Arch. Psychiat. Nervenkr. 216, 172—187 (1972)

Die diagnostische Effizienz von drei psychologischen Verfahren zur Auslese hirngeschädigter Patienten

W. Hartje und B. Orgass

Abteilung Neurologie der Medizinischen Fakultät
der Rheinisch-Westfälischen Technischen Hochschule Aachen
(Vorstand: Prof. Dr. K. Poeck)

Eingegangen am 27. März/10. Juni 1972

Diagnostic Efficiency of Three Psychological Methods for the Detection of Brain-Damaged Patients

Summary. Three psychological methods for the detection of brain damage in patients were examined for diagnostic efficiency in a neurological setting: these were the Visual Retention Test (Benton, 1968), the Visual Organization Test (Hooper, 1958) and the comparison between estimated premorbid and actual postmorbid IQ.

The sensitivity of each test was checked individually and then that of an optimal combination. The diagnostic efficiency was compared with the results obtained by guessing based only on a knowledge of the incidence of brain damage in the population being examined. There was also some attempt to establish whether the risks associated with a wrong diagnosis are less when psychodiagnostic techniques are applied than when the diagnosis is guessed at.

The results suggest that these psychological screening methods for the detection of brain damage cannot be used to any practical advantage in a neurological hospital. This is partly because subjects with brain damage make up the vast majority of any population of neurological patients, and partly because the risk associated with false negative diagnosis is relatively high in the case of brain damage. The results obtained by other investigators are shown to be in agreement with these observations if population base-rates and cost-efficiency are taken into account.

Key words: Psychological Screening — Brain-Damage — Diagnostic Efficiency — Population Base-Rates — Cost-Efficiency.

Zusammenfassung. Drei psychologische Methoden zur Erfassung hirngeschädigter Patienten wurden auf praktische Bewährung in einer neurologischen Klinik überprüft: der Visual Retention Test (Benton, 1968), der Visual Organization Test (Hooper, 1958) und der Vergleich zwischen prä- und postmorbidem IQ.

Außer der Trennschärfe der Einzelmerkmale wurde auch die Trennschärfe einer optimalen Kombination der Merkmale untersucht. Die Effizienz der Tests wurde verglichen mit dem Erfolg bei „gezieltem Raten", das nur die Kenntnis der Grundhäufigkeiten hirngeschädigter und nichthirngeschädigter Patienten in der unter-

suchten Population voraussetzt. Es wurde geprüft, ob die Anwendung der psychodiagnostischen Verfahren die mit falschen diagnostischen Entscheidungen verbundenen Risiken gegenüber gezieltem Raten verringert.

In einer neurologischen Klinik ist die Anwendung solcher psychodiagnostischen Verfahren zur Auslese hirngeschädigter Patienten ohne praktischen Nutzen. Das liegt einmal daran, daß in der zu untersuchenden Population die hirngeschädigten Patienten stark überwiegen, zum anderen daran, daß falsche Testbefunde bei Hirngeschädigten mit relativ hohen Risiken verbunden sind. Es wird gezeigt, daß die Resultate anderer Untersucher zu demselben Schluß führen, wenn diese Faktoren berücksichtigt werden.

Schlüsselwörter: Psychologische Verfahren — Auslese von Hirngeschädigten — Diagnostische Bewährung — Basisraten — Kosteneffizienz.

Einer der häufigsten Gründe, aus denen neuro-psychiatrische Patienten zur psychologischen Untersuchung vorgestellt werden, ist der Verdacht auf eine hirnorganische Schädigung. Mit psychologischen Methoden kann man zwar nur indirekte Anzeichen einer Hirnschädigung erfassen. Diese Methoden gelten jedoch als diagnostisch nützlich, weil sie hirnorganisch bedingte Leistungsmängel aufdecken können, die sonst unbemerkt blieben. Es gibt im wesentlichen drei Möglichkeiten, solche Leistungsmängel nachzuweisen:

1. den globalen Vergleich der gegenwärtigen Leistungen eines Patienten mit seinem prämorbiden Leistungsniveau;
2. den Vergleich bestimmter Leistungen, die besonders empfindlich auf Hirnschädigungen reagieren, mit anderen Leistungen, die dafür relativ unempfindlich sind;
3. den Nachweis spezieller, umschriebener Leistungsdefekte, die praktisch nur nach einer Hirnschädigung auftreten.

Der globale Vergleich prä- und postmorbider Leistungen ist methodisch am einfachsten, er erscheint aber auch von vorneherein am wenigsten effizient. Das prämorbide Leistungsniveau eines Patienten ist in der Regel nicht genau bekannt und kann nur aufgrund von Bildungsgang und Berufsstellung geschätzt werden. Eine solche Schätzung muß konservativ sein und wird eine Schätzung des Mindestniveaus darstellen, das der Patient besessen haben muß, um einen bestimmten Bildungsstand oder eine bestimmte berufliche Qualifikation zu erreichen. Deshalb kann man auf diese Weise nur eine sehr auffällige Leistungsminderung sicher erfassen.

Die beiden anderen Ansätze scheinen mehr zu versprechen. Aus ihnen wurden denn auch seit etwa 40 Jahren zahlreiche spezielle Tests oder besondere Auswertungstechniken entwickelt, die zum psychodiagnostischen Nachweis hirnorganischer Schäden dienen. Als bekannteste Beispiele seien hier nur Lauretta Benders „Visual Motor Gestalt Test" und Wechslers „Abbauquotient" genannt.

Problemstellung

Eine Durchsicht der Literatur zeigt, wie problematisch die Diagnose einer Hirnschädigung mit psychologischen Methoden immer noch ist. Von den zahlreichen vorgeschlagenen Verfahren haben sich die meisten nicht bewährt und kamen außer Gebrauch. Über die heute noch verwendeten Methoden liegen annähernd gleichviele positive wie negative Urteile vor. Die unterschiedliche Beurteilung beruht z. T. darauf, daß die diagnostische Bewährung dieser Verfahren nicht unter gleichartigen Bedingungen geprüft wurde. Nicht selten erfolgen Bewährungskontrollen unter Bedingungen, die von vornherein viel günstiger sind als die Verhältnisse bei der praktischen Anwendung solcher Ausleseverfahren.

Ein kritischer Punkt ist die Wahl einer adäquaten Kontrollgruppe. Gesunde Versuchspersonen, seien es Studenten, Pflegepersonal oder auch zufällig ausgewählte Stichproben aus der Gesamtbevölkerung, erlauben sicher keinen realistischen Vergleich. Sie unterscheiden sich von neuropsychiatrischen Patienten von vornherein im allgemeinen Wohlbefinden, in der Belastbarkeit, der Motivation und anderen Faktoren, die sich auf das Leistungsniveau auswirken. Dasselbe gilt auch für manche Patientengruppen, besonders für ambulante Patienten ohne stärkere Beschwerden.

Aber auch die hirngeschädigte Vergleichsgruppe kann mehr oder minder adäquat gewählt sein. Besteht sie beispielsweise aus Anstaltspatienten mit deutlichen Abbauerscheinungen oder aus schwer hirngeschädigten Patienten, bei denen die Diagnose nie fraglich war, dann wird man mit jedem beliebigen psychodiagnostischen Verfahren eine unrealistisch hohe Trefferzahl erzielen.

Das Ergebnis solcher Untersuchungen kann nur dann gültig sein, wenn die Bewährungskontrolle unter denselben Bedingungen erfolgte, unter denen das diagnostische Verfahren in der Praxis angewandt wird. Das verlangt nicht nur eine adäquate Wahl der Kontroll- und der Vergleichsgruppe; auch das Verhältnis zwischen der Größe beider Gruppen kann das Ergebnis entscheidend beeinflussen. Diesen wichtigen Punkt haben Meehl u. Rosen (1955) ausführlich erörtert. Sie haben das übliche Vorgehen kritisiert, ohne weitere Überlegung einfach gleichgroße Gruppen zu vergleichen. Ihre Untersuchung demonstriert, wie sehr sich die prozentuale Häufigkeit richtiger Testbefunde ändern kann, wenn die Proportion zwischen Hirngeschädigten und Nichthirngeschädigten variiert wird. Die praktische Bewährung eines Ausleseverfahrens läßt sich nur dann zutreffend beurteilen, wenn bei der Kontrolluntersuchung beide Untergruppen mit demselben Anteil vertreten sind wie unter den Patienten, bei denen man das Verfahren in der Praxis anwendet.

Das von Meehl u. Rosen erörterte Problem schließt ein, daß man keine allgemeingültige Aussage über die praktische Bewährung solcher Auslese-

verfahren machen kann. Ein und dasselbe Verfahren kann von unterschiedlichem Nutzen sein, wenn es auf Patientengruppen angewandt wird, deren Zusammensetzung verschieden ist. Wir haben deshalb die Effizienz verschiedener psychologischer Methoden zur Auslese hirngeschädigter Patienten unter den Bedingungen geprüft, unter denen sie in einer neurologischen Klinik verwendet werden, um zu untersuchen, ob sie hier nachweisbar nützlich sind.

Psychodiagnostische Verfahren

Es wurden drei verschiedenartige Methoden überprüft:

1. Das (mutmaßliche) prämorbide Intelligenzniveau wurde mit dem bei einer psychologischen Untersuchung in der Klinik ermittelten Intelligenzquotienten (IQ) verglichen. Zu diesem Zweck wurde der prämorbide Mindest-IQ (IQ′) eines jeden Patienten aufgrund seiner Schulbildung geschätzt. Der Schätzung lagen publizierte Angaben über den Durchschnitts-IQ bei unterschiedlichem Schulabschluß zugrunde. Als Schätzwerte wurden benutzt:

bei Sonderschulabschluß	IQ′ = 75
bei Volksschülern, die 1—2 Klassen wiederholen mußten	IQ′ = 85
bei regulärem Volksschulabschluß oder höherer Schulbildung bis zur Mittleren Reife	IQ′ = 100
bei Abitur	IQ′ = 110

Außerdem war während des Klinikaufenthalts bei allen Patienten der Hawie-IQ ermittelt worden. Die Differenz IQ—IQ′ diente als globales Maß für einen eventuellen Intelligenzabbau; sie wird im folgenden als IQ-Differenz (DIQ) bezeichnet. Das rechnerische Vorzeichen dieser Differenz wurde beibehalten.

2. Von jedem Patienten lag ferner das Ergebnis des Visual Retention Tests von Benton (1968) vor. Es wurde stets die bekannte Standardform des Tests benutzt (Zeichenform, Instruktion A, Vorlagenserie C). Beim Ergebnis wurde die empfohlene Korrektur für Alter und Intelligenzniveau berücksichtigt, als Maß für die prämorbide Intelligenz diente dabei der oben genannte Schätzwert (IQ′).

Es wurden beide für diesen Test vorgesehene Leistungsmaße berechnet: die Differenz zwischen der normalen und der erreichten Anzahl der insgesamt richtig reproduzierten Vorlagen (BT_r) und die entsprechende Differenz für die Anzahl der Einzelfehler (BT_f).

3. Außerdem wurde der weniger bekannte Visual Organization Test von Hooper (1958) verwendet. Dieser Test wurde speziell zur Unterscheidung zwischen hirnorganischen und funktionellen Leistungsstörungen entwickelt. Der Testautor schreibt ihm eine hohe Empfindlichkeit für Leistungsmängel zu, die auf „(patho-) physiologischen Veränderungen corticaler Strukturen" beruhen.

Der Test besteht aus 30 Vorlagen. Jede zeigt die Abbildung eines bekannten Objekts, die willkürlich in 2—5 Teile zerschnitten ist. Die zu einem Objekt gehörenden Teile sind auf der Vorlage in anscheinend zufälliger Anordnung wiedergegeben. Der Patient soll herausfinden, welches Objekt durch Zusammenfügen der Teile entstehen würde. Der Test wird ohne Zeitbegrenzung gegeben. Als Gesamtergebnis können maximal 30 Punkte erreicht werden. Eine Korrektur nach Alter oder Intelligenzniveau ist nicht vorgesehen.

Versuchspersonen. Der statistischen Analyse liegen die Testbefunde von 192 Patienten unserer Klinik zugrunde, die während eines Zeitraums von etwa 2 Jahren

psychologisch untersucht wurden. Anlaß zur psychologischen Untersuchung war in allen Fällen der klinisch begründete Verdacht auf eine Hirnschädigung[1].

Diese Patienten bilden eine nahezu auslesefreie Stichprobe aus jener Gruppe, die in unserer Klinik mit dieser Fragestellung zur psychologischen Untersuchung kommt. Ausgeschieden wurden nur wenige Patienten, die nicht regelrecht untersucht werden konnten (z.B. wegen Bewußtseinstrübung oder mangelnder Kooperation) oder von denen nicht alle erforderlichen Testbefunde erhoben worden waren (z.B. wegen vorzeitiger Entlassung). Unberücksichtigt blieben ferner Patienten, bei denen im klinischen Entlassungsbericht keine eindeutige Entscheidung getroffen worden war, ob eine Schädigung der Großhirnhemisphären vorlag oder nicht.

Diese Entscheidung mußte sich mindestens auf den neurologischen Befund sowie auf die Ergebnisse einer technischen Zusatzuntersuchung stützen. In den allermeisten Fällen beruhte sie auf neurologischer Untersuchung, EEG und neuroradiologischem Befund. Rein psychopathologische Zeichen (Leistungsversagen, Wesensänderung) sowie die psychologischen Untersuchungsergebnisse blieben dabei außer Betracht.

Tabelle 1. *Ätiologie der Hirnschädigungen (Gruppe HS,* n = *148)*

Atrophisch	51	Traumatisch	19
Vasculär	35	Sonstige	24
Neoplastisch	19		

Tabelle 2. *Vergleich der Gruppen nach Geschlecht, Alter und IQ*

Gruppe	Geschlecht		Alter		Hawie-IQ	
	w	m	M	s	M	s
HS ($n = 148$)	30%	70%	43,23	13,29	89,49	13,20
NH ($n = 44$)	23%	77%	38,89	14,53	91,25	10,76

Auf Grund der Schlußdiagnose im Entlassungsbericht wurden zwei Untergruppen gebildet. Die Gruppe HS umfaßt 148 Patienten, bei denen sich der Verdacht auf eine Hirnschädigung bestätigt hatte. Einen Überblick über die Ätiologie der Hirnschäden gibt Tab.1. Die Gruppe NH umfaßt 44 Patienten, bei denen die Symptome, die den Verdacht auf eine cerebrale Schädigung erregt hatten, auf andere Ursachen zurückgeführt werden konnten (meistens auf neurotische Wesenszüge oder neurologische Erkrankungen, die nicht das Großhirn betrafen).

Hinsichtlich des Geschlechts, des Alters und des Hawie-IQ bestanden zwischen beiden Gruppen keine statistisch signifikanten Unterschiede (Tab.2). Das Vorwiegen männlicher Patienten in beiden Gruppen ergibt sich aus der größeren Aufnahmekapazität der Männerabteilung unserer Klinik. Insgesamt dürften die 192 Patienten eine typische Stichprobe jener Fälle bilden, die in neurologischen Kliniken wegen des Verdachts auf eine Hirnschädigung psychologisch untersucht werden.

1 Mit „Hirnschädigung" ist hier stets eine Schädigung der Großhirnhemisphären gemeint.

Methoden und Ergebnisse

Trennschärfe der Einzelmerkmale. Einen ersten Eindruck von der Trennschärfe der Einzelmerkmale erhält man durch Vergleich der Mittelwerte für die beiden Gruppen HS und NH. Die Differenzen wurden mit dem t-Test zufallskritisch geprüft. Da die Richtung der Abweichungen theoretisch vorhergesagt werden kann, wurde einseitig geprüft. Ausgangsdaten und Ergebnisse gibt Tab. 3 wieder.

Tabelle 3. *Vergleich der Mittelwerte beider Gruppen für die vier Einzelmerkmale*

	Gruppe HS		Gruppe NH		t	p (einseitig)
	M	s	M	s		
DIQ	− 8,58	12,28	− 3,75	9,73	2,38	≤ 0,01
BT_r	− 2,07	2,00	− 1,50	1,97	2,17	≤ 0,05
BT_f	3,83	4,20	2,30	3,80	2,80	≤ 0,01
VOT	19,09	5,00	20,31	4,68	1,43	n.s.

Beim VOT von Hooper ist der Mittelwertsunterschied zwischen den beiden Gruppen statistisch nicht signifikant. Bei den übrigen drei Merkmalen bestehen zwar gesicherte Unterschiede, verglichen mit den Merkmalsstreuungen (s) sind die Mittelwertsdifferenzen aber recht gering. Offenbar trennt keines von diesen Merkmalen sehr scharf zwischen hirngeschädigten und nichthirngeschädigten Patienten.

Beim Benton-Test und beim VOT von Hooper läßt sich das anschaulich demonstrieren. Für beide Tests haben die Autoren kritische Grenzwerte angegeben, nach denen die Testbefunde als auffällig oder als normal klassifiziert werden. Als Hinweis auf eine Hirnschädigung gelten

beim Merkmal BT_r: mindestens 2 richtige Reproduktionen weniger als erwartet,
beim Merkmal BT_f: mindestens 3 Fehler mehr als erwartet,
beim VOT: weniger als 20 Punkte.

Legt man diese Grenzwerte zugrunde, dann verteilen sich in unserer Stichprobe die auffälligen (positiven) und die normalen (negativen) Testbefunde auf die beiden Patientengruppen wie in Tab. 4, Spalten (1) und (2).

In Anlehnung an die Terminologie von Meehl u. Rosen (1955) nennen wir die zutreffenden Testbefunde *„Treffer" (Hits)* und übernehmen für die prozentualen Trefferhäufigkeiten ihr Symbol H (Hit Rate). In Spalte (5) der Tabelle sind aufgeführt: die prozentualen Trefferhäufigkeiten bei den positiven Testbefunden (H_P), bei den negativen Testbefunden (H_N) und bei den Testbefunden insgesamt (H_T = Total Hit Rate).

Tabelle 4. *Trefferhäufigkeiten (H) beim Benton-Test und beim VOT von Hooper*

	Testbefund	Klinische Diagnose HS (1)	NH (2)	Summe der Testbefunde (3)	Davon richtig[a] (4)	Treffer in % (5)
BT_r	Positiv	95	21	116	95	$82 = H_P$
	Negativ	53	23	76	23	$30 = H_N$
	Zusammen	148	44	192	118	$61 = H_T$
BT_f	Positiv	92	18	110	92	$84 = H_P$
	Negativ	56	26	82	26	$32 = H_N$
	Zusammen	148	44	192	118	$61 = H_T$
VOT	Positiv	74	13	87	74	$85 = H_P$
	Negativ	74	31	105	31	$30 = H_N$
	Zusammen	148	44	192	105	$55 = H_T$

[a] Bei positiven Testbefunden der Wert aus Spalte (1), bei negativen der Wert aus Spalte (2).

Beachtet man vorläufig nur die *Gesamttrefferhäufigkeit (H_T)*, so beträgt diese zwischen 55% und 61%. Das ist sicher kein sehr befriedigendes Ergebnis. Immerhin liegt H_T aber über 50%, was den Schluß nahelegt, daß man mit den Tests wenigstens etwas mehr Treffer erzielt als bei bloßem Raten. Meehl u. Rosen haben aber gezeigt, daß eine solche Überlegung nicht ganz realistisch ist. Sie würde nur gelten, wenn wir über das Patientengut, das in unserer Klinik zur psychologischen Untersuchung kommt, buchstäblich nichts wüßten und daher die Existenz wie die Nichtexistenz einer Hirnschädigung bei diesen Patienten für gleich wahrscheinlich halten müßten. Wir hätten uns aber unschwer die Information verschaffen können, daß 77% dieser Patienten tatsächlich hirngeschädigt sind. Dann hätte uns allein die Kenntnis dieser Grundhäufigkeit (Base Rate) der Hirngeschädigten unter unseren Probanden ein weit erfolgreicheres Raten erlaubt.

Wenn aus einer Gesamtgruppe zwei Untergruppen ungleichen Umfangs auszulesen sind, dann erzielt man beim Raten stets mehr als 50% Treffer, wenn man bei jedem Individuum vermutet, es entstamme der größeren Untergruppe. Wollte man die Kenntnis, daß in unserer Stichprobe die Hirngeschädigten überwiegen, sinngemäß nutzen, dann müßte man bei reinen Vermutungsdiagnosen jeden zur psychologischen Untersuchung vorgestellten Patienten als „hirngeschädigt" bezeichnen. Dieses „gezielte Raten" ergäbe Trefferhäufigkeiten wie in Tab. 5.

Man könnte also auf diese Weise eine Gesamttrefferhäufigkeit von 77% erzielen, ohne einen einzigen Patienten psychologisch untersucht

Tabelle 5. *Trefferhäufigkeiten beim gezielten Raten*

Vermutung	Klinische Diagnose HS	NH	Summe	Davon richtig	Treffer in %
HS	148	44	192	148	$77 = H_P$
NH	—	—	—	—	$- = H_N$
Zusammen	148	44	192	148	$77 = H_T$

zu haben. Es dürfte nicht unbillig sein, den diagnostischen Erfolg der Tests an dieser Trefferhäufigkeit zu messen, statt an der von 50%, die nur für völlig blindes Raten gilt. Für alle drei Testmerkmale liegt H_T aber beträchtlich niedriger als 77%.

Anders verhält es sich, wenn man die Trefferhäufigkeiten H_P und H_N in den Tab. 4 und 5 vergleicht. Bei den Testbefunden liegt H_P zwischen 82% und 85%, beim gezielten Raten erhält man nur 77%. H_N liegt bei den Testbefunden zwischen 30% und 32%. Beim Raten kann dieser Wert nicht berechnet werden, weil die Vermutung „nichthirngeschädigt" prinzipiell nicht ausgesprochen wird. Man kann H_N hier aber so behandeln als sei der Wert Null.

Die geringe Erhöhung von H_P bei den Tests wird dadurch erkauft, daß die absolute Zahl der richtig identifizierten Hirngeschädigten sich erheblich vermindert (Spalte 1 in Tab. 4 und 5). Beim gezielten Raten erfaßt man alle 148 Hirngeschädigten, mit den Tests dagegen nur 74 bis 95 — das sind 50—64% aller Hirngeschädigten. Die übrigen haben negative Testbefunde. Bei H_N ist der Zuwachs zahlenmäßig größer. Praktisch bleiben die Werte, die H_N bei den Tests erreicht, jedoch höchst unbefriedigend: Nur knapp ein Drittel der Patienten mit negativen Testbefunden hat tatsächlich keine Hirnschädigung. Negative Testbefunde führen also weitaus häufiger zur falschen als zur richtigen diagnostischen Schlußfolgerung.

Die Effizienz der beiden Tests läßt sich auch dadurch nicht überzeugend verbessern, daß man für die Klassifikation der Befunde als „positiv" oder „negativ" einen anderen kritischen Grenzwert wählt als die Testautoren vorgesehen haben. Wird der Grenzwert systematisch variiert, so erhält man bei beiden Tests dann die höchste Gesamttrefferhäufigkeit, wenn jeder Befund als positiv klassifiziert wird. Das entspricht aber genau dem Vorgehen beim gezielten Raten. Das gleiche Resultat ergab sich auch für die IQ-Differenz.

Trennschärfe einer optimalen Kombination der Merkmale. Es wurde geprüft, ob sich eventuell durch Kombination der Einzelmerkmale ein günstigeres Ergebnis erzielen läßt. Zu diesem Zweck wurde eine Diskriminanzanalyse durchgeführt. Mit diesem statistischen Verfahren läßt sich objektiv ermitteln, welche Gewichte den Einzelmerkmalen zugeteilt werden müssen, um eine optimale Kombination mit der höchstmöglichen

Trennschärfe zu erhalten (vgl. Weber, 1967). Für die Merkmale BT_f, VOT und DIQ ergab sich die lineare Diskriminanzfunktion[2]

$$X = 5{,}94\, x_{BT(f)} + 0{,}94\, x_{VOT} - x_{DIQ}\,.$$

Der Wert X stellt die bestmögliche Kombination der drei Einzelwerte dar und muß in unserer Stichprobe die schärfste Trennung zwischen den beiden Untergruppen erlauben, die mit diesen Tests überhaupt möglich ist.

Tabelle 6. *Häufigkeitsverteilung der Diskriminanzwerte (X) und zugehörige Trefferhäufigkeiten (H)*

X	Gruppe		Treffer in %		
	HS	NH	H_P	H_N	H_T
141—150	3		100	23	24
131—140	—		100	23	24
121—130	2		100	24	26
111—120	3		100	24	27
101—110	5		100	25	30
91—100	4	1	94	25	31
81— 90	10	3	87	25	35
71— 80	10	1	88	26	40
61— 70	18	4	86	27	47
51— 60	21	5	84	29	55
41— 50	16	2	85	33	62
31— 40	13	6	83	34	66
21— 30	20	9	80	36	72
11— 20	11	6	79	37	74
1— 10	10	5	78	50	77
−9 bis 0	2	1	77	100	78
−19 bis −10	—	1	77	—	77

Die Häufigkeitsverteilung der X-Werte für beide Patientengruppen ist in Tab. 6 wiedergegeben. Zusätzlich wurde berechnet, welche Trefferhäufigkeiten H_P, H_N und H_T man jeweils erhalten würde, wenn man irgendeinen der aufgeführten X-Werte als kritische Grenze zwischen positiven und negativen Testbefunden betrachtet. Wie Tab. 6 zeigt, wächst die Gesamttrefferhäufigkeit H_T je weiter man den Grenzwert nach unten verlegt, d.h. je mehr Patienten aus beiden Gruppen man positive Testbefunde zuerkennt. Die Gesamttrefferhäufigkeit von 77%, die man beim gezielten Raten erreicht, wird aber nur in einem Falle ganz geringfügig übertroffen: Wenn man alle Werte von $X = -9$ und größer als positive Befunde und alle von $X = -10$ und kleiner als negative

2 Das Merkmal BT_r konnte bei dieser Analyse nicht berücksichtigt werden, weil es sich als nicht normalverteilt erwies.

Befunde wertet, dann beträgt die Gesamttrefferhäufigkeit 78% statt 77%. Dieser minimale Gewinn steht aber in offenkundigem Mißverhältnis zu dem Mehraufwand, den Routineuntersuchungen mit dem Hawie, dem Benton-Test und dem VOT von Hooper gegenüber gezielten Raten erfordern.

Die Trefferhäufigkeit H_P läßt sich vergrößern, wenn man die kritische Grenze weit nach oben verlegt. Man müßte dann aber in Kauf nehmen, daß nur noch ein kleiner Teil der Hirngeschädigten ausgelesen wird. Beispielsweise wird eine Trefferhäufigkeit H_P von 94% erzielt, wenn man nur die X-Werte von 91 und größer als positive Befunde ansieht. Man würde dann jedoch nur noch 17 von insgesamt 148 Hirngeschädigten erfassen. Die Trefferhäufigkeit H_N dagegen läßt sich auf sinnvolle Weise überhaupt nicht vergrößern. Sie beträgt zwar nominell 50%, wenn man alle X-Werte von 1 und größer als positiven Befund betrachtet — jedoch erhalten dann auch nur noch 2 Patienten aus der Gruppe NH den zutreffenden negativen Befund.

Selbst wenn man die drei Testmerkmale auf die bestmögliche Weise kombiniert, bleibt der diagnostische Erfolg der pyschologischen Verfahren offensichtlich unbefriedigend.

Die Risikominderung bei einer Klassifikation nach Testbefunden gegenüber gezieltem Raten. Rimm (1963) hat den Überlegungen von Meehl u. Rosen einen weiteren wichtigen Aspekt hinzugefügt: Die realen Nachteile, die aus falschen positiven und die aus falschen negativen Befunden erwachsen, haben häufig ein sehr unterschiedliches Gewicht. Will man den praktischen Nutzen irgendeines Ausleseverfahrens realistisch beurteilen, muß man auch dies berücksichtigen.

Psychologische Verfahren zur Diagnose von Hirnschäden werden im allgemeinen dann angewandt, wenn die medizinischen Routineuntersuchungen keinen klaren Befund ergaben und man vor der Entscheidung steht, ob zusätzlich schwierigere, kostspielige oder für den Patienten nicht ganz risikofreie Untersuchungen durchgeführt werden müssen. Ist der psychologische Befund negativ, wird man auf solche Untersuchungen eher verzichten; ist er positiv, dann wird das meistens den Ausschlag geben, die Untersuchungen doch vorzunehmen, um den Fall diagnostisch abzuklären. Demnach wird ein falsch positiver Testbefund zwar unnötige Untersuchungen veranlassen, die zeitraubend, teuer und u. U. nicht ganz ohne Risiko sind. Ein falsch negativer Testbefund hätte jedoch zur Folge, daß eine tatsächlich bestehende Hirnschädigung nicht erkannt wird und eine vielleicht mögliche Behandlung unterbleibt. Wägt man alle Nachteile gegeneinander ab, so sind sie bei einem falsch negativen Testbefund sicher größer als bei einem falsch positiven.

Bei gezieltem Raten würde man das Risiko, das mit falsch negativen Befunden verbunden ist, völlig ausschließen; man müßte aber bei allen

nichthirngeschädigten Patienten die Nachteile falsch positiver Diagnosen in Kauf nehmen. Bei Anwendung der Tests werden dagegen die Risiken falscher positiver Entscheidungen etwas reduziert, dafür wachsen die Risiken, die aus falschen negativen Entscheidungen entstehen. Es ist die Frage, ob das Raten oder ob Entscheidungen auf Grund der Testbefunde insgesamt das größere Risiko mit sich bringen. Rimm (1963) hat eine Rechenformel angegeben, mit der quantitativ abgeschätzt werden kann, in welchem Maße das Gesamtrisiko durch Anwendung eines Auslesetests gegenüber gezieltem Raten verringert wird.

In sehr vielen Fällen, in denen Auslesetests verwendet werden (z.B. bei der Eignungsauslese), besteht der Nachteil, der mit unzutreffenden Testbefunden verbunden ist, in unnötigen Geldausgaben (z.B. den Ausbildungskosten). Rimm definiert deshalb die Risiken als Kosten und die Risikominderung als Kostenersparnis (cost-efficiency). Rimm weist aber selbst darauf hin, daß seine Überlegungen keineswegs nur auf finanzielle Nachteile anwendbar sind, sondern auf Risiken jeder Art. Voraussetzung ist nur, daß sich die mit falschen positiven und mit falschen negativen Testbefunden verbundenen Nachteile in ein quantitatives Verhältnis setzen lassen. Wir erläutern deshalb hier seine Formel in einer adäquaten allgemeinen Terminologie.

In unserem Fall, in dem die Probanden überwiegen, die einen positiven Testbefund haben müßten, lautet die Rechenformel[3] in ihrer übersichtlichsten Form:

$$C = q_2 - q_1 \cdot \frac{P}{Q} \cdot \frac{1}{R}.$$

Darin bedeutet

- C die relative Verminderung des Gesamtrisikos bei Anwendung des Tests gegenüber dem Gesamtrisiko bei gezieltem Raten
- q_2 den proportionalen Anteil der richtigen (negativen) Testbefunde bei den nichthirngeschädigten Patienten
- q_1 den proportionalen Anteil der falschen (negativen) Testbefunde bei den Hirngeschädigten
- P den proportionalen Anteil der Hirngeschädigten in der untersuchten Gesamtgruppe
- Q den entsprechenden Anteil der nichthirngeschädigten Patienten
- R das Verhältnis der mit falsch positiven Befunden verbundenen Risiken zu den mit falsch negativen Befunden verbundenen.

In unserem Fall muß man wohl annehmen, daß ein falsch negativer Befund mindestens doppelt so schwerwiegende Folgen hat wie ein falsch positiver Befund; demnach ist $R = \frac{1}{2}$ und $\frac{1}{R} = 2$. Das Verhältnis $\frac{P}{Q}$ ist in unserer Stichprobe $\frac{0{,}77}{0{,}33} = 3{,}35$. Berechnet man nun aus Tab. 4, Spalten (1) und (2), die Werte q_1 und q_2 für das Merkmal BT_r und setzt sie in die Formel ein, so erhält man für dieses Merkmal $C = -1{,}88$.

Das negative Vorzeichen dieses Wertes zeigt, daß bei Entscheidungen, die sich auf das Testmerkmal BT_r stützen, das Gesamtrisiko noch größer

3 Wenn die Probanden überwiegen, die negative Testbefunde haben müßten, gilt dagegen die Formel von Rimm (1963, S. 90).

ist als beim gezielten Raten. Wir erhielten mit den Daten aus Tab. 4 auch bei den übrigen Testmerkmalen stets negative Werte von C. Nicht viel anders ist es bei der Kombination aus den Einzelmerkmalen: Wenn man in Tab. 6 die Grenze zwischen positiven und negativen Testbefunden von den höchsten X-Werten ausgehend schrittweise nach unten verlegt, so bleibt C immer negativ, bis die Grenze zwischen $X = -10$ und $X = -9$ liegt. In diesem einen Falle beträgt $C = +0{,}02$ und zeigt eine minimale Verminderung des Gesamtrisikos an. Der Unterschied zum gezielten Raten ist aber praktisch völlig bedeutungslos. Er besteht lediglich darin, daß auch ein nichthirngeschädigter Patient richtig klassifiziert wird. Wenn die Voraussetzung richtig ist, daß falsch negative Befunde (mindestens) doppelt so schwerwiegende Folgen haben wie falsch positive, und wenn das Verhältnis $\frac{P}{Q}$ in unserer Stichprobe annähernd repräsentativ ist, dann können diagnostische Entscheidungen auf Grund der Tests das Gesamtrisiko gegenüber gezieltem Raten bei einem solchen Patientengut nicht verringern, sondern praktisch nur vergrößern.

Aus der Berechnungsformel für C ist leicht zu erkennen, woran das liegt: q_1, der Anteil falscher Befunde bei den Hirngeschädigten, muß in unserem Falle mit einem sehr großen Faktor (nämlich $2 \cdot 3{,}35 = 6{,}7$) multipliziert werden. Selbst wenn q_2, der Anteil richtiger Befunde bei den nichthirngeschädigten Patienten sehr hoch ist, muß q_1 sehr klein sein, damit C ein positives Vorzeichen erhält und eine Risikominderung ausdrückt. Es läßt sich leicht errechnen, daß unter diesen Umständen selbst ein Test, der 100% der nichthirngeschädigten Patienten richtig klassifiziert, schon bei nur 15% falscher Befunde in der Gruppe der Hirngeschädigten einen negativen Wert von C ergibt. Man muß daraus folgern, daß es keinen psychologischen Test gibt, mit dem man unter diesen Umständen den Erfolg beim gezielten Raten übertreffen könnte.

Das wäre nur möglich, wenn der Ausdruck $\frac{P}{Q} \cdot \frac{1}{R}$ einen weniger hohen Wert hätte als nach den bisherigen Voraussetzungen. Tatsächlich könnten unsere Annahmen in gewissen Grenzen falsch sein. Der Wert P und der komplementäre Wert $Q = P - 1$ wurden aus unserer Stichprobe berechnet, sie sind deshalb mit einem Zufallsfehler behaftet. Die wahren Werte für das entsprechende Patientengut könnten weniger extrem sein. Die untere Grenze des 99%-Vertrauensbereichs für unseren Wert $P = 77\%$ liegt etwa bei 68,5%, der Komplementärwert Q ist dann 31,5%. Diese günstigeren Schätzungen ergeben für $\frac{P}{Q}$ einen Wert von rund 2. Probeweise könnte man ferner unterstellen, daß die mit falsch negativen Testbefunden verbundenen Risiken zu hoch eingeschätzt wurden und sich zu den Risiken bei falsch positiven Befunden nur verhalten wie 3:2. Unter diesen Voraussetzungen, die sicher die günstigsten sind, die man in unserem Fall noch akzeptieren könnte, wird $\frac{1}{R} = 1{,}5$ und $\frac{P}{Q} \cdot \frac{1}{R} = 3$.

Selbst wenn man mit diesem wesentlich kleineren Faktor rechnet, muß man jedoch noch hohe Anforderungen an einen Auslesetest stellen. Allein um das Gesamtrisiko gegenüber gezieltem Raten nicht zu erhöhen

und wenigstens einen Wert von $C = 0$ zu erreichen, würde man einen Test brauchen, der beispielsweise eine der folgenden Bedingungen erfüllt:

Wenn richtige Befunde bei HS:	dann richtige Befunde bei NH:
90 %	30 %
85 %	45 %
80 %	60 %
75 %	75 %
70 %	90 %
66,7 %	100 %

Keines unserer Einzelmerkmale genügt diesen Bedingungen auch nur annähernd. Selbst für die bestmögliche Kombination der Merkmale findet man in Tab. 6 keinen kritischen Grenzwert, bei dem sie erfüllt werden — außer dem einen zwischen $X = -10$ und $X = -9$, dessen praktische Bedeutungslosigkeit bereits demonstriert wurde.

Diskussion

Es bleibt zu prüfen, ob die ungünstigen Ergebnisse unserer Untersuchung im Widerspruch zu den Resultaten anderer Untersucher stehen. Leider gibt es nicht viele Bewährungsstudien, in denen die für einen Vergleich nötigen Daten publiziert wurden. Meistens beschränken sich die Angaben auf Mittelwerte und Standardabweichungen.

Für das Merkmal BT_r wird in drei Untersuchungen der Anteil richtiger Befunde bei den verglichenen Gruppen mitgeteilt:

Benton (1968) verglich 100 Hirngeschädigte mit 100 nichthirngeschädigten Patienten. Die Kontrollgruppe bestand hauptsächlich aus Patienten mit Bandscheibenschäden oder Psychoneurosen, bei denen eine Hirnschädigung von vornherein höchst unwahrscheinlich war. Brilliant u. Gynther (1963) untersuchten die Trennschärfe dieses Merkmals an einer unausgelesenen Stichprobe von 110 neuropsychiatrischen Patienten, unter denen sich 34 Hirngeschädigte befanden. Bei den übrigen Patienten mit psychiatrischen Erkrankungen bestand offenbar niemals ein begründeter Verdacht auf eine Hirnschädigung. Cronholm u. Schalling (1963) verglichen 29 hirngeschädigte und 56 nichthirngeschädigte Patienten einer Rehabilitationsklinik. Auch hier war die Zugehörigkeit eines jeden Patienten zu der einen oder anderen Gruppe von vornherein sicher.

Die Ergebnisse dieser Untersuchungen werden in Tab. 7 unseren Resultaten gegenübergestellt. Die Klassifikation der Testbefunde als positiv oder negativ beruht bei allen drei Untersuchungen auf demselben kritischen Grenzwert, den wir verwendeten; dasselbe gilt auch für die übrigen Testmerkmale, die in Tab. 7 aufgeführt sind.

Tabelle 7. *Häufigkeit richtiger Testbefunde bei den verschiedenen Untersuchungen*

Merkmal	Untersucher	Richtige Testbefunde (%)			Korrigiert für $P = 77\%$
		HS (1)	NH (2)	Insges. (3)	(4)
BT_r	Benton	69	84	76	72
	Brilliant-Gynther	62	85	76	67
	Cronholm-Schalling	45	72	62	51
	Hartje-Orgass	64	52	61	61
BT_f	Brilliant-Gynther	47	95	66	58
	Hartje-Orgass	62	59	61	61
VOT	Hooper (1952)	79	100	95	84
	Hooper (1958)	64	81	78	68
	Hartje-Orgass	50	70	55	55

Mit dem Merkmal BT_r erzielten wir bei den Hirngeschädigten 64% richtige Befunde. Verglichen mit den Werten der anderen Autoren ist dieses Ergebnis recht günstig (Spalte 1). Demgegenüber erhielten wir nur 52% richtige Befunde bei den nichthirngeschädigten Patienten, während die anderen Untersucher Werte zwischen 72% und 84% mitteilen (Spalte 2). Der Unterschied dürfte sich dadurch erklären, daß bei den anderen Bewährungsstudien die Kontrollgruppe zu einem großen Teil aus Patienten bestand, bei denen von vornherein kein begründeter Verdacht auf eine Hirnschädigung vorlag.

Der Prozentsatz insgesamt richtiger Befunde (Spalte 3) bei den verschiedenen Untersuchungen läßt sich nicht unmittelbar vergleichen, weil der Anteil der Hirngeschädigten sehr verschieden war. Man kann jedoch errechnen, wie hoch der Prozentsatz der insgesamt richtigen Testbefunde jeweils gewesen wäre, wenn die Grundhäufigkeit der Hirngeschädigten (P) stets 77% betragen hätte wie bei uns (Spalte 4). In diesem Fall liegt unser Wert von 61% etwa in der Mitte zwischen dem höchsten und dem niedrigsten Wert.

Für das Merkmal BT_f teilen nur Brilliant u. Gynther vergleichbare Daten mit. Ihre Ergebnisse sind insgesamt etwas ungünstiger als unsere Resultate. Nur für die richtigen Befunde bei den nichthirngeschädigten Patienten (Spalte 2) liegt ihr Prozentsatz wesentlich höher — höchstwahrscheinlich aus den bereits erwähnten Gründen.

Für den VOT macht nur der Testautor selbst vergleichbare Angaben. Hooper (1952) verglich 70 Hirngeschädigte mit 140 psychiatrischen Patienten sowie 70 Patienten eines Allgemeinkrankenhauses. Die 210

Kontrollpatienten waren so ausgewählt, daß eine Hirnschädigung von vornherein mit hoher Wahrscheinlichkeit ausgeschlossen werden konnte, während sie bei der Vergleichsgruppe unbezweifelbar sicher war. Diese Patienten hatten entweder ausgedehnte diffuse Hirnveränderungen oder umschriebene traumatische Läsionen; es dürften also überwiegend gravierende Schädigungen gewesen sein. Die auffallend günstigen Ergebnisse dieser Kontrollstudie sind daher wohl kaum repräsentativ für den VOT. Ganz sicher gilt dies für die 100% richtigen Befunde bei den nichthirngeschädigten Patienten: Hooper (1958) gibt selbst an, daß er sogar bei normalen Collegestudenten 6% falsche Befunde erhielt. Bei einer weiteren Kontrolluntersuchung an einer unausgelesenen Stichprobe von 240 Patienten aus einer Heil- und Pflegeanstalt, darunter 39 mit chronischer Hirnschädigung, erzielte Hooper (1958) weniger ungewöhnliche Resultate. Diese Studie fällt allerdings insofern aus dem Rahmen, als die Hirngeschädigten ausnahmslos eine pyschotische Begleitstörung (psychotic reaction) hatten. Insgesamt sehen wir keinen Widerspruch zwischen Hoopers Ergebnissen und unseren weniger günstigen Befunden.

Zusammenfassend darf man folgern, daß unsere Untersuchung kein ungerechtfertigt negatives Bild von der diagnostischen Bewährung der überprüften Testmerkmale vermittelt. Die Ergebnisse der anderen Untersucher führen zu denselben Schlüssen. Selbst wenn man bei jedem Testmerkmal nur die günstigsten Resultate herausgreift, wird nur in einem Falle ein etwas höherer Anteil an insgesamt richtigen Befunden erreicht als beim gezielten Raten: Bei der Untersuchung von Hooper (1952) beträgt dieser Anteil, bezogen auf eine klinische Population mit 77% Hirngeschädigten, 84% gegenüber 77% beim Raten. Unter den Voraussetzungen, die wir für wahrscheinlich halten (S. 182), tritt dadurch noch keinerlei Verminderung des Gesamtrisikos ein. Sogar wenn man von günstigeren Voraussetzungen (S. 183) ausgeht, bleiben die Ergebnisse dieser einen Untersuchung die einzigen, die den Erfolg beim Raten übertreffen.

Unsere Ergebnisse wie die der anderen Untersucher lassen bei realistischer Beurteilung nur den Schluß zu, daß diese psychodiagnostischen Verfahren bei einem Patientengut, dessen Zusammensetzung auch nur annähernd unserer Stichprobe entspricht, ohne den geringsten praktischen Nutzen sind. Ihre Anwendung in einer neurologischen Klinik ist sogar nachteilig, falls die Testbefunde die Entscheidung beeinflussen, ob bei einem Patienten weitere klinisch-diagnostische Maßnahmen nötig sind, um den Verdacht auf eine Hirnschädigung abzuklären. Das liegt vor allem an der Zusammensetzung einer solchen Population und an dem hohen Risiko, daß mit falschen Befunden bei Hirngeschädigten verbunden ist. Man kann nicht ausschließen, daß dieselben Tests, auf andere Populationen angewandt, diagnostisch nützlich sein könnten.

Literatur

Bender, L.: A visual motor Gestalt test and its clinical use. Res. Monogr. Amer. Orthopsychiat. Ass., No. 3 (1938).

Benton, A. L.: Der Benton-Test, 3. Aufl. Bern-Stuttgart: Huber 1968.

Brilliant, P. J., Gynther, M. D.: Relationships between performance on three tests for organicity and selected patient variables. J. cons. Psychol. **27**, 474—479 (1963).

Cronholm, B., Schalling, D.: Intellectual deterioration after focal brain injury. Arch. Surg. **86**, 670—687 (1963).

Hooper, H. E.: Use of the Hooper Visual Organization Test in the differentiation of organic brain pathology from normal, psychoneurotic, and schizophrenic reactions. Amer. Psychol. **7**, 350 (1952).

Hooper, H. E.: The Hooper Visual Organization Test. Beverly Hills, Calif.: Western Psychological Services 1958.

Meehl, P. E., Rosen, A.: Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychol. Bull. **52**, 194—216 (1955).

Rimm, D.: Cost efficiency and test prediction. J. cons. Psychol. **27**, 89—91 (1963).

Weber, E.: Grundriß der biologischen Statistik, 6. Aufl. Stuttgart: Fischer 1967.

Wechsler, D.: Die Messung der Intelligenz Erwachsener, 3. Aufl. Bern-Stuttgart: Huber 1964.

Dipl.-Psychol. Dr. phil. W. Hartje
Dipl.-Psychol. B. Orgass
Abt. Neurologie der Medizinischen
Fakultät an der Rhein.-Westf.
Technischen Hochschule Aachen
D-5100 Aachen, Goethestraße 27/29
Deutschland