

Ist IQ-Diagnostik noch zeitgemäß – oder, was ist dran am ‚Flynn-Effekt‘?

(Vortrag bei der 16. Bundeskonferenz für Schulpsychologie im September 2004 in Nürnberg)

Im Jahre 1947 erregte ein amerikanischer Psychologe Aufsehen, als er in seinem Buch mit dem Titel „Nuremberg Diary“ die Ergebnisse von Intelligenztests der im Nürnberger Kriegsverbrechergefängnis einsitzenden Nazigrößen veröffentlichte, denen 1945 und 1946 wegen ihrer Kriegsverbrechen der Prozess vor einem internationalen Tribunal gemacht wurde. Er stellte dabei fest, dass alle einen, an den amerikanischen Normen des Wechslertests gemessenen, Intelligenzquotienten von mehr als 110 IQ hatten. Dieser hohe IQ hinderte sie allerdings nicht daran, die schlimmsten Kriegsverbrechen des 20. Jahrhunderts zu begehen - vielleicht war er dazu sogar behilflich. Was ich mit diesem fast makaberen Beispiel zum Ausdruck bringen möchte ist, dass IQ-Diagnostik wertneutrale, quantifizierte Kompetenzen misst, die sozial-schädlich oder auch prosozial zum Wohle der Menschheit und für mehr Chancengerechtigkeit eingesetzt werden können.

Im Jahre **1987** hatte **J.R. Flynn** (*University of Otago, New Zealand*) wissenschaftliches Aufsehen erregt, als er im *Psychological Bulletin* als Ergebnis eines internationalen Vergleiches von 14 Nationen ‚massive IQ-Gewinne‘ feststellte und IQ-Tests bezüglich dessen, was sie zu messen vorgeben, in Frage stellte. Bereits 10 Jahre vor Flynn hatten zwei deutsche Wissenschaftler (**Royl & Schwarzer 1976**) anhand empirischer Studien von einer ‚säkularen Akzeleration der PSB-Intelligenz‘ gesprochen. 15 Jahre nach Flynn wurde von **C. Zerahn-Hartung et al.** unter dem Titel ‚Normverschiebungen bei Rechtschreibleistungen und sprachfreier Intelligenz‘ in der Zeitschrift *„Praxis der Kinderpsychologie und Kinderpsychiatrie“* (**2002**) u.a. behauptet, dass der erwartete Flynn-Effekt von 0,33 IQ-Punkten pro Jahr durch den CFT 20 weit übertroffen würde (angeblich + 0,6 IQ pro Jahr). Dies führte zu einer Verunsicherung, insbesondere bei den schul-psychologischen Praktikern.

In meinem Referat werde ich

1. zunächst einige bekannte Untersuchungen zu ‚Normverschiebungen‘ bei Intelligenz- bzw. Begabungstests in den vergangenen 50 Jahren in der BRD bilanzieren,
2. den viel diskutierten Flynn-Effekt methodenkritisch bewerten,
3. die Schlussfolgerungen von Zerahn-Hartung et al. (2002) zum CFT 20 überprüfen und auf Schwachstellen des zugrunde liegenden Datenmaterials hinweisen,
4. neuere Befunde zum Grundintelligenztest CFT darstellen und anhand dieser sehr umfangreichen empirischen Untersuchungen den von Zerahn-Hartung behaupteten ‚Flynn-Effekt‘ kritisch überprüfen.
5. Abschließend werde ich auch transparent machen, in wie weit sprachfreie IQ-Diagnostik gerade in der schulpsychologischen Praxis noch immer ihre Berechtigung hat und welche Weiterentwicklungen es gibt.

Zu 1.

Die erste große Untersuchung führte Ende der 50er Jahre Prof. **Arnold**, Universität Würzburg, durch und veröffentlichte sie in seinem Buch *„Begabungswandel und Erziehungsfragen“* (1960). Arnold stellte anhand von Untersuchungsdaten der Arbeitsverwaltung, die bei den ‚Einfachen Eignungsuntersuchungen‘ (EEU) im Zweiten Weltkrieg und nach dem Krieg Tests durchführte, aufgrund einer großen Probandenzahl fest, dass ein Begabungswandel *weg vom Wort und hin zur Zahl* stattgefunden habe. Verringerte Werte in verbalen Testverfahren standen Steigerungen in zahlengebundenen Verfahren gegenüber.

Die zweite mir bekannte Untersuchung mit dem *Prüfsystem für die Schul- und Bildungsberatung (PSB)* stammt von **Royl & Schwarzer** aus dem Jahre 1976. Diese hatten festgestellt, dass die durchschnittliche Testleistung im PSB, dessen Normierung aus den

Jahren 1966/67 stammte, in den folgenden zehn Jahren bedeutsam zugenommen hatte. Mit der Publikation „Zur säkularen Akzeleration der PSB-Intelligenz“ hatten sie bewirkt, dass dieses Verfahren für Baden-Württemberg für die Klassenstufen 4-6 nachnormiert werden musste, da es dort sehr häufig für Schullaufbahnberatungen beim Übertritt von der Grundschule in die weiterführenden Schulen eingesetzt wurde.

Ich selbst hatte zur gleichen Zeit festgestellt, dass die ebenfalls im wesentlichen aus der gleichen Zeit stammenden Normierungsdaten des *Grundintelligenztests Skala 2 (CFT 2)* zu leicht waren und eine Nachnormierung geboten erschien. Diese fand im Jahre 1976/77 bundesweit an 4.500 Schülern statt. Daraus und aus anderen Validierungsbe-funden entstand dann der Grundintelligenztest CFT 20, der ohne weitere Normenkorrek-tur im Jahre 1998 in die 4. Auflage ging.¹

Die Ursachen für diese Entwicklung in den 70er Jahren mit bedeutsamen Punktezu-wachs in den Tests wurden übereinstimmend darin gefunden, dass Ende der 60er Jahre und vor allem in den 70er Jahren durch die bundesweiten großen Schulentwicklungspro-jekte und andere Maßnahmen zur lernzielorientierten Leistungs-messung mit *Multiple-Choice-Aufgaben* die sog. Testsophistication enorm gesteigert wurde. *Nicht etwa die In-telligenz hatte zugenommen, sondern die Erfahrungen beim Umgang mit Tests.* Nach ei-nem Sättigungsgrad auf relativ hohem Niveau an Testerfahr-ungen blieben die Ergebnis-se in IQ- bzw. Begabungstests danach ziemlich konstant. Vielleicht bis zum 1. *Großen IQ-Test von Günter Jauch im TV des Jahres 2002.* Dabei wurden - bei selten hoher Ein-schaltquote von rd. 13 Millionen Bundesbürgern aller Alters-gruppen - Testaufgaben aus allen gängigen psychologischen Intelligenztests vorgeführt und danach noch etwa 600.000 CDs mit denselben Aufgaben verkauft. Eine wahrhaft gigantische Zahl, die Testautoren und Verlagen großes Kopfzerbrechen bereitete.

Zunächst aber nochmals zurück in die Zeit vor 1980, aus der die Daten stammen, über die **Flynn (1987)** berichtete.

Zu 2.

Methodenkritische Anmerkungen zum viel diskutierten Flynn-Effekt.

Unter dem Titel “Massive IQ gains in 14 Nations: what IQ tests really measure.”, erschie-nen 1987 im Psychol. Bulletin, übte der neuseeländische Wissenschaftler *J.R. Flynn* hef-tige Kritik an den gängigen Intelligenztests. Seine Test-Daten stammten aus 14 Nationen und er bewirkte damit ein mittleres wissenschaftliches Erdbeben, dessen Auswirkungen bis heute spürbar sind. Insider wiesen schon länger darauf hin, dass gegen Flynns Ana-lyse berechnete methodische Einwände bestehen.

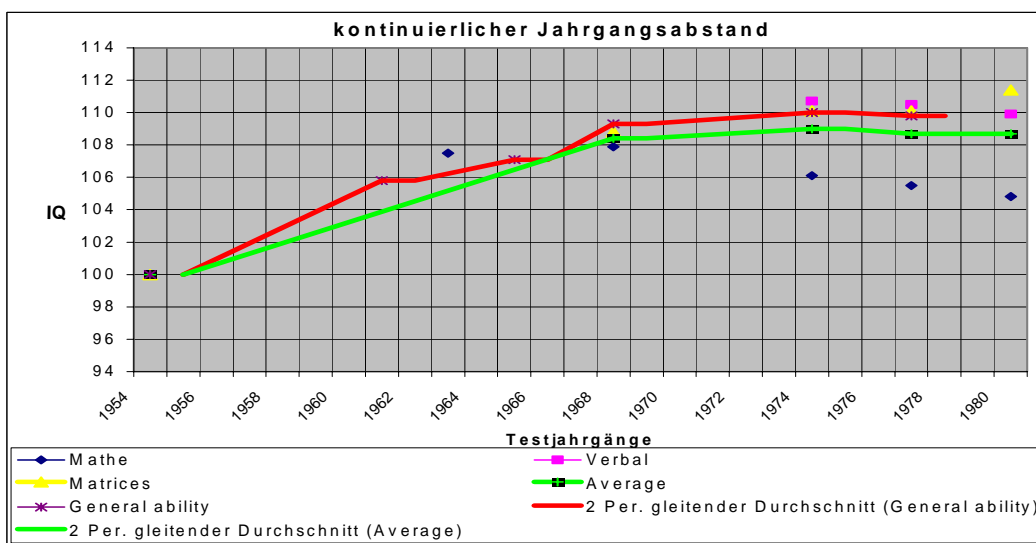
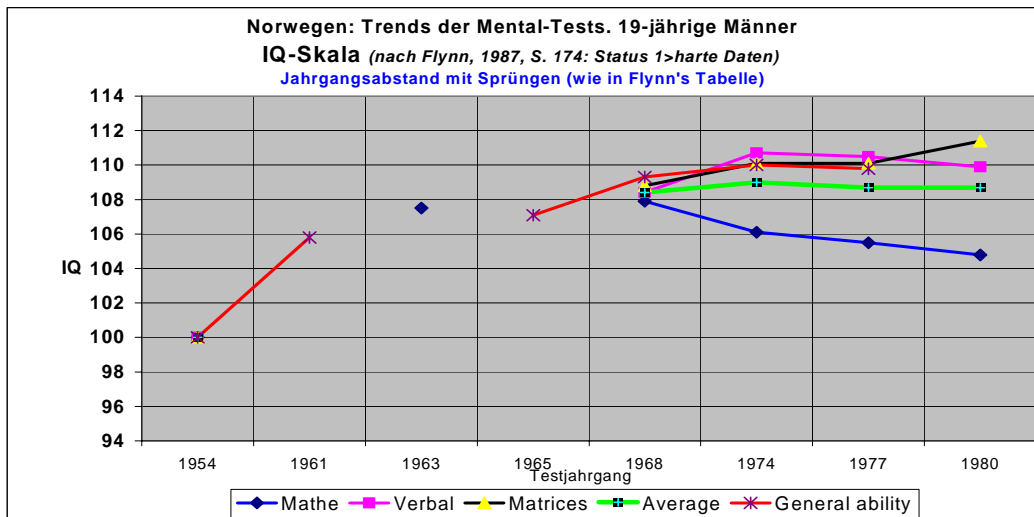
In einer enormen Fleißarbeit hatte *Flynn* Testdaten aus 14 Nationen gesammelt und nach einem einheitlichen Schema bewertet. In seiner Summary resumierte Flynn folgen-de Ergebnisse (eigene Übersetzung): „Die Daten aus 14 Nationen enthüllen, dass der IQ-Gewinn in einer einzigen Generation von 5 bis 25 Punkten reicht. Einige der größten Gewinne ereignen sich bei kulturell weniger abhängigen Tests wie bei *Tests, welche die Fluid Intelligence* erfassen. So zeigen die Norwegischen Daten, dass eine Nation signifi-kante IQ-Gewinne bei einem kulturell weniger abhängigen Test aufweisen kann, bei gleichzeitigem IQ-Verlust in anderen Tests. Die holländischen Daten bestätigten, dass die Existenz unbekannter Umweltfaktoren so stark sein kann, dass sie 15 von 20 Punk-ten IQ-Gewinn erklären kann. Die Hypothese, die am besten die Ergebnisse erklären kann, ist, dass *IQ-Tests nicht Intelligenz messen, sondern eher ein Korrelat mit einer schwachen Beziehung zur Intelligenz.* Diese Hypothese kann auch die unterschiedlichen Trends bei den einzelnen mentalen Tests erklären, so wie die Kombination von IQ-

¹ Diese Daten sind bis heute weitgehend gültig, denn zwischenzeitliche Normenkontrollen an meh-reren tausend Schülern aus Hamburg und Baden-Württemberg zeigten, dass in diesem Test im Zeitraum von rd. 20 bis 25 Jahren keine weiteren bedeutsamen Normverschiebungen stattfanden (Näheres hierzu siehe Seite 7ff).

Gewinnen und Verlusten bei schulischen Fähigkeitstests in den Vereinigten Staaten.“
(Flynn, 1987, S.171)

Im Durchschnitt geht Flynn von einem IQ-Gewinn von 0,33 Punkten pro Jahr aus. Als Testentwickler weiß ich nur zu genau, dass man den Wahrheitsgehalt solcher Aussagen nur nachvollziehen kann, wenn man die ‚Urdaten‘ einer kritischen Prüfung unterzieht. Dies ist im vorliegenden Fall - selbst wenn man die Originalarbeit heranzieht - nur bedingt möglich, denn auch Flynn musste sich auf die Daten anderer verlassen, die zwar weltweit erhoben wurden, jedoch nicht alle notwendigen Parameter enthalten. Dabei fiel mir besonders auf, dass der häufig verwendete Raven Progressive Matrices Test teilweise stark modifiziert worden war: So wurde er in **Norwegen mit 36 Aufgaben mit ansteigendem Schwierigkeitsgrad ohne Subtests** administriert, in den **Niederlanden mit 40 Aufgaben, die am besten diskriminieren sollen**, und in einigen anderen Nationen mit allen **60 Aufgaben**. Dabei wurde in Belgien zwar die volle 60 Item-Version verwendet, jedoch die Bewertung verändert, indem falsche Aufgabenlösungen mit einem Strafpunkt bewertet wurden, in Frankreich wurde ungeschultes Militärpersonal bei der Testung der Rekruten eingesetzt und bei den meisten Nationen fehlte der weibliche Bevölkerungsteil. Teilweise lagen auch nur Angaben zu den Quartils- bzw. Perzentilleistungen vor, die von Flynn dann in eine IQ-Skala transformiert wurden – ein äußerst fragwürdiges Verfahren – usw.

Als Ergebnisbeispiel möchte ich die norwegischen Daten heranziehen:



Der Matrizen test nach Raven umfasste nur 36 Items in einem Aufgabenpool mit ansteigendem Schwierigkeitsgrad. Von 1968 bis 1980, in einem Zeitraum von 12 Jahren, wurde lediglich ein IQ-Gewinn von 2,6 Punkten gemessen. Dies sind 0,217 IQ-Punkte pro Jahr. Für die abfallende Kurve der Mathematiktestleistungen machte Flynn folgende Tatsache verantwortlich (nach Rist 1982): "Rist notes that the mathematic test losses began, when students trained in the new math began to reach military age". Also ein vermuteter Einfluss durch den veränderten Lehrplan in Mathematik.²

Abschließende Bewertung zu Flynn:

Die Schlussfolgerung von Flynn, nach welcher der *Raven Progressive Matrices Test* keinesfalls Intelligenz misst, sondern eher ein Korrelat mit einer schwachen ursächlichen Beziehung zur Intelligenz, mag aufgrund der vorgelegten Befunde aus vielen Nationen durchaus zutreffen. Die weitergehende Behauptung jedoch, dass dasselbe auch für alle IQ-Tests gilt, lässt sich durch folgende Gründe widerlegen:

- a) Die Raven-Intelligenz beruht auf einem einzigen Aufgabendesign, nämlich den Matrizenaufgaben.
- b) In der Untersuchung von Flynn sind keine Ergebnisse des Grundintelligenztests erhalten. Nur kurz erwähnt wird eine Untersuchung von Cattell mit dessen Culture Free Intelligence Test, die jedoch nur bis an das Jahr 1950 reichte, andererseits jedoch erbrachte, dass während des 13-jährigen Beobachtungszeitraums nur eine unbedeutende Steigerung von 0,91 IQ-Punkten, also lediglich 0,07 IQ-Punkten pro Jahr, festgestellt wurde. Diese Daten mit dem Cattell-Test wurden jedoch von Flynn verworfen, obwohl die Untersuchungsqualität mit dem ‚Status 1‘ bewertet wurde. Der Grund: Sie stammen aus der Vor-50er Generation.
- c) Bezeichnend für meine Kritik an solchen globalen Feststellungen, wie sie Flynn trifft, sind die post hoc nicht mehr oder nur unvollständig erfüllten Bedingungen für die Repräsentativität der miteinander verglichenen Stichproben oder Erhebungen. Repräsentativität ist aber eine unverzichtbare Voraussetzung für derartige Pretest-Posttest Vergleiche, genauso wie bei der Standardisierung bzw. Normierung eines Testverfahrens. Dazu gehören u.a. vergleichbare Altersgruppen, gleiche Schulartanteile, gleiche Berufsgruppenanteile, gleiches Bildungsniveau, gleiche regionalspezifische Zusammensetzung, gleiche Migrantenanteile und freiwillige oder angeordnete Testteilnahme mit oder ohne Selektionsdruck (Testangst oder keine, Anstrengungsbereitschaft...), vor allem aber gleiche Testbedingungen (Tageszeit, qualifiziertes Testpersonal oder nur angelernte Hilfskräfte). Die Kontrolle dieser Bedingungen war weitgehend nicht oder nur eingeschränkt gewährleistet.

Zu 3.

Kritische Bewertung der Schlussfolgerungen von Zerahn-Hartung (2002): Schwachstellen des zugrunde liegenden Datenmaterials.

In der interessanten Arbeit mit dem Titel: „Normverschiebungen bei Rechtschreibleistungen und sprachfreier Intelligenz“ wurde untersucht, ob sich in einem Generationenvergleich die Rechtschreibleistung und die Grundintelligenz verändert haben. Ergebnis:

² Es ist schon fast eine Selbstverständlichkeit, davon auszugehen, dass exogene Einflüsse sich auch im Sinne einer Leistungssteigerung auf sog. ‚Fluid-Ability-Tests‘ auswirken können. So wie zu vermuten ist, dass sich ein ‚Jauch-Effekt‘ nach besagter TV-IQ-Testung von Millionen Deutschen auf ähnliche Intelligenztestaufgaben im Anschluss an deren TV-Erfahrung positiv auswirken wird, genauso können sich Transferwirkungen durch testnahe schulische Unterrichtseinheiten positiv bemerkbar machen. Als Beispiel: Zu Beginn der 80er Jahre bot sich mir in Hamburger vierten Grundschulklassen bei fast 2000 Schülern die Gelegenheit, die CFT20-Ergebnisse auf Subtestbasis zwischen Schülern mit und ohne ‚Neue Mathematik‘ zu vergleichen. Neue Mathematik enthielt eine Unterrichtseinheit Topologien. Topologische Schlussfolgerungen sind jedoch im Subtest 4 des CFT20 enthalten. Die messbare Folge war eine durchschnittliche Mehrleistung bei dieser Schülergruppe um 1-2 Rohwerte. Dieser Übungserfolg ist jedoch nicht von Dauer. Solche möglichen Einflussgrößen bedürfen immer einer ergänzenden explorativen Diagnostik, genauso wie es selbstverständlich ist, nach Testvorerfahrungen zu fragen.

Die Rechtschreibleistungen haben sich sehr stark verschlechtert, die Grundintelligenz (general fluid ability, gemessen mit dem CFT 20 -Teil1) habe sich bedeutsam gesteigert. Im folgenden geht es um die Steigerung der Intelligenzleistung: [Meine Kritik an der Untersuchung von Zerahn-Hartung et al. \(2002\)](#) richtet sich insbesondere darauf, dass der CFT 20-Teil 1 einen doppelten ‚Flynn-Effekt‘ nach 20 Jahren erbringe.

Kritikpunkte an der Stichprobe 1995 von Zerahn-Hartung et al. (2002):
a) Unterschiedliche Altersbereiche
b) Nicht repräsentative Schulartenanteile
⇒ Gesamtstichprobe und Anteil der Tpn ‚ohne Hauptschulabschluss‘
⇒ Messfehler für die Gruppe ‚ohne Hauptschulabschluss‘
c) Fehlende Ausländer
d) Fehlende Tpn aus ländlichen Regionen (nur Raum Mannheim-Heidelberg)
e) Keine Angaben über Testvorerfahrungen der Tpn

Von den einzelnen Kritikpunkten kann ich an dieser Stelle aus Platzgründen nur die Positionen a) bis c) differenzierter begründen:

Zu a): Unterschiedliche Altersbereiche mit nicht repräsentativen Schulabschlussanteilen

Bei einer den Handbuchnormen in etwa entsprechenden Einteilung in zwei Altersgruppen treten in dieser Versuchsstichprobe bedeutsame Unterschiede auf:

CFT20/2-Normen (1977/1987)		Versuchsstichprobe 1995	
Altersbereich	RW-Mittelwert (s)	Altersbereich	RW-Mittelwert (s)
15;1-19;0 Jahre	30,8 (6,4)	16-19 Jahre	33,75 (5,76)
20-29 (CFT2 extrapol.)	30-31 (5,8)	20-29 Jahre	36,18 (4,96)

Die wesentlichen Schlüsse für eine von den AutorInnen postulierte Normverschiebung werden aus einem Vergleich der 15;1 bis 19;0-Jährigen aus der Normierungsstichprobe 1977 und der Gesamtstichprobe des Versuchs 1995 gezogen, die aber den Altersbereich von *16 bis 30 Jahren* umfasst. Dieser soll unkorrigiert 34,5 RW bzw. rd. 110 IQ-Werte betragen, also um 3,7 RW bzw. 10 IQ über den Normierungsdaten von 1977 liegen. Hauptgrund der Differenzen dürfte wahrscheinlich in einer mangelnden Repräsentativität der Berufsgruppen in der Versuchsstichprobe, insbesondere bei den 20-29 Jährigen, liegen.

Warum die Trennung der Altersgruppen trotz meiner Einwände, die ich vor der Publikation mitgeteilt hatte, unberücksichtigt blieb, ist nicht nachvollziehbar.

Das Versuchsergebnis von Zerahn-Hartung mit einer bedeutsamen Steigerung des IQ-Wertes von der Gruppe 16-19 Jahre zur Gruppe 20-29 Jahre um rd. 7 Punkte (RW-Diff.= + 2.42) kann nur als statistisches Artefakt erklärt werden, da er wahrscheinlich durch eine *mangelnde Repräsentativität der Berufsgruppen in der Versuchsstichprobe* zustande kam. Da mir die dazu erforderlichen ungefilterten Daten nicht zur Verfügung gestellt wurden, vermute ich, dass besonders in der Stichprobe der 20- bis 29-Jährigen *Berufe ohne qualifizierten Schulabschluss* zu schwach vertreten sind. Bei einer Differenzierung nach Ausbildungsrichtungen für gewerbliche, kaufmännische und hauswirtschaftliche Bereiche fällt auf, dass *der hauswirtschaftliche mit 1,8% zu schwach vertreten ist.*

Zu b): Nicht repräsentative Schulartanteile

Dieser Einwand meinerseits wurde von den AutorInnen aufgegriffen und ein Korrekturwert von $-0,4$ RW in eine Korrekturtabelle einbezogen. Eine solche Berechnung ist aber nur dann zulässig, wenn alle gewichteten Gruppen ausreichend besetzt sind. Dies ist mit einem $N= 11$ bei den Tpn ohne Hauptschulabschluss jedoch nicht gegeben. Deshalb fehlt den Gewichtungsberechnungen eine ausreichende empirisch-statistische Basis.

Nach eigenen neueren Untersuchungen liegt der Wert für diese Gruppe, die sich etwa zur Hälfte auf ehemalige Sonder-/FörderschülerInnen bezieht, mit großer Wahrscheinlichkeit erheblich niedriger als der ermittelte Wert von $27,8$ RW (IQ=94). Er dürfte höchstens bei einem RW von 23 liegen, was eine weitere Korrektur um $-0,3$ RW bedeutete.

Zu c): Fehlende Ausländeranteile

Dies dürfte sich besonders auf die Gruppen mit und ohne Hauptschulabschluss im Sinne einer Erhöhung des durchschnittlichen CFT- Ergebnisses auswirken, denn der Ausländeranteil reicht in Großstädten bis 25% (allgemeinbildende und berufliche Schulen Baden-Württembergs $12-13\%$; Statistisches Bundesamt 2003, S.103 bzw. S.19). Die Auswirkung auf Teil1 des CFT 20 ist dabei besonders hoch (fehlender Ausländeranteil bedeutet Erhöhung der Testdurchschnittswerte in der Stichprobe), während sich für Teil 2 die Testleistungen den vergleichbaren deutschen sozio-ökonomischen Gruppen angleichen (s. Weiß, 1998, Seite 66ff), sich also nicht mittelwertssteigernd auswirken würden. Aber Teil 2 wurde in der Versuchsstichprobe von Zerahn-Hartung nicht erhoben.

Bilanz der Rohwertkorrektur für die Stichprobe 1995 (Zerahn-Hartung):

a) Unterschiedliche Altersbereiche (nur bedingt vergleichbare Schulabschlussgruppen)	
b) Korrektur für nicht repräsentative Schulartenanteile	
⇒ Gesamtstichprobe mit empirischem Wert für ‚ohne Hauptschulabschluss‘	- 0,4 RW
⇒ Messfehlerkorrektur für Gruppe ‚ohne Hauptschulabschluss‘ (Mindestkorrektur)	- 0,3 RW
c) Korrektur wegen fehlender Ausländer (Mindestkorrektur)	- 0,4 RW
d) Korrektur ‚fehlende ländliche Regionen‘	- 0,4 RW
e) Korrekturschätzung durch Testvorerfahrungen bei etwa 10% der Tpn	- 0,2 RW
Summe	- 1,7 RW

Bezogen auf den berechneten Wert für die mit der Normierungsstichprobe in etwa vergleichbaren Altersgruppe der 16 bis 19-Jährigen, der bei $33,7$ RW liegt, kann man als realistisch einen durchschnittlichen Rohwert von $33,7 - 1,7 = 32,0$ ansehen. Dies läge nur noch um $1,2$ RW über dem 1977 ermittelten Rohwert von $30,8$. Nach der Altersnormtabelle im Handbuch auf Seite 85 entspricht dies einem IQ-Wert von rd. 103 bzw. einem T-Wert von 52 . Dies wurde nach obigen Einwänden auf der Basis einer geschätzten Mindestkorrektur für die 16 bis 19-Jährigen berechnet.

Zu 4.

Neuere Befunde zum Grundintelligenztest CFT:

a) Hamburger Untersuchungen in 5. Klassen

In den vergangenen Jahren konnten aus mehreren Untersuchungsprojekten bei jüngeren Probanden für den CFT 20 (4. und 5. Klassen) keine Veränderungen im Vergleich zu den Normierungsdaten vor 20 Jahren entnommen werden. Allein für den CFT 20 wurde dabei die Normenstabilität an **12.300 Schülern** aus Hamburg voll bestätigt (s. Lehmann & Peek, 1996).

Kontrolle der Normen aus 1977 durch Daten aus 1996 für die 5. Klassenstufe (alle Schularten)		
Mittelwertsvergleich für den Rohwert aus Teil 1 des CFT 20		
CFT 20 Teil 1	Normierungsdaten 1977	Kontrolldaten 1996 (Hamburg)
Erhebungsumfang N	521	12.330
Erhebungsmonat	Februar	September
Ausländeranteil	ca.10 % (wie BRD)	20%
Durchschnittlicher Rohwert Teil 1 = real	27,6	26,5
nach Korrektur durch Alter + Ausländeranteil	(27,6)	27,52

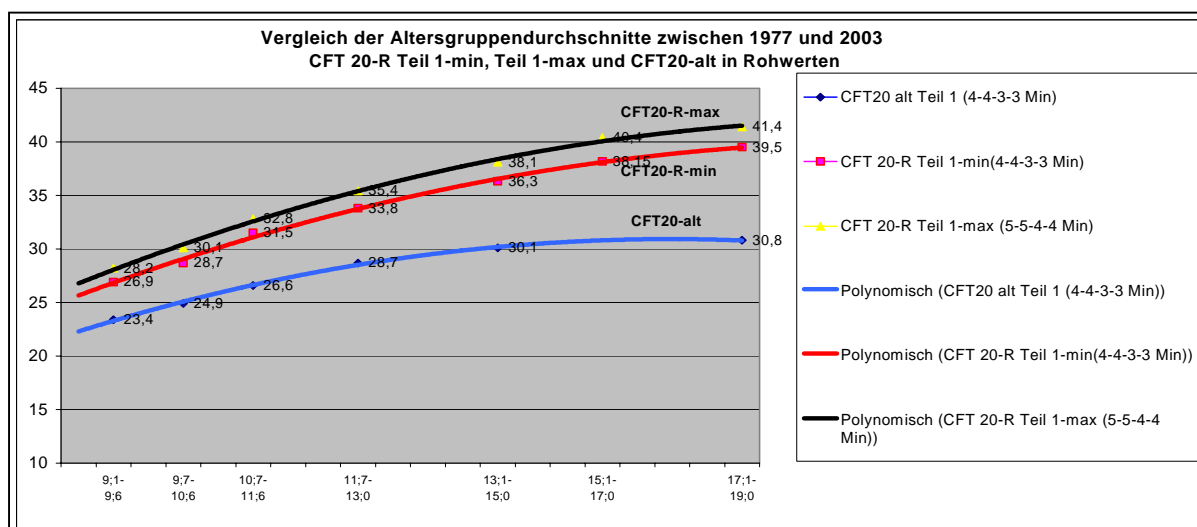
Ergebnis:

Der anlässlich der Eichung des Tests im Jahre 1977 ermittelte durchschnittliche Rohwert für den Teil 1 des CFT 20 beträgt für die 5. Klassenstufe über alle Schularten 27,6 Punkte (s. Handanweisung Seite 49). Für den Stadtbezirk Hamburg wurden für die 5. Klassen aller Schularten aus zwei Schulamtsbezirken (N = 2.034) im Jahre 1995 26,4 Punkte erzielt (Lehman & Peek, 1996, S.29). Nach den CFT1 -Teilanalysen des Forschungsprojekts mit der Totalerhebung in allen 5. Klassen aller Schularten (N = 12.330) im Jahre 1996 wurden 26,5 Rohwerte erzielt. Die 1995 und 1996 ermittelten Mittelwerte für den Teil 1 liegen also um 1,1 bzw. 1,2 Rohwerte niedriger als die 19 Jahre früher ermittelten Werte aus der Eichstichprobe.

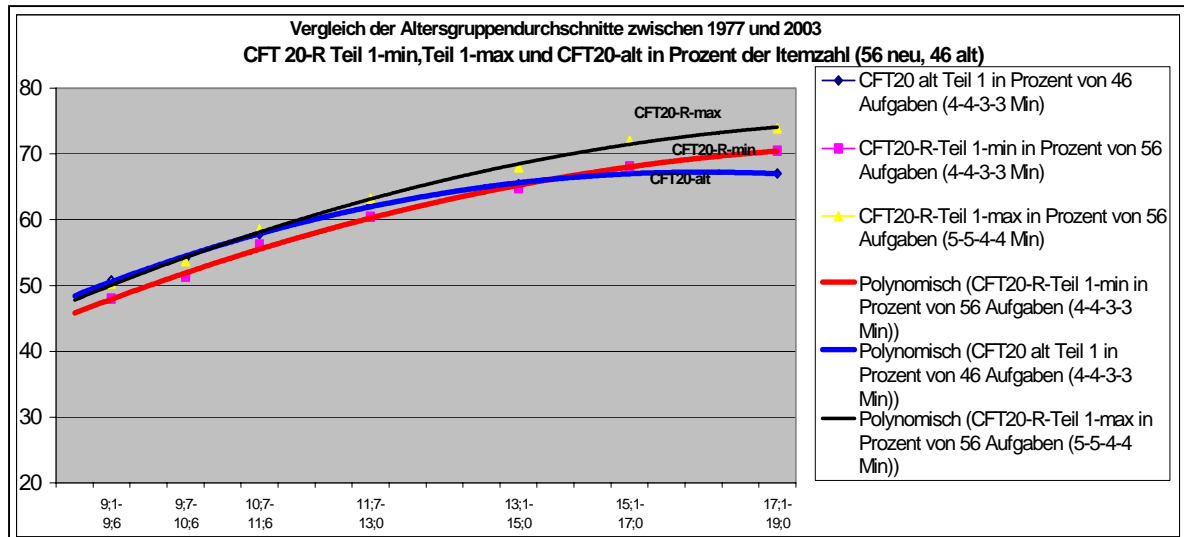
Nach der erforderlichen Alterskorrektur und der Korrektur durch den erhöhten Ausländeranteil kann festgestellt werden, dass die Kontrolldaten sowohl aus der Voruntersuchung 1995 als auch aus der Erhebung in den Hamburger Schulen aus dem Jahre 1996 nur unwesentlich von den Normierungsdaten aus dem Jahre 1977 abweichen. *Die durchschnittlichen Rohwerte für den Teil 1 können mit 27,4 bzw. 27,5 als nahezu identisch betrachtet werden.*

b) Ergebnisse aus der Normierung des weiterentwickelten CFT 20-R + WS/ZF

⇒ Vergleich der Altersgruppen-Durchschnitte zwischen CFT 20 (1977) und CFT 20- R (2003)



Obiger Alterskurvenvergleich könnte fälschlicherweise zu der Annahme führen, dass ein IQ-Gewinn in den vergangenen 25 Jahre stattgefunden habe. Man muss jedoch dabei die unterschiedlichen Itemzahlen der beiden Testversionen berücksichtigen, denn durch die Erhöhung der Itemzahl in Teil 1 des CFT20-R von 46 auf 56 und etwa gleichbleibendem Schwierigkeitsgrad bei mehr als der Hälfte der Items innerhalb der Subtests, ist natürlich die Chance gegeben, auch mehr Treffer zu erzielen. Deshalb sind die Leistungen in Bezug auf die vorgegebene Itemzahl in den beiden Testversionen (1977 versus 2003) zu vergleichen:



Bei diesem Prozentvergleich können mit Ausnahme der 17-19-Jährigen keine besonderen Unterschiede mehr festgestellt werden. Dieser Effekt war jedoch Konstruktionsziel für den neuen Grundintelligenztest CFT 20-R, weil damit der Nachweis erbracht wurde, dass diese Neuentwicklung besser in der Lage ist, bei älteren Tpn und höheren Intelligenzgruppen die Differenzierung im oberen Leistungsbereich zu erhöhen.

Nach diesen Ergebnissen konnte **keine Bestätigung für einen ‚Flynn-Effekt‘** in der behaupteten Höhe gefunden werden.

b) Auch für Wortschatz- und Zahlenfolgentest (Ergänzungstests zum CFT 20) ergaben sich im Zeitraum von 1986 bis 2003 nach vorläufigen (noch nicht vollständig auf Repräsentativität untersuchten) Ergebnissen bei rd. 3000 Schülern keine bedeutsamen Normverschiebungen.

Konsequenzen:

Eine Neunormierung ist aufgrund meiner Ergebnisse und Korrekturhinweise für die Altersgruppen 8;5 bis 15 Jahren nicht erforderlich. Sie wäre auch für die 16- bis 19-Jährigen nicht geboten, weil der Differenzwert von 1,2 RW (s. Zerahn-Hartung 2002, S. 6) innerhalb des Messfehlerbereichs liegt. Auch eine eigene Normtabelle für Personen mit deutscher Muttersprache ist nicht zu empfehlen, zumal derartige ethnische Normierungen konträr zu den Integrationsbemühungen stehen und eine Stigmatisierung bestimmter Migrantengruppen verstärken. Hier bietet der CFT 20 durch den Teil 2, der im Projekt von Zerahn-Hartung nicht durchgeführt wurde, eine sehr gute Basis zur Kompensation sprachlicher Probleme beim Verstehen der Testinstruktion (siehe Handbuch zum CFT20, Seite 71 unten). Im übrigen gilt dies in gleicher Weise für deutschsprachige Personen aus einem spracharmen Milieu. Die Differenz von rd. drei IQ-Werten läge - auf einen Zeitraum von 20 Jahren verteilt - mit 0,15 IQ-Punkten pro Jahr unter dem sog. Flynn-Effekt mit 0,33 IQ und nicht darüber, wie in der Arbeit von Zerahn-Hartung u.a. dargestellt.

Die Ergebnisse der experimentellen Studie zur Veränderung der Rechtschreibleistung werden durch meine Korrektur der sprachfreien Intelligenztestleistung nicht bedeutsam geschmälert. Denn dort ist der Verlust an Rechtschreibkompetenz in einem Zeitraum von 30 Jahren ohnehin eklatant und wegen der Stichprobenszusammensetzung mit Überrepräsentation

tion der sozio-ökonomisch ‚höheren‘ Schichten eigentlich geringer zu erwarten. Die abwertende Interpretation zum ‚Culture Fair Konzept‘ und der ‚Fluiden Intelligenz‘ des CFT 20 (Zerahn-Hartung et al. 2002, S. 13) ist allerdings zu relativieren, vor allem hinsichtlich des behaupteten Lerneffekts durch elektronische Medien. Der vermutete positive Effekt auf die kognitive Leistung müsste bereits wesentlich früher auftreten, denn aus der Medienforschung wissen wir (Weiß 2000), dass eine enorme Konsumsteigerung bereits in der Grundschule oder Sekundarstufe I beobachtet wurde. Und genau in diesem Altersbereich fanden mit Ausnahme leichter Verbesserungen beim Subtest ‚Ähnlichkeiten‘ auch im CFT 1 keine kognitiven Leistungssteigerungen in den vergangenen 25 Jahren statt.³ Diese Aufgabenform, in der es darum geht, Ähnlichkeiten bei Figuren zu erkennen, ist aber beim CFT 20 nicht mehr enthalten.

Fazit: Es ergibt sich keine zwingende Notwendigkeit für eine Neunormierung des CFT 20. Dies trifft nach den referierten Kontrolluntersuchungen für die Kurzform (Teil 1) zu. Es kann jedoch mit großer Wahrscheinlichkeit angenommen werden, dass dies auch für die Langform gilt. Hinweise dazu gibt es aus Untersuchungen in Baden-Württemberg aus dem Jahr 1989/90 für die 4. Grundschulklassen⁴ sowie aus den Vergleichen mit den Normierungsdaten zur revidierten Fassung des CFT 20-R aus dem Jahre 2003.

zu 5.

Die eingangs gestellte Frage, ob wir immer klüger werden, kann nach diesen Ergebnissen sicher nicht mit ja beantwortet werden.

Inwiefern hat die sprachfreie IQ-Diagnostik in der schulpsychologischen Praxis noch immer ihre Berechtigung und welche Weiterentwicklungen gibt es? Aus Zeitgründen kann ich diese Frage nur im Hinblick auf das Problem der Normenstabilität und der Anforderung einer fairen Intelligenzmessung beantworten, nicht jedoch hinsichtlich genereller Validitätsprobleme.

Testintelligenz und Schule:

*Systemisches Denken in der Schule muss gut ausgeprägt sein, damit ich in diesem hochkomplexen System mit vielen Untersystemen erfolgreich sein kann. Jeder Lehrer, jedes Fach, jede Klasse stellt ein eigenes System dar. Je absoluter ein solches System behandelt wird oder sich ausbildet (z. B. neigen viele Lehrer dazu, ihr eigenes Fach als absolut wichtig anzusehen), je mehr der Lehrer mit seinem eigenen Fachsystem identifiziert ist und je weniger er in der Lage ist, sich in andersgeartete ‚Schüler-Denksysteme‘ hinein zu versetzen, umso mehr wird es zu ‚Missverständnissen‘, Fehlbeurteilungen und Konflikten kommen (mangelnde ‚intrapersonale Intelligenz‘ nach Gardner). Früher sagte man, ein guter Schüler muss gut angepasst sein, damit er erfolgreich ist. Nach meinem jetzigen Wissen über Denksysteme in Verbindung auch mit dem, was man mit *Fluid-Ability* meint, oder auch Gardner mit seinen *sieben Intelligenzen* und der darin enthaltenen Dimension ‚interpersonale Intelligenz‘ erklärt, ist die Fähigkeit in einem System zu hoher Performance zu kommen davon abhängig, wie rasch jemand in der Lage ist, sich in einem neuen System zurecht zu finden und seine mehr oder weniger komplexen Strukturen zu erkennen. Gelingen dem Schüler diese dauernden Anpassungen und Anpassungsstrategien nicht, so wird er vom System Schule ausgestoßen. Die Selbstattribuierung des Schülers „Ich bin nicht genügend befähigt“, wäre besser zu ersetzen durch „zu wenig systemangepasst“.*

³ Für den **CFT 1** wurden im Vergleich zu den Daten aus 1976 bei etwa 1200 Kindern aus Sachsen und Baden-Württemberg für die wichtigsten Problemlösungsbereiche des Tests ebenfalls keine bedeutsamen Veränderungen gefunden. Zitat aus dem Handbuch: **„Entgegen der Erwartung hat sich eine deutliche Bestätigung für die Gültigkeit der Normen herausgestellt“** (Weiß & Osterland, 1997, Seite 3). Wir konnten keinen sog. ‚Flynn-Effekt‘ feststellen.

⁴ siehe Weiß, 1997, S.58 sowie Weiß u. Osterland, 1996.

Schule ist also ein eigenes System, in dem man am besten reüssiert, wenn man eine hohe ‚Schulintelligenz‘ besitzt. Es gilt aber auch:

„Jeder Test ist ein eigenes System“.

Übertragen auf einen kognitiven Test bedeutet dies, dass seine Gestaltung und Darbietung für die unterschiedlichen Voraussetzungen der Menschen, die sich diesem Test unterziehen sollen oder möchten, so gerecht elaboriert wird, dass mehr oder weniger zufällige Handicaps den Zugang zu diesem Testsystem nicht erschweren. Es sollte zum einen eine angemessene Einarbeitungsphase vorhanden sein, damit Tpn, die noch keine Berührung mit diesem oder einem ähnlichen Verfahren hatten, sich an das neue „System“ gewöhnen können, d.h., eine noch nicht vorhandene „Testsophistication“ sollte erst geschaffen werden. Mit anderen Worten, der „Nürnberger Trichter“ muss erst eine genügend weite Öffnung in das Gehirn besitzen, bevor man oben etwas hineinschütten kann, um überhaupt messen zu können, wie das alles darin verarbeitet wird. Im CFT 20 habe ich deshalb den Teil 1 vorgeschaltet, um in der Lage zu sein, feststellen zu können, wie und wie erfolgreich im Teil 2 die Aufgaben gelöst werden. Zum anderen wird dadurch eine Chance gewährt, ohne das Medium Sprache zu beherrschen, die Problemstellungen des Tests einwandfrei wahrnehmen zu können. Probanden sollen ja nicht für mangelnde Testerfahrung und Sprachkenntnisse bestraft werden.

Erst dann ist man in der Lage, die Ausprägung der individuellen Fähigkeit, Sprünge zwischen den Denksystemen des Tests, durch unterschiedliche Aufgabenarten wie Einzelaufgaben, erfassen und in seinem Umfang bestimmen zu können. Nur so kann dies zuverlässig geschehen. Und nur dann kann man einen Test auch als gerecht bzw. fair bezeichnen.

In früheren Faktorenanalysen bei der Entwicklung der CFT-Reihen habe ich immer wieder einen Faktor gefunden, den wir Interferenz-Faktor nannten. Man kann ihn auch mit ‚Störanfälligkeit‘ oder ‚Störbarkeit‘ umschreiben. Sein Anteil an der testspezifischen Varianzaufklärung betrug etwa 20-25 %. Gute Fähigkeiten in diesem Sinne kommen aber nur dann zustande, wenn auch ein ausreichend großer kortikaler Arbeitsspeicher vorhanden ist, in dem Zwischenergebnisse so lange gespeichert werden können, bis sie für den nächsten Lösungsschritt gebraucht werden. Ich vermute, dass eine erhöhte Störanfälligkeit bei Testpersonen verhindert, dass wichtige zwischengespeicherte Informationen wieder abgerufen werden können, weil sie durch andere irrelevante Einflüsse überlagert werden. Bei pathologischen Intelligenzhemmungen bzw. extrem erhöhter Störanfälligkeit persönlichkeitspezifischer Genese, ist dies augenscheinlich. Im heutigen ‚Normalfall‘ treten aber immer häufiger Störanfälligkeiten zutage, die auf exogene Einflüsse zurückzuführen sind: Traumatische Ereignisse, Überflutung durch AV-Medien, mediale Gewaltbelastungen, massive Konflikte in der Klasse, gestörte Beziehungen zwischen Lehrer und Schüler, Schulangst u.a. Auch diese sind relevante Einflussgrößen, die vorher durch entsprechende Interviewmethoden eingegrenzt und bestimmt werden müssen.

Das Problem der mangelnden Übereinstimmung von *Schulintelligenz* und *Testintelligenz* bleibt dennoch bestehen. Eine weitere Diskussion an dieser Stelle würde jedoch den Rahmen hier sprengen. (Eine etwas amüsante Empfehlung für den (die) Leser(in): lesen Sie nochmals die ersten beiden Absätze unter ‚*Jeder Test ist ein eigenes System*‘ bis zu ‚...fair bezeichnen‘, und setzen Sie anstelle des Wortes *Test* das Wort *Schule* ein, dann werden Sie merken, warum Intelligenztestleistung und Schulleistung oftmals nicht übereinstimmen).

Nach der PISA-Studie wurde besonders in der IGLU-Studie hervorgehoben, dass eine Verbesserung der Chancengerechtigkeit in unserem Bildungssystem für Migrantenkinder und andere sozial benachteiligte Schüler dringend geboten ist. In diesem Zusammenhang spielen m.E. kulturfaire Verfahren in der Schullaufbahnberatung eine herausragende Rolle. Deshalb möchte ich abschließend noch eine Bemerkung machen zum Grundintelligenztest CFT20:

Weiterentwicklungen des Testkonzeptes zum CFT 20-R.

Zur Verbesserung der Differenzierung im oberen Leistungsbereich und Erhöhung der ‚Testdecke‘ wurden von mir bereits im Jahre 1994 für den CFT 20 22 neue Items mit überwiegend höherem Schwierigkeitsgrad konstruiert. Diese Items wurden in das Testheft integriert und dann als ‚Forschungsversion‘ an mehreren hundert Testpersonen evaluiert. Unterschied-

liche Testzeiten im Teil 1 mit eigener Normierung ermöglichen zudem eine Verlängerung der Bearbeitungszeit bei langsamer arbeitenden Tpn. Für diese Testerweiterung wurden inzwischen die repräsentativen Normierungsuntersuchungen bei rd. 4.400 Schülern aller Schularten aus sechs Bundesländern abgeschlossen (2003). Dieser CFT 20-R kann ebenso wie die bisherige Standardform des CFT 20 auch als Computerprogramm eingesetzt werden (erscheint 2005). In Versuchsreihen mit drei Feldexperimenten in den Schularten Gymnasium, Hauptschule/Werkrealschule und Förderschule fand ich keine wesentlichen Veränderungen der Testkennwerte zwischen **Paper-Pencil und PC-Testung**. Sie sind weitgehend äquivalent. Mittels Einzelfallbeobachtungen konnte ich darüber hinaus feststellen, dass die PC-Version durch die singuläre Darstellung der Einzelitems auf dem Bildschirm gerade für ADS-Kinder und –Jugendliche größere Vorteile bietet, weil sie dabei nicht durch benachbarte andere Items irritiert werden. Mit systematischen Vergleichen in einer Therapieeinrichtung (PIA-Schorndorf) wurden bereits begonnen.

Literatur

- Arnold, W (1960):** Begabungswandel und Erziehungsfragen, Reinhard, München.
- Flynn, J. R. (1987):** Massive IQ gains in 14 Nations: what IQ tests really measure. Psychol. Bulletin 101:171-191.
- Flynn, J. R. (1994):** IQ gains over time. In R.J. Sternberg (Ed.), Encyclopedia of human intelligence. New York: Macmillan, S. 617-623.
- Gilbert, G. M. (1962).** Nürnberger Tagebuch. Gespräche mit Angeklagten. Fischer.
- Lehman, R.H. und Peek, R (1996):** Aspekte der Lernausgangslage von Schülern und Schülerinnen der fünften Jahrgangsstufe an Hamburger Schulen. Forschungsbericht. Humboldt-Universität Berlin.
- Royl, W. & Schwarzer, R. (1976):** Zur säkularen Akzeleration der PSB-Intelligenz. Diagnostika, 22, 99-104.
- Statistisches Bundesamt (2003):** Bildung und Kultur, Allgemeinbildende Schulen (Reihe 1) und Berufliche Schulen (Reihe 2), Fachserie 11, Schuljahr 2002/03, Wiesbaden
- Weiß, R. H. (1998):** Grundintelligenztest Skala 2 (CFT 20). 4. überarbeitete Auflage, Göttingen: Hogrefe.
- Weiß, R. H. & Osterland, J. (1997):** Grundintelligenztest Skala 1. 5., revidierte Auflage, Göttingen: Hogrefe.
- Weiß, R. H. (2000):** Gewalt, Medien und Aggressivität bei Schülern. Hogrefe: Göttingen.
- Zerahn-Hartung, C. (2002):** Normverschiebungen bei Rechtschreibleistungen und sprachfreier Intelligenz. Praxis der Kinderpsychologie und Kinderpsychiatrie.

Anschrift des Verfassers
Dr. Rudolf H. Weiß
Dipl. Psych. (BdP)
Drosselweg 13
D-71549 Auenwald

vorm. St. Prof.
beim Oberschulamt Stuttgart
Leitender Schulpsychologe
e-mail: RHWEISS@t-online.de
Fax: 07191-58615