Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

Vergleich des Modularen Kurzintelligenztests M-KIT mit dem IST-Screening: Leistung und Akzeptanz

MASTER-THESIS

Eingereicht per: 01/2026

Autorin
Sarah Müller

Betreuungsperson
Michael Dantlgraber

Praxispartner
Michael Dantlgraber

Zeichenzahl: 143’070

Zusammenfassung

Intelligenztests werden zunehmend digital und in ökonomischen Kurzformaten eingesetzt. Vor diesem Hintergrund untersucht die vorliegende Arbeit, inwiefern sich zwei Online-Intelligenztestverfahren unterscheiden. Verglichen wird das IST-Screening und das Modul A des Modularen Kurzintelligenztests (M-KIT) hinsichtlich Testleistung, subjektiver Akzeptanz sowie möglicher Geschlechtsunterschiede. Beide Verfahren sind Kurzversionen und wurden unter vergleichbaren Bedingungen im unbeaufsichtigten Online-Setting durchgeführt.

Die Studie folgte einem quantitativen Within-Subjects-Design mit einer Stichprobe von N = 59 Personen. Alle Teilnehmenden bearbeiteten beide Intelligenztests sowie jeweils einen Akzeptanzfragebogen. Zusätzlich wurde die Testreihenfolge variiert, um mögliche Reihenfolgeeffekte zu berücksichtigen. Die Auswertung erfolgte mittels Korrelationsanalysen, t-Tests und explorativer Zusatzanalysen.

Die Ergebnisse zeigen einen starken positiven Zusammenhang zwischen den Leistungen im IST-Screening und im M-KIT, was darauf hindeutet, dass beide Verfahren einen gemeinsamen Kern kognitiver Leistungsfähigkeit erfassen. Geschlechtsunterschiede in der Testleistung fielen insgesamt gering aus und erreichten kein statistisches Signifikanzniveau. Altersbezogene Effekte zeigten sich explorativ, insbesondere beim M-KIT. Die subjektive Akzeptanz der beiden Testverfahren war insgesamt vergleichbar, mit einem leichten Vorteil des M-KIT in einzelnen Akzeptanzdimensionen. Reihenfolgeeffekte konnten ausgeschlossen werden.

Die Ergebnisse unterstreichen die grundsätzliche Eignung beider Kurztests für den Online-Einsatz und verdeutlichen zugleich die Bedeutung einer differenzierten Interpretation im Hinblick auf Testkonstruktion, Fairness und Akzeptanz.

Abstract

Intelligence tests are increasingly administered in digital formats and in economically efficient short versions. Against this background, the present study examines the extent to which two online intelligence test procedures differ. The IST-Screening and Module A of the Modular Short Intelligence Test (M-KIT) are compared with regard to test performance, subjective acceptance, and potential gender differences. Both instruments are short forms and were administered under comparable conditions in an unsupervised online setting.

The study followed a quantitative within-subjects design with a sample of N = 59 participants. All participants completed both intelligence tests as well as a corresponding acceptance questionnaire for each test. In addition, the order of test administration was varied in order to control for possible order effects. Data analysis was conducted using correlation analyses, t-tests, and exploratory additional analyses.

The results indicate a strong positive association between performance on the IST-Screening and the M-KIT, suggesting that both instruments assess a common core of cognitive ability. Gender differences in test performance were generally small and did not reach statistical significance. Age-related effects emerged exploratorily, particularly for the M-KIT. Overall, subjective acceptance of the two test procedures was comparable, with a slight advantage for the M-KIT in certain acceptance dimensions. No order effects were found.

These findings underscore the general suitability of both short tests for online use, while simultaneously highlighting the importance of a differentiated interpretation with regard to test construction, fairness, and acceptance.

Inhaltsverzeichnis

1 Einleitung

Intelligenztests zählen zu den zentralen Instrumenten der psychologischen Diagnostik und werden in einer Vielzahl von Anwendungsfeldern eingesetzt, darunter Bildung, Berufsberatung, klinische Psychologie und Personalpsychologie. Empirische Forschung zeigt seit Jahrzehnten, dass kognitive Fähigkeiten in engem Zusammenhang mit schulischem Erfolg, beruflicher Leistung und allgemeiner Problemlösekompetenz stehen (Carroll, 1993; Neisser et al., 1996). Entsprechend hoch ist die praktische Relevanz standardisierter Intelligenztests für diagnostische und selektive Entscheidungen.

Gleichzeitig sind Intelligenztests Gegenstand anhaltender wissenschaftlicher und gesellschaftlicher Diskussionen. Neben klassischen psychometrischen Gütekriterien wie Objektivität, Reliabilität und Validität rücken zunehmend Fragen der Fairness, der Geschlechtsneutralität und der subjektiven Akzeptanz durch die Testpersonen in den Fokus (Kersting, 1998; Moosbrugger & Kelava, 2020). Diese Aspekte gewinnen insbesondere vor dem Hintergrund einer erhöhten Sensibilisierung für Chancengleichheit sowie der breiten Anwendung von Testverfahren in hochrelevanten Entscheidungskontexten an Bedeutung.

Parallel dazu hat sich die Intelligenzdiagnostik in den vergangenen Jahren stark verändert. Während Intelligenztests traditionell als Papier-Bleistift-Verfahren in kontrollierten Präsenzsettings durchgeführt wurden, kommen heute zunehmend digitale und ortsunabhängige Online-Testverfahren zum Einsatz (Goldberg, 2018). Diese Entwicklung eröffnet neue Möglichkeiten, wie etwa eine hohe Standardisierung, flexible Einsatzformen und automatisierte Auswertungen. Gleichzeitig bringt sie neue Herausforderungen mit sich, da Online-Testungen unter weniger kontrollierten Bedingungen stattfinden und stärker von Benutzerfreundlichkeit, Instruktionsklarheit und subjektiver Wahrnehmung abhängen (Goldberg, 2018; Kersting, 1998).

In diesem Zusammenhang rückt die Akzeptanz von Testverfahren zunehmend in den Mittelpunkt der Forschung. Studien zeigen, dass die subjektive Bewertung eines Tests – etwa hinsichtlich Fairness, Verständlichkeit oder Belastung – die Motivation der Testpersonen und die Qualität der Bearbeitung beeinflussen kann (Kersting, 1998; Schuler, 2014). Insbesondere im Online-Setting, in dem soziale Rückmeldung und direkte Unterstützung durch Testleitende weitgehend fehlen, kann eine geringe Akzeptanz zu oberflächlicher Bearbeitung oder erhöhten Abbruchraten führen (Goldberg, 2018).

Ein weiterer zentraler Aspekt betrifft mögliche geschlechtsbezogene Leistungsunterschiede in Intelligenztests. Die empirische Literatur zeigt übereinstimmend, dass sich Frauen und Männer in der allgemeinen Intelligenz nicht systematisch unterscheiden (Hyde, 2005; Neisser et al., 1996). In bestimmten Aufgabenformaten lassen sich jedoch gelegentlich geringe Leistungsunterschiede beobachten, die häufig auf Aufgabenstruktur, Lerngelegenheiten oder soziale Faktoren zurückgeführt werden (Hyde, 2005; Matlin & Halpern, 1988). Für die Testentwicklung ergibt sich daraus der Anspruch, Aufgabenformate möglichst konstruktzentriert und frei von geschlechtsbezogenen Verzerrungen zu gestalten (Kersting, 1998; Messick, 1989).

Diese Entwicklungen spiegeln sich in unterschiedlichen Testtraditionen wider. Klassische Verfahren wie das IST-Screening basieren auf einer langen Entwicklungsgeschichte und kombinieren sprachliche, numerische und figurale Aufgabenformate (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Moderne Verfahren wie der Modulare Kurzintelligenztest (M-KIT) verfolgen hingegen einen konstruktzentrierten Ansatz mit Fokus auf fluide Intelligenz, reduzierter Sprachabhängigkeit und einem expliziten Anspruch auf Geschlechterfairness sowie hohe Akzeptanz (Dantlgraber, 2015). Die Entwickler:innen des M-KIT vertreten die Position, dass durch diese Konstruktionsprinzipien geschlechtsbezogene Verzerrungen minimiert werden können.

Für die Langversion des M-KIT liegen bereits empirische Befunde vor, die geringe Geschlechtsunterschiede und gute psychometrische Eigenschaften nahelegen (Dantlgraber, 2015). Unklar ist bislang, ob sich dieser Anspruch auch für die Kurzversion des Tests im Online-Setting bestätigt. Gerade bei Kurzversionen stellt sich aus testtheoretischer Sicht die Frage, ob zentrale Qualitätsmerkmale – insbesondere Fairness und Akzeptanz – in gleichem Masse gewährleistet sind wie bei umfangreichen Testversionen. Zudem fehlen bislang direkte Vergleichsstudien zwischen modernen, digital optimierten Kurztests und etablierten klassischen Verfahren unter vergleichbaren Bedingungen.

Die vorliegende Arbeit setzt an dieser Forschungslücke an. Ziel ist es, das IST-Screening und den Modularen Kurzintelligenztest (M-KIT) als zwei unterschiedliche Online-Intelligenztestverfahren systematisch miteinander zu vergleichen. Im Fokus stehen dabei die erfasste Testleistung, die subjektive Akzeptanz der Verfahren sowie mögliche geschlechtsbezogene Leistungsunterschiede. Ergänzend wird geprüft, ob die Reihenfolge der Testdurchführung einen Einfluss auf die Ergebnisse hat, um methodische Verzerrungen auszuschliessen.

Vor diesem Hintergrund werden im Anschluss die Zielsetzung der Studie, die zentrale Forschungsfrage sowie die daraus abgeleiteten Hypothesen dargestellt.

1.1 Zielsetzung

Die vorliegende Studie verfolgt das Ziel, das IST-Screening und das Modul A des Modularen Kurzintelligenztests (M-KIT) als zwei unterschiedliche Online-Intelligenztestverfahren systematisch zu vergleichen. Im Zentrum stehen dabei zwei Aspekte:

(1) die kognitive Testleistung der Teilnehmenden sowie
(2) die subjektive Akzeptanz der beiden Verfahren.

Untersucht wird insbesondere, ob sich zwischen den beiden Tests Unterschiede in der Leistung und der wahrgenommenen Akzeptanz zeigen und inwieweit mögliche geschlechtsbezogene Leistungsunterschiede zwischen den Verfahren variieren. Vor dem Hintergrund der konstruktzentrierten und explizit geschlechterfairen Entwicklung des M-KIT wird geprüft, ob dieses Verfahren im digitalen Kontext geringere Geschlechtsunterschiede und eine höhere Akzeptanz aufweist als das IST-Screening.

Ergänzend wird analysiert, ob die Reihenfolge der Testdurchführung einen Einfluss auf die Leistungsergebnisse hat. Diese Analyse dient der methodischen Absicherung der Untersuchung und soll sicherstellen, dass beobachtete Unterschiede nicht auf Ermüdungs- oder Übungseffekte zurückzuführen sind.

Die Studie leistet damit einen Beitrag zum Verständnis, inwieweit moderne, digital optimierte Intelligenztests den Anforderungen an diagnostische Qualität, Fairness und Akzeptanz im Online-Setting gerecht werden.

1.2 Forschungsfrage

Aus den beschriebenen Überlegungen ergibt sich die folgende Forschungsfrage:

Inwiefern unterscheiden sich die Akzeptanz und die Leistungsergebnisse des M-KIT (Modul A) und des IST-Screenings, insbesondere im Hinblick auf mögliche Geschlechtsunterschiede?

Diese Forschungsfrage bildet den konzeptionellen Rahmen der Untersuchung und bestimmt die Ableitung der Hypothesen.

1.3 Hypothesen

Basierend auf theoretischen Annahmen, Manualangaben und bisherigen Forschungsergebnissen werden die folgenden Hypothesen formuliert:

H1 – Geschlechtsunterschiede in der Testleistung

Es wird erwartet, dass sich geschlechtsspezifische Unterschiede in den Testergebnissen zeigen. Aufgrund der ausgewogenen und geschlechterfairen Konstruktion sollten Geschlechtsunterschiede im M-KIT geringer ausfallen als im IST-Screening.

H2 – Akzeptanzunterschiede zwischen den Testverfahren

Der M-KIT wird in der Gesamtakzeptanz sowie in den Dimensionen Verständlichkeit, Fairness und Belastung signifikant höher bewertet als das IST-Screening.

H3 – Reihenfolgeeffekte (methodische Kontrollhypothese)

Die Reihenfolge der Testdurchführung hat keinen bedeutsamen Einfluss auf die Leistungsergebnisse. Diese Hypothese dient der Absicherung der internen Validität des Forschungsdesigns und soll sicherstellen, dass mögliche Unterschiede nicht durch Ermüdung oder Übungseffekte verursacht werden.

2 Theoretischer Hintergrund

2.1 Intelligenz als psychologisches Konstrukt

Intelligenz gehört zu den zentralen Konstrukten der psychologischen Forschung. Trotz ihrer hohen praktischen Bedeutung besteht jedoch bis heute keine einheitliche Definition, die alle theoretischen Perspektiven gleichermassen abdeckt. Vielmehr existieren unterschiedliche Ansätze, die Intelligenz entweder als allgemeine kognitive Fähigkeit, als Zusammenspiel spezifischer Einzelfähigkeiten oder als funktionale Anpassungsleistung an neue Anforderungen verstehen (Neisser et al., 1996; Sternberg, 2021).

Ein früher und bis heute einflussreicher Ansatz ist die Annahme einer allgemeinen Intelligenz, des sogenannten g-Faktors. Dieser geht davon aus, dass Leistungen in unterschiedlichen kognitiven Aufgabenbereichen positiv miteinander korrelieren und durch eine gemeinsame zugrunde liegende Fähigkeit erklärt werden können (Spearman, 1904). Empirische Studien konnten diesen Generalfaktor wiederholt mithilfe faktorenanalytischer Verfahren nachweisen (Carroll, 1993). Gleichzeitig wird kritisch diskutiert, inwieweit ein solches eindimensionales Konzept der Vielfalt individueller Leistungsprofile gerecht wird.

Als Reaktion auf diese Kritik entwickelten sich mehrdimensionale und hierarchische Modelle der Intelligenz. Besonders etabliert ist das Drei-Schichten-Modell von Carroll (1993), das spezifische Einzelfähigkeiten, breitere Fähigkeitsdimensionen sowie einen übergeordneten Generalfaktor unterscheidet. Dieses Modell wurde später mit der Theorie fluider und kristallisierter Intelligenz (Cattell, 1963) zum Cattell-Horn-Carroll (CHC)-Modell integriert (McGrew, 2009). Das CHC-Modell gilt heute als eines der umfassendsten und empirisch am besten fundierten Rahmenmodelle kognitiver Fähigkeiten (McGrew, 2009).

Neben strukturellen Modellen existieren funktionale Definitionen von Intelligenz, die stärker deren praktische Bedeutung betonen (Neisser et al., 1996). Neisser et al. (1996) beschreiben Intelligenz als Fähigkeit, aus Erfahrung zu lernen, Probleme zu lösen und sich an neue oder veränderte Anforderungen anzupassen. Diese Perspektive rückt weniger die formale Struktur kognitiver Fähigkeiten in den Vordergrund als vielmehr deren Nutzen in realen Lebens- und Entscheidungssituationen.

Unabhängig von diesen theoretischen Differenzen besteht in der psychologischen Diagnostik weitgehender Konsens darüber, dass Intelligenz ein latentes Konstrukt darstellt, das nicht direkt beobachtbar ist. Stattdessen wird es über standardisierte Testverfahren operationalisiert. Die Aussagekraft solcher Verfahren hängt entscheidend davon ab, inwieweit sie das intendierte Konstrukt valide, reliabel und fair erfassen. Diese Anforderungen sind insbesondere dann relevant, wenn Testergebnisse als Grundlage für bedeutsame Entscheidungen dienen, etwa im Bildungs- oder Personalbereich (Messick, 1989; Schuler, 2014).

Im Alltag wird Intelligenz häufig mit schulischer oder beruflicher Leistung gleichgesetzt. Eine solche Gleichsetzung greift jedoch zu kurz, da Leistung stets auch von Motivation, Interessen, sozialen Rahmenbedingungen und situativen Faktoren beeinflusst wird (Neisser et al., 1996; Schuler, 2014). Intelligenztests zielen demgegenüber darauf ab, das individuelle Leistungspotenzial unter möglichst standardisierten Bedingungen zu erfassen. Gerade diese Trennung von Potenzial und realisierter Leistung bildet eine zentrale Grundlage für den diagnostischen Einsatz von Intelligenztests und erklärt ihre anhaltende Bedeutung trotz wiederkehrender Kritik.

Für die vorliegende Arbeit ist diese theoretische Einordnung insofern relevant, als sie den Rahmen für den Vergleich zweier unterschiedlicher Intelligenztestverfahren bildet. Sowohl das IST-Screening als auch der Modulare Kurzintelligenztest (M-KIT) verfolgen das Ziel, kognitive Leistungsfähigkeit abzubilden, unterscheiden sich jedoch hinsichtlich ihres theoretischen Fokus und ihrer testmethodischen Umsetzung. Diese Unterschiede sind für die Interpretation der empirischen Ergebnisse von zentraler Bedeutung und werden in den folgenden Abschnitten näher betrachtet.

2.2 Fluide Intelligenz als Kern moderner Testdiagnostik

Die Unterscheidung zwischen fluider und kristalliner Intelligenz stellt einen wichtigen Bezugspunkt in der modernen Intelligenzforschung dar. Sie geht ursprünglich auf Cattell zurück und wurde später

durch Horn weiter ausgearbeitet (Cattell, 1963; Horn & Cattell, 1966). Fluide Intelligenz bezeichnet die Fähigkeit, neuartige Probleme zu analysieren und zu lösen, ohne auf erlerntes Wissen oder spezifische Erfahrungen zurückzugreifen. Kristalline Intelligenz hingegen umfasst Wissen und Fertigkeiten, die im Verlauf des Lebens erworben werden und stark von Bildung sowie kulturellem Kontext geprägt sind (Jensen, 1998; McGrew, 2009).

Fluide Intelligenz zeigt sich insbesondere in Aufgaben, die abstraktes Schlussfolgern, Mustererkennung oder logisches Denken erfordern. Typische Aufgabenformate sind figural-analoge Schlussfolgerungen, Matrizen oder serielle Fortsetzungen. Da diese Aufgaben weitgehend ohne sprachliche Inhalte oder schulisches Vorwissen auskommen, gelten sie als weniger abhängig von Bildungsstand oder Sprachkompetenz als kristalline Aufgaben (Carroll, 1993; McGrew, 2009).

Aus diagnostischer Sicht wird fluide Intelligenz daher häufig als besonders geeignet angesehen, um vergleichbare Leistungsmassstäbe über unterschiedliche Personengruppen hinweg zu erfassen (McGrew, 2009; Schuler, 2014). Testverfahren mit stark kristallinen Anteilen sind anfälliger für Einflüsse von Bildungsbiografien, kulturellen Erfahrungen oder sprachlichen Voraussetzungen (Jensen, 1998). Fluide Aufgaben zielen dagegen stärker auf grundlegende kognitive Prozesse ab und gelten als weniger verzerrungsanfällig (Kersting, 2008).

Gleichzeitig ist fluide Intelligenz nicht frei von Einflüssen externer Faktoren. Auch bei darauf bezogenen Aufgaben spielen Aspekte wie Motivation, Konzentration oder Vertrautheit mit dem Aufgabenformat eine Rolle (Kersting, 2008). Zudem können bei figuralen Aufgaben visuelle Anforderungen oder räumliches Vorstellungsvermögen selbst leistungsrelevant sein (McGrew, 2009). Dennoch wird fluide Intelligenz im Vergleich zu kristallinen Fähigkeiten insgesamt als robuster gegenüber systematischen Verzerrungen betrachtet (McGrew, 2009).

Ein weiterer Aspekt betrifft altersbezogene Unterschiede. Während fluide Intelligenz typischerweise im jungen Erwachsenenalter ihren Höhepunkt erreicht und im weiteren Lebensverlauf tendenziell abnimmt, bleibt kristalline Intelligenz häufig stabil oder nimmt weiter zu (Cattell, 1963; Horn & Cattell, 1966). Diese unterschiedlichen Entwicklungsverläufe sind insbesondere für Studien mit einer breiten Altersstreuung relevant, da sie die Interpretation von Testergebnissen beeinflussen können (McGrew, 2009). In Online-Studien mit heterogenen Stichproben kommt diesem Aspekt eine besondere Bedeutung zu.

Im Kontext digitaler Testungen ist die Fokussierung auf fluide Intelligenz noch relevanter. Online-Testverfahren werden häufig unter weniger kontrollierten Bedingungen bearbeitet, etwa in unterschiedlichen Umgebungen oder mit variierender technischer Ausstattung (Goldberg, 2018). Verfahren, die stark auf Sprachverständnis oder Vorwissen angewiesen sind, reagieren auf solche Unterschiede besonders sensibel (Schuler, 2014). Fluide Testformate gelten demgegenüber als besser geeignet für heterogene Online-Settings (McGrew, 2009; Schuler, 2014).

Der Modulare Kurzintelligenztest (M-KIT) knüpft explizit an dieses theoretische Verständnis an (Dantlgraber, 2015). Er wurde mit dem Ziel entwickelt, fluide Intelligenz ökonomisch und konstruktzentriert zu erfassen und dabei mögliche Verzerrungsquellen zu reduzieren (Dantlgraber, 2015). Die Fokussierung auf Aufgabenformate mit besonders reduziertem Vorwissen stellt eine bewusste konzeptionelle Entscheidung dar und bildet eine zentrale Grundlage für den Vergleich mit dem IST-Screening in der vorliegenden Studie.

2.3 Psychometrische Testgütekriterien

Die Qualität psychologischer Testverfahren wird traditionell anhand der klassischen Hauptgütekriterien Objektivität, Reliabilität und Validität beurteilt (Moosbrugger & Kelava, 2020; Schuler, 2014). Diese Kriterien bilden den normativen Rahmen der psychologischen Diagnostik und sind Voraussetzung dafür, dass Testergebnisse sinnvoll interpretiert und für diagnostische Entscheidungen herangezogen werden können (Moosbrugger & Kelava, 2020).

Objektivität beschreibt, inwieweit Testergebnisse unabhängig von der Person sind, die den Test durchführt, auswertet oder interpretiert (Moosbrugger & Kelava, 2020). In digitalen Testverfahren ist die Durchführungs- und Auswertungsobjektivität in der Regel hoch, da Instruktionen standardisiert präsentiert und Ergebnisse automatisiert berechnet werden (Goldberg, 2018).

Interpretationsobjektivität hängt hingegen weiterhin von klaren Normen und transparenten Auswertungsvorgaben ab (Schuler, 2014).

Reliabilität bezeichnet die Zuverlässigkeit eines Tests, also den Grad, zu dem ein Verfahren frei von zufälligen Messfehlern ist (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2020). Ein reliabler Test liefert bei wiederholter Messung unter vergleichbaren Bedingungen ähnliche Ergebnisse. In der Intelligenzdiagnostik werden häufig interne Konsistenzen oder Retest-Reliabilitäten herangezogen, um die Messgenauigkeit zu beurteilen (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2020). Hohe Reliabilitätswerte gelten als notwendige, jedoch nicht hinreichende Voraussetzung für valide Messungen (Messick, 1989).

Validität beschreibt, inwieweit ein Test tatsächlich das misst, was er zu messen vorgibt (Messick, 1989). Für Intelligenztests sind insbesondere die Konstruktvalidität (Passung zur theoretischen Annahme von Intelligenz), die Kriteriumsvalidität (Zusammenhang mit externen Leistungsindikatoren) sowie die Inhaltsvalidität (angemessene Abdeckung des Zielkonstrukts) relevant (Moosbrugger & Kelava, 2020). Moderne Testentwicklungen orientieren sich häufig an etablierten Intelligenzmodellen wie dem CHC-Modell, um eine klare theoretische Verankerung sicherzustellen (McGrew, 2009).

Neben diesen klassischen Hauptgütekriterien gewinnt das Konzept der Fairness als Nebengütekriterium zunehmend an Bedeutung (Kersting, 1998, 2008). Ein Test gilt als unfair, wenn systematische Leistungsunterschiede zwischen Gruppen auftreten, die nicht auf tatsächliche Unterschiede im gemessenen Konstrukt zurückzuführen sind (Kersting, 1998). Solche Verzerrungen können unter anderem durch sprachliche Anforderungen, kulturell geprägte Inhalte oder teststrategische Vorteile bestimmter Gruppen entstehen (Schuler, 2014). Fairness wird daher als integraler Bestandteil diagnostischer Qualität verstanden.

Im Kontext digitaler Intelligenzdiagnostik kommen weitere Aspekte hinzu. Die Gestaltung der Benutzeroberfläche, die Verständlichkeit der Instruktionen und die subjektive Belastung können das Testverhalten beeinflussen und damit indirekt die Aussagekraft der Ergebnisse verändern (Kersting, 2008). Auch wenn diese Faktoren nicht zu den klassischen Gütekriterien zählen, stehen sie in engem Zusammenhang mit der diagnostischen Qualität und werden zunehmend unter dem Begriff der Testakzeptanz diskutiert (Kersting, 2008).

Für die vorliegende Studie bilden die aufgeführten Gütekriterien den Bewertungsrahmen, innerhalb dessen die beiden Testverfahren eingeordnet werden. Sie ermöglichen es, Leistungsunterschiede, Akzeptanzbewertungen und potenzielle Geschlechtsunterschiede nicht isoliert, sondern im Kontext etablierter diagnostischer Qualitätsstandards zu interpretieren.

2.4 Geschlechtsunterschiede in Intelligenztests

Die Frage nach Geschlechtsunterschieden in der Intelligenz zählt seit Beginn der Intelligenzforschung zu den am intensivsten diskutierten Themen (Neisser et al., 1996). Dabei ist zwischen der allgemeinen Intelligenz als übergeordnetem Konstrukt und spezifischen Fähigkeitsbereichen zu unterscheiden. In der Forschung besteht heute weitgehender Konsens darüber, dass sich Frauen und Männer in der allgemeinen Intelligenz nicht systematisch unterscheiden (Neisser et al., 1996). Meta-Analysen zeigen, dass Mittelwertsunterschiede, sofern sie auftreten, sehr gering sind und in der Regel keine praktische Relevanz besitzen (Hyde, 2005; Neisser et al., 1996).

Gleichzeitig lassen sich in einzelnen Fähigkeitsbereichen wiederholt geschlechtsspezifische Unterschiede beobachten (Hyde, 2005). So schneiden Männer in bestimmten visuell-räumlichen Aufgaben im Durchschnitt etwas besser ab, während Frauen in sprachlichen Aufgaben häufig leichte Vorteile zeigen (Hyde, 2005; Matlin & Halpern, 1988). Diese Unterschiede sind jedoch stark kontextabhängig und variieren je nach Aufgabenformat, Testkonstruktion und Stichprobe. Zudem überlappen die Leistungsverteilungen von Frauen und Männern in allen Bereichen in hohem Masse, sodass individuelle Unterschiede innerhalb der Geschlechter deutlich grösser sind als Unterschiede zwischen den Geschlechtern (Hyde, 2005).

Zur Erklärung dieser Befunde existieren unterschiedliche Ansätze. Biologische Erklärungsmodelle verweisen unter anderem auf hormonelle Einflüsse oder neuroanatomische Unterschiede. Diese Ansätze sind jedoch umstritten, da biologische Effekte meist klein ausfallen und in ihrer Wirkung stark von Umweltfaktoren beeinflusst werden (Hyde, 2005). Demgegenüber betonen sozial-kognitive

Erklärungsansätze die Bedeutung von Sozialisation, Rollenbildern und Erwartungshaltungen. Geschlechtsspezifische Leistungsunterschiede werden hier als Ergebnis unterschiedlicher Lerngelegenheiten, Interessenentwicklung oder sozialer Zuschreibungen interpretiert (Hyde, 2005).

Ein weiterer zentraler Erklärungsansatz betrifft testmethodische Faktoren (Kersting, 1998). Zahlreiche Studien zeigen, dass Geschlechtsunterschiede weniger vom gemessenen Konstrukt selbst als vielmehr von der konkreten Ausgestaltung der Testaufgaben abhängen (Kersting, 1998, 2008). Sprachliche Anforderungen, Aufgabenformate, Zeitdruck oder visuelle Gestaltung können geschlechtsspezifische Effekte verstärken oder abschwächen. So zeigen beispielsweise stark sprachgebundene Aufgaben häufiger Vorteile für Frauen, während Aufgaben mit ausgeprägtem räumlichem oder mechanischem Anteil eher Vorteile für Männer aufweisen (Schuler, 2014).

Vor diesem Hintergrund gewinnt der Begriff der Testfairness an Bedeutung (Kersting, 2008; Messick, 1989). Ein Test gilt als fair, wenn er für unterschiedliche Gruppen dieselbe Bedeutung hat und keine systematischen Verzerrungen aufweist, die nicht durch Unterschiede im gemessenen Konstrukt erklärbar sind (Kersting, 1998; Messick, 1989). Fairness bedeutet dabei nicht zwangsläufig, dass Frauen und Männer identische Mittelwerte erzielen müssen. Vielmehr geht es darum, dass Unterschiede, sofern sie auftreten, auf tatsächliche Fähigkeitsunterschiede zurückzuführen sind und nicht auf testfremde Anforderungen (Hyde, 2005; Kersting, 1998).

Moderne Testentwicklungen verfolgen daher zunehmend das Ziel, potenzielle Bias-Quellen bereits auf Ebene der Testkonstruktion zu minimieren. Dazu zählen unter anderem die Reduktion sprachlicher Anforderungen, die Verwendung abstrakter Aufgabenformate sowie eine sorgfältige Prüfung einzelner Items auf gruppenspezifische Verzerrungen (Kersting, 2008). Insbesondere Tests, die auf fluide Intelligenz abzielen, werden häufig als vergleichsweise geschlechtsneutral betrachtet, da sie weniger stark an schulische Inhalte oder kulturelle Erfahrungen gebunden sind (Kersting, 2008; McGrew, 2009).

Gleichzeitig ist zu beachten, dass auch fluide Tests nicht vollständig frei von geschlechtsspezifischen Effekten sind (Hyde, 2005). Visuelle Anforderungen wie räumliches Vorstellungsvermögen oder Wahrnehmungsgeschwindigkeit, Wortschatz oder mathematisches Vorwissen, sowie Strategien bei der Aufgabenbearbeitung können ebenfalls geschlechtstypisch variieren. Aus diesem Grund ist es erforderlich, Fairness nicht nur theoretisch zu postulieren, sondern empirisch zu überprüfen (Dantlgraber, 2015). Dies gilt insbesondere für neue oder verkürzte Testversionen, bei denen einzelne Aufgaben ein höheres Gewicht für das Gesamtergebnis haben (Kersting, 2008; McGrew, 2009).

Für die vorliegende Arbeit ist diese Differenzierung zentral. Der Modulare Kurzintelligenztest (M-KIT) erhebt den Anspruch, geschlechtsfair konstruiert zu sein (Dantlgraber, 2015). Während dieser Anspruch für die Langversion des Tests bereits empirisch untersucht wurde, ist bislang unklar, ob sich vergleichbare Befunde auch für die Kurzversion im Online-Setting zeigen. Die Untersuchung möglicher Geschlechtsunterschiede stellt daher einen wesentlichen Bestandteil der vorliegenden Studie dar und bildet die Grundlage für Hypothese 1 (Geschlechtsunterschiede in der Testleistung).

2.5 Online Testung und digitale Intelligenzdiagnostik

Parallel zur theoretischen Weiterentwicklung der Intelligenzforschung hat sich in den letzten Jahren die Art der Testdurchführung grundlegend verändert. Intelligenztests werden zunehmend digital und ortsunabhängig durchgeführt, insbesondere in der Eignungsdiagnostik, im Bildungsbereich und in der Forschung (Goldberg, 2018; Steiner & Lieberei, 2024). Online-Testverfahren bieten zahlreiche Vorteile, darunter hohe Standardisierung, flexible Einsatzmöglichkeiten und automatisierte Auswertung (Goldberg, 2018; Steiner & Lieberei, 2024).

Gleichzeitig bringt die digitale Testung spezifische Herausforderungen mit sich (Goldberg, 2018). Während Präsenztests in kontrollierten Umgebungen stattfinden, sind Online-Testungen durch eine höhere Varianz der Testbedingungen gekennzeichnet. Unterschiedliche technische Ausstattungen, Ablenkungen im häuslichen Umfeld sowie eine stärkere Selbststeuerung der Testpersonen können das Testverhalten beeinflussen (Goldberg, 2018). Diese Faktoren wirken sich potenziell auf Motivation, Konzentration und Bearbeitungsstrategie aus (Goldberg, 2018; Kersting, 2008).

Studien zeigen, dass digitale Testformate das Antwortverhalten verändern können, insbesondere wenn klassische Papier-Bleistift-Tests lediglich digitalisiert werden, ohne an die neue Umgebung

angepasst zu sein (Goldberg, 2018). Unübersichtliche Layouts, ungewohnte Navigation oder unklare Instruktionen können zu erhöhter kognitiver Belastung führen und die Leistung beeinträchtigen (Kersting, 2008). Moderne digitale Tests versuchen daher, Benutzeroberfläche, Instruktionsdesign und Aufgabenpräsentation gezielt auf das Online-Setting abzustimmen (Goldberg, 2018).

Ein weiterer Aspekt betrifft die Standardisierung. Digitale Tests ermöglichen eine sehr präzise Zeitmessung und eine einheitliche Präsentation der Items (Goldberg, 2018). Gleichzeitig können Unterschiede in digitaler Kompetenz oder technischer Vertrautheit neue Ungleichheiten erzeugen (Goldberg, 2018; Kersting, 2008). Diese Aspekte sind besonders relevant für heterogene Stichproben und unterstreichen die Bedeutung einer benutzerfreundlichen und intuitiven Testgestaltung.

In der vorliegenden Studie kommt dem digitalen Kontext eine hohe Bedeutung zu, da beide untersuchten Verfahren online durchgeführt werden, jedoch aus unterschiedlichen Testtraditionen stammen. Während moderne Verfahren explizit für digitale Anwendungen konzipiert wurden, basieren klassische Tests häufig auf analogen Vorläufern (Dantlgraber, 2015; Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Diese Unterschiede können sich sowohl auf die Testleistung als auch auf die subjektive Wahrnehmung der Tests auswirken und bilden einen wichtigen Hintergrund für die Untersuchung von Akzeptanzunterschieden (Kersting, 2008).

2.6 Testakzeptanz: Bedeutung, Modelle und Einflussfaktoren

Neben den klassischen Hauptgütekriterien wie Objektivität, Reliabilität und Validität hat als Nebengütekriterium die Akzeptanz von Testverfahren in den vergangenen Jahrzehnten zunehmend an Bedeutung gewonnen (Kersting, 1998, 2008). Testakzeptanz beschreibt die subjektive Bewertung eines Testverfahrens durch die Testpersonen und umfasst Aspekte wie Verständlichkeit, Fairnesswahrnehmung, Belastung, Transparenz und allgemeine Zustimmung zum Einsatz des Tests (Kersting, 1998). Insbesondere in Anwendungsfeldern, in denen Testergebnisse als Grundlage für bedeutsame Entscheidungen dienen, spielt die Akzeptanz eine zentrale Rolle (Schuler, 2014).

Forschungsarbeiten zeigen, dass eine geringe Akzeptanz nicht nur negative Einstellungen gegenüber einem Test hervorruft, sondern auch das Bearbeitungsverhalten beeinflussen kann (Kersting, 1998; Schuler, 2014). Testpersonen, die ein Verfahren als unfair, unverständlich oder unangemessen empfinden, zeigen häufiger geringeres Engagement, oberflächliche Bearbeitung oder erhöhte Abbruchraten (Goldberg, 2018; Kersting, 1998). Damit wirkt sich die subjektive Wahrnehmung eines Tests potenziell auch auf die Qualität der erhobenen Daten aus (Kersting, 1998).

Testakzeptanz ist dabei von verwandten Konzepten abzugrenzen (Kersting, 1998). Sie ist nicht mit Motivation gleichzusetzen, da auch motivierte Personen einen Test kritisch bewerten können. Ebenso ist Akzeptanz nicht identisch mit Testfairness, da ein objektiv fair konstruiertes Verfahren subjektiv dennoch als unfair wahrgenommen werden kann (Kersting, 1998; Messick, 1989). Akzeptanz stellt somit eine eigenständige Dimension dar, die sowohl von testbezogenen als auch von personenspezifischen Faktoren beeinflusst wird (Kersting, 1998; Moosbrugger & Kelava, 2020).

Zu den testbezogenen Einflussfaktoren zählen unter anderem die verwendeten Aufgabenformate, die Länge des Tests, die Klarheit der Instruktionen sowie die wahrgenommene Relevanz der Aufgaben (Kersting, 2008). Längere Testverfahren werden häufiger als belastend empfunden, während kurze, übersichtliche Tests tendenziell höhere Akzeptanzwerte erzielen (Rammstedt & Beierlein, 2014). Gleichzeitig kann eine starke Verkürzung auch Skepsis hervorrufen, etwa hinsichtlich der Aussagekraft der Ergebnisse (Ziegler, Kemper & Kruyen, 2014). Die Wahrnehmung eines angemessenen Verhältnisses zwischen Testdauer und diagnostischem Nutzen spielt daher eine zentrale Rolle (Kersting, 2008).

Personenspezifische Faktoren umfassen unter anderem Vorerfahrungen mit Testverfahren, Einstellungen gegenüber Leistungsdiagnostik sowie situative Aspekte wie Zeitdruck oder emotionale Verfassung (Kersting, 1998). Auch die Bedeutung, die dem Testergebnis beigemessen wird, beeinflusst die Akzeptanz (Schuler, 2014). In Hochrisiko-Situationen, etwa bei Auswahlentscheidungen, wird ein Test häufig kritischer bewertet als in rein wissenschaftlichen Kontexten (Schuler, 2014).

Im Kontext von Online-Testungen gewinnt die Testakzeptanz zusätzlich an Relevanz (Goldberg, 2018; Steiner & Lieberei, 2024). Digitale Testverfahren werden häufig unbeaufsichtigt durchgeführt, wodurch

soziale Kontrolle und direkte Rückmeldung durch Testleitende entfallen (Goldberg, 2018). Gleichzeitig sind Online-Tests stärker von Benutzerfreundlichkeit, technischer Stabilität und klaren Instruktionen abhängig (Goldberg, 2018; Kersting, 2008). Unklare Abläufe oder technische Schwierigkeiten können die Akzeptanz deutlich reduzieren und zu Frustration führen (Kersting, 2008).

Zugleich bieten Online-Tests auch das Potenzial zur Steigerung der Akzeptanz (Goldberg, 2018; Steiner & Lieberei, 2024). Flexible Durchführungszeiten, ortsunabhängige Bearbeitung und eine moderne Gestaltung können als positiv wahrgenommen werden (Goldberg, 2018). Besonders kürzere, übersichtlich strukturierte Tests werden im digitalen Kontext häufig bevorzugt, da sie sich besser in den Alltag integrieren lassen und weniger Ermüdung hervorrufen (Rammstedt & Beierlein, 2014).

Für die vorliegende Studie ist die Betrachtung der Testakzeptanz aus mehreren Gründen zentral (Kersting, 2008). Zum einen stellt Akzeptanz eine wichtige Ergänzung zu leistungsbezogenen Kennwerten dar und erlaubt eine differenziertere Bewertung der untersuchten Testverfahren (Kersting, 2008; Moosbrugger & Kelava, 2020). Zum anderen ist insbesondere im Vergleich zwischen einem etablierten klassischen Verfahren und einem modernen, digital ausgerichteten Kurztest von Interesse, ob sich Unterschiede in der subjektiven Wahrnehmung zeigen (Dantlgraber, 2015; Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Die Analyse der Testakzeptanz liefert somit einen wesentlichen Beitrag zum Verständnis der Einsatzmöglichkeiten und Grenzen moderner Online-Intelligenztests (Goldberg, 2018; Kersting, 2008) und besitzt – wie die Analyse der Geschlechtsunterschiede – eine hohe praxisbezogene Relevanz.

2.7 Testfairness und Bias in psychologischen Testverfahren

Testfairness stellt ein zentrales Qualitätskriterium psychologischer Diagnostik dar und gewinnt insbesondere im Kontext standardisierter Leistungs- und Eignungstests zunehmend an Bedeutung (Messick, 1989; Moosbrugger & Kelava, 2020). Ein Test gilt als fair, wenn Personen mit gleicher Ausprägung des gemessenen Konstrukts – unabhängig von Gruppenzugehörigkeit wie Geschlecht, Herkunft oder Bildung – die gleiche Chance haben, eine vergleichbare Testleistung zu erzielen (Messick, 1989). Fairness ist eng mit dem Anspruch verbunden, valide und gerechte diagnostische Entscheidungen zu ermöglichen (Messick, 1989; Schuler, 2014).

In der psychologischen Testtheorie wird Fairness nicht als eigenständiges Gütekriterium im klassischen Sinne verstanden, sondern als integrativer Bestandteil von Validität (Messick, 1989; Moosbrugger & Kelava, 2020). Ein Verfahren ist dann unfair, wenn systematische Leistungsunterschiede auftreten, die nicht auf tatsächliche Unterschiede im Zielkonstrukt, sondern auf testfremde Einflüsse zurückzuführen sind (Kersting, 1998; Messick, 1989; Schuler, 2014). Solche Verzerrungen werden unter dem Begriff Bias zusammengefasst.

Ein Bias kann auf unterschiedlichen Ebenen entstehen. Ein Inhaltsbias liegt vor, wenn Aufgabenformate spezifische Wissensbestände oder kulturell geprägte Erfahrungen voraussetzen, die nicht für alle Testpersonen gleichermassen verfügbar sind (Kersting, 1998; Messick, 1989; Moosbrugger & Kelava, 2020). Ein Sprachbias kann auftreten, wenn sprachliche Komplexität oder Wortschatzanforderungen über das eigentlich zu messende Konstrukt hinausgehen. Konstruktbias entsteht, wenn ein Test unterschiedliche kognitive Prozesse bei verschiedenen Gruppen anspricht und dadurch das intendierte Konstrukt nicht äquivalent erfasst (Messick, 1989).

Ein weiterer wichtiger Aspekt betrifft formatbedingte Verzerrungen (Kersting, 1998). Empirische Forschung zeigt, dass bestimmte Aufgabenformate systematisch mit geschlechtsbezogenen Leistungsunterschieden assoziiert sind, ohne dass dies auf Unterschiede in der allgemeinen Intelligenz zurückgeführt werden kann (Hyde, 2005; Neisser et al., 1996). Verbal geprägte Aufgaben begünstigen tendenziell Frauen, während räumlich-figurale Aufgaben häufiger mit Vorteilen für Männer verbunden sind (Hyde, 2005; Kersting, 1998). Solche Effekte sind besonders relevant für Tests, die unterschiedliche Aufgabenformate kombinieren (Kersting, 1998; Schuler, 2014).

Moderne Testentwicklung reagiert auf diese Befunde mit einem verstärkt konstruktzentrierten Ansatz (Dantlgraber, 2015; Moosbrugger & Kelava, 2020). Ziel ist es, Aufgaben so zu gestalten, dass sie möglichst direkt auf das intendierte kognitive Konstrukt abzielen und testfremde Anforderungen minimieren. Dies umfasst unter anderem die Reduktion sprachlicher Belastung, die Vermeidung kulturell spezifischer Inhalte sowie eine klare, einheitliche Aufgabenstruktur (Dantlgraber, 2015; Kersting, 2008). In diesem Zusammenhang gewinnt auch die Analyse von Differential Item

Functioning (DIF) an Bedeutung, mit der überprüft wird, ob einzelne Items für verschiedene Gruppen unterschiedliche Schwierigkeiten aufweisen, obwohl die zugrunde liegende Fähigkeit gleich ausgeprägt ist (Moosbrugger & Kelava, 2020).

Im digitalen Kontext treten zusätzliche Fairnessaspekte hinzu (Goldberg, 2018; Steiner & Lieberei, 2024). Während digitale Testverfahren eine hohe Standardisierung ermöglichen, können Unterschiede in technischer Ausstattung, digitaler Erfahrung oder Benutzerfreundlichkeit der Oberfläche neue Verzerrungsquellen darstellen (Goldberg, 2018). Eine unübersichtliche Darstellung oder komplexe Navigation kann bestimmte Gruppen stärker belasten und somit indirekt Leistungsunterschiede erzeugen (Goldberg, 2018; Steiner & Lieberei, 2024). Fairness im digitalen Setting erfordert daher nicht nur inhaltlich ausgewogene Aufgaben, sondern auch eine nutzerzentrierte Gestaltung der Testumgebung (Kersting, 2008).

Testfairness steht zudem in engem Zusammenhang mit der subjektiven Wahrnehmung der Testpersonen (Kersting, 1998, 2008). Studien zeigen, dass als unfair empfundene Tests nicht nur die Akzeptanz senken, sondern auch Motivation und Bearbeitungsqualität negativ beeinflussen können (Kersting, 2008; Schuler, 2014). Damit wirkt Fairness sowohl objektiv auf die Messgenauigkeit als auch subjektiv auf das Erleben der Testpersonen (Messick, 1989). Diese doppelte Wirkung ist insbesondere für Online-Testungen relevant, bei denen Testpersonen stärker auf ihre eigene Einschätzung des Verfahrens angewiesen sind (Goldberg, 2018; Kersting, 2008).

Für die vorliegende Studie ist das Fairnesskonzept von zentraler Bedeutung, da sich das IST-Screening und der M-KIT hinsichtlich ihrer Aufgabenstruktur, ihres Konstruktfokus und ihres Entwicklungsanspruchs unterscheiden. Während klassische Tests häufig eine Kombination verschiedener Aufgabenformate einsetzen, verfolgen moderne Verfahren explizit das Ziel, geschlechtsbezogene und andere gruppenspezifische Verzerrungen schon bei der Entwicklung eines einzelnen Aufgabenformats zu reduzieren (Dantlgraber, 2015; Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Theoretisch ist daher zu erwarten, dass sich Unterschiede in der Testkonstruktion sowohl in geschlechtsbezogenen Leistungsunterschieden als auch in der subjektiven Akzeptanz widerspiegeln.

Diese Überlegungen bilden die konzeptionelle Brücke zwischen den Hypothesen H1 und H2 und liefern die theoretische Grundlage für den empirischen Vergleich der beiden Testverfahren im digitalen Kontext.

2.8 Kurz- und Langversionen von Intelligenztests

In der psychologischen Diagnostik werden Intelligenztests häufig sowohl in umfassenden Langversionen als auch in ökonomischeren Kurzversionen angeboten. Diese Differenzierung ist vor allem durch praktische Anforderungen motiviert: In vielen Anwendungsfeldern – etwa in der Eignungsdiagnostik, in Online-Studien oder in frühen Screening-Phasen – stehen begrenzte Zeitressourcen, eingeschränkte Aufmerksamkeit der Testpersonen oder organisatorische Rahmenbedingungen im Vordergrund. Kurzversionen sollen unter diesen Bedingungen eine valide Einschätzung kognitiver Leistungsfähigkeit ermöglichen, ohne den diagnostischen Aufwand unverhältnismässig zu erhöhen (Moosbrugger & Kelava, 2020; Rammstedt & Beierlein, 2014).

Aus testtheoretischer Perspektive sind Kurzversionen jedoch nicht als verkleinerte Äquivalente von Langversionen, sondern als eigenständige Messinstrumente mit spezifischen Stärken und Limitationen zu verstehen (Rammstedt & Beierlein, 2014; Ziegler et al., 2014). Ein zentrales Konzept in diesem Zusammenhang ist der Trade-off zwischen Testökonomie und Messpräzision (Lienert & Raatz, 1998; Moosbrugger & Kelava, 2020). In der klassischen Testtheorie gilt, dass die Reliabilität eines Tests in hohem Masse von der Anzahl der enthaltenen Items abhängt. Mit abnehmender Itemzahl steigt der Einfluss zufälliger Messfehler, wodurch die Messgenauigkeit und die Differenzierungsfähigkeit eines Tests eingeschränkt sein können (Lienert & Raatz, 1998; Bortz & Döring, 2016).

Neben der Reliabilität betrifft dieser Trade-off auch die inhaltliche Repräsentativität des gemessenen Konstrukts (Moosbrugger & Kelava, 2020; Ziegler et al., 2014). Während Langversionen durch eine grössere Itempopulation unterschiedliche Facetten eines Konstrukts abbilden und potenzielle Verzerrungen einzelner Aufgabenformate ausgleichen können, hängt die diagnostische Qualität von Kurzversionen stärker von der konkreten Itemauswahl ab (Rammstedt & Beierlein, 2014; Ziegler et al.,

2014). Insbesondere bei komplexen Konstrukten wie Intelligenz besteht das Risiko, dass bestimmte kognitive Prozesse oder Aufgabenformate über- oder unterrepräsentiert sind (Moosbrugger & Kelava, 2020).

Diese Problematik gewinnt zusätzlich an Bedeutung im Hinblick auf Fairness und gruppenspezifische Verzerrungen (Kersting, 1998; Messick, 1989). Forschung zur Testfairness zeigt, dass systematische Leistungsunterschiede zwischen Gruppen nicht allein auf Unterschiede im gemessenen Konstrukt zurückzuführen sind, sondern häufig durch testfremde Anforderungen wie Sprachlastigkeit, Aufgabenformate oder kulturelle Erfahrungen beeinflusst werden (Kersting, 1998; Messick, 1989). In umfangreichen Langversionen können sich solche Effekte über viele Items hinweg teilweise nivellieren. In Kurzversionen hingegen haben einzelne Items ein höheres Gewicht für den Gesamtwert, wodurch potenzielle Bias-Effekte stärker ins Gewicht fallen können (Rammstedt & Beierlein, 2014; Ziegler et al., 2014).

Vor diesem Hintergrund betonen mehrere Autor:innen, dass psychometrische Befunde aus Langversionen nicht automatisch auf Kurzversionen übertragbar sind (Rammstedt & Beierlein, 2014; Ziegler et al., 2014). Kurztests benötigen eine eigenständige empirische Prüfung ihrer Reliabilität, Validität und Fairness im jeweiligen Anwendungskontext (Moosbrugger & Kelava, 2020; Ziegler et al., 2014). Dies gilt insbesondere für Merkmale wie geschlechtsbezogene Leistungsunterschiede, da bereits geringe formatbedingte Effekte bei kurzen Skalen einen disproportionalen Einfluss auf die Ergebnisse haben können (Kersting, 1998; Ziegler et al., 2014).

Gleichzeitig weisen Kurzversionen auch spezifische Vorteile auf, die insbesondere im digitalen Kontext relevant sind (Goldberg, 2018; Rammstedt & Beierlein, 2014). Kürzere Bearbeitungszeiten können die Akzeptanz der Testverfahren erhöhen, Ermüdung reduzieren und die Motivation der Testpersonen fördern (Kersting, 2008; Rammstedt & Beierlein, 2014). Studien zeigen, dass insbesondere bei Online-Erhebungen eine hohe Testökonomie mit geringeren Abbruchraten und einer sorgfältigeren Bearbeitung einhergehen kann (Goldberg, 2018; Rammstedt & Beierlein, 2014). Damit stehen Kurzversionen nicht per se in einem Qualitätskonflikt, sondern erfordern eine differenzierte Bewertung entlang mehrerer diagnostischer Kriterien (Moosbrugger & Kelava, 2020; Ziegler et al., 2014).

Moderne Testentwicklungen versuchen, diesen Zielkonflikt durch konstruktzentrierte Ansätze zu adressieren. Durch eine gezielte Auswahl von Items, die zentrale kognitive Prozesse möglichst direkt erfassen, sollen auch in Kurzversionen valide und faire Messungen ermöglicht werden (Dantlgraber, 2015; Moosbrugger & Kelava, 2020). Der modulare Aufbau des Modularen Kurzintelligenztests (M-KIT) folgt genau diesem Ansatz (Dantlgraber, 2015). Laut Manual wurde der M-KIT explizit mit dem Ziel entwickelt, eine ökonomische Erfassung fluider Intelligenz zu ermöglichen und gleichzeitig potenzielle Verzerrungsquellen – insbesondere sprachliche und bildungsabhängige Anforderungen – zu minimieren (Dantlgraber, 2015).

Während für die Langversion des M-KIT empirische Hinweise auf gute psychometrische Eigenschaften und geringe geschlechtsbezogene Leistungsunterschiede vorliegen (Dantlgraber, 2015), ist bislang unklar, ob sich diese Befunde auch auf die Kurzversion im Online-Setting übertragen lassen (Ziegler et al., 2014). Gerade im digitalen Kontext, in dem Testbedingungen weniger kontrolliert sind und subjektive Wahrnehmung eine grössere Rolle spielt, ist eine eigenständige Überprüfung zentraler Qualitätsmerkmale erforderlich (Rammstedt & Beierlein, 2014; Ziegler et al., 2014).

Die vorliegende Studie setzt an dieser empirischen Leerstelle an. Durch den direkten Vergleich der Kurzversion des M-KIT mit einem etablierten klassischen Verfahren wird geprüft, ob der Anspruch konstruktzentrierter, geschlechtsfairer Testentwicklung auch unter den Bedingungen einer ökonomischen Online-Testung eingelöst werden kann. Damit leistet die Arbeit einen Beitrag zur differenzierten Bewertung moderner Kurztests in der angewandten Intelligenzdiagnostik.

2.9 Das IST-Screening: Aufbau, theoretische Einbettung und psychometrische Eigenschaften

Das IST-Screening ist ein etabliertes Kurzverfahren zur Erfassung kognitiver Leistungsfähigkeit und basiert konzeptionell auf dem Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R), einem der am häufigsten eingesetzten Intelligenztests im deutschsprachigen Raum (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012) . Als Screening-Instrument wurde das Verfahren entwickelt, um

in begrenzter Zeit eine ökonomische Einschätzung schlussfolgernden Denkens zu ermöglichen, insbesondere in eignungsdiagnostischen und beratungspsychologischen Kontexten (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012).

2.9.1 Theoretischer Hintergrund

Das IST-Screening steht in der Tradition klassischer Intelligenzmodelle, die Intelligenz als ein Zusammenspiel mehrerer kognitiver Teilfähigkeiten verstehen. Es orientiert sich an Amthauers Konzept der Intelligenzstruktur, das davon ausgeht, dass unterschiedliche Fähigkeitsbereiche – insbesondere verbale, numerische und figurale Leistungen – gemeinsam zur allgemeinen Intelligenz beitragen (Kersting, 2000). Dieses Verständnis ist kompatibel mit hierarchischen Modellen der Intelligenz, in denen sich verschiedene spezifische Fähigkeiten auf einen gemeinsamen allgemeinen Faktor (g) zurückführen lassen (Carroll, 1993; Spearman, 1904).

Im Gegensatz zu stärker konstruktzentrierten Verfahren, die primär fluide Intelligenz erfassen, integriert das IST-Screening sowohl fluide als auch kristallisierte Anteile. Diese theoretische Ausrichtung erlaubt eine breite Abbildung kognitiver Leistungsfähigkeit, bringt jedoch zugleich eine stärkere Abhängigkeit von Vorwissen, sprachlicher Kompetenz und erlernten Problemlösestrategien mit sich (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012; McGrew, 2009).

2.9.2 Aufgabenformate und Teststruktur

Das IST-Screening umfasst drei zentrale Aufgabenbereiche:

- Analogien (verbal-logisch):
  Diese Aufgaben erfordern das Erkennen semantischer oder formaler Beziehungen zwischen Begriffen. Sie setzen sprachliches Verständnis und einen gewissen Wortschatz voraus und aktivieren sowohl schlussfolgernde als auch wissensbasierte Prozesse.

- Zahlenreihen (numerisch-schlussfolgernd):
  In diesem Aufgabenformat müssen regelhafte Muster in Zahlenfolgen erkannt und fortgesetzt werden. Die Bearbeitung erfordert logisches Denken, kann jedoch auch durch mathematische Strategien und numerische Übung beeinflusst werden.

- Matrizen (figural-logisch):
  Diese Aufgaben zielen auf visuell-räumliches Schliessen ab. Die Testpersonen müssen Beziehungen zwischen grafischen Elementen erkennen und logisch fortsetzen. Matrizen gelten als relativ spracharm, erfordern jedoch visuell-analytische Fähigkeiten.

Durch die Kombination dieser drei Aufgabentypen soll ein breites Spektrum schlussfolgernder Denkprozesse erfasst werden (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Die Aufgaben sind zeitlich begrenzt, was zusätzliche Anforderungen an Verarbeitungsgeschwindigkeit und Arbeitsgedächtnis stellt.

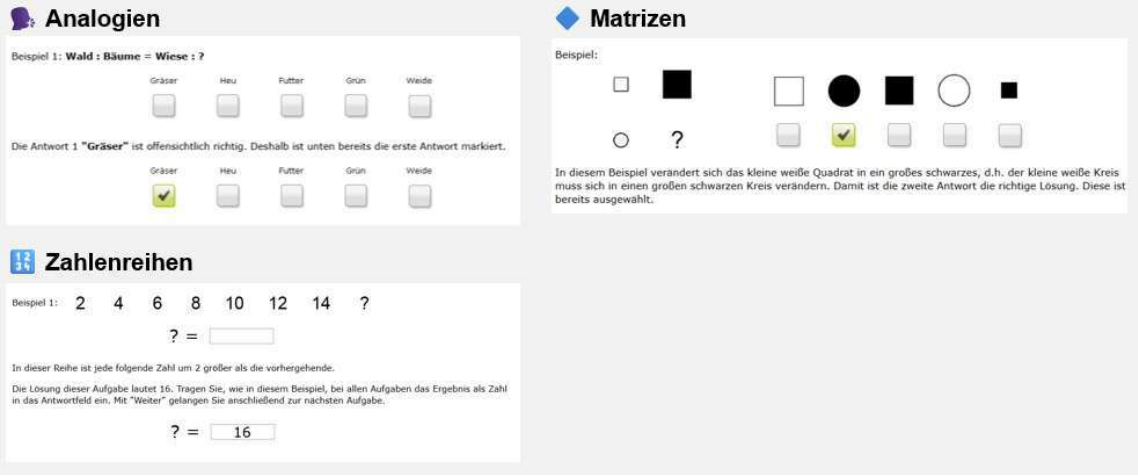

Abbildung 1: Beispielhafte Aufgabenformate des IST-Screenings (Analogien, Zahlenreihen, Matrizen).

Abbildung 1 veranschaulicht die drei im IST-Screening eingesetzten Aufgabenformate. Die Beispiele verdeutlichen die Kombination aus sprachlichen, numerischen und figuralen Anforderungen, die für das Verfahren charakteristisch sind.

2.9.3 Psychometrische Eigenschaften

Gemäss Manual weist das IST-Screening zufriedenstellende bis gute psychometrische Kennwerte auf (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Die interne Konsistenz der Subtests sowie des Gesamtwerts liegt im akzeptablen bis guten Bereich, was auf eine zuverlässige Messung der intendierten Fähigkeiten hinweist (Lienert & Raatz, 1998). Die faktorielle Struktur zeigt, dass die drei Subtests auf einen gemeinsamen übergeordneten Faktor laden, zugleich jedoch eigenständige Varianzanteile aufweisen (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012).

Hinsichtlich der Validität zeigt das IST-Screening erwartungskonforme Zusammenhänge mit schulischen und beruflichen Leistungsindikatoren sowie mit umfassenderen Intelligenztests. Die Konstruktvalidität wird dadurch gestützt, dass die Aufgabenformate klar den theoretisch postulierten Fähigkeitsbereichen zugeordnet werden können (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012).

Als Screening-Verfahren ist das IST-Screening jedoch nicht auf eine differenzierte Fähigkeitsdiagnostik ausgelegt, sondern auf eine ökonomische Gesamteinschätzung kognitiver Leistungsfähigkeit (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Entsprechend ist die inhaltliche Breite höher als die Tiefe einzelner Fähigkeitsbereiche.

2.9.4 Fairness und mögliche Bias-Quellen

Aufgrund seiner Aufgabenstruktur weist das IST-Screening potenzielle Fairnessaspekte auf, die in der Literatur diskutiert werden. Verbale Analogien können durch sprachlichen Hintergrund und Bildung beeinflusst sein, Zahlenreihen durch mathematische Lerngelegenheiten und Matrizen durch visuell-räumliche Übung (Hyde, 2005; Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Diese Aspekte bedeuten nicht zwangsläufig eine mangelnde Fairness des Verfahrens, verdeutlichen jedoch, dass unterschiedliche Aufgabenformate verschiedene kognitive und erfahrungsbasierte Ressourcen aktivieren (Kersting, 1998; Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012).

Empirische Befunde zu geschlechtsbezogenen Leistungsunterschieden im IST-Screening zeigen keine einheitlichen oder systematischen Effekte, jedoch lassen sich – abhängig vom Subtest – Unterschiede in einzelnen Aufgabenbereichen beobachten. Diese Befunde werden in der Regel auf die unterschiedliche Aufgabenstruktur und nicht auf Unterschiede in der allgemeinen Intelligenz zurückgeführt.

2.9.5 Online Durchführung

Das IST-Screening wurde ursprünglich für die Durchführung in Präsenzsettings entwickelt (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Inzwischen wird das Verfahren häufig auch digital eingesetzt. Die Online-Version folgt inhaltlich der klassischen Struktur, wobei die Aufgaben am Bildschirm präsentiert und die Bearbeitungszeit kontrolliert wird.

Die digitale Durchführung ermöglicht eine standardisierte Präsentation und automatisierte Auswertung, stellt jedoch zusätzliche Anforderungen an Benutzerführung und visuelle Gestaltung. Da das Verfahren nicht primär für digitale Umgebungen konzipiert wurde, können Unterschiede in der subjektiven Wahrnehmung im Vergleich zu modern entwickelten Online-Tests auftreten (Dantlgraber, 2015; Goldberg, 2018). Diese Aspekte sind insbesondere für die Bewertung der Testakzeptanz relevant und werden in der vorliegenden Studie berücksichtigt.

2.10 Der Modulare Kurzintelligenztest (M-KIT): Aufbau, theoretische Einbettung und psychometrische Eigenschaften

Der Modulare Kurzintelligenztest (M-KIT) ist ein modernes Intelligenztestverfahren, das explizit auch für digitale Anwendungen und für eine ökonomische, faire Erfassung kognitiver Leistungsfähigkeit entwickelt wurde. Der Test folgt einem modularen Aufbau, der eine flexible Zusammenstellung unterschiedlicher Aufgabenbereiche erlaubt. Ziel des Verfahrens ist es, fluide Intelligenz konstruktzentriert, möglichst frei von Vorwissen und unabhängig von sprachlichen oder kulturellen Einflüssen zu erfassen (Dantlgraber, 2015).

Im Unterschied zu klassischen Intelligenztests, die häufig mehrere Fähigkeitsbereiche kombinieren, wurde der M-KIT mit dem Anspruch entwickelt, zentrale Prozesse schlussfolgernden Denkens möglichst direkt abzubilden. Der Fokus liegt auf der Messung fluider Intelligenz im Sinne des Cattell-Horn-Carroll-Modells (Carroll, 1993; Cattell, 1963; McGrew, 2009). Damit ordnet sich der M-KIT klar in die Tradition moderner, konstruktzentrierter Testentwicklung ein.

2.10.1 Theoretischer Hintergrund

Die theoretische Grundlage des M-KIT bildet das CHC-Modell der Intelligenz, insbesondere die fluide Intelligenz (Gf). Fluide Intelligenz beschreibt die breite Fähigkeit, neue Probleme zu analysieren, Muster zu erkennen und logische Beziehungen abzuleiten, ohne auf erlerntes Wissen zurückzugreifen (Cattell, 1963; McGrew, 2009). Diese Fähigkeit gilt als zentral für allgemeine Problemlösekompetenz und als besonders geeignet für faire Leistungsdiagnostik (Dantlgraber, 2015; Messick, 1989).

Die Entwickler des M-KIT betonen, dass fluide Intelligenz weniger abhängig von formaler Bildung, Sprachkompetenz oder kulturellem Hintergrund ist als kristallisierte Fähigkeitsanteile. Entsprechend wurde der Test so konstruiert, dass sprachliche Anforderungen minimiert und Aufgabenformate gewählt wurden, die primär auf visuell-analytische und logisch-schlussfolgernde Prozesse abzielen (Dantlgraber, 2015).

Diese theoretische Ausrichtung unterscheidet den M-KIT von klassischen Testverfahren und ist insbesondere im Kontext digitaler Diagnostik relevant, da Online-Tests häufig in heterogenen Stichproben eingesetzt werden (Steiner & Lieberei, 2024).

2.10.2 Aufgabenformate und Teststruktur (Modul A)

In der vorliegenden Untersuchung wurde Modul A des M-KIT eingesetzt. Dieses Modul besteht aus drei Aufgabentypen, die unterschiedliche Aspekte fluider Intelligenz erfassen, jedoch auf einen gemeinsamen zugrunde liegenden kognitiven Prozess ausgerichtet sind.

- Wortfolgen:
  In diesen Aufgaben müssen regelhafte Beziehungen zwischen abstrakten Begriffen erkannt werden. Der Fokus liegt auf logischem Denken, nicht auf Wortschatz oder semantischem Wissen. Die Aufgaben weisen eine geringe Sprachlastigkeit auf und sind so gestaltet, dass sie auch ohne vertiefte sprachliche Kenntnisse bearbeitet werden können.

- Bildteile:
  Diese Aufgaben erfassen figural-analytisches Denken. Testpersonen müssen visuelle Muster analysieren und passende Elemente identifizieren. Die Aufgaben sind klar strukturiert und verzichten auf kulturell oder inhaltlich geprägte Darstellungen.

- Zahlenvergleiche:
  Die Zahlenvergleiche messen Verarbeitungsgeschwindigkeit und basale numerische Diskriminationsfähigkeit und erfordern zur Lösung nur geringes mathematisches Vorwissen.

Alle Aufgabenformate sind zeitlich begrenzt und digital standardisiert. Das Manual betont, dass die einzelnen Subtests bewusst so gewählt wurden, dass sie unterschiedliche Oberflächenmerkmale aufweisen, jedoch dieselben zentralen kognitiven Prozesse aktivieren: Mustererkennung, Regelableitung und kognitive Flexibilität – alles zentrale Bestandteile der fluiden Intelligenz (Dantlgraber, 2015).

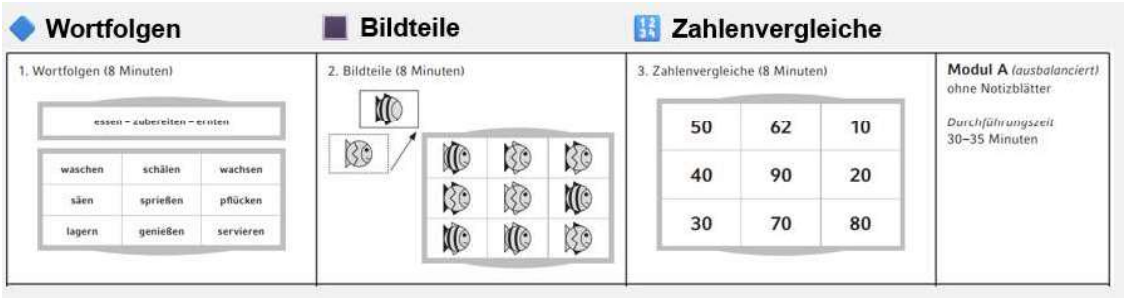

Abbildung 2: Beispielhafte Aufgabenformate des Modularen Kurzintelligenztests (M-KIT), Modul A.

Abbildung 2 zeigt exemplarisch die Aufgabenformate des M-KIT (Modul A). Die visuelle Gestaltung verdeutlichen den konstruktzentrierten Fokus auf fluide Intelligenz.

2.10.3 Psychometrische Eigenschaften

Die psychometrischen Kennwerte des M-KIT werden im Manual ausführlich dokumentiert (Dantlgraber, 2015). Für Modul A wird eine interne Konsistenz von etwa $\alpha = .80$ berichtet, was für ein kurzes Modul als gut einzustufen ist. Für den Gesamttest ergeben sich sehr hohe Reliabilitätswerte ($\alpha > .90$), was auf eine präzise Messung fluider Intelligenz hinweist.

Faktoranalytische Untersuchungen zeigen eine klare einfaktorielle Struktur, die mit dem theoretischen Anspruch einer konstruktzentrierten Erfassung fluider Intelligenz übereinstimmt. Die Konstruktvalidität wird zudem durch einen hohe Zusammenhang mit dem Modul „schlussfolgerndes Denken“ des I-S-T 2000 R (der Langfassung des IST-Screenings) gestützt (Dantlgraber, 2015).

Hinsichtlich der Kriteriumsvalidität zeigen sich erwartungskonforme Zusammenhänge mit schulischen und akademischen Leistungsindikatoren sowie mit weiteren kognitiven Leistungsmassen. Die Ergebnisse sprechen dafür, dass der M-KIT zentrale Aspekte allgemeiner kognitiver Leistungsfähigkeit zuverlässig abbildet.

2.10.4 Fairness und Geschlechtsneutralität

Ein zentrales Entwicklungsziel des M-KIT ist die Reduktion potenzieller Bias-Quellen. Das Manual berichtet, dass bei der Itemkonstruktion gezielt darauf geachtet wurde, sprachliche, kulturelle und bildungsabhängige Anforderungen zu minimieren. Empirische Analysen zeigen geringe bis vernachlässigbare geschlechtsbezogene Leistungsunterschiede, sowohl auf Subtest- als auch auf Gesamttestebene (Dantlgraber, 2015).

Darüber hinaus wurden Analysen zu Differential Item Functioning (DIF) durchgeführt, die keine systematischen Verzerrungen einzelner Items zugunsten einer bestimmten Geschlechtsgruppe zeigten. Diese Befunde sprechen für eine geschlechtsneutrale Konstruktion des Verfahrens und unterstützen den Anspruch, fluide Intelligenz unabhängig von gruppenspezifischen Merkmalen zu erfassen (Dantlgraber, 2015).

2.10.5 Akzeptanz und Benutzerfreundlichkeit

Neben psychometrischen Kriterien spielt beim M-KIT die subjektive Wahrnehmung durch die Testpersonen eine zentrale Rolle. Laut Manual berichten Testpersonen von einer hohen Verständlichkeit der Instruktionen, einer klaren Aufgabenstruktur und einer als angemessen empfundenen Belastung. Die Bearbeitungsdauer wird überwiegend als passend eingeschätzt, und der Test wird insgesamt als modern und professionell gestaltet wahrgenommen (Dantlgraber, 2015).

Diese hohe Akzeptanz wird im Manual explizit als Qualitätsmerkmal hervorgehoben. Die Entwickler argumentieren, dass eine positive Testerfahrung die Motivation zur sorgfältigen Bearbeitung fördert und damit indirekt zur diagnostischen Qualität beiträgt.

2.10.6 Online Durchführung

Der M-KIT wurde von Beginn an auch für die digitale Durchführung konzipiert (Dantlgraber, 2015). Die Benutzeroberfläche ist auf Online-Settings abgestimmt, mit klarer Navigation, übersichtlicher Darstellung und automatisierter Zeitmessung. Instruktionen werden standardisiert präsentiert, und der Testablauf ist intuitiv gestaltet.

Diese digitale Optimierung unterscheidet den M-KIT von klassischen Verfahren, die ursprünglich ausschliesslich für Präsenzsettings entwickelt wurden. Im Kontext der vorliegenden Studie ist dieser Aspekt besonders relevant, da Unterschiede in der digitalen Nutzererfahrung sowohl die Leistung als auch die Akzeptanz beeinflussen können.

2.11 Diagnostische Konsequenzen moderner Online-Kurztests

Die zunehmende Verbreitung von Online-Kurztests in der Intelligenzdiagnostik wirft grundlegende diagnostische und ethische Fragen auf (Steiner & Lieberei, 2024). Einerseits ermöglichen solche Verfahren eine effiziente, standardisierte und ortsunabhängige Erfassung kognitiver Leistungsfähigkeit. Andererseits verändern sie die Bedingungen, unter denen diagnostische Entscheidungen getroffen werden, und stellen traditionelle Annahmen der Testanwendung in Frage (Goldberg, 2018).

Ein zentrales Spannungsfeld moderner Online-Kurztests liegt im Verhältnis von Testökonomie und diagnostischer Tiefe (Moosbrugger & Kelava, 2020; Rammstedt & Beierlein, 2014). Während Kurztests eine rasche Einschätzung ermöglichen, erfassen sie zwangsläufig nur einen begrenzten Ausschnitt kognitiver Fähigkeiten. Dies erfordert eine sorgfältige Abwägung zwischen Effizienzgewinnen und potenziellen Informationsverlusten. Die diagnostische Aussagekraft eines Kurztests ist daher stets im Kontext seines Einsatzbereichs zu beurteilen (Rammstedt & Beierlein, 2014; Ziegler et al., 2014).

Ein weiteres Spannungsfeld betrifft die Frage der Fairness. Online-Tests werden häufig mit dem Anspruch entwickelt, möglichst unabhängig von sprachlichen, kulturellen oder geschlechtsspezifischen Faktoren zu sein (Dantlgraber, 2015; McGrew, 2009). Gleichzeitig können digitale Rahmenbedingungen neue Ungleichheiten erzeugen, etwa durch unterschiedliche technische Voraussetzungen, variierende Testumgebungen oder Unterschiede im Umgang mit digitalen Formaten (Goldberg, 2018; Steiner & Lieberei, 2024). Fairness ist daher nicht allein eine Eigenschaft des Testinhalts, sondern auch des Anwendungskontexts.

Darüber hinaus gewinnt die subjektive Wahrnehmung der Testpersonen im Online-Setting an Bedeutung. Ohne direkte Testaufsicht hängt die Qualität der Datenerhebung in hohem Masse vom Engagement und der Motivation der Teilnehmenden ab (Goldberg, 2018). Akzeptanz wird damit zu einer praktischen Voraussetzung für valide Testergebnisse. Ein Test, der als unfair, unverständlich oder übermässig belastend erlebt wird, birgt das Risiko reduzierter Anstrengung oder vorzeitiger Abbrüche.

Aus diagnostischer Sicht ergibt sich daraus die Notwendigkeit, Online-Kurztests nicht isoliert anhand klassischer Gütekriterien zu bewerten (Moosbrugger & Kelava, 2020). Vielmehr sollten leistungsbezogene Kennwerte, Fairnessüberlegungen und Akzeptanzaspekte gemeinsam betrachtet werden. Dies gilt insbesondere in Anwendungsfeldern wie der Personaldiagnostik oder

Bildungsberatung, in denen Testergebnisse weitreichende Konsequenzen für die getesteten Personen haben können (Schuler, 2014).

Die vorliegende Arbeit positioniert sich vor diesem Hintergrund als Beitrag zur differenzierten Betrachtung moderner Online-Intelligenztests. Durch den Vergleich zweier Kurzverfahren unter vergleichbaren Bedingungen wird aufgezeigt, dass diagnostische Qualität nicht allein an Effizienz oder Testergebnissen festgemacht werden kann. Vielmehr erfordert der verantwortungsvolle Einsatz solcher Verfahren eine reflektierte Auseinandersetzung mit ihren Stärken, Grenzen und Wirkungen auf die Testpersonen.

2.12 Vergleich des IST-Screenings und des M-KIT

Das IST-Screening und der Modulare Kurzintelligenztest (M-KIT) verfolgen beide das Ziel, kognitive Leistungsfähigkeit standardisiert zu erfassen. Trotz dieses gemeinsamen Grundanliegens unterscheiden sich die beiden Verfahren in mehreren zentralen Aspekten, die für die Interpretation der Ergebnisse der vorliegenden Studie von Bedeutung sind. Ein systematischer Vergleich dieser Unterschiede ist erforderlich, um Leistungsunterschiede, Akzeptanzbewertungen und mögliche Fairnessaspekte angemessen einordnen zu können.

Ein wesentlicher Unterschied betrifft die theoretische Ausrichtung der beiden Testverfahren. Das IST-Screening basiert auf einer breit angelegten Konzeption von Intelligenz und erfasst verschiedene Fähigkeitsbereiche mithilfe unterschiedlicher Aufgabenformate (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Diese Vorgehensweise folgt der klassischen Tradition der Intelligenzdiagnostik und zielt darauf ab, ein möglichst umfassendes Leistungsbild zu gewinnen. Der M-KIT verfolgt demgegenüber einen stärker fokussierten Ansatz und konzentriert sich auf die Erfassung fluider Intelligenz (Dantlgraber, 2015). Durch diese konstruktzentrierte Ausrichtung sollen grundlegende kognitive Prozesse möglichst direkt gemessen werden.

Auch hinsichtlich der Aufgabenformate zeigen sich deutliche Unterschiede. Das IST-Screening kombiniert sprachliche, numerische und figural-anschauliche Aufgaben (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Diese Vielfalt erlaubt eine breite Abdeckung unterschiedlicher Fähigkeitsbereiche, geht jedoch mit einer stärkeren Abhängigkeit von sprachlichen Kompetenzen, schulischen Vorerfahrungen und Bildungsbiografien einher. Der M-KIT verwendet ebenfalls sprach-, visuell- und zahlenbasierte Aufgaben, die jedoch bei der Bearbeitung weniger Vorwissen voraussetzen (Dantlgraber, 2015). Diese Reduktion solcher Anforderungen wird in der Testentwicklung häufig als Vorteil im Hinblick auf Vergleichbarkeit und potenzielle Fairness betrachtet (Kersting, 2008; McGrew, 2009).

Im Hinblick auf psychometrische Eigenschaften unterscheiden sich die Verfahren ebenfalls. Für das IST-Screening liegen umfangreiche empirische Befunde zu Reliabilität und Validität vor, die auf einer langen Anwendungstradition beruhen. Der M-KIT ist ein vergleichsweise neues Verfahren, für das insbesondere für die Langversion bereits gute psychometrische Kennwerte berichtet werden. Für die Kurzversion ist die empirische Evidenz hingegen begrenzter, weshalb eine eigenständige Untersuchung ihrer Eigenschaften erforderlich ist. Dies betrifft insbesondere Aspekte wie Reliabilität, Fairness und Akzeptanz im Online-Setting.

Auch der Umgang mit potenziellen Bias-Quellen unterscheidet sich zwischen den beiden Verfahren. Das IST-Screening wurde zu einer Zeit entwickelt, in der Fragen der Geschlechtsneutralität und Testfairness eine geringere Rolle spielten als in der heutigen Testentwicklung. Obwohl das Verfahren kontinuierlich überarbeitet wurde, bleibt seine Grundstruktur stärker an klassische Testtraditionen gebunden. Der M-KIT wurde demgegenüber mit einem expliziten Anspruch auf Fairness konzipiert. Durch die Fokussierung auf fluide Intelligenz und die Reduktion von Vorwissen sollen geschlechts- und bildungsbezogene Verzerrungen möglichst minimiert werden.

Schliesslich können sich die beiden Verfahren auch in ihrer subjektiven Wahrnehmung unterscheiden. Während das IST-Screening durch seine grössere Anforderungsvielfalt als abwechslungsreich erlebt werden kann, besteht zugleich die Gefahr einer höheren kognitiven Belastung. Der M-KIT könnte aufgrund seiner klaren Struktur als weniger belastend wahrgenommen werden, gleichzeitig aber auch Skepsis hinsichtlich seiner diagnostischen Aussagekraft hervorrufen. Diese unterschiedlichen Wahrnehmungen sind insbesondere im Online-Setting relevant, da sie das Bearbeitungsverhalten und damit indirekt auch die Testergebnisse beeinflussen können.

Zusammenfassend lassen sich IST-Screening und M-KIT als Vertreter zweier unterschiedlicher diagnostischer Ansätze verstehen. Das IST-Screening steht für eine klassische, breit angelegte Intelligenzdiagnostik mit langjähriger Anwendungstradition. Der M-KIT repräsentiert einen modernen, konstruktzentrierten Ansatz, der auf Testökonomie, digitale Einsatzmöglichkeiten und Fairness ausgerichtet ist. Diese Unterschiede bilden die Grundlage für den in der vorliegenden Arbeit vorgenommenen Vergleich und sind zentral für die Einordnung der empirischen Befunde.

2.13 Theoretische Integration und Bedeutung für die vorliegende Studie

Die vorangegangenen Kapitel haben zentrale theoretische Perspektiven der Intelligenzdiagnostik, der Testfairness sowie der Testakzeptanz im Kontext digitaler Verfahren dargestellt. Im Folgenden werden diese Ansätze zusammengeführt, um die theoretische Ausgangslage der vorliegenden Studie zu präzisieren und ihre Relevanz im bestehenden Forschungsfeld einzuordnen.

Moderne Intelligenzdiagnostik steht zunehmend vor der Herausforderung, valide und ökonomische Verfahren bereitzustellen, die unter unterschiedlichen Rahmenbedingungen einsetzbar sind (Moosbrugger & Kelava, 2020). Insbesondere im digitalen Kontext gewinnen Kurztests an Bedeutung, da sie eine zeiteffiziente Erfassung kognitiver Leistungsfähigkeit ermöglichen und sich gut in Online-Settings integrieren lassen (Goldberg, 2018; Steiner & Lieberei, 2024). Gleichzeitig stellen sich jedoch Fragen nach der diagnostischen Qualität, der Fairness sowie der subjektiven Wahrnehmung solcher Verfahren durch die Testpersonen (Kersting, 2008; Rammstedt & Beierlein, 2014).

Die theoretischen Ausführungen zur fluider Intelligenz verdeutlichen, dass diese als relativ kultur- und bildungsunabhängige Fähigkeit gilt und daher häufig als geeignete Zielgrösse für faire Intelligenztests herangezogen wird (McGrew, 2009). Testverfahren, die primär auf fluide Intelligenz abzielen, verfolgen den Anspruch, Verzerrungen durch sprachliche Anforderungen, schulische Vorerfahrungen oder geschlechtsspezifische Sozialisationseffekte zu reduzieren (Kersting, 2008). Dieser Anspruch ist insbesondere im Kontext von Testfairness und Chancengleichheit von zentraler Bedeutung (Messick, 1989).

Gleichzeitig zeigen psychometrische Modelle, dass Fairness kein eindimensionales Merkmal darstellt, sondern sich aus mehreren Aspekten zusammensetzt (Messick, 1989; Moosbrugger & Kelava, 2020). Neben der inhaltlichen Konstruktion der Aufgaben spielen auch formale Merkmale wie Instruktionsklarheit, Aufgabenformat, Bearbeitungsbedingungen und Testdauer eine Rolle. Diese Faktoren beeinflussen nicht nur die gemessene Leistung, sondern auch die subjektive Wahrnehmung des Testverfahrens (Kersting, 1998, 2008).

An dieser Stelle gewinnt das Konzept der Testakzeptanz an Bedeutung. Akzeptanz stellt eine zentrale Schnittstelle zwischen objektiver Testgüte und subjektiver Testerfahrung dar (Kersting, 2008). Sie beeinflusst Motivation, Bearbeitungsengagement und potenziell auch die Qualität der erhobenen Leistungsdaten (Kersting, 1998). Gerade bei unbeaufsichtigten Online-Tests ist davon auszugehen, dass Akzeptanz einen relevanten Einfluss auf das Bearbeitungsverhalten der Teilnehmenden hat (Goldberg, 2018).

Die bisherige Forschung zeigt, dass sich Akzeptanz nicht automatisch aus guten psychometrischen Kennwerten ergibt. Ein Test kann objektiv valide und reliabel sein, jedoch von den Testpersonen als unfair, belastend oder wenig verständlich wahrgenommen werden (Kersting, 1998). Umgekehrt bedeutet eine hohe Akzeptanz nicht zwangsläufig eine hohe diagnostische Qualität (Moosbrugger & Kelava, 2020). Aus theoretischer Sicht ist Akzeptanz daher als eigenständige, aber eng mit der Testgüte verbundene Dimension zu betrachten.

Vor diesem Hintergrund erscheint es sinnvoll, Intelligenztests nicht isoliert hinsichtlich ihrer Leistungsdaten zu vergleichen, sondern auch die subjektive Bewertung durch die Testpersonen systematisch zu berücksichtigen. Dies gilt insbesondere für den Vergleich unterschiedlicher Testkonzepte, wie etwa breit angelegter Intelligenztests und stärker konstruktzentrierter Verfahren.

Die vorliegende Studie knüpft an diese theoretischen Überlegungen an, indem sie zwei etablierte Intelligenztestverfahren in ihrer Online-Kurzversion miteinander vergleicht. Dabei wird nicht nur die Beziehung zwischen den Testergebnissen untersucht, sondern auch geprüft, inwieweit sich die Verfahren hinsichtlich wahrgenommener Fairness und Akzeptanz unterscheiden. Durch die

Kombination leistungsbezogener und subjektiver Daten wird ein integrativer Zugang zur Bewertung moderner Online-Intelligenztests gewählt.

Insgesamt trägt die theoretische Integration dazu bei, die Forschungsfrage der Arbeit klar zu verorten: Es geht nicht allein um die Frage, ob sich zwei Testverfahren in ihren Ergebnissen unterscheiden, sondern darum, wie diese Unterschiede im Spannungsfeld von diagnostischer Qualität, Fairnessanspruch und subjektiver Wahrnehmung zu verstehen sind.

3 Methodik

3.1 Studiendesign

Die vorliegende Untersuchung folgt einem quantitativen, vergleichenden Within-Subjects-Design. Alle Teilnehmenden bearbeiteten sowohl das IST-Screening als auch das Modul A des Modularen Kurzintelligenztests (M-KIT). Dieses Design ermöglicht es, Unterschiede in der Testleistung sowie in der subjektiven Akzeptanz der beiden Verfahren innerhalb derselben Stichprobe zu untersuchen. Durch den intraindividuellen Vergleich können interindividuelle Störvariablen, wie etwa Unterschiede in allgemeinen kognitiven Fähigkeiten, Motivation oder Bildungshintergrund, weitgehend kontrolliert werden (Döring & Bortz, 2016).

Das gewählte Studiendesign ist insbesondere für vergleichende Fragestellungen in der psychologischen Diagnostik geeignet, da es eine höhere statistische Sensitivität aufweist als Between-Subjects-Designs. Leistungsunterschiede zwischen den Testverfahren lassen sich somit präziser auf testinhärente Merkmale wie Aufgabenstruktur oder theoretische Ausrichtung zurückführen und weniger auf Unterschiede zwischen den untersuchten Personen. Die höhere Sensitivität ergibt sich aus der Reduktion der Fehlervarianz durch den intraindividuellen Vergleich (Cohen, 2009; Döring & Bortz, 2016).

Die Reihenfolge der Testdurchführung wurde variiert, um mögliche Reihenfolge-, Übungs- oder Ermüdungseffekte zu kontrollieren. Ein Teil der Teilnehmenden absolvierte zunächst das IST-Screening, während der andere Teil mit dem M-KIT begann. Die Testreihenfolge diente ausschliesslich als methodische Kontrollvariable und war nicht Bestandteil der inhaltlichen Fragestellungen. Durch diese Variation sollte sichergestellt werden, dass potenzielle Leistungseinflüsse nicht systematisch an eines der beiden Testverfahren gekoppelt sind (Döring & Bortz, 2016).

Ergänzend zur Erfassung der Testleistung wurden nach jedem Intelligenztest Akzeptanzdaten erhoben. Dieses Vorgehen erlaubt es, leistungsbezogene Ergebnisse unmittelbar mit subjektiven Bewertungen der jeweiligen Testsituation zu verknüpfen. Die Kombination objektiver Leistungsdaten mit Akzeptanzmessungen entspricht dem Anspruch, Intelligenztests im Online-Setting nicht allein anhand klassischer Gütekriterien, sondern auch im Hinblick auf ihre Zumutbarkeit und Nutzerperspektive zu bewerten (Döring & Bortz, 2016).

3.2 Stichprobe

Die finale Stichprobe umfasste N = 59 Personen, die beide Intelligenztests vollständig bearbeitet haben. Davon waren 32 Frauen (54.2 %) und 27 Männer (45.8 %). Das Alter der Teilnehmenden lag zwischen 15 und 68 Jahren (M = 31.42, SD = 15.53), was auf eine vergleichsweise breite Altersstreuung hinweist. Diese Altersheterogenität ermöglichte explorative Analysen altersbezogener Effekte, auch wenn die Stichprobe nicht gezielt für Altersvergleiche rekrutiert wurde.

Die Rekrutierung erfolgte über das persönliche Umfeld der Autorin im Sinne eines Convenience Sampling. Die Teilnahme war freiwillig und unentgeltlich. In einzelnen Fällen erfolgte die Teilnahme im Rahmen persönlicher Gefälligkeiten, etwa durch Unterstützung im privaten Umfeld, ohne dass eine formelle Entschädigung oder leistungsbezogene Anreize gewährt wurden. Dieses Vorgehen ist in empirischen Online-Studien verbreitet, geht jedoch mit Einschränkungen hinsichtlich der Repräsentativität der Stichprobe einher.

Es ist davon auszugehen, dass Personen, die bereit waren, an einer zeitlich anspruchsvollen Online-Studie teilzunehmen, eine gewisse Offenheit gegenüber psychologischen Testverfahren oder Online-Erhebungen aufweisen. Gleichzeitig könnten Personen mit geringerer Motivation, begrenzter zeitlicher Verfügbarkeit oder einer kritischeren Haltung gegenüber Intelligenztests von einer Teilnahme abgesehen haben. Diese Selbstselektion stellt eine potenzielle Verzerrungsquelle dar und schränkt die Generalisierbarkeit der Ergebnisse ein.

Die Datenerhebung fand in einem unbeaufsichtigten Online-Setting statt. Die Teilnehmenden bearbeiteten die Tests an einem selbstgewählten Ort, in der Regel zu Hause. Dadurch konnten zwar flexible Teilnahmebedingungen ermöglicht werden, gleichzeitig liessen sich situative Einflüsse wie Ablenkungen, Pausen oder unterschiedliche technische Voraussetzungen nicht kontrollieren. Diese

Aspekte sind insbesondere bei leistungsdiagnostischen Verfahren relevant und wurden bei der Interpretation der Ergebnisse berücksichtigt.

Die Testreihenfolge war nicht vollständig ausgeglichen. Denn 23 Personen (39.0 %) bearbeiteten zuerst das IST-Screening, während 36 Personen (61.0 %) mit dem M-KIT begannen. Trotz dieser leicht ungleichen Verteilung wurde die Testreihenfolge in den statistischen Analysen als Kontrollvariable berücksichtigt, um mögliche Verzerrungen durch Reihenfolgeeffekte zu identifizieren.

Insgesamt eignet sich die Stichprobe trotz der genannten Einschränkungen für die Zielsetzung der Studie. Der Fokus lag nicht auf der Repräsentativität, sondern auf dem methodisch kontrollierten Vergleich zweier Intelligenztestverfahren unter möglichst vergleichbaren Bedingungen im Online-Setting.

3.3 Durchführung der Datenerhebung

Die Datenerhebung erfolgte vollständig im unbeaufsichtigten Online-Setting. Nach der Rekrutierung erhielten die Teilnehmenden einen Zugangslink zur Studie sowie standardisierte Instruktionen zur Durchführung der Tests. Der vollständige Einladungstext zur Studienteilnahme ist in Anhang C dokumentiert. Sie wurden darauf hingewiesen, dass die Studie in einer möglichst ruhigen Umgebung und ohne Unterbrechungen zu bearbeiten ist. Eine Kontrolle der tatsächlichen Durchführungsbedingungen war aufgrund des Online-Settings jedoch nicht möglich.

In der Studie wurden das IST-Screening (Standard A) und das Modul A des Modularen Kurzintelligenztests (M-KIT) eingesetzt. Beide Tests wurden unter vergleichbaren Rahmenbedingungen durchgeführt und wiesen eine ähnliche Bearbeitungsdauer auf. Die durchschnittliche Bearbeitungsdauer betrug etwa 28 Minuten pro Intelligenztest, die Bearbeitung des jeweiligen Akzeptanzfragebogens rund 2 Minuten.

Ziel war es, die beiden Testverfahren unter möglichst gleichen Bedingungen zu vergleichen und Unterschiede nicht durch strukturelle oder zeitliche Faktoren zu verzerren.

Die Reihenfolge der Testdurchführung wurde variiert. Ein Teil der Teilnehmenden absolvierte zunächst das IST-Screening und anschliessend den M-KIT, während der andere Teil mit dem M-KIT begann und danach das IST-Screening bearbeitete. Durch diese Variation sollte geprüft werden, ob die Reihenfolge der Tests einen Einfluss auf die Testleistung oder die Akzeptanz hatte. Nach jedem Intelligenztest wurde eine kurze Akzeptanzbefragung durchgeführt, um die subjektive Wahrnehmung der jeweiligen Testsituation unmittelbar zu erfassen.

Das unbeaufsichtigte Online-Setting bringt sowohl Vorteile als auch methodische Herausforderungen mit sich. Einerseits ermöglicht es eine flexible Teilnahme unabhängig von Ort und Zeit und erleichtert die Rekrutierung einer heterogenen Stichprobe. Andererseits können externe Einflüsse wie Ablenkungen, Unterbrechungen oder individuelle Pausen nicht kontrolliert werden. Auch Unterschiede in der technischen Ausstattung der Teilnehmenden, etwa Bildschirmgrösse oder Eingabegeräte, könnten die Bearbeitung beeinflusst haben.

Trotz dieser Einschränkungen wurde versucht, durch klare Instruktionen, eine einheitliche Testabfolge und eine übersichtliche Struktur der Studie möglichst vergleichbare Bedingungen für alle Teilnehmenden zu schaffen. Die standardisierte Online-Durchführung stellt zudem einen realitätsnahen Anwendungsfall moderner Intelligenzdiagnostik dar, da entsprechende Tests zunehmend digital und unbeaufsichtigt eingesetzt werden.

Insgesamt ermöglichte die gewählte Durchführung eine ökonomische und zugleich methodisch kontrollierte Datenerhebung, die den Vergleich der beiden Testverfahren unter praxisnahen Bedingungen erlaubte.

3.4 Erhebungsinstrumente

Zur Erfassung der kognitiven Testleistung wurden das IST-Screening (Standard A) sowie der Modulare Kurzintelligenztest (M-KIT, Modul A) eingesetzt. Beide Verfahren wurden in digitaler Form durchgeführt. Die theoretische Einbettung, der Aufbau sowie die psychometrischen Eigenschaften der

Testverfahren werden ausführlich in Kapitel 2 dargestellt. In der Methodik wird daher auf eine erneute Beschreibung der Tests verzichtet und der Fokus auf die Durchführung und Auswertung gelegt.

Die subjektive Akzeptanz der beiden Testverfahren wurde mithilfe eines von der Autorin dieser Arbeit gekürzten und adaptierten Fragebogens auf Basis des AKZEPT-L nach Kersting erhoben (Kersting, 2008). Aus dem Originalinstrument wurden gezielt jene Items ausgewählt, die für den Vergleich der beiden Online-Testformate, die Überprüfung der Hypothese H2 sowie die Bewertung der Testgestaltung im digitalen Kontext inhaltlich relevant waren.

Berücksichtigt wurden insbesondere Items zur Verständlichkeit der Aufgaben und Instruktionen, zum wahrgenommenen Fairnessempfinden, zur subjektiven Belastung beziehungsweise Anstrengung sowie zur professionellen Gestaltung und zum Gesamteindruck des jeweiligen Testverfahrens. Items mit berufsspezifischem Bezug, zur Testauswertung oder mit inhaltlich redundanter Belastungserfassung wurden ausgeschlossen. Ziel dieser Reduktion war es, die Bearbeitungsdauer des Fragebogens zu begrenzen und den Fokus der Untersuchung auf die zentralen Akzeptanzdimensionen beizubehalten.

Tabelle 1 gibt eine Übersicht über die Originalformulierungen der im Akzeptanzfragebogen eingesetzten Items sowie das jeweilige Antwortformat.

Tabelle 1: Originalformulierungen der Items des eingesetzten Akzeptanzfragebogens

Formulierung	Antwortformat
Die Testaufgaben waren klar und verständlich.	1–6 Likert
Während der Bearbeitung wusste ich jederzeit, was ich tun musste.	1–6 Likert
Die Bearbeitung war sehr belastend.	1–6 Likert
Ich habe mich während der Testung überfordert gefühlt.	1–6 Likert
Die getesteten Fähigkeiten sind auch fürs Berufsleben relevant.	1–6 Likert
Der Test schien mir professionell konstruiert zu sein.	1–6 Likert
Die Bearbeitung der Aufgaben hat Spass gemacht.	1–6 Likert

Negativ formulierte Items zur Belastung wurden für die Auswertung rekodiert, sodass höhere Werte durchgängig eine höhere Akzeptanz anzeigen.

Zusätzlich vergaben die Teilnehmenden für jedes Testverfahren eine Schulnote nach dem in der Schweiz üblichen Schulnotensystem als globale Akzeptanzbewertung. Diese diente als ergänzendes, intuitives Mass zur Erfassung des Gesamteindrucks und ermöglichte eine übergreifende Einschätzung der Akzeptanz.

Alle Items wurden auf einer sechsstufigen Likert-Skala beantwortet (1 = stimme überhaupt nicht zu, 6 = stimme voll und ganz zu). Eine neutrale Antwortoption wurde bewusst nicht angeboten, um klare Urteile zu fördern und Tendenzen zur Mitte zu reduzieren. Die Itemreihenfolge war für das IST-Screening und den M-KIT identisch, um Vergleichbarkeit zwischen den beiden Testverfahren sicherzustellen. Der vollständige Akzeptanzfragebogen ist in Anhang D dokumentiert.

3.5 Anonymität, Code-System und Datenzuordnung

Zur anonymen Zuordnung der Testergebnisse und Akzeptanzbewertungen verwendeten die Teilnehmenden einen selbstgewählten persönlichen Code. Dieser Code bestand aus einer vierstelligen Kombination aus Buchstaben und Zahlen und erlaubte keine Rückschlüsse auf die Identität der Teilnehmenden. Die Eingabe desselben Codes in allen vier Untersuchungsteilen stellte sicher, dass die Daten korrekt zusammengeführt werden konnten, ohne personenbezogene Informationen zu erfassen.

Dieses Vorgehen ermöglichte eine datenschutzkonforme Verknüpfung der einzelnen Erhebungsteile und reduzierte zugleich das Risiko von Zuordnungsfehlern. Die Verwendung selbstgewählter Codes ist ein gängiges Verfahren in anonymen Online-Studien und trägt zur Wahrung der Privatsphäre der Teilnehmenden bei.

Die Teilnahme an der Studie war grundsätzlich anonym. Den Teilnehmenden wurde jedoch die Möglichkeit eingeräumt, ihr individuelles Testergebnis nach Abschluss der Untersuchung zu erhalten. In diesem Fall konnten sie den verwendeten Code freiwillig per E-Mail an die Autorin senden. In diesem Schritt war die vollständige Anonymität gegenüber der Autorin aufgehoben. Dies erfolgte ausschliesslich auf Wunsch der Teilnehmenden und hatte keinen Einfluss auf die Datenauswertung oder die Berücksichtigung der Daten in der Studie.

3.6 Datenaufbereitung und Datenbereinigung

Zur Vorbereitung der statistischen Analysen wurden die Daten aus den vier Erhebungsteilen (zwei Intelligenztests und zwei Akzeptanzfragebögen) in einem zentralen Datensatz zusammengeführt. Die Zuordnung der einzelnen Datensätze erfolgte anhand des von den Teilnehmenden vergebenen persönlichen Codes.

Die Daten wurden in einer Excel-Datei so aufbereitet, dass für jede teilnehmende Person alle relevanten Variablen (Testleistungen, Akzeptanzwerte, soziodemografische Angaben und Testreihenfolge) in einer Zeile abgebildet waren. Diese Strukturierung erleichterte die anschliessende statistische Auswertung und reduzierte das Risiko fehlerhafter Zuordnungen. Der aufbereitete Datensatz wurde anschliessend für die Analyse in die Statistiksoftware jamovi importiert.

Die Rohdaten wurden nach Abschluss der Datenerhebung hinsichtlich Vollständigkeit, Plausibilität und Bearbeitungsqualität überprüft. Datensätze mit fehlenden Angaben oder auffälligen Bearbeitungsmustern (z. B. sehr kurze Bearbeitungszeiten, unrealistische Leistungswerte oder Hinweise auf nicht regelkonforme Bearbeitung) wurden ausgeschlossen. Ziel dieser Datenbereinigung war es, die Validität der Analysen sicherzustellen und Verzerrungen durch unsorgfältige Bearbeitung zu minimieren.

Als leistungsbezogenes Ausschlusskriterium wurde ein Rohwert unter 12 Punkten im M-KIT definiert.Das betrifft Personen, die im Schnitt weniger als vier Aufgaben pro Aufgabenformat gelöst hatten. Diese Grenze wurde gewählt, um Bearbeitungen zu identifizieren, bei denen von unzureichendem Instruktionsverständnis oder mangelnder Mitarbeit auszugehen war. Zusätzlich wurden Datensätze ausgeschlossen, bei denen die Bearbeitungsdauer eines Intelligenztests unter 10 Minuten lag. Diese Kriterien orientierten sich an den vorab kommunizierten Instruktionen zur sorgfältigen Bearbeitung der Tests und entsprechen gängigen Vorgehensweisen in der Leistungsdiagnostik.

Nach Anwendung aller Ausschlusskriterien verblieb eine finale Stichprobe von N = 59 Personen.

3.7 Statistische Analysen

Die statistischen Analysen wurden mit der Software jamovi durchgeführt. Vor der Durchführung inferenzstatistischer Tests wurden die Verteilungseigenschaften der zentralen Variablen mithilfe visueller Inspektionen (Histogramme, Q-Q-Plots) sowie Shapiro-Wilk-Tests überprüft, um die Voraussetzungen parametrischer Verfahren zu beurteilen.

Zur Untersuchung des Zusammenhangs zwischen den Testergebnissen des IST-Screenings und des M-KIT wurde eine Pearson-Korrelation berechnet. Diese Analyse diente der Prüfung, inwieweit beide Testverfahren vergleichbare Aspekte kognitiver Leistungsfähigkeit erfassen.

Geschlechtsunterschiede in der Testleistung wurden mithilfe von t-Tests für unabhängige Stichproben analysiert. Der Vergleich der Akzeptanzbewertungen zwischen dem IST-Screening und dem M-KIT erfolgte mittels t-Tests für verbundene Stichproben, da beide Testverfahren von denselben Teilnehmenden bearbeitet wurden und ein Within-Subjects-Design vorlag.

Zur Kontrolle möglicher Reihenfolgeeffekte wurde ein weiterer t-Test für unabhängige Stichproben durchgeführt, bei dem die Leistung der Gruppe mit zuerst bearbeitetem IST-Screening mit jener der Gruppe mit zuerst bearbeitetem M-KIT verglichen wurde.

Zusätzlich wurden für alle relevanten Mittelwertvergleiche Effektstärken nach Cohen (*d*) berichtet, um neben der statistischen Signifikanz auch die praktische Bedeutung der Ergebnisse einordnen zu

können (Cohen, 2009). Das Signifikanzniveau wurde für alle Analysen auf $\alpha = .05$ festgelegt, was dem in der psychologischen Forschung üblichen Konventionswert entspricht.

3.8 Ethische Aspekte

Die Teilnahme an der Studie war freiwillig. Vor Beginn der Untersuchung gaben alle Teilnehmenden eine elektronische Einverständniserklärung ab. Es wurden keine personenbezogenen Daten erhoben, die eine Identifikation der Teilnehmenden erlauben.

Den Teilnehmenden wurde zudem kommuniziert, dass sie die Bearbeitung der Studie jederzeit ohne Angabe von Gründen abbrechen können. Da Intelligenztests als potenziell leistungsbezogen und belastend wahrgenommen werden können, wurde auf eine transparente Information bezüglich des Ablaufes und Zielsetzung der Studie geachtet.

Da es sich um ein anonymes Studierendenprojekt ohne invasive oder gesundheitlich belastende Interventionen handelt, war keine Genehmigung durch eine Ethikkommission erforderlich.

4 Ergebnisse

In diesem Kapitel werden die Ergebnisse der empirischen Untersuchung dargestellt. Der Aufbau orientiert sich an der Forschungsfrage sowie den formulierten Hypothesen. Zunächst wird die finale Stichprobe beschrieben. Anschliessend werden die deskriptiven Befunde zu den Testleistungen und Akzeptanzbewertungen präsentiert. Darauf aufbauend folgen inferenzstatistische Analysen zu Geschlechtsunterschieden, Akzeptanzunterschieden sowie möglichen Reihenfolgeeffekten. Ergänzend werden Zusammenhänge zwischen den Testergebnissen sowie explorative Zusatzanalysen zu Alterseffekten berichtet.

Alle Auswertungen basieren auf der bereinigten Stichprobe mit N = 59 Personen. Die Darstellung der Ergebnisse erfolgt wertungsfrei. Eine inhaltliche Einordnung und Interpretation der Befunde wird in Kapitel 5 vorgenommen.

4.1 Zusammenhang zwischen den Testergebnissen (IST-Screening und M-KIT)

Zur Untersuchung des Zusammenhangs zwischen den Testergebnissen des IST-Screenings und des Modul A des M-KIT wurde eine Pearson-Korrelation berechnet. Abbildung 3 zeigt ein Streudiagramm der Rohwerte.

Zwischen den beiden Testleistungen zeigte sich ein starker positiver Zusammenhang ($r = .694$, $p < .001$). Höhere Rohwerte im IST-Screening (Standard A) gingen mit höheren Gesamtwerten im M-KIT (Modul A) einher.

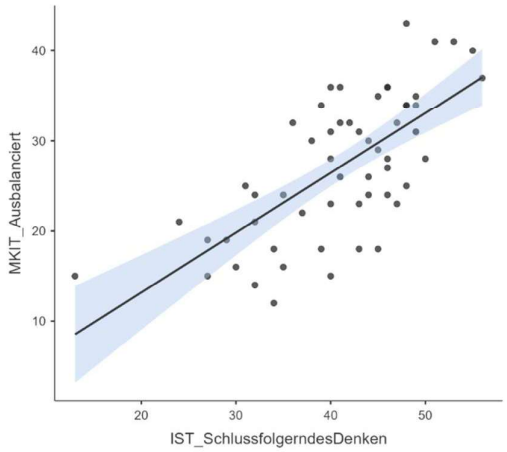

Abbildung 3: Zusammenhang zwischen den Testergebnissen

Die visuelle Inspektion des Streudiagramms zeigt eine gleichmässig ansteigende Punktwolke ohne erkennbare Clusterbildungen oder systematische Abweichungen. Auffällige Ausreisser, die den Korrelationskoeffizienten wesentlich beeinflusst hätten, waren nicht erkennbar. Die Streuung der Datenpunkte nahm mit zunehmender Testleistung weder deutlich zu noch ab, was auf eine konsistente Beziehung über den gesamten Wertebereich hinweg hindeutet.

Das Konfidenzintervall der Regressionslinie fiel vergleichsweise schmal aus, was auf eine stabile Schätzung des Zusammenhangs in der vorliegenden Stichprobe hinweist.

4.2 Geschlechtsunterschiede in der Testleistung (Hypothese 1)

Zur Überprüfung der Hypothese H1 wurden mögliche Geschlechtsunterschiede in der Testleistung mithilfe von t-Tests für unabhängige Stichproben getrennt für das IST-Screening (Standard A) und den M-KIT (Modul A) untersucht. Die deskriptiven Kennwerte der Testleistungen nach Geschlecht sind in Tabelle 2 dargestellt.

Tabelle 2: Geschlechtsunterschiede in der Testleistung

Test	Geschlecht	M	SD	t	p	d
IST-Screening	Frauen	39.4	8.61			
	Männer	43.3	7.12	1.85	.069	0.48
M-KIT	Frauen	26.8	7.56			
	Männer	27.9	8.14	0.54	.592	0.14

Für das IST-Screening erzielten Männer im Mittel höhere Rohwerte (M = 43.3) als Frauen (M = 39.4). Der Unterschied verfehlte das Signifikanzniveau knapp (p = .069) und zeigte eine Effektstärke im mittleren Bereich (d = 0.48).

Trotz dieses numerischen Unterschieds zeigten die Verteilungen der Testleistungen eine deutliche Überlappung zwischen den Geschlechtern. Die Streuung der Werte war in beiden Gruppen vergleichbar, was darauf hinweist, dass die individuellen Leistungsunterschiede innerhalb der Gruppen grösser waren als die Unterschiede zwischen den Geschlechtern.

Für den M-KIT ergaben sich nahezu identische Mittelwerte für Frauen und Männer. Der Gruppenvergleich zeigte keinen statistisch signifikanten Unterschied (p = .592). Die Effektstärke fiel mit d = 0.14 sehr gering aus, was bezüglich der vorliegenden Stichprobe nur auf minimale Leistungsunterschiede zwischen den Geschlechtern hinweist.

Insgesamt konnten für keines der beiden Testverfahren statistisch signifikante Geschlechtsunterschiede in der Testleistung festgestellt werden. Die Hypothese H1 wurde somit nicht bestätigt.

4.2.1 Explorative Zusatzanalyse: Zusammenhang zwischen Alter und Testleistung

Ergänzend wurde der Zusammenhang zwischen dem Alter der Teilnehmenden und der Testleistung explorativ untersucht. Dafür wurden Spearman-Korrelationen berechnet.

Für das IST-Screening ergab sich kein signifikanter Zusammenhang zwischen Alter und Testleistung (r = .165, p = .212). Die Korrelation fiel schwach aus.

Für den M-KIT zeigte sich hingegen ein signifikanter positiver Zusammenhang zwischen Alter und Testleistung (r = .395, p = .002). Höhere Alterswerte gingen dabei mit höheren Testleistungen einher.

Die Stärke des Zusammenhangs unterschied sich somit zwischen den beiden Testverfahren. Während beim IST-Screening nur eine schwache Beziehung beobachtet wurde, zeigte sich beim M-KIT ein moderater Zusammenhang. Die Streuung der Leistungswerte innerhalb der Altersbereiche war jedoch in beiden Tests hoch, was darauf hinweist, dass das Alter allein nur einen begrenzten Anteil der Leistungsvarianz erklärt.

4.3 Akzeptanzunterschiede zwischen den Testverfahren (Hypothese 2)

Zur Überprüfung der Hypothese H2 wurde untersucht, ob sich das IST-Screening und der M-KIT hinsichtlich der subjektiven Akzeptanz unterscheiden. Da alle Teilnehmenden beide Testverfahren bearbeiteten, erfolgten die Vergleiche mittels t-Tests für verbundene Stichproben.

4.3.1 Gesamtakzeptanz und Geschlechtsunterschiede

Auf Ebene der Gesamtakzeptanz zeigten sich sehr ähnliche Mittelwerte für beide Testverfahren. Der M-KIT wurde im Mittel geringfügig höher bewertet (M = 29.7) als das IST-Screening (M = 28.8). Dieser Unterschied war nicht statistisch signifikant.

Die Verteilung der Gesamtakzeptanzwerte zeigte für beide Testverfahren eine vergleichbare Streuung. Es ergaben sich keine Hinweise auf ausgeprägte Boden- oder Deckeneffekte. Die Bewertungen lagen überwiegend im mittleren bis oberen Skalenbereich, was auf eine insgesamt positive Einschätzung beider Testverfahren hinweist.

Hinsichtlich der Gesamtakzeptanz zeigten sich auf deskriptiver Ebene keine relevanten Unterschiede zwischen den Geschlechtern. Männer und Frauen bewerteten sowohl das IST-Screening als auch den M-KIT nahezu identisch. Die deskriptiven Kennwerte sind in Tabelle 3 dargestellt.

Tabelle 3: Gesamtakzeptanz nach Testverfahren und Geschlecht

Test	Geschlecht	N	M	SD
IST-Screening	Männer	27	28.5	5.85
	Frauen	32	29.0	4.59
M-KIT	Männer	27	29.6	3.80
	Frauen	32	29.7	5.88

4.3.2 Akzeptanzdimensionen (Einzelvergleiche)

Auf Ebene der einzelnen Akzeptanzdimensionen zeigten sich überwiegend ähnliche Bewertungen für das IST-Screening und den M-KIT. In den Dimensionen Verständlichkeit, Klarheit, Belastung, Überforderung sowie Professionalität lagen die Mittelwerte der beiden Testverfahren jeweils nahe beieinander. Die statistischen Tests ergaben in diesen Bereichen keine signifikanten Unterschiede (alle p-Werte > .05).

Ein signifikanter Unterschied zeigte sich in der Dimension *Spass*. Diese wurde für den M-KIT höher bewertet als für das IST-Screening (p = .040). Die zugehörige Effektstärke fiel mit d = -0.27 klein aus.

Für die Dimension *Relevanz* ergab sich ein Unterschied, der das konventionelle Signifikanzniveau knapp verfehlte (p = .055). Die deskriptiven Kennwerte weisen jedoch auf eine höhere durchschnittliche Bewertung des M-KIT im Vergleich zum IST-Screening hin. Eine Übersicht der Ergebnisse findet sich in Tabelle 4.

Tabelle 4: Vergleich der Akzeptanzdimensionen zwischen IST-Screening und M-KIT

Dimension	Test	M	SD	t	p	d
Verständlichkeit	IST-Screening	5.07	0.93			
	M-KIT	5.12	1.07	-0.31	.761	-0.04
Klarheit	IST- Screening	4.64	1.16			
	M-KIT	4.80	1.26	-0.89	.375	-0.12
Belastung (rev.)	IST-Screening	3.85	1.26			
	M-KIT	3.80	1.28	0.30	.766	0.04
Überforderung (rev.)	IST-Screening	3.69	1.10			
	M-KIT	3.85	1.24	-0.99	.327	-0.13
Relevanz	IST-Screening	2.81	1.37			
	M-KIT	3.14	1.27	−1.96	.055	−0.25
Professionalität	IST-Screening	4.76	0.95			
	M-KIT	4.68	1.06	0.64	.527	0.08
Spass	IST-Screening	3.95	1.44			
	M-KIT	4.29	1.23	−2.10	.040	−0.27

4.3.3 Zusammenfassung Hypothese H2

Die Ergebnisse zeigen keine ausgeprägten Unterschiede in der Gesamtakzeptanz zwischen dem IST-Screening und dem M-KIT. Ein signifikanter Unterschied zugunsten des M-KIT zeigte sich lediglich in der Dimension *Spass*. Die Hypothese H2 wurde daher teilweise bestätigt.

4.3.4 Explorative Zusatzanalyse: Zusammenhang zwischen Alter und Akzeptanz

Ergänzend wurde der Zusammenhang zwischen dem Alter der Teilnehmenden und der Akzeptanz der beiden Testverfahren explorativ untersucht. Für beide Testverfahren zeigte sich ein signifikanter positiver Zusammenhang zwischen Alter und Akzeptanz. Für das IST-Screening ergab sich eine moderate positive Korrelation (r = .38, p = .003), ebenso für den M-KIT (r = .40, p = .002). Ältere Teilnehmende bewerteten beide Testverfahren somit tendenziell positiver als jüngere.

Da das Alter als kontinuierliche Variable erfasst wurde, beziehen sich diese Ergebnisse auf eine allgemeine Tendenz innerhalb der Stichprobe und nicht auf klar abgegrenzte Altersgruppen.

4.4 Reihenfolgeeffekte (Hypothese 3)

Zur Überprüfung der Hypothese H3 wurde untersucht, ob die Reihenfolge der Testdurchführung einen Einfluss auf die Testleistung hatte. Dazu wurden die Leistungen von Teilnehmenden, die mit dem IST-Screening begannen, mit jenen verglichen, die zuerst den M-KIT bearbeiteten. Die Vergleiche erfolgten mittels t-Tests für unabhängige Stichproben.

Für das IST-Screening zeigten sich keine signifikanten Leistungsunterschiede zwischen den beiden Reihenfolgegruppen. Teilnehmende, die mit dem IST-Screening begannen, erzielten im Mittel vergleichbare Leistungen wie Teilnehmende, die zunächst den M-KIT bearbeiteten (M_1 = 42.9, M_2 = 40.1). Der statistische Test ergab keinen signifikanten Unterschied (p = .211) bei einer kleinen Effektstärke (d = .34).

Auch für den M-KIT ergaben sich keine signifikanten Unterschiede in Abhängigkeit von der Testreihenfolge. Die mittleren Leistungen der beiden Gruppen lagen nahe beieinander (M_1 = 28.7, M_2 = 26.4). Der Unterschied erwies sich ebenfalls als nicht signifikant (p = .274) und ging mit einer kleinen Effektstärke einher (d = .30).

Insgesamt ergaben sich keine Hinweise auf systematische Reihenfolgeeffekte in der Testleistung. Die Hypothese H3 wurde somit bestätigt.

4.5 Zusammenfassung der Ergebnisse

Zusammenfassend zeigen die Ergebnisse einen starken Zusammenhang zwischen den Leistungen im IST-Screening und im M-KIT. Geschlechtsunterschiede in der Testleistung fielen insgesamt gering aus und erreichten kein statistisches Signifikanzniveau. Altersbezogene Effekte zeigten sich explorativ, wobei diese testabhängig auftraten. Die subjektive Akzeptanz der beiden Testverfahren war insgesamt vergleichbar, mit einzelnen Unterschieden auf Ebene spezifischer Akzeptanzdimensionen. Reihenfolgeeffekte konnten ausgeschlossen werden.

Die inhaltliche Einordnung und Diskussion dieser Befunde erfolgt im folgenden Kapitel.

5 Diskussion

In diesem Kapitel werden die Ergebnisse der vorliegenden Studie im Kontext der theoretischen Grundlagen und des bestehenden Forschungsstands diskutiert. Ziel ist es, die empirischen Befunde kritisch einzuordnen und ihre Bedeutung im Hinblick auf die Forschungsfrage sowie die formulierten Hypothesen zu reflektieren. Dabei werden sowohl inhaltliche als auch methodische Aspekte berücksichtigt.

Ein zentrales Anliegen der Arbeit war der Vergleich zweier Online-Intelligenztestverfahren, des IST-Screenings (Standard A) und des Modul A des Modularen Kurzintelligenztests (M-KIT), unter vergleichbaren Bedingungen. Beides sind Kurzversionen längerer Verfahren (I-S-T 2000 R und M-KIT Gesamttest), die im Online-Setting durchgeführt wurden.

Die Diskussion folgt der Struktur der Forschungsfragen und Hypothesen. Zunächst werden die Leistungsbefunde der beiden Testverfahren eingeordnet. Anschliessend werden mögliche Geschlechtsunterschiede diskutiert. In den folgenden Abschnitten werden die Ergebnisse zur Testakzeptanz sowie methodische Limitationen reflektiert und Implikationen für Forschung und Praxis abgeleitet.

5.1 Zielsetzung und Einordnung

Für die Interpretation der Ergebnisse ist zentral, dass in der vorliegenden Studie die Kurzversionen des IST-Screenings und des M-KIT untersucht wurden. Beide Testverfahren wurden im Online-Setting durchgeführt und wiesen vergleichbare Rahmenbedingungen auf. Unterschiede in den Ergebnissen können daher nicht auf unterschiedliche Testdauer oder strukturell ungleiche Durchführung zurückgeführt werden, sondern sind vor dem Hintergrund der jeweiligen Testkonstruktion, Aufgabenformate und theoretischen Ausrichtung zu interpretieren.

Die beiden Verfahren verfolgen unterschiedliche diagnostische Ansätze. Das IST-Screening ist in der Tradition klassischer Intelligenztests verankert und kombiniert verschiedene Aufgabenformate, um ein breiteres Spektrum kognitiver Fähigkeiten abzubilden (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Der M-KIT verfolgt demgegenüber einen konstruktzentrierten Ansatz mit Fokus auf fluide Intelligenz und abstraktes Schlussfolgern (Dantlgraber, 2015). Diese Unterschiede sind bei der Einordnung der Ergebnisse stets mitzudenken und bilden den Rahmen für die nachfolgende Diskussion.

Ziel der Diskussion ist es daher nicht, eines der Verfahren als generell überlegen darzustellen, sondern die Befunde differenziert zu interpretieren und im Hinblick auf die jeweilige diagnostische Zielsetzung einzuordnen. Insbesondere soll geprüft werden, inwieweit sich der Anspruch des M-KIT auf Fairness und Akzeptanz auch in der Kurzversion im Online-Setting empirisch stützen lässt.

5.2 Leistungsunterschiede

Die Ergebnisse zeigen, dass die Leistungen im IST-Screening und im M-KIT insgesamt in einem vergleichbaren Bereich liegen. Dies deutet darauf hin, dass beide Kurztests grundsätzlich geeignet sind, kognitive Leistungsfähigkeit auch im Online-Setting valide zu erfassen. Gleichzeitig zeigen sich Unterschiede in der Ausprägung und Streuung der Leistungswerte, die nicht als Widerspruch, sondern als Ausdruck unterschiedlicher diagnostischer Schwerpunkte zu verstehen sind.

Das IST-Screening kombiniert sprachliche, numerische und figural-anschauliche Aufgaben und erfasst damit verschiedene Fähigkeitsbereiche innerhalb eines Tests (Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W., 2012). Diese Breite ermöglicht es, unterschiedliche kognitive Stärken zu berücksichtigen, erhöht jedoch zugleich die Abhängigkeit von sprachlichen Kompetenzen und bildungsbezogenen Vorerfahrungen. Der M-KIT fokussiert hingegen auf fluide Intelligenz und verwendet überwiegend Aufgabenformate, die etwas weniger Vorwissen erfordern (Dantlgraber, 2015). Dadurch werden grundlegende Schlussfolgerungsprozesse stärker gewichtet.

Vor diesem Hintergrund ist es plausibel, dass sich Leistungsunterschiede zwischen den beiden Tests zeigen, ohne dass diese als Hinweis auf eine unterschiedliche Testqualität interpretiert werden müssen. Vielmehr erfassen die Verfahren unterschiedliche Facetten kognitiver Leistungsfähigkeit. Personen mit Stärken im abstrakten Schlussfolgern könnten im M-KIT höhere Werte erzielen,

während Personen mit ausgeprägten sprachlichen oder numerischen Fähigkeiten im IST-Screening profitieren könnten.

Ein weiterer Aspekt betrifft die Kurzteststruktur beider Verfahren. Da die Kurzversionen mit einer begrenzten Anzahl von Items arbeiten, erhalten einzelne Aufgaben ein höheres Gewicht für das Gesamtergebnis. Dadurch können zufällige Leistungsschwankungen, etwa aufgrund von Konzentration oder situativen Einflüssen, stärker ins Gewicht fallen als bei umfangreicheren Testbatterien. Dieser Effekt ist bei beiden untersuchten Kurztests zu berücksichtigen und stellt eine generelle Einschränkung ökonomischer Testverfahren dar.

Insgesamt sprechen die Ergebnisse dafür, dass sowohl das IST-Screening als auch der M-KIT in ihren valide Leistungsinformationen liefern. Die Unterschiede zwischen den Verfahren sind vor allem auf ihre theoretische Ausrichtung und Aufgabenstruktur zurückzuführen und unterstreichen die Bedeutung einer zielgerichteten Auswahl von Testverfahren in Abhängigkeit vom diagnostischen Kontext.

5.3 Geschlechtsunterschiede in der Testleistung

Ein zentrales Ziel der vorliegenden Studie bestand darin, mögliche Geschlechtsunterschiede in den Leistungen des IST-Screenings und des Modularen Kurzintelligenztests (M-KIT) zu untersuchen. Auf theoretischer Ebene wurde angenommen, dass der M-KIT aufgrund seines Fokus auf fluide Intelligenz und der weitgehenden Reduktion von interessensbezogenem Vorwissen geringere geschlechtsbezogene Leistungsunterschiede aufweisen sollte als das IST-Screening.

Die Ergebnisse zeigen insgesamt nur geringe Unterschiede zwischen Frauen und Männern. Für keines der beiden Testverfahren konnten statistisch signifikante Geschlechtsunterschiede nachgewiesen werden. Dieser Befund steht im Einklang mit der breiten empirischen Literatur, die davon ausgeht, dass sich Frauen und Männer in der allgemeinen Intelligenz nicht systematisch unterscheiden und dass beobachtete Differenzen in der Regel klein ausfallen (Hyde, 2005; Neisser et al., 1996).

Die Hypothese H1, wonach sich geschlechtsspezifische Leistungsunterschiede zwischen den beiden Testverfahren zeigen und diese beim M-KIT geringer ausfallen sollten, konnte somit nicht bestätigt werden. Für den M-KIT ergaben sich nahezu identische Mittelwerte für Frauen und Männer, was sich auch in einer sehr kleinen Effektstärke widerspiegelte ($d = 0.14$). Dieses Ergebnis spricht gegen das Vorliegen systematischer Geschlechtsunterschiede in der Testleistung und stützt den Anspruch des M-KIT, geschlechtsneutral konstruiert zu sein.

Für das IST-Screening zeigte sich hingegen ein numerischer Unterschied zugunsten der männlichen Teilnehmenden. Die Effektstärke lag mit $d = 0.48$ im mittleren Bereich, erreichte jedoch das konventionelle Signifikanzniveau knapp nicht ($p = .069$). Dieses Ergebnismuster verdient eine differenzierte Betrachtung. Einerseits ist der Befund aus inferenzstatistischer Sicht nicht signifikant und kann daher nicht als Nachweis eines Geschlechtsunterschieds interpretiert werden. Andererseits deutet die Effektstärke darauf hin, dass geschlechtsbezogene Leistungsunterschiede beim IST-Screening zumindest nicht ausgeschlossen werden können.

Ein möglicher Erklärungsansatz liegt in der begrenzten Stichprobengrösse der vorliegenden Studie. Mit N = 59 war die statistische Power möglicherweise nicht ausreichend, um einen mittleren Effekt zuverlässig nachzuweisen. Für die Detektion eines Effekts dieser Grössenordnung wäre bei einem Signifikanzniveau von $\alpha = .05$ und einer Power von .80 eine grössere Stichprobe erforderlich (Cohen, 2009). Vor diesem Hintergrund ist es möglich, dass ein tatsächlich vorhandener Unterschied statistisch nicht signifikant wurde.

Gleichzeitig ist zu berücksichtigen, dass p-Werte in der Nähe der Signifikanzschwelle keine eindeutige Interpretation erlauben. In der aktuellen methodischen Diskussion wird betont, dass sogenannte „marginal signifikante“ Befunde weder als klarer Effekt noch als Beleg für die Abwesenheit eines Effekts verstanden werden sollten. Aus dieser Perspektive ist der vorliegende Befund als Hinweis auf eine mögliche Tendenz zu interpretieren, der jedoch einer Replikation in grösseren Stichproben bedarf.

Auffällig ist, dass sich geschlechtsspezifische Unterschiede je nach Testverfahren unterschiedlich darstellen. Während sich beim IST-Screening tendenziell grössere Unterschiede zeigen, fallen diese

beim M-KIT praktisch nicht ins Gewicht. Dieses Muster lässt sich plausibel mit den unterschiedlichen Aufgabenformaten erklären.

Vor dem Hintergrund der Kurzversionen ist dieser Befund besonders relevant. Da einzelne Items ein höheres Gewicht für das Gesamtergebnis haben, könnten potenzielle Bias-Effekte theoretisch stärker zum Tragen kommen als bei umfangreicheren Testbatterien. Dass beim M-KIT dennoch keine nennenswerten Geschlechtsunterschiede beobachtet wurden, kann als Hinweis darauf interpretiert werden, dass der Fairnessanspruch dieses Verfahrens in der vorliegenden Stichprobe weitgehend eingelöst wird.

Abschliessend ist festzuhalten, dass die fehlende Signifikanz geschlechtsspezifischer Unterschiede nicht mit dem Nachweis ihrer Abwesenheit gleichzusetzen ist. Die Interpretation von Nullbefunden erfordert eine sorgfältige Abwägung von statistischer Power, Effektstärken und theoretischer Plausibilität. Zukünftige Studien mit grösseren und gezielt geplanten Stichproben sollten prüfen, ob sich die hier beobachteten Tendenzen replizieren lassen und in welchem Ausmass sie mit den jeweiligen Testkonstruktionen zusammenhängen.

5.4 Alterseffekte in der Testleistung

Neben Leistungsunterschieden zwischen den Testverfahren und möglichen Geschlechtsunterschieden wurde in der vorliegenden Studie explorativ untersucht, ob sich altersbezogene Zusammenhänge in der Testleistung zeigen. Die Betrachtung von Alterseffekten ist insbesondere im Kontext fluider Intelligenz von Bedeutung, da zahlreiche Studien darauf hinweisen, dass fluide kognitive Fähigkeiten im Lebensverlauf anderen Veränderungen unterliegen als kristalline Fähigkeiten.

Theoretisch wird davon ausgegangen, dass fluide Intelligenz im jungen Erwachsenenalter ihren Höhepunkt erreicht und im weiteren Lebensverlauf tendenziell abnimmt (Horn & Cattell, 1966). Dieser Verlauf wird unter anderem mit altersbedingten Veränderungen in der Verarbeitungsgeschwindigkeit, der Arbeitsgedächtniskapazität und der kognitiven Flexibilität in Verbindung gebracht. Kristalline Fähigkeiten hingegen, die stärker auf Wissen und Erfahrung beruhen, bleiben häufig stabil oder nehmen weiter zu. Vor diesem Hintergrund wären altersbezogene Effekte insbesondere bei Tests zu erwarten, die primär fluide Intelligenz erfassen.

Ein explorativ untersuchter, aber bemerkenswerter Befund der vorliegenden Studie betrifft den Zusammenhang zwischen Alter und Testleistung. Während sich beim IST-Screening kein signifikanter Zusammenhang zeigte ($r = .165$, $p = .212$), ergab sich beim Modularen Kurzintelligenztest (M-KIT) ein moderater positiver Zusammenhang zwischen Alter und Testleistung ($r = .395$, $p = .002$). Höhere Alterswerte gingen dabei mit höheren Leistungen im M-KIT einher.

Dieser Befund weicht von klassischen Annahmen der Intelligenzforschung ab, wonach fluide Intelligenz im Verlauf des Erwachsenenalters typischerweise eher stabil bleibt oder abnimmt. Da der M-KIT explizit auf die Erfassung fluider Intelligenz ausgerichtet ist, wäre theoretisch eher ein negativer oder kein Zusammenhang mit dem Alter zu erwarten gewesen. Vor diesem Hintergrund erscheint eine differenzierte Einordnung erforderlich.

Ein möglicher Erklärungsansatz liegt in Selektionseffekten des verwendeten Convenience Samplings. Es ist plausibel anzunehmen, dass ältere Personen, die bereit waren, an einer zeitlich anspruchsvollen Online-Studie mit zwei Intelligenztests teilzunehmen, überdurchschnittlich motiviert, kognitiv fit und leistungsbereit waren. Jüngere Teilnehmende könnten demgegenüber eine heterogenere Gruppe dargestellt haben, die sowohl hoch- als auch niedrigmotivierte Personen umfasst. Diese unterschiedliche Selbstselektion könnte den positiven Zusammenhang zwischen Alter und Testleistung im M-KIT erklären.

Ein weiterer Erklärungsansatz betrifft Unterschiede in der Testbearbeitung. Ältere Teilnehmende könnten aufgrund grösserer Lebens- und Berufserfahrung über ausgeprägtere metakognitive Strategien, eine höhere Aufmerksamkeitsregulation oder eine systematischere Herangehensweise an Problemlösungsaufgaben verfügen. Solche Faktoren sind zwar nicht Bestandteil der fluiden Intelligenz im engeren Sinne, könnten die Testleistung jedoch indirekt beeinflussen – insbesondere in einem unbeaufsichtigten Online-Setting ohne Kontrolle der Bearbeitungsbedingungen.

Darüber hinaus ist zu berücksichtigen, dass der M-KIT in der praktischen Anwendung möglicherweise nicht ausschliesslich fluide Intelligenz erfasst, sondern auch strategische oder erfahrungsbasierte Komponenten beinhaltet, die mit dem Alter zunehmen können. Auch bei konstruktzentrierter Testentwicklung lässt sich eine vollständige Trennung fluider und kristalliner Anteile empirisch nicht immer eindeutig realisieren.

Das Fehlen eines Alterseffekts beim IST-Screening lässt sich plausibel damit erklären, dass die Kombination unterschiedlicher Aufgabenformate altersbezogene Effekte nivelliert. Während einzelne Subtests potenziell unterschiedliche Altersverläufe aufweisen, könnten sich diese im Gesamtscore gegenseitig ausgleichen. Das IST-Screening erscheint aufgrund seiner breiteren Konzeption weniger sensitiv für altersbezogene Unterschiede, während der M-KIT als fokussierter Test stärker auf altersbezogene Variationen reagiert.

Die beobachteten Alterseffekte sind jedoch vorsichtig zu interpretieren. Die Stichprobe war nicht gezielt auf Altersvergleiche ausgelegt, und individuelle Unterschiede in Bildung, beruflicher Erfahrung oder Testvertrautheit könnten altersbezogene Zusammenhänge überlagern. Zudem war die Streuung der Leistungswerte innerhalb der Altersbereiche in beiden Tests hoch, was darauf hinweist, dass das Alter allein nur einen begrenzten Anteil der Leistungsvarianz erklärt.

Im Kontext der Kurzversionen gewinnt dieser Befund zusätzliche Relevanz. Da Kurztests mit einer begrenzten Anzahl von Items arbeiten, können altersbezogene Unterschiede in einzelnen kognitiven Prozessen einen stärkeren Einfluss auf das Gesamtergebnis haben als bei umfangreicheren Testbatterien. Dies unterstreicht die Bedeutung einer differenzierten Interpretation von Testergebnissen, insbesondere bei heterogenen Altersgruppen.

Insgesamt verdeutlichen die explorativen Befunde die Notwendigkeit, altersbezogene Effekte in der Online-Intelligenzdiagnostik sorgfältig zu berücksichtigen. Die Ergebnisse erlauben keine generalisierenden Aussagen über altersbedingte Leistungsunterschiede, liefern jedoch wertvolle Hinweise darauf, dass Alter als Einflussfaktor in der Interpretation von Online-Intelligenztests eine Rolle spielen kann. Zukünftige Studien sollten diesen Zusammenhang mit gezielt altersgeschichteten Stichproben und stärker kontrollierten Designs weiter untersuchen. Eine zusätzliche Analyse auf Aufgabenebene könnte zudem Aufschluss darüber geben, welche Aufgabenformate besonders sensitiv für altersbezogene Effekte sind.

5.5 Testakzeptanz als ergänzende Perspektive der Intelligenzdiagnostik

Die Hypothese H2, wonach der Modulare Kurzintelligenztest (M-KIT) hinsichtlich der subjektiven Akzeptanz höher bewertet wird als das IST-Screening, konnte in der vorliegenden Untersuchung nicht bestätigt werden. Die Ergebnisse zeigen vielmehr, dass sich beide Testverfahren auf Ebene der Gesamtakzeptanz nur geringfügig unterscheiden. Auch in den einzelnen Akzeptanzdimensionen ergaben sich überwiegend vergleichbare Bewertungen.

Dieser Befund ist insofern bemerkenswert, als der M-KIT mit dem Anspruch entwickelt wurde, durch eine konstruktzentrierte Aufgabenstruktur, reduzierte Sprachabhängigkeit und eine klare digitale Umsetzung eine hohe Akzeptanz zu erzielen (Dantlgraber, 2015). Gleichzeitig legen die Ergebnisse nahe, dass auch das IST-Screening im Online-Setting von den Teilnehmenden als angemessen, verständlich und professionell wahrgenommen wurde. Die Akzeptanzwerte beider Verfahren lagen im mittleren bis oberen Skalenbereich, was insgesamt auf eine positive Bewertung der Testsituation hinweist.

Eine mögliche Interpretation dieses Ergebnisses ist, dass Unterschiede auf konzeptioneller Ebene nicht zwangsläufig zu deutlich wahrnehmbaren Unterschieden in der subjektiven Akzeptanz führen. Obwohl sich IST-Screening und M-KIT hinsichtlich theoretischer Ausrichtung und Aufgabenformate unterscheiden, scheinen beide Verfahren aus Sicht der Testpersonen vergleichbar gut handhabbar zu sein. Das Ausbleiben eines Akzeptanzvorteils des M-KIT kann daher weniger als Schwäche dieses Verfahrens verstanden werden, sondern vielmehr als Hinweis darauf, dass auch etablierte klassische Tests im digitalen Kontext akzeptabel umgesetzt werden können.

Gleichzeitig sind methodische Aspekte der Akzeptanzerhebung zu berücksichtigen. Die subjektive Akzeptanz wurde mithilfe eines von der Autorin dieser Arbeit gekürzten und adaptierten Fragebogens auf Basis des AKZEPT-L erfasst. Durch die Reduktion der Itemanzahl konnten einzelne

Akzeptanzdimensionen nur eingeschränkt differenziert abgebildet werden. Es ist daher nicht auszuschliessen, dass feinere Unterschiede zwischen den Testverfahren mit einem umfangreicheren Instrument deutlicher hervorgetreten wären. Zudem erfolgte die Erhebung der Akzeptanz retrospektiv nach Abschluss der Testbearbeitung, wodurch situative Schwankungen während der Bearbeitung oder unmittelbare Reaktionen auf einzelne Aufgabenformate nicht erfasst werden konnten.

Ein weiterer relevanter Faktor betrifft die Zusammensetzung der Stichprobe. Die Teilnehmenden wurden im Rahmen eines Convenience Samplings rekrutiert und erklärten sich bereit, an einer zeitlich anspruchsvollen Online-Studie mit zwei Intelligenztests teilzunehmen. Es ist anzunehmen, dass diese Personen eine gewisse Offenheit gegenüber psychologischen Testverfahren mitbringen und insgesamt eine eher positive Haltung gegenüber Testungen aufweisen. In stärker selektiven oder hochrelevanten Anwendungskontexten, etwa in der Personalauswahl, könnten Akzeptanzurteile kritischer ausfallen und sich Unterschiede zwischen Testverfahren deutlicher zeigen.

Auf Ebene einzelner Akzeptanzdimensionen zeigte sich ein signifikanter Unterschied zugunsten des M-KIT in der Dimension *Spass*. Dieser Befund deutet darauf hin, dass spezifische Gestaltungsmerkmale der Tests – etwa Aufgabenformat oder visuelle Umsetzung – die subjektive Testerfahrung beeinflussen können, ohne sich zwingend in der globalen Akzeptanz niederzuschlagen. Damit wird deutlich, dass Akzeptanz kein eindimensionales Konstrukt darstellt, sondern aus mehreren Teilaspekten besteht, die von Testpersonen unterschiedlich gewichtet werden können (Kersting, 2008).

Aus diagnostischer Perspektive ist es wichtig, Akzeptanz nicht mit diagnostischer Qualität gleichzusetzen (Kersting, 2008). Ein Test kann als angenehm oder ansprechend erlebt werden, ohne zwingend überlegene psychometrische Eigenschaften aufzuweisen. Umgekehrt kann ein diagnostisch hochwertiges Verfahren auch dann wertvolle Informationen liefern, wenn es subjektiv als fordernd oder wenig unterhaltsam wahrgenommen wird. Akzeptanz ist daher als ergänzende, jedoch nicht hinreichende Bedingung für qualitativ hochwertige Diagnostik zu verstehen.

Gerade im unbeaufsichtigten Online-Setting kommt der Testakzeptanz dennoch eine besondere Bedeutung zu. Ohne direkte Testaufsicht hängt die Qualität der erhobenen Leistungsdaten in hohem Masse vom Engagement der Testpersonen ab. Eine grundsätzlich positive Wahrnehmung des Testverfahrens kann dazu beitragen, Motivation und Bearbeitungsqualität aufrechtzuerhalten, insbesondere bei ökonomischen Kurztests, bei denen einzelne Aufgaben einen relativ grossen Einfluss auf das Gesamtergebnis haben.

Vor diesem Hintergrund erscheint es sinnvoll, Akzeptanz systematisch in die Evaluation von Intelligenztests einzubeziehen, insbesondere bei digitalen Verfahren und in Anwendungsfeldern mit hoher praktischer Relevanz. Die vorliegende Studie zeigt, dass sich Akzeptanzaspekte auch dann sinnvoll erfassen lassen, wenn sich die Leistungsdaten nur geringfügig unterscheiden. Damit erweitert sie den diagnostischen Blick über rein leistungsbezogene Kriterien hinaus.

Zusammenfassend verdeutlichen die Ergebnisse zur Testakzeptanz, dass subjektive Bewertungen eine eigenständige Informationsquelle darstellen, die jedoch stets im Zusammenspiel mit objektiven Leistungs- und Gütekriterien interpretiert werden sollte. Akzeptanz allein erlaubt keine abschliessende Bewertung der Qualität eines Testverfahrens, kann jedoch wichtige Hinweise auf dessen Zumutbarkeit, Anwendungsrisiken und potenzielle Akzeptanzbarrieren liefern. In diesem Sinne ergänzt die Akzeptanzperspektive die klassische Intelligenzdiagnostik, ohne diese zu ersetzen.

5.6 Zusammenhang zwischen Alter und Testakzeptanz

Neben der Betrachtung von Alterseffekten in der Testleistung wurde in der vorliegenden Studie auch untersucht, ob das Alter der Teilnehmenden mit der subjektiven Akzeptanz der Testverfahren zusammenhängt. Diese Fragestellung ist insbesondere im Kontext der Online-Intelligenzdiagnostik relevant, da digitale Testformate von unterschiedlichen Altersgruppen potenziell unterschiedlich wahrgenommen werden können.

Die Ergebnisse zeigen einen positiven Zusammenhang zwischen dem Alter der Teilnehmenden und der Akzeptanz beider Testverfahren. Ältere Personen bewerteten sowohl das Modul A des Modularen Kurzintelligenztests (M-KIT) als auch das IST-Screening (Standard A) tendenziell positiver als jüngere

Teilnehmende. Dieser Befund deutet darauf hin, dass die subjektive Wahrnehmung der Testsituation mit zunehmendem Alter günstiger ausfällt.

Ein möglicher Erklärungsansatz liegt darin, dass ältere Personen strukturierte Testsituationen als weniger belastend oder als vertrauter empfinden. Mit zunehmender Berufs- und Lebenserfahrung könnten standardisierte Leistungssituationen stärker als sachlich und kontrollierbar wahrgenommen werden, während sie bei jüngeren Personen eher Leistungsdruck oder Vergleichsorientierung auslösen. Auch unterschiedliche Erwartungshaltungen gegenüber Intelligenztests könnten eine Rolle spielen.

Darüber hinaus ist denkbar, dass ältere Teilnehmende eine geringere Tendenz zum sozialen Vergleich aufweisen und Testergebnisse weniger stark als Bewertung der eigenen Person interpretieren. Jüngere Personen hingegen könnten sensibler auf die implizite Leistungsbewertung reagieren, was sich in einer kritischeren oder ambivalenteren Akzeptanz niederschlagen kann. Solche altersbezogenen Unterschiede in der subjektiven Bedeutung der Testsituation könnten erklären, warum sich die Akzeptanz mit zunehmendem Alter erhöht.

Ein weiterer Aspekt betrifft die Gestaltung der Testverfahren selbst. Beide Tests sind klar strukturiert, folgen einem transparenten Ablauf und stellen keine besonderen technischen Anforderungen. Diese Eigenschaften könnten insbesondere für ältere Teilnehmende zu einer positiven Nutzererfahrung beitragen, da sie Orientierung bieten und Unsicherheiten im Umgang mit dem Test reduzieren. Gleichzeitig scheint die digitale Umsetzung der Tests keine altersbedingten Akzeptanzbarrieren erzeugt zu haben.

Es ist jedoch zu betonen, dass es sich bei der Analyse des Zusammenhangs zwischen Alter und Testakzeptanz um eine explorative Zusatzanalyse handelt. Die Stichprobe war nicht gezielt auf Altersvergleiche ausgelegt, und potenzielle Einflussfaktoren wie digitale Vorerfahrung, Bildung oder frühere Testerfahrungen wurden nicht differenziert erfasst. Die dargestellten Erklärungen sind daher als vorsichtige Interpretationsansätze zu verstehen und nicht als höchst stringente Schlussfolgerungen.

Zusammenfassend zeigen die Ergebnisse, dass ältere Teilnehmende die untersuchten Kurzversionen (IST-Screening und Modul A des M-KIT) tendenziell positiver bewerten als jüngere Personen. Dieser Befund ergänzt die Diskussion der Akzeptanz insgesamt und unterstreicht, dass subjektive Bewertungen von Testverfahren nicht unabhängig von personenspezifischen Merkmalen wie dem Alter betrachtet werden sollten. Gleichzeitig ergeben sich daraus Ansatzpunkte für zukünftige Forschung, die den Zusammenhang zwischen Alter, Testerfahrung und Akzeptanz systematischer untersuchen können.

5.7 Reihenfolgeeffekte

Ein weiterer Untersuchungsfokus der vorliegenden Studie betraf mögliche Reihenfolgeeffekte bei der Testdurchführung. Ziel war es zu prüfen, ob die Reihenfolge, in der das IST-Screening und der Modulare Kurzintelligenztest (M-KIT) bearbeitet wurden, einen Einfluss auf die Testleistung hatten. Solche Effekte können insbesondere in Studien mit mehreren Testverfahren relevant sein, da Ermüdung, Übung oder veränderte Motivation die Ergebnisse verzerren können.

Die Ergebnisse zur Hypothese H3 zeigten, dass die Reihenfolge der Testdurchführung keinen signifikanten Einfluss auf die Leistung hatte. Teilnehmende, die zunächst das IST-Screening und anschliessend den M-KIT bearbeiteten, erzielten vergleichbare Leistungen wie Teilnehmende, die die umgekehrte Reihenfolge durchliefen. Dieser Befund spricht gegen das Vorliegen systematischer Ermüdungs- oder Übungseffekte und stärkt die interne Validität der Untersuchung.

Gerade im Kontext unbeaufsichtigter Online-Studien ist dieser Befund von besonderer Bedeutung. Online-Testungen werden häufig kritisch betrachtet, da Testleitende keinen direkten Einfluss auf Pausen, Konzentration oder Bearbeitungsstrategien haben. Insbesondere bei längeren oder sequenziellen Erhebungen besteht die Befürchtung, dass nachlassende Aufmerksamkeit oder mentale Ermüdung die Leistung im zweiten Test beeinträchtigen könnten. Die vorliegenden Ergebnisse liefern hierfür keine Hinweise.

Ein möglicher Erklärungsansatz liegt in der Struktur der Untersuchung. Beide Testverfahren waren in ihrer Kurzversion konzipiert und wiesen vergleichbare Rahmenbedingungen auf. Die Tests folgten klaren Instruktionen und wurden jeweils durch eine Akzeptanzbefragung abgeschlossen, was den Teilnehmenden eine kurze kognitive Zäsur zwischen den Leistungstests ermöglichte. Diese Struktur könnte dazu beigetragen haben, potenzielle Ermüdungseffekte zu reduzieren.

Darüber hinaus ist zu berücksichtigen, dass sich die beiden Testverfahren in ihren Aufgabenformaten unterscheiden. Während das IST-Screening verschiedene Fähigkeitsbereiche kombiniert, fokussiert der M-KIT stärker auf abstraktes Schlussfolgern. Diese Unterschiede könnten dazu geführt haben, dass sich die Bearbeitung der beiden Tests subjektiv weniger monoton anfühlte, was wiederum die Wahrscheinlichkeit von Ermüdungseffekten reduziert haben könnte.

Der fehlende Reihenfolgeeffekt lässt zudem darauf schliessen, dass sich keine relevanten Übungseffekte zwischen den Tests ergeben haben. Obwohl beide Verfahren kognitive Leistungsfähigkeit erfassen, unterscheiden sie sich in ihrer konkreten Aufgabenstruktur ausreichend, sodass eine direkte Übertragung von Lösungsstrategien vom ersten auf den zweiten Test unwahrscheinlich ist. Dies ist insbesondere im Vergleich zweier Intelligenztests von Bedeutung, da inhaltliche Überschneidungen potenziell zu Lern- oder Trainingseffekten führen könnten.

Insgesamt sprechen die Ergebnisse dafür, dass die gewählte Untersuchungsstruktur methodisch angemessen war. Die fehlenden Reihenfolgeeffekte stärken die Aussagekraft der Leistungsbefunde und sprechen dafür, dass die beobachteten Unterschiede zwischen den Testverfahren nicht auf artefaktbedingte Verzerrungen zurückzuführen sind. Damit liefert die Analyse der Reihenfolgeeffekte einen wichtigen Beitrag zur Bewertung der internen Validität der Studie und unterstützt die Interpretation der Ergebnisse als primär test- und konstruktbedingt.

5.8 Methodische Limitationen

Trotz der sorgfältigen Planung und Durchführung der vorliegenden Studie sind verschiedene methodische Limitationen zu berücksichtigen, die die Interpretation der Ergebnisse einschränken. Diese betreffen insbesondere die Stichprobe, das Studiendesign, die Durchführung im unbeaufsichtigten Online-Setting sowie die eingesetzten Messinstrumente. Die kritische Reflexion dieser Aspekte ist notwendig, um die Aussagekraft der Befunde realistisch einzuordnen und ihre Generalisierbarkeit angemessen zu bewerten.

5.8.1 Stichprobe und Rekrutierung

Eine zentrale Limitation ergibt sich aus der Zusammensetzung der Stichprobe. Die Teilnehmenden wurden im Rahmen eines Convenience Samplings rekrutiert, wodurch keine Repräsentativität für die Allgemeinbevölkerung oder spezifische Zielgruppen beansprucht werden kann. Insbesondere ist davon auszugehen, dass Personen, die bereit waren, an einer zeitlich anspruchsvollen Online-Studie mit zwei Intelligenztests teilzunehmen, überdurchschnittlich motiviert, leistungsbereit oder grundsätzlich offener gegenüber psychologischer Diagnostik sind.

Diese Selbstselektion könnte dazu geführt haben, dass Personen mit stärkerer Testaversion, geringerem Vertrauen in Intelligenztests oder niedrigerer intrinsischer Motivation unterrepräsentiert sind. Gerade im Kontext der Untersuchung von Testakzeptanz ist dieser Aspekt relevant, da die erhobenen Akzeptanzwerte potenziell günstiger ausfallen könnten als in einer zufällig gezogenen Stichprobe. Die Ergebnisse zur Akzeptanz sollten daher nicht ohne Weiteres auf andere Kontexte übertragen werden, insbesondere nicht auf hochrelevante Entscheidungssituationen wie Personalauswahl oder Leistungsselektion.

Darüber hinaus weist die Stichprobe eine breite Altersstreuung auf, ohne dass die Rekrutierung gezielt altersstratifiziert erfolgte. Zwar ermöglichte dies explorative Analysen zu Alterseffekten, gleichzeitig erschwert es jedoch die kontrollierte Interpretation altersbezogener Unterschiede. Alter ist eng mit weiteren Variablen wie Bildung, beruflicher Erfahrung oder digitaler Kompetenz verknüpft, die in der vorliegenden Studie nicht systematisch erfasst wurden und potenzielle Konfundierungen darstellen.

5.8.2 Studiendesign und Testreihenfolge

Das gewählte Within-Subjects-Design stellt grundsätzlich eine Stärke der Studie dar, da interindividuelle Unterschiede weitgehend kontrolliert werden können. Gleichzeitig bringt dieses Design spezifische Herausforderungen mit sich. Auch wenn keine signifikanten Reihenfolgeeffekte nachgewiesen wurden, kann nicht vollständig ausgeschlossen werden, dass subtile Ermüdungs-, Übungs- oder Motivationseffekte auf individueller Ebene aufgetreten sind, die durch die statistischen Analysen nicht erfasst wurden.

Zudem war die Testreihenfolge nicht vollständig balanciert. Ein grösserer Teil der Teilnehmenden absolvierte zuerst den M-KIT und anschliessend das IST-Screening. Auch wenn dieser Umstand in den Analysen berücksichtigt wurde, könnte die ungleiche Verteilung die Sensitivität der Reihenfolgeanalyse reduziert haben. Insbesondere bei kleineren Stichproben können solche Ungleichgewichte dazu führen, dass kleinere Effekte statistisch nicht erkennbar sind.

Ein weiterer Aspekt betrifft die Gesamtdauer der Untersuchung. Die Bearbeitung zweier Intelligenztests innerhalb einer Sitzung stellt eine erhebliche kognitive Beanspruchung dar. Auch wenn die Tests als Kurzversionen konzipiert sind, kann nicht ausgeschlossen werden, dass Ermüdung oder nachlassende Konzentration die Bearbeitungsqualität einzelner Teilnehmender beeinflusst haben, insbesondere gegen Ende der Untersuchung.

5.8.3 Online-Setting und fehlende Kontrolle der Testbedingungen

Eine wesentliche Limitation ergibt sich aus der unbeaufsichtigten Online-Durchführung der Tests. Obwohl standardisierte Instruktionen bereitgestellt wurden, konnte nicht kontrolliert werden, unter welchen Bedingungen die Teilnehmenden die Tests bearbeiteten. Unterschiede in der Umgebung (z. B. Ablenkungen, Unterbrechungen), der technischen Ausstattung (z. B. Bildschirmgrösse, Eingabegeräte) oder der individuellen Pausengestaltung könnten sowohl die Testleistung als auch die subjektive Akzeptanz beeinflusst haben.

Gerade bei leistungsdiagnostischen Verfahren stellt die fehlende Kontrolle der Bearbeitungsbedingungen eine potenzielle Verzerrungsquelle dar. Auch wenn Online-Testungen zunehmend verbreitet sind und eine hohe ökologische Validität aufweisen, bleibt die Frage offen, inwieweit Unterschiede in den Rahmenbedingungen systematische Effekte erzeugt haben könnten. Diese Limitation betrifft beide Testverfahren gleichermassen, schränkt jedoch die Vergleichbarkeit mit Studien in kontrollierten Präsenzsettings ein.

5.8.4 Messinstrumente und Operationalisierung der Akzeptanz

Die Erfassung der Testakzeptanz erfolgte mithilfe eines gekürzten und adaptierten Fragebogens auf Basis des AKZEPT-L (Kersting, 2008). Diese Reduktion war aus praktischen Gründen sinnvoll, um die Belastung der Teilnehmenden zu begrenzen und die Durchführbarkeit der Studie sicherzustellen. Gleichzeitig ist davon auszugehen, dass die Reliabilität und Differenziertheit der Akzeptanzmessung gegenüber dem Originalinstrument eingeschränkt ist.

Einzelne Akzeptanzdimensionen wurden nicht mit vielen Items abgebildet, was die Sensitivität für feine Unterschiede zwischen den Testverfahren reduziert haben könnte. Zudem wurde die Akzeptanz ausschliesslich retrospektiv nach der Testbearbeitung erfasst. Situative Schwankungen während der Bearbeitung oder unmittelbare emotionale Reaktionen auf einzelne Aufgabenformate konnten dadurch nicht differenziert abgebildet werden.

Auch die Verwendung einer Schulnote als globale Akzeptanzbewertung ist kritisch zu betrachten. Zwar stellt sie ein intuitives und leicht verständliches Mass dar, sie unterliegt jedoch individuellen Bewertungsstandards und kann interindividuell unterschiedlich interpretiert werden. Die Schulnote sollte daher primär als ergänzende Information und nicht als eigenständiges Qualitätskriterium verstanden werden.

5.8.5 Statistische Power und explorative Analysen

Die Stichprobengrösse der vorliegenden Studie war ausreichend, um mittlere Effekte zu detektieren, jedoch begrenzt im Hinblick auf die Identifikation kleiner Effekte. Dies ist insbesondere im Kontext der

Untersuchung von Geschlechts- und Alterseffekten relevant, da in der Intelligenzforschung häufig nur geringe Mittelwertsunterschiede auftreten. Nicht signifikante Befunde sollten daher nicht als Beleg für das vollständige Fehlen entsprechender Effekte interpretiert werden. Insbesondere der knapp nichtsignifikante Geschlechtseffekt beim IST-Screening ($d = 0.48$, $p = 0.069$) weckt das Interesse an einer vertieften Untersuchung.

Darüber hinaus hatten mehrere Analysen explorativen Charakter, insbesondere die Untersuchungen zu Alterseffekten und Zusammenhängen zwischen Akzeptanz und personenbezogenen Variablen. Diese Analysen liefern wertvolle Hinweise, erlauben jedoch keine kausalen Schlussfolgerungen. Die Ergebnisse sind daher primär als hypothesengenerierend zu verstehen und bedürfen einer Replikation in grösseren und gezielt geplanten Stichproben.

5.8.6 Gesamtwürdigung der Limitationen

Insgesamt zeigen die methodischen Limitationen, dass die Ergebnisse der vorliegenden Studie mit Vorsicht zu interpretieren sind. Viele der genannten Einschränkungen sind typisch für empirische Online-Forschung und betreffen insbesondere Studien, die leistungsdiagnostische Verfahren unter realitätsnahen Bedingungen untersuchen. Die kritische Reflexion dieser Aspekte schmälert nicht den Erkenntniswert der Arbeit, sondern trägt im Gegenteil zu einer transparenten und verantwortungsvollen Einordnung der Befunde bei.

Gleichzeitig verdeutlichen die Limitationen die Notwendigkeit weiterführender Forschung, insbesondere mit grösseren, differenzierter rekrutierten Stichproben, stärker kontrollierten Designs und erweiterten Akzeptanzmessungen. Die vorliegende Studie kann in diesem Sinne als explorativer Beitrag verstanden werden, der zentrale Fragestellungen adressiert und Ansatzpunkte für zukünftige Untersuchungen liefert.

6 Fazit

Ziel der vorliegenden Arbeit war es, zwei Online-Intelligenztestverfahren – das IST-Screening und das Modul A des Modularen Kurzintelligenztests (M-KIT) – systematisch miteinander zu vergleichen. Im Fokus standen dabei die Kurzversionen beider Verfahren, die unter vergleichbaren Bedingungen im Online-Setting durchgeführt wurden. Untersucht wurden Unterschiede in der Testleistung, mögliche Geschlechts- und Alterseffekte sowie die subjektive Akzeptanz der Testverfahren.

Die Ergebnisse zeigen, dass beide Kurztests grundsätzlich geeignet sind, kognitive Leistungsfähigkeit im Online-Setting zu erfassen. Zwischen den Testergebnissen des IST-Screenings und des M-KIT bestand ein deutlicher Zusammenhang, was darauf hindeutet, dass beide Verfahren einen gemeinsamen Kern kognitiver Leistungsfähigkeit abbilden. Gleichzeitig wurden Unterschiede sichtbar, die sich plausibel durch die unterschiedliche theoretische Ausrichtung und Aufgabenstruktur der Tests erklären lassen. Während das IST-Screening mehrere Fähigkeitsbereiche kombiniert, fokussiert der M-KIT gezielt auf fluide Intelligenz und abstraktes Schlussfolgern. Die Befunde sprechen somit nicht für eine generelle Überlegenheit eines Verfahrens, sondern verdeutlichen die Bedeutung der diagnostischen Zielsetzung bei der Auswahl eines Tests.

In Bezug auf Geschlechtsunterschiede zeigten sich insgesamt nur geringe Effekte. Diese Ergebnisse stehen im Einklang mit der bestehenden Forschung, die davon ausgeht, dass sich Frauen und Männer in der allgemeinen Intelligenz nicht systematisch unterscheiden. Auffällig war jedoch, dass sich beim M-KIT in der vorliegenden Stichprobe geringere geschlechtsbezogene Unterschiede zeigten (d = 0.14) als beim IST-Screening (d = 0.48). Trotz fehlender Signifikanz beider Effekte kann dies als Hinweis interpretiert werden, dass der konstruktzentrierte Ansatz und die Reduktion von aufgabenrelevantem Vorwissen zur Geschlechtsneutralität beitragen. Besonders interessant ist dieser Befund vor dem Hintergrund, dass es sich um Kurzversionen handelt, bei denen einzelne Items ein höheres Gewicht für das Gesamtergebnis haben.

Auch Alterseffekte können differenziert betrachtet werden. Während sich in der Testleistung altersbezogene Unterschiede insbesondere beim M-KIT zeigten, war die Akzeptanz der Testverfahren mit zunehmendem Alter tendenziell höher. Ältere Teilnehmende bewerteten beide Tests insgesamt positiver als jüngere Personen. Diese Befunde unterstreichen, dass Leistungs- und Akzeptanzaspekte nicht zwingend parallel verlaufen und dass subjektive Bewertungen von Testverfahren von personenspezifischen Faktoren beeinflusst werden.

Die Analyse der Testakzeptanz zeigte insgesamt eine gute Akzeptanz beider Verfahren. Trotz unterschiedlicher Aufgabenformate und theoretischer Ausrichtungen wurden sowohl das IST-Screening als auch der M-KIT von den Teilnehmenden als angemessen und durchführbar wahrgenommen. Dies ist insbesondere im Kontext unbeaufsichtigter Online-Testungen von Bedeutung, da Akzeptanz einen wichtigen Einfluss auf Motivation und Bearbeitungsqualität haben kann. Unterschiede in der Akzeptanz fielen insgesamt gering aus und scheinen weniger von der Teststruktur als von individuellen Einstellungen gegenüber Leistungsdiagnostik geprägt zu sein.

Die Untersuchung der Reihenfolgeeffekte ergab keine Hinweise auf systematische Ermüdungs- oder Übungseffekte. Die Reihenfolge der Testdurchführung hatte keinen signifikanten Einfluss auf die Testergebnisse. Dieser Befund stärkt die interne Validität der Studie und spricht dafür, dass die gewählte Untersuchungsstruktur methodisch angemessen war.

Trotz dieser Befunde sind die Ergebnisse vor dem Hintergrund mehrerer methodischer Limitationen zu interpretieren. Die Stichprobe war nicht repräsentativ, die Tests wurden unbeaufsichtigt durchgeführt, und einzelne Analysen hatten explorativen Charakter. Zudem lassen sich die Ergebnisse der Kurzversionen nicht ohne Weiteres auf die Langversionen der Tests übertragen. Diese Einschränkungen relativieren die Generalisierbarkeit der Befunde, mindern jedoch nicht ihren Erkenntniswert im untersuchten Kontext.

Zusammenfassend leistet die vorliegende Arbeit einen Beitrag zur Evaluation moderner Online-Intelligenzdiagnostik. Sie zeigt, dass sich zentrale Entwicklungsansprüche wie Fairness und Akzeptanz auch bei Kurzversionen von Intelligenztests empirisch untersuchen lassen. Gleichzeitig verdeutlicht sie, dass Unterschiede zwischen Testverfahren differenziert interpretiert werden müssen und nicht allein auf Effizienz oder Testdauer reduziert werden können. Für zukünftige Forschung ergeben sich Ansatzpunkte, insbesondere im Hinblick auf grössere und differenziertere Stichproben,

die gezielte Untersuchung von Altersgruppen sowie den Vergleich von Kurz- und Langversionen innerhalb desselben Testverfahrens.

6.1 Ausblick und Implikationen

Die Ergebnisse der vorliegenden Arbeit liefern verschiedene Ansatzpunkte für weiterführende Forschung sowie für die praktische Anwendung von Online-Intelligenztests. Dabei zeigen sich sowohl methodische als auch inhaltliche Fragestellungen, die über den Rahmen der vorliegenden Untersuchung hinausweisen.

Aus Forschungsperspektive verdeutlichen die Befunde zunächst die Bedeutung einer eigenständigen Betrachtung von Kurzversionen psychologischer Testverfahren. Während zu den Langversionen des IST-Screenings und des Modul A des Modularen Kurzintelligenztests bereits empirische Erkenntnisse vorliegen, zeigen die Ergebnisse dieser Arbeit, dass Aussagen zur Fairness, Akzeptanz und Leistungsstruktur nicht ohne Weiteres von Lang- auf Kurzversionen übertragen werden können. Zukünftige Studien sollten daher systematisch untersuchen, inwieweit sich psychometrische Eigenschaften, Geschlechtsneutralität und Akzeptanz zwischen Kurz- und Langversionen desselben Testverfahrens unterscheiden.

Darüber hinaus erscheint eine differenziertere Analyse altersbezogener Effekte sinnvoll. Die vorliegenden Ergebnisse deuten darauf hin, dass Alter sowohl mit der Testleistung als auch mit der Akzeptanz von Intelligenztests in Zusammenhang stehen kann. Künftige Forschungsarbeiten könnten diesen Aspekt gezielt vertiefen, etwa durch altersstratifizierte Stichproben oder längsschnittliche Designs, um Entwicklungsverläufe und altersbedingte Veränderungen genauer abzubilden. Auch der Einbezug weiterer Variablen wie digitale Vorerfahrung oder beruflicher Hintergrund könnte dazu beitragen, altersbezogene Effekte besser zu verstehen.

Ein weiterer relevanter Forschungsansatz betrifft die Kombination leistungsbezogener und subjektiver Kriterien. Die vorliegende Arbeit zeigt, dass Akzeptanz eine eigenständige und relevante Dimension der Testbewertung darstellt. Zukünftige Studien könnten untersuchen, wie Akzeptanz, Motivation und Bearbeitungsqualität miteinander zusammenhängen und in welchem Ausmass subjektive Bewertungen die Validität von Testergebnissen beeinflussen. Insbesondere im Online-Setting erscheint eine stärkere Integration dieser Perspektiven vielversprechend.

Neben den Implikationen für die Forschung ergeben sich auch Konsequenzen für die praktische Anwendung von Intelligenztests. Die Ergebnisse sprechen dafür, dass Kurzversionen von Intelligenztests im Online-Setting grundsätzlich einsetzbar sind, sofern ihre Einsatzgrenzen berücksichtigt werden. Sie ermöglichen eine ökonomische Erfassung kognitiver Leistungsfähigkeit, sollten jedoch nicht als vollwertiger Ersatz für umfassendere diagnostische Verfahren verstanden werden. Insbesondere in Kontexten mit weitreichenden Entscheidungsfolgen ist eine sorgfältige Auswahl und Kombination diagnostischer Instrumente erforderlich.

Der Vergleich zwischen IST-Screening und dem Modul A des M-KIT verdeutlicht zudem, dass die Auswahl eines Testverfahrens stets an der diagnostischen Zielsetzung orientiert sein sollte. Während breit angelegte Verfahren ein umfassenderes Leistungsprofil liefern können, bieten konstruktzentrierte Tests Vorteile im Hinblick auf Fairness und Vergleichbarkeit. Praktikerinnen und Praktiker sind daher gefordert, nicht allein auf Effizienz oder Testdauer zu fokussieren, sondern auch theoretische Ausrichtung, Aufgabenformate und Akzeptanz der Testpersonen in ihre Entscheidung einzubeziehen.

Schliesslich unterstreichen die Ergebnisse die Bedeutung der Testakzeptanz als Qualitätsmerkmal moderner Diagnostik. Eine hohe Akzeptanz kann dazu beitragen, Motivation und Bearbeitungsqualität zu fördern und damit indirekt auch die Aussagekraft der Testergebnisse zu erhöhen. In der Praxis sollte Akzeptanz daher nicht nur als Nebenaspekt, sondern als integraler Bestandteil der Testevaluation verstanden werden.

Insgesamt zeigt der Ausblick, dass die vorliegende Arbeit über den konkreten Vergleich zweier Testverfahren hinaus relevante Impulse für die Weiterentwicklung der Online-Intelligenzdiagnostik liefert. Sie macht deutlich, dass moderne Testverfahren nicht nur an psychometrischen Gütekriterien gemessen werden sollten, sondern auch an ihrer Fairness, Akzeptanz und Passung zum jeweiligen Anwendungskontext.

Literaturverzeichnis

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies* (1. Auflage). Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22. https://doi.org/10.1037/h0046743

Cohen, J. (2009). *Statistical power analysis for the behavioral sciences* (2. ed., reprint.). New York, NY: Psychology Press.

Dantlgraber, M. (2015). *M-KIT. Modularer Kurzintelligenztest. Manual*. Testmanual. Bern: Hogrefe AG.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (Springer-Lehrbuch). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41089-5

Goldberg, Lewis R. (2018). Doing it all online: The challenges of internet-based testing. *Journal of Personality Assessment*, *100*(1), 1–6.

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. https://doi.org/10.1037/h0023816

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*(6), 581–592. https://doi.org/10.1037/0003-066X.60.6.581

Jensen, A. R. (1998). *The g factor: the science of mental ability* (Human evolution, behavior, and intelligence) (1. publ.). Westport, Conn.: Praeger.

Kersting, M. (1998). Differentielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren. *Zeitschrift für Arbeits- und Organisationspsychologie*, *42*, 61–75.

Kersting, M. (2000). Rezension des “Intelligenz-Struktur-Test 2000” von R. Amthauer, B. Brocke, D. Liepmann und A. Beauducel. *Zeitschrift für Arbeits- und Organisationspsychologie A&O*, *44*(2), 96–101. https://doi.org/10.1026//0932-4089.44.2.96

Kersting, M. (2008). Zur Akzeptanz von Intelligenz- und Leistungstests. *Report Psychologie*, *33*, 420–433.

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

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

Matlin, M. W. & Halpern, D. F. (1988). Sex Differences in Cognitive Abilities. *The American Journal of Psychology*, *101*(3), 451. https://doi.org/10.2307/1423092

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. https://doi.org/10.1016/j.intell.2008.08.004

Messick, S. (1989). Validity. In Linn, Robert L. (Hrsg.), *Educational Measurement* (3. Auflage, S. 13–103). New York: Macmillan.

Moosbrugger, H. & Kelava, A. (Hrsg.). (2020). *Testtheorie und Fragebogenkonstruktion* (Lehrbuch) (3., vollständig neu bearbeitete, erweiterte und aktualisierte Auflage.). Berlin: Springer.

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

Rammstedt, B. & Beierlein, C. (2014). Can't We Make It Any Shorter?: The Limits of Personality Assessment and Ways to Overcome Them. *Journal of Individual Differences*, *35*(4), 212–220. https://doi.org/10.1027/1614-0001/a000141

Schuler, H. (2014). *Psychologische Personalauswahl* (3. Auflage). Göttingen: Hogrefe.

Spearman, C. (1904). „General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201. https://doi.org/10.2307/1412107

Steiner, H. & Lieberei, W. (2024). Einsatzbereiche von Online-Tests. In H. Steiner (Hrsg.), *Online-Assessment* (S. 3–20). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-68684-3_1

Sternberg, R. J. (2021). "Social policy and intelligence" Redux: a tribute to Edward Zigler. *Development and Psychopathology*, *33*(2), 522–532. https://doi.org/10.1017/S0954579420000693

Ziegler, M., Kemper, C. J. & Kruyen, P. (2014). Short Scales – Five Misunderstandings and Ways to Overcome Them. *Journal of Individual Differences*, *35*(4), 185–189. https://doi.org/10.1027/1614-0001/a000148

Abbildungsverzeichnis

Tabellenverzeichnis

Hilfsmittelverzeichnis mit Verwendungszweck

KI-Assistenzsystem	Teile / Stelle(n) in der Arbeit	Einsatz
ChatGPT	Zusammenfassung (deutsch)	Sprachliche Überarbeitung und Kürzung des Abstracts auf die vorgegebene Zeichenzahl
DeepL	Abstract (englisch)	Übersetzung des deutschsprachigen Abstracts ins Englische sowie grammatikalische und stilistische Überprüfung
ChatGPT	Einleitung	Unterstützung bei sprachlicher Überarbeitung (Rechtschreibung, Stil, Verständlichkeit)
ChatGPT	Methodik / Auswertung	Methodische Rückversicherung zur Wahl geeigneter statistischer Verfahren (z. B. t-Tests, Korrelation)
Claude	Ergebnisteil / Diskussion	Sprachliche Präzisierung und stilistische Überarbeitung einzelner Textpassagen
DeepL	Theorieteil	Übersetzung englischsprachiger Fachartikel ins Deutsche sowie sprachliche Verständnishilfe