfähigkeit unmittelbar nach dem Aufstehen. In *Biologische Rhytmen und Arbeit* (S. 59-64). Wien: Springer.

Gittler, G. (2009). *WIT-2 Testbeurteilung nach TBS-TK.* Unveröffentliches Manusskript.

Holland, P. W., & Wainer, H. (Frühling 1993). Differential Item Functioning. *Journal of Educational Measurement*, S. 88-92.

Kersting, M. (2007). Wenn Tests in die Jahre kommen. Probleme des Einsatzes überalterter Testverfahren. In C. Lorei (Hrsg.), *Polizei und Psychologie* (S. 565-577). Frankfurt: Verlag für Polizeiwissenschaft.

Kersting, M. (2008). *Qualität in der Diagnostik und Personalauswahl: Der DIN Ansatz.* Göttingen: Hogrefe.

Kersting, M., Althoff, K., & Jäger, A. O. (2008). WIT-2. Wilde-Intelligenz-Test 2. Göttingen: Hogrefe.

Kersting, M., Häcker, H., & Hornke, L. (2011). Qualitätsstandards in der Diagnostik. In F. Hornke, M. Amelang, & M. Kersting (Hrsg.), *Grundlagen und Anwendungsfelderpsychologischer Diagnostik* (Enzyklopädie der Psychologie, Serie: Psychologische Diagnostik Ausg., Bd. I, S. 1-86). Göttingen: Hogrefe.

Kubinger, K. (2009). *Psychologische Diagnostik Theorie und Praxis psychologischen Diagnostizierens.* Göttingen: Hogrefe.

Lepach, A. C., Reimers, W., Pauls, F., Petermann, F., & Daseking, M. (10. März 2015). Geschlechtseffekte bei Intelligenz- und Gedächtnisleistungen. *Zeitschrift für Neuropsychologie*(26), S. 5-16. doi:10.1024/1016-264X/a000144

Lienert, G. A. (1969). *Testaufbau und Testanalyse* (3 Ausg.). Weinheim: Beltz.

Liernert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse* (6 Ausg.). Weinheim: Beltz.

Lorenz, K. (1959). Gestaltwahrnehmung als Quelle wissenschaftlicher Erkenntnis. *Zeitschrift für Experimentelle und Angewandte Psychologie*.

Maiden, I. (2003). Rainmaker. Auf *Dance of Death*. London, England: S. Harris, & K. Shirley.

Messick, S. (1989). Validity. In R. L. Linn (Hrsg.), *Educational measurement* (3 Ausg., S. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inference from persons` responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, S. 741-749. doi:10.1037/003-066X.50.9.741

Petermann, F. (1. September 2006). Intelligenzdiagnostik. *Kindheit und Entwicklung, 15*(2), S. 71-75. doi:10.1026/0942-5403.15.2.71

Rasch, D., Kubinger, K. D., & Moder, K. (28. April 2011). The two-sample t test: pre-testing its assumptions does not pay off. *Statistical Papers, 52*(1), S. 219–231. doi:DOI 10.1007/s00362-009-0224-x

Reischies, F. M., & Lindenberger, U. (2010). Grenzen und Potentiale kognitiver Leistungsfähigkeit im Alter. In *Die Berliner Altersstudie* (S. 375-401). Akademie Verlag.

Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika, 1*(33??).

Schermelleh-Engel, K., Kelava, A., & Moosbrugger, H. (2006). Gütekriterien. In F. Petermann, & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 420-433). Göttingen: Hogrefe.

Schnell, R., Hill, P. B., & Esser, E. (2008). *Methoden der Empirischen Sozialforschung* (8. Ausg.). R. Oldenburg: München.

Weise, G. (1975). *Psychologische Leistungstests.* Göttingen: Hogrefe.

11 TABELLENVERZEICHNIS

12 ANHANG

A. VERFAHRENSMATERIAL : BEGRÜßUNG UND INSTRUKTION ZUR STUDIENTEILNAHME

Vielen Dank für Ihre Teilnahme an dieser Untersuchung. Diese Studie dient einer Diplomarbeit von der psychologischen Fakultät der Universität Wien. Sie sollen gleich zwei verschiedene Aufgaben bearbeiten.

- In der ersten Aufgabe bekommen Sie eine Geschichte vorgelegt, welche Sie sich einprägen sollen. Später werden Ihnen Fragen zu dieser Geschichte gestellt
- Die zweite Aufgabe beinhaltet eingekleidete Rechenaufgaben, welche gelöst werden sollen.

Bitte nehmen Sie sich jetzt das Aufgabenheft zur Hand, blättern Sie bitte nicht um und warten Sie auf die Instruktionen des Versuchsleiters.

Vielen Dank!

B. Form A Instruktion

WIT-2 FA	**MF – Merkfähigkeit – I E**	Seite 27

Bitte legen Sie die Stifte jetzt aus der Hand.

Ihnen wird jetzt eine Geschichte samt Abbildungen vorgelegt. Sie sollen die Informationen aus der Geschichte und den Abbildungen so weit *lernen,* dass Sie sich später an alle Einzelheiten zuverlässig erinnern können. Prägen Sie sich also alles möglichst genau ein, jedes Detail ist von Bedeutung. Lesen Sie die Geschichte bitte zunächst *ganz* durch. Dann fangen Sie wieder am Anfang an und merken sich möglichst alle Einzelheiten, wie z.B. :Namen, Symbole, Objekte, Zeitangaben, etc.

Für das Einprägen haben Sie *vier Minuten Zeit.* Das reicht nach unseren Erfahrungen aus. Sie können sich also unbesorgt Ihrer Aufgabe zuwenden. Sie werden informiert, wenn die Zeit um ist.

Sie dürfen sich bei dieser Aufgabe keinerlei Notizen machen.

Gibt es noch Fragen?

STOPP! Bitte erst nach Aufforderung umblättern!

C. ZUSAMMENFASSUNG

Diese Diplomarbeit befasst sich mit den Testwerten des Moduls *M5-Merkfähigkeit* des Wilde-Intelligenz-Test-2 (WIT-2). Untersucht werden die interne Konsistenz und die Trennschärfen des Gesamtmoduls, der jeweiligen Itembündel sowie der spezifischen Items. Die spezifische Gestaltung des Originalmoduls deutet darauf hin, dass es zu Ungenauigkeiten der Erfassung der Merkfähigkeit kommen kann. Deshalb wird eine alternative Version des Testmoduls erstellt und empirisch überprüft, ob durch Modifikationen der Modulinstruktion eine präzisere Messung durchgeführt werden kann. Es wird gezeigt, dass die Testpersonen statistisch signifikant im Durchschnitt bessere Leistungen in der neuerstellten Version gegenüber der Originalversion erbringen. Zusätzlich kann durchgängig eine Erhöhung der mittleren Trennschärfen und der internen Konsistenz auf Ebene der jeweiligen Itembündel identifiziert werden. Auch auf Ebene des Gesamtmoduls lassen sich ähnliche Ergebnisse feststellen.

D. ABSTRACT

This thesis deals with the test values of the module *M5-Memory* of the Wilde Intelligence Test 2 (WIT-2). The study examines the internal consistency and discriminatory power of the entire module, the respective item-clusters and the specific items. The specific design of the original module suggests that there may be inaccuracies in the measurement of memory. Therefore, an alternative version of the test module is created and it is empirically checked whether a more precise measurement can be achieved by modification of the instruction of the module. Results indicate that the subjects perform on average better in the modified version compared to the original version. In addition, an increase in average discriminatory power and internal consistency at the level of respective item-clusters is identified. Furthermore, similar results are obtained at the level of the entire module itself.

E. LEBENSLAUF

Persönliche Daten:

Name:	Roman Göttner
Anschrift:	Taubnesselweg 14
	22549 Hamburg
Telefon:	0049/4080030622
E-Mail:	goettner.roman@web.de
Geburtsdatum und -ort:	11.03.1988 in Hamburg

Ausbildung:

09/1994 – 07/1998	Grundschule Klein-Flottbeker-Weg, Deutschland
09/1998 – 05/2007	Gymnasium Hochrad, Deutschland
seit 10/2008	Diplomstudiengang Psychologie an der Universität Wien

Sprachkenntnisse:

Deutsch	Muttersprache
Englisch	fließend in Wort und Schrift
Spanisch	sehr gute Kenntnisse in Wort und Schrift
Rumänisch	Grundkenntnisse
Italienisch	Grundkenntnisse
Russisch	Grundkenntnisse