BERICHTE

AUS DER

ABTEILUNG ANGEWANDTE PSYCHOLOGIE

Das Ausmass des "IQ-Gewinns"
im deutschen Sprachraum
von ca. 1956 bis ca. 1983.
Zur Diskussion um die HAWIK-R-Normen

Urs Schallberger

Nr. 27, 1991

UNIVERSITÄT ZÜRICH
PSYCHOLOGISCHES INSTITUT
Angewandte Psychologie
Universitätstrasse 84
CH-8006 Zürich

PSYCHOLOGISCHES INSTITUT DER UNIVERSITÄT ZÜRICH
[redacted]

Inhaltsübersicht

	Vorwort	S. 5
	Einleitung	S. 7
1.	"IQ-Gewinn": Begriffliches, Relevanz und Methodisches	S. 8
2.	Was lehren die HAWIK- und die HAWIK-R-Normen über den "IQ-Gewinn" im deutschen Sprachraum?	S. 12
	2.1 Vorblick	S. 12
	2.2 Untersuchungen zur Ueberhöhung der HAWIK-IQs	S. 13
	2.3 Vergleichsuntersuchungen von HAWIK und HAWIK-R	S. 15
	2.4 Alternativerklärungen zur Interpretation der IQ-Differenzen als "IQ-Gewinn"	S. 19
	2.5 Zusammenfassung und Diskussion	S. 22
3.	Einige intelligenzdiagnostische Folgeprobleme	S. 28
4.	Das Problem der Erklärung	S. 30
	Schlussbemerkungen	S. 34
	Literaturverzeichnis	S. 37
	Anhang: Zum Abschnitt "West Germany and Switzerland" in Flynn (1987)	S. 41

Zusammenfassung

Im Jahre 1983 erschien der HAWIK-R, eine revidierte Version des 1956 publizierten Hamburg-Wechsler-Intelligenztests für Kinder (HAWIK). Im Zentrum des Papiers steht die Frage, was die gegenwärtige Befundlage zu diesen beiden Testversionen über das Ausmass des "IQ-Gewinns" zwischen ca. 1956 und ca. 1983 zu lehren vermag. Ausgangspunkt bildet dabei eine Uebersicht über die Methoden zur Bestimmung solcher "IQ-Gewinne" (Leistungssteigerungen zwischen Geburtskohorten). Eine detaillierte Diskussion von drei Studien, die einen Rückschluss auf die Ueberhöhung der HAWIK-IQs zulassen, ferner von sechs Vergleichsuntersuchungen von HAWIK und HAWIK-R und schliesslich von Alternativinterpretationen legt die Vermutung nahe, dass mit einem "IQ-Gewinn" von 15 bis 20 Punkten zu rechnen ist. Dabei ergeben sich wesentliche Unterschiede zwischen Verbal- und Handlungsteil sowie zwischen den einzelnen Subtests. Ein Vergleich mit Ergebnissen aus andern Ländern zeigt, dass diese auf den ersten Blick erstaunliche Leistungszunahme nicht aussergewöhnlich ist. Abschliessend werden einige Konsequenzen für den Einsatz von Intelligenztests in Praxis und Forschung gezogen und mögliche Aspekte einer Erklärung diskutiert.

Vorwort

Der hier vorgelegte Text hat eine längere Vorgeschichte: Im Jahre 1985 verfasste ich ein Manuskript mit dem Titel *"HAWIK und HAWIK-R: Ein empirischer Vergleich. Ueber einen Fall phänomenaler Leistungsakzeleration und die Problematik von Intelligenztestnormen"* (Schallberger, 1985). Dieses Manuskript wird auch heute noch etwa verlangt, und zwar weil es 1987 in einem Artikel im Psychological Bulletin ausführlich zitiert worden ist (Flynn, 1987). Es ist seit einigen Monaten vergriffen. Ich habe mich entschlossen, das umfangreiche Originalmanuskript nicht mehr neu zu vervielfältigen und statt dessen diesen Text herauszugeben. Die beiden Texte haben folgenden Zusammenhang:

Das Manuskript von 1985 behandelte im wesentlichen zwei Themen: 1.) Eine empirische Untersuchung, in deren Rahmen HAWIK und HAWIK-R verglichen wurde. 2.) Der Versuch einer Abschätzung der Grösse des "IQ-Gewinns" im deutschen Sprachraum von ca. 1956 bis ca. 1983 auf der Basis eigener und fremder Befunde zum HAWIK und zum HAWIK-R. Die Ausführungen zum ersten Thema, also zum empirischen Vergleich der beiden Tests, sind inzwischen in der Zeitschrift Diagnostica erschienen (Schallberger, 1987). Die Erörterungen zum zweiten Thema hingegen, also zum Ausmass des "IQ-Gewinns", auf die sich Flynn (1987) bezog, wurden dort nur gestreift und sind in der ursprünglichen Ausführlichkeit noch nirgends publiziert worden. Sie sind aber in der Zwischenzeit durch das Erscheinen neuer Untersuchungen zwar nicht veraltet, aber ergänzungsfähig. Im Hinblick auf ein Referat habe ich diese Ergänzungen im Frühjahr 1990 vorgenommen.[1] Der hier vorgelegte Text ist aus diesem Referat entstanden. Er enthält neben allgemeinen Erörterungen zum Thema "IQ-Gewinn" die empirischen Argumentationen aus dem Manuskript von 1985, ergänzt durch die Diskussion relevanter neuerer Arbeiten.

Die Schlussfolgerungen der beiden Papiere stimmen übrigens überein, mit *einer* Ausnahme: 1985 nahm ich aufgrund eines Vergleichs meiner Befunde mit solchen aus den USA noch an, der festgestellte überraschend hohe "IQ-Gewinn" sei spezifisch für Deutschland und am plausibelsten durch die besondere Nachkriegssituation dieses Landes zu erklären (ähnlich in Schallberger, 1987). Wie aber neuere Befunde aus andern Ländern deutlich machen, ist das Phänomen nicht wirklich auf Deutschland beschränkt.

Abschliessend sei noch darauf hingewiesen, dass die oben erwähnte Darstellung meiner Arbeit in Flynn (1987) mehrere Unrichtigkeiten enthält (Editorial note, 1987). Die einzelnen Punkte, um die es geht, sind für den interessierten Leser im Anhang dieses Papiers zusammengestellt.

---

1 Dieses Referat wurde am 3. Mai 1990 im Rahmen des Kinder- und Jugendpsychiatrischen Kolloqiums der Psychiatrischen Universitäts-Poliklinik für Kinder und Jugendliche der Universität Zürich gehalten.

Einleitung

Im Jahre 1983 erschien die revidierte Version des Hamburg-Wechsler-Intelligenztests für Kinder, HAWIK-R genannt (Tewes, 1983). Die Notwendigkeit einer solchen revidierten Fassung war schon lange hervorgehoben worden, wies doch die 1956 erschienene Originalversion, der HAWIK (Hardesty & Priester, 1956), verschiedene gravierende Mängel auf (siehe z.B. das Uebersichtsreferat von Tewes & Titze, 1983). Die seit seinem Erscheinen erhobenen kritischen Stimmen zum HAWIK-R machen deutlich, dass die Revision nicht alle diese Probleme behoben hat (z.B. Jäger, 1984). Einer der Gründe dürfte darin liegen, dass sie sich im Rahmen des Copyrights der amerikanischen Originalversion, der Wechsler Intelligence Scale for Children WISC bzw. WISC-R (Wechsler, 1974), zu bewegen hatte, das grundsätzliche Testkonzept daher unangetastet bleiben musste (Tewes, 1983, S. 4).

Unabhängig von der Frage des Testkonzepts ist allerdings die *Normenfrage*, die in der diagnostischen Praxis und in den Fachzeitschriften seit dem Erscheinen des HAWIK-R immer wieder zu Stellungnahmen und Diskussionen Anlass gab und gibt und mit der sich auch dieser Text befasst. Worum geht es?

Schon seit geraumer Zeit gibt es verschiedene empirische Hinweise darauf, dass die Anwendung der HAWIK-Normentabellen überhöhte Wertpunkte und IQs liefert.[2] Als Grund für diese "Normverschiebung" (Kornmann & Müller-Edenborn, 1975) wird allgemein ein *Kohorteneffekt* angenommen: In der seit der Eichung des HAWIK (ca. 1956) verflossenen Zeit ist die Leistungsfähigkeit späterer Geburtskohorten derart angestiegen, dass der IQ des Durchschnittskindes nicht mehr – wie es definitionsgemäss der Fall sein sollte – 100 beträgt, sondern eine Zahl über 100 erreicht. Flynn (1984, 1987) bezeichnet dieses Phänomen als *"IQ-gain" ("IQ-Gewinn")*. Dieser Begriff ist zwar aus verschiedenen Gründen problematisch, auf die im einzelnen zurückzukommen sein wird. Wegen seiner Anschaulichkeit soll er hier trotzdem – wenn auch in Anführungszeichen – weiter verwendet werden.

Mit Selbstverständlichkeit hatte man nun vom HAWIK-R erwartet, dass er (unabhängig von allen anderen Problemen) zumindest dieses Problem künstlich überhöhter Normwerte nicht mehr aufweist, da seine Normentabellen auf einer Neueichung beruhen. Diese Erwartung bestätigte sich insofern, als die Wertpunkte und IQs des HAWIK-R tatsächlich bereits eindrucksmässig deutlich niedriger ausfielen, als man es sich vom HAWIK her gewohnt war. Was jedoch irritierte, war die Grösse der Differenz: Man sprach von 10 bis 20 IQ-Punkten – eine Grössenordnung, die auch durch empirische Vergleichsuntersuchungen

---

2 "IQs" (Intelligenzquotienten) und "Wertpunkte" sind Gattungen von Testnormen; HAWIK(-R)-IQs sind durch den Mittelwert 100 und die Standardabweichung 15 definiert; HAWIK(-R)-Wertpunkte durch den Mittelwert 10 und die Standardabweichung 3.

bestätigt wurde. Gemäss einer weit verbreiteten Meinung ist nun aber diese Differenz viel zu gross, um noch als Kompensation des "IQ-Gewinns" seit der Eichung des HAWIK angesehen werden zu können. Eine Leistungssteigerung dieses Ausmasses scheint undenkbar. Der naheliegende Schluss unter diesen Umständen war: Die Normentabellen des HAWIK-R liefern *zu tiefe* Wertpunkte und IQs! Und man suchte nach Gründen für dieses neue Normenproblem!

Ziel dieses Papiers ist es, diese Diskussion um die HAWIK-R-Normen aufzuarbeiten. Dabei liegt das *Hauptinteresse* nicht auf dem eher technischen Normenproblem eines spezifischen Tests, sondern auf der damit – wie oben ausgeführt – eng verzahnten *Frage nach dem Ausmass des "IQ-Gewinns"* von ca. 1956 (genauer: der Eichung des HAWIK) bis ca. 1983 (der Eichung des HAWIK-R). Der *Aufbau* gestaltet sich wie folgt: Um die angesprochene "Verzahnung" der beiden Fragenkomplexe noch etwas deutlicher werden zu lassen, folgen im nächsten Abschnitt zunächst einige begriffliche und methodische Erläuterungen zum Konzept des "IQ-Gewinns" (Abschnitt 1). Auf diesem Hintergrund wird dann auf die verschiedenen Argumentationen und die empirische Befundlage eingegangen (Abschnitt 2). Anschliessend werden einige Folgerungen für den Einsatz von Intelligenztests in der Diagnostik und in der Forschung gezogen (Abschnitt 3). Abschliessend wird noch kurz die Frage nach einer Erklärung des Phänomens "IQ-Gewinn" diskutiert (Abschnitt 4).

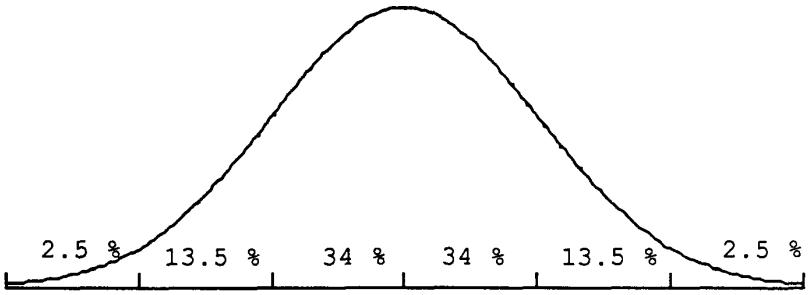
1. "IQ-Gewinn": Begriffliches, Relevanz und Methodisches

Es wurde bereits oben erwähnt, dass der Begriff "IQ-Gewinn" nicht sehr glücklich gewählt sei. Einer der Gründe liegt darin, dass er aufgrund eines naiven Verständnisses von "IQ" ("Intelligenz-Quotient") leicht Assoziationen zum Begriff "Intelligenz-Gewinn" wecken kann. Wenn man aber beachtet, dass es sich beim IQ einfach um einen Typus von Testnormen handelt, wird deutlich, dass sich der Begriff "IQ-Gewinn" zunächst einmal nur auf *Veränderungen durchschnittlicher Testleistungen* bezieht. Die Frage, ob es sich dabei um den Ausdruck eines tatsächlichen Intelligenzgewinns handelt, bleibt hingegen völlig offen. Dies ergibt sich bereits aus der statistischen Definition des IQs: In vielen gängigen Intelligenztests entstehen IQ-Tabellen aus einer einfachen Lineartransformation der Rohwerte (= Anzahl gelöster Aufgaben).[3] Basis bilden dabei die Verteilungskennwerte dieser Rohwerte in der Eichstichprobe (vgl. unterer Teil von Abb. 1). Durch die Lineartransformation wird der durchschnittlichen Leistung "m" ein IQ von 100 zugeordnet, der Leistung, die eine Standardabweichung "s" darunter liegt, ein IQ von 85, etc. Ziel dieses Unterfangens ist es, dem Diagnostiker den Vergleich einer individuellen Leistung mit derjenigen einer Referenzgruppe (Eichstichprobe) zu erleichtern.

---

3 Die folgenden Argumentationen gelten aber auch analog bei anderen Vorgehensweisen.

Abb. 1: Definition des IQ durch die Verteilung der Rohwerte in der Eichstichprobe



Rohwerte:	m-2s	m-s	m	m+s	m+2s
IQ:	70	85	100	115	130

Erläuterungen: m = Mittelwert aus der Eichstichprobe
s = Standardabweichung aus der Eichstichprobe

Würde man nochmals eine Repräsentativstichprobe testen und für alle Probanden mit Hilfe der Normentabellen IQs bestimmen, müsste sich also – dies folgt aus der Logik des beschriebenen Prozeders – ein mittlerer IQ von 100 und eine Standardabweichung von 15 ergeben. Solange man nicht damit rechnet, dass sich die Leistungsfähigkeit in Intelligenztestaufgaben zwischen Geburtskohorten wesentlich verändert, kann man auf derartige Kontrolluntersuchungen verzichten. Dies war im deutschen Sprachraum jahrzehntelang der Fall. Die Eichuntersuchungen vieler Tests liegen viele Jahre zurück.

Was geschieht nun aber, wenn sich die Durchschnittsleistung zwischen den Geburtskohorten effektiv verändert? Dann verschiebt sich die Kurve in Abbildung 1 nach rechts oder links. Da aber die Umrechnungstabelle "Rohwerte - IQs" dieselbe bleibt und dieser Veränderung nicht Rechnung trägt, resultiert ein IQ-Mittelwert über oder unter 100; je nach der Richtung der Differenz wäre dann von einem "IQ-Gewinn" oder – analog – von einem "IQ-Verlust" zu sprechen. Hier zeigt sich somit ein weiterer Grund, warum der Begriff "IQ-Gewinn" unglücklich ist. Denn per definitionem sollte der Durchschnitts-IQ in einer Repräsentativstichprobe bei 100 liegen. Liegt er über 100, dann hat die neue Generation nicht etwa IQ-Punkte gewonnen, sondern der IQ dieses Tests hat schlicht seine (definitionsgemässe) Bedeutung verloren!

Die Frage, die sich nun aufdrängt, ist, ob überhaupt und in welchen Grössenordnungen solche Leistungsverschiebungen auftreten. Muss man mit Verschiebungen deutlicher Grössenordnung rechnen, ergeben sich im wesentlichen zwei Familien von Folgeproblemen: Das eine dieser Probleme betrifft die diagnostische Praxis: Solche Leistungsverschiebungen

("IQ-Gewinne" oder "IQ-Verluste") haben zur Folge, dass der Nutzen von Testergebnissen fraglich wird: Wenn unklar ist, ob ein IQ von 100 tatsächlich der Durchschnittsleistung entspricht, kann ein Diagnostiker mit einem IQ nicht mehr viel anfangen. Das andere Problem ist eher intelligenztheoretischer und entwicklungspsychologischer Natur und mündet in die Frage nach der Erklärung solcher Leistungsverschiebungen aus (siehe Abschnitte 3 und 4).

Wie kann man nun das Ausmass solcher Leistungsverschiebungen einigermassen gesichert feststellen? Es gibt im wesentlichen vier Typen von Untersuchungen, die dazu Informationen liefern:[4]

Typ 1: Die beste Möglichkeit wären *Reihenuntersuchungen* in möglichst *unselegierten Geburtskohorten*, also die Untersuchung ganzer Populationen. Wenn man dann Unterschiede zwischen den Durchschnittsleistungen feststellen würde, wäre dies ein klarer Beweis für einen "IQ-Gewinn" oder "IQ-Verlust".

Diesem Ideal sehr nahe kommt die Anwendung von Intelligenztests bei Eingangsprüfungen in Armeen, und zwar in Ländern mit allgemeiner Dienstpflicht. Derartige Resultate wurden in den letzten Jahren zum Beispiel aus Holland, Belgien und Norwegen berichtet. Im deutschen Sprachraum fehlt eine Untersuchung dieser Art.

Typ 2: Ein ebenfalls recht gutes Vorgehen würde die *Neueichung* eines im wesentlichen unveränderten *alten Tests* darstellen. Der Vergleich der Ergebnisse der neuen Eichstichprobe mit jenen der alten Eichstichprobe würde dann erlauben, das Ausmass des "IQ-Gewinns" oder "-Verlustes" abzuschätzen (wie oben demonstriert). Dabei besteht aber eine grössere Unsicherheit als bei Untersuchungen des Typs 1: Denn bei der Erklärung allfälliger Unterschiede muss man auch die Möglichkeit kleinerer oder grösserer Stichprobenunterschiede in Rechnung stellen, die "IQ-Gewinne" oder "-Verluste" lediglich vortäuschen. Jede Repräsentativstichprobe enthält ja ihrem Wesen nach zufällige Abweichungen von den Verhältnissen in der jeweiligen Grundgesamtheit.

Derartige Daten gibt es beispielsweise aus Neuseeland, Kanada und den USA. Im deutschen Sprachraum fehlt bislang eine derartige Neustandardisierung. Gegenwärtig läuft in unserer Abteilung aber eine Untersuchung dieser Art: Es geht dabei um den PSB-Horn und den Vergleich der Testleistungen zwischen ca. 1978 (Schallberger, Vetsch, Orth, Schädelin & Wiederkehr, 1981) und ca. 1990 in der Schweiz.

Typ 3: Vom Typ 2 streng zu unterscheiden sind Untersuchungen des Typs 3, obschon sie auf den ersten Blick recht ähnlich aussehen. Es geht dabei um Untersuchungen an irgendwelchen geographisch und/oder klinisch speziellen Stichproben, die nicht den

---

4 Die nachstehend ohne Quellenangabe angeführten Beispiele von Untersuchungen finden sich in Flynn (1987) dokumentiert.

Anspruch erheben können, wirklich mit der ursprünglichen Eichstichprobe vergleichbar zu sein. Wenn man jedoch mit mehr oder weniger guten Gründen annehmen kann, dass die untersuchte Stichprobe in etwa durchschnittlich leistungsfähig ist, dann kann der effektiv gefundene Durchschnitts-IQ zumindest als grober Indikator für einen "IQ-Gewinn" oder "-Verlust" genommen werden. Es bleibt aber immer die Möglichkeit, dass allfällige Unterschiede nur Besonderheiten der Untersuchungsstichprobe (z.B. Besonderheiten der sozioökonomischen Zusammensetzung etc.) widerspiegeln. Zusätzlich stellt sich übrigens wie bei Typ 2 die Frage nach der Repräsentativität der ursprünglichen Eichstichprobe. Informationen aus Untersuchungen dieses Typs haben damit – je nach Qualität der Untersuchungsstichprobe – für die hier thematische Frage möglicherweise einen gewissen Indikatorwert – jedoch kaum mehr.

Solche Untersuchungen gibt es aus vielen Ländern zu vielen Tests. Jene, die – mit den obigen Einschränkungen – eventuell einen Rückschluss auf die Ueberhöhung des HAWIK-IQs zulassen, werden in Abschnitt 2.2 besprochen.

Typ 4: Der letzte Untersuchungstyp ist von der Logik der Argumentation her der komplexeste, gleichzeitig aber durchführungstechnisch der einfachste. Voraussetzung für Untersuchungen des Typs 2 (Schluss auf den "IQ-Gewinn" aus der Neueichung eines Tests) ist, wie dort gesagt wurde, dass der neu geeichte Test nicht wesentlich verändert wird. Nur dann kann man die Testleistungen (Rohwerte) der beiden Eichstichproben direkt vergleichen. Die meisten Tests werden nun aber sinnvollerweise nicht einfach unverändert neu geeicht, sondern vor einer allfälligen Neueichung revidiert. Beispielsweise werden veraltete Aufgaben ersetzt, die Anzahl der Aufgaben verändert, einzelne Aufgabentypen hinsichtlich Durchführung ökonomischer gestaltet etc. Damit verbietet sich aber ein unmittelbarer Vergleich der Ergebnisse der beiden Eichstichproben, sie resultieren ja wegen jener Veränderungen schon a priori in verschiedenen Rohwerten. Hingegen ist ein *indirekter Vergleich der zwei Eichstichproben* möglich: Werden die beiden Testversionen nämlich denselben Probanden vorgelegt und bestimmt man die IQs im alten und im neuen Test, so wird dieselbe individuelle Leistungsfähigkeit sowohl an der Leistungsfähigkeit der alten Eichstichprobe als auch an der Leistungsfähigkeit der neuen Eichstichprobe gemessen. Die mittlere IQ-Differenz zwischen den beiden Testversionen gibt damit Aufschluss über das Ausmass der "Entwertung" oder "Aufwertung" ein und derselben Leistung im Zeitraum zwischen den beiden Eichungen. Diese Ent- oder Aufwertung widerspiegelt aber nichts anderes als das gesuchte Ausmass des "IQ-Gewinns" bzw. "IQ-Verlusts" in der Eichpopulation im genannten Zeitraum. Dabei müssen allerdings einige untersuchungstechnische Bedingungen erfüllt sein. Die wichtigsten sind die folgenden: 1.) Die beiden Tests müssen psychologisch – trotz Revision – einigermassen äquivalent sein. 2.) Die Untersuchungsstichprobe muss zwar nicht repräsentativ, aber doch im wesentlichen das in den Eichstichproben angezielte Leistungsspektrum umfassen. 3.) Zur Kontrolle der

Lerneffekte durch das zweimalige Testen muss die Reihenfolge der beiden Testversionen ausbalanciert werden. Zusätzlich hängt natürlich die Aussagekraft einer Untersuchung dieses Typs für das Problem des "IQ-Gewinns oder -Verlusts" – wie beim zweiten Typ – von der Güte der Eichstichproben ab.

Solche Untersuchungen gibt es aus einer Reihe von Ländern, vor allem aus den USA, Japan, aber auch Australien. Sie setzen voraus, dass eine revidierte Fassung eines alten Intelligenztests erscheint. Dies war nun im deutschen Sprachraum im Jahre 1983 der Fall; damals erschien eben der HAWIK-R, eine revidierte, aber vom Testkonzept her unveränderte Fassung des alten, 1956 erschienen HAWIK. Damit eröffnete sich die Möglichkeit, auch für den deutschen Sprachraum etwas einigermassen Stichhaltiges über das Ausmass des "IQ-Gewinns" festzustellen. Die entsprechenden Studien bilden Thema von Abschnitt 2.3.

2. Was lehren die HAWIK- und die HAWIK-R-Normen für den "IQ-Gewinn" im deutschen Sprachraum?

2.1 Vorblick

Im folgenden sollen nun die Daten und Argumentationen zu den HAWIK(-R)-Normen diskutiert werden, die im Zusammenhang mit der Frage nach dem "IQ-Gewinn" von direkter oder indirekter Bedeutung sind. In Abschnitt 2.2 geht es um die Untersuchungen vom Typ 3, also reine HAWIK-Untersuchungen. In Abschnitt 2.3 werden die verfügbaren Untersuchungen des Typs 4, also Vergleichsuntersuchungen von HAWIK und HAWIK-R, vorgestellt. In Abschnitt 2.4 wird die Plausibilität verschiedener Interpretationen der entsprechenden Befunde untersucht. In Abschnitt 2.5 wird eine Bilanz gezogen.

Der folgende Text ist in Abschnitte von Normaldruck und Kleindruck gegliedert. In Kleindruck gehaltene Textteile enthalten die Detaildiskussionen der einzelnen Untersuchungen[5]. Wer sich für die Einzelheiten dieser Diskussionen nicht interessiert, kann diese Passagen überspringen. Die entsprechenden Schlussfolgerungen finden sich im normal gedruckten Text jeweils zusammengefasst.

---

5 Soweit sie Untersuchungen betreffen, die 1985 oder früher erschienen sind, stammen die Diskussionen aus Schallberger (1985).

2.2 Untersuchungen zur Ueberhöhung des HAWIK-IQs

Obwohl sie keine stichhaltigen Schlüsse zulassen, sei zunächst auf Untersuchungen des Typs 3 eingegangen, also solche, die auf HAWIK-Aufnahmen in speziellen Stichproben beruhen. Dazu sind vor allem zwei Studien zu nennen. Die eine stammt aus der Schweiz (Schallberger, Gysin, Mattes & Sigrist, 1981, im folgenden "Zürcher Eichung" genannt). Sie ergab einen um ca. 15 Punkte überhöhten IQ.[6] Die andere Studie betrifft österreichische Kinder (Schubert & Berlach, 1982). Sie spricht gemäss der Meinung der Autorinnen für eine Ueberhöhung um 6 IQ-Punkte. Eine genauere Analyse der Studie lässt aber vermuten, dass sie die effektiven Verhältnisse deutlich unterschätzt und dass eher mit ca. 14 Punkten zu rechnen ist.

Zu den zwei Studien im einzelnen:

Bei der Zürcher Eichung (Schallberger, Gysin et al., 1981) handelt es sich um eine Repräsentativstudie, die auf den Testergebnissen von je ca. 150 9- und 12-jährigen Zürcher Kindern aus dem Jahre 1977 basiert. Ob bzw. in welchem Ausmass die festgestellte Ueberhöhung des HAWIK-IQs um ca. 15 Punkte als Ausdruck eines "IQ-Gewinns" angesehen werden kann, musste dabei offen bleiben. Voraussetzung für einen derartigen Schluss wäre ja der Nachweis, dass die Zürcher Kinder von ca. 1956 ungefähr dieselben Testleistungen erbracht hatten (hätten) wie jene der HAWIK-Eichstichprobe.

Schubert & Berlach (1982) werteten die HAWIK-Protokolle von 2318 sog. "heilpädagogischen Fällen" aus, die an der Universitäts-Kinderklinik Wien untersucht worden waren. In ihren "neuen Richtlinien zur Interpretation des HAWIK" sprechen die Autorinnen von einer Ueberhöhung des HAWIK-IQs um 6 Punkte. Mindestens drei Gründe sprechen aber dafür, dass diese Schätzung zu tief angesetzt ist: Erstens wird durch die gleichzeitige Verrechnung von Daten aus dem Zeitraum von 1962 bis 1979 nicht berücksichtigt, dass in diesen Jahren – wie die entsprechenden Abbildungen in Schubert & Berlach (1982) zeigen – eine ständige Leistungsverbesserung stattgefunden hat. Zweitens wurde offenbar bei einem Teil der Testprotokolle (jenen aus der Zeit nach 1975) eine strengere Bewertung angewendet, als das HAWIK-Handbuch vorsieht. Und drittens wurde kein Versuch gemacht, die Besonderheit heilpädagogischer Fälle in Rechnung zu stellen.

Versucht man den Einfluss dieser Störmomente abzuschätzen, ergibt sich folgendes: Nimmt man aus Schubert & Berlach (1982) aufgrund der dortigen Abbildungen 1 und 2 nur den Mittelwert der Jahre 1973 bis 1975, also die jüngsten Daten (Einwand 1), die aber noch vor der Anwendung einer strengeren Auswertung erhoben worden waren (Einwand 2), resultiert ein durchschnittlicher Gesamt-IQ von ca. 109 Punkten. Schwieriger ist die Abschätzung der Rolle der Besonderheit der untersuchten Gruppe (Einwand 3). Bei 300

---

6 Die Reaktion der Praxis auf die Zürcher Normen war übrigens dieselbe wie jene auf die HAWIK-R-Normen 12 Jahre später: Man betrachtete sie als im Vergleich zu den alten HAWIK-Normen viel zu streng und vermutete das Vorliegen von Eichungsfehlern. Dies war dann Anlass für eine Kontrolluntersuchung (Schallberger, 1983) und eigentlicher Startpunkt für mein Interesse an der Frage des "IQ-Gewinns".

schulpsychologischen Beratungsfällen aus dem Kanton Zürich, die aus denselben Gründen getestet worden waren wie die Probanden der Wiener Studie, ergab sich im Vergleich zur Zürcher Eichstichprobe ein im Mittel um ca. 5 Punkte niedrigerer IQ (Schallberger, 1983). (Eine mehr als doppelt so hohe Differenz liesse sich aus einem Vergleich der Untersuchungen von Pfeiffer & Zielinski (1976) und Meckel (1977) herauslesen, wobei aber auch regionale Unterschiede in diesen Vergleich einfliessen.) Zählt man diese 5 Punkte zu den 109 Punkten hinzu, ergibt sich die oben angegebene Ueberhöhung von 14 Punkten. Abgesehen von der Problematik der eben gemachten Korrektur ist hier wie bei der Zürcher Eichung aber letztlich unklar, welche Rolle regionalen Unterschieden zukommt.

Die Aussagekraft beider Untersuchungen leidet darunter, dass sie Kinder aus einem anderen und sehr viel spezifischeren geographischen Raum betreffen als mit der Eichstichprobe des HAWIK angezielt worden war (Bundesrepublik Deutschland). Leider fehlen aber analoge Untersuchungen aus Deutschland. Eine indirekte Möglichkeit zur Abschätzung der Situation bietet die Untersuchung von Pfeiffer & Zielinski (1976). Sie legt aufgrund verschiedener Vergleiche die Vermutung nahe, dass der HAWIK in der Mitte der siebziger Jahre in Deutschland möglicherweise einen um ca. 12 Punkte zu hohen IQ lieferte.

Die Untersuchung von Pfeiffer & Zielinski (1976) betraf sämtliche Schüler der 4. Klasse eines Heidelberger Vororts. Dabei wurde gleichzeitig der HAWIK und der PSB (Horn, 1969) angewendet. Die Differenz des durchschnittlichen HAWIK-IQs (113.3) und des PSB-IQs (107.5) beträgt ca. 6 Punkte. Im gleichen Jahr erschien eine Neueichung des PSB (Royl & Schwarzer, 1976), die durch einen Vergleich mit den Original-PSB-Normentabellen erlaubt, die damalige Ueberhöhung der PSB-Ergebnisse abzuschätzen. Diese scheint ca. 6 IQ-Punkte betragen zu haben. Beides zusammen ergibt 12 Punkte als Schätzung für die Ueberhöhung des HAWIK-IQs.[7] Auch diese Schätzung steht aber natürlich auf sehr schwachen Füssen; regionale Differenzen und allfällige Unterschiede der Eichstichproben von Horn (1969) und Royl & Schwarzer (1976) sind die beiden augenfälligsten Störfaktoren.

Zusammenfassend ist festzuhalten: Die Untersuchungen des Typs 3, die zum HAWIK vorliegen, lassen wegen der vielfältigen Störmomente keinen Schluss auf das Ausmass des "IQ-Gewinns" seit der HAWIK-Eichung zu; sie haben höchstens einen – wie in Abschnitt 1 gesagt wurde – gewissen Indikatorwert. Das durch diese Untersuchungen Indizierte stimmt nun aber doch in bemerkenswerter Weise überein: Die drei Untersuchungen sprechen – bei aller Problematik – dafür, dass möglicherweise bereits in der Mitte der siebziger Jahre eine Ueberhöhung der HAWIK-IQs um ca. 12 bis 15 Punkte vorlag.

---

7 Am Rande sei hier noch auf eine Besonderheit dieser Studie hingewiesen: Hier ist nämlich – im Unterschied zu den meisten anderen in diesem Papier genannten Studien – der Handlungs-IQ nicht wesentlich stärker überhöht ist als der Gesamt-IQ.

2.3 Vergleichsuntersuchungen von HAWIK und HAWIK-R

Informativere Befunde als die eben beschriebenen Untersuchungen liefern Untersuchungen des Typs 4; das heisst (im gegebenen Zusammenhang) Vergleichsuntersuchungen von HAWIK und HAWIK-R. Bis heute existieren sechs Untersuchungen dieser Art. Sie stammen aus den Jahren 1983 bis 1989, aus Deutschland und der Schweiz und betreffen sehr verschiedenartige Untersuchungsgruppen (siehe Tabelle 1). In Anbetracht der weiten Verbreitung des HAWIK und der Wichtigkeit der Frage bereits für die diagnostische Praxis ist diese Anzahl eigentlich erstaunlich gering.[8,9] Auffällig ist auch, dass im allgemeinen mit recht kleinen Stichprobengrössen gearbeitet wird, was recht grosse Stichprobenfehler impliziert, sodass den einzelnen Untersuchungen nicht allzuviel Gewicht beigemessen werden darf.

Tabelle 1: Die Ergebnisse der Vergleichsuntersuchungen von HAWIK und HAWIK-R: Originalbefunde

				IQ-Differenzen		
Untersuchung	n	Alter	IQ[a]	Verbal	Handlung	Gesamt
(1) Mayer (1983)	87	?	70.2	7.1	15.4	12.4
(2) Eggert et al. (1984)	25	?	83.0	9.2	18.4	15.1
(3) Bründler (unpubl.)	21	7-15 J.	89.0	11.5	25.4	20.7
(4) Boehnke (1986)	17	10-12 J.	101.5	12.9	17.4	17.8
(5) Schallberger (1987)	124	7-10 J.	96.5	11.4	27.3	21.5
(6) Gutezeit (1989)	20	8-15 J	90.5	12.15	15.75	16.0

Anmerkung: [a] HAWIK-R-IQ, wegen der Lerneffekte nur ungefährer Anhaltspunkt über die durchschnittliche Leistungsfähigkeit der Gruppen

In der Spalte ganz rechts von Tabelle 1 sind die Differenzen von HAWIK und HAWIK-R im Gesamt-IQ aufgeführt. Sie schwanken zwischen 12.4 und 21.5 Punkten. Gemäss dem Testkonzept Wechslers liefert aber der HAWIK nicht nur einen Gesamt-IQ sondern auch zwei Teil-IQs, den sogenannten Verbal-IQ und den Handlungs-IQ. Im Verbal-IQ sind jene Aufgabentypen zusammengefasst, die sprachliche und rechnerische Leistungen verlangen, also Leistungen, die deutlich bildungsabhängig sind. Hier ergeben sich Differenzen

---

8 Die Zusammenstellung in Tabelle 1 basiert auf einer eigenen Durchsicht verfügbarer Zeitschriften und einer Literatursuche in der Datenbank PSYNDEX.

9 Die Situation der Praxis nach dem Erscheinen des HAWIK-R entspricht jener, die 1974 in den USA entstanden war, als der WISC-R den WISC – die 1949 publizierte amerikanische Originalversion des HAWIK – ersetzte. Swerdlik (1977) berichtete bereits drei Jahre später über ein Dutzend Vergleichsuntersuchungen!

zwischen 7.1 und 12.9 IQ-Punkten. Beim Handlungs-IQ geht es – nach Wechsler – eher um handlungsorientierte, praktische Leistungen. Die entsprechenden Tests haben einen mehr spielerischen Charakter, beispielsweise findet sich hier eine Art Zusammensetzspiel oder die Aufgabe, die Bilder eines kurzen Comics in die richtige Reihenfolge zu bringen. Hier ergeben sich Differenzen zwischen 15.4 und 25.4 IQ-Punkten.

Die Unterschiede zwischen den Differenzen aus den verschiedenen Untersuchungen sind somit recht gross, und es stellt sich die Frage, ob und wie aus diesen Zahlen präzisere, einigermassen verlässliche Schätzungen der IQ-Differenzen von HAWIK und HAWIK-R abgeleitet werden können. Vertieft man sich zu diesem Zweck eingehender in die aufgeführten Studien, fällt auf, dass einzelne der Untersuchungen unter gravierenden untersuchungstechnischen Mängeln leiden. Ein (zentrales) Beispiel eines solchen Problems ist etwa die Frage eines Lern- bzw. Uebungseffekts: Bei den Untersuchungen von Mayer (1983), Eggert et al. (1984) und Gutezeit (1989) wurde mit einer festen Testreihenfolge gearbeitet, und zwar wurde zuerst der HAWIK und dann der HAWIK-R durchgeführt. Nun weiss man aber aus andern Untersuchungen, dass die Durchführung von zwei so ähnlichen Tests hintereinander Uebungseffekte zur Folge hat, die das zweite Testergebnis künstlich erhöhen. In den genannten Fällen profitiert also der HAWIK-R von diesem Uebungsgewinn, was zu einer künstlichen Senkung der Differenzen führt (siehe Schallberger, 1987). Versucht man den Effekt dieser Uebungsgewinne sowie anderer Probleme (siehe im Detail weiter unten) abzuschätzen und die Differenzen entsprechend zu korrigieren, so ergeben sich die Zahlen in Tabelle 2.

Tabelle 2: Die Ergebnisse der Vergleichsuntersuchungen von HAWIK und HAWIK-R: Teilweise korrigierte Befunde

			IQ-Differenzen		
Untersuchung	n	Alter	Verbal	Handlung	Gesamt
(1a) Mayer (1983)	87	?	11.4	22.4	18.2
Korrektur:			(7.1+0.3+4)	(15.4+3+4)	(12.4+1.8+4)
(2a) Eggert et al. (1984)	25	?	9.9	24.4	18.7
Korrektur:			(9.2+0.65)	(18.4+5.95)	(15.1+3.6)
(3) Bründler (unpubl.)	21	7-15 J.	11.5	25.4	20.7
(4) Boehnke (1986)	17	10-12 J.	12.9	17.4	17.8
(5) Schallberger (1987)	124	7-10 J.	11.4	27.3	21.5
(6a) Gutezeit (1989)	20	8-15 J	12.45	18.75	17.5
Korrektur:			(12.15+0.3)	(15.75+3)	(16.0+1.8)
Gewichtete Mittelwerte	294		11.4	24.3	19.7

Zu den Untersuchungen und Korrekturen im einzelnen:

ad (1) und (2): Mayer (1983) untersuchte 87 Sonderschüler mit dem HAWIK-R und verglich die Ergebnisse mit früher aufgenommenen HAWIK-Protokollen, die Archiven entnommen wurden. Eggert et al. (1984) testeten 25 sprachgestörte Kinder zunächst mit dem HAWIK und 14 Tage später mit dem HAWIK-R. In beiden Fällen wurde somit mit einer festen Testreihenfolge gearbeitet. Wie in Schallberger (1987) gezeigt werden konnte, werden dadurch in der zweiten Testaufnahme beträchtliche Lerneffekte wirksam: Nach drei bis vier Wochen betrugen sie beim Verbal-IQ ca. 0.65 IQ-Punkte, beim Handlungs-IQ 5.95 IQ-Punkte und beim Gesamt-IQ 3.6 Punkte. Dieser Punktegewinn begünstigt in beiden Untersuchungen den – im Vergleich zum HAWIK niedrigere Werte liefernden – HAWIK-R, was insbesondere bei Eggert et al. zu einer substantiellen Verkleinerung der Differenz geführt haben dürfte. Erstaunlicherweise zeigt diese Studie trotzdem noch grössere Differenzen als jene von Mayer. Dafür könnten Stichprobenunterschiede verantwortlich sein. Wahrscheinlich ist aber auch ein Regressionseffekt wirksam: Wie Titze & Tewes (1984, S. 59) in einer Beschreibung dieser Studie ausführen, basierte die Zuweisung der untersuchten Sonderschüler in diesen Schultyp – und damit auch die Selektion der Probanden für die Untersuchungsgruppe – oft auf dem tiefen HAWIK-Ergebnis. Unter solchen Umständen tritt aber das bekannte Phänomen der Regression zur Mitte auf, das hier – weil der Populationsmittelwert des HAWIK-R niedriger ist als jener des HAWIK – ebenfalls zu einer künstlichen Verkleinerung der beobachteten Differenzen führt.

Die Zeilen (1a) und (2a) in Tabelle 2 enthalten einen groben Versuch, die genannten Fehlerquellen zu berücksichtigen: In Zeile (2a) (Eggert et al., 14 Tage Testabstand) wurde – als vorsichtige Schätzung – der oben berichtete, nach 3 bis 4 Wochen beobachtete Lerneffekt eingesetzt. Bei Mayer ist wegen nicht bekanntem Testabstand die Rolle des Lerneffekts weniger gut abschätzbar. Der Regressionseffekt wurde in Schallberger (1987, S. 10) aufgrund von Modellrechnungen auf ca. 8 Punkte angesetzt. Die beiden Korrekturglieder in Zeile (1a) entsprechen – als wiederum vorsichtige Schätzungen – je ca. der Hälfte der verfügbaren Schätzungen des Lern- und des Regressionseffekts.

ad (3): Bründler (persönliche Mitteilung 1985) wertete eine Sammlung von 21 HAWIK- und HAWIK-R-Protokollpaaren aus, die ihm von Mitgliedern der Vereinigung Schweizerischer Kinder- und Jugendpsychologen zur Verfügung gestellt wurden. In 12 Fällen wurde zuerst der HAWIK, in 9 Fällen zuerst der HAWIK-R eingesetzt. Die Reihenfolge der Testversionen ist somit ungefähr ausbalanciert. Geht man davon aus, dass die Lerneffekte unabhängig von der Testreihenfolge sind, HAWIK und HAWIK-R also – bei entsprechender Reihenfolge – in gleichem Masse betroffen sind (vgl. Schallberger, 1987), ergeben die in der Gesamtgruppe berechneten und in Tabelle 1 angegebenen Differenzen direkt eine brauchbare Schätzung der IQ-Unterschiede. Daher wurden sie unverändert von Tabelle 1 in Tabelle 2 übernommen.

ad (4): Boehnke (1986) untersuchte eine fünfte Grundschulklasse aus Berlin mit beiden Tests (Abstand: 14 Tage), wobei eine Gruppe von 9 Schülern zuerst den HAWIK, eine Gruppe von 8 Schülern zuerst den HAWIK-R und eine dritte, hier nicht interessierende Gruppe (n = 10) zwei Mischformen aus den beiden Tests bearbeiteten. Die in Tabelle 1 enthaltenen Zahlen wurden vom Autor berechnet; sie betreffen nur die beiden

ersten Gruppen.[10] In dieser reduzierten Stichprobe ist die Testreihenfolge im wesentlichen ausgeglichen. Wie bei Studie (3) kann somit direkt mit den IQ-Differenzen argumentiert werden.

ad (5): Die Untersuchung von Schallberger (1987) betraf eine tendenziell repräsentative Schülerstichprobe aus dem Kt. Solothurn (Schweiz). Die Testreihenfolge (Testabstand: 3 bis 4 Wochen) wurde auch hier ausbalanciert. Von der Stichprobengrösse, der Untersuchungsgruppe und dem Design her dürfte es sich um die bislang aussagekräftigste Studie handeln.

ad (6): Gutezeit (1989) verglich die HAWIK-R-Ergebnisse von 20 Schülern mit schweren Lesestörungen mit früher aufgenommenen, Archiven entnommenen HAWIK-Ergebnissen. Wiederum liegt eine feste Testreihenfolge mit einer Begünstigung des HAWIK-R-Ergebnisses von – wegen unklarem Testabstand – unbestimmbarem Ausmass vor. Setzt man wie in Zeile (1a) die Hälfte des Lerngewinns aus der Untersuchung (5) ein, ergeben sich die in Zeile (6a) von Tabelle 2 festgehaltenen Werte.

Die in Tabelle 2 aufgeführten, wenn nötig korrigierten Werte stimmen nun untereinander wesentlich besser überein als die Zahlen in Tabelle 1. Dies gilt zumindest für den Gesamt-IQ (Differenzen zwischen 17.5 und 21.5) und den Verbal-IQ (Differenzen zwischen 9.9 und 12.9). Bedeutendere Abweichungen ergeben sich beim Handlungs-IQ; hier liegen die Differenzen zwischen 17.4 und 27.3. Der Unterschied der beiden Extremwerte erreicht dabei Signifikanz ($t=2.541$, $df = 131$, $p < .05$). Eine abschliessende Erklärung dieses Unterschieds ist nicht sichtbar; Stichprobenunterschiede hinsichtlich Nationalität oder Alter könnten dafür verantwortlich sein. Eine Inspektion der Subtestergebnisse der Untersuchungen (4) und (5), die die Extremwerte liefern, zeigte, dass der Unterschied vor allem durch die Subtests BE (Bilderergänzen) und BO (Bilderordnen) bedingt ist.

Zwei hier möglicherweise relevante Beobachtungen seien wenigstens genannt: 1.) Die Untersuchungen (3) und (5), die die höchsten Handlungs-IQ-Differenzen liefern, stammen beide aus der Schweiz. Dabei wurde - im Unterschied zu den späteren Schweizer Untersuchungen (Bründler & Schallberger, 1988) - der Original-HAWIK-R verwendet, der im Test Bilderergänzen noch zwei für Schweizer Kinder ungewohnte Bilder enthält (z.B. einen Briefkasten nach deutscher Norm). 2.) Einer der Unterschiede zwischen den Untersuchungen (3) und (5), die die kleinste und die grösste Handlungs-IQ-Differenz zeigen, liegt im erfassten Altersbereich (10-12 vs. 7-10 Jahre). Ein Hinweis, dass die Altersdifferenz eventuell eine Rolle spielt, ergab sich aus einer entsprechenden Reanalyse der Zürcher Eichung (Schallberger, Gysin et al., 1981): Während die 9-jährigen im Handlungsteil durchschnittlich 120 IQ-Punkte erreichten, waren es bei den 12-jährigen 114 Punkte. Bei beiden Stichproben ergab sich im übrigen eine Ueberhöhung des Gesamt-IQs um ca. 15 Punkte.

Bildet man nun – wie in solchen Metaanalysen üblich – die mit der jeweiligen Stichprobengrösse gewichteten Mittelwerte der sechs Studien, erhält man wohl die beste im Moment verfügbare Schätzung der IQ-Differenzen von HAWIK und HAWIK-R. Dabei ergeben sich folgende Werte: *Verbal-IQ: 11.4, Handlungs-IQ: 24.3, Gesamt-IQ: 19.7.*

---

[10] Ich danke Herrn Boehnke für die Ueberlassung der Daten.

Anzumerken ist, dass diese Werte Durchschnittscharakter haben; es ist nicht auszuschliessen, dass bestimmte altersmässig oder klinisch definierte Gruppen etc. abweichende Differenzen zeigen. Dafür fehlen im Moment aber zwingende Hinweise.

Wie bereits die Unterschiede zwischen den Verhältnissen im Verbal- und im Handlungsteil zeigen, sind die Differenzen von HAWIK und HAWIK-R nicht bei allen Aufgabentypen gleich gross. Dies gilt auch auf *Subtestebene*. Darauf wird in Abschnitt 2.5 zurückzukommen sein.

2.4 Alternativerklärungen zur Interpretation der IQ-Differenzen von HAWIK und HAWIK-R als "IQ-Gewinn"

Wie einleitend erwähnt, herrscht nun in der diagnostischen Praxis und in der Fachliteratur die Ansicht vor, IQ-Differenzen der im vorigen Abschnitt festgestellten Grössenordnung seien *zu hoch,* um noch als Effekt des "IQ-Gewinns" von ca. 1956 bis ca. 1983 interpretiert werden zu können. Derartige Leistungszunahmen seien zu unwahrscheinlich, die Differenzen müssten mindestens zum Teil anders erklärt werden. Dabei tauchen drei verschiedene Argumentationslinien auf, nämlich:

1.) HAWIK und HAWIK-R messen nicht dasselbe. Eine Interpretation des Unterschieds als "IQ-Gewinn" ist deswegen nicht möglich.

2.) Der HAWIK-R liefert wegen Eichungsfehlern zu tiefe IQs, die die Differenz der beiden Tests künstlich aufblähen.

3.) Der HAWIK lieferte (ebenfalls wegen Eichungsfehlern) von Anfang an zu hohe IQs, was ebenfalls dazu führt, dass die Differenzen von HAWIK und HAWIK-R künstlich erhöht ausfallen.

Diese Argumente tauchen in verschiedenen Varianten auf. Sie seien im folgenden auf ihre Plausibilität hin untersucht:

ad 1: Die erste Gruppe von Einwänden bezieht sich auf Veränderungen von Testmaterial und -durchführung, die zur Folge haben sollen, dass HAWIK und HAWIK-R generell oder in einer bestimmten Subpopulation Verschiedenes messen (z.B. Boehnke, 1986). Ein Vergleich der IQs als Mass für den "IQ-Gewinn" wäre dann besonders bei nicht repräsentativen Stichproben problematisch. Dieser Einwand ist aber auf der Basis der heutigen Befundlage nicht stichhaltig, da die Korrelationen der beiden Testformen recht hoch sind.

Empirische Indizien für eine derartige Bedeutungsverschiebung finden sich höchstens auf der Ebene einzelner Subtests (Boehnke, 1986; Schallberger, 1987). Die verfügbaren Korrelationskoeffizienten zwischen den Gesamt-IQs der beiden Versionen lauten: Eggert et al. (1984): .75; Boehnke (1986; eigene Berechnung ohne Gruppe mit Mischformen, siehe die Besprechung der Untersuchung weiter oben): .8; Schallberger (1979):

.69. Sie erreichen somit eine Höhe, die (knapp) jener entspricht, die man bei einer Wiederholung eines derart komplexen Tests in dieser Altersgruppe erwarten kann. Wirklich bedeutende Veränderungen des diagnostischen Gehalts liegen somit kaum vor.

Eine Variante dieses ersten Einwandes besagt, dass die Veränderungen beim HAWIK-R besonders leistungsschwache Kinder noch zusätzlich benachteiligen (z.B. Gutezeit, 1989). Dass dies nicht zutrifft, ergibt sich bereits aus den Tabellen 1 und 2: Untersuchungen an Kindern mit tiefem Durchschnitts-HAWIK-IQ liefern keineswegs grössere IQ-Differenzen. Das Argument wurde kürzlich von Koop & Röttger (1989) in neuer Form vorgebracht; sie glauben nachweisen zu können, dass die Differenzen zwischen HAWIK-R und anderen Tests bei leistungsschwachen Probanden sehr viel grösser seien als bei durchschnittlichen Kindern. Dieser Befund erwies sich aber bei einer Reanalyse der Daten als Folge eines Auswertungsfehlers (Schallberger, im Druck).

ad 2: Der zweite Einwand ist der am weitesten verbreitete: Nach diesem Einwand sind die Differenzen von HAWIK und HAWIK-R kein Mass für den IQ-Gewinn, weil beim HAWIK-R Eichungsfehler vorliegen. Der HAWIK-R liefere generell zu tiefe IQs und würde die Leistungsfähigkeit der Probanden unterschätzen, weil der Test an einer durchschnittlich zu leistungsfähigen Stichprobe und/oder unter Anwendung zu günstiger Testbedingungen geeicht worden sei (ausführlich z.B. Boehnke, 1986). Im gegebenen Rahmen geht es dabei nicht um die einzelnen Vermutungen, in die sich diese Auffassung konkretisiert, sondern lediglich um empirische Indizien, die für oder gegen zu tiefe HAWIK-R-IQs sprechen. Betrachtet man die heute verfügbaren Indizien dieser Art, muss man zum Schluss kommen, dass der HAWIK-R kaum wesentlich zu streng ist und wirklich gravierende Eichungsfehler nicht vorliegen. Es könnte sich höchstens um einzelne IQ-Punkte handeln, und das vor allem bei jüngeren Kindern. Jedenfalls bietet der Rekurs auf Eichungsfehler beim HAWIK-R keine plausible Erklärung für die massiven Differenzen von HAWIK und HAWIK-R.

Endgültige Klarheit könnte hier natürlich nur eine Untersuchung an einer repräsentativen Stichprobe deutscher Kinder liefern, was einer Neueichung gleichkäme. Nahe an dieses Ideal kommt die Schweizer Eichung, die von Bründler & Schallberger (1988) durchgeführt wurde. Die Testaufnahmen (durchgeführt 1985/86) betrafen 8-, 9-, 10- und 12jährige Kinder. Von den somit total 12 bestimmbaren IQs lagen 10 im Bereich von 98 und 102. Die zwei anderen betrafen den Handlungs-IQ (95) und als Folge den Gesamt-IQ (96) bei 8jährigen Kindern. In diesem engen Streubereich ergab sich aber eine systematische Tendenz, dass die Durchschnitts-IQs mit dem Alter leicht ansteigen (z.B. Gesamt-IQ: 96, 98, 99, 101). Vielleicht kann man daraus lesen, dass der HAWIK-R jüngere Kinder etwas strenger bewertet als ältere, als Beweis ist der Befund aber sicher nicht anzusehen. Weiter stellt sich noch die entscheidende Frage, ob das Schweizer Durchschnittskind nicht eventuell leistungsfähiger ist als das deutsche. Dagegen spricht folgender Sachverhalt: Royl & Schwarzer (1976) legten eine Neueichung des PSB (Horn, 1969) vor, die durch einen Vergleich mit den Original-Normentabellen erlaubt, die Ueberhöhung der PSB-Ergebnisse in Deutschland im Jahre 1973/74 abzuschätzen, was ca. 6 IQ-Punkte ergibt. Bei einer Schweizer Eichung des PSB-Horn im Jahre 1978 ergaben

sich 7.5 IQ-Punkte (Schallberger, Vetsch et al., 1981), was unter Berücksichtigung der möglichen Rolle des "IQ-Gewinns" zwischen 1973/74 und 1978 eher gegen die obige Annahme spricht.

Auch HAWIK-R-Mittelwerte aus Deutschland lassen auf keine systematische Unterschätzungstendenz schliessen: So fand z.B. Eggert et al. (1984) bei Sprachheilschülern 82.96, Gutezeit (1989) bei Schülern mit schweren Lesestörungen 90.5, Boehnke (1986) bei "einer vermutlich leicht überdurchschnittlich intelligenten Schulklasse" (a.a.O., S. 39) 105.11, Titze (1989, S. 93f.) bei einer bisher noch nicht erwähnten Stichprobe von 300 Schülern von 8 bis 14 Jahren, unter denen " ... Gymnasiasten und Realschüler im Vergleich zu den Hauptschülern etwas überrepräsentiert" sind, 108.94 IQ-Punkte, etc. – alles von der statistischen Definition des IQs her höchst sinnvolle Ergebnisse.

Die einzige Untersuchung, die ein unplausibel tiefes Ergebnis lieferte, ist jene von Mayer (1983): Er fand bei Sonderschülern einen durchschnittlichen Gesamt-IQ von 70.2. Um dieses Ergebnis evaluieren zu können, wäre es nützlich, die einzelnen Testprofile zu analysieren. Schallberger (1983) zeigte, dass bei Problemschülern mit sehr tiefen IQs sehr oft extreme Zackenprofile vorliegen; ein Sachverhalt, der den Schluss auf eine generell tiefe Leistungsfähigkeit ("Intelligenz") aber verbietet.

Ein völlig anderes Argument braucht Boehnke (1986) zur Ableitung der Annahme, der HAWIK-R unterschätze den IQ um ca. 7 Punkte. Er stützt sich dabei auf die Tatsache, dass der Zahlensymbol-Test in beiden Versionen identisch ist, trotzdem jedoch einen Wertepunktunterschied ergibt. Bei diesem Test – Kodieren von Zahlen in Symbole – sei aber eine Leistungssteigerung unplausibel. Die Wertepunktdifferenz in diesem Test wird daher als Indikator für die Grösse des Eichungsfehlers in IQ-Punkten verwendet. Beide Argumente sind schwer nachvollziehbar. Es gibt wohl kaum einen stichhaltigen Grund für die Unplausibilität einer Leistungsveränderung in diesem Test. Zudem gibt es aber auch keinen Grund zur Annahme, der Zahlensymbol-Test widerspiegle denselben Unterschied, der in den Gesamt-IQs zu beobachten wäre.

Ein letzter zu erwähnender Einwand im Rahmen dieser Argumentationslinie bezieht sich auf die Tatsache, dass die Eichstichprobe des HAWIK-R keine Sonderschüler enthielt. Für die Frage des "IQ-Gewinns" ist dies eher ein Vorteil, weil dasselbe auch für die Eichstichprobe des HAWIK – wie der meisten Intelligenztests – zutrifft (Hardesty & Priester, 1956). Daraus resultiert übrigens auch keine wirklich ernsthafte Unterschätzungstendenz. Mit Probeberechnungen lässt sich leicht zeigen, dass ein proportional richtiger Einbezug dieser Gruppe selbst bei extremen Annahmen die Normwerte nur geringfügig verändert.

ad 3: Eine dritte Alternative zur Erklärung der vollen Differenz von HAWIK und HAWIK-R als "IQ-Gewinn" wäre der Rekurs auf die Vermutung, dass schon beim HAWIK Eichungsfehler vorlagen, dass er also heute unter anderem deswegen zu hohe IQs liefert, weil er dies mindestens in abgeschächter Form von Anfang getan hat und damit die Differenzen von HAWIK und HAWIK-R künstlich aufbläht. Obwohl logisch ebenso korrekt wie die obige, zweite Erklärung, wurde dieses Argument ausser bei Schallberger (1985, 1987) noch nie in Betracht gezogen, geschweige denn systematisch geprüft. Dabei wirkt sich allerdings ungünstig aus, dass Repräsentativuntersuchungen mit dem HAWIK

aus den fünfziger Jahren fehlen. Jedoch gibt es andere Studien, die zumindest auf das Vorliegen von Indizien hin untersucht werden können.

Dass der HAWIK von Anfang an überhöhte IQs zeigte (z.B. weil er an einer zu wenig leistungsfähigen Stichprobe geeicht worden ist), scheint allein aufgrund der Beschreibung von Priester (1958) unwahrscheinlich. In der Eichstichprobe waren nämlich Kinder aus städtischen Verhältnissen übervertreten. Hingegen können die Ergebnisse früher Vergleichsuntersuchungen des HAWIK mit anderen Tests (über deren Normenqualität allerdings auch nichts bekannt ist) in diesem Sinne interpretiert werden: Priester & Kukulka (1958) verglichen 1956 anhand der Testleistungen von 50 Hamburger Volksschülern HAWIK und Binetarium (Norden, 1953) und fanden im ersten einen um 2.28 IQ-Punkte höheren IQ (HAWIK: 107.94, Binetarium: 105.66). Priester (1959) verglich bei je 50 Hilfsschülern, Volksschülern und Oberschülern (ebenfalls aus Hamburg) den HAWIK mit dem Stanford-Intelligenztest (SIT, Lückert, 1957). In allen drei Gruppen war der HAWIK-IQ höher: 0.6 Punkte bei den Sonderschülern, 4.2 Punkte bei den Volksschülern (102.3 vs. 98.1) und 5.8 Punkte bei den Oberschülern. In drei der vier Gruppen, auf die sich die genannten Untersuchungen beziehen, ist also der HAWIK-IQ um einzelne Punkte höher als derjenige des Vergleichstests. Bemerkenswert ist dabei aber, dass es keinen Hinweis darauf gibt, dass der Handlungs-IQ stärker überhöht war als der Verbal-IQ, im Gegenteil: In zwei der vier Gruppen war der Handlungs-IQ um 6 bis 9 Punkte tiefer und in den beiden andern Gruppen lediglich um 1.6 bzw. 2.6 Punkte höher.

Eine analoge Differenz wie Priester (1959) von knapp 3 Punkten zwischen HAWIK und SIT fand auch Amelang (1966) in einer 7 Jahre später publizierten Untersuchung bei 114 Kindern des Landkreises Marburg. Leider fehlen in dieser Untersuchung Angaben über die Teil-IQs.[11]

Die verfügbaren Indizien sprechen somit dagegen, dass beim HAWIK mit wirklich gravierenden Eichungsfehlern zu rechnen ist. Sie schliessen aber nicht aus, dass der HAWIK bereits in den fünfziger Jahren um einzelne Punkte zu hohe IQs lieferte. Hingegen lassen sie keinen Hinweis darauf erkennen, dass die Ueberhöhung des HAWIK-IQs gegenüber dem Verbal-IQ bereits in den fünfziger Jahren festzustellen war.

2.5 Zusammenfassung und Diskussion

In diesem Abschnitt 2 wurden Befunde und Argumente diskutiert, die im Zusammenhang mit der Frage nach dem Ausmass des "IQ-Gewinns" von ca. 1956 bis ca. 1983 von direkter oder indirekter Bedeutung sind. Diese Diskussion lässt sich wie folgt zusammenfassen:

Die drei referierten HAWIK-Untersuchungen (Abschnitt 2.2) sprechen dafür, dass bereits in der Mitte der siebziger Jahre möglicherweise eine Ueberhöhung des HAWIK-Gesamt-IQs um ca. 12 bis 15 Punkte vorlag. Die Beweiskraft jeder einzelnen Untersuchung ist zwar schwach, da es sich um Untersuchungen an speziellen Stichproben handelt (Untersuchungen

---

11 Bemerkenswert ist, dass in allen drei Vergleichsuntersuchungen, von denen hier die Rede ist, die Testreihenfolge variiert worden war!

des Typs 3, siehe Abschnitt 1). Dass sie jedoch alle in dieselbe Richtung weisen, ist immerhin bemerkenswert.

Aussagekräftiger ist die kleine Metaanalyse von sechs Vergleichsuntersuchungen von HAWIK und HAWIK-R in Abschnitt 2.3 (Untersuchungen des Typs 4, siehe Abschnitt 1). Sie ergab, dass die HAWIK-R Gesamt-IQs um ca. 20 Punkte tiefer auszufallen scheinen als jene des HAWIK, ferner, dass beim Handlungs-IQ mit wesentlich grösseren Differenzen zu rechnen ist als beim Verbal-IQ (24 versus 11 IQ-Punkte). Auch diese Vergleichsuntersuchungen sind nicht problemlos, und zwar aus untersuchungstechnischen Gründen (Versuchspläne, Stichproben, Stichprobengrössen). Der Versuch, diese Unzulänglichkeiten mit Schätzungen zu kompensieren, ergab aber doch recht gut übereinstimmende Zahlen.

Schliesslich (Abschnitt 2.4) wurde noch versucht, die Plausibilität von Argumenten zu überprüfen, die die Interpretation dieser Differenzen als Mass für den "IQ-Gewinn" im Zeitraum zwischen den beiden Testeichungen problematisieren. Eine Veränderung des diagnostischen Gehalts des Tests durch die Revision – wie bisweilen vermutet wird – liegt aller Wahrscheinlichkeit nach nicht vor. Weniger eindeutig ist die Situation in bezug auf mögliche Eichungsfehler, und zwar sowohl beim HAWIK-R als beim HAWIK. Jedoch dürfte es sich in beiden Fällen höchstens um einzelne IQ-Punkte handeln. Ferner gibt es keinen Anhaltspunkt dafür, dass die massive Diskrepanz zwischen den Differenzen im Verbalteil und im Handlungsteil etwas mit Eichungsfehlern zu tun hat.

Nimmt man nun all diese Punkte zusammen, so scheint es *auf der Basis des gegenwärtigen Wissens* kein wirklich überzeugendes Gegenargument gegen die Richtigkeit der Vermutung zu geben, dass *die Differenzen von HAWIK und HAWIK-R zumindest zum wesentlichen Teil als Spiegel des "IQ-Gewinns" von ca. 1956 bis ca. 1983* zu betrachten sind.

Der "IQ-Gewinn" von ca. 1956 bis ca. 1983 scheint also tatsächlich von einer beträchtlichen Grössenordnung zu sein. Führt man – willkürlich, aber möglicherweise gerechtfertigt – zweimal ca. 2 IQ-Punkte der oben genannten Differenz im Gesamt-IQ auf Eichungsfehler zurück, verbleibt immer noch ein "IQ-Gewinn" von ca. 15 Punkten, also eine Leistungssteigerung im Gegenwert von einer Standardabweichung. Dies entspricht beispielsweise dem Leistungsunterschied zwischen einem 10jährigen und einem 12jährigen Durchschnittskind, oder zwischen einem Durchschnittskind und dem Grenzwert der Schweizerischen Invaliden-Versicherung für eine leichte geistige Behinderung, oder zwischen dem durchschnittlichen Schüler einer weiterführenden Schule und dem durchschnittlichen jugendlichen Erwerbstätigen ohne Berufslehre (Schallberger, 1988).

Eine Leistungssteigerung dieser Grössenordnung übertrifft wohl das, was man gemeinhin auf der Basis gewohnter Vorstellungen erwartet. Der Befund ist aber nicht mehr so erstaunlich, wenn man etwa die Uebersicht in Flynn (1987) betrachtet. Sie zeigt nämlich,

dass Studien mit verschiedenen Tests und mit zum Teil aussagekräftigeren Untersuchungstypen in mehreren andern Ländern der industrialisierten Welt Ergebnisse vergleichbarer Grössenordnung erbrachten. In dieselbe Richtung weist Flieller, Saintigny & Schaeffer (1986). Dies kann wohl als zusätzliche Absicherung der obigen Befunde angesehen werden.

Bei der Uebersicht von Flynn (1987) ist zwar eine gewisse Vorsicht am Platz; wie im Vorwort erwähnt und im Anhang erläutert sind zumindest bei der Widergabe der Befunde von Schallberger (1985) über Westdeutschland und die Schweiz Unkorrektheiten unterlaufen. Trotzdem folgen hier einige Hinweise:

Flynn (1987) argumentiert in seiner Ergebnisübersicht (a.a.O., Tabelle 15, S. 185) mit "IQ-Gewinnen" pro Jahr. Geht man – wie Flynn dies tut – davon aus, dass die oben gefundenen 20 Punkte im Gesamttest ganz als "IQ-Gewinn" zu betrachten sind, ergibt sich für die 27 Jahre zwischen ca. 1956 und ca. 1983 eine jährliche Zunahme von 0.74 Punkten; würde man mögliche Eichungsfehler berücksichtigen und wie oben mit ca. 15 Punkten rechnen, ergäbe dies 0.56 Punkte. Vergleicht man diese Werte mit Befunden mit Wechsler-Tests bei Kindern und Jugendlichen aus andern Ländern, so sind sie nicht extrem auffällig, liegen diese doch zwischen 0.835 (Japan) und 0.300 (USA, 2 bis 18jährige).

Die genannte Untersuchung von Flieller, Saintigny & Schaeffer (1986) betraf 8jährige Kinder aus Frankreich. Der verwendete Test (Test Mosaique) ist zwar von anderer Art als der HAWIK, enthält aber ebenfalls sowohl verbale als auch nicht-verbale Aufgaben. Die Periode der Untersuchung (Typ 2) umfasste 40 Jahre, der "IQ-Gewinn" 1.6 Standardabweichungen, was eine jährliche Zunahme von 0.6 IQ-Punkten ergibt.

Die internationale Befundlage macht auch eine früher vertretene Annahme (vgl. Schallberger, 1985, 1987) hinfällig: Gemäss der Logik der Untersuchungen des Typs 4 sagen die besprochenen Vergleichsuntersuchungen genau genommen nur etwas über die Population aus, aus denen die Eichstichproben stammen, also über das Gebiet von *Westdeutschland.* Aufgrund eines Vergleichs bloss mit Daten aus den USA, die für einen deutlich kleineren "IQ-Gewinn" sprachen, vermutete ich daher, es handle sich hier um ein spezifisch deutsches Phänomen, das mit der besonderen Nachkriegssituation dieses Landes zu tun habe. Dieser Rekurs auf eine besondere Situation in Westdeutschland scheint nun aber nicht notwendig zu sein. Dies bedeutet wohl auch, dass die hier dargestellten Ergebnisse zumindest der Tendenz nach auch auf die andern *deutschsprachigen Länder* übertragbar sind. Die Richtigkeit dieser Verallgemeinerung wäre allerdings nur dann überprüfbar, wenn für andere deutschsprachige Regionen einigermassen repräsentative HAWIK-Datensammlungen aus den fünfziger Jahren gefunden werden könnten. Diese würden erlauben abzuklären, ob die HAWIK-Eichung auch den durchschnittlichen Leistungsstandard von Kindern aus andern deutschsprachigen Regionen adäquat widerspiegelte. Zumindest für die Schweiz haben Bründler & Schallberger (1988) die analoge Frage für die HAWIK-R-Eichung ja bereits positiv beantworten können (siehe oben, Abschnitt 2.4).

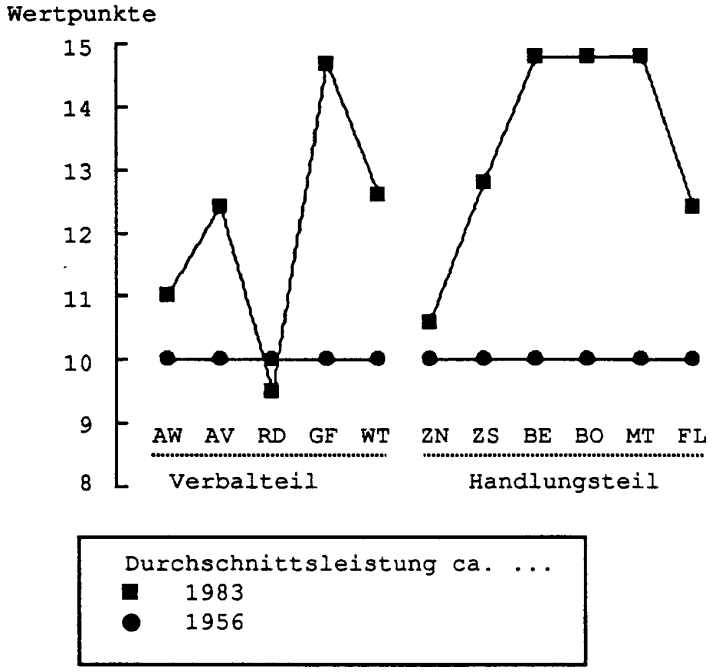
Um das Phänomen des "IQ-Gewinns" in seiner vollen Tragweite zu sehen, sind noch zwei Ergänzungen notwendig. Die eine betrifft den *differentiellen*, die andere den *dynamischen* Charakter des Phänomens:

Mit dem Attribut *"differentiell"* wird der Umstand angesprochen, dass die Leistungssteigerungen, die den "IQ-Gewinn" ausmachen, von Aufgabentyp zu Aufgabentyp variieren. Dies äussert sich bereits im Unterschied zwischen den Verhältnissen im Verbal- und im Handlungsteil. Während die IQ-Differenz im Verbalteil "nur" ca. 11 IQ-Punkte beträgt, findet sich im Handlungsteil eine solche von ca. 24 Punkten. Auch diese Diskrepanz zugunsten des Handlungsteils bzw. nonverbaler Tests scheint ein internationales Phänomen zu sein.

In allen von Flynn (1987) verarbeiteten Untersuchungen mit Wechsler-Tests, in denen Informationen zu Teil-IQs vorlagen, fiel der IQ-Gewinn im Handlungsteil z.T. leicht, z.T. deutlich höher aus als im Verbalteil (a.a.O., Tab. 17, S. 186). Analoge Befunde ergaben sich mit andern Tests: Sieben Datensammlungen mit verbal-mathematischen Tests ergaben "IQ-Gewinne" pro Jahr zwischen -0.133 und 0.628 (Median: .370); 13 Datensammlungen mit nonverbalen Tests (v.a. Ravens Progressive Matrizen) Werte zwischen 0.181 und 1.25 (Median: .588) (a.a.O., Tab. 15, S. 185). Operiert man mit den in Tab. 2 festgehaltenen gewichteten mittleren IQ-Differenzen von 11.4 (Verbalteil) und 24.3 (Handlungsteil) so ergeben sich "Jahreszuwachsraten" von .42 bzw. .90 – also keine aussergewöhnlichen Werte, vor allem, wenn man noch das über mögliche Eichungsfehler Gesagte berücksichtigt.

Man kann die Differenzierung aber auch noch über die Testteile hinaus weitertreiben und auf die Ebene der einzelnen Subtests gehen. Aus methodischen Gründen ist es hier allerdings unmöglich, alle sechs Vergleichsuntersuchungen aus Abschnitt 2.3 gleichzeitig zu verarbeiten. Die Abbildung 2, die daher allein auf den Daten von Schallberger (1987, Tabelle 4) beruht, kann daher nur als grober Anhaltspunkt dienen. Das horizontale "Profil" (= Effekt der Testnormierung) entspricht der Durchschnittsleistung von ca. 1956, das Zackenprofil – Effekt von differentiellen Leistungsveränderungen – der Durchschnittsleistung von ca. 1983. Die Abbildung macht augenscheinlich, dass sich hinter dem globalen "IQ-Gewinn", aber auch hinter dem Gewinn an Punkten im Verbal- und Handlungs-IQ, ein höchst differenziertes Muster an aufgabenspezifischen Leistungsverschiebungen ergibt. Bei einzelnen Aufgabentypen finden sich massive Leistungsverbesserungen, bei anderen praktisch keine oder sogar eine leichte Verschlechterung. Das ergibt einen letzten Grund dafür, warum der Name "IQ-Gewinn" – wie eingangs gesagt – im Grunde unglücklich ist: Der Name täuscht ein homogenes Phänomen vor, was offensichtlich nicht der Fall ist: Dahinter verbirgt sich vielmehr ein komplexes Muster an grossen und kleinen Leistungsverbesserungen, eventuell sogar Leistungsverschlechterungen.

Abb. 2: Die Leistungsverschiebungen auf der Ebene der Subtests (Erläuterungen im Text)



Versucht man die in Abbildung 2 sichtbaren – eventuell mit Eichungs- und Stichprobenfehlern kontaminierten – Leistungsverschiebungen zu interpretieren, ergibt sich etwa folgendes Bild der Hintergründe des gesamthaften "IQ-Gewinns":

- Es gibt keinen Aufgabentyp, bei dem eine substantielle Leistungs*abnahme* feststellbar wäre. Die Leistung der Kinder zu Beginn der achtziger Jahre ist in allen Aufgabentypen mindestens so gut wie jene der Kinder in den fünfziger Jahren. In diesem Rahmen sind aber doch drei Gruppen von Subtests zu unterscheiden:
- Praktisch gleich geblieben sind die Leistungen im Bereich der Merkfähigkeit (Zahlennachsprechen, "ZN") und ebenso in Aufgaben, die in Form von Sätzchenaufgaben das rechnerische Denken prüfen ("RD"). Interessanterweise trifft dasselbe auf allgemeine Wissensfragen ("AW", Allgemeines Wissen) zu.
- Mittlere Zunahmen – etwa in der Grössenordnung des Gesamttests – finden sich beim Test Allgemeines Verständnis ("AV"), wo es um das Verständnis von Problemen aus dem Alltag geht, und im Wortschatztest ("WT"). Ebenfalls eine mittlere Leistungszunahme zeigt sich beim Zahlen-Symbol-Test ("ZS"), einer psychomotorischen Aufgabe, deren Lösung auch in erheblichem Ausmass Konzentrationsfähigkeit voraussetzt, sowie im Test Figurenlegen ("FL"), einer Art Zusammensetzspiel.

- Die grössten Leistungszunahmen finden sich in den Aufgabentypen, bei denen es vor allem darauf ankommt, abstrakt-logisch zu denken, bzw. Wesentliches von Unwesentlichem zu unterscheiden. Ein Beispiel ist das Finden von Oberbegriffen zu zwei vorgegebenen Begriffen ("GF", Gemeinsamkeiten finden). Die anderen Vertreter dieser Gruppe finden sich im Handlungsteil: Beim Bilderergänzen ("BE") geht es darum, in schematischen Zeichnungen bekannter Gegenstände fehlende wichtige Teile zu identifizieren, beim Bilderordnen ("BO"), die einzelnen Zeichnungen einer Bildergeschichte in die richtige Reihenfolge zu bringen, und schliesslich beim Mosaik-Test ("MT") um das Nachbilden abstrakter Muster mit verschiedenfarbigen Klötzchen.

Entsprechend dem komplexen Charakter der Aufgabentypen fällt es nicht leicht, hinter diesen differentiellen Leistungsverschiebungen ein generelles Prinzip zu sehen. Immerhin fällt auf, dass Aufgaben, die explizit auch schulische Bildungsinhalte (Sprache, Rechnen) mitprüfen, im allgemeinen geringere Leistungszunahmen zeigen als Aufgaben, zu deren Lösung kein "Schulwissen" erforderlich ist. Auch dieser Sachverhalt findet sich in der Uebersicht von Flynn (1987) bestätigt, zeigen dort doch sogenannte "culturally reduced tests" (z.B. Ravens Progressive Matrizen) deutlich grössere Leistungszunahmen als verbale Tests.

Neben diesem differentiellen Aspekt des Phänomens des "IQ-Gewinns" ist schliesslich auch sein *dynamischer Charakter* zu beachten. Das damit Bezeichnete ist zwar mit den vorliegenden Daten nicht belegbar, aber wohl evident. Es geht um die Variabilität der Leistungsveränderungen, die hinter dem Phänomen des "IQ-Gewinns" stehen. Die hier berichteten Befunde beziehen sich auf die 27 Jahre der Periode von ca. 1956 bis ca. 1983. Unter der Voraussetzung, dass in dieser Periode ein "IQ-Gewinn" von ca. 15 bis 20 Punkten zu registrieren war, ergibt sich ein solcher von durchschnittlich 0.56 bis 0.74 Punkten pro Jahr. Ob die Leistung aber wirklich stetig von Jahr zu Jahr im gleichen Ausmass zunahm, ist natürlich unklar.[12] Es ist sogar unklar, ob es tatsächlich über die ganze Periode eine Leistungszunahme gab; es könnte ja auch sein, dass in irgendeinem Zeitpunkt zwischen 1956 und 1983 ein höherer "IQ-Gewinn" feststellbar gewesen wäre und dann eine Trendumkehr eingetreten ist, also ein "IQ-Verlust" stattfand. Und schliesslich ist auch unbekannt, was seit 1983 der Fall ist, ob die Testleistungen – gesamthaft oder in Teilleistungen – zunehmen oder abnehmen. Faktische Befunde über den "IQ-Gewinn" haben also ausgesprochen historischen Charakter und besagen nur etwas über eine spezifische Zeitspanne. Darüber hinaus lehren sie nur, dass sich Testleistungen innert relativ kurzer Zeitspannen von Geburtskohorte zu Geburtskohorte verändern *können*, und zwar in einem recht beträchtlichen Ausmass, das zudem von Aufgabentyp zu Aufgabentyp variiert.

---

12 Lynn & Hampson (1986) berichten zum Beispiel von Daten aus Japan, die für die jüngste Zeit eine Abnahme der dortigen "IQ-Gewinn"-Rate indizieren.

3. Einige intelligenzdiagnostische Folgeprobleme

Entsprechend seinem Ausmass stellt nun das Phänomen des "IQ-Gewinns" für die Intelligenzdiagnostik erhebliche Folgeprobleme. Zum Abschluss seien vier dieser Folgeprobleme illustriert. Sie betreffen zum einen die diagnostische Praxis (Punkt 1 und 2), zum andern die Verwendung von Intelligenztests in der Forschung (Punkt 3 und 4) und sind in Form von Leitsätzen formuliert. Um einem Missverständnis vorzubeugen: In den folgenden Punkten wird zwar immer wieder der HAWIK als Beispiel herangezogen. Die angesprochenen Probleme sind aber keineswegs auf den HAWIK beschränkt, sondern betreffen alle Intelligenztests – ja sogar alle Typen von Tests[13]. Die Probleme wären nur dann zu vermeiden, wenn alle (Intelligenz-)Tests regelmässig und in relativ kurzen Abständen (5 bis 10 Jahre) neu geeicht würden.

1. Die Interpretation von IQs älterer Intelligenztests ist ein riskantes Unterfangen. Dies ist ein naheliegender Schluss aus den mitgeteilten Befunden, dessen Begründung auf der Hand liegt: Je älter ein Test ist, d.h. je weiter seine Eichung zurückliegt, umso mehr muss angenommen werden, dass sich inzwischen gegenüber der Eichstichprobe Leistungsverschiebungen ergeben haben. Da gemäss dem über den dynamischen Charakter von Leistungsverschiebungen Gesagten ohne spezielle Untersuchungen nicht zu wissen ist, in welche Richtung und in welchem Ausmass solche Leistungsverschiebungen stattgefunden haben, weiss man bei einem älteren Test nie genau, was ein konkreter "IQ" eigentlich bedeutet.

Dabei ist noch folgendes anzumerken: Offenbar wurde in den letzten Jahrzehnten nicht in vollem Umfang realisiert, dass sich die in der diagnostischen Praxis gefundenen IQs in den letzten Jahrzehnten ständig erhöht haben. Sonst wäre das Erstaunen über die deutlich tieferen, aber realistischeren HAWIK-R- IQs nicht so gross gewesen. Das bedeutet mit aller Wahrscheinlichkeit auch, dass sich das Bezugssystem zur Beurteilung von IQs laufend verändert hat, und dass man nicht mehr gewohnt ist, mit tiefen IQs sachlich umzugehen. Tatsächlich kann man in der Praxis oft beobachten, dass ein IQ von 90 oft schon als auffällig beurteilt wird, ganz zu schweigen von einem IQ von 80. Dabei sind solche Werte gerade in einem kinderpsychiatrischen oder sonderpädagogischen Kontext von der statistischen Definition des IQs her recht oft zu erwarten (siehe dazu auch Schallberger, 1983).

In diesem Zusammenhang ist auch noch zu erwähnen, dass das Phänomen des "IQ-Gewinns" ein neues Argument gegen die Verwendung von IQs liefert, deren grosser

---

13 So berichten z.B. Bengel & Wittmann (1983) über analoge "Normverschiebungen" beim Freiburger Persönlichkeitsinventar (FPI).

Variationsbereich von ca. 100 Punkten (50 bis 150) ja in keinem Verhältnis zur Reliabilität der zugrundeliegenden Instrumente steht. Bereits kleine Leistungsveränderungen beeinflussen so differenzierte Testergebnisse merklich. Es wäre damit auch von daher viel sinnvoller, mit einem gröberen Normenraster zu arbeiten.

2. Die Interpretation von Profilen älterer Intelligenztests ist ebenfalls ein riskantes Unterfangen. Analoges gilt natürlich auch für die Interpretation von Profilen. In dieser Beziehung ist gerade beim HAWIK viel Anschauungsmaterial verfügbar. So wurden zum Beispiel die individuellen Differenzen von Handlungs- und Verbal-IQ als diagnostisch bedeutsamer Sachverhalt angesehen. Das dürfte 1956 auch tatsächlich der Fall gewesen sein und ist beim HAWIK-R mit seiner neuen Eichung auch tatsächlich wieder so. Beim HAWIK war eine solche Differenz in den letzten Jahrzehnten aber – wie oben gezeigt wurde – völlig belanglos. Da die Leistungssteigerung im Handlungsteil wesentlich grösser war, war ein Handlungs-IQ, der deutlich grösser ausfiel als der Verbal-IQ, der Normalfall. Extrem auffällig war aber, wenn die beiden IQs gleich gross ausfielen. Nur konnte man dies nicht wissen ...

Genau dasselbe gilt natürlich auch für das in Subtests differenzierte Profil: Das Profil des Durchschnittskindes von ca. 1983 (Abbildung 2) wies ganz auffällige Ausschläge auf, die im Einzelfall wohl zu Interpretationen verleitet haben. Dabei war dieses Profil belanglos, einfach das typische Profil eines Durchschnittskindes, entstanden durch die Leistungsverschiebungen seit der Eichung des Tests. Ein ausgewogenes Profil, das war der interpretationsbedürftige Sonderfall!

3. Klinische Forschung mit älteren Intelligenztests ohne Kontrollgruppen ist wertlos und irreführend. Die Einschränkung "Forschung mit älteren Intelligenztests" bedeutet nicht etwa, dass andere Forschungstypen problemlos ohne Kontrollgruppe auskommen, sondern nur, dass wir uns hier auf Implikationen des Phänomens des "IQ-Gewinns" beschränken. Auch die Begründung dieser These wurde schon gegeben: Bei älteren Tests weiss man nie genau, ob ein IQ oder ein Profil irgendetwas für den untersuchten Einzelfall oder die untersuchte klinische Gruppe Spezifisches charakterisiert oder aber einfach ein Effekt allgemeiner Leistungsverschiebungen darstellt. Ob das eine oder andere der Fall ist, kann nur mittels einer guten Kontrollgruppe festgestellt werden.

Das Gemeinte kann sehr schön am sogenannten "Problemkinderprofil" illustriert werden: In Untersuchungen an Legasthenikern, Kindern mit Lern- und Verhaltensstörungen und an Kindern mit leichter frühkindlicher Hirnschädigung zeigte sich immer wieder ein ähnliches HAWIK-Profil (siehe Schneider, 1982, bzw. die Zusammenstellung in Schallberger, 1983). Es wurde wie folgt beschrieben: Eine M-Form mit tiefgestelltem rechten Bein im Verbalteil und ein asymmetrischer "Berg" im Handlungsteil. Die wiederholte Entdeckung dieser Profilform verhinderte weitgehend, dass sie als differential-diagnostisches Kennzeichen für

eine der genannten Störungen vorgeschlagen wurde. Da es sich aber bei allen untersuchten Gruppen um "Problemkinder" handelte, betrachtete man sie als eine Art "Problemkinderkurve". Der Vergleich mit den Ergebnissen der Zürcher Eichung (Schallberger, Gysin et al., 1981) zeigte dann aber, dass diese Interpretation nicht zutraf; die Problemkinderkurve entspricht vielmehr weitgehend dem Durchschnittsprofil "normaler" Kinder, das durch Leistungsverschiebungen entstanden ist. Hätte man bei nur einer der genannten Untersuchungen an Problemkindern eine Kontrollgruppe verwendet, wäre dies schon früher aufgefallen.

4. Bei Untersuchungen, in denen verschiedene Intelligenztests angewendet werden, ist bei der Interpretation der Unterschiede genauestens auf das Eichdatum zu achten: Auch der Grund für dieses Postulat dürfte offensichtlich sein: Je älter ein Test ist, umso eher muss man mit Leistungsverschiebungen rechnen, die die gefundenen Testergebnisse uninterpretierbar machen, je jünger der Test ist, umso eher kann man den Testergebnissen trauen. Wendet man nun in einer Untersuchung verschiedene Tests an und interpretiert man Unterschiede und Uebereinstimmungen ohne auf das Eichdatum zu achten, begibt man sich leicht auf Glatteis. Flynn (1984) hat einige viel zitierte amerikanische Untersuchungen unter diesem Aspekt neu unter die Lupe genommen; zum Teil hätten sich die inhaltlichen Schlussfolgerungen genau umgekehrt, wären die Eichdaten berücksichtigt worden.

Eine besondere heimtückische Variante dieser Falle des Phänomens des IQ-Gewinns stellen Längsschnittstudien an Kindern dar, in denen man zum Teil gezwungen sein kann, bei einer bestimmten Altersstufe das Testinstrument zu wechseln. Stammen die Eichungen der beiden sich ablösenden Tests nicht ungefähr aus demselben Zeitraum, kann man damit höchst merkwürdige Befunde produzieren. Analoge Probleme ergeben sich etwa beim Vergleich der Testergebnisse von Kindern und Eltern oder zwischen Geschwistern.

5. Zum Problem der Erklärung

Natürlich stellt sich nun noch das Problem der *Erklärung* des Phänomens des "IQ-Gewinns". Diese Frage hat zwei Hauptkomponenten, nämlich zum einen die Frage nach den Ursachen der empirisch festgestellten Leistungszunahme, zum andern aber auch die Frage, was diese Leistungszunahmen im Zusammenhang mit dem Konstrukt "Intelligenz" bedeuten. Diesen beiden Fragen sei hier noch in aller Kürze nachgegangen. Ausführlichere, in wesentlichen Punkten divergierende Diskussionen finden sich vor allem bei Flynn (1987) und bei Flieller (1989).

Zur ersten Teilfrage: Ueber die *Ursachen* der phänomenalen, weltweit zu beobachtenden Leistungssteigerungen in Intelligenztests kann im heutigen Zeitpunkt nur spekuliert werden, da wirklich stichhaltige empirische Befunde weitgehend fehlen. Im wesentlichen scheint es aber vier mögliche Erklärungen zu geben, die sich allerdings gegenseitig nicht ausschliessen.

Es ist sogar zu vermuten, dass diese vier Erklärungen alle gleichzeitig eine gewisse, wenn auch beschränkte Gültigkeit beanspruchen können und dass sie erst zusammen eine wirklich hinreichende Erklärung ergeben:

1.) Eine erste Erklärung bezieht sich gewissermassen auf Aeusserlichkeiten, nämlich die Zunahme der Testgewandtheit oder auf grössere Vertrautheit mit testartigen Aufgaben. Dies mag sicher eine Rolle spielen,[14] erklärt aber kaum die volle Differenz, wie Flieller (1989) durch eine Rasch-Analyse zumindest nahezulegen vermag. Auf einem solchen Hintergrund wäre beispielsweise auch nicht einzusehen, warum es so grosse und systematische Unterschiede zwischen den verschiedenen Aufgabentypen gibt. Weiter müsste man einen gewissen "Plateau-Effekt" erwarten; wie Trainingsstudien zeigen, ist der überhaupt mögliche Effekt steigender Testgewandtheit oder -vertrautheit in seiner Grösse begrenzt.

In dieselbe Kategorie gehört die Vermutung, der Antwortstil habe sich verändert, die Bereitschaft zu raten habe zugenommen, woraus dann per Zufall höhere Punktzahlen entstehen. Brand, Freshwater & Dockrell (1989) vertreten die Auffassung, massive "IQ-Gewinne" seien daher nur in Multiple-Choice-Tests zu beobachten – eine Auffassung, die durch die vorliegende Studie widerlegt wird. Es finden sich hier auch grosse Leistungszunahmen in Subtests, deren Aufgaben keine Verwandtschaft mit Multiple-Choice-Aufgaben aufweisen (siehe zu jener Vermutung auch Flieller et al., 1989).

2.) Eine zweite Erklärung beruft sich auf sozioökonomische Veränderungen in modernen Gesellschaften. Es ist ja schon lange bekannt, dass deutliche Leistungsunterschiede zwischen Kindern aus verschiedenen sozialen Schichten festzustellen sind. Nun haben sich aber in den letzten Jahrzehnten die Lebensbedingungen vor allem in den unteren Schichten gewaltig verbessert, ebenso das Ausmass der Teilhabe an Kultur und Bildung (vgl. Beck, 1986). Durch diese sozioökonomischen Verschiebungen sei auch – so lautet das Argument – das durchschnittliche intellektuelle Leistungsniveau in der Gesamtheit der Bevölkerung angestiegen.

Auch diese Erklärung reicht nicht aus. In Studien, wo dies kontrolliert werden kann, reichen die Veränderungen der sozioökonomischen Struktur bei weitem nicht aus, um die gefundenen "IQ-Gewinne" zu erklären (Flieller et al., 1986; Flynn, 1987). Zudem ist in allen Schichten und Bildungsgruppen ein "IQ-Gewinn" feststellbar.[15] Beim Motor dieses Phänomens muss es sich also um etwas handeln, das alle Bevölkerungskreise betrifft und nicht nur einen Teil.

---

14 Eine Bestätigung findet sich in den sogenannten Uebungs- oder Lerneffekten bei Testwiederholungen, von denen in Abschnitt 2.3 die Rede war.

15 Es gibt aber Anzeichen dafür, dass der "IQ-Gewinn" je nach Schicht variiert. So fanden Flieller et al. (1986) für den Zeitraum 1944 bis 1984 bei Kindern von Landwirten grössere "IQ-Gewinne" als in allen anderen Herkunftsgruppen.

3.) Eine dritte Erklärung verweist im Zusammenhang mit einem solchen generellen Faktor auf das Schulsystem (inkl. Kindergarten); dieses sei besser geworden, bewirke eine breitere Förderung der intellektuellen Leistungsfähigkeit. Auch hierzu gibt es empirische Hinweise (siehe z.B. Frey, 1977). Ein Problem bei dieser Erklärung ist aber, dass sie nicht gut verstehen lässt, warum – wie im 2. Abschnitt hervorgehoben wurde – die explizit schulabhängigen Leistungen deutlich weniger angestiegen sind als jene, die kein Vorwissen dieser Art erfordern. Flynn (1984) stellt für die USA sogar fest, dass es ein Kontinuum zwischen Leistungszunahme und -abnahme gibt, wenn man die Aufgabentypen nach ihrer Bildungsabhängigkeit ordnet: Die grössten Zunahmen ergeben sich bei sprachfreien, von expliziten Bildungsinhalten völlig unabhängigen Aufgabentypen. Hingegen scheinen in den USA Aufgabentypen, deren Lösung Bildungsinhalte voraussetzen, sogar Abnahmen zu zeigen.

4.) Es ist also auch noch eine vierte Erklärung notwendig, nämlich die, dass sich in den letzten Jahrzehnten in den alltäglichen Lebensbedingungen, in den alltäglichen Entwicklungsbedingungen etwas massiv verändert haben muss, etwas, was die Entwicklung der intellektuellen Leistungen wesentlich stärker stimuliert als in früheren Generationen. Mögliche Konkretisierungen für diese Pauschalaussage findet man natürlich rasch auf allen Ebenen, von den Spielsachen über die Medien bis hin zu den vielfältigen kognitiven Reizen der modernen Lebenswelt, die fast permanent wirksam sind; bei einer Eisenbahn- oder Autofahrt, beim Gang durch eine Einkaufsstrasse, etc. Die Rolle dieser basalen und allgegenwärtigen Stimulierung, dieses im Unterschied zu bewusst konzipierten Erziehungs- und Bildungsmassnahmen eigentlich "nicht-intendierten" Trainings, scheint bisher unterschätzt worden zu sein.[16]

Wie weiter oben bereits erwähnt wurde, handelt es sich bei diesen vier Erklärungen um plausible Vermutungen, für die auch empirische Indizien vorliegen. Um ihre relative Bedeutung besser abschätzen zu können, ist aber noch intensive empirische Forschung notwendig. Sehr nützlich wären dabei ältere Datensammlungen, denen zeitgenössische gegenübergestellt werden könnten. Im deutschen Sprachraum scheinen aber solche Datensammlungen eher rar zu sein.

Die andere Frage, die – wie angekündigt – zum Abschluss noch gestreift werden soll, betrifft *die Bedeutung der Befunde zum "IQ-Gewinn" für die Intelligenztheorie.* Unter diesem Aspekt drängen sich vor allem einige Paradoxien in den Blick, die intelligenztheoretisch nicht einfach zu lösende Probleme stellen:

---

16 Der Vollständigkeit halber sei hier noch auf Lynn (1990) hingewiesen. Dieser Autor interpretiert den "IQ-Gewinn" als Effekt besserer Ernährung und der dadurch bewirkten Verbesserung der neurologischen Entwicklung und Funktionstüchtigkeit des Gehirns.

Eine erste dieser Paradoxien ergibt sich aus dem Vergleich der Befunde zu Generationenunterschieden mit jenen zur Erblichkeit. Wie sämtliche verfügbaren Studien zeigen, weist das Merkmal "Intelligenz", operationalisiert über Intelligenztests, eine sehr hohe Heritabilität auf. Generationenunterschiede im beschriebenen Ausmass wirken aus dieser Sicht sehr irritierend; es wäre ja offensichtlich nicht sinnvoll, sie durch eine Veränderung des genetischen Potentials erklären zu wollen. Das Paradox ist allerdings nicht auf Intelligenz beschränkt: Auch das Merkmal Körpergrösse zeigt ja beispielsweise gleichzeitig hohe Erblichkeit *und* grosse Kohortenunterschiede.

Eine zweite Paradoxie ergibt sich aus dem differentiellen Charakter der Leistungszunahmen: Gemäss der traditionellen Auffassung hält man Testaufgaben, die mit Bildungsinhalten gesättigt sind ("kristallisierte Intelligenz"), für stärker kulturabhängig als rein logisch-sprachfreie Aufgaben ("flüssige Intelligenz"), die geradezu als kultur-frei, oder zumindest "culture-fair" oder "culturally reduced" angesehen wurden. Die oben dargestellten Befunde weisen aber – zumindest in bezug auf den Vergleich verschiedener Generationen – in eine genau gegenteilige Richtung.

Ein letztes hier erwähntes Paradox ergibt sich schliesslich aus der Frage, ob die hier besprochenen Leistungszunahmen in Intelligenztests etwas mit Intelligenzzunahmen zu tun haben – eine Frage, die bisher ja bewusst ausgeklammert worden ist. Die Antwort auf diese Frage hängt offensichtlich vom zugrundegelegten Intelligenzbegriff ab. Geht man von der bekannten "operationalen Definition" aus – Intelligenz ist das, was die Intelligenztests messen –, dann wäre die Antwort natürlich "ja": Genauso wie die Kinder in den letzten Jahrzehnten grösser geworden sind, oder die modernen Läufer schneller, wären auch die jüngeren Generationen "intelligenter" geworden. Nun geht aber ein solches Verständnis des Intelligenzbegriffs gänzlich an den Auffassungen vorbei, die im Alltag und auch in der Wissenschaft damit üblicherweise verbunden werden. Nach diesen Auffassungen handelt es sich bei "Intelligenz" nicht um einen Beobachtungsbegriff wie z.B. Grösse oder Schnelligkeit, deren Ausprägungen sich am Beobachtbaren "ablesen" lassen, sondern um einen Dispositionsbegriff, d.h. um einen Begriff, der etwas bezeichnet, das sich zwar in beobachtbaren Sachverhalten manifestiert, jedoch nicht damit identisch ist. Konkreter: Intelligenz bezeichnet nach üblichem Verständnis nicht das Intelligenztestverhalten an sich, sondern eine Disposition, die – neben anderem – die Qualität der Intelligenztestleistungen (mit-)bestimmt. Nun scheint die plausibelste Deutung des Phänomens des "IQ-Gewinns" (oder "IQ-Verlusts") in einem Rekurs auf variierende Trainingsbedingungen für dieses Leistungsverhalten zu bestehen (siehe oben). Ob man aber solche Effekte des Trainings auch zur "Disposition" rechnen darf oder soll, ist letztlich eine begriffliche und nicht eine empirische Frage. Rein intuitiv scheint es wenig sinnvoll zu sein; das würde dann bedeuten, dass "IQ-Gewinne" und "IQ-Verluste" nichts mit Intelligenzgewinn oder -verlust zu tun haben, die Antwort auf unsere Frage also "nein" lauten müsste. Leider kommt man aber

auch mit dieser Lösung in grosse Schwierigkeiten: Ein Intelligenzbegriff, der auf alle Trainingseffekte verzichtet, degeneriert nämlich zum Mythos, denn Kinder, die ohne jede intellektuelle Stimulierung oder (spezifisch menschliches) intellektuelles Training aufwachsen – oft "wilde Kinder" genannt – zeigen nämlich alle Zeichen extremen Schwachsinns (Zimmer, 1989). Muss man also verschiedene Typen von Training unterscheiden? Und wo ist der qualitative Sprung zwischen diesen Typen? Die Frage nach der Bedeutung des IQ-Gewinns und IQ-Verlusts für die Frage nach einem Intelligenzgewinn oder -verlust führt also wieder einmal vor Augen, wie komplex der scheinbar so "handliche" Intelligenzbegriff eigentlich ist!

Schlussbemerkungen

Im Zentrum des hier vorgelegten Papiers stand der Versuch, die gegenwärtige Befundlage zum HAWIK und zum HAWIK-R im Hinblick auf die Frage nach dem Ausmass des "IQ-Gewinns" zwischen den beiden Testeichungen auszuwerten. In Anbetracht der Ungewissheiten, die dabei in Kauf genommen werden mussten, taucht vielleicht die Frage auf, ob sich dieses Unterfangen und der dabei betriebene Aufwand überhaupt gelohnt hat. Solange nicht mehr wirklich aussagekräftige Vergleichsuntersuchungen vorliegen und solange nicht mehr Informationen zur Validität der HAWIK-Eichung zur Verfügung stehen – lohnt sich da überhaupt eine derart akribische Diskussion der vorhandenen Daten, wie sie hier vorgenommen wurde? Dazu möchte ich folgende zwei Punkte festhalten:

1.) Wie im Vorwort schon erwähnt, stammen viele der hier wiedergegebenen Ueberlegungen und Argumentationen aus dem Jahre 1985. Es ist kein Zufall, dass sie damals nicht zu einer definitiven Publikation ausgearbeitet wurden. Denn ich erwartete, dass bald neue Vergleichsuntersuchungen zu einer Klärung der Situation führen würden. Ferner suchte ich nach weiteren Informationen zur HAWIK- und zur HAWIK-R-Eichung. Die Ausbeute war geringer als erwartet. Die neuen Untersuchungen wie jene von Boehnke (1986), Bründler & Schallberger (1988), Gutezeit (1989) und Titze (1989) stützten aber die Schlussfolgerungen, die sich bereits 1985 aufgedrängt hatten, was als eine zusätzliche Absicherung betrachtet werden kann.

2.) Kohortenunterschiede in Intelligenztestleistungen des zur Diskussion stehenden Ausmasses scheinen mir ein sehr bemerkenswertes Phänomen zu sein. Sie provozieren eine Fülle interessanter Fragen kognitionspsychologischer, entwicklungspsychologischer, ökopsychologischer und epochalpsychologischer Natur. Dazu kommt die Relevanz für den Einsatz von Tests in Diagnostik und Forschung (siehe oben, Abschnitt 3). Und schliesslich können diesbezügliche Informationen auch einen Beitrag zur aktuellen Diskussion um Generationenunterschiede in Schulleistungen und Schulverhalten bilden (vgl. z.B. Schallberger, 1990). Auf dem Hintergrund fehlender anderer Erkenntnisgrundlagen scheint

es daher sinnvoll zu sein, auch suboptimale Daten und unvollständige Informationen zu verwerten. Dies geschieht in diesem Text, und zwar – wie ich hoffe – mit der gebührenden Vorsicht. Das Ganze bleibt aber natürlich eine Art "Indizienbeweis".

Literaturverzeichnis

Amelang, M. (1966) Zum Vergleich des Begabungs-Test-Systems mit dem HAWIK und dem Stanford-Intelligenz-Test-Lückert. *Psychologische Beiträge, 9*, 525-535.

Beck, U. (1986) *Risikogesellschaft. Auf dem Weg in eine andere Moderne*. Frankfurt a. M.: Suhrkamp.

Bengel, J. & Wittmann, W. (1983) Urteilsfehler aufgrund zeitlicher Veränderungen von Testnormen am Beispiel des FPI. *Diagnostica, 29*, 101-117.

Boehnke, K. (1986) Probleme der Intelligenzmessung bei Kindern mit dem HAWIK-R. *Praxis der Kinderpsychologie und Kinderpsychiatrie, 35*, 34-41.

Brand, C.R., Freshwater, S. & Dockrell, W. B. (1989) Has there been a "massive" rise in IQ levels in the West? Evidence from Scottish children. *Irish Journal of Psychology, 10*, 388-393.

Bründler, M. & Schallberger, U. (1988) *HAWIK-R. Ergänzungsband für die deutschsprachige Schweiz*. Bern: Huber.

Editorial note. (1987) Correction to Flynn. *Psychological Bulletin, 101*, 427.

Eggert, D., Liman, E. & Schirmacher, A. (1984) Vergleich des Hamburg-Wechsler-Intelligenztests für Kinder (HAWIK) mit der revidierten Version (HAWIK-R) bei sprachbehinderten Kindern. *Zeitschrift für Heilpädagogik, 35*, 54-58.

Flieller, A. (1989) Application du modèle de Rasch à un problème de comparaison de générations. *Bulletin de psychologie, XLII*, 86-91.

Flieller, A. (1989) Les comparaisons de cohortes et de générations dans l'étude psychométrique de l'intelligence. *Psychologie scolaire, 68*, 47-64.

Flieller, A., Jautz, M. & Kop, J.-L. (1989) Les réponses au test Mosaïque à quarante ans d'intervalle. *Enfance, 42*, 7-22.

Flieller, A., Saintigny, N. & Schaeffer, R. (1986) L'évolution du niveau intellectuel des enfants de 8 ans sur une période de 40 ans (1944-1984). *L'orientation scolaire et professionnelle, 15*, 61-83.

Flynn, J. R. (1984) The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29 - 51.

Flynn, J. R. (1987) Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191 (siehe dazu auch: Editorial note, 1987).

Frey, H. (1977) Einfluss von herkömmlichem und neuem Mathematikunterricht auf Intelligenztestleistungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 9*, 185-191.

Gutezeit, G. (1989) Ergebnisse mit der Revision des Hamburg-Wechsler-Intelligenztests für Kinder von 1983 (HAWIK-R) bei Schülern mit schweren Lesestörungen. *Zeitschrift für Kinder- und Jugendpsychiatrie, 17*, 70-78.

Hardesty, F. P. & Priester, H. J. (1956) *Handbuch für den Hamburg-Wechsler-Intelligenztests für Kinder* (3. Auflage 1966). Bern: Huber.

Horn, W. (1969) *Prüfsystem für Schul- und Bildungsberatung (PSB)*. Göttingen: Hogrefe.

Jäger, R. S. (1984) Hamburg-Wechsler-Intelligenztest für Kinder Revision 1983 (HAWIK-R). Testbesprechung. *Zeitschrift für Differentielle und Diagnostische Psychologie, 5*, 247-249.

Koop, J. & Röttger, K. (1989) Der HAWIK-R im Vergleich zu anderen Intelligenztests. *Diagnostica, 35*, 201-210.

Kornmann, R. & Müller-Edenborn, E. (1975) Normverschiebung oder diagnostisch relevante Testprofile beim HAWIK? *Praxis der Kinderpsychologie und Kinderpsychiatrie, 24,* 48-51.

Kurth, E. (1970) Erhöhung der Leistungsnormen bei den Farbigen Progressiven Matrizen. *Zeitschrift für Psychologie, 77,* 85-90.

Lückert, H. R. (1957) *Stanford-Intelligenztest.* Göttingen: Hogrefe.

Lynn, R. (1990) The role of nutrition in secular increases in intelligence. *Personality and Individual Differences, 11,* 273-285.

Lynn, R. & Hampson, S. (1986) The rise of national intelligence: Evidence from Britain, Japan and the USA. *Personality and Individual Differences, 7,* 23-32.

Mayer, W. (1983) *Der Vergleich zweier HAWIK-Versionen bei lernbehinderten Sonderschülern.* Unveröffentlichte Dissertation, Medizinische Hochschule Hannover. (zitiert nach Tewes & Titze 1983).

Meckel, H. (1977) Kritische Darstellung empirisch gewonnener HAWIK-Daten im Vergleich zur Standardisierung. *Psychologie in Erziehung und Unterricht, 24,* 20-27.

Norden, I. (1953) *Das Binetarium.* Göttingen: Hogrefe.

Pfeiffer, G. & Zielinski, W. (1976) Messen HAWIK und PSB das Gleiche? *Diagnostica, 22,* 8-13.

Priester, H. J. (1958) *Die Standardisierung des Hamburg-Wechsler-Intelligenztests für Kinder.* Bern: Huber.

Priester, H. J. (1959) Untersuchungen zum Vergleich zwischen Stanford-Intelligenztest (SIT) und Hamburg-Wechsler-Intelligenztest für Kinder (HAWIK). *Diagnostica, 5,* 49-54.

Priester, H. J. & Kukulka, R. (1958) Vergleichsuntersuchungen zum Hamburg-Wechsler-Intelligenztest für Kinder (HAWIK) und Binet-Bobertag und zum HAWIK und dem Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE) im Bezug auf die Intelligenzquotienten und die Benutzung dieser Tests als Paralleltests. *Diagnostica, 4,* 6-16.

Royl, W. & Schwarzer, R. (1976) Zur säkularen Akzeleration der PSB-Intelligenz. *Diagnostica, 22,* 99-109.

Schallberger, U. (1983) *Ergänzende Untersuchungen zu den Zürcher HAWIK-Normen aufgrund von Testprotokollen aus schulpsychologischen Diensten.* (Sind die Zürcher Normen zu streng? Gibt es ein Problemkinderprofil?) Berichte aus der Abteilung Angewandte Psychologie, Nr. 17. Zürich: Psychologisches Institut der Universität.

Schallberger, U. (1985) *HAWIK und HAWIK-R: Ein empirischer Vergleich.* Oder: Ueber einen Fall phänomenaler Leistungsakzeleration und die Problematik von Intelligenztestnormen. Unveröffentlichtes Manuskript. Zürich: Psychologisches Institut der Universität.

Schallberger, U. (1987) *HAWIK und HAWIK-R: Ein empirischer Vergleich.* Diagnostica, 33, 1-13.

Schallberger, U. (1988) Berufsausbildung und Intelligenzentwicklung. In: Häfeli, K., Kraft, U. & Schallberger, U.: *Berufsausbildung und Persönlichkeitsentwicklung* (S. 148-167). Bern: Huber.

Schallberger, U. (1990) Was Intelligenztests über Generationenunterschiede lehren. *Schweizer Schule, 77,* Nr. 10, S. 3 - 8.

Schallberger, U. (im Druck) Benachteiligt der HAWIK-R leistungsschwache Kinder? *Diagnostica, 37.*

Schallberger, U., Gysin, H.R., Mattes, W. & Sigrist, M. (1981) *Anwendung des HAWIK bei Deutschschweizer Kindern.* Bern: Huber.

Schallberger, U., Vetsch, M., Orth, P., Schädelin, I. & Wiederkehr, J. (1981) *Schweizer Eichung des PSB Horn.* Schriftenreihe des Schweiz. Verbandes für Berufsberatung Nr. 39. Zürich: Verlag des SVB.

Schneider, R. (1982) Gibt es eine "Problemkinderkurve" im HAWIK? *Praxis der Kinderpsychologie und Kinderpsychiatrie, 31,* 286-290.

Schubert, M. Th. & Berlach, G. (1982) Neue Richtlinien zur Interpretation des Hamburg-Wechsler-Intelligenztests für Kinder (HAWIK). *Zeitschrift für Klinische Psychologie, 11,* 253-279.

Swerdlik, M. E. (1977) The question of the comparability of the WISC and WISC-R: Review of the research and implications for schoolpsychologists. *Psychology in the Schools, 14,* 260-270.

Tewes, U. (1983) *HAWIK-R. Handbuch und Testanweisung.* Bern: Huber.

Tewes, U. & Titze, I. (1983) Untersuchungen zur Anwendung des HAWIK in der klinischen und sonderpädagogischen Diagnostik. *Zeitschrift für Differentielle und Diagnostische Psychologie, 4,* 179-201.

Titze, I. (1989) Ueber den Zusammenhang von Persönlichkeitsmerkmalen und Intelligenz bei Kindern. *Zeitschrift für Differentielle und Diagnostische Psychologie, 10,* 91-101.

Titze, I. & Tewes, U. (1984) *Messung der Intelligenz bei Kindern mit dem HAWIK-R.* Bern: Huber.

Wechsler, D. (1974) *Manual for the Wechsler Intelligence Scale for Children - Revised.* New York: Psychological Corporation.

Zimmer, D. E. (1989) *Experimente des Lebens. Wilde Kinder, Zwillinge, Kibbuzniks und andere aufschlussreiche Wesen.* Zürich: Haffmans.

Anhang: Zum Abschnitt "West Germany and Switzerland" in Flynn (1987)

Flynns Darstellung im Abschnitt "West Germany and Switzerland" des Artikels "Massive IQ gains in 14 nations" (Flynn, 1987, S. 182f.) basiert weitgehend auf Schallberger (1985). Diese Darstellung enthält verschiedene Unrichtigkeiten mehr oder weniger gravierender Natur, die im folgenden *als Ergänzung zu Flynn (1987)* zusammengestellt seien.

Flynns Darstellung meiner Befunde erfolgte entlang einer zusammenfassenden Tabelle, die hier daher zuerst wiedergegeben sei:

Table 12
Four Samples from West Germany and Switzerland: Years Tested and Levels of Performance
(aus Flynn, 1987, S. 182):

Sample	Age (years)	Year	IQ Verbal	IQ Performance	IQ Full scale
West German, WISC	7-15[a]	1954[b]	100	100	100
Zurich, WISC	9 and 12	1977	108	120	115
West German, WISC-R	7-15[a]	1981[b]	111[c]	125[c]	120[c]
Solothurn, WISC	8-9	1984	107	122	116

Note: Data from Schallberger (1985). WISC=Wechsler Intelligence Scale for Children. WISC-R = WISC-Revised.
[a] At least half of the 257 subjects used to compare the West German WISC and WISC-R samples were 7-9-years old. [b] Estimated as tested 2 years before publication. [c] Estimates of how the WISC-R sample would have performed against the West German WISC norms.

Die Zahlen in dieser Tabelle haben folgenden Hintergrund:

Die erste Zeile *(West German, WISC)* repräsentiert die 1956 publizierte HAWIK-Eichung. Flynn nimmt dabei – wie auch beim HAWIK-R – an, dass die Eichung zwei Jahre vor der Publikation erfolgte, daher die Angabe "1954" (bzw. "1981" in der dritten Zeile).

Die zweite Zeile *(Zurich, WISC)* gibt die Ergebnisse der Zürcher Eichung des HAWIK (Schallberger, Gysin et al., 1981; Datenaufnahme: 1977) wieder, allerdings nur jene der 9-jährigen Kinder und nicht wie angegeben der 9 und 12-jährigen. Die Ergebnisse der 12-jährigen Kinder lauteten übrigens: Verbal-IQ: 110; Handlungs-IQ: 114; Gesamt-IQ: 114.

Die dritte Zeile *(West-German, WISC-R)* basiert auf einem Analogon zu Tabelle 2 dieses Textes in Schallberger (1985), die naturgemäss die Studien von Boehnke (1986) und Gutezeit (1989) noch nicht enthielt. Die dritte Zeile enthält somit die gewichteten Mittelwerte aus den Zeilen (1a), (2a), (3) und (5) aus Tabelle 2 dieses Berichts. Zu diesen Mittelwerten wurde von Flynn 100 addiert, um den Charakter als Mass für den "IQ-Gewinn" deutlicher zu machen.

Die vierte Zeile schliesslich *(Solothurn, WISC)* betrifft die HAWIK-Ergebnisse aus der eigenen Vergleichsuntersuchung von HAWIK und HAWIK-R, die in Schallberger (1987) publiziert worden ist (und die auch in der Zeile "West German, WISC-R" enthalten ist). Es handelt sich dabei um eine Schätzung der Durchschnitt-IQs im HAWIK, wenn man die (geschätzten) Auswirkungen von Lerneffekten subtrahiert.

Flynns verbale Darstellung dieser Befunde ist in folgenden Punkten zu korrigieren:

1.) Im dritten Absatz des Unterabschnitts "Samples and tests" wird die Solothurner Studie als reine HAWIK-Untersuchung beschrieben. Effektiv war die Solothurner Studie – wie oben erwähnt – eine Vergleichsstudie von HAWIK und HAWIK-R. Sie ist identisch mit jener vierten Studie, die im ersten Absatz des Unterabschnitts "Samples and tests" genannt wird.

Im selben Absatz wird gesagt, die Zürcher Eichung basiere auf einer Stichprobe "... deemed representative based on test scores". Effektiv basierte diese Stichprobe auf einer Analyse der sozioökonomischen Struktur des Kantons Zürich (vgl. Schallberger, Gysin et al., 1981).

3.) Im Unterabschnitt "Results" finden sich die Aussagen:

> As for Switzerland, Schallberger believes that the fact the 1977 Zurich sample outscored the 1954 West German WISC sample argues for a massive Swiss gain of 15 points over that period. This assumes of course that Swiss and West German children were roughly equal in 1954. The samples internal to Switzerland show 1984 Solothurn only 1 point above 1977 Zurich.

Und am Anfang des Unterabschnitts "Assessment" wird gesagt:

> The case for Swiss IQ gains rests on the assumption that Swiss and West German children were roughly equal in 1954. However, strong data from elsewhere ... [usw., total 14 Zeilen, die sich mit jener "assumption" befassen, der Verf.]. The estimate of little gain after 1977 is no better. To compare samples from Zurich to Solothurn is rather like comparing Boston with rural Ohio ... [usw., total 6 Zeilen, der Verf.].

Flynn kritisiert hier eine Strohpuppe, denn ich habe nie versucht, einen Schweizer "IQ-Gewinn" abzuschätzen, und zwar aus den Gründen, die Flynn selber angibt. Ebenso habe ich die Differenz von einem Punkt zwischen Zürich und Solothurn nie als Indikator für einen geringen "IQ-Gewinn" von 1977 bis 1984 gewertet, wiederum aus denselben Gründen, auf die sich Flynn bezieht (auch wenn der Vergleich des Unterschieds zwischen den Kantonen Zürich und Solothurn mit jenem zwischen Boston und "rural Ohio" wohl etwas übertrieben ist).

Die Tatsache, dass Flynn die Zeilen "Zurich" und "Solothurn" in seine Tabelle 15 aufnahm, hat also nichts mit meinen Ueberlegungen zu tun.

3.) Im Unterabschnitt "Results" (zweitletzter Absatz) wird im Zusammenhang mit dem sonst korrekt wiedergegebenen Ergebnis der Vergleichsuntersuchungen von HAWIK und HAWIK-R mit keinem Wort auf die verfügbaren Indizien für Eichungsfehler eingegangen.

Nachdem ich diese Punkte Flynn mit Schreiben vom 2. März 1987 mitgeteilt hatte, erschien in der folgenden Nummer des Psychological Bulletin folgende Richtigstellung (Editorial note, 1987):

> **Correction to Flynn**
>
> In the article "Massive IQ Gains in 14 Nations: What IQ Tests Really Measure", by James R. Flynn (Psychological Bulletin, 1987, Vol. 101, No. 2, pp. 171-191), it was wrongly stated that U. Schallberger put forward a hypothesis or hypotheses concerning the magnitude of Swiss IQ gains; in fact, he did not. Flynn wishes to acknowledge this error and stress that his critique of these hypotheses does not apply to Schallberger. Schallberger had earlier come to the same conclusion as Flynn, namely, that the Swiss have probably made sizable IQ gains, but all estimates of magnitude would have only speculative status.

Wie die Behauptungen über "meine" Hypothesen zustande gekommen sind, konnte sich Flynn - wie er mir brieflich mitteilte – selber nicht erklären ...