



universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Der Flynn-Effekt: Domänenspezifität und
Einflüsse von Itemobsoleszenz in einer präregistrierten
querschnittlichen Untersuchung von HAWIE-R und WIE“

verfasst von / submitted by

Lea Haase, BSc

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

Wien, 2020 / Vienna 2020

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

UA 066 840

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

Masterstudium Psychologie UG2002

Betreut von / Supervisor:

Mag. Dr. Jakob Pietschnig

Inhaltsverzeichnis

Einleitung	4
Konstrukt der Intelligenz	4
Messung der Intelligenz	7
Flynn-Effekt	11
Erklärungsansätze	19
Fragestellung	24
Methode	26
Untersuchungsdesign	26
Stichprobe	27
Durchführung	28
Erhebungsinstrumente	28
Statistische Analyse	32
Ergebnisse	35
Deskriptive Statistik	35
Hypothesenprüfung	36
Inhaltliche Äquivalenz	39
Diskussion	40
Kein Einfluss obsoleter Items auf den kristallinen Flynn-Effekt	40
Indirekter Hinweis auf einen domänenspezifischen Flynn-Effekt	43
Multikausales Zusammenwirken verschiedener Erklärungsmodelle	52
Inhaltliche Äquivalenz der Untertests angenommen	54
Beobachtung von Teststrategieverhalten	55
Limitation	55
Conclusio	58
Literaturverzeichnis	60
Anhang	68

Einleitung

Menschen sind von Natur aus sehr verschieden und im Alltag werden Einschätzungen über Mitmenschen ganz selbstverständlich getroffen. Das entspricht dem grundsätzlichen Wunsch nach Versteh- und Erklärbarkeit des eigenen Umfelds. Neben der Persönlichkeit einer Person steht in der westlichen Gesellschaft vor allem die Leistung - und damit untrennbar verbunden auch die Intelligenz - im Vordergrund des allgemeinen Interesses. Denn die Intelligenz steht in engem Zusammenhang mit vielfältigen praktischen Leistungen einer Person. Dazu zählen Aspekte wie der Bildungserfolg, die weitere berufliche Karriere oder das Ausmaß an Sozialkompetenz (Gottfredson, 1997). Bei Abweichungen vom normalen Entwicklungsverlauf werden bereits Kleinkinder diagnostisch untersucht. Dabei wird die Intelligenz routinemäßig erfasst und bildet oft eine Basis anhand derer spezifisches Verhalten eingeschätzt und das weitere Vorgehen abgewogen wird (Daseking et al., 2007).

Zuletzt verdeutlicht auch die Medienpräsenz und der politische Einfluss internationaler Studienergebnisse zum Intelligenzvergleich wie beispielsweise die der PISA-Studien den Stellenwert in der Gesellschaft (Pietschnig et al., 2018). Auf die Intelligenz eines Menschen wird im nichtwissenschaftlichen Kontext häufig intuitiv über das Auftreten, Verhalten und den gesellschaftlichen Status geschlossen. Aus wissenschaftlicher Perspektive hingegen ist eine einheitliche Definition der Intelligenz nicht so einfach. Seit Beginn der Intelligenzforschung wird versucht die Fülle individueller kognitiver Leistungsaspekte zu klassifizieren und zu sortieren (Neisser et al., 1996). Daraus hat sich nach wie vor noch keine allgemein akzeptierte Formulierung ergeben. Wohl aber hat sich inzwischen ein inhaltlicher Konsens über das Konstrukt heraus kristallisiert (Gottfredson, 1997).

Konstrukt der Intelligenz

Den meisten Definitionsversuchen ist die Annahme gleich, dass Intelligenz nicht als eine einzelne, isolierte Fähigkeit angesehen werden kann, wie zum Beispiel Buchstabieren, Rätsel lösen oder Fahrradfahren. Es wird stattdessen davon ausgegangen, dass die Intelligenz die latente kognitive Grundlage darstellt, die

einzelne Fähigkeiten ermöglicht. Demzufolge kann sie durch viele differenzierte Verhaltensweisen erkennbar werden (Flynn, 2007; Gottfredson, 1997; Von Aster et al., 2006). Dazu zählen unter anderem Fähigkeiten wie die Anpassungsfähigkeit einer Person an ihre Umgebung, das Verstehen komplexer Zusammenhänge und das Lösen abstrakter Problemstellungen. Auch das logische Schlussfolgern gehört dazu. Die individuelle Ausprägung fällt dabei unterschiedlich aus und kann sich sowohl in den verschiedenen Aspekten der Intelligenz, als auch abhängig von Situation und Tagesform unterscheiden (Neisser et al., 1996).

Der Entwickler der bekannten Wechsler Intelligenztests, David Wechsler, versteht die menschliche Intelligenz ebenfalls als ein komplexes Konstrukt. Dies drückt sich im Alltag in unterschiedlicher Weise aus und charakterisiert das Verhalten einer Person. Er beschreibt Intelligenz als die Fähigkeit „zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umgebung wirkungsvoll auseinander zu setzen“ (Wechsler, 1956, S.13, zitiert nach Aster et al., 2006, S.8). Die Intelligenz sollte ihm zufolge, über die Abfrage verschiedener spezifischer Fähigkeiten erfolgen. Erst durch die Gesamtheit der Aufgaben sei ein Rückschluss auf die allgemeine Ausprägung zulässig.

Bereits in den 1930er Jahren wurde beobachtet, dass Menschen, die in einer speziellen kognitiven Aufgabe gut abschneiden, auch in anderen Aufgaben zu guten Leistungen tendieren. Daraus wurde die Annahme abgeleitet, dass in diesem Phänomen die bereits erwähnte Basiskomponente der Intelligenz sichtbar wird. Diese Entdeckung brachte daraufhin die nötige empirische Fundierung. Inzwischen wird von einer sogenannten *generellen Intelligenzkomponente* ausgegangen, auch *g-Faktor* oder *psychometrisches g* genannt, die der individuellen kognitiven Fähigkeitsansammlung zu gewissen Anteilen zugrunde liegt (Pietschnig & Voracek, 2015; Spearman, 1927). Mit Hilfe eines faktorenanalytischen Verfahrens wurden dem g-Faktor Mitte des 20. Jahrhunderts acht Sekundärfaktoren der Intelligenz untergeordnet (Horn, 1965). Unter Zusammenschluss mehrerer Theorien bildete sich daraus schließlich die Cattell-Horn-Carroll Theorie (CHC-Modell), welche den derzeitigen Konsens darstellt. In diesem hierarchischen Intelligenzmodell steht das psychometrische g auf oberster Ebene (Stratum I). Diesem untergeordnet (auf Stratum II)

befinden sich, je nach Auslegung, zwischen sieben und zehn umfassende kognitive Fähigkeiten. Dazu gehört unter anderem das Kurzzeit- und Langzeitgedächtnis, die Verarbeitungsgeschwindigkeit, die visuelle Verarbeitung sowie die fluide und kristalline Intelligenz. Auf dritter Stufe (Stratum III) werden noch bis zu 70 weitere eingegrenzte Fähigkeiten erfasst (McGrew, 2009).

Fluide und kristalline Intelligenz

Neben der Untersuchung des g-Faktors erhalten einzelne Intelligenzaspekte wie die fluide und kristalline Intelligenz auch immer mehr Aufmerksamkeit in der aktuellen Forschung (u.a. Höltl, 2019; Lynn, 2009a; Pietschnig & Voracek, 2015; Williams, 2013).

Beide umfassen Fähigkeiten wie das Erschließen von Zusammenhängen, das Erstellen von Verknüpfungen, die Abstraktionsfähigkeit sowie das logische Schlussfolgern und Aufrechterhalten der Aufmerksamkeit. Während die kristalline Intelligenz jedoch durch das Sammeln von Erfahrungen zunehmend ansteigt, gilt die fluide Intelligenz als kultur- und lernunabhängig. Fluide Intelligenz kann durch abstrakte Aufgaben gemessen werden, die universelle, symbolische, figürliche oder auch semantische Inhalte umfassen. Bei Aufgaben zur Erfassung der kristallinen Komponente fließen stets Aspekte mit ein, die Schulwissen, Erfahrungen aus der Vergangenheit oder das Anwenden erlernter Lösungsstrategien erfordern (Horn, 1965). Denn im Gegensatz zur fluiden Intelligenz können Aufgaben zur kristallinen Intelligenz nur durch bestimmtes Vorwissen gelöst werden. Logisches Denken allein reicht dazu nicht aus. Die meisten kristallinen Aufgaben haben zudem einen verbalen Schwerpunkt (Pietschnig & Voracek, 2015) und hängen stärker mit dem bereits beschriebenen g-Faktor zusammen (Johnson et al., 2004). Während Fähigkeiten des kristallinen Intelligenzspektrums über die Lebensspanne hinweg weiter ansteigen, sinkt die fluide Intelligenz mit der Zeit wieder ab (Dickinson & Hiscock, 2010; Schaie, 1994).

Mit der Definition und Differenzierung des Intelligenzbegriffs geht zeitgleich auch die Forschung nach der bestmöglichen Methode zur Erfassung des Konstrukts einher.

Messung der Intelligenz

Die Entwicklung von Testformaten zur Messung der Intelligenz reicht weit zurück und ist hochentwickelt. Im Gegensatz zur Begriffsdefinition besteht unter Forschenden kaum Zweifel darüber, ob Intelligenztests das gewünschte Konstrukt auch tatsächlich erfassen. Sie gehören zu den präzisesten Testinstrumenten der empirischen Psychologie (Gottfredson, 1997). Zwei Testarten gelten als Vorreiter der heutigen Intelligenztestung: Die in den USA entwickelten Stanford-Binet- und Wechsler Intelligenztests (Flynn, 1984).

Generell gibt es unterschiedliche Herangehensweisen zur Erfassung der Intelligenz, je nach Zweck und Anwendungsbereich. Zum einen gibt es Tests, die nur einen spezifischen Aspekt der Intelligenz durch eine bestimmte Art von Aufgabe abfragen. Zum anderen gibt es sogenannte Intelligenztestbatterien, die den Anspruch haben, durch die umfangreiche Abfrage von Einzelleistungen eine generellere Intelligenzeinschätzung liefern zu können. Sie zeichnen sich daher durch viele verschiedene Aufgaben aus, die auf unterschiedliche Weise einzelne Aspekte der Intelligenz abfragen. Durch die Auswertung können sowohl Teilleistungen als auch eine Gesamtbeurteilung rückgemeldet werden. Sie sind meist ein guter Indikator für das psychometrische g . Zu dieser Form gehört auch das Wechsler Intelligenztestformat (Neisser et al., 1996; Tewes, 1994; Von Aster et al., 2006; Weiss et al., 2016).

Wechsler Intelligenztest

Dieser Test gilt global betrachtet als einer der populärsten Testbatterien zur Erfassung der Intelligenz (Boake, 2002; Petermann & Köller, 2008). Er steht übersetzt in verschiedenen Sprachen und in unterschiedlichen Formaten speziell für Kinder und Erwachsene zur Verfügung. Auf Grund der kontinuierlichen Anwendung und Popularität des Wechsler Intelligenztests kann auf die hohe Akzeptanz des Messinstruments in der wissenschaftlichen Gesellschaft geschlossen werden (Molz et al., 2010).

Versionen und Aufbau. Der Originaltest für Erwachsene entstand bereits 1939 unter dem Namen *Wechsler-Bellevue Intelligence Scale* in den USA (Boake, 2002). Seither gab es mehrfache Revisionen aus deren englischen Ausgaben stets auch deutschsprachige Neuauflagen abgeleitet wurden. Die einzelnen Versionen unterscheiden sich jeweils durch die Adaption der kulturabhängigen und aktualisierungsbedürftigen Inhalte, der generelle Skalenaufbau ist hingegen gleich (Tewes, 1994).

Erst 35 Jahre nach der ersten deutschsprachigen Fassung, dem HAWIE (engl. WAIS), wurde der *Hamburg-Wechsler Intelligenztests für Erwachsene - Revision* (HAWIE-R) veröffentlicht. Dessen Grundlage stellt die amerikanische Revision WAIS-R dar (Tewes, 1994). Schon weitere 15 Jahre später folgte der gründlich überarbeitete *Wechsler Intelligenztest für Erwachsene* (WIE) als übersetzte Adaption des WAIS-III (Von Aster et al., 2006). Die Versionen HAWIE-R und WIE wurden für die vorliegende Untersuchung verwendet und werden daher noch mehrfach thematisiert. Zum aktuellen Zeitpunkt stellt die Nachfolgeversion des WIE, der *Wechsler Adult Intelligence Scale - Fourth edition* (WAIS-IV), die aktuellste deutschsprachige Version dar (Petermann & Wechsler, 2012).

Die einzelnen Untertests der Wechsler Intelligenztests werden entweder dem *Verbal-* oder dem *Handlungsteil* zugeordnet. Zusätzlich besteht auch eine Einteilung in vier verschiedene Skalen (Intelligenzaspekte): *Sprachliches Verständnis*, *Wahrnehmungsorganisation*, *Arbeitsgedächtnis* und *Arbeitsgeschwindigkeit*. Diese Unterteilung ermöglicht neben der Beurteilung der Gesamtleistung auch differenziertere Aussagen (Boake, 2002; Von Aster et al., 2006). Die einzelnen Untertests weisen hohe Interkorrelationen auf (Flynn, 2007). Damit entsprechen sie den Annahmen des zuvor beschriebenen hierarchischen Intelligenzkonzepts mit zugrundeliegendem g-Faktor (McGrew, 2009; Spearman, 1927).

Intelligenzkonstrukte der relevanten Untertests. Laut Testmanual gehören die für die aktuelle Untersuchung ausgewählten drei Untertests alle dem Verbalteil an (Tewes, 1994; Von Aster et al., 2006). Die Gesamtleistung aus allen Aufgaben dieses Testteils wird in großen Studien zur Beurteilung der kristallinen Intelligenz genutzt (fluide Intelligenz durch den Handlungsteil) (Flynn, 1987; Pietschnig &

Voracek, 2015). Bei der differenzierten Betrachtung der einzelnen Untertestinhalte weicht die Literatur teilweise von der im Manual getroffenen Zuordnung der Untertests zu einzelnen Intelligenzdomänen ab.

Dies trifft nicht auf den ausgewählten Untertest *Allgemeines Wissen* (AW) zu. Wechsler (1939, zitiert nach Tewes, 1994) ordnete die Leistung in diesem Untertest als stark kultur- und bildungsabhängig ein. Im Manual wird er der Skala Sprachliches Verständnis zugewiesen (Tewes, 1994; Von Aster et al., 2006). Auch explorative faktorenanalytische Überprüfungen bestätigen die Ladung auf diesen Faktor (Niileksela & Reynolds, 2019; Weiss et al., 2013). Da sich die Inhalte ziemlich genau mit der gängigen Auffassung zur kristallinen Intelligenz decken (Horn, 1965; Pietschnig & Voracek, 2015), besteht kaum Zweifel über die erfassten Intelligenzanteile dieser Aufgabe.

Der zweite Untertest, *Rechnerisches Denken* (RD), wird in der Testbeschreibung zur Skala des Arbeitsgedächtnisses geordnet (Tewes, 1994; Von Aster et al., 2006). Wechsler (1939, zitiert nach Tewes, 1994) zufolge wird hiermit die Flexibilität des Geistes abgefragt. In diesen Beschreibungen wird demnach die fluide Tendenz der Aufgabe betont. Zugleich wird aber auch angemerkt, dass noch kein Konsens darüber besteht, ob die Aufgabe mehr Konzentration, Schulbildung oder Sprachverständnis erfasst (Tewes, 1994). In externer Literatur wird die Verbal-Lastigkeit der Aufgabe kritisiert. Die abstrakte Fähigkeit zur Problemlösung werde demnach nicht gut erfasst (Flynn, 1987). Explorative Faktorenanalysen (EFAs) finden neben der Ladung auf den Faktor des Arbeitsgedächtnisses vergleichbar hohe Ladungen auf Sprachliches Verständnis (kristalline Intelligenz) und auf einen fünften (neuen) Faktor. Dieser wird als Fluides Schlussfolgern benannt (Niileksela & Reynolds, 2019). Die Mehrdimensionalität der Aufgabe wird in einer anderen EFA im klassischen Vierfaktorenmodell bestätigt. Unter Hinzugabe des Skalenfaktors Fluides Schlussfolgern wird die höchste Ladung darauf erzielt. Zusätzlich wird eine moderate Ladung auf das Arbeitsgedächtnis beschrieben. Dort wird resümiert, dass die Aufgabe primär fluides Schlussfolgern erfasst und sekundär das Arbeits-, das Kurzzeitgedächtnis, das sprachliche Verständnis und den g-Faktor (Weiss et al., 2013). Damit

kann RD als einziger Untertest aus der Wechsler-Intelligenztestbatterie nicht eindeutig eindimensional zugeordnet werden.

Die Aufgabe *Gemeinsamkeiten Finden* (GF) wird (wie auch AW) der Skala Sprachliches Verständnis zugeordnet (Tewes, 1994; Von Aster et al., 2006). Wechsler (1939, zitiert nach Tewes, 1994) betont jedoch die Besonderheit dieses Untertests. Neben der standardisierten Auswertung kann dieser auch qualitative Rückschlüsse über das logische Denkmuster und die Fähigkeit zur Abstraktion liefern. Auf Grund der abstrakten, imaginativen Anforderung des Untertests wird in verschiedenen Studien dazu tendiert, ihn der fluiden Intelligenz zuzuordnen (Flynn, 1999; Satzger et al., 1996). Auf Grund der sowohl kristallinen als auch fluiden Komponenten wird er auch als Maß für abstraktes verbales Denken beschrieben (Weiss et al., 2016). In den bereits beschriebenen EFAs wurde jedoch die kristalline Manualzuordnung (eindimensionale Ladung auf Sprachliches Verständnis) mit hoher Ladung eindeutig bestätigt (Niileksela & Reynolds, 2019; Weiss et al., 2013).

Intelligenzquotient (IQ)

Die erfasste Intelligenz wird nach aktuellem Standard durch den IQ angegeben. Dieser dient dazu, die erhobene Intelligenz über verschiedene Tests hinweg und im Bezug zur Population vergleichbar zu machen, indem das Testergebnis in den einheitlichen IQ-Wert transformiert wird. Dazu wird festgelegt, dass die durchschnittliche IQ-Leistung einen Wert von 100 beträgt und die Standardabweichung 15. Diese standardisierte Skala gibt Auskunft über die Leistung im Verhältnis zur eigenen Kohorte (Neisser et al., 1996).

Der IQ ist explizit vom bereits definierten g-Faktor zu unterscheiden. Während g auf der Korrelation zwischen kognitiven Aufgaben beruht, wird der IQ durch die Summierung der Leistung in diesen Aufgaben und durch den Normierungsvergleich bestimmt. So wird zum Beispiel durch die Bestimmung des Gesamt-IQs in einem Wechsler Intelligenztests nicht isoliert g erfasst, sondern eine Kombination aus spezifischen kognitiven Fähigkeiten sowie einem Anteil an g (Colom et al., 2002). Wie hoch dieser ist unterscheidet sich je nach Aufgabe (Johnson et al., 2004).

Jeder neue Test wird durch die Erhebung einer umfangreichen und populationsrepräsentativen Stichprobe standardisiert. Dabei wird festgelegt, welcher Rohwert des Tests welchem IQ-Wert zugeordnet werden muss, sodass der Mittelwert der Rohwertscores der Normierungsstichprobe bei 100 IQ-Punkten liegt. Der IQ einer Testperson oder einer Stichprobe spiegelt dann die Leistung im Vergleich zur Normierungsstichprobe wider (Pietschnig & Voracek, 2015). Es zeigte sich jedoch relativ schnell, dass die ursprünglich getroffene Zuordnung von Testwert zu IQ über die Jahre hinweg stets an Gültigkeit verlor. Die Nutzung von veralteten Normtabellen zur Einordnung des IQ führen daher tendenziell zur Überschätzung der Fähigkeiten (Satzger et al., 1996). Um die Veränderung der durchschnittlichen Intelligenzleistung wieder auszugleichen, werden Intelligenztests daher regelmäßig an einer neuen Normierungsstichprobe restandardisiert, sodass der Mittelwert wieder bei 100 IQ-Punkten liegt. Ohne Aktualisierung würde eine Person, die mit einer veralteten IQ-Skala eingeschätzt wird, einen höheren Wert erzielen als mit einer neu normierten Skala (Neisser et al., 1996; Rodgers, 1999). Genau dieses Phänomen wirkt schon seit knapp einem Jahrhundert wichtige Fragen im Forschungsgebiet der differentiellen Psychologie auf und ist inzwischen als *Flynn-Effekt* bekannt.

Flynn-Effekt

Unter diesem Effekt versteht man die steigende Leistung in Intelligenztests der allgemeinen Bevölkerung über Generationen und Zeit hinweg (Flynn, 2007; Pietschnig et al., 2018; Pietschnig & Voracek, 2015). Die ursprünglich definierten Normierungen eines Testinstruments werden dadurch obsolet, haben also keine gültige Aussagekraft mehr (Trahan et al., 2014).

Schon seit es psychometrische Tests zur Erfassung der kognitiven Fähigkeiten gibt wurden steigende IQ-Leistungen über verschiedene Zeiträume beobachtet. Bereits in den 1930er Jahren fielen Leistungsunterschiede zwischen originalen und revidierten IQ-Skalen bei erneuter Testung derselben Kinder mit zwei Stanford-Binet-Testversionen auf. Die Unterschiede wurden zu diesem Zeitpunkt aber noch als Folge ungeeigneter Stichprobenauswahl interpretiert und als Artefakte abgehandelt (Merrill, 1938). Erst in den 1970er Jahren wurden Leistungsunterschiede in Intelligenzaufgaben als Effekt zwischen unterschiedlichen Kohorten zugeordnet (Schaie

& Strother, 1968). Erstmals systematisch analysiert und beschrieben wurde das Phänomen der IQ-Testnormveränderungen über Generationen hinweg von James R. Flynn. Er verglich die durchschnittliche Leistung der Normierungss Stichproben unterschiedlicher Wechsler und Stanford-Binet Intelligenztestversionen. Dabei konnte er zeigen, dass die mittlere IQ-Leistung (Median) über 46 Jahre zunehmend weiter angestiegen war. Pro Jahr um 0.33 IQ-Punkte bzw. 3.3 IQ-Punkte pro Jahrzehnt (Flynn, 1984).

Seitdem rückte das Phänomen mehr und mehr in den Fokus der Wissenschaft. Ende der 1990er Jahre wurde das Vorkommen kontinuierlich steigender IQ-Leistungen von der *American Psychological Association* als eines von sieben zu priorisierenden Forschungsangelegenheiten der differentiellen Psychologie kategorisiert. Der Ursprung und die Bedeutung dieser Anstiege sollten im Mittelpunkt weiterer Untersuchungen stehen (Neisser et al., 1996). Inzwischen wurden IQ-Zuwachsraten international intensiv erforscht. Eine erste umfangreich vergleichende Metaanalyse zum Flynn-Effekt bietet sogar einen weltweiten Überblick aus 271 Einzelstudien dazu. Hierbei wurden IQ-Leistungen über einen Zeitraum von über 100 Jahren (1909 bis 2013) betrachtet. Miteinbezogen wurden Daten aus 31 Ländern von sechs verschiedenen Kontinenten. Die Ergebnisse zeichnen ein klares Bild von einer kontinuierlichen, nicht linearen, globalen Zuwachsrate. Diese umfasst für den Gesamt-IQ 0.28 IQ-Punkte pro Jahr, also 2.8 IQ-Punkte pro Jahrzehnt (Pietschnig & Voracek, 2015). Eine andere Metaanalyse von 2014 inkludierte 285 amerikanische Studien ab dem Jahr 1951 und geht von einem etwas geringeren, durchschnittlichen positiven Flynn-Effekt von 2.3 IQ-Punkten pro Dekade aus. Unter isolierter Betrachtung der 53 Studien, die Stanford-Binet und Wechsler Intelligenztests (ab 1972) angewendet haben, steigt der mittlere IQ-Zuwachs auf 2.9 IQ-Punkte an (Trahan et al., 2014). Damit ist der Wert vergleichbar mit den Ergebnissen der zuvor beschriebenen Studien (Flynn, 1984; Pietschnig & Voracek, 2015).

Auf Grund der beobachteten Testnormverschiebungen und der damit verbundenen ungültigen IQ-Einschätzungen wurde in der Praxis zwischenzeitlich zu einer Korrektur geraten. Ergebnisse aus einem Test mit veralteten Normtabellen sollten

um jeweils 0.3 IQ-Punkte pro Jahr nach oben korrigiert werden (Flynn, 2009). Eine Adaption der Ergebnisse um einen solchen fixen IQ-Wert kann die Verzerrung durch den Flynn-Effekt jedoch nicht ausgleichen. Aktueller Literatur zufolge kann nicht von einem linearen IQ-Zuwachs (bzw. IQ-Veränderung) über die Zeit hinweg ausgegangen werden (Pietschnig & Voracek, 2015; Truscott & Frank, 2001). Ein einfaches Vorgehen wie dieses würde die Obsoleszenz der Normtabellen daher nicht beheben, sondern ebenfalls zu Verzerrung führen. Bisher stellt nur die regelmäßige Renormierung der Testinstrumente an einer repräsentativen Stichprobe ein sinnvolles Mittel dar, um die Gültigkeit gemessener IQ-Werte zu gewährleisten.

Die Interpretation, dass die beobachteten IQ-Veränderungen über die Zeit einem tatsächlichen Intelligenzanstieg in der Bevölkerung entsprechen, ist kritischen Stimmen zu voreilig. Denn dies sei theoretisch nicht die einzig zulässige Schlussfolgerung (Rodgers, 1999). Um die Hypothese zu fundieren, wird immer wieder nach Anzeichen für steigende kognitive Fähigkeiten in der Bevölkerung gesucht. So wurde beispielsweise beobachtet, dass die zugelassenen Kandidaten professioneller Schachtunier zunehmend jünger werden (Howard, 2005). Auch die Befragung von Lehrpersonal in Singapur ergab Hinweise für zunehmend intelligentere Kinder (Cocodia et al., 2003). Solche Ergebnisse geben Evidenz für die tatsächliche Existenz des Flynn-Effekts und unterstreichen dadurch die Validität von Intelligenztests (Howard, 2005). Die Forschung ist sich daher überwiegend einig, dass genug empirische Fundierung für die Interpretation des Effekts als kognitiven Leistungszuwachs vorliegt. Das Phänomen kann jedoch auch anfällig für Einflüsse der Testung selbst sein. Die verschiedenen Herangehensweisen zur Untersuchung des Flynn-Effekts unterliegen dabei weitgehend den gleichen Qualitätsansprüchen.

Messung des Flynn-Effekts

Einige Studien zum Flynn-Effekt vergleichen die IQ-Leistung derselben Testversion anhand von mehreren äquivalenten Stichproben zu verschiedenen Messzeitpunkten längsschnittlich. Die Schwierigkeit liegt dabei darin, eine tatsächlich vergleichbare Stichprobe zu finden, bei der die IQ-Differenz eindeutig als kognitiver Leistungsanstieg interpretiert werden kann und nicht auf andere, externe Faktoren zurückgeführt werden muss. Auf Grund ihrer guten Vergleichbarkeit wurden daher

gerne die jährlich erhobenen Datensätze aus den Eignungsprozessen der Bundesheere eines Landes genutzt wie zum Beispiel in Studien aus Finnland (Dutton & Lynn, 2013) oder Dänemark (Teasdale & Owen, 2005). Allerdings bestehen diese Datensätze nur aus männlichen Probanden. Immer öfter werden die Ergebnisse vieler Primärstudien auch metaanalytisch aufbereitet (u.a. Pietschnig & Gittler, 2017; Pietschnig & Voracek, 2015; Trahan et al., 2014; Wongupparaj et al., 2017). Diese Studien geben einen guten Überblick und machen Einzelergebnisse dadurch besser einschätzbar.

Querschnittliche Untersuchung. Die Normierungsstichproben können aber auch indirekt, anhand einer einzigen Stichprobe, querschnittlich verglichen werden (u.a. Erzberger & Engel, 2010; Hölzl, 2019; Satzger et al., 1996). Bei dieser gängigen Methode werden zwei inhaltlich äquivalente Intelligenztests mit Normierungen aus unterschiedlichen Jahren an denselben Personen angewendet. Die mittlere IQ-Differenz misst dann die Leistungsveränderung innerhalb der Zeitspanne, die zwischen den beiden Normierungen der jeweiligen Testversionen liegt. Ein IQ-Anstieg wird sichtbar, wenn die durchschnittliche Leistung derselben Stichprobe, bewertet nach den älteren Testnormen, besser ausfällt als nach den neueren Normierungen. Die Differenz zwischen den Mittelwerten gibt dann das Ausmaß des Zugewinns an. Die Methode beruht allerdings auf der Annahme, dass die Normierungsstichprobe zum Zeitpunkt ihrer Erhebung repräsentativ für die jeweilige Population war. Ebenso wird die inhaltliche Äquivalenz der beiden Testformate angenommen. Die Reliabilität einer Untersuchung hängt hierbei sowie generell stets von der Messgenauigkeit bei der Erhebung der Leistung mit dem jeweiligen Test als auch von der Testqualität ab (Flynn, 1984; Satzger et al., 1996).

Inhaltliche Äquivalenz der Wechsler Intelligenztestversionen. Beim Vergleich der IQ-Leistung anhand originaler und revidierter Testversionen sollten die Testinhalte zuvor geprüft werden, um Verzerrung zu vermeiden. Im Zuge der Überarbeitung könnten einzelne Fragen sowie ganze Skalen verändert oder ersetzt worden sein. Das kann wiederum Auswirkungen auf die inhaltliche Äquivalenz der Tests haben, also ob nach wie vor dasselbe Konstrukt gemessen wird. Es ist große Sorgfalt dabei gefragt, dies zu prüfen und nur gleichwertige Tests in

Vergleichsstudien einzubeziehen (Pietschnig & Voracek, 2015). Bei zu großen Unterschieden zwischen den Tests, könnten die interessierenden Generationseffekte von Effekten geänderter Testformate maskiert werden. Wenn ein Test aktualisiert wird, werden die originalen und neu revidierten Testversionen querschnittlich verglichen, um die zu erwartende Leistungsdifferenz zu bestimmen. Die Ergebnisse dieser Vergleiche laufen zum Teil in die Berechnungen anderer Flynn-Effekt Studien (Metaanalysen) mit ein. Die Schlüsse aus der Analyse dieser Daten könnten jedoch problematisch sein, wenn sich das Verhältnis der erfassten Intelligenzdomänen zwischen den Testversionen verändert hat (Weiss et al., 2016). Denn der Flynn-Effekt verläuft nicht für jeden Intelligenzaspekt gleich (Pietschnig & Voracek, 2015). Angenommen es fließen beispielsweise mehr fluide und weniger kristalline Intelligenzanteile durch die revidierten Untertests in den Gesamt-IQ mit ein als in der Originalversion. Dann könnten Leistungsdifferenzen, die dem Flynn-Effekt zugeschrieben werden, eigentlich durch die Änderung der gemessenen Konstrukte oder der Testanforderungen zustande gekommen sein. Diese Kritik gibt es explizit auch für Wechsler Intelligenztestversionen (Weiss et al., 2016).

Aktuellen faktorenanalytischen Untersuchungen zufolge ist die Konstruktvalidität zwischen amerikanischen Wechsler Testversionen aber durchaus gegeben. Der Vergleich der Konstrukte und das Generalisieren von Forschungsergebnissen über verschiedene Versionen hinweg wird als unbedenklich eingeordnet. Trotz Differenzen zwischen Untertests zur Skala Sprachliches Verständnis wird auch der Verbal-IQ als vergleichbar eingestuft (Niileksela & Reynolds, 2019). Auch aus Vergleichsstudien deutschsprachiger Wechsler Intelligenztestversionen gehen ähnliche Funde hervor. Der Vergleich des HAWIE mit dem revidierten HAWIE-R ergibt ebenfalls Evidenz für die inhaltliche Äquivalenz der beiden. Dies bezieht sich auch hier insbesondere auf den Gesamt-IQ mit einem hohen Korrelationskoeffizienten von $r = .97$. Nach Durchführung der Gesamttests kann daher davon ausgegangen werden, dass das gleiche Fähigkeitskonstrukt abgeprüft wurde (Satzger et al., 1996). Die Untersuchung zur Äquivalenz zwischen HAWIE-R und WIE zieht einen ähnlichen Schluss. Ein hoher Korrelationskoeffizient von $r = .91$ für den Gesamt-IQ gibt ebenfalls ausreichend Hinweis auf eine hohe inhaltliche Äquivalenz der beiden

Gesamttestungen. Diese wird der Studie nach auch für den Handlungs- und Verbal-IQ, sowie für die meisten Untertests bestätigt (Erzberger & Engel, 2010).

Spezifika des Flynn-Effekts

Die globale Analyse der IQ-Testnormverschiebungen über ein Jahrhundert hinweg lässt, wie bereits erwähnt, ein ungleichmäßiges Muster erkennen. Der IQ scheint über Generationen hinweg nicht linear anzusteigen. Schwankungen im Verlauf lassen sich zum Teil mit einschneidenden Geschichtsentwicklungen in Verbindung bringen. So entwickelte sich der Gesamt-IQ zwischen den beiden Weltkriegen besonders stark. Mit Beginn des 2. Weltkrieges flachte dieser Trend deutlich ab. Erst ab den 1950er Jahren kam es wieder zu einer stärkeren Zuwachsperiode, die bis in die 1970er Jahre anhielt. Seither wird wieder eine leichte Abschwächung des Anstiegs verzeichnet (Pietschnig & Voracek, 2015).

Nicht nur über die Zeit hinweg werden Unterschiede beobachtet. Auch zwischen einzelnen Ländern fallen die Zuwachsraten zum Teil sehr unterschiedlich aus (Pietschnig et al., 2010). Berichte über eine positive IQ-Entwicklung kommen beispielsweise aus der USA (Flynn, 1987), Österreich (Pietschnig et al., 2011), Tschechien (Laciga & Cígler, 2017), Spanien (Colom et al., 2005), Brasilien (Colom et al., 2007) und Kenia (Daley et al., 2003). Im letzten Jahrzehnt wird teilweise aber auch von Stagnation und Umkehrung des beobachteten IQ-Zuwachses berichtet. So melden Dänemark (Teasdale & Owen, 2005), Finnland (Dutton & Lynn, 2013), Norwegen, Niederlande, Estland (Dutton et al., 2016), Frankreich (Dutton & Lynn, 2015) und Kuwait (Dutton et al., 2017) Ergebnisse über einen negativen Flynn-Effekt. Noch immer überwiegen aber die Funde über positive Intelligenzentwicklungen aus vielen Ländern. Negative Funde machen bisher eine deutlich kleinere Menge aus (Dutton et al., 2016). Die globale Betrachtung weist insgesamt aber auf einen kontinuierlich schwächer werdenden Anstieg der IQ-Werte hin (Pietschnig & Voracek, 2015).

Differenzen in der Stärke des Flynn-Effekts lassen sich außerdem über die verschiedenen Fähigkeitslevel und Altersstufen erkennen (Platt et al., 2019). Die Betrachtung des gesamten Intelligenzspektrums zeigt stärkere Leistungsanstiege im

schwächeren bis mittleren IQ-Bereich (Colom et al., 2005). Die Analyse der Altersunterschiede lässt auf tendenziell stärkere Effekte bei Erwachsenen als bei Kindern schließen (Pietschnig & Voracek, 2015). Für Geschlechtsunterschiede gibt es hingegen kaum Evidenz (Pietschnig et al., 2011).

Forschungen zum Zusammenhang der IQ-Anstiege mit dem g-Faktor liefern Evidenz für eine negative Assoziation zwischen der Größe des Flynn-Effekts und der Ladung einer Intelligenzaufgabe auf den g-Faktor (Must et al., 2003; Te Nijenhuis & van der Flier, 2013). Je weniger eine Testaufgabe den g-Faktor erfasst, desto größer fällt demnach die gemessene Leistungszunahme aus. Das lässt die Schlussfolgerung zu, dass der Flynn-Effekt eher mit einer zunehmenden Differenzierung der Fähigkeiten als mit dem g-Faktor zusammenhängt (Woodley & Madison, 2013). Nicht die generelle Intelligenz scheint zu steigen, sondern viel mehr die spezifischen Anteile. Zu Aspekten, die besonders stark steigen, gehört unter anderem die fluide Intelligenz (u.a. Flynn, 1987) sowie die Fähigkeit zum abstrahierenden Denken (Flynn, 2007).

Der Fokus der Forschung zum Flynn-Effekt lag bisher oft auf der Analyse des Gesamt-IQ Verlaufs. Die spezifische Untersuchung auf Ebene unterschiedlicher Intelligenzdomänen gewinnt jedoch zunehmend an Bedeutung. Dabei ist der Befund, dass Stärke und Richtung des Effekts zwischen unterschiedlichen Intelligenzdomänen variiert, keine Neuheit (Flynn, 1987; Pietschnig & Gittler, 2015; Pietschnig & Voracek, 2015).

Domänenspezifischer Flynn-Effekt

Schon Anfang des 20. Jahrhunderts fielen bei der Analyse von Daten aus 14 Ländern dementsprechende Differenzen auf. Das Ausmaß des IQ-Anstiegs aus den Intelligenzserfassungen zwischen 1932 und 1984 (v.a. nach 1950) differiert in Abhängigkeit von der Art der Aufgabe. Der Median zeigt für kulturunabhängige, abstrakte Problemlösungsaufgaben (fluide Intelligenz) einen Anstieg um 5.8 IQ-Punkte pro Jahrzehnt an. Bei wissensbasierten Aufgaben (kristalline Intelligenz) wird hingegen ein deutlich geringerer Zuwachs von 3.7 IQ-Punkten pro Dekade gemessen (Flynn, 1987). Auch der internationale Datenvergleich über 100 Jahre hinweg (1909 - 2013)

zeichnet ein ähnliches Muster. Die stärksten Zuwachsraten wurden bei fluiden Intelligenzaufgaben mit 4.1 IQ-Punkten pro Jahrzehnt gefunden. Bei kristallinen Aufgaben belief sich die Steigerung hingegen nur auf 2.1 IQ-Punkte (Pietschnig & Voracek, 2015). Global betrachtet nahm die kristalline Intelligenz demzufolge über drei Dekaden hinweg weiter leicht zu, jedoch nur halb so stark wie der fluide IQ.

Einzelne Studienergebnisse aus verschiedenen Ländern verdeutlichen die bereits thematisierten Unterschiede in der Ausprägung des Flynn-Effekts auch auf domänenspezifischer Ebene. Beispielsweise konnte eine Überblicksstudie aus dem angloamerikanischen Raum kaum Evidenz für einen kristallinen IQ-Anstieg finden. Dieser blieb zwischen 1979 und 2008 vorwiegend stabil, während der fluide IQ weiter anstieg (Lynn, 2009a). Ergebnisse einer französischen Studie verzeichneten, neben einem generellen Abwärtstrend, den stärksten Rückgang in der verbalen Intelligenzkomponente um 4 IQ-Punkte in zehn Jahren (Dutton & Lynn, 2015). Auch bei Untersuchungen in Estland fiel der Intelligenzzuwachs bei verbalen (eher kristallinen) Aufgaben nur halb so groß aus, wie bei abstrakten (eher fluiden) Aufgaben (Must et al., 2003). Es wurde bereits geschlossen, dass der Effekt bei erlernten (kristallinen) Inhalten womöglich generell schwächer ausfällt (Flynn, 1987). Eine Analyse deutschsprachiger Studien spricht jedoch wiederum für vergleichbar starke Effekte in fluiden und kristallinen Intelligenzmessungen und macht die zuvor getroffene Annahme damit nicht verallgemeinerbar. Generell hat die Untersuchung des Flynn-Effekts auf Ebene der kristallinen Intelligenzmessung bisher deutlich weniger Aufmerksamkeit erlangt als die Untersuchung der fluiden Intelligenzanstiege (Pietschnig et al., 2010).

Auch bei der Analyse von Wechsler Intelligenztestversionen gibt es Evidenz für die Domänenspezifität des Flynn-Effekts. So basiert der Gesamt-IQ Zuwachs stets auf unterschiedlich starken Zuwachsraten auf Ebene der einzelnen Untertests (u.a. Erzberger & Engel, 2010; Flynn, 2007; Khaleefa et al., 2009; Satzger et al., 1996; Truscott & Frank, 2001). So berichtet Wechsler (1991, zitiert nach Daseking et al., 2007) bereits von tendenziell niedrigeren Anstiegen unter anderem für die bereits dargestellten Untertests AW und RD. Im Unterschied dazu verzeichnen viele

Studien besonders starke Leistungsschwankungen für GF (Erzberger & Engel, 2010; Flynn, 2007; Khaleefa et al., 2009; Satzger et al., 1996).

Die differenzierte Untersuchung der dargestellten Spezifika des Flynn-Effekts sowie die Forschung zur Ursache der Testnormverschiebungen ist nach wie vor noch nicht abgeschlossen.

Erklärungsansätze

In den letzten Jahrzehnten rückt die Frage nach den Gründen für die Anstiege und vereinzelt Gegenentwicklungen des globalen IQs immer mehr in den Fokus. Viele verschiedene, mögliche Einflussfaktoren stehen im Zentrum der Forschung. Dazu gehören sowohl umweltbezogene Einflüsse und biologische Aspekte als auch komplexere Ursachengefüge. Diskutiert wurden bereits Einflüsse des technologischen Fortschritts (Neisser, 1997), der Geburtenreihenfolge (Zajonc & Mullally, 1997) oder des heterogener werdenden Genpools (Mingroni, 2004). Deren Evidenz wird jedoch als mäßig bewertet (Pietschnig & Voracek, 2015). Die starken IQ-Zuwachsraten über relativ kurze Zeit scheinen sich außerdem zu schnell zu entwickeln als dass rein genetische Ursachen als Erklärung standhalten könnten. Wahrscheinlicher scheinen daher vor allem umweltbezogene Einflüsse (Te Nijenhuis & van der Flier, 2013). Die meiste wissenschaftliche Anerkennung erhalten derzeit vor allem folgende Einflussfaktoren und Modelle.

Bildung

Die positive Auswirkung verbesserter Bildung auf die steigende Intelligenz (Teasdale & Owen, 2005) über Generationen hinweg ist intuitiv naheliegend. Bildung scheint jedoch nur einen gewissen Anteil des Einflusses auf die IQ-Testnormverschiebungen auszumachen. Vollständig erklären kann sie das komplexe Muster des Flynn-Effekts nicht. Offen bleibt bei diesem Ansatz unter anderem, warum die stärkeren Zuwachsraten gerade in der eher lernunabhängigen, fluiden Intelligenz zu finden sind (Pietschnig & Voracek, 2015).

Ernährung

Ebenso plausibel klingt die förderliche Wirkung besserer Ernährung auf die Entwicklung kognitiver Fähigkeiten. Ausgewogenere Nährstoffzufuhr bereits vor und kurz nach der Geburt könnte die geistige Entwicklung über viele Jahre hinweg tatsächlich verbessert haben (Colom et al., 2005, 2007; Lynn, 2009b). Funde über stärker ausgeprägte IQ-Zuwachsraten im schwachen bis mittleren Intelligenzbereich decken sich ebenfalls mit der Ernährungshypothese (Colom et al., 2005).

Pathogener Stress

Hinzu kommen Überlegungen zum Einfluss geringeren pathogenen Stresses. Die verbesserte Gesundheitsvorsorge und die damit verbundene geringere Prävalenz von Infektionskrankheiten wird mit der steigenden Intelligenzleistung in Verbindung gebracht. Ähnlich wie beim Ernährungsansatz wird hierbei argumentiert, dass sich das Gehirn in Abwesenheit anderer Stressoren besser entwickeln kann (Eppig et al., 2010).

Testverhalten

Bei dieser Hypothese wird ein geändertes Testverhalten für das zunehmend bessere Abschneiden in fluiden (Must & Must, 2013) und kristallinen (Pietschnig et al., 2013) Intelligenzaufgaben verantwortlich gemacht. Dies wurde mittels Archivdaten, gesunder und klinischer Stichproben, analysiert. Bis in die Anfänge des 21. Jahrhunderts zeigten die Daten, dass Antwortbögen zunehmend öfter vollständig bearbeitet wurden. Ebenso minimierte sich die Anzahl der ausgelassenen Multiple-Choice Antworten (Must & Must, 2013). Die Annahme dazu ist, dass Personen über diesen Zeitraum immer mehr dazu tendierten die schwierigsten Fragen wenigstens zu erraten. Abnehmende Reliabilitäten können als Evidenz für ein verändertes Antwortverhalten gedeutet werden (Pietschnig et al., 2013). Gegen diesen Erklärungsansatz sprechen jedoch steigende IQ-Funde bereits weit vor der Verbreitung von Multiple-Choice Testaufgaben (Pietschnig & Voracek, 2015). Zudem wäre die geringere Auslassung von Antworten gleichzeitig eine rein logische Folge der gestiegenen Intelligenzleistung (Pietschnig et al., 2013). Auch dieser Faktor könnte demnach, wenn überhaupt, nur Anteile am Effekt ausmachen und müsste seinen

maximalen Einfluss auf die IQ-Anstiege bereits überschritten haben (Pietschnig & Voracek, 2015).

Soziale Multiplikatoren

Ein vielversprechendes Modell ist das der individuellen und sozialen Multiplikatoren. Hier wird von einem wechselseitigen und sich potenzierenden Zusammenspiel zwischen genetischen und umweltbezogenen Faktoren als Ursache für den Flynn-Effekt ausgegangen. Schon leicht überdurchschnittliche kognitive Fähigkeiten einer Person führen häufig dazu, dass sich ihre Umweltbegebenheiten für sie verbessern. Zum Beispiel in Form von individueller Förderung eines Talents oder einer besseren Schulform. Solche Veränderungen können daraufhin das Intelligenzwachstum des Individuums weiter ankurbeln. Dies wiederum kann zur Folge haben, dass sich die Umwelt weiter positiv an sie anpasst. Dieser Mechanismus führt unter Umständen zu einem sich multiplizierenden Effekt. Zum anderen wirken soziale Multiplikatoren. Schon ein leichter IQ-Anstieg in der Gesellschaft kann ein Individuum zur Verbesserung der eigenen Leistung motivieren. Dies erhöht die durchschnittliche Intelligenz weiter. Personen mit höherem IQ im Umfeld anderer Personen wirken dann als Multiplikatoren der Intelligenz (Dickens & Flynn, 2001). Diese simultan verlaufenden, selbstverstärkenden Prozesse hätten das Potential einige Aspekte der starken IQ-Zunahmen über die Jahrzehnte zu erklären. Es mangelt jedoch bisher an empirischen Untersuchungen zur Prüfung dieser Hypothese (Nisbett et al., 2012). Auch aktuellere Daten sprechen für einen Erklärungsanteil der sozialen Multiplikatoren an den domänenspezifischen Unterschieden, jedoch steht die direkte Testung nach wie vor aus (Pietschnig & Voracek, 2015).

Lebensgeschichtliche Entwicklung

Ein anderes Modell beschäftigt sich mit der Auswirkung der langsameren Lebensentwicklung auf den Flynn-Effekt. Durch die Verbesserung von Ernährung, medizinischer Versorgung und kleineren Familienstrukturen könnte mehr Zeit in Bildung und das Erlernen spezifischer kognitiver Fähigkeiten investiert werden. Der Fokus würde sich hin zu langfristigen Zielen verlagern und das Thema der Fortpflanzung erst später im Leben in den Vordergrund rücken (Woodley, 2012). Evidenz zur Assoziation des Flynn-Effekts mit zunehmender Fähigkeitsspezialisierung

bestärken diese Hypothese. Trotzdem benötigt auch dieser Ansatz noch weitere Fundierung (Woodley & Madison, 2013).

Bisher gibt es keine Theorie, die für sich genommen genug Evidenz für die vollständige Aufklärung des komplexen Phänomens liefert. Als wahrscheinlich gilt vor allem das multikausale Zusammenwirken verschiedener vor allem umweltbezogener Ursachen auf die Entwicklung des IQs. So unterstützen Daten verschiedener Studien oftmals unterschiedliche Faktoren zu unterschiedlichen Anteilen (u.a. Daley et al., 2003; Pietschnig & Voracek, 2015).

Obsolete Items

Ob genetische Faktoren oder Einwirkungen aus der Umwelt nun den größeren Einfluss auf den Flynn-Effekt haben, wurde viel diskutiert. Der potenzielle Einfluss des Testinstruments selbst blieb dabei weitgehend unbeachtet. Dabei wurde bereits vor über 20 Jahren der Einfluss der Inhalte auf das Abschneiden in einem Intelligenztest thematisiert und eine Prüfung für notwendig befunden. Fragen, deren Antworten zum Zeitpunkt der Normierung noch knifflig waren, könnten über die Zeit in den generellen Wissenskonsens eingegangen sein. Daraufhin wäre die Beantwortung heute leichter als damals (Rodgers, 1999).

In die umgekehrte Richtung deutet eine Hypothese, die letztes Jahr im Zuge einer Masterarbeit an der Universität Wien aufgestellt wurde. Domänenspezifische Leistungsunterschiede könnten dadurch zustande kommen, dass veraltete Fragen nicht zunehmend leichter, sondern schwerer zu beantworten sind. Damals relevantes Wissen, könnte heute obsolet oder sogar falsch sein. Das würde vor allem die Inhalte aus kristallinen Intelligenzaufgaben betreffen. Denn auf Grund ihrer wissensbasierten Inhalte, die sich über die Zeit eventuell verändern, sind es meist diese Aufgaben, die aktualisiert werden müssen. Abstraktere Fragen aus fluiden Aufgaben können hingegen meist unverändert bleiben, da der Lösungsweg gleichbleibt. Veraltete Frageninhalte könnten den kristallinen Flynn-Effekt dadurch maskieren, dass ihre Antworten auf Grund von veraltetem Wissen nicht mehr korrekt oder zumindest schwerer zu beantworten sind (Höttl, 2019). Hat sich seit der Testerstellung beispielsweise der Name einer Hauptstadt geändert oder ein wissenschaftliches

Ergebnis revidiert, erfasst die Aufgabe nicht mehr sein ursprüngliches Ziel (Prüfung des deklarativen Wissens).

Der Einfluss obsoleter Items als mögliche Ursachen für den domänenspezifisch unterschiedlich ausfallenden Flynn-Effekt wurde in der Originalstudie mit einem querschnittlichen Studiendesign geprüft. Dazu wurden drei Untertests von zwei Intelligenztestversionen (IST-70 und IST-2000R) herangezogen, deren Normierungen 30 Jahre auseinanderliegen. Es wurde die mittlere Leistung zwischen den alten und neuen Items der kristallinen Untertests berechnet, die jeweils nur in einer Version vorkamen. Das Ergebnis zeigte, dass die IQ-Leistung in den veralteten Fragen tatsächlich signifikant geringer ausfiel als in der revidierten Form. Dies wurde als Hinweis für die Maskierung des kristallinen Intelligenzzuwachses durch obsoletere Items interpretiert. Es wurde daraus geschlossen, dass der kristalline Flynn-Effekt womöglich gar nicht stark von dem des fluiden abweicht, sondern durch die Maskierung eventuell unterschätzt wird (Hörtl, 2019). Damit würde es sich bei den vielfach erfassten Unterschieden in der Literatur eher um Artefakte als um eine domänenspezifische IQ-Leistungsdifferenz handeln. Die weitere Analyse dieses bisher einmaligen Fundes steht allerdings noch aus und ist daher Teil der vorliegenden Masterarbeit.

Fragestellung

Die Untersuchung des Flynn-Effekts ist von großer Bedeutung, da die IQ-Testnormverschiebungen konkrete Auswirkungen auf die Validität von Intelligenztests haben. Das Wissen über das Ausmaß der Veränderungen ist wichtig für die Beurteilung und Renormierung von Intelligenztests (Pietschnig et al., 2010). Andernfalls könnten wichtige intelligenzbezogene Entscheidungen auf Grundlage ungültiger Testergebnisse getroffen werden. Beispielsweise könnte der IQ einer Person durch die Anwendung veralteter Intelligenztests überschätzt werden, während ihr bei der Intelligenzmessung mit aktuellen IQ-Normen eine Lernstörung diagnostiziert worden wäre. Ihr IQ könnte also unauffällig erscheinen, obwohl eigentlich eine Lernstörung vorliegt. Damit würde die Aktualität der Testnormierung über den Erhalt notwendiger Fördermaßnahmen entscheiden (Truscott & Frank, 2001).

Dieses Forschungsfeld erfreut sich bereits immer größerer Aufmerksamkeit. Spätestens seit den Befunden zur möglichen Umkehr der Intelligenzentwicklung ist das Interesse der breiten Gesellschaft geweckt worden und die Besorgnis groß. Doch noch immer ist der Ursprung des Phänomens nicht ausreichend aufgeklärt. Dabei könnten wissenschaftlich fundierte Ergebnisse wichtigen Entscheidungstragenden (z.B. in der Politik) helfen, sinnvolle Prozesse in Gang zu setzen, um die zukünftige Weiterentwicklung der globalen Intelligenz gezielt zu stabilisieren und zu verbessern.

Domänenspezifische Unterschiede im Rahmen des Flynn-Effekts treten trotz ihrer ebenso frühen Entdeckung nach wie vor noch seltener in den Fokus der Wissenschaft. Noch immer gibt es keinen Konsens darüber, weshalb der Effekt in der fluiden Intelligenzdomäne meist deutlich stärker ausfällt als in der kristallinen Intelligenzdomäne. Umso relevanter ist die weitere Auseinandersetzung mit den ersten Ergebnissen eines Erklärungsversuches zum potentiell maskierenden Effekt obsoleter Items auf den kristallinen Flynn-Effekt (Höttl, 2019). Während die Untersuchung der Hypothese mit dem IST-Intelligenztest dort bereits Evidenz zeigte, steht die Prüfung mit anderen bekannten Intelligenztest wie dem Wechsler-Intelligenztest noch aus.

Daten aus dieser Intelligenztestbatterie flossen auf Grund ihrer internationalen Verbreitung schon von Beginn an in die Untersuchung des Flynn-Effekts mit ein und bestimmen somit die bisherigen Forschungsergebnisse maßgeblich mit (u.a. Dutton & Lynn, 2015; Erzberger & Engel, 2010; Flynn, 1984; Pietschnig & Voracek, 2015; Satzger et al., 1996; Trahan et al., 2014; Weiss et al., 2016). Umso wichtiger ist es daher, mögliche maskierende Effekte in diesem Testformat ebenso zu prüfen. Ansonsten könnten diese den kristallinen Leistungsanstieg unbemerkt verdecken. Ein solcher Fund könnte die in der Vergangenheit gefunden domänenspezifischen Unterschiede relativieren.

Die vorliegende querschnittliche Studie umfasst zwei Ziele. Zum einen soll der domänenspezifische Flynn-Effekt im deutschsprachigen Raum untersucht werden. Zum anderen soll die soeben beschriebenen Hypothese von Hölzl (2019) anhand von zwei Wechsler Intelligenztestversionen konzeptuell repliziert werden. Dazu werden die bereits vorgestellten drei Untertests aus dem HAWIE-R (Tewes, 1994) und dem WIE (Von Aster et al., 2006) herangezogen: Der Untertest AW als Messinstrument der kristallinen Intelligenzdomäne, RD zur Erfassung der fluiden Intelligenzkomponente und GF zur Messung einer Mischform aus fluiden und kristallinen Intelligenzanteilen.

Zunächst wird der Flynn-Effekt generell untersucht. Dazu wird angenommen, dass sich die durchschnittliche IQ-Ausprägung in den beiden Testversionen (HAWIE-R und WIE) signifikant voneinander unterscheidet, wenn ein Flynn Effekt vorliegt (*Hypothese 1*). Des Weiteren wird explizit auf domänenspezifische Leistungsunterschiede getestet. Ein entsprechender Effekt würde sich durch die signifikante Interaktion zwischen den ausgewählten Untertests (AW, GF und RD) und den Testversionen zeigen (*Hypothese 2*). Zuletzt wird der Einfluss obsoletter Items auf den Flynn-Effekt analysiert. Ein solcher Einfluss liegt vor, wenn sich die durchschnittliche Leistung der abweichenden Items zwischen den Testversionen unterscheidet (*Hypothese 3*).

Methode

Im Folgenden wird der Versuchsplan der vorliegenden Studie dargestellt. Dazu werden im folgenden Abschnitt das Untersuchungsdesign, die Stichprobe, die Durchführung und die verwendeten Erhebungsinstrumente beschrieben. Zuletzt wird auf die statistische Datenanalyse eingegangen.

Untersuchungsdesign

Es handelt sich bei dieser Studie um ein querschnittliches Experiment. Die Normierungsstichproben von zwei Wechsler Intelligenztestversionen wurden indirekt in einem within-subject Design verglichen. Es handelte sich dabei nicht um ein klassisches Retestverfahren, in dem die beiden Testversionen der Stichprobe jeweils in kompletter Form nacheinander dargeboten werden. Stattdessen wurde eine kombinierte Testdurchführung (z.B. Höttl, 2019; Satzger et al., 1996) angewendet. Im Rahmen dieses Vorgehens wurden die Aufgaben beider Testversionen manuell so zusammengestellt, dass eine gleichzeitige Testung möglich war. Dazu wurden sechs angepasste Testhefte erstellt, deren Inhalt sich mit den originalen Untertesthalten deckt, mit denen identische Inhalte jedoch nicht doppelt abfragt wurden. Dies änderte teilweise die Reihenfolge der Fragen innerhalb eines Untertests. So enthält eine Hälfte der Testhefte die Untertests in unveränderter Form des HAWIE-R. Die andere Hälfte enthält sie in Originalreihenfolge des WIE. Pro Untertest wurden anschließend zusätzlich die Fragen aus der jeweils anderen Testversion angehängt, die nur in einer Version vorkommen (N-Items, Beschreibung der Item-Kategorisierung folgt). Um eine Verzerrung durch Reihenfolgeeffekte zu vermeiden, wurde die Reihenfolge der Untertestdarbietung zusätzlich balanciert. Durch dieses Vorgehen konnte nach der Durchführung von jeder Testperson ein vollständiger Rohwertscore, pro Untertest und Testversion, berechnet werden: Insgesamt sechs pro Person. Die Zuordnung der Teilnehmenden zu einem der daraus resultierenden sechs Testhefte erfolgte randomisiert (per Losverfahren). Die Antworten wurden am Laptop mitgeschrieben und anonymisiert ausgewertet.

Das geplante Vorgehen wurde bereits am 18.11.2019 auf OSF präregistriert, um die Transparenz der wissenschaftlichen Arbeit zu erhöhen (Link zur dazugehörigen Registrierung: <https://osf.io/v4kjs>).

Stichprobe

Die Stichprobe umfasst Männer und Frauen mit österreichischer und deutscher Staatsbürgerschaft. Zwei weitere Inklusionskriterien sind zum einen der Altersrahmen von 18 bis 30 Jahren, zum anderen die Matura bzw. das Abitur als minimalstes Bildungsniveau. Die notwendige Stichprobengröße von 90 Personen für die geplante Untersuchung wurde mit Hilfe einer Poweranalyse (*G*Power* Software) berechnet. Dazu wurde eine zu erreichende Testmacht von .80 festgelegt, mit der ein mittlerer Effekt nach Cohen ($d = .3$) bei einer Alpha-Fehlerwahrscheinlichkeit von $\alpha = .05$ bei zweiseitiger Testung messbar werden sollte.

Abweichend von der Präregistrierung wurde die Datenerhebung schon Mitte März 2020, wegen geltender Kontaktbeschränkungen auf Grund der SARS-COV-2 Pandemie, vorzeitig abgebrochen. Zu diesem Zeitpunkt waren persönliche Befragungen nicht mehr möglich. Daher beläuft sich die endgültige Stichprobe auf $N = 85$ und weicht damit um 5 Personen von der berechneten Mindestgröße ab. Davon erhielten 43 Personen die Fragenformulierung des HAWIE-R und 42 die des WIE. Daraus ergeben sich die Teilstichproben $N_{HAWIE-R} = 43$ und $N_{WIE} = 42$.

Das Alter der Teilnehmenden variierte wie präregistriert zwischen 18 und 30 Jahren mit einem Mittelwert von $M = 25.4$ ($SD = 3.08$). Davon ordneten sich 56 dem weiblichen (65.9%) und 29 dem männlichen Geschlecht (Sex) (34.1%) zu. Die Nationalität teilte sich ebenfalls zu etwa zwei Drittel und ein Drittel auf. 55 Personen gaben eine deutsche (64.7%) und 30 eine österreichische Staatsbürgerschaft (35.5%) an. Für den Bildungsgrad ergaben sich drei Stufen. Die Mehrheit der Teilnehmenden wies einen Bachelorabschluss auf (44 Personen, 51.8%). Ansonsten gaben ähnlich viele Personen an eine Matura bzw. ein Abitur (19 Personen, 22.4%) oder einen Master- bzw. Diplomabschluss (22 Personen, 25.9%) absolviert zu haben.

Durchführung

Die Rekrutierung der Stichprobe erfolgte im Bekanntenkreis, sowie durch Aushänge an der Fakultät für Psychologie der Universität Wien und durch Aufrufe in den dazugehörigen universitären Foren auf Facebook. Als Anreiz wurde die Verlosung eines Gutscheins beworben. Unangekündigt erhielt jede Versuchsperson im Anschluss an die Teilnahme zusätzlich einen Schokoladenriegel als Dankeschön. Die Erhebung begann im Dezember 2019 und lief über dreieinhalb Monate. Die verbale Individualtestung fand in ruhigen Umgebungen im Privaten sowie an der Universität statt und wurde durchgehend von derselben Testleiterin administriert. Ein Durchlauf dauerte jeweils zwischen 40 bis 60 Minuten. Dabei wurden die Anweisungen aus den Testmanualen beachtet. Wie präregistriert, wurde die normalerweise geltende Umkehr- und Abbruchregel (Tewes, 1994; Von Aster et al., 2006) nicht angewendet, um von jeder Person Werte für jedes Item zu generieren. Dies war zur Beantwortung der dritten Hypothese notwendig.

Erhebungsinstrumente

Zur Untersuchung des Flynn-Effekts wurden, wie bereits dargestellt, drei Untertests aus den deutschsprachigen Testversionen HAWIE-R und WIE aus dem Wechsler Intelligenztestformat ausgewählt (Tewes, 1994; Von Aster et al., 2006). Die Erhebung der jeweiligen Normierungsstichproben fand 1988 (HAWIE-R) und 2002 (WIE) statt. Damit liegen (anders als präregistriert) 14 Jahre dazwischen (Erzberger & Engel, 2010; Pietschnig & Voracek, 2015).

Administration und Auswertung des Wechsler Intelligenztests

Der Test besteht aus einem offenen Antwortformat und wird immer mündlich durchgeführt. Die Antworten werden von der Testleitung auf einem Protokollbogen notiert und anhand der Auswertungsrichtlinien im Manual ausgewertet. Auf die meisten Fragen gibt es nur eine richtige Lösung, die jeweils mit null oder einem Punkt bewertet wird. Bei einigen Fragen in Untertests des Verbalteils gibt es mehrere richtige Optionen und feste Bewertungskriterien, welche Art von Antwort 0, 1 oder 2 Punkte erhält. Die erfasste Leistung wird zunächst für jede Frage einzeln als Rohwert notiert und anschließend pro Untertest als Rohwertscore aufsummiert.

Anschließend wird dieser anhand der Normtabelle in einen Wertpunktscore umgewandelt. Von hier aus kann durch weitere Transformation der Gesamt-, Verbal- und Handlungs-IQ separat abgelesen werden.

Güte des HAWIE-R. Diese Testversion umfasst insgesamt elf Untertests und ist für die Altersspanne von 16 bis 72 Jahren normiert (Tewes, 1994). Angewendet wurde der Test in ganz Deutschland, sowie in Österreich und der Schweiz. Normiert wurde er jedoch nur für die alten Bundesländer Deutschlands (Erzberger & Engel, 2010; Von Aster et al., 2006).

Als Reliabilitätsmaß wird im Manual eine hohe innere Konsistenz von Alpha .95 für den Gesamttest für die Altersgruppe 20 bis 24 angeführt. Für die ausgewählten Untertests werden dort Werte zwischen .73 (RD), .81 (GF) und .83 (AW) angegeben. Der Standardmessfehler wird für den Gesamttest der oben genannten Altersgruppe mit 3.35 angeführt und variiert zwischen 1.08 und 1.53 in den drei Untertests. Die Auswertungsobjektivität des Wechsler Intelligenztests wird auf Grund des offenen Antwortformats laut Manual kritisiert. Es liegen Werte zum Vergleich der Objektivität zwischen erfahrenen und unerfahrenen Durchführenden bei der Beurteilung der verbalen Untertests vor. Diese bieten nur einen groben Überblick, da hierzu nur die Testergebnisse von vier Personen mit 16 Auswertenden verglichen wurden. Sie weisen eine geringere Streuung zwischen den Erfahrenen auf. Um eine gute Qualität der Ergebnisse zu garantieren, wird die Durchführung daher nur qualifiziertem Personal empfohlen (Tewes, 1994).

Güte des WIE. Diese Revision umfasst neben den bestehenden elf, drei neue Untertests. Einige Fragen wurden außerdem aktualisiert und neue hinzugefügt sowie der Normierungsbereich auf 89 Jahre erweitert. Die Renormierung war zudem dringend notwendig, da die Normierungen des HAWIE-R seit der deutschen Wiedervereinigung nicht mehr aktuell waren. Nachdem der HAWIE-R trotz fehlender entsprechender Normierung im gesamten DACH-Raum bereits zum Einsatz kam, wurde der WIE diesmal explizit für den Einsatz auch in Österreich und der Schweiz ausgelegt. Sämtliche Materialien, Inhalte und Anweisungen wurden dazu überarbeitet (Erzberger & Engel, 2010; Von Aster et al., 2006).

Speziell für die verwendeten Untertests werden im Manual für die Altersgruppe 20 bis 24 hohe Split-Half-Reliabilitäten angegeben: AW .90, RD .91 und GF .81. Die Standardfehler bewegen sich für die genutzten Untertests zwischen 1.13 und 1.53 in der entsprechenden Altersgruppe. Die Kritik der Auswertungsobjektivität besteht auch hier (Von Aster et al., 2006).

Darstellung der ausgewählten Untertests

Im Folgenden wird der Inhalt der drei Untertests sowie deren konzeptuelle Zuordnung zu den Intelligenzdomänen dargestellt.

AW. Hier werden Fragen zu erlerntem Wissen aus den Richtungen Naturwissenschaft, Geschichte und Gesellschaft erfragt. Zum Teil wird nach der Rolle bestimmter Personen, Namen von Hauptstädten oder bestimmten Prozessabläufen gefragt. Die Antworten werden bei diesem Untertest generell im Nachhinein mit den Bewertungsrichtlinien verglichen und mit Null oder einem Punkt bewertet. Im HAWIE-R umfasst dieser Untertest 24, im WIE 28 Fragen (Tewes, 1994; Von Aster et al., 2006). Auf Grund der Bildungsabhängigkeit und der stark verbalen Komponente wurde dieser Untertest, in Übereinstimmung mit dem Manual, in der vorliegenden Studie zur Messung der kristallinen Intelligenz eingesetzt.

RD. Bei diesem Untertest werden den Teilnehmenden verschiedene mathematische Textaufgaben vorgelesen. Diese müssen je nach Frage innerhalb von 15, 30, 60 oder 120 Sekunden beantwortet werden. Die Zahlen bewegen sich im Hunderterraum und die Anwendung einfacher Grundrechenregeln ist notwendig. Auf Nachfrage darf die gesamte Frage von der Testleitung erneut vorgelesen werden, die Nennung einer bestimmten Zahl als Hilfestellung ist jedoch nicht erlaubt. Bei manchen Items wird bei schneller Lösung in unter 10 bzw. 15 Sekunden ein extra Punkt vergeben. Erfolgt die Beantwortung erst nach dem Gesamtzeitlimit erhält die Person auch bei richtiger Antwort keinen Punkt mehr für diese Frage. Im WIE umfasst diese Aufgabe 20 Fragen, während es im HAWIE-R nur 14 sind (Tewes, 1994; Von Aster et al., 2006). In dieser Untersuchung wurde der Untertest zur Erfassung der fluiden Intelligenz herangezogen. Dies stützte sich auf die im Manual

angegeben Zuordnung zur Skala des Arbeitsgedächtnisses sowie auf die abstrakten Problemlösungskomponente der Berechnungen.

GF. In diesem Untertest werden pro Frage zwei Begriffe genannt. Die befragte Person soll die Gemeinsamkeit dazu nennen bzw. erklären. Darunter befinden sich hauptsächlich Gegenstände und Aspekte des Alltags. Wie auch beim Untertest AW gibt es zur inhaltlichen Bewertung der genannten Lösung Auswertungsrichtlinien. Hierbei wird nach Abstraktionsgrad der Antwort unterschieden. Wird die wichtigste Gemeinsamkeit genannt, werden zwei Punkte vergeben. Bei nebensächlicheren gemeinsamen Eigenschaften nur ein Punkt. Im HAWIE-R enthält dieser Untertest 16, im WIE 19 Fragen (Tewes, 1994; Von Aster et al., 2006). In dieser Untersuchung wurde in Anlehnung an das Manual angenommen, dass der Untertest sowohl wissensbasierte (kristalline) als auch abstrakte, logische (fluide) Komponenten erfasst und somit zur Erfassung einer Mischform zwischen fluider und kristalliner Intelligenz geeignet ist.

Angepasste Testhefte für das kombinierte Vorgehen

Für die Konzeption der bereits beschriebenen Testhefte wurden alle 121 Items aus den Untertests der beiden Testversionen in drei Kategorien eingeteilt. Je nachdem, ob die Fragen in den Versionen gleich, in ähnlicher Form oder lediglich in einer der beiden vorkommen. Das entspricht dem gleichen Vorgehen wie in der Originalstudie (Hörtl, 2019). Die daraus resultierten drei Item-Kategorien werden im Folgenden kurz dargestellt.

Äquivalente Items (E-Items). Dazu zählen Fragen, die in beiden Testversionen identisch sind. Bei der Durchführung wurden sie allen 85 Personen in gleichem Wortlaut vorgelesen.

Angenommen-äquivalente Items (A-Items). Dies sind Fragen, die inhaltlich zwar identisch in beiden Versionen vorkommen, deren Wortlaut jedoch voneinander abweicht. Um die inhaltliche Äquivalenz nach der Durchführung statistisch überprüfen zu können, erhielten 43 Personen die Fragen im Wortlaut des HAWIE-R und 42 in dem des WIE.

Nicht-äquivalente Items (N-Items). Hierbei handelt es sich um Fragen, die nur in einer Testversion vorkommen. Dazu gehören zum einen Fragen aus dem HAWIE-R, die im Zuge der Testrevidierung nicht mehr in den WIE mitaufgenommen wurden. Zum anderen, um neu konzipierte Fragen aus dem WIE, die es im HAWIE-R noch nicht gab. Wie bereits beschrieben, wurden diese Items in den Testheften an die jeweils andere Untertestversion angehängt und allen 85 Personen dargeboten.

Statistische Analyse

An dieser Stelle wird das Vorgehen der Datenaufbereitung und die zur Beantwortung der Hypothesen ausgewählten Analyseverfahren dargestellt.

Datenaufbereitung in SPSS

Die beantworteten Fragen wurden mit den jeweiligen Bewertungsrichtlinien verglichen und auf einer Skala von 0 bis 1 bzw. 0 bis 2 Punkten (je nach Untertest) bewertet. Die Lösungen aus den 121 Items wurden als Rohwerte pro Versuchsperson und Item einzeln in einem SPSS Dokument angelegt. Anschließend wurden pro Untertest und Testversion die einzelnen Rohwertpunkte zu sechs Rohwertscores aufsummiert.

Zur Beantwortung der ersten beiden Hypothesen (H1 und H2) zum Flynn-Effekt, allgemein sowie domänenspezifisch, war die Bildung von IQ-Werten notwendig. Auf Grund spezifischer Testcharakteristika konnten diese aber nicht abgeleitet werden. Daher wurden (abweichend von der Präregistrierung) relative Häufigkeiten aus den Rohwertscores berechnet, um die Leistungen zwischen den beiden Testversionen trotzdem vergleichbar zu machen.

Zur Prüfung der dritten Hypothese (H3), zu möglichen Einflüssen obsoleter Items, wurden nur die Rohwertpunkte der einzelnen N-Items pro Untertest und Testversion zu eigenen Rohwertscores summiert. Diese wurden

(präregistrierungskonform) in relative Häufigkeiten transformiert. Schlussendlich erfolgte das gleiche Vorgehen noch mit den A-Items, um deren inhaltliche Äquivalenz prüfen zu können.

Berechnungen zur Beantwortung der drei Hypothesen

Im folgenden Abschnitt wird das statistische Vorgehen zur Beantwortung der Hypothesen dargestellt.

Untersuchung auf einen Flynn-Effekt (H1). Hierfür wurde eine zweifaktorielle ANOVA (2x3) mit Messwiederholung angewendet. Statt der IQ-Leistung als Zielvariable (AV) wurde die Leistung in Prozent als AV explorativ ausgewählt. Die dazugehörigen zwei Faktoren stellten zum einen die Testversion mit zwei Stufen (HAWIE-R und WIE), zum anderen den Untertest mit drei Stufen (AW, RD, GF) dar. Unter Annahme gleicher Itemschwierigkeiten zwischen den Testversionen, kann die explorative ANOVA interpretiert werden. Die Effektstärke, partielles η^2 , wurde dabei nach Cohen (1988) interpretiert. Die Grenze für einen kleinen Effekt lag demnach bei .01, für einen mittleren Effekt bei .06 und für einen großen Effekt bei .14.

Untersuchung auf domänenspezifische Unterschiede (H2). Der signifikante Interaktionseffekt zwischen Untertest und Testversion aus der zweifaktoriellen ANOVA (2x3), wurde mit einer Simple Effect Analyse untersucht. Dazu wurden drei abhängige t -Tests berechnet. Diese dienten der Analyse der Richtung und Stärke der Differenz zwischen den Testversionen pro Untertest separat, um Hinweise für domänenspezifische Leistungsunterschiede zu analysieren.

Untersuchung auf Einflüsse obsoleter Items (H3). Hierzu wurde die Mittelwertdifferenz der Rohwertscores der N-Items pro Untertest zwischen den Testversionen mit Hilfe von drei abhängigen t -Tests berechnet. Die Analyse der Richtung und Stärke der Differenzen sollte Evidenz für einen möglichen Einfluss obsoleter Items auf die durchschnittliche Leistung in Prozent liefern. Die Effektstärken (Cohens d) der Mittelwertdifferenzen aus dem Vergleich der N-Items wurden anschließend explorativ mit der Formel $\left(\frac{d * 15}{14}\right) * 10$ in IQ-Punktdifferenzen pro Jahrzehnt umgewandelt. Die Formel ergibt sich aus der IQ-Metrik mit $M = 100$ und $SD = 15$. Da ein

Cohens $d = 1$ einer Standardabweichung entspricht, muss d für die Umrechnung in die IQ-Differenz zwischen den Messzeitpunkten mit 15 multipliziert werden. Um die Veränderung pro Jahr zu generieren, wird der Wert durch den Zeitraum zwischen den beiden Normierungen (hier 14 Jahre) dividiert. Zur besseren Vergleichbarkeit mit Angaben zur IQ-Veränderungen pro Jahrzehnt aus anderen Studien wurde die jährliche Veränderung anschließend noch mit 10 multipliziert.

Prüfung auf inhaltliche Äquivalenz

Es wurde geprüft, ob der unterschiedliche Wortlaut der A-Items, wie angenommen, keinen Einfluss auf die durchschnittliche Leistung in Prozent hatte. Dazu wurden, anders als präregistriert, drei unabhängige t -Tests mit der Testversion als unabhängige Variable (UV) berechnet.

Ergebnisse

In diesem Teil wird auf die deskriptive Statistik der Daten eingegangen und die berechneten Ergebnisse pro Hypothese erläutert. Ebenso wird die Voraussetzungsprüfung dargestellt.

Deskriptive Statistik

Die Betrachtung der Mittelwerte zwischen den Testversionen lässt Unterschiede erkennen. So liegt der numerische Wert für RD in der revidierten Testversion (WIE) höher als in der älteren Testversion (HAWIE-R). Gegenätzlich fällt der Wert für GF aus. Die Werte beim Untertest AW sind hingegen identisch. Die Spannweite fällt für den Untertest GF numerisch kleiner aus als für AW und RD (vgl. Tab.1).

Tabelle 1

Deskriptive Statistik der relativen Häufigkeiten pro Untertest und Testversion

	M (SD) ^a	Spannweite	Min	Max
<i>Allgemeines Wissen (AW)</i>				
HAWIE-R	.67 (0.12)	.54	.38	.92
WIE	.67 (0.13)	.68	.25	.93
<i>Rechnerisches Denken (RD)</i>				
HAWIE-R	.70 (0.14)	.63	.37	1
WIE	.74 (0.14)	.59	.36	.95
<i>Gemeinsamkeiten Finden (GF)</i>				
HAWIE-R	.84 (0.09)	.34	.63	.97
WIE	.80 (0.08)	.39	.58	.97

Anmerkung. ^a Die Mittelwerte beziehen sich auf die gesamte Stichprobe von $N = 85$

Hypothesenprüfung

Anschließend werden die Ergebnisse aus den Berechnungen der drei Hypothesen dargestellt.

Zum Flynn-Effekt (H1)

Im Folgenden werden die explorativ berechneten Ergebnisse mit der prozentualen Leistung als AV beschrieben.

Die Betrachtung der Voraussetzungen für die Durchführung der zweifaktoriellen ANOVA (2x3) mit Messwiederholung ergab zwei Verletzungen. Da der Mauchly-Test auf Sphärizität hoch signifikant ausfiel ($p = .001$), kann diese nicht angenommen werden. Es wurde daher die Greenhouse-Geisser Korrektur interpretiert. Zudem befinden sich vereinzelt Ausreißer in den Daten. Dabei handelt es sich um unterschiedliche Personen, deren Leistung in jeweils einem Untertest auffällig abwich. Dies deutet auf unsystematische Abweichungen hin. Da kein spezifisches Vorgehen mit Ausreißern präregistriert wurde, blieben sie im Datensatz erhalten.

Die Analyse des Haupteffekts ergab einen großen Effekt (partiell η^2 von $\eta^2 = .44$) für den Einfluss des jeweiligen Untertests auf die prozentuale Leistung ($F(1.86, 156.13) = 65.98, p < .001, N = 85$). Für den Einfluss der Testversion auf die prozentuale Leistung zeigte sich hingegen kein Effekt (partiell $\eta^2 < .001, F(1, 84) = .013, p = .908, N = 85$). Der Interaktionseffekt zwischen den beiden Faktoren Testversion und Untertest auf die prozentuale Leistung wurde als groß (partiell $\eta^2 = .21$) interpretiert ($F(1.72, 144.19) = 22.79, p < .001, N = 85$).

Die graphische Darstellung der Interaktion (vgl. Abbildung 1) gab einen Hinweis auf eine hybride Interaktion. Während für den Untertest AW kein prozentualer Leistungsunterschied sichtbar wurde, zeigte sich für RD ein höherer Mittelwert in der revidierten Testversion (WIE). Für GF hingegen lag der Mittelwert in der älteren Testversion (HAWIE-R) höher.

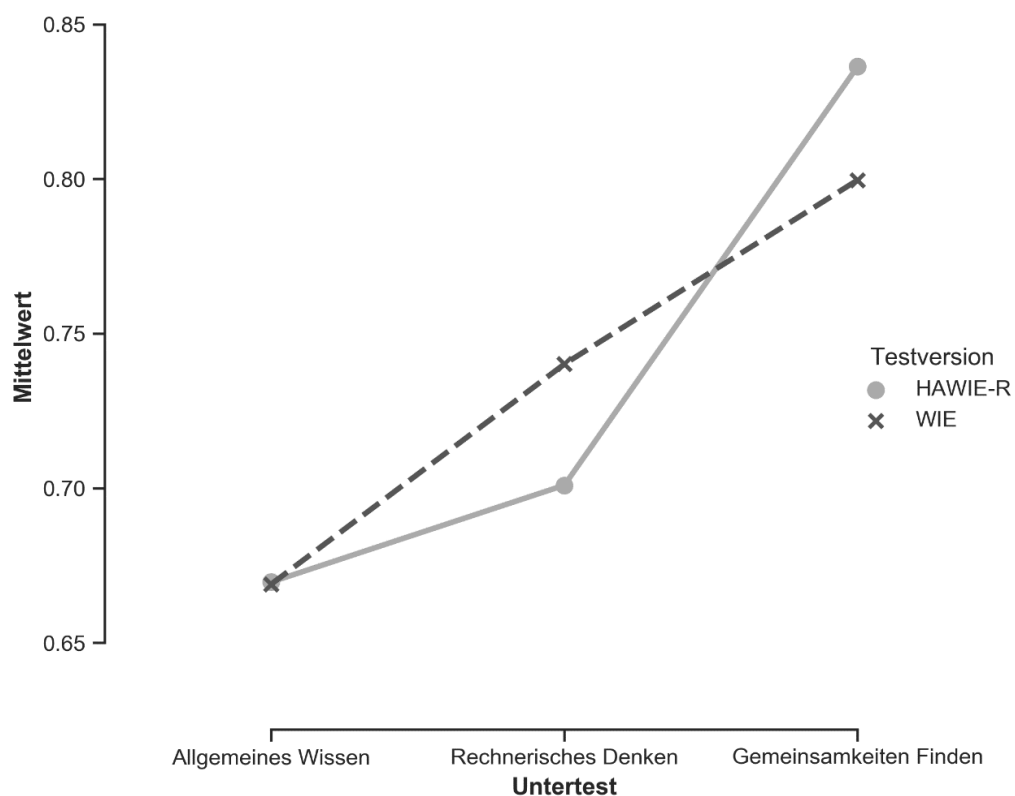


Abbildung 1. Graphik zum Interaktionseffekt zwischen den Faktoren Testversion und Untertest.

Die erste Hypothese war auf Grund der Formulierung nicht prüfbar. Denn die explorative Berechnung der ANOVA fand nicht mehr auf Ebene der IQ-Werte statt. Der signifikante Interaktionseffekt führte jedoch zur Annahme der zweiten Hypothese. Das machte die Simple Effects Analyse mittels abhängiger t -Tests nötig, um Aussagen über die Richtung der Differenzen zwischen Untertest und Testversion treffen zu können.

Zum domänenspezifischen Flynn-Effekt (H2)

Die Ergebnisse der gerade erwähnten drei abhängigen t -Tests werden pro Untertest aufgezählt.

- (A) Für den Untertest AW ergab sich ein trivialer Effekt ($d = -0.01$) der Differenz der prozentualen Leistung zwischen den Testversionen HAWIE-R (M

= .67, $SD = 0.12$) und WIE ($M = .67$, $SD = 0.13$), $t(84) = -0.07$, $p = .945$, $N = 85$.

(B) Für den Untertest RD hingegen wurde ein kleiner Effekt ($d = 0.28$) der Differenz der prozentualen Leistung zwischen den Testversionen gefunden. Der Mittelwert im HAWIE-R fiel kleiner ($M = .70$, $SD = 0.14$) aus als der im WIE ($M = .74$, $SD = 0.14$), $t(84) = 4.44$, $p < .001$, $N = 85$.

(C) Für den Untertest GF zeigte sich ein kleiner Effekt ($d = -0.43$) der Differenz der prozentualen Leistung zwischen den Testversionen. Der Mittelwert im HAWIE-R ($M = .84$, $SD = 0.09$) war größer als der im WIE ($M = .80$, $SD = 0.08$), $t(84) = -8.12$, $p < .001$, $N = 85$.

Für alle drei Untertests ergab sich eine durchwegs hohe Korrelation der durchschnittlichen prozentualen Leistung zwischen den beiden Versionen mit $r = .73$ ($p < .001$, $N = 85$) für den Untertest AW, $r = .83$ ($p < .001$, $N = 85$) für RD und $r = .88$ ($p < .001$, $N = 85$) für GF.

Zur Itemobsoleszenz (H3)

Die dritte Hypothese auf signifikante Differenzen der N-Items zwischen den Testversionen konnte für alle Untertests angenommen werden. Die Voraussetzung gleicher Itemschwierigkeiten zwischen den Testversionen konnte daher nicht erfüllt werden. Die Ergebnisse der drei abhängigen t -Tests pro Untertest nacheinander dargestellt.

(A) Die Betrachtung der N-Items des Untertests AW zeigte einen kleinen Effekt ($d = -0.25$) der Differenz der prozentualen Leistung zwischen den Testversionen an. Dieser fiel knapp nicht signifikant ($p = .052$) aus. Der Mittelwert war im HAWIE-R ($M = .64$, $SD = 0.20$) größer als im WIE ($M = .59$, $SD = 0.17$), $t(84) = -1.97$, $N = 85$.

(B) Für die N-Items des Untertests RD ergab sich ein kleiner Effekt ($d = -0.26$) der Differenz der prozentualen Leistung zwischen den Testversionen. Der

Mittelwert war im HAWIE-R ($M = .76$, $SD = 0.15$) größer als im WIE ($M = .72$, $SD = 0.17$), $t(84) = -2.11$, $p = .038$, $N = 85$.

- (C) Für die N-Items des Untertests GF ergab sich ein großer Effekt ($d = -1.08$) der Differenz der prozentualen Leistung zwischen den Testversionen. Der Mittelwert war im HAWIE-R ($M = .92$, $SD = 0.09$) größer als im WIE ($M = .82$, $SD = 0.10$), $t(84) = -7.52$, $p < .001$, $N = 85$.

Die explorative Umrechnung der Effektstärken (Cohens d) mit bereits genannter Formel, ergab für alle drei Untertests eine Zunahme in IQ-Punkten pro Jahrzehnt: Für AW betrug dies 2.7 IQ-Punkte, für RD 2.8 und für GF 11.6 IQ-Punkte.

Inhaltliche Äquivalenz

Die Ergebnisse der unabhängigen t -Tests zwischen den A-Items pro Untertest und Version werden im Folgenden pro Untertest erläutert. Alle drei wurden mit den Teilstichproben $N_{HAWIE-R} = 43$ und $N_{WIE} = 42$ berechnet. Die Annahme, dass die A-Items zwischen den beiden Testversionen inhaltlich äquivalent waren, erwies sich als berechtigt.

- (A) Die Analyse der A-Items des Untertests AW zeigte einen trivialen Effekt der Differenz ($d = 0.08$) zwischen den Mittelwerten der Testversionen HAWIE-R ($M = .70$, $SD = 0.23$) und WIE ($M = .71$, $SD = 0.18$) an. Dieser war nicht signifikant $t(83) = 0.37$, $p = .711$.
- (B) Für die A-Items des Untertests RD gab es einen kleinen Effekt ($d = 0.13$) der Differenz zwischen den Mittelwerten der Testversionen HAWIE-R ($M = .71$, $SD = 0.17$) und WIE ($M = .73$, $SD = 0.13$). Dieser war nicht signifikant $t(83) = 0.57$, $p = .567$.
- (C) Für die Mittelwerte der A-Items des Untertests GF ergab sich ein trivialer Effekt ($d = -0.04$) der Differenz zwischen den Testversionen HAWIE-R ($M = .89$, $SD = 0.21$) und WIE ($M = .88$, $SD = 0.23$). Auch dieser war nicht signifikant $t(83) = -0.18$, $p = .857$.

Diskussion

Das Ziel der aktuellen Studie war zum einen die querschnittliche Analyse fluider und kristalliner Testnormverschiebungen, zum anderen die konzeptuelle Replikation eines maskierenden Einflusses obsoleter Items auf den kristallinen Flynn-Effekt aus der Masterarbeit von Höttl (2019). Im folgenden Teil werden die Ergebnisse der aktuellen Studie in den Rahmen des aktuellen Wissensstandes eingeordnet und interpretiert. Abschließend werden Grenzen der vorliegenden Arbeit genannt und Implikationen für die zukünftige Forschung abgeleitet.

Kein Einfluss obsoleter Items auf den kristallinen Flynn-Effekt

In der vorliegenden Untersuchung konnte der Einfluss obsoleter Items auf den kristallinen Flynn-Effekt in den ausgewählten Wechsler Intelligenztests nicht konzeptuell repliziert werden. Stattdessen gaben die Berechnungen zur dritten Hypothese Evidenz für einen gegenläufigen Effekt im Vergleich zur Originalstudie (Höttl, 2019). Dort wurde beobachtet, dass die Leistung in den Items der älteren Intelligenztestversion niedriger ausfiel als in der moderneren Version. Dies wurde als Effekt obsoleter Items interpretiert. Im Gegensatz dazu zeigen die Mittelwertdifferenzen zwischen den nicht-äquivalenten Fragen der beiden Testversionen in dieser Studie einen umgekehrten Trend. In allen Untertests wurde sichtbar, dass die mittlere Leistung in den N-Items des neueren Tests (WIE) niedriger lag als in der älteren Testversion (HAWIE-R). Das könnte darauf hinweisen, dass die Lösung der älteren Items den Teilnehmenden in dieser Studie leichter fiel als die richtige Beantwortung der aktuelleren Fragen.

Ein Grund für den konträren Fund könnte in der deutlich kürzeren Zeitspanne zwischen den indirekt verglichenen Normierungsstichproben liegen. In der Originaluntersuchung lagen über 30 Jahre zwischen dem älteren und dem revidierten Intelligenztest (Höttl, 2019). Während der Normierungsabstand der hier verglichenen Wechsler Testversionen lediglich 14 Jahre betrug (Erzberger & Engel, 2010). Damit war der aktuell betrachtete Zeitraum nicht einmal halb so groß. Es wäre plausibel, dass ein Intelligenztest, der erst nach drei Jahrzehnten wieder aktualisiert wird, gravierendere inhaltliche Veränderungen aufweist als nach nur 14 Jahren. Außerdem

war die ältere Version bei Höttl (2019) insgesamt deutlich älter als die hier genutzte. Es wäre also logisch, dass kristalline Inhalte des IST-70, die 1970 normiert wurden, obsoleter sind als Wissensfragen des HAWIE-R, dessen Normierung von 1988 stammt (Erzberger & Engel, 2010). Höttl (2019) berichtete tatsächlich von Items, deren Inhalt heute als überholt gilt. Wie zum Beispiel die Nennung von Teilen eines elektronischen Geräts, welches inzwischen ganz anders konzipiert ist. Die passende Antwort dazu zu finden, dürfte heute deutlich schwerer geworden sein als es damals ursprünglich vorgesehen war. Für Fragen mit veraltetem Inhalt, erhält eine Person also unter Umständen keine Punkte, obwohl ihr Ausmaß an kristalliner Intelligenz im Vergleich zu ihrer Kohorte möglicherweise nicht unterdurchschnittlich abscheiden würde. Zur korrekten Beantwortung der alten Frage wäre heutzutage aber ein unverhältnismäßig detailliertes Wissen zu veralteten Inhalten nötig gewesen, welches für die ursprüngliche Normierungsstichprobe damals viel naheliegender war. Solche Items erfassen die Intelligenz folglich nicht mehr auf die vorgesehene Weise (sind zu schwer) und könnten den kristallinen IQ negativ verzerrt haben. In der vorliegenden Studie gab es jedoch keine Evidenz für diese Hypothese.

Beschreibung augenscheinlich auffälliger Items

Bei der Betrachtung der Frageninhalte der beiden Wechsler Intelligenztests erschienen im kristallinen Untertest AW des HAWIE-R nur wenige der N-Items auffällig. Ein Item aus dem HAWIE-R kommt dem in der Originalstudie beschriebenen Problem am nächsten. Dort wird nach einer Anzahl von Objekten gefragt, die nach aktuellem Stand der Wissenschaft wieder um eine Ziffer nach unten korrigiert wurde. Bei der Erhebung nannten die Teilnehmenden meist beide Zahlen und berichteten von der vor Kurzem erfolgten Aktualisierung. Sie wählten dann aber die heute gültige Anzahl als endgültige Antwort aus und erhielten dafür keinen Punkt. Nur Teilnehmende, die nichts von der Aktualisierung wussten, bekamen den Punkt. Dabei entsprach letztere Leistung eigentlich weniger kristalliner Intelligenz als die Leistung einer Person, der die Aktualisierung bekannt war. Denn diese hatte die Entwicklung in der Wissenschaft verfolgt und verfügte über detaillierteres Allgemeinwissen.

Zudem befanden sich in der veralteten Form des Untertests AW noch vier spezifische Fragen zur deutschen Geografie. Wie bereits erwähnt, wurde der HAWIE-R trotz fehlender Normierungen auch in der Schweiz und Österreich angewendet (Erzberger & Engel, 2010). Bei der Durchführung fiel auf, dass sich österreichische Teilnehmende erwartungsgemäß schwerer bei der Lösung taten. Es ist anzunehmen, dass die Itemschwierigkeit für sie nicht die Gleiche war wie für Teilnehmende, die das deutsche Schulsystem besucht hatten. Eine mögliche Verzerrung in den aktuellen Daten wäre jedoch nur durch einen systematischen Vergleich einer österreichischen und deutschen Stichprobe statistisch prüfbar gewesen.

Des Weiteren fiel in allen drei Untertests jeweils eine Frage auf, in der eine Formulierung veraltet war und zu leichten Schwierigkeiten während der Erhebung führte. Einzelnen Personen war das jeweilige Wort nicht bekannt. Nach kurzem Überlegen (bzw. Raten) kamen die meisten jedoch zum richtigen Schluss und beantworteten die Frage korrekt. Die Tendenz zum Erraten der richtigen Lösungen fiel in der gesamten Testung auf. Einige Antworten mit bereits genannten Schwierigkeiten wurden dadurch doch noch richtig gelöst. Die so gewonnenen Punkte entsprachen jedoch eigentlich nicht der interessierten kognitiven Fähigkeit, die der jeweilige Untertest messen sollte. Da die Untertests AW und GF keine Lösungsgeschwindigkeit erfassen, führte das Überlegen bzw. Raten nicht zu Punktabzügen. Das Verhalten könnte die Leistung dort leicht positiv verzerrt haben. In RD könnte das Verständnisproblem den gemessenen Lösungszeitraum verlängert und die Chance auf einen Zeitbonuspunkt, unabhängig von der eigentlichen Rechenleistung, geschmälert haben.

Zusammenfassend lässt sich sagen, dass im HAWIE-R augenscheinlich ein paar auffällige Items zu erkennen waren, die die korrekte Beantwortung erschwert haben könnten. Die dritte Hypothese (zu signifikanten Differenzen zwischen den Untertests und Testversionen) fiel jedoch in die entgegengesetzte Richtung signifikant aus als in der Originalstudie. Es konnte daher kein empirischer Hinweis zur Bestätigung der Hypothese des Einflusses obsoleter Items nachgewiesen werden. Die angenommene Hypothese kann stattdessen als indirekter Hinweis auf einen positiven Flynn Effekt interpretiert werden.

Indirekter Hinweis auf einen domänenspezifischen Flynn-Effekt

Die Ergebnisse aus den Berechnungen zur dritten Hypothese deuten darauf hin, dass die prozentuale Leistung in der älteren Testversion jeweils höher lag als in der aktuelleren Form. Das spricht für einen Leistungszuwachs (positiven Flynn-Effekt) in allen drei Aufgaben. In den Untertests AW und RD fiel dieser deutlich geringer aus als in GF. Damit spiegeln sie ein häufig gefundenes Muster in bisheriger Literatur wider (u.a. Flynn, 2007). Durch die Transformation der standardisierten Effektstärken (Cohens d) in IQ-Punktdifferenzen pro Jahrzehnt konnten die hier gefundenen Effekte im Folgenden mit bisherigen Forschungsergebnissen verglichen werden.

Für AW, der als Indikator für die Entwicklung der kristallinen Intelligenz ausgewählt wurde, betrug der IQ-Zuwachs pro Jahrzehnt 2.7 Punkte. Nahezu identisch fiel der Effekt für RD, als Maß für fluide Intelligenz, mit 2.8 Punkten aus. Für die Mischform GF mit fluiden und kristallinen Intelligenzkomponenten betrug die Steigerung sogar 11.6 IQ-Punkte pro Jahrzehnt und ist damit in etwa viermal so groß.

Konsistentes Muster in Wechsler Intelligenztest Studien

Bei vielen Studien, deren Ergebnisse in metaanalytische Untersuchungen des Flynn-Effekts eingegangen sind, handelte es sich um Äquivalenzprüfungen zwischen älteren und revidierten Testversionen. Viele davon verglichen die Kinderversion der Wechsler Intelligenztests (Carlton & Sapp, 1997; Dumont & Faro, 1993; Flynn, 2007; Gunter et al., 1995; Horn-Alsberge, 1999; Liu & Lynn, 2013; Sanborn et al., 2003; Te Nijenhuis et al., 2012; Truscott & Frank, 2001; Vance et al., 1996). Inhaltlich sind diese Wechsler Intelligenztests altersentsprechend an die junge Zielgruppe angepasst, der Skalenaufbau und die meisten Untertests entsprechen jedoch den Erwachsenenversionen. Überwiegend wurden dabei die Versionen WISC-R und WISC-III verglichen. Im Erwachsenenbereich gehen die Vergleiche vom WAIS bis zum WAIS-III (Erzberger & Engel, 2010; Fitzgerald et al., 2007; Khaleefa et al., 2009; Quereshi & Erstad, 1990; Satzger et al., 1996; Scott & Poncy, 1999; Simon & Clopton, 1984). Auf Skalenebene zeichnete sich bei den Ergebnissen der aufgeführten Studien ein überwiegend ähnliches Bild ab. Der Handlungs-IQ wies

gegenüber den anderen Skalen fast immer den stärksten Flynn-Effekt auf, dicht gefolgt vom Gesamt-IQ. Je nach Untersuchung fiel der Zuwachs im Verbal-IQ deutlich oder nur geringfügig schwächer aus, teilweise auch negativ. Dies gilt sowohl für die Ergebnisse aus den genannten Studien zu Wechsler Intelligenztests für Erwachsene als auch für die Kinderversionen.

Bei Studien, die neben den drei übergeordneten IQs auch Angaben zur Leistungsdifferenz auf Untertestebene trafen, zog sich die untertestspezifische Abweichung von GF durch die Ergebnisse (Carlton & Sapp, 1997; Erzberger & Engel, 2010; Flynn, 2007; Horn-Alsberge, 1999; Quereschi & Erstad, 1990; Satzger et al., 1996; Simon & Clopton, 1984; Truscott & Frank, 2001). Über verschiedene Testversionen, Länder und Zeitpunkte hinweg wies dieser Untertest im Vergleich zu den Differenzen der übrigen Untertests stets einen relativ großen Flynn-Effekt auf. Im Verhältnis zu den Untertests AW und RD fiel der Zuwachs für GF häufig überproportional groß aus. Dieses Phänomen deckt sich mit den Ergebnissen der vorliegenden Untersuchung.

In einer Dissertation zur Analyse der amerikanischen Äquivalente der hier verwendeten Testversionen betrug der Abstand zwischen den IQ-Mittelwerten pro Dekade für AW (0.1 IQ-Punkte) und GF (einen IQ-Punkt) sogar das Zehnfache. Für RD belief sich der Zuwachs auf 0.26 IQ-Punkte (Horn-Alsberge, 1999). In einem Äquivalenzvergleich zu den gleichen deutschsprachigen Versionen wie in dieser Untersuchung, betrug die verhältnismäßig große Leistungszunahme für GF 1.72 IQ-Punkte pro Jahrzehnt. Die Steigerung für RD lag hingegen lediglich bei 0.07 IQ-Punkten und für AW bei 0.68 IQ-Punkten (Erzberger & Engel, 2010). Beide Studien weisen damit den gleichen Trend wie in der vorliegenden Studie auf.

Zumindest strukturell ähnlich sah der Verlauf auch beim Vergleich der Vorgängerversionen HAWIE und HAWIE-R aus. Für RD und AW fiel die kognitive Leistungsveränderung negativ aus, während sie für GF leicht anstieg. Gemäß den Daten sank die Leistung für RD um - 0.27 IQ-Punkte und für AW um - 0.03 IQ-Punkte pro Dekade. Während der IQ für GF um 0.22 IQ-Punkte pro Jahrzehnt anstieg und sich damit auch in dieser Studie wieder von anderen Untertests abhob (Satzger et

al., 1996). Ähnlich dazu verlief der Flynn-Effekt für GF in einer Studie aus dem Sudan. Dort erweckte er durch die stärkste Leistungsabnahme auf Untertestebene Aufsehen. Diese entsprach einem IQ-Rückgang um - 1.7 IQ-Punkte pro Jahrzehnt, während der Effekt für AW mit 2.6 IQ-Punkten positiv ausfiel und damit dem Wert der vorliegenden Studie ähnelt. Der Effekt für RD kam jedoch nur auf eine Steigung von 0.5 IQ-Punkten pro Jahrzehnt (Khaleefa et al., 2009)

Ebenfalls deutlich fiel der Abstand einer amerikanischen Untersuchung zu WISC-R und WISC-III aus. Während der Flynn-Effekt für AW und RD als einzige Untertests nicht signifikant ausfiel, wurde im Gegensatz dazu für GF der zweitgrößte Anstieg mit 1.01 IQ-Punkten pro Jahrzehnt erfasst (Truscott & Frank, 2001). Auch in einer Datenanalyse des kognitiven Leistungszuwachses über vier Revisionen (WISC bis WISC-IV) hinweg zeigte sich Ähnliches. Für AW und RD wurde die gleiche IQ-Zuwachsrate von jeweils 0.4 IQ-Punkten pro Jahrzehnt erfasst. Für GF fiel der Anstieg mit 4.4 IQ-Punkten pro Dekade hingegen viermal so groß aus (Flynn, 2007). Damit gleicht die dortige untertestspezifische Differenz im Flynn-Effekt den Ergebnissen der aktuellen Studie am meisten.

Der überdurchschnittlich hohe Flynn-Effekt für den Untertest GF, den viele Untersuchungen beschrieben, kann jedoch nicht verallgemeinert werden. Laut aktueller Studienlage wurden sowohl positive wie negative IQ-Entwicklungen erfasst. Was sich jedoch wiederholt, ist der Extremwert für GF. Sein Wert hebt sich jeweils deutlich von anderen Untertests, wie AW und RD, ab. Insgesamt ist das Muster unterschiedlich starker IQ-Zunahmen je nach spezifischer Fähigkeit nicht neu. So ging auch Schallberger (1991, zitiert nach Erzberger & Engel, 2010, S.31) bereits von einem „differentiellen Charakter des Phänomens der [internationalen] IQ-Zunahme aus“.

Vergleich mit Überblicksstudien zu domänenspezifischen Unterschieden

Die Stärke des Flynn-Effekts für AW fällt im Vergleich zur erfassten kristallinen Intelligenzentwicklung anderer Studien stärker aus. Eine ältere Untersuchung, die als eine der ersten umfassenden gilt, gab noch einen höheren Anstieg um 3.7 IQ-Punkte in der kristallinen Intelligenzdomäne an (Flynn, 1987). Während die

Metaanalyse mit globalen Vergleichswerten einen schwächeren Anstieg von nur 2.1 IQ-Punkten pro Dekade berichtete (Pietschnig & Voracek, 2015). Die Ergebnisse der erwähnten Masterarbeit (Hörtl, 2019), deren Erklärungshypothese hier überprüft wurde, deuteten sogar auf einen negativen Effekt mit einer starken Abnahme von - 3.6 IQ-Punkten pro Jahrzehnt hin. Die aktuellen Daten decken sich damit am ehesten mit den Ergebnissen der Studie aus dem deutschsprachigen Raum. Deren Befunde ließen auf eine vergleichbar große IQ-Steigerung in fluider und kristalliner Intelligenz schließen (Pietschnig et al., 2010).

Das Ergebnis von RD ähnelt dem gefundenen fluiden Flynn-Effekt in der Originalstudie (Hörtl, 2019) mit 2.9 IQ-Punkten pro Jahrzehnt. Im Vergleich zum aktuellen Wert fielen die fluiden Zuwachsraten größerer Flynn-Effekt Studien jedoch stärker aus. In der globalen Metaanalyse betrug der Anstieg 4.1 IQ-Punkte pro Dekade (Pietschnig & Voracek, 2015). Die länderübergreifende Studie von Flynn (1987) gab sogar 5.8 IQ-Punkte pro Jahrzehnt an. Das etwas niedrigere aktuelle Ergebnis spiegelt interessanter Weise eher die Zuwachsrate des durchschnittlichen Gesamt-IQs von knapp 3 IQ-Punkten pro Dekade aus bekannten Studien wider (Flynn, 1984; Pietschnig & Voracek, 2015; Trahan et al., 2014).

Der in der vorliegenden Studie erfasste, viermal so große Flynn-Effekt für den Untertest GF, sowie dessen konsistente Sonderposition ist erstaunlich. Mehrheitlich stärkere fluide und geringere kristalline Flynn-Effekte aus bisheriger Literatur hätten für einen Effekt, dessen Stärke zwischen den beiden Domänen liegt, gesprochen. Ähnlich zeigten es die Ergebnisse von Hörtl (2019): Für den Untertest, der fluide Intelligenz erfassen sollte, wurde eine ähnlich große Zunahme wie in dieser Studie dokumentiert. In den Untertests, die zur Analyse kristalliner Komponenten ausgewählt wurden, sank der IQ hingegen um 0.5 und 3.6 IQ-Punkte pro Dekade. Die hier vorliegenden Daten zeigten jedoch keinen Hinweis auf ein ähnliches Muster. Während der Zuwachs in der fluiden eher moderat und in der kristallinen Domäne fast identisch dazu abschnitt, übertraf der große Flynn-Effekt für GF die beiden anderen deutlich.

Im Folgenden werden mögliche Einflüsse auf die Stärke der gefundenen domänenspezifischen Flynn-Effekte sowie Erklärungsansätze für das untertestspezifische Muster aus anderen Welcher Intelligenztest Studien erläutert.

Untertests erfassen zumindest anteilig kristalline Intelligenz

Alle hier verwendeten Untertests erfordern verbale Kompetenzen und gehören dementsprechend zum Verbalteil des Wechsler Intelligenztests (Tewes, 1994; Von Aster et al., 2006). Darum muss angenommen werden, dass jede Aufgabe zumindest zu einem kleinen Anteil kristalline Intelligenzaspekte miterfasst.

Generell wurde in der Literatur stets ein stärkerer IQ-Zuwachs im Handlungsteil der Wechsler Intelligenztests und sonstigen nonverbalen Intelligenztests beschrieben (u.a. Erzberger & Engel, 2010; Lynn, 2013; Must et al., 2003; Satzger et al., 1996). Ein tendenziell geringerer Flynn-Effekt in allen drei Aufgaben hätte sich demnach mit der Literatur gedeckt. Die starke Abweichung des Effekts für GF wird dadurch zwar nicht erklärbar, aber der geringe Unterschied zwischen den hier als fluid und kristallin eingestuften Untertests könnte sich daraus ableiten. Es wäre plausibel, dass der fluide Effekt aus RD in einer anderen rechenbasierten Aufgabe aus dem Handlungsteil, ohne verbale Textaufgabenkomponente, stärker vom kristallinen Effekt abgewichen wäre. In der Untersuchung von Hölzl (2019) wurde die fluide Intelligenzdomäne durch eine abstraktere Aufgabe abgefragt, in der Zahlenreihen ergänzt werden mussten. Obwohl der verbale Anteil darin geringer war, fiel der Effekt dort trotzdem nicht deutlich stärker aus als in dieser Studie. Das spricht wiederum eher gegen die eben getroffene Annahme. Der Vergleich der Stärke des Flynn-Effekts zwischen Aufgaben des Handlungs- und Verbalteils wurde in dieser Studie außerdem nicht untersucht.

Untertests laden ähnlich auf den g-Faktor

Der negative Zusammenhang des g-Faktors mit der IQ-Steigerung über die Zeit wurde mehrfach dokumentiert (u.a. Must et al., 2003; Te Nijenhuis & van der Flier, 2013). Ebenso gibt es Belege für die Tendenz, dass vor allem kristalline Intelligenzaufgaben hoch auf den g-Faktor laden (Johnson et al., 2004). Dieser Zusammenhang wurde in der Literatur bereits zur Erklärung des geringeren kristallinen

Flynn-Effekts herangezogen. In der aktuellen Studie weicht der Effekt zwischen der kristallinen (AW) und fluiden (RD) Aufgabe jedoch kaum voneinander ab. Eine EFA mit amerikanischen Wechsler Intelligenztestformaten verzeichnete bei allen der hier besprochenen Untertests vergleichbar hohe g-Ladungen (AW mit .75, RD mit .56 und GF mit .63; Johnson et al., 2004). Dieser Befund wäre durch die verbale Komponente aller Untertests erklärbar, würde aber eher für generell kleinere Effekte mit einer geringen Abweichung untereinander sprechen. Das trifft für AW und RD zwar zu, jedoch wäre für RD mit dem numerisch kleinsten Zusammenhang mit g der größte Effekt zu erwarten. Für GF wäre demnach ein Effekt, der dem Wert für AW ähnelt, schlüssig. Dessen starke Abweichung in den aktuellen Daten wird dadurch folglich auch nicht erklärbar.

Es lässt sich resümieren, dass der hohe verbale Aufgabenanteil und die Ladungen auf das psychometrische g den nahezu identisch ausfallenden Flynn-Effekt für AW und RD eventuell erklären könnte. Die Ursache für den starken Flynn-Effekt für GF bleibt jedoch nach wie vor unklar.

Erklärungsansätze aus anderen Wechsler Intelligenztest Studien

Eine explizite Erklärung für die stets starke Abweichung von GF wurde in Studien bisher eher selten angeführt. Viele Studien hatten den Flynn-Effekt als solches nicht als Untersuchungshypothese, sondern bewerteten die Vergleichbarkeit zweier Versionen. In Studien speziell zum Flynn-Effekt stand außerdem meist die Interpretation der Ergebnisse auf Ebene der Gesamt-, Handlungs- und Verbal-IQs (bzw. deren Interpretation als Hinweis auf die Entwicklung der fluiden und kristallinen Intelligenz) im Vordergrund.

In der Untersuchung zu den gleichen Testversionen wie in dieser Studie wurden die untertestspezifischen Unterschiede vor allem durch spezifisch bessere Bildung im Bereich des prozeduralen Wissens erklärt (Erzberger & Engel, 2010). Dies bezog sich auf eine Hypothese von Flynn (2007). Dabei wurde ein gesellschaftlicher Wandel in Richtung eines zunehmend wissenschaftlicheren Denkansatzes für die spezifisch starke Zunahme von GF verantwortlich gemacht. Eine solche Veränderung könnte demzufolge dazu führen, dass die Abstraktionsfähigkeit und das

hypothetisch logische Denken im Alltag an Relevanz gewinnen. Zugleich wurde angenommen, dass die wissenschaftliche Betrachtung von Problemen (mehr: abstrakte Kategorisierung, losgelöst von konkreten Sachlagen) in der Schule zunehmend mehr gefördert wird. Diese beiden Entwicklungen könnten sich gegenseitig verstärken (Flynn, 2007; Flynn & Weiss, 2007). Laut den Autoren wird diese Entwicklung durch die zunehmende Anwendung von flexiblen Führungs- und Problemlösefähigkeiten von Universitätsabgehenden in der Arbeitswelt sichtbar. Die reine Durchführung nach Schema F (bzw. Befehlsausführung) nimmt der Hypothese nach an Wichtigkeit ab (Flynn & Weiss, 2007). Zusammen mit dem sich selbst verstärkenden Modell der sozialen Multiplikatoren könnte das Konzept den starken Anstieg in der Aufgabe GF womöglich erklären (Dickens & Flynn, 2001).

Die generelle Abwesenheit eines klaren Trendunterschieds in der Stärke des Effektes zwischen den fluiden und kristallinen Untertests wird in der Literatur mit der Überlappung der Intelligenzkonstrukte in den einzelnen Aufgaben der Wechsler Intelligenztests erklärt. Hierzu wurde GF allerdings zusammen mit RD zu den fluiden Intelligenzaufgaben und AW zu den kristallinen Aufgaben gezählt (Truscott & Frank, 2001). Auch in anderen Studien wurde argumentiert, dass sich der Effekt für GF so deutlich von den anderen Aufgaben des Verbalteils (und insgesamt) abhebt, da er in großem Maße logisches Denken erfasst (Trahan et al., 2014). Damit wurde auch hier die gerade beschriebene Hypothese von Flynn (2007) unterstützt. GF scheint in jedem Fall eine Ausnahme darzustellen. Im Folgenden wird die Zuordnung der Untertests zu einzelnen Intelligenzdomänen mit bisheriger Literatur verglichen.

Untertestzuordnung zu den Intelligenzdomänen

In anderen Studien wich die Zuordnung der gewählten Untertests RD und GF zu den einzelnen Intelligenzdomänen teilweise von der hier auf Grund der Testinhalte und Manualangaben getroffenen Einteilung ab.

Der indirekte Hinweis auf den positiven Flynn-Effekt in RD aus der vorliegenden Untersuchung sollte Informationen über die Entwicklung der fluiden Intelligenz geben. Auf Grund der sprachlichen Komponente wurde der Untertest aber auch schon zu den kristallinen Aufgaben gezählt (Satzger et al., 1996). Andere Studien

ordnen ihn wiederum auch der fluiden Intelligenzdomäne zu (Trahan et al., 2014; Truscott & Frank, 2001). Laut Manual sollte der Untertest das Arbeitsgedächtnis abprüfen. Folglich würde der hier gefundene positive Flynn-Effekt für RD einen gegenläufigen Trend zu den Ergebnissen einer anderen aktuellen Studie abbilden. Dort wurde ein negativer Flynn-Effekt für die Bereiche der Aufmerksamkeitsspanne und des Arbeitsgedächtnisses dokumentiert (Graves et al., 2019). Auch die Analyse einer metaanalytischen Untersuchung ließ eher auf eine negative Leistungsentwicklung im Bereich des verbalen Arbeitsgedächtnisses schließen (Wongupparaj et al., 2017)

Diese eindimensionale Zuordnung zum Arbeitsgedächtnis wird dem Untertest womöglich aber nicht gerecht. Stattdessen scheint das Ergebnis aus dieser Aufgabe vielmehr ein Resultat aus mehreren Fähigkeitskomponenten zu sein. Neben dem Arbeitsgedächtnis scheint fluides Schlussfolgern, kristalline Intelligenz sowie g erfasst zu werden (Niileksela & Reynolds, 2019; Weiss et al., 2013). Diese faktorenanalytisch mehrfach gemessene Mehrdimensionalität des Untertests könnte auch die Ähnlichkeit des Flynn-Effekt in RD mit der Stärke des Gesamt-IQ Anstiegs mit den bereits genannten Überblicksstudien erklären. Mit 2.8 IQ-Punkten pro Dekade entspricht er eher den dort angegebenen Flynn-Effekten für den Gesamt-IQ von ca. 3 IQ-Punkten pro Jahrzehnt als den dort angegebenen fluiden IQ-Zuwächsen (Flynn, 1984; Pietschnig & Voracek, 2015; Trahan et al., 2014). Der Anstieg für RD könnte sich durch dämpfende Einflüsse der verbalen und Arbeitsgedächtniskomponente und verstärkende Einflüsse der fluiden Intelligenzanteile mittig eingependelt haben. Dies wäre ähnlich zum Prozess bei der Summierung der einzelnen Untertestleistungen zum Gesamt-IQ. Der zumeist stärker ausfallende Flynn-Effekt für fluide Intelligenz (hier: das logische Erschließen des Rechenwegs) könnte durch einen schwachen Anstieg der Leistung des Arbeitsgedächtnisses (hier: das Abspeichern der auditiven Informationen) gesenkt worden sein.

Die augenscheinliche Beobachtung während der Testdurchführung würde diese Interpretation unterstützen: Viele Teilnehmende baten um mehrfache Wiederholung der Frage und äußerten den Wunsch, sich Notizen machen zu können. Dies könnte für eine geringere Arbeitsgedächtniskapazität sprechen. Der Zeitverlust wirkt

sich außerdem negativ auf die Punktzahl aus. Eine zusätzliche negative Verzerrung durch den verbalen Anteil der Aufgabe (hier: die Verpackung der Rechnung in eine Textaufgabe) kann nicht ausgeschlossen werden. Die Interpretation von RD als fluiden Flynn-Effekt, könnte daher womöglich um die genannten Intelligenzaspekte erweitert werden.

Der bereits diskutierte stark positive Flynn-Effekt in GF wurde in der aktuellen Untersuchung als Hinweis für die IQ-Entwicklung in einer Kombination an fluiden und kristallinen Intelligenzanteilen herangezogen. Die kristalline Komponente ist, wie auch für RD, durch den verbalen Anteil der Aufgabe unstrittig. In der Literatur wird er jedoch, im Gegensatz zur Manualangabe (Skala Sprachliches Verständnis) wie bereits erwähnt, auch der fluiden Intelligenz zugeordnet (Satzger et al., 1996; Trahan et al., 2014; Truscott & Frank, 2001). Dies wird mit dem abstrakten Charakter der Aufgabe erklärt (Trahan et al., 2014; Weiss et al., 2016). Demzufolge scheint GF deutlich mehr fluide Intelligenzkomponenten zu erfassen als für die vorliegende Erhebung angenommen wurde. Dies würde sich mit der zuvor beschriebenen Erklärungshypothese der differenzierten Fähigkeitsspezialisierung im Bereich des abstrakten Schlussfolgerns decken. Die angeführten explorativen Faktorenanalysen, die die Mehrdimensionalität von RD darlegten, geben hingegen keine empirische Evidenz für eine fluide Komponente in GF. In ihren Modellen wurde RD stets eindimensional zu der im Manual angegebenen Skala des sprachlichen Verständnisses geordnet (Niileksela & Reynolds, 2019; Weiss et al., 2013).

Insgesamt legt der Literaturabgleich der getroffenen Zuordnung der drei Untertests sowohl Übereinstimmungen als auch Abweichungen nahe. So wurde AW durchgehend als gutes Maß für kristalline Intelligenz eingeschätzt, während aus anderen Studien Evidenz für die Mehrdimensionalität des Untertests RD vorliegt. Welche Intelligenzdomäne nun genau zu welchem Anteil gemessen wird, variiert je nach Untersuchung. Neben GF könnte daher auch RD eine Art Mischform abbilden. Die Zuordnung von GF bleibt nach wie vor widersprüchlich. Auf Grundlage der Manualangaben und EFA-Ergebnissen wird dessen kristalline Komponente aber weiterhin als plausibler eingeschätzt. Die abstrakte Komponente des Untertests ist dem Aufbau der Aufgabe nach jedoch auch nicht von der Hand zu weisen. Damit bleibt

GF eine Sonderrolle, wie bereits Wechsler (1939, zitiert nach Tewes, 1994) selbst resümierte. Da die dargestellten Hypothesen und Evidenzen dazu in der vorliegenden Arbeit nicht empirisch überprüft wurden, wird die Interpretation der Untertests hier wie präregistriert beibehalten.

Multikausales Zusammenwirken verschiedener Erklärungsmodelle

Die Komplexität des Flynn-Effekts führt dazu, dass ein Erklärungsansatz bisher nie genug Erklärungswert aufbringen konnte, um das gesamte Phänomen schlüssig abzubilden. Die meisten Erklärungen erscheinen nur unter bestimmten Bedingungen plausibel. Oft können einige Aspekte durch einen Ansatz erklärt werden, andere bleiben aber außen vor (wie z.B. beim Faktor Bildung). Wieder andere Hypothesen sind schwer empirisch prüfbar. Dazu zählt die Hypothese der sozialen Multiplikatoren. Außerdem fällt die Aussagekraft teilweise von Studie zu Studie unterschiedlich aus. Es könnte daher geschlussfolgert werden, dass die Inkonsistenz des Flynn-Effekts über einzelne Erhebungen hinweg die meist belegte Annahme über das Phänomen darstellt (Williams, 2013). Dass der Einfluss obsoleter Items in dieser Studie nicht konzeptuell replizierbar war, stellt wieder ein Beispiel mehr hierfür dar. Dieser bisher wenig untersuchte Ansatz ist einer von vielen, der versucht, Anteile des Effekts bzw. dessen domänenspezifischen Ausprägungsbesonderheiten zu beschreiben.

Lange schien eine kausale Erklärung generell schwer greifbar zu sein. Das grundlegende Verständnis des Phänomens fehlte. Noch 1999 kritisierte Rodgers, dass der Ursprung des Flynn-Effekts in der Wissenschaft zu schnell als gegeben hingenommen wurde, obwohl die potenziellen Erklärungen zu wenig durch systematische Analysen des Effektverlaufes untermauert wurden. Ihm zufolge wäre eine Erklärungshypothese erst dann wirklich aussagekräftig (Rodgers, 1999). Ähnliche Kritik wurde von Seiten der American Psychological Association laut. Sie betonte die Wichtigkeit der genaueren Analyse und Ursachenforschung (Neisser et al., 1996).

In den letzten zwanzig Jahren ist das Forschungsfeld zum Flynn-Effekt jedoch massiv gewachsen und der Grad an Expertise hat stetig zugenommen. Aus der ganzen Welt wurden spezifische Analysen veröffentlicht. Großprojekte wie die

Metaanalyse von Pietschnig und Voracek (2015) versuchten die Fülle an Einzelkenntnissen zu bündeln. Diese Arbeit hat weiter zur systematischen Bewertbarkeit der möglichen Erklärungsansätze beigetragen. Ihre Schlussfolgerung ist, dass vor allem hybride Faktoren, die mit dem Modell zur langsameren lebensgeschichtlichen Entwicklung (Woodley, 2012) in Einklang stehen, genug Evidenz aufweisen, um zum Flynn-Effekt beizutragen. Dazu gehören die bereits beschriebenen Aspekte der besseren und längeren Bildung, die gesündere Ernährung und sowie die Abnahme pathogenen Stresses. Auch andere führende Forschungsmitglieder sind sich einig, dass die Hauptgründe für den IQ-Anstieg über die Zeit vor allem in Umweltfaktoren wie besserer Gesundheit, Bildung und Ernährung sowie dem gestiegenen Lebensstandard liegen (Rindermann et al., 2017).

Über die unterschiedlich starken Zuwachsraten je nach Intelligenzdomäne gibt es noch weniger Konsens. Wie bereits erläutert, könnten diese am ehesten durch den Effekt sozialer Multiplikatoren sowie eines allgemeinen wirtschaftlichen Wohlstands mitbestimmt werden. Die allmähliche Stagnation des Leistungszuwachses könnte durch Deckeneffekte und die generelle Abnahme der Erträge der Faktoren, die für den positiven Flynn-Effekt verantwortlich sind, zustande kommen (Pietschnig & Voracek, 2015). Die Abnahmen und die Umkehr werden hingegen vor allem mit dem negativen Zusammenhang mit dem g-Faktor in Verbindung gebracht. Demnach könnten die Abnahmen der generellen Intelligenzkomponente (g) bisher von Zunahmen spezieller kognitiver Fähigkeiten verdeckt worden sein. Da diese aber ein Maß erreicht haben, an dem weitere Steigerung nur noch schwer möglich ist, wird der Rückgang in psychometrischem g durch einen stagnierenden oder sogar negativen Flynn-Effekt sichtbar (Pietschnig et al., 2018; Pietschnig & Voracek, 2015).

Die hier gefundenen domänenspezifischen Unterschiede auf Untertestebene könnten diese Systematik abbilden. Die starke Zunahme in GF könnte einer Spezialisierung der Fähigkeit zum abstrahierenden Denken (Flynn, 2007) zu Grunde liegen, deren Anstieg durch den Effekt der sozialen Multiplikatoren (Dickens & Flynn, 2001) deutlich verstärkt wurde. Während die Leistungen in Untertests wie AW und RD keinem solchen Verstärkungseffekt unterlagen. Bei der Berechnung des

Gesamt-IQs pendeln sich die einzelnen Extremwerte mit den milderen Anstiegen dann auf einem moderaten IQ-Zuwachs ein.

Inhaltliche Äquivalenz der Untertests angenommen

Aus den aktuellen Studienergebnissen geht hervor, dass die A-Items, deren Formulierung zwischen den Versionen leicht abwich, als inhaltlich gleichwertig gelten können. Es hatte demnach keine signifikante Auswirkung auf die Leistung, ob eine Person den Wortlaut aus dem HAWIE-R oder WIE vorgelesen bekam. Bereits vor der Durchführung der vorliegenden Studie wurde die Prüfung der inhaltlichen Äquivalenz der verschiedenen Untertestversionen als notwendig befunden. Denn nur wenn diese Voraussetzung erfüllt ist, können die Differenzen zwischen den Testversionen als unterschiedlich starke Leistung angesehen werden. Andernfalls könnten Einflüsse geänderter Testformate die eigentlich gesuchten Effekte, wie zuvor beschrieben, maskieren (Pietschnig & Voracek, 2015; Weiss et al., 2016). Es kann daher angenommen werden, dass die indirekten Hinweise auf domänenspezifische Flynn-Effekte in diesem Experiment nicht durch veränderte Fragenformulierungen zwischen den Versionen beeinflusst wurde.

Die Interkorrelationen auf Ebene der prozentualen kognitiven Leistung in den Untertests wurden bei allen Aufgaben als zufriedenstellend bewertet. Mit Korrelationskoeffizienten zwischen .73 und .88 kann von einer hohen Konstruktvalidität der beiden Testversionen ausgegangen werden. Ähnliche Werte wurden auch in den anderen beiden Studien mit deutschsprachigen Wechsler Testversionen angegeben. Dort variierten die Korrelationskoeffizienten pro Untertest zwischen .81 und .83 (Satzger et al., 1996) sowie .85 und .86 (Erzberger & Engel, 2010). Dieser Befund spricht für eine gute a priori getroffene Einschätzung der ähnlichen Items als inhaltlich äquivalent und die gültige Interpretierbarkeit der berechneten Ergebnisse. Die inhaltliche Äquivalenz kann angenommen werden. Mit hoher Wahrscheinlichkeit messen die gleichnamigen Untertests beider Testversionen das gleiche Konstrukt. Die zuvor diskutierten, augenscheinlich auffälligen Items im HAWIE-R haben die Vergleichbarkeit folglich nicht gemindert.

Beobachtung von Testrateverhalten

Vielleicht führte die freiwillige Teilnahme der Stichprobe an der Intelligenztestung zu einem generell sorgloseren Umgang mit Antworten, da keine negative Konsequenz zu erwarten war. Womöglich könnte aber auch ein systematisch verändertes Verhalten in Testsituationen sichtbar geworden sein. Dies würde zu einer bereits dargestellten Erklärungshypothese passen, die von einem verstärkenden Effekt geänderten Testrateverhaltens auf den Flynn-Effekt ausgeht (Must & Must, 2013; Pietschnig et al., 2013). Jedoch wird dort eher von Rateverhalten in Multiple-Choice Aufgaben ausgegangen und der Einfluss auf den Flynn-Effekt generell angezweifelt bzw. eine bereits ausgeschöpfte positive Wirkung angenommen. Es muss außerdem angemerkt werden, dass die statistische Prüfung dieser These mit den aktuellen Daten nicht möglich war. Es handelt sich hierbei um mit den Beobachtungen konsistente Ergebnisse, die aber nicht kausal als Hypothesentest integriert werden konnten.

Limitation

Der Vergleich der Wertpunkttabellen zwischen dem ersten deutschsprachigen Wechsler Intelligenztest HAWIE und dem Nachfolger HAWIE-R stellte noch kein Problem dar (Satzger et al., 1996). Mit der dritten revidierten Testversion (WIE) war dies auf Grund bestimmter Testcharakteristika in der vorliegenden Studie jedoch nicht mehr möglich. Tatsächlich wurde der Wertpunktvergleich zwischen HAWIE-R und WIE schon in einer älteren Studie (Erzberger & Engel, 2010) als problematisch bewertet. Dort wurde vermerkt, dass die originalen, ausschließlich leistungsorientierten Wertpunktskalen des HAWIE-R nicht mit denen des WIE vergleichbar waren. Im Zuge der Neuauflage des Testinstruments wurden die Wertpunkte diesmal altersabhängig bestimmt. Infolgedessen berechneten Erzberger und Engel (2010) in ihrer Untersuchung Wertpunkte ohne Altersbezug aus einer Teilstichprobe. Diese wurden dort zur weiteren Berechnung verwendet. Zusätzlich wurde wie auch bei Satzger et al. (1996) eine Umrechnungsskala für die Praxis erstellt. Mit Hilfe dieser Skala können IQ-Werte aus der älteren Testversion (z.B. aus Vortestungen) in entsprechende IQ-Werte der revidierten Version transformiert werden. Dadurch wird die Intelligenz einer Person über verschiedene

Normierungszeitpunkte hinweg vergleichbar, ohne dass der Flynn-Effekt zu verzerrten Werten führt. Da hierbei die Normverschiebung dementsprechend schon ausgeglichen wurde, wies die Skala keinen Nutzen für die aktuelle Datenanalyse auf. In dieser Studie wurde die beschriebene Problematik durch die explorative Analyse auf Ebene relativer Häufigkeiten gelöst. Dieses Vorgehen wich von der Präregistrierung ab, stellte aber die Vergleichbarkeit der Leistungen wieder her.

Die erste Hypothese zur signifikanten Differenz der durchschnittlichen IQ-Werte zwischen den Testversionen hätte eigentlich Aufschluss über einen vorliegenden Flynn-Effekt geben können. Sie konnte auf Grund der erläuterten neuartigen Wertpunktberechnung des WIE aber so nicht geprüft werden. Durch das explorative Vorgehen konnten jedoch keine durchschnittlichen IQ-Werte mehr verglichen werden. Anders als in der Pilotstudie von Höttl (2019) konnte aus der Berechnung einer ANOVA in dieser Studie kein Flynn-Effekt abgeleitet werden. Die Interpretation der explorativ untersuchten Haupt- und Interaktionseffekte wäre nur unter der Annahme gleicher Itemschwierigkeiten zwischen den beiden Testversionen gültig gewesen. Dies widerlegte jedoch die bereits beschriebene signifikante dritte Hypothese. Folglich waren die ersten beiden Hypothesen nicht interpretierbar. Die Richtungen der Mittelwertdifferenzen hätten ansonsten auf einen kleinen negativen Effekt im Untertest GF, einen kleinen positiven Effekt für RD sowie keinem Flynn-Effekt für den kristallinen Untertest AW gesprochen. Stattdessen wurde indirekt (Vergleich der N-Items) auf den Flynn-Effekt geschlossen. Die Transformation der gemessenen Effektgrößen machte einen Vergleich auf IQ-Ebene möglich, wenn auch auf anderem Weg als in der Pilotstudie und Präregistrierung.

Stichprobenspezifika

Der Literatur nach wird bei Stichproben höheren Leistungsniveaus ein eher kleinerer Flynn-Effekt gemessen. Die stärksten Effekte werden vor allem aus dem niedrigeren bis mittleren Intelligenzspektrum berichtet (Colom et al., 2005). Im Vergleich mit den Ergebnissen erwähnter anderer Wechsler Intelligenztest Studien auf Untertestebene fallen die Flynn-Effekte in dieser Untersuchung jedoch durchaus eher groß aus. Da aus der Literatur hervor geht, dass sich der Flynn-Effekt bei Erwachsenen meist am stärksten auswirkt (Pietschnig & Voracek, 2015), könnten die

tendenziell größeren Leistungsdifferenzen auch am Alter der Stichprobe (18-30) liegen. Die dargestellten Vergleichswerte kamen zumeist aus Untersuchungen im Kindesalter.

Mit der Matura (bzw. dem Abitur) als Mindestvoraussetzung wurde eine bereits hohe intellektuelle Grundfähigkeit erwartet. Dies wurde aufgrund des studentischen Umfelds am praktikabelsten bewertet. Mit einer hoch gebildeten Stichprobe geht jedoch die Gefahr von Deckeneffekten einher. Sind die Items einer Aufgabe zu einfach für die Teilnehmenden, kann das Messinstrument die wahren Intelligenzunterschiede nicht mehr genau messen. Werden die meisten Aufgaben von fast allen richtig gelöst, werden die Differenzen zu gering. Es entsteht kein Effekt. Während die Spannweite der Untertests AW und RD relativ breit ausfällt, lässt sich für GF eine deutlich geringere Differenz der Gesamtscores erkennen (vgl. Tabelle 1). Die relativen Häufigkeiten erreichen aber keinen auffällig hohen Wert (kaum volle Punktzahlen) der Grund zur Annahme für Deckeneffekte geben würde. Vielmehr könnte die Einschränkung des Wertebereichs (auch: restriction of range) durch die Homogenität der Stichprobe die Interkorrelationen zwischen den Untertests der beiden Versionen etwas gemindert haben. Womöglich wären diese in einer heterogenen Stichprobe noch größer ausgefallen. Gleichzeitig läge mit zunehmender Heterogenität der Stichprobe aber wieder das Risiko unerwünschter konfundierender Variablen höher.

Auswertungsobjektivität

Bei der Auswertung einiger verbaler Untertests bleibt der durchführenden Testleitung ein gewisser Ermessensspielraum. Äquivalent richtige Antworten, die nicht in den Bewertungsrichtlinien genannt werden, dürfen zugelassen werden (Tewes, 1994; Von Aster et al., 2006). Das führt zu einem erhöhten Einfluss des Durchführenden und kann die Auswertungsobjektivität beeinträchtigen (Molz et al., 2010). Während die Beurteilung in RD keine Interpretation zulässt, trifft die dargestellte Problematik auf die beiden anderen Untertests zu. Durch Unerfahrenheit mit der Auswertung kann es zu Verzerrung kommen. Die Durchführung wurde hier aber, wie im Manual empfohlen, von psychologischem Fachpersonal mit praktischen Vorkenntnissen in der Auswertung von Wechsler Intelligenztests

vorgenommen. Vorbeugend wurde für die äquivalent zugelassenen Antworten im Untertest GF eine Tabelle mit synonymen Lösungen und entsprechender Punktevergabe angelegt. Damit wurde sichergestellt, dass in allen 85 Auswertungsbögen die gleiche Punktzahl für die gleiche, als äquivalent eingeschätzte Lösung vergeben wurde. Eine Verzerrung durch mangelnde Auswertungsobjektivität wird daher als unwahrscheinlich eingeschätzt.

Conclusio

Abschließend werden die Ergebnisse sowie deren Bedeutung resümiert. Die konzeptuelle Replikation der Untersuchung von Höttl (2019) konnte keinen Einfluss obsoleter Items auf den kristallinen Flynn-Effekt in den Wechsler Intelligenztestversionen HAWIE-R und WIE finden. Dieser Befund trägt dazu bei, die Anzahl an Replikationen in der psychologischen Forschung zu erhöhen. Dadurch kann die Aussagekraft der bisher einmaligen Beobachtung besser eingeschätzt werden. Da der Flynn-Effekt in der Literatur bereits viele unterschiedliche Facetten gezeigt hat, wären weitere konzeptuelle Replikationen zu dieser Hypothese von großem Nutzen. Weil die Aufgabeninhalte vor allem in kristallinen Intelligenztests sehr unterschiedlich ausfallen können, lässt sich dieser Fund sonst schwer generalisieren. In einem anderen Intelligenztest oder einer anderen Wechsler Testversion kann eine Maskierung des kristallinen Flynn-Effekts durch veraltete Inhalte nach wie vor nicht ausgeschlossen werden. Dazu interessant wäre der Vergleich zwischen der Originalfassung des IST und des IST-70, also mit einer älteren Testversion als in der Studie von Höttl (2019). Damit würde erkennbar, ob der Effekt über alle IST-Versionen hinweg verläuft oder nur versionsspezifisch ist. Auch die Analyse weiterer Wechsler Intelligenztestversionen wäre von Wert. Daraus könnte abgeleitet werden, ob die Hypothese in anderen Wechsler Testversionen repliziert werden kann oder ob der Einfluss bei dieser Art Intelligenztest generell nicht auftritt.

Die Daten zeigten außerdem indirekt Evidenz für einen domänenspezifischen Flynn-Effekt. Die Untertests waren konzeptuell so ausgewählt worden, dass ihre Ergebnisse Rückschlüsse auf fluide und kristalline IQ-Veränderungen geben konnten. Der Untertest AW gab Hinweis auf eine verhältnismäßig große kristalline Leistungssteigerung, die kaum von der Steigung im fluiden Untertest RD abwich. Dieses

Ergebnis wurde bisher nur selten erfasst. Speziell aus dem deutschsprachigen Raum liegt jedoch Evidenz dazu vor (Pietschnig et al., 2010). Für einen fluiden Flynn-Effekt fiel RD eher moderat aus. Dies erwies sich auf Untertestebene jedoch konsistent mit anderen Studienergebnissen zu RD. Ebenso deckt sich der starke Flynn-Effekt für die Mischform GF in seinem generellen Trend, einen Extremwert unter den Untertests der Wechsler Intelligenztestbatterien darzustellen, mit anderen Studien. Es wird argumentiert, dass sich darin eine zunehmende Spezialisierung einzelner kognitiver Fähigkeiten abbilden könnte (Woodley & Madison, 2013). Ob der Untertest nun mehr fluide oder kristalline Intelligenz erfasst, konnte in dieser Studie nicht überprüft werden. Dazu sowie zur dokumentierten Mehrdimensionalität von RD wäre weitere Forschung interessant.

Für die zukünftige Entwicklung des Flynn-Effekts prognostizieren Expert*innen vor allem weitere Anstiege in entwicklungsschwächeren Ländern und voranschreitende Stagnationen in westlichen Ländern (Rindermann et al., 2017). Die Daten aus dieser Studie zeichnen dieses Bild jedoch noch nicht ab. Vielmehr ähnelt der positive Flynn-Effekt in allen drei Untertests den nach wie vor überwiegenden positiven Funden weltweit (Dutton et al., 2016; Flynn, 1987; Pietschnig & Voracek, 2015; Trahan et al., 2014).

Da die Frage nach der Ursache der domänenspezifischen Unterschiede weiter unklar bleibt, sollte sich der Fokus zukünftiger Arbeiten weg von reiner Exploration hin zu mehr quantitativ hochwertiger Überprüfung bisheriger Erklärungsansätze bewegen. Mit mehr Replikationen und Metaanalysen könnte die Qualitätssicherung der differentiellen Intelligenzforschung vorangetrieben werden (Pietschnig et al., 2019). Dadurch könnte die Fülle an unterschiedlichen Funden zunehmend auf die Ergebnisse mit der meisten Evidenz reduziert und die Frage nach der Ursache des Flynn-Effekts eindeutig beantwortet werden. Dies würde die Ableitung zielführender Handlungsansätze zur Prävention zukünftiger IQ Über- bzw. Unterschätzungen in der Praxis und der damit verbundenen Auswirkungen für die Individuen ermöglichen.

Literaturverzeichnis

- Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology*, 24, 383–405. <https://doi.org/10.1076/jcen.24.3.383.981>
- Carlton, M., & Sapp, G. L. (1997). Comparison of WISC-R and WISC-III scores of urban exceptional students. *Psychological Reports*, 80, 755–760.
- Cocodia, E. A., Kim, J. S., Shin, H. S., Kim, J. W., Ee, J., Wee, M. S. W., & Howard, R. W. (2003). Evidence that rising population intelligence is impacting in formal education. *Personality and Individual Differences*, 35, 797–810. [https://doi.org/10.1016/S0191-8869\(02\)00284-2](https://doi.org/10.1016/S0191-8869(02)00284-2)
- Cohen, J. (1988). *Standard power analysis for the behavioral sciences*. Hillsdale NJ Lawrence Erlbaum.
- Colom, R., Abad, F. J., García, L. F., & Juan-Espinosa, M. (2002). Education, Wechsler's full scale IQ, and g. *Intelligence*, 30, 449–462. [https://doi.org/10.1016/S0160-2896\(02\)00122-8](https://doi.org/10.1016/S0160-2896(02)00122-8)
- Colom, R., Flores-Mendoza, C. E., & Abad, F. J. (2007). Generational changes on the Draw-a-Man test: A comparison of Brazilian urban and rural children tested in 1930, 2002 and 2004. *Journal of Biosocial Science*, 39, 79–89. <https://doi.org/10.1017/S0021932005001173>
- Colom, R., Lluís-Font, J. M., & Andrés-Pueyo, A. (2005). The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: Supporting evidence for the nutrition hypothesis. *Intelligence*, 33, 83–91. <https://doi.org/10.1016/j.intell.2004.07.010>
- Daley, T. C., Whaley, S. E., Sigman, M. D., Espinosa, M. P., & Neumann, C. (2003). IQ on the rise: The Flynn effect in rural Kenyan children. *Psychological Science*, 14, 215–219. <https://doi.org/doi:10.1111/1467-9280.02434>
- Daseking, M., Petermann, U., & Petermann, F. (2007). Intelligenzdiagnostik mit dem HAWIK-IV. *Kindheit und Entwicklung*, 16, 250–259. <https://doi.org/10.1026/0942-5403.16.4.250>
- Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, 108, 346–369. <https://doi.org/10.1037//0033-295X.108.2.346>
- Dickinson, M. D., & Hiscock, M. (2010). Age-related IQ decline is reduced markedly

- after adjustment for the Flynn effect. *Journal of Clinical and Experimental Neuropsychology*, 32, 865–870. <https://doi.org/10.1080/13803391003596413>
- Dumont, R., & Faro, C. (1993). A WISC-III short form for learning-disabled students. *Psychology in the Schools*, 30, 212–219. [https://doi.org/10.1002/1520-6807\(199307\)30:3%3C212::AID-PITS2310300303%3E3.0.CO;2-V](https://doi.org/10.1002/1520-6807(199307)30:3%3C212::AID-PITS2310300303%3E3.0.CO;2-V)
- Dutton, E., Bakhiet, S. F., Essa, Y. A. S., Blahmar, T. A., & Hakami, S. M. A. (2017). A Negative Flynn effect in Kuwait: The same effect as in Europe but with seemingly different causes. *Personality and Individual Differences*, 114, 69–72. <https://doi.org/10.1016/j.paid.2017.03.060>
- Dutton, E., & Lynn, R. (2013). A negative Flynn effect in Finland, 1997-2009. *Intelligence*, 41, 817–820. <https://doi.org/10.1016/j.intell.2013.05.008>
- Dutton, E., & Lynn, R. (2015). A negative Flynn effect in France, 1999 to 2008-9. *Intelligence*, 51, 67–70. <https://doi.org/10.1016/j.intell.2015.05.005>
- Dutton, E., van der Linden, D., & Lynn, R. (2016). The negative Flynn effect: A systematic literature review. *Intelligence*, 59, 163–169. <https://doi.org/10.1016/j.intell.2016.10.002>
- Eppig, C., Fincher, C. L., & Thornhill, R. (2010). Parasite prevalence and the worldwide distribution of cognitive ability. *Proceedings of the Royal Society B: Biological Sciences*, 277, 3801–3808. <https://doi.org/10.1098/rspb.2010.0973>
- Erzberger, C. S., & Engel, R. R. (2010). Zur Äquivalenz der Normen des Wechsler-Intelligenztests für Erwachsene (WIE) mit denen des Hamburg-Wechsler-Intelligenztests für Erwachsene - Revision (HAWIE-R). *Zeitschrift für Neuropsychologie*, 21, 25–37. <https://doi.org/10.1024/1016-264X/a000002>
- Fitzgerald, S., Gray, N. S., & Snowden, R. J. (2007). A comparison of WAIS-R and WAIS-III in the lower IQ range: Implications for learning disability diagnosis. *Journal of Applied Research in Intellectual Disabilities*, 20, 323–330. <https://doi.org/10.1111/j.1468-3148.2006.00349.x>
- Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, 95, 29–51. <https://doi.org/10.1037/0033-2909.95.1.29>
- Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, 101, 171–191. <https://doi.org/10.1037/0033-2909.101.2.171>
- Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time.

- American Psychologist*, 54, 5–20. <https://doi.org/10.1037/0003-066X.54.1.5>
- Flynn, J. R. (2007). *What is intelligence? Beyond the Flynn effect*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511605253.003>
- Flynn, J. R. (2009). The WAIS-III and WAIS-IV: Daubert motions favor the certainly false over the approximately true. *Applied Neuropsychology*, 16, 98–104. <https://doi.org/10.1080/09084280902864360>
- Flynn, J. R., & Weiss, L. G. (2007). American IQ gains from 1932 to 2002 : The WISC subtests and educational progress. *International Journal of Testing*, 7, 209–224. <https://doi.org/10.1080/15305050701193587>
- Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography. *Intelligence*, 24, 13–23. <https://doi.org/10.1111/j.1432-1033.1976.tb11027.x>
- Graves, L. V., Drozdick, L., Courville, T., Farrer, T. J., Gilbert, P. E., & Delis, D. C. (2019). Cohort differences on the CVLT-II and CVLT3: Evidence of a negative Flynn effect on the attention/working memory and learning trials. *The Clinical Neuropsychologist*, 1–18. <https://doi.org/10.1080/13854046.2019.1699605>
- Gunter, C. M., Sapp, G. L., & Green, A. C. (1995). Comparison of scores on WISC-III and WISC-R of urban learning disabled students. *Psychological Reports*, 77, 473–474. <https://doi.org/10.2466/pr0.1995.77.2.473>
- Hötl, F. (2019). *Einflüsse von Itemobsoleszenz auf generationsspezifische Testnormverschiebungen: eine querschnittliche Untersuchung des Flynn Effekts*. Unveröff. Masterarbeit, Universität Wien.
- Horn-Alsberge, M. M. (1999). *Stability of WISC-R and WISC-III IQs and subtest scores*. Doctoral diss., St. John's University (New York).
- Horn, J. L. (1965). *Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities*. Doctoral diss., University of Illinois.
- Howard, R. W. (2005). Objective evidence of rising population ability: a detailed examination of longitudinal chess data. *Personality and Individual Differences*, 38, 347–363. <https://doi.org/10.1016/j.paid.2004.04.013>
- Johnson, W., Bouchard, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: consistent results from three test batteries. *Intelligence*, 32, 95–107. [https://doi.org/10.1016/S0160-2896\(03\)00062-X](https://doi.org/10.1016/S0160-2896(03)00062-X)
- Khaleefa, O., Sulman, A., & Lynn, R. (2009). An increase of intelligence in Sudan,

- 1987-2007. *Journal of Biosocial Science*, 41, 279–283.
<https://doi.org/10.1017/S0021932008003180>
- Laciga, J., & Cígler, H. (2017). The Flynn effect in the Czech Republic. *Intelligence*, 61, 7–10. <https://doi.org/10.1016/j.intell.2016.11.005>
- Liu, J., & Lynn, R. (2013). An increase of intelligence in China 1986-2012. *Intelligence*, 41, 479–481. <https://doi.org/10.1016/j.intell.2013.06.017>
- Lynn, R. (2009a). Fluid intelligence but not vocabulary has increased in Britain, 1979-2008. *Intelligence*, 37, 249–255.
<https://doi.org/10.1016/j.intell.2008.09.007>
- Lynn, R. (2009b). What has caused the Flynn effect? Secular increases in the development quotients of infants. *Intelligence*, 37, 16–24.
<https://doi.org/10.1016/j.intell.2008.07.008>
- Lynn, R. (2013). Who discovered the Flynn effect ? A review of early studies of the secular increase of intelligence. *Intelligence*, 41, 765–769.
<https://doi.org/10.1016/j.intell.2013.03.008>
- 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, 1–10. <https://doi.org/10.1016/j.intell.2008.08.004>
- Merrill, M. A. (1938). The significance of IQ's on the revised Stanford-Binet scales. *Journal of Educational Psychology*, 29, 641–651.
<https://doi.org/10.1037/h0071291>
- Mingroni, M. A. (2004). The secular rise in IQ: Giving heterosis a closer look. *Intelligence*, 32, 65–83. [https://doi.org/10.1016/S0160-2896\(03\)00058-8](https://doi.org/10.1016/S0160-2896(03)00058-8)
- Molz, C., Schulze, R., Schroeders, U., & Wilhelm, O. (2010). TBS- TK Rezension: „Wechsler Intelligenztest für Erwachsene WIE. Deutschsprachige Bearbeitung und Adaptation des WAIS-III von David Wechsler“. *Psychologische Rundschau*, 61, 229–230. <https://doi.org/10.1026/0033-3042/a000042>
- Must, O., & Must, A. (2013). Changes in test-taking patterns over time. *Intelligence*, 41, 780–790. <https://doi.org/10.1016/j.intell.2013.04.005>
- Must, O., Must, A., & Raudik, V. (2003). The secular rise in IQs : In Estonia, the Flynn effect is not a Jensen effect. *Intelligence*, 31, 461–471.
[https://doi.org/10.1016/S0160-2896\(03\)00013-8](https://doi.org/10.1016/S0160-2896(03)00013-8)
- Neisser, U. (1997). Rising scores on intelligence tests. *American Scientist*, 85, 440–

447. <https://www.jstor.org/stable/27856851>
- Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, 51, 77–101. <https://doi.org/10.1037/0003-066X.51.2.77>
- Niileksela, C. R., & Reynolds, M. R. (2019). Enduring the tests of age and time: Wechsler constructs across versions and revisions. *Intelligence*, 77, 101403. <https://doi.org/10.1016/j.intell.2019.101403>
- Nisbett, R. E., Aronson, J., Blair, C., Dickens, W. T., Flynn, J. R., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist*, 67, 130–159. <https://doi.org/10.1037/a0026699>
- Petermann, F., & Köller, O. (2008). Intelligenzdiagnostik. *Diagnostica*, 54, 171–173. <https://doi.org/10.1026/0012-1924.54.4.171>
- Petermann, F., & Wechsler, D. (2012). *Wechsler Adult Intelligence Scale - Fourth edition (WAIS-IV). Deutschsprachige Adaptation nach David Wechsler*. Person.
- Pietschnig, J., & Gittler, G. (2015). A reversal of the Flynn effect for spatial perception in German-speaking countries: Evidence from a cross-temporal IRT-based meta-analysis (1977–2014). *Intelligence*, 53, 145–153. <https://doi.org/10.1016/j.intell.2015.10.004>
- Pietschnig, J., & Gittler, G. (2017). Is ability-based emotional intelligence impervious to the Flynn effect? A cross-temporal meta-analysis (2001–2015). *Intelligence*, 61, 37–45. <https://doi.org/10.1016/j.intell.2016.12.006>
- Pietschnig, J., Siegel, M., Eder, J. S. N., & Gittler, G. (2019). Effect declines are systematic, strong, and ubiquitous: A meta-meta-analysis of the decline effect in intelligence research. *Frontiers in Psychology*, 10, 1–13. <https://doi.org/10.3389/fpsyg.2019.02874>
- Pietschnig, J., Tran, U. S., & Voracek, M. (2013). Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypothesis perhaps. *Intelligence*, 41, 791–801. <https://doi.org/10.1016/j.intell.2013.06.005>
- Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909–2013). *Perspectives on Psychological Science*, 10, 282–306. <https://doi.org/10.1177/1745691615577701>

- Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Pervasiveness of the IQ rise: A cross-temporal meta-analysis. *PLoS ONE*, 5, e14406.
<https://doi.org/10.1371/journal.pone.0014406>
- Pietschnig, J., Voracek, M., & Formann, A. K. (2011). Female Flynn effects: No sex differences in generational IQ gains. *Personality and Individual Differences*, 50, 759–762. <https://doi.org/10.1016/j.paid.2010.12.019>
- Pietschnig, J., Voracek, M., & Gittler, G. (2018). Is the Flynn effect related to migration? Meta-analytic evidence for correlates of stagnation and reversal of generational IQ test score changes. *Politische Psychologie*, 2, 267–283.
- Platt, J. M., Keyes, K. M., McLaughlin, K. A., & Kaufman, A. S. (2019). The Flynn effect for fluid IQ may not generalize to all ages or ability levels: A population-based study of 10,000 US adolescents. *Intelligence*, 77, 101385.
<https://doi.org/10.1016/j.intell.2019.101385>
- Quereschi, M. Y., & Erstad, D. (1990). A comparison of the WAIS and the WAIS-R for ages 61-91 years. *A Journal of Consulting and Clinical Psychology*, 2, 293–297.
<https://doi.org/10.1037/1040-3590.2.3.293>
- Rindermann, H., Becker, D., & Coyle, T. R. (2017). Survey of expert opinion on intelligence: The FLynn effect and the future of intelligence. *Personality and Individual Differences*, 106, 242–247. <https://doi.org/10.1016/j.paid.2016.10.061>
- Rodgers, J. L. (1999). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, 26, 337–356.
[https://doi.org/10.1016/S0160-2896\(99\)00004-5](https://doi.org/10.1016/S0160-2896(99)00004-5)
- Sanborn, K. J., Truscott, S. D., Phelps, L., & McDougal, J. L. (2003). Does the Flynn effect differ by IQ level in samples of students classified as learning disabled? *Journal of Psychoeducational Assessment*, 21, 145–159.
<https://doi.org/10.1177/073428290302100203>
- Satzger, W., Dragon, E., & Engel, R. R. (1996). Zur Normäquivalenz von HAWIE-R und HAWIE. *Diagnostica*, 43, 119–138.
- Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist*, 49, 304–313. <https://doi.org/10.1037//0003-066x.49.4.304>
- Schaie, K. W., & Strother, C. R. (1968). A cross-sequential study of age changes in cognitive behavior. *Psychological Bulletin*, 70, 671–680.
<https://doi.org/10.1037/h0026811>

- Scott, R., & Poncy, B. (1999). The „Flynn effect“: How applicable is it to longitudinal IQ assessment of american university students? *Mankind Quarterly*, 39, 385–397. <https://doi.org/10.46469/mq.1999.39.4.1>
- Simon, C. L., & Clopton, J. R. (1984). Comparison of WAIS and WAIS-R scores of mildly and moderately mentally retarded adults. *American Journal of Mental Deficiency*, 89, 301–303.
- Spearman, C. E. (1927). *The abilities of man*. MacMillan.
<https://doi.org/10.1038/120181a0>
- Te Nijenhuis, J., Cho, S. H., Murphy, R., & Lee, K. H. (2012). The Flynn effect in Korea: Large gains. *Personality and Individual Differences*, 53, 147–151.
<https://doi.org/10.1016/j.paid.2011.03.022>
- Te Nijenhuis, J., & van der Flier, H. (2013). Is the Flynn effect on g?: A meta-analysis. *Intelligence*, 41, 802–807. <https://doi.org/10.1016/j.intell.2013.03.001>
- Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. *Personality and Individual Differences*, 39, 837–843. <https://doi.org/10.1016/j.paid.2005.01.029>
- Tewes, U. (1994). *Hamburg Wechsler Intelligenztest für Erwachsene, Revision 1991 (HAWIE-R). Handbuch und Testanweisung* (2. Aufl.). Verlag Hans Huber.
- Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn effect: A meta-analysis. *Psychological bulletin*, 140, 1332–1360.
<https://doi.org/10.1037/a0037173>
- Truscott, S. D., & Frank, A. J. (2001). Does the Flynn effect affect IQ scores of students classified as LD? *Journal of School Psychology*, 39, 319–334.
[https://doi.org/10.1016/S0022-4405\(01\)00071-1](https://doi.org/10.1016/S0022-4405(01)00071-1)
- Vance, H., Maddux, C. D., Fuller, G. B., & Awadh, A. M. (1996). A longitudinal comparison of WISC-III and WISC-R scores of special education students. *Psychology in the Schools*, 33, 113–118. [https://doi.org/10.1002/\(sici\)1520-6807\(199604\)33:2<113::aid-pits3>3.0.co;2-s](https://doi.org/10.1002/(sici)1520-6807(199604)33:2<113::aid-pits3>3.0.co;2-s)
- Von Aster, M., Neubauer, A., & Horn, R. (2006). *Wechsler Intelligenztest für Erwachsene (WIE). Manual*. Harcourt Test Services.
- Weiss, L. G., Gregoire, J., & Zhu, J. (2016). Flaws in Flynn effect research with the Wechsler scales. *Journal of Psychoeducational Assessment*, 34, 411–420.
<https://doi.org/10.1177/0734282915621222>

- Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). WAIS-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment*, 31, 94–113.
<https://doi.org/10.1177/0734282913478030>
- Williams, R. L. (2013). Overview of the Flynn effect. *Intelligence*, 41, 753–764.
<https://doi.org/10.1016/j.intell.2013.04.010>
- Wongupparaj, P., Wongupparaj, R., Kumari, V., & Morris, R. G. (2017). The Flynn effect for verbal and visuospatial short-term and working memory: A cross-temporal meta-analysis. *Intelligence*, 64, 71–80.
<https://doi.org/10.1016/j.intell.2017.07.006>
- Woodley, M. A. (2012). A life history model of the Lynn-Flynn effect. *Personality and Individual Differences*, 53, 152–156. <https://doi.org/10.1016/j.paid.2011.03.028>
- Woodley, M. A., & Madison, G. (2013). Establishing an association between the Flynn effect and ability differentiation. *Personality and Individual Differences*, 55, 387–390. <https://doi.org/10.1016/j.paid.2013.03.016>
- Zajonc, R. B., & Mullen, P. R. (1997). Birth order: Reconciling conflicting effects. *American Psychologist*, 52, 685–699. <https://doi.org/10.1037/0003-066X.52.7.685>

Anhang

Zusammenfassung

Seit den 1930er Jahren werden IQ-Testnormverschiebungen bei der Intelligenzermessung festgestellt. Dieses Phänomen ist als *Flynn-Effekt* bekannt und umfasst global unterschiedlich starke nicht lineare IQ-Anstiege (zuletzt auch Stagnationen und Abnahmen) über Generationen und Zeit hinweg. Auch für verschiedene Intelligenzdomänen wurden unterschiedlich starke IQ-Entwicklungen erfasst. Der fluide IQ-Zuwachs fiel meist stärker aus als der kristalline. Die Ursache für den Flynn-Effekt ist noch nicht abschließend geklärt. Das Zusammenwirken hauptsächlich umweltbasierter Einflussfaktoren wird der Studienlage nach als am plausibelsten bewertet. Gründe für die domänenspezifischen Unterschiede sind ebenso noch wenig belegt. Daher hatte die vorliegende Studie zum Ziel, eine kürzlich aufgestellte Erklärungshypothese zum maskierenden Einfluss von Itemobsoleszenz auf den kristallinen Flynn-Effekt konzeptuell zu replizieren, sowie die IQ-Entwicklung im deutschsprachigen Raum querschnittlich auf einen domänenspezifischen Flynn-Effekt zu untersuchen. Dazu wurden der Stichprobe ($N = 85$) drei Untertests aus den *Wechsler Intelligenztestversionen HAWIE-R* (1988) und *WIE* (2002) vorgegeben (Normierungsabstand 14 Jahre). Die Ergebnisse auf Ebene relativer Häufigkeiten lieferten indirekte Hinweise auf einen positiven domänenspezifischen Flynn-Effekt. Im Untertest *Allgemeines Wissen* (AW) mit 2.7, bei *Rechnerisches Denken* (RD) mit 2.8 und *Gemeinsamkeiten Finden* (GF) mit 11.6 IQ-Punkten pro Dekade. Damit lieferte die Studie Evidenz für ähnlich starke IQ-Anstiege in der fluiden (RD) und kristallinen (AW) Aufgabe und widerspricht damit den viel belegten Unterschieden, ähnelt aber den Ergebnissen einer anderen deutschsprachigen Untersuchung. Im Vergleich zur Literatur fiel der fluide Flynn-Effekt moderat aus, der starke Flynn-Effekt der Mischform GF erwies sich auf Untertestebene als konsistent. Die erläuterte Erklärungshypothese konnte nicht konzeptuell repliziert werden.

Abstract

IQ-test norm changes in intelligence measures have been observed since the 1930s. This phenomenon is known as the *Flynn-effect* and includes globally varying degrees of non-linear IQ-increases (recently also stagnations and decreases) over generations and time. Different magnitudes of IQ-increases were also recorded for different intelligence-domains. The fluid IQ-increase was mostly stronger than the crystallized. The cause of the Flynn-effect is still not conclusively understood. The interaction of mainly environmental factors is considered to be the most plausible. Possible reasons for the domain-specific differences are also not yet proven. Therefore, the present study aimed to replicate conceptually a recently established hypothesis on the masking influence of Itemobsolescence on the crystallized Flynn-effect. As well as to examine the IQ-development in German-speaking countries cross-sectionally for a domain-specific Flynn-effect. For this purpose, the sample ($N = 85$) was given three subtests from the *Wechsler Adult Intelligence Scales HAWIE-R* (1988) and *WIE* (2002) (standardization interval: 14 years). The results on the level of relative frequencies provided indirect evidence for a positive domain-specific Flynn-effect. In the subtest *Information* (AW) with 2.7, in *Arithmetic* (RD) with 2.8 and *Similarities* (GF) with 11.6 IQ-points per decade. Thus, the study provided evidence for similarly strong IQ-increases in the fluid (RD) and crystallized (AW) domain, contradicting the much documented differences, but resembling the results of another German-language study. Relative to the literature, the fluid Flynn-effect was moderate, and the strong Flynn-effect of the mixed form GF proved consistent at the subtest level. The mentioned hypothesis could not be replicated.