Boehnke, Klaus

# Probleme der Intelligenzmessung bei Kindern mit dem HAWIK-R

*Praxis der Kinderpsychologie und Kinderpsychiatrie 35 (1986) 2, S. 34-41*

# Praxis der Kinderpsychologie und Kinderpsychiatrie

Ergebnisse aus Psychoanalyse, Psychologie und Familientherapie



35. Jahrgang / 1986

VERLAG FÜR MEDIZINISCHE PSYCHOLOGIE IM VERLAG
VANDENHOECK & RUPRECHT IN GÖTTINGEN UND ZÜRICH

# Probleme der Intelligenzmessung bei Kindern mit dem HAWIK-R [1]

Von Klaus Boehnke


## Zusammenfassung

Die seit 1983 erhältliche revidierte Fassung des Hamburg-Wechsler-Intelligenztests für Kinder (HAWIK-R) (*Tewes*, 1983) diagnostiziert nach einer Untersuchung von *Eggert, Liman* und *Schirmacher* (1984) IQ-Werte, die um 15 Punkte unter denen des HAWIK (Bondy, 1956) liegen. Die vorliegende Arbeit diskutiert vier Hypothesen der Erklärung dieser großen IQ-Unterschiede: (a) Zero-Shift, (b) Eichstichprobenverzerrungen beim HAWIK-R, (c) Konzentrationsabfall beim HAWIK-R und (d) inhaltliche Unterschiede. Eine empirische Untersuchung an einer Grundschulklasse bestätigt die hohen Differenzen, 16 IQ-Punkte Unterschied wurden gefunden. Die einzelnen Ergebnisse der Studie legen den Schluß nahe, daß alle vier postulierten Ursachen – auch Eichstichprobenverzerrungen beim HAWIK-R – zu den hohen Differenzen beitragen, am wenigsten noch ein Konzentrationsabfall. Der Artikel faßt das Material der empirischen Untersuchung dahingehend zusammen, daß vorläufig kein Anlaß besteht, anzunehmen, der HAWIK-R könne die „wahre" Intelligenz verläßlicher messen als der HAWIK.


---

Seit 1983 ist eine neue Version des Hamburg-Wechsler-Intelligenztests für Kinder, der HAWIK-R (*Tewes*, 1983) erhältlich. Es handelt sich um eine Bearbeitung der Wechsler Intelligence Scale for Children – Revised (*Wechsler*, 1974). Gleichzeitig soll der HAWIK-R den in der klinischen, schulpsychologischen und sonderpädagogischen Diagnostik weit verbreiteten HAWIK (*Bondy*, 1956) ablösen.

Erste Untersuchungen, in denen Kinder sowohl mit dem HAWIK als auch mit dem HAWIK-R getestet wurden, zeigen deutliche IQ-Differenzen zwischen beiden Tests. Eine Untersuchung von *Eggert, Liman* und *Schirrmacher* (1984) berichtet z. B. um ca. 15 Punkte niedrigere Testwerte beim HAWIK als beim HAWIK-R. Die hier vorgelegte Arbeit versucht zunächst die testpraktische Bedeutung des HAWIK in einem wichtigen Bereich der schulpsychologischen Diagnostik, nämlich dem Sonderschulaufnahmeverfahren (SAU) zu skizzieren. Danach werden vier Hypothesen erläutert, die möglicherweise geeignet sind, die hohen Diskrepanzen zwischen HAWIK- und HAWIK-R-Werten zu erklären. Zum Schluß wird eine kleine Untersuchung vorgestellt, die Material zu den Hypothesen liefert, ohne jedoch für sich in Anspruch zu nehmen, eine Überprüfung der Hypothesen möglich zu machen.

## Testpraktische Bedeutung des HAWIK

In den Schulpsychologischen Diensten ist es üblich, bei Schülern mit einem IQ unter 85 die Überweisung auf die Sonderschule ernsthaft in Betracht zu ziehen (vgl. *Ahrbeck, Lommatzsch & Schuck,* 1984). Auslösend für die Einleitung eines SAU sind die in der Regel zwar nicht Intelligenzdefizite, sondern Schulschwierigkeiten (vgl. hierzu allgemein *Barkey, Langfeldt & Neumann,* 1976) und, wie *Maxeiner, Lauff* und *Homfeldt* (1979) zeigen, Verhaltensauffälligkeiten, man kann aber davon ausgehen, daß ein IQ unter 85 die Befunde besonders von nicht-objektivierten Diagnostika stark überformt. Dieser Effekt wird auch dadurch kaum gemildert, daß es z. B. in Berlin für ein mit der Empfehlung der Sonderschulüberweisung abschließendes Gutachten immer noch eines IQ-Befunds unter 85 mittels eines nonverbalen Intelligenztests bedarf, etwa der Standard Progressive Matrices (*Raven,* 1971).

Nach den Gesetzmäßigkeiten der Normierung des HAWIK würde eine Sonderschulüberweisung nur nach dem Kriterium $IQ \leq 85$ zu jeweils ca. 16% Sonderschülern pro Jahrgang führen, wenn die gesamte Kohorte getestet würde. Diese Größenordnung aber wurde in der Geschichte der Bundesrepublik nie erreicht (vgl. *Sander,* 1978). Ein Schrankenwert, der sich an den Empfehlungen des *Deutschen Bildungsrats* (1974) orientiert – dort werden maximal 6% Sonderschüler pro Jahrgang gefordert – läge bei einem IQ von 76. Bildungsgeschichtlich stieg der prozentuale Anteil von Sonderschülern an der Gesamtschülerzahl bis Mitte der 70er Jahre stetig an (vgl. *Sander,* 1978). Seit dieser Zeit sinkt der Anteil der Sonderschüler langsam aber kontinuierlich (vgl. *Statistisches Bundesamt,* 1978–1984 oder *Der Senator für Schulwesen, Jugend und Sport,* 1984). Über die Gründe für die sinkenden Anteile von Sonderschülern in den letzten Jahren kann nur spekuliert werden, in Berlin liegt der Anteil derzeit bei 3,8% [2] (vgl. *Der Senator für Schulwesen, Jugend und*

[1] Für die Möglichkeit ausführlicher Diskussionen eines ersten Entwurfs dieses Textes danke ich Dr. R. Oesterreich, Dr. J. Schreckling und Dipl.-Psych. D. Widowski.

[2] Dieser Prozentwert entspräche einem Schrankenwert von 74.

Prax. Kinderpsychol. Kinderpsychiat. 35: 34–41 (1986), ISSN 0032-7034

*Sport,* 1985). Diese vergleichsweise niedrige Zahl ist sicher auch ein Reflex einer in den 70er Jahren zu beobachtenden schulpolitischen Tendenz gegen eine erweiterte Aussonderung von Kindern, hin zu einer Integration behinderter Kinder, seien sie nun lern-, geistig- oder körperbehindert, in die Regelschule (vgl. hierzu den Sammelband von *Deppe-Wolfinger,* 1983). Diese schulpolitische Akzentsetzung dürfte zu einem gewissen Maß auch die Urteile in der Einzelfalldiagnostik in den Schulpsychologischen Diensten beeinflußt haben.

Als ein weiterer Grund für die sinkende Anzahl von Sonderschülern wird auch die Normveraltung vieler Intelligenztests, speziell des HAWIKs genannt (vgl. zum Problem der Normveraltung *Kubinger,* 1983). Es wird angenommen, daß der Test für die heutige Schülergeneration, ob eines gesellschaftlich gewachsenen Bildungs- und Ausbildungsniveaus, im Durchschnitt leichter geworden ist, als er es für Schüler der Eichstichprobe Mitte der 50er Jahre war. Man würde also weniger Schüler mit einem IQ unter 85 diagnostisch ermitteln können, als „in Wahrheit" vorhanden sind. Man könnte, wenn der Intelligenztest als Hauptkriterium für die Sonderschulüberweisung dient, weniger Kinder aussondern (!), als wenn man gültige Normen hätte.

## Erklärungsmöglichkeiten für Testwertdiskrepanzen zwischen HAWIK und HAWIK-R

Sicher dürfte eine Revision des HAWIK überfällig gewesen sein, doch muß auch gesehen werden, daß der HAWIK-R zu einem Zeitpunkt auf den Markt kam, zu dem auch gesamtpolitisch eine Wende hin zu mehr Auslese stattfand. Und mehr – negative – Auslese könnte der HAWIK-R in der Tat ermöglichen. Erste Vergleichsstudien zeigen, wie oben bereits erwähnt, daß sich die Werte von HAWIK und HAWIK-R im Durchschnitt um ca. 15 IQ-Punkte unterscheiden, die HAWIK-R-Werte liegen um etwa diesen Betrag niedriger als die HAWIK-Werte (vgl. *Eggert, Liman & Schirrmacher,* 1984). Die Testautoren selbst sprechen von 10 Punkten, die sie aufgrund einer Studie von *Mayer* (1983) bei Lernbehinderten in etwa erwarten (vgl. *Titze & Tewes,* 1984). Wie kann es zu diesen Diskrepanzen kommen?

Vier Erklärungsmöglichkeiten bieten sich an:

(a) Die Diskrepanz muß als Folge eines „Zero-Shift" gewertet werden: die Testaufgaben des HAWIK werden heute von einer größeren Anzahl von Schülern gelöst als zur Zeit des Tests. Dies führt dann zu einer im Vergleich zur Normstichprobe zu hohen Bewertung. Durch die Bearbeitung und Neueichung werden vom HAWIK-R nur die Maßstäbe zurechtgerückt. Die Differenz zwischen den IQ-Werten beider Tests ergibt sich aus dem Bildungszuwachs der heutigen Jugend gegenüber der Jugend der 50er Jahre.

(b) Die große Diskrepanz ist eine Folge von systematischen Verzerrungen in der Eichstichprobe des HAWIK-R. Diese Erklärung für die hohen Diskrepanzen böte sich dann an, wenn belegt werden könnte, daß die Eichstichprobe des HAWIK-R im Vergleich zur Durchschnittsbevölkerung der einbezogenen Altersgruppe zu „gut" ist. Eine zu gute, zu intelligente Eichstichprobe hätte dann nämlich eine systematisch zu niedrige Bewertung aller später mit dem HAWIK-R getesteten Probanden zur Folge.

(c) Da der HAWIK-R gegenüber dem HAWIK erheblich verlängert wurde, mißt er in stärkerem Maß nicht nur Intelligenz, sondern auch Konzentrationsfähigkeit und Ausdauer. Da die Ausdauer von Kindern und Jugendlichen aber u.U. für eine ca. zweistündige Testsitzung nicht ausreicht, könnte dies zwangsläufig zu niedrigeren Rohwerten führen als beim HAWIK, da der HAWIK meist nicht länger als eineinhalb Stunden dauert. Da es sich bei beiden Intelligenztests um geeichte Verfahren handelt, fassen sich Wertpunkt- bzw. IQ-Unterschiede mit dieser Hypothese allerdings nur dann plausibel erklären, wenn bei der Eichung des HAWIK-R, abweichend von der Standardtestsituation eine substantielle Anzahl von Testungen über mehr als eine Sitzung verteilt wurden. Nur dann hat dies den Effekt, daß Konzentrationsabfälle in der Eichstichprobe eine geringere Rolle spielen, als sie es in der „normalen" Testsituation tun. Nur in diesem Fall wäre erneut von einer zu „guten" Eichstichprobe des HAWIK-R auszugehen.

(d) Vierte Ursache für die Diskrepanz könnte eine inhaltliche Unterschiedlichkeit der beiden Verfahren sein. Völlig unabhängig davon, ob der HAWIK – in den 50er Jahren angewendet – dieselbe Art von Intelligenz gemessen hat, wie der HAWIK-R heute, könnte es sein, daß, wenn beide Verfahren in den 80er Jahren benutzt werden, mit beiden Instrumenten unterschiedliche Fähigkeiten gemessen werden. Hierfür würden im Rahmen einer quantitativen Analyse niedrige Korrelationen zwischen gleichen Subtests des HAWIK und des HAWIK-R sprechen. Lägen bei niedrigen Korrelationen gleichzeitig hohe Wertpunkte für HAWIK-Subtests vor, so spräche dies dafür, daß der HAWIK eine „leichtere" Variante von Intelligenz mißt, der HAWIK-R eine schwerere. Im folgenden sollen diese vier Erklärungsansätze diskutiert werden:

*Zu (a)*: Ein Zero-Shift ist recht wahrscheinlich. Er wurde auch bei der Neueichung der Wechsler Intelligence Scale for Children (*Wechsler,* 1949) festgestellt. *Quattrocchi* und *Sherets* (1980) berichten in ihrem Review von Arbeiten zum WISC-R (*Wechsler,* 1974) ebenfalls im Durchschnitt niedrigere Werte bei der WISC-R-Scala als bei der WISC-Skala. Die gefundenen IQ-Differenzen liegen aber – von Studie zu Studie verschieden – bei höchstens etwa 10 Punkten. Nimmt man diese Werte als Maßstab (der Durchschnitt liegt bei etwa sechs Punkten), so bliebe zu fragen, warum es für die bundesdeutsche Revision des HAWIK zu höheren Diskrepanzen kommt.

*Zu (b)*: Eine Analyse der Eichprozedur ist nur schwer möglich, da *Titze* und *Tewes* (1984) zwar sagen, der HAWIK-R sei einer der bestgeeichten Tests überhaupt, über die Eichstichprobe aber außer Schultypaufschlüsselungen noch weniger Information geben als *Bondy* (1956) für

den HAWIK, für den zumindest noch ein gesonderter Bericht über die Standardisierung vorliegt (*Priester,* 1958). Durch Diskussionen mit Berliner Eichhelferinnen[3] ist jedoch eine, wenn auch begrenzte, Einschätzung der Güte der Eichprozedur möglich.

Zunächst ein Rückblick auf die Normierung des HAWIK. Die regionale Stratifizierung war bei der HAWIK-Eichung nicht gelungen, Kinder aus ländlichen Gebieten kamen nur aus Norddeutschland (vgl. *Bondy,* 1956). Eine adäquate Alters-, Geschlechts- und Schultypstratifizierung scheint beim HAWIK jedoch gewährleistet.

Bei der Normierung des HAWIK-R ist die regionale Stratifizierung auf den ersten Blick besser gelungen. Es wurden Eichhelfer an zehn Hochschulorten angeworben, die dann je eine Teilstichprobe am Hochschulort (Großstadt) und in einer hochschulortnahen Kleinstadt oder Umlandgemeinde zu testen hatten. Anders als bei der HAWIK-Eichung, bei der die Schulen, an denen Schüler getestet werden sollten, vorgegeben waren, erfolgte die Auswahl der Schulen – unter Alters- und Schultypquotenvorgabe – durch die Eichhelfer selbst. Dieses Selektionsprinzip könnte ein erster Grund für eine zu „gute" Eichstichprobe beim HAWIK-R sein: Eichhelfer werden – zumal wenn sie, wie bei der HAWIK-R-Eichung, pro Test und nicht nach Stundenlohn bezahlt werden – selbstverständlich darum bemüht sein, sich ihre Aufgabe zu erleichtern, sie werden sich im Prinzip „brave" Schüler aussuchen, bei klassenweisen Untersuchungen Klassen mit möglichst wenigen verhaltensauffälligen oder aus sonstigen Gründen problematischen Schülern. Auch werden Schulleiter, die von den Eichhelfern gebeten werden, eine Klasse für die Eichstichprobe zur Verfügung zu stellen, regelmäßig ein zumindest „unauffällige", oftmals sogar ihre beste Klasse in der einbezogenen Altersgruppe zur Verfügung stellen, ohne dies den Eichhelfern mitzuteilen. Ein solcher Mechanismus führt auch in größeren Stichproben noch zu erheblichen Verzerrungen (vgl. hierzu etwa *Kastner,* 1985 oder *Luck & Boehnke,* 1985). Brave Schüler werden nämlich in Testsituationen oder mehr Konzentration und Leistungsmotivation mitbringen als in irgendeiner Art und Weise „auffällige" Schüler. Brave Schüler, so darf man annehmen, werden im statistischen Mittel auch höhere IQ-Werte erreichen (vgl. hierzu auch *Kubinger,* 1983).

Auch die regionale Stratifizierung der HAWIK-R-Eichstichprobe birgt ihre Probleme. Zwar ist die regionale Ausgewogenheit sehr viel besser gewährleistet als beim HAWIK, die Auswahl von Universitätsstädten bzw. deren Umlandgemeinden birgt aber erneut die Gefahr einer zu „guten" Eichstichprobe. Es scheint nicht unplausibel, anzunehmen, daß in Universitätsstädten und deren Umlandgemeinden das durchschnittliche Bildungsniveau höher ist als im Gesamtdurchschnitt von Städten bzw. Landgemeinden. Trifft diese Annahme zu, würde das höhere Bildungsniveau vermutlich auch zu einer höheren Fähigkeit zur Lösung von Intelligenztestaufgaben und damit letztlich zu einer zu guten Eichstichprobe führen.

Drittes Argument für die Annahme eines positiven Bias in der Eichstichprobe des HAWIK-R ist die behördliche Praxis der Bewilligung von Forschungsvorhaben an Schulen: Da Kinder und Jugendliche organisatorisch am leichtesten an Schulen getestet werden können, liegt es nahe, unter Maßgabe der Altersquoten jeweils ganze Schulklassen zu testen. Dies aber war 1956 bürokratisch noch sehr viel einfacher als 1983. War es damals vielfach noch möglich, nach einfacher Rücksprache mit dem Lehrer ohne explizites Einverständnis der Eltern zu testen, so war 1983 im Regelfall die Zustimmung der obersten Schulbehörde einzuholen. Die Zustimmung dieser Behörde wiederum wurde (wie z. B. in Berlin) oft nur gegeben, wenn die Testleiter sich verpflichteten, in jedem Einzelfall das schriftliche Einverständnis der Erziehungsberechtigten einzuholen. Diese verschärfte Genehmigungspraxis dürfte erneut die positive Selektivität der Stichprobe verstärkt haben, da die Eichhelfer hierdurch darauf angewiesen waren, sich Schulklassen zu suchen, in denen mit einer hohen Teilnahmezustimmung zu rechnen war. Solche Schulklassen dürften aber, wie auch Erfahrungen aus anderen Forschungsvorhaben zeigen, eher aus Schulen in Mittelschichtsgebieten kommen als aus Unterschichtsregionen einer Stadt (vgl. hierzu *Boehnke & Scherrinsky,* 1985). Ein Mittelschichtsbias in der Teilnehmerauswahl aber dürfte erneut auch zu einem Bildungsbias und damit zu einer zu guten Eichstichprobe geführt haben. Es muß allerdings hinzugefügt werden, daß das Problem der Bewilligung nur bei schriftlichem Einverständnis der Eltern heute für alle schulbasierten Untersuchungen gilt. Bei der Eichung von Intelligenztests muß dieser Unterschied zu Untersuchungen aus den 50er Jahren, als man bestimmte Erhebungen noch als Quasi-Reihenuntersuchungen im Unterricht durchführen konnte, mitbedacht werden.

Das Ausmaß der skizzierten möglichen Verzerrungen in der Eichstichprobe des HAWIK-R ist kaum einzuschätzen. Der über die Diskrepanz zwischen WISC und WISC-R hinausgehende Unterschied zwischen IQ-Wert nach HAWIK und HAWIK-R deutet aber darauf hin, daß positive Verzerrungen in der Eichstichprobe eine gewisse Rolle spielen könnten.

*Zu (c)*: Auch für eine Bedeutsamkeit der Annahme eines Konzentrationsabfalls beim HAWIK-R gibt es einige Hinweise aus Details der Durchführungsanweisungen für Eichhelfer. Diese erhielten nämlich die Instruktion, bei Schulschluß – falls vom Kind verlangt – oder bei erkennbaren Konzentrationsmängeln die Testsitzung zu unterbrechen und am nächsten Tag fortzusetzen. Da Testsitzungen besonders bei schwach intelligenten Schülern überdurchschnittlich lange dauern, dürfte diese Instruktion erneut zu mehr „ungerechtfertigt" verbesserten Testleistungen – in diesem Fall bei schwach Intelligenten geführt haben, als wenn ununterbrochene Testsitzungen vorgeschrieben gewesen wären. Das Ausmaß dieses möglichen Eichfehlers ist a priori kaum einzuschätzen, da

[3] Für die Möglichkeit ausführlicher Diskussionen bedanke ich mich bei Frau Dipl.-Psych. S. Sawitzki und Frau Dipl.-Psych. C. Münch.

nicht bekannt ist, welcher Prozentsatz von Kindern in getrennten Testsitzungen getestet wurde.

*Zu (d)*: Die Frage der Bedeutsamkeit von inhaltlichen Unterschieden zwischen den beiden Tests läßt sich an dieser Stelle qualitativ nicht ausführlich diskutieren. Eine oberflächliche Analyse der Subtestkorrelationen für 10jährige läßt jedoch die Vermutung zu, daß zumindest der Subtest ‚Rechnerisches Denken (RD)' bei beiden Tests unterschiedliche Fähigkeiten mißt. Mit LISREL V (*Jöreskog & Sörbom*, 1981) durchgeführte konfirmatorische Faktorenanalysen dieser Interkorrelationsmatrizen bestätigen diese Vermutung. Postuliert man für die beiden Tests jeweils zwei korrelierte Faktoren (Verbale Intelligenz und Handlungsintelligenz), so ergeben sich die in Tabelle 1 dokumentierten Ladungen.

*Tab. 1: Ladungen für korrelierte Faktoren (Verbale Intelligenz und Handlungsintelligenz) beim HAWIK und HAWIK-R*[a]

| | Verbale Intelligenz | | Handlungsintelligenz | |
|---|---|---|---|---|
| | HAWIK | HAWIK-R | HAWIK | HAWIK-R |
| AW[b] | .88 | .84 | | |
| AV | .73 | .68 | | |
| RD | .58 | .70 | | |
| GF | .71 | .80 | | |
| WT | .84 | .81 | | |
| ZS | | | .50 | .44 |
| BE | | | .64 | .61 |
| BO | | | .69 | .55 |
| MT | | | .69 | .82 |
| FL | | | .69 | .76 |

[a] Ladungen berechnet mit dem Maximum-Likelihood-Schätzalgorithmus auf der Basis der Interkorrelationsmatrizen der Eichstichproben der 10jährigen ohne Subtest ‚Zahlen-Nachsprechen (ZN)'.

[b] AW = Allgemeines Wissen, AV = Allgemeines Verständnis, RD = Rechnerisches Denken, GF = Gemeinsamkeiten finden, WT = Wortschatztest, ZS = Zahlen-Symbol-Test, BE = Bilder ergänzen, BO = Bilder ordnen, MT = Mosaiktest, FL = Figurenlegen.

Diese Ladungen machen deutlich, daß bei den Subtests ‚RD', ‚Bilderordnen (BO)' und ‚Mosaiktest (MT)' von gewissen inhaltlichen Veränderungen auszugehen ist, da hier Ladungsdifferenzen von mehr als .10 vorliegen.

Die im folgenden geschilderte empirische Untersuchung versucht, in einem exemplarischen Design Material zu den vier geschilderten Erklärungsmöglichkeiten zu sammeln. Sie nimmt nicht für sich in Anspruch Hypothesen zu testen, hierzu ist bereits der Stichprobenumfang – eine Schulklasse – viel zu klein.

## Methode

Untersucht wurde eine fünfte Grundschulklasse[4] aus dem Berliner Bezirk Neukölln (Einzugsbereich Gropiusstadt – eines der größten geschlossenen Neubaugebiete Berlins)[5] in einem zweifach balancierten Design. Die Klasse hatte 28 Schüler, neun davon wurden zuerst mit dem HAWIK (ohne Subtest ‚Zahlennachsprechen [ZN]'), dann mit dem HAWIK-R (ebenfalls ohne Subtest ‚ZN') getestet (Gruppe A). Bei neun Schülern war die Testreihenfolge umgekehrt (Gruppe C). Um auch Effekte der Sitzungsdauer testbar zu machen, wurde für eine dritte Gruppe (Gruppe B) eine Testmischform erstellt. Per Zufall wurden fünf Subtests aus dem HAWIK ausgewählt und mit den anderen fünf Subtests aus dem HAWIK-R zusammengefügt; die Testmischform bestand also je zur Hälfte aus HAWIK- und HAWIK-R-Subtests. Diese Testform wurde bei der ersten Testsitzung benutzt, bei der zweiten Testsitzung wurden die jeweils noch nicht benutzten Subtests einbezogen. Die Testsitzungen erfolgten in 14tägigem Abstand[6]. Die Teilnehmer wurden den drei Gruppen per Zufall zugewiesen: Gruppe A: n = 9, Gruppe B: n = 10, Gruppe C: n = 9 (8)[7]. Neben den Testscores wurden Geschlecht und Alter der Teilnehmer sowie die Deutsch- und Mathematik-Note erhoben. Zur Prüfung der internen Validität der Untersuchung wurden auch noch der Wochentag und die Tageszeit der Testung sowie Geschlecht und Alter des Versuchsleiters dokumentiert. Die Stichprobe umfaßte 13 Mädchen und 15 Jungen, der Altersrange lag zwischen 10 Jahren und 10 Monaten und 12;11 mit einem Mittelwert von 11;7. Die Testung der Teilnehmer wurde anonym durchgeführt, die Wiederauffindung erfolgte über Codenamen. Testleiter waren die Teilnehmer eines Hauptstudiumsseminars am Institut für Psychologie der Technischen Universität Berlin[8]. Die Testleiterschulung war ein Teil des Seminars, so daß ein etwa gleicher Schulungsstand für alle Testleiter gewährleistet war. Die Entscheidung über die Bewertung einzelner Testantworten erfolgte in Zweifelsfällen im Seminar. Für jeden Teilnehmer wurden pro Test Rohpunkte, Wertpunkte und Gesamt-IQ ermittelt. Für die Teilnehmer, die mit der Mischform getestet wurden, bedeutete dies, daß der pro Test zu ermittelnde IQ erst nach Beendigung beider

[4] Die Grundschule endet in Berlin erst mit Klasse 6.

[5] Die Durchführung der Untersuchung wurde ermöglicht durch die Schulpsychologische Beratungsstelle Berlin-Neukölln und die Klassenlehrerin der untersuchten Klasse, Frau Schörke. Die Leiterin der Beratungsstelle, Frau Dipl.-Psych. C. Wiedorn, und Dipl.-Psych. R. Füllert hatten maßgeblichen Anteil an einer ausführlichen Testleiterschulung.

[6] Insgesamt fünf Schüler aus allen drei Gruppen konnten wegen eines kurzfristig anberaumten Wandertages erst nach drei Wochen zum zweiten Mal getestet werden; das verlängerte Testintervall hatte jedoch keinen Einfluß auf die Testwerte (p. > .25).

[7] Ein Teilnehmer der Gruppe C konnte – ebenfalls bedingt durch die Verschiebung der Testwiederholung – gar nicht mehr an der zweiten Testung teilnehmen. Dieser Ausfall dürfte jedoch die Ergebnisse nicht verfälschen, da der Teilnehmer bei der Erstmessung mit dem HAWIK-R genau den durchschnittlichen IQ seiner Treatmentgruppe erreichte.

[8] Für die Durchführung der Testsitzungen und ihre Auswertung bedanke ich mich bei allen Teilnehmern des zweisemestrigen Seminars ‚Pädagogisch-psychologische Diagnostik'.

Testsitzungen errechenbar war, da nach einer Testsitzung jeweils erst die Ergebnisse von fünf Subtests des HAWIK und des HAWIK-R vorlagen.

## Ergebnisse

Im Vordergrund der Auswertung standen zunächst Fragen der internen Validität. Es zeigte sich, daß Testungseffekte wie Wochentag und Tageszeit der Testung sowie Geschlecht und Alter des Versuchsleiters nicht nachzuweisen waren. Varianzanalytische Überprüfungen, bei denen sowohl der IQ nach HAWIK als auch der nach HAWIK-R die abhängige Variable waren, erbrachte jeweils nicht-signifikante Ergebnisse ($p. > .25$), so daß man nach *Bortz* (1985) davon ausgehen kann, daß die Nullhypothese beibehalten werden kann. Ebenfalls ohne Einfluß auf den durchschnittlichen IQ der drei Testgruppen[9] war die Testreihenfolge ($F = .47$, $p = .631$). Bei einzelnen Subtests gab es deutliche Lerneffekte: die Gruppe derer, die den Subtest ‚Bilder ergänzen (BE)' des HAWIK bei der ersten Sitzung erhielten, gaben im Durchschnitt 64,4% richtige Antworten, diejenigen, die diesen Subtest in der zweiten Testsitzung, also nach Vortestung mit dem HAWIK-R, erhielten, erreichten im Durchschnitt 68,1% richtige Lösungen. Beim entsprechenden Subtest des HAWIK-R ist der Lerneffekt ähnlich groß – 70,8% vs. 73,7% –, so daß davon auszugehen ist, daß Lerneffekte bei einzelnen Tests zwar vorliegen, diese aber durch das balancierte Design nicht zugunsten der Gruppenmittelwerte eines der beiden Tests durchschlagen können: von den HAWIK-R-Getesteten, wie von den HAWIK-Getesteten, erhielten jeweils gleichviele Testpersonen den Test in der ersten Sitzung und in der Wiederholungssitzung[10].

Ebenfalls kein Einfluß auf die Testergebnisse konnte für die Testdauer nachgewiesen werden; diese Frage war für die Bewertung des HAWIK-R von Bedeutung, ein Einfluß der – experimentell variierten – Testdauer auf die Testwerte des HAWIK-R hätte als Indiz für die Bedeutsamkeit der oben erläuterten Konzentrationsabfallüberlegungen interpretiert werden können. Zwischen der Gruppe, die den Test im Standardformat erhielten und der Gruppe, die mit der Mischform getestet wurde, ließ sich jedoch kein signifikanter Unterschied im IQ nach HAWIK-R feststellen ($F = 2.56$, $p = .122$). Eine gewisse Tendenz zu niedrigeren Testwerten durch verlängerte Testsitzungen läßt sich jedoch nicht ausschließen: Die Teilnehmer, die den HAWIK-R in der kürzeren Mischform erhielten, erreichten in der Tat höhere IQ-Werte als diejenigen, die den Test in seiner Standardform erhielten – 110 vs. 102 – doch ist dieser Unterschied statistisch nicht bedeutsam. Interessant ist auch die Tatsache, daß die Streuung der IQ-Werte in der mit der Standardform getesteten Teilstichprobe wesentlich höher ist – $s = 13.9$ – als in der mit der Mischform getesteten Gruppe – $s = 9.3$.

Die Korrelation zwischen IQ und Geschlecht (♂ = 0, ♀ = 1) lag bei $r = -.03$ beim HAWIK und bei $r = -.09$ beim HAWIK-R. Beide Werte sind nicht signifikant. Die Korrelationen mit der Deutsch- bzw. der Mathematik-Note lagen für den HAWIK bei $r = -.35$ ($p = .048$) bzw. $r = -.54$ ($p = .002$) und für den HAWIK-R bei $r = -.29$ ($p = .082$) bzw. $r = -.47$ ($p = .007$). Die Tatsache, daß keine Effekte der Testbedingungen nachweisbar sind, ermöglicht eine valide Gegenüberstellung der Testergebnisse von HAWIK und HAWIK-R.

Berichtet wird zunächst das Ergebnis eines einfachen Mittelwertsvergleichs mit dem t-Test für abhängige Stichproben über die Gesamt-IQ-Werte. Der durchschnittliche mit dem HAWIK ermittelte IQ lag bei 121.11 bei einem Range von 83 bis 155. Der durchschnittliche mit dem HAWIK-R ermittelte Wert lag bei 105.11 mit einem Range von 70 bis 126. Die Differenz hat einen t-Wert von $t = 9.7$ ($p < .001$). Multivariate Varianzanalysen mit Meßwiederholung mit den zehn Subtest-Wertpunkten als abhängigen Variablen zeigen allerdings, daß nicht alle Subtest in gleichem Ausmaß zu diesem hohen IQ-Unterschied beitragen. Auch multivariat ist zwar der Unterschied zwischen HAWIK und HAWIK-R statistisch bedeutsam ($F = 24.02$, $p = .001$), bei univariater Partitionierung zeigen sich jedoch in zwei Subtests keine signifikanten Unterschiede: ‚RD' ($F = 3.08$, $p = .092$) und ‚BE' ($F = 2.69$, $p = .114$). In allen anderen Subtests liegen die HAWIK-R-Wertpunkte signifikant niedriger als die HAWIK-Wertpunkte. Besonders ausgeprägt sind die Unterschiede bei ‚MT' ($F = 196.92$, $p < .001$) und beim Subtest ‚Figuren legen (FL)' ($F = 28.99$, $p < .001$). Tabelle 2 gibt einen Überblick über die durchschnittlichen Wertpunkte für alle Untertests des HAWIK und des HAWIK-R.

*Tab. 2: Durchschnittliche Wertpunkte in den Subtests des HAWIK und des HAWIK-R*[a]

| | HAWIK | HAWIK-R |
|---|---|---|
| AW | 11.6 | 10.1 |
| AV | 13.7 | 11.8 |
| RD | 9.7 | 8.8 |
| GF | 14.9 | 11.5 |
| WT | 13.6 | 10.8 |
| ZS | 12.5 | 11.6 |
| BE | 11.4 | 10.4 |
| BO | 13.1 | 10.7 |
| MT | 14.3 | 10.1 |
| FL | 13.1 | 10.7 |

[a] Alle Wertpunkt-Differenzen außer ‚RD' und ‚BE' sind mindestens auf dem 5%-Niveau signifikant.

Die hohen Wertpunkt-Diskrepanzen scheinen zunächst recht eindeutig für die Zero-Shift-Hypothese zu sprechen. Diese Hypothese kann jedoch nicht bei allen

[9] Der durchschnittliche IQ wurde berechnet als arithmetisches Mittel der Gesamt-IQ-Werte von HAWIK und HAWIK-R.

[10] Vgl. Anmerkung 7.

Subtests gleichermaßen als plausibel angenommen werden. Dies zeigt sich, wenn man die sogenannten Rohschwierigkeiten (= Prozentwerte der gelösten Aufgaben) der Subtests des HAWIK in der hier untersuchten Stichprobe mit den Rohschwierigkeiten in der Eichstichprobe (vgl. *Priester,* 1958) vergleicht. Tabelle 3 gibt einen Überblick über die entsprechenden Werte. Die Werte der Eichstichprobe wurden durch Interpolation und Umrechnung in Prozentwerte aus den Tabellen 15 und 16 bei *Priester* (1958, 45–46) errechnet. Zur Information gibt die Tabelle auch die Rohschwierigkeiten für die Untertests des HAWIK-R in der hier untersuchten Stichprobe; die Rohschwierigkeiten der Eichstichprobe des HAWIK-R wurden einer breiteren wissenschaftlichen Öffentlichkeit bisher noch nicht zugänglich gemacht.

*Tab. 3: Schwierigkeitsgrade der Subtests des HAWIK und des HAWIK-R*

| | HAWIK | | | HAWIK-R |
|---|---|---|---|---|
| | hier untersuchte Stichprobe | Eichstichprobe | Streuungseinheit in der Eichstichprobe | hier untersuchte Stichprobe |
| AW | .53 | .49 | .12 | .40 |
| AV | .58 | .44 | .14 | .69 |
| RD | .68 | .69 | .12 | .46 |
| GF | .54 | .36 | .13 | .51 |
| WT | .70 | .55 | .15 | .45 |
| ZS | .51 | .44 | .11 | .54 |
| BE | .64 | .59 | .14 | .77 |
| BO | .57 | .44 | .13 | .69 |
| MT | .70 | .47 | .21 | .65 |
| FL | .77 | .63 | .14 | .52 |
| Gesamttest | .62 | .51 | .14 | .57 |

Die Tabelle zeigt, daß für die Subtests ‚Allgemeines Verständnis (AV)', ‚Gemeinsamkeiten finden (GF)', ‚Wortschatzrest (WT)', ‚BO', ‚MT' und ‚FL' die in dieser Studie getesteten Schüler Rohpunkte erreichen, die um eine oder mehr als eine Standardabweichung höher liegen als die Eichstichprobe. Für die genannten Subtests scheint die Zero-Shift-Hypothese – auch wenn man bedenkt, daß es sich um eine leicht überdurchschnittlich intelligente Schulklasse handeln dürfte – plausibel. Keine Gültigkeit kann sie jedoch für die Subtests ‚Allgemeines Wissen (AW)', ‚RD)', ‚Zahlen-Symboltest (ZS)' und ‚BE' haben. Bei zwei dieser vier Tests sind jedoch, wie in Tabelle 2 dokumentiert ist, die Wertpunktunterschiede signifikant: ‚AW' und ‚ZS'. Nachdem die Zero-Shift-Hypothese für diese Subtests nicht plausibel ist und auch die Hypothese eines Konzentrationsabfalls als Ursache für niedrigere Testwerte im HAWIK-R nicht bestätigt werden konnte, bleiben noch zwei Erklärungsansätze für die Unterschiede bei diesen Subtests: (a) Qualitative Unterschiede der gemessenen Fähigkeiten und (b) Eichstichprobenverzerrungen. Tabelle 4 gibt die Korrelationen zwischen gleichnamigen Subtests von HAWIK und HAWIK-R und zusätzlich die 7- bis 9-Monats-Stabilitäten der Untertests beim HAWIK-R. Diese Stabilitäten sind die nach Fisher's-Z-Korrektur gemittelten Stabilitäten, die *Tewes* (1983, 43) für 8-, 11- und 14jährige gibt.

*Tab. 4: Korrelationen zwischen gleichnamigen Subtests des HAWIK und HAWIK-R und Stabilitäten des HAWIK-R*

| | Korrelationen | | Stabilitäten |
|---|---|---|---|
| | Rohwerte | Wertpunkte | |
| AW | .47 | .34 | .56 |
| AV | .30 | .39 | .51 |
| RD | .09 | .18 | .49 |
| GF | .58 | .61 | .31 |
| WT | .57 | .44 | .45 |
| ZS | .48 | .49 | .58 |
| BE | .16 | .25 | .35 |
| BO | .58 | .59 | .37 |
| MT | .80 | .82 | .58 |
| FL | .61 | .63 | .39 |

Vergleicht man die in Tabelle 4 angegebenen 7- bis 9-Monats-Stabilitäten mit den Zwei-Wochen-Subtest-Korrelationen, so ist folgerichtig anzunehmen, daß – gleiche Inhalte vorausgesetzt – die Zwei-Wochen-Stabilitäten, als die die Korrelationen im Falle inhaltlicher Äquivalenz interpretiert werden könnten, höher liegen müßten als die 7- bis 9-Monats-Stabilitäten. Die Tatsache, daß dies bei fünf Subtests nicht der Fall ist, legt zunächst die Vermutung nahe, daß bei den ‚AW', ‚AV', ‚RD', ‚ZS' und ‚BE' im HAWIK und HAWIK-R unterschiedliche Inhalte erfaßt werden. Bei den Subtests ‚AW', ‚AV', ‚RD' und ‚BE' ist die Annahme unterschiedlicher Inhalte auch durchaus plausibel, sie haben maximal fünf identische Items (vgl. *Tewes,* 1983, 22–26). Beim Subtest ‚RD' gab auch die Interkorrelationsmatrix der Eichstichprobe Hinweise auf inhaltliche Unterschiede. Beim Zahlen-Symbol-Test hat die Erklärung jedoch keine Gültigkeit; die Frage von Verzerrungen in der Eichstichprobe läßt sich an diesem Test am besten diskutieren: da der Test in beiden Testfassungen identisch ist, scheiden Inhaltsunterschiede als Erklärung für signifikante Wertpunktunterschiede aus. Die Annahme eines Zero-Shift läßt sich theoretisch und empirisch kaum belegen. Bildungsgeschichtlich ist die Annahme wenig plausibel, daß sich die Schnelligkeit im Kodieren von Zahlen in Symbole zwischen den 50er Jahren und heute wesentlich geändert haben soll. Empirisch zeigt sich auch kein bedeutsamer Unterschied zwischen der Eichstichprobe des HAWIK und der hier untersuchten Schulklasse. Rohpunktunterschiede zwischen beiden Tests, die etwa durch die unterschiedliche Positionierung dieses Subtests innerhalb der Testsitzung bei HAWIK und HAWIK-R entstehen könnten, sind ebenfalls nicht festzustellen. Der empirische Rohpunktunterschied – $\bar{x} = 47{,}1$ (HAWIK) und $\bar{x} = 50{,}2$ (HAWIK-R) – ist gering und erhöht im übrigen die durchschnittlichen Wertpunkte beim HAWIK-R, mi-

nimiert also eher die IQ-Diskrepanzen zwischen beiden Tests. Obwohl aber inhaltliche Unterschiede, Zero-Shift oder Testdauereffekte als Erklärung ausscheiden, unterscheiden sich die Wertpunktergebnisse dieses Subtests zwischen HAWIK und HAWIK-R signifikant: 12.5 vs. 11.6 (vgl. Tabelle 2). Rechnet man diese Werte exemplarisch auf den Gesamt-IQ hoch, multipliziert also die Wertpunkte des einen Subtests mit der Anzahl der im Gesamttest enthaltenen Subtests (10) und rechnet die so erreichten Wertpunkte in IQ-Punkte um, so ergibt sich ein IQ-Unterschied von sieben: beim HAWIK-R ergäbe sich ein Gesamt-IQ von 112, beim HAWIK ein IQ von 119.

## Zusammenfassende Diskussion

Die zusammenfassende Diskussion muß zunächst hervorheben, daß die IQ-Diskrepanz in der hier untersuchten Stichprobe mit 16 Punkten höher liegt als der Wert von 10, den *Titze* und *Tewes* (1984) selbst erwarten. Die Erhebung bestätigt damit die Ergebnisse der Studie von *Eggert, Liman* und *Schirrmacher* (1984), die eine Differenz von 15.12 Punkten berichten.

Vier Erklärungsmöglichkeiten für das Zustandekommen der hohen Diskrepanzen waren formuliert worden: (a) die Zero-Shift-Annahme, (b) die Annahme von Verzerrungen in der Eichstichprobe, (c) die Annahme eines Konzentrationsabfalls durch überproportionale Testverlängerung beim HAWIK-R und (d) die Annahme inhaltlicher Unterschiede.

Die Annahme, daß ein Konzentrationsabfall beim HAWIK-R wesentlich negativ zu Buche schlagen könnte, ließ sich nicht bestätigen: die Testergebnisse, die mit dem HAWIK-R in Standardtestsituationen erreicht wurden, unterschieden sich nicht statistisch bedeutsam von denen, die in verkürzten Testsitzungen erzielt wurden.

Die Zero-Shift-Annahme ließ sich am deutlichsten bei den Subtests ‚AV', ‚GF', ‚WT', ‚BO', ‚MT' und ‚FL' stützen, wobei beim Subtest ‚AV' zusätzlich noch von einer inhaltlichen Veränderung ausgegangen werden muß.

Von inhaltlichen Veränderungen ohne den schlüssigen Beleg eines Zero-Shift kann bei den Subtests ‚RD', ‚AW' und ‚BE' ausgegangen werden, eine Tatsache, die sich für den Subtest ‚RD' auch bereits in den Eichstichproben andeutet. Diese Tasache dürfte besonders für die Differentialdiagnose im klinischen Bereich von Bedeutung sein. Die Frage, was die Tests jeweils psychologisch unterscheidet, bedarf jedoch zusätzlicher Forschung.

Am Subtest ‚ZS' wird deutlich, daß neben Zero-Shift und inhaltlichen Unterschieden auch positive Abweichungen von der Repräsentativität bei der Eichstichprobe des HAWIK-R (= zu hohe durchschnittliche Intelligenz) von Bedeutung sein könnten. Obwohl bei diesem Test weder inhaltliche Unterschiede noch Zero-Shift als plausible Erklärung in Frage kommen, gibt es signifikante Wertpunktunterschiede. Akzeptiert man, entsprechend den Axiomen der klassischen Testtheorie den Stichprobenmittelwert als Punktschätzung der „wahren" Intelligenz der untersuchten Schulklasse, so ließe sich die Hypothese formulieren, daß der HAWIK-R, bedingt durch ein positives Eichstichprobenbias, um ca. 7 IQ-Punkte zu niedrige Werte ermittelt. Wägt man nun die Bedeutsamkeit eines möglichen Bias mit der des Zero-Shift ab, so ließe sich angesichts einer empirischen Differenz von 16 Punkten, die in dieser Stichprobe ermittelt wurden, formulieren: der HAWIK überschätzt die „wahre" Intelligenz etwa in demselben Maße, wie der HAWIK-R sie unterschätzt. Die von *Titze* und *Tewes* (1984) berichtete Untersuchung von *Mayer* (1983) spricht – wenn man die gleiche Hochrechnungsprozedur anwendet – sogar für eine noch größere Unterschätzung der „wahren" Intelligenz durch den HAWIK-R bei Lernbehinderten. Selbstverständlich ist eine Punktschätzung der Unterschätzung angesichts der sehr kleinen Stichprobe, die hier untersucht wurde, mit Fehlern behaftet. Kritisch ließe sich auch einwenden, daß exakte Hypothesen auf der Basis eines im Vergleich zu anderen Subtests recht instabilen Merkmals wie der Fähigkeit schnell Zahlen in Symbole zu kodieren, kaum formuliert werden können. Die aufgestellte Hypothese von Verzerrungen in der Eichstichprobe bleibt zudem insofern Spekulation, als Annahmen über die Eichprozedur aus Erfahrungen von Eichhelferinnen in Berlin verallgemeinert wurden. Die Studie versteht sich deshalb auch nur als Materialsammlung zu verschiedenen Erklärungsansätzen für große IQ-Diskrepanzen zwischen HAWIK und HAWIK-R. Weitere Veröffentlichungen seitens des Testautors über den technischen Ablauf der Eichung und deren Repräsentativitätsprüfungen, die für den HAWIK z.T. vorliegen (vgl. *Priester,* 1958), sind auch für den HAWIK-R sehr wünschenswert.

Insgesamt jedoch läßt sich festhalten, daß es gewisse Hinweise dafür gibt, daß eine Punktschätzung der „wahren" Intelligenz mit dem HAWIK-R kaum besser möglich ist als mit dem HAWIK. Die Tatsache, daß der HAWIK-R die „wahre" Intelligenz eines Kindes möglicherweise um eine Reihe von Punkten unterschätzt, sollte nun allerdings den Praktiker nicht dazu verleiten, zum HAWIK-R-Testergebnis einfach ein paar Punkte hinzuzuzählen und dann davon auszugehen, die „wahre" Intelligenz eines Kindes sei ermittelt. Vielmehr zeigt die hier vorgestellte Untersuchung, wie wenig sich „wahre" Intelligenz schon allein wegen vieler Unwägbarkeiten im Eichprozeß überhaupt ermitteln läßt.

Besonders in einer Zeit, in der Auslese und Elitenförderung schul- und gesellschaftspolitisch stärker diskutiert werden, ist es dringlich, hervorzuheben, wie wichtig es ist, Diagnostik in den Dienst der Förderung zu stellen. *Guthke* (1972) und in jüngerer Zeit *Kornmann* (1983) haben hierzu wichtige Vorschläge gemacht. *Feuerstein* (1968) hat gezeigt, daß sich auch klassische Intelligenztests durchaus im Rahmen einer förderungsorientierten Begabungsdiagnostik einsetzen lassen. Hier hat auch der HAWIK-R seinen Platz und sollte, wenn z.B. ein Sonderschulaufnahmeverfahren tatsächlich unumgänglich erscheint, auch (nur) so verwendet werden.

## Summary

*Problems of the Measurement of Intelligence in Children by Means of the HAWIK-R*

According to a study by *Eggert, Liman* and *Schirmacher* (1984), the revised version of the Hamburg-Wechsler-Intelligence-Scale for Children (HAWIK-R), introduced by *lewes* (1983), leads to IQ-scores which are up to 15 points below the scores found when using the HAWIK (*Bondy,* 1956). The present study discusses four hypotheses, which possibly explain the large discrepancies: (a) zero-shift, (b) a bias in the normative sample of the HAWIK-R, (c) a decline in concentration because of an increase in the number of items of the HAWIK-R and (d) contentwise differences. An empirical study of one class from elementary school confirms high IQ-score differences, 16 points are reported. The detailed results of the study lead to the conclusion that all four causes postulated – including a bias in the normative sample – are presumably responsible for the large differences, least probable being a decline in concentration. In summary, the empirical material suggests that there is little evidence as of yet that the HAWIK-R measures „true" intelligence more reliably than the HAWIK.


## Literatur


*Ahrbeck, B., Lommatzsch, E.-M. & Schuck, K.-D.* (1984): Der neue HAWIK – ein neues Verfahren der sonderpädagogischen Diagnostik? In: Zeitschrift für Heilpädagogik, 35, 51–54. – *Barkey, P., Langfeldt, H.-P. & Neumann, G.* (1976): Pädagogisch-psychologische Diagnostika am Beispiel von Lernschwierigkeiten. Bern: Huber. – *Boehnke, K. & Scherrinsky, K.* (1985): Die ersten zwei Erhebungswellen im Berliner Jugendlängsschnitt – Eine Bilanz der Stichprobenentwicklung. Berlin: Berichte aus der Arbeitsgruppe TUdrop/Jugendforschung (49/85). – *Bondy, C.* (Hg.) (1956): Hamburg-Wechsler-Intelligenztest für Kinder. Bern: Huber. – *Bortz, J.* (1985): Lehrbuch der Statistik. Berlin: Springer. – *Deppe-Wolfinger, H.* (Hg.) (1983): Behindert und abgeschoben – Zum Verhältnis von Behinderung und Gesellschaft. Weinheim: Beltz. – *Der Senator für Schulwesen, Jugend und Sport* (Hg.) (1984): Das Schuljahr 1983/84 in Zahlen. Berlin: o.V. – *Der Senator für Schulwesen, Jugend und Sport* (Hg.) (1985): Das Schuljahr 1984/85 in Zahlen. Berlin: o.V. – *Deutscher Bildungsrat* (Hg.) (1974): Empfehlungen der Bildungskommission zur pädagogischen Förderung behinderter und von Behinderung bedrohter Kinder und Jugendlicher. Stuttgart: Klett. – *Eggert, D., Liman, E. & Schirmacher, A.* (1984): Vergleich des Hamburg-Wechsler-Intelligenztests für Kinder (HAWIK) mit seiner revidierten Fassung bei sprachbehinderten Kindern. In: Zeitschrift für Heilpädagogik, 35, 54–58. – *Feuerstein, R.* (1968): A dynamic approach to the causation, prevention and alleviation of retarded performance. In: *Haywood, H.C. (ed.): Social Cultural Aspects of Mental Retardation. New York: Appleton-Century-Crofts.* – *Guthke, J.* (1972): Zur Diagnostik der intellektuellen Lernfähigkeit. Berlin (DDR): Verlag der Wissenschaften. – *Jöreskog, K.D. & Sörbom, D.* (1981): LISREL V – Analysis of linear structural relationship by maximum likelihood and least square methods. Uppsala: University of Uppsala Research Report 81–8. – *Kastner, P.* (1985): Die Revision des Fragebogens zum Drogengebrauch im Berliner Jugendlängsschnitt, die Einführung eines diskriminationsvaliden Fragebogens zum Sportverhalten und epidemiologische Daten zum Drogengebrauch bei Jugendlichen 1982 und 1983. Berlin: Berichte aus der Arbeitsgruppe TUdrop/Jugendforschung (53/85). – *Kornmann, R.* (1983): Variationen von Testbedingungen als förderungsdiagnostischer Ansatz. In: *Trolldenier, H.-P. & Meißner, B.* (Hg.): Texte zur Schulpsychologie und Bildungsberatung, Band 4. Braunschweig: Pedersen. – *Kubinger, K.-D.* (Hg.) (1983): Der HAWIK – Möglichkeiten und Grenzen seiner Anwendung. Weinheim: Beltz. – *Luck, H. & Boehnke, K.* (1985): Die Einwohnermeldeamtsstichprobe im Berliner Jugendlängsschnitt. Berlin: Berichte aus der Arbeitsgruppe TUdrop/Jugendforschung (47/85). – *Maxeiner, J., Lauff, W. & Homfeldt, H.G.* (1979): Lehrer-Schüler-Interaktion und Schulerfolg. Weinheim: Beltz. – *Mayer, W.* (1983): Der Vergleich zweier HAWIK-Versionen bei lernbehinderten Sonderschülern. Hannover: Medizinische Hochschule, phil. Diss. – *Priester, H.-J.* (1958): Die Standardisierung des Hamburg-Wechsler-Intelligenztests für Kinder. Bern: Huber. – *Quattrocchi, M. & Sherets, S.* (1980): WISC-R: The First Five Years. In: Psychology in the Schools, 17, 297–312. – *Raven, J.C.* (1971): Standard Progressive Matrices. London: Lewis & Co. – *Sander, A.* (1978): Das Sonderschulwesen in der Bundesrepublik Deutschland. In: *Klauser, K.J. & Reinartz, A.* (Hg.): Sonderpädagogik in allgemeinen Schulen. Berlin: Marhold. – *Statistisches Bundesamt* (Hg.) 1978–1984: Statistisches Jahrbuch der Bundesrepublik Deutschland. Stuttgart: Kohlhammer. – *Tewes, U.* (1983): Hamburg-Wechsler-Intelligenztest für Kinder – Revision 1983. Bern: Huber. – *Titze, I. & Tewes, U.* (1984): Messung der Intelligenz bei Kindern mit dem HAWIK-R. Bern: Huber. – *Wechsler, D.* (1949): Manual for the Wechsler Intelligence Scale for Children. New York: Psychological Corporation. – *Wechsler, D.* (1974): Manual for the Wechsler Intelligence Scale for Children – Revised. New York: Psychological Corporation.


Anschr. d. Verf.: Dipl.-Psych. Klaus Boehnke, Inst. f. Allgemeine u. Vergleichende Erziehungswissenschaften der FU Berlin, Fabeckstr. 13, 1000 Berlin 33.