HATHAWAY, S. R., & MONACHESI, E. D., The personalities of predelinquent boys, *Journal of criminal Law and Criminology*, 1957, 48, 149–163.
HATHAWAY, S. R., & MONACHESI, E. D., *Adolescent personality and behavior*, Minneapolis, University of Minnesota Press, 1963.
QUAY, H. G. (Ed.), *Juvenile delinquency*, Princeton, Van Nostrand, 1965.
SCHMIDT, L. R., STEIGERWALD, F., & SCHNEIDER, H., Diskriminanzanalytische Untersuchungen mit dem MMPI-Saarbrücken zur Drogenabhängigkeit und Delinquenz, in: L. ECKENSBERGER (Ed.), *Bericht über den 28. Kongreß der Deutschen Gesellschaft für Psychologie Saarbrücken 1972*, Göttingen, Hogrefe, 1974 (im Druck).
SCHMITT, G., & STEIGERWALD, F., Zur Diagnostik von jugendlichen Straftätern mit Hilfe einer neuen MMPI-Skala (Dz), in: G. NASS (Ed.), *Kriminalität – vorbeugen und behandeln*, Köln, Heymanns, 1971, p. 115–137.
STEIGERWALD, F., & SCHMIDT, L. R., Vergleich der Persönlichkeitsstrukturen von Drogenabhängigen und Delinquenten im MMPI-Saarbrücken, 1974 (in Vorbereitung).
TIATOR, G., Die Klassifizierung delinquenter Mädchen mit dem MMPI-Saarbrücken, Diplom-Arbeit am Psychologischen Institut der Universität des Saarlandes, Saarbrücken, 1969.

(Eingegangen am 21. September 1973)

Anschrift der Verfasser: F. Steigerwald, Prof. Dr. L. Schmidt, Universitäts-Nervenklinik, Abt. für Med. Psychologie, 665 Homburg/Saar.

H. Lukesch

Testkriterien des Depressionsinventars von A. T. Beck

(Aus dem Fachbereich Erziehungswissenschaft der Universität Konstanz)

1. Einleitung

1.1. Vorbemerkung

Im Rahmen der klinischen Forschung und Diagnostik werden in zunehmendem Maße neben „objektiven Tests" (z. B. physiologische Messungen) auch „subjektive Tests zur Erfassung der Persönlichkeit" (*Mittenecker*, 1964) verwendet. Unter dieser Bezeichnung faßt man üblicherweise alle Fragebogenverfahren zusammen. Dabei macht es prinzipiell keinen Unterschied aus, ob die so erhaltenen Informationen („questionnaire-data" nach *R. B. Cattell*, 1966, p. 571) aus Selbstauskünften oder aus Fremdbeobachtungen stammen[1]), in beiden Fällen sind sie anfällig für vielerlei Verfälschungstendenzen (z. B. „social desirability", *Edwards*, 1953; „bewußte Verfälschungen", *Hoeth* et al., 1967; „Antworttendenzen", *Carl*, 1968; *Cronbach*, 1942) und daher auch die Bezeichnung

[1]) Die hier getroffene Einteilung wie auch die weitere Anwendung der Begriffe Validität, Reliabilität oder Objektivität deckt sich dabei nicht mit der von *Pöldinger, Blaser & Gehring* (1969) vorgeschlagenen, sondern sie ist an der Standardnomenklatur über Gütekriterien psychologischer Tests orientiert (*Lienert*, 1967; *Michel*, 1964).

„subjektive Verfahren". Bei den sog. „objektiven Verfahren" (eigentliche „testdata" nach *R. B. Cattell*, 1966, p. 575) treten diese Probleme nicht oder zumindest nicht in derselben Weise auf. Um welche Art von Verfahren es sich aber auch immer handeln mag, die Bezeichnung „Test" verdienen sie erst dann, wenn der erfolgreiche Nachweis über das Vorhandensein notwendiger Gütekriterien erfolgt ist (*APA*, 1954; *Lienert*, 1967, p. 12 f.; *Selg & Bauer*, 1971, p. 65 f.).

1.2. Problemstellung

Ein Anstoß zur Konstruktion von Verfahren zur Erfassung der Depressivität lag in der oft gefundenen Nichtübereinstimmung der Urteile von Psychiatern (*Ash*, 1949; *Beck*, 1961; *Beck & Ward*, 1961; *Fogel* et al., 1966; *Ward* et al., 1962), obwohl auch bisweilen gegenteilige Befunde vorliegen (*Schmidt & Fonda*, 1956). So meint z. B. *Beck* (1961, p. 168) über seine Untersuchung: „Although there was a high percentage of exact or near agreement on the clinical ratings of the depth of depression, that method of assessment could not be fully relied on to provide a stable criterion of depression because it was highly subject to inconsistencies on the part of clinicians. Furthermore, there was no assurance that other clinicians assessing depression would use the same indices." Aufgrund dieser Tatsachen wurden eine stattliche Anzahl standardisierter Selbst- und Fremdbeurteilungsskalen zur Erfassung der Depression entworfen (*Lukesch*, 1974). Eines der beliebtesten Verfahren davon stellt das Depressionsinventar von *Beck* et al. (1961) dar. Es wurde nicht nur in verschiedenen anglo-amerikanischen Untersuchungen verwendet (*Beck* et al., 1962; *Metcalfe & Goldman*, 1965; *Schwab* et al., 1967 a, b; *Weckowitz* et al., 1967), sondern liegt auch in einer erweiterten französischen Form vor (*Delay* et al., 1963; *Pichot*, 1964, 1969 a, b; *Pichot & Lemperiere*, 1964; *Pichot* et al., 1966), wurde ins Tschechische übersetzt (*Vinar*, 1966; *Vinar & Grof*, 1969) und schließlich auch in einer deutschen Version vorgelegt (*Blaser* et al., 1968 a, b; *Pöldinger* et al., 1969 a, b).

Der Fragebogen besteht aus Fragen zu 21 verschiedenen Bereichen (vgl. Tab. 1). Die genaue Itemformulierung und Testdurchführung sind bei *Blaser* et al. (1968) beschrieben und sollen hier nicht wiederholt werden. Es soll hingegen versucht werden, die Gütekriterien dieses Verfahrens zum einen an Hand

Tabelle 1: *Symptome, auf welche sich die Fragen des DI beziehen*

A Traurigkeit	L Soziale Isolierung
B Pessimismus	M Entschlußunfähigkeit
C Versagen	N Körperbild
D Unzufriedenheit	O Arbeitsunfähigkeit
E Schuldgefühle	P Schlafstörungen
F Strafbedürfnis	Q Ermüdbarkeit
G Selbsthaß	R Appetitverlust
H Selbstanklagen	S Gewichtsverlust
I Selbstmordimpulse	T Hypochondrie
J Weinen	U Libidoverlust
K Reizbarkeit	

der in der Literatur berichteten Ergebnisse darzustellen und zum anderen aufgrund eigener Untersuchungen. Letztere Ergebnisse beruhen auf drei verschiedenen Stichproben:

(1) Gruppe 1 waren depressive Patienten, die bei ihrer Aufnahme in die Klinik getestet wurden. Suizidversuche und Aufnahmen zum Wochenende, die vor einer Testung hätten behandelt werden müssen, wurden von der Untersuchung ausgeschlossen (N = 30; ♀ = 19; ♂ = 11; Alter: 24—53).

(2) Gruppe 2 waren nicht-depressive Probanden. Diese Stichprobe war so gebildet worden, daß sie zur ersten möglichst parallelisiert war (N = 30; ♀ = 18; ♂ = 12; Alter: 23—55)[2]).

(3) Gruppe 3 war eine „anfallende" Stichprobe von Psychologiestudenten, die zu Beginn eines psychologischen Praktikums getestet wurden (N = 42; Alter: 20—36).

Die Gütekriterien des Verfahrens (Objektivität, Itemcharakteristiken, Reliabilität, Validität, Normen) sollen im einzelnen kritisch dargestellt werden, um seine Brauchbarkeit abschätzen zu können.

2. *Gütekriterien des DI*

2.1. Objektivität

Sieht man einmal von dem „subjektiven" Charakter von Fragebögen überhaupt und den damit zusammenhängenden Verfälschungsmöglichkeiten ab, so kann man unter Objektivität „den Grad, in dem die Ergebnisse eines Testes unabhängig vom Untersucher sind" (*Lienert*, 1967, p. 13) verstehen. Es werden üblicherweise drei Aspekte dieser so definierten Objektivität unterschieden, nämlich die Durchführungs-, Auswertungs- und Interpretationsobjektivität.

2.1.1. Durchführungsobjektivität. Speziell ist damit gemeint, daß die Testergebnisse vom Verhalten des Untersuchers nicht beeinflußt werden sollen. Dies versucht man durch standardisierte Instruktionen zu erreichen. Bei dem DI steht man aber vor der schwierigen Aufgabe, Depressive zu einer Reaktion zu veranlassen, die der eigenen Einschätzung ihres Erlebens entsprechen soll. Daß hierbei die verschiedensten Schwierigkeiten entstehen können (z. B. Verständnisschwierigkeiten, mangelnde Kooperationsbereitschaft), ist eine belegbare und erlebbare Tatsache (*Blaser*, 1968 b, p. 302). Da für die verschiedenen hier auftretenden Möglichkeiten keine gezielten Instruktionen vorliegen, bleibt dieser Aspekt der Objektivität ein nicht immer erreichbares Desiderat.

2.1.2. Auswertungsobjektivität. Da es hier nur um die richtige Protokollierung und Addition von Einzelwerten geht, ist dieser Punkt – wie bei den anderen Fragebogenverfahren auch – kein Problem.

2.1.3. Interpretationsobjektivität. Damit ist gemeint, daß von demselben Ausgangsmaterial (auch von verschiedenen Untersuchern) dieselben diagnostischen Schlüsse gezogen werden. Inwieweit dies bei diesem Verfahren zutrifft, kann erst später im Zusammenhang mit Validitätsfragen besprochen werden. Es erscheint aber einleuchtend, daß dieser Test allein keine Entscheidungsgrundlage z. B. für die Zuweisung zu einer speziellen Therapie darstellt (*Blaser* et al., 1968 a, p. 33; 1968 b, p. 314). Dies vor allem auch, da das oberflächliche klinische Erscheinungsbild von neurotisch Depressiven bisweilen auf eine tiefere Depression schließen läßt als das von psychotischen Depressiven. Die Information,

[2]) Diese Daten wurden von Frl. Reiter im Rahmen ihrer Dissertation erhoben (*Reiter*, 1972).

die man aus diesem Test erhält, ist also nicht ausreichend für eine eindeutige Interpretation (etwa in der Art, daß Patienten ab einem bestimmten Teilungspunkt mit einer gewissen Fehlerwahrscheinlichkeit einer bestimmten nosologischen Kategorie zuzuordnen wären und einer bestimmten Behandlung bedürften).

2.2. Itemcharakteristiken

Im Zuge der klassischen Testkonstruktion (*Gulliksen*, 1950) wurden, um ungeeignete Items auszuschließen, der Schwierigkeitsindex (Mittelwert jedes Testitems) und die Trennschärfe (Korrelation des Testitems mit dem Gesamttestwert) bestimmt. Auf die verschiedenen Selektionstechniken soll nicht detailliert eingegangen werden (*Lienert*, 1967, p. 137 f.).

2.2.1. Schwierigkeitsindex. Bei der Konstruktion eines Tests werden Items mit einem mittleren Schwierigkeitsindex bevorzugt beibehalten: Items, denen nämlich jeder zustimmt oder die jeder ablehnt, können nicht mehr zwischen den Pbn differenzieren (in Ausnahmefällen behält man extreme Items bei; etwa leichte Aufgaben bei Intelligenztests, um die Pbn nicht abzuschrecken, oder sehr schwere, um auch in oberen Bereichen differenzieren zu können).

Da Übersichten über die Itemmittelwerte bei diesem Verfahren in der Literatur nicht berichtet werden, seien die eigenen Ergebnisse wiedergegeben (vgl. Tab. 2, möglich sind Mittelwerte [$\bar{X}$] zwischen 0 und 3).

Tabelle 2: *Itemmittelwerte und -streuungen bei den verschiedenen Stichproben*[a]

Item	Gruppen 1, 2, 3 N = 102		Gruppe 1 N = 30		Gruppe 2 N = 30		Gruppe 3 N = 42	
	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s
A	0,47	0,84	1,47	0,92	0,00	0,00	0,10	0,29
B	0,40	0,80	1,27	1,00	0,07	0,25	0,02	0,15
C	0,50	0,83	1,20	0,98	0,13	0,43	0,26	0,58
D	0,45	0,67	1,17	0,69	0,17	0,37	0,14	0,35
E	0,35	0,76	0,97	0,98	0,00	0,00	0,17	0,57
F	0,25	0,43	0,57	0,50	0,10	0,30	0,12	0,32
G	0,30	0,64	0,83	0,90	0,03	0,18	0,12	0,32
H	0,81	0,75	1,47	0,72	0,53	0,50	0,55	0,62
I	0,46	0,75	1,00	0,82	0,13	0,34	0,31	0,71
J	0,59	0,95	1,33	0,91	0,17	0,37	0,36	0,97
K	0,68	0,90	1,33	0,91	0,27	0,51	0,50	0,85
L	0,40	0,70	0,90	0,94	0,13	0,34	0,24	0,48
M	0,49	0,75	1,27	0,85	0,17	0,37	0,17	0,37
N	0,40	0,77	1,10	0,98	0,10	0,30	0,12	0,45
O	0,75	0,81	1,37	0,57	0,27	0,44	0,65	0,78
P	0,89	1,04	1,73	0,96	0,67	0,91	0,45	0,79
Q	0,75	0,88	1,50	0,81	0,33	0,47	0,50	0,82
R	0,45	0,72	0,83	0,78	0,27	0,51	0,31	0,71
S	0,27	0,70	0,63	1,02	0,03	0,18	0,19	0,55
T	0,52	0,84	1,43	0,88	0,30	0,59	0,02	0,15
U	0,44	0,85	1,33	1,04	0,17	0,45	0,00	0,00

[a] Die Berechnungen wurden mit dem Programm LAS-BC von Herrn *B. Cloetta* und dem Programm FCTOAN des Rechenzentrums der Universität Konstanz durchgeführt.

Wie aus der Tabelle 2 hervorgeht, handelt es sich bei den Items um „schwere" Fragen, d. h. solchen, denen nicht leicht zugestimmt wird. Während bei den depressiven Pbn (Gruppe 1) die Itemmittelwerte um den Wert von 1 variieren, zeigen die nicht-depressiven Pbn (Gruppen 2 und 3) Itemmittelwerte, die nur wenig größer als Null sind. Dies bedeutet u. a., daß der Test allein aufgrund dieser Werte nicht geeignet ist, den „Depressionsgrad" bei Normalen differenziert festzustellen, sondern eher nur bei Depressiven. Die Itemantworten sind nämlich so konstruiert, daß auch besonders schwere Fälle des subjektiven Erlebens der Depressivität noch erfaßt werden können. Dies würde aber der Intention der Testautoren entsprechen, denn das Verfahren soll ja bei klinischen Fällen und nicht bei normalen Pbn differenzieren.

2.2.2. Trennschärfeindex. Ein positiver Trennschärfeindex bedeutet bekanntlich, daß Pbn mit einem hohen Wert bei diesem Item auch ein hohes Gesamtresultat bekommen bzw. solche mit einem niedrigen Wert bei diesem Item ein niedriges. Testitems mit negativen Trennschärfekoeffizienten schließt man üblicherweise bei der Testkonstruktion aus, denn wollte man jene beibehalten, so würde man in Kauf nehmen, daß Pbn, die eine stark für Depression sprechende Itemantwort geben, ein niedriges Gesamttestresultat erhalten. In den Tabellen 3 und 4 sind die Ergebnisse aus der Literatur und aus der eigenen Untersuchung wiedergegeben. Ein Vergleich der beiden Tabellen zeigt, daß die von den anderen Autoren berichteten Trennschärfekoeffizienten durchwegs po-

Tabelle 3: *In der Literatur berichtete Trennschärfekoeffizienten des DI*

Item	*Delay* et al. (1963) N = 79	*Beck* (1967) N = 606	*Vinar & Grof* (1969) N = 90
A	.42	.68	.73
B	.57	.68	.68
C	.38	.62	.48
D	.51	.68	.51
E	.34	.61	.67
F	.39	.50	.54
G	.51	.57	.68
H	.42	.51	.46
I	.46	.60	.61
J	.40	.51	.54
K	.19*	.31	.49
L	.35	.60	.61
M	.57	.63	.63
N	.32	.51	.56
O	.40	.54	.58
P	.32	.50	.48
Q	.38	.54	.57
R	.36	.54	.48
S	.16*	.32	.29
T	.33	.38	.43
U	(.?)	.51	.50

*) Korrelationskoeffizienten, die nicht mindestens bei p = .05 signifikant sind.

Tabelle 4: *Trennschärfekoeffizienten des DI bei den verschiedenen Stichproben*

Item	Gruppen 1, 2, 3 N = 102	Gruppe 1 N = 30	Gruppe 2 N = 30	Gruppe 3 N = 42
A	.76	.34*	.00*	.17*
B	.74	.52	.15*	—.01*
C	.61	.48	.19*	.04*
D	.72	.45	—.11*	.24*
E	.54	.21*	.00*	.18*
F	.44	.07*	—.11*	.04*
G	.62	.46	.28*	.24*
H	.55	.18*	.10*	.13*
I	.55	.37	.28*	.30*
J	.50	.06*	.30*	.20*
K	.48	.09*	.07*	.27*
L	.54	.46	.12*	.12*
M	.73	.54	.55	.02*
N	.59	.27*	.05*	.02*
O	.58	.32*	.32*	.42*
P	.57	.36	.13*	.30*
Q	.62	.29*	.03*	.49
R	.32	—.33*	.39	.48
S	.26	—.21*	.21*	.15*
T	.65	.08*	.59	—.09*
U	.63	.14*	.07*	.00*

*) Korrelationskoeffizienten, die nicht mindestens bei p = .05 signifikant sind.

sitiv und bis auf zwei Ausnahmen alle statistisch signifikant sind. Allerdings weisen die drei Versionen zum Teil signifikant voneinander abweichende Werte auf; dies dürfte zum einen auf kulturspezifische Differenzen, zum anderen durch die Übersetzungen bedingt sein.

In unserer Untersuchung konnten diese Ergebnisse nicht repliziert werden (vgl. Tab. 4). Während zwar für die zusammengefaßten Stichproben alle Trennschärfekoeffizienten positiv und statistisch signifikant sind, sind bei den einzelnen Gruppen nur mehr wenige statistisch von Null unterschieden, einige weisen sogar negative Trennschärfen auf. Dieses Resultat bedeutet, daß aufgrund der untersuchten Pbn behauptet werden kann, daß der Test zwar gut zwischen Depressiven und Nicht-Depressiven zu differenzieren vermag, nicht aber innerhalb der einzelnen Gruppen, selbst nicht bei der Gruppe der klinisch Depressiven. Der Test ist aber primär nicht dazu geschaffen, um „Kranke" von „Normalen" zu unterscheiden, sondern um den Ausprägungsgrad der Depression bei verschieden stark Depressiven zuverlässig zu messen (*Beck*, 1967, p. 187).

Wollte man sich allein auf die hier wiedergegebenen Resultate verlassen, so müßte man eine neue Testfassung ohne die Items R und S konzipieren. Auch müßte man überlegen, ob die Items, die keine signifikante Korrelation mit dem Gesamttestwert aufweisen, nicht revidiert werden sollten. Da die anderen Untersuchungen (vgl. Tab. 3) positivere Ergebnisse gebracht haben, ist eventuell zu schließen, daß die Güte des Tests durch die Übersetzung gelitten hat und daß deshalb bisweilen andere Formulierungen bei den Itemantworten vorteil-

hafter wären. Vorerst ist aber auch zu schauen, ob sich auch bei weiteren und größeren Stichproben diese negativen Resultate replizieren lassen.

2.3. Zuverlässigkeit (Reliabilität)

Unter Reliabilität (von *Pöldinger* et al., 1969 a, p. 81, mit Konstruktvalidität konfundiert) wird die formale Meßgenauigkeit eines Tests verstanden, und zwar unter Absehung von dem, was der Test zu erfassen vorgibt. Als Methoden zur Erfassung der Reliabilität dienen die Paralleltestreliabilität, Testwiederholungsmethode, Halbierungsreliabilität und Schätzungen der inneren Konsistenz eines Tests. Es ist dabei eine offene Frage, ob durch diese Methoden immer dasselbe Konzept erfaßt wird oder verschiedene. Zumindest sind es verschiedene Aspekte der Meßgenauigkeit, die durch diese Verfahren geprüft werden.

2.3.1. Paralleltestreliabilität. Angaben zur Paralleltestreliabilität können für das vorliegende Verfahren nicht gemacht werden, da es nur in einer Version vorliegt.

2.3.2. Wiederholungsreliabilität. Bei der Korrelation von Testergebnissen, die zu zwei aufeinanderfolgenden Zeitpunkten aufgenommen wurde, kommen mehrere Faktoren ins Spiel, welche die endgültige Höhe des so erhaltenen „Koeffizienten der zeitlichen Stabilität" (*APA*, 1954) beeinflussen. Werden z. B. Patienten zu Beginn und zu Ende der Therapie getestet, so ergibt sich durch den Therapieeinfluß eine Fluktuation der Testresultate. Natürlich nimmt der Einfluß unkontrollierbarer Effekte mit der Zeit, die zwischen den Testungen liegt, ebenfalls zu. Diese Veränderungen müssen die Höhe des Wiederholungskoeffizienten nicht unbedingt beeinflussen: ändert sie sich bei allen Pbn um einen konstanten Betrag, so kann der Koeffizient trotzdem maximal werden. Da klinische Untersuchungen vor allem an den Resultaten der Therapie interessiert sind, werden solche Berechnungen nicht vorgenommen. Es wird hingegen geprüft, ob die Veränderung der Testwerte den Therapieerfolg widerspiegelt (*Beck*, 1967, p. 199; *Blaser* et al., 1968 a, p. 310; *Nussbaum* et al., 1963); diese Frage, ob das Verfahren sensitiv genug ist, um therapiebedingte Veränderungen zu erfassen, gehört aber hauptsächlich zur Validitätsproblematik. Numerische Angaben zur Wiederholungsreliabilität liegen leider nicht vor.

2.3.3. Testhalbierungsreliabilität. Teilt man einen Test – beispielsweise nach den geradzahligen und den ungeradzahligen Testitems – in zwei Hälften, so kann man durch die Korrelation dieser beiden Werte bestimmen, inwieweit die beiden Testhälften einander gleichwertig sind. Da diese Reliabilitätsschätzung nur auf jeweils der Hälfte der Items basiert und bekanntlich die Reliabilität auch eine Funktion der Anzahl der Testitems ist (*Magnusson*, 1969, p. 81), wertet man den so erhaltenen Koeffizienten nach der Methode von *Spearman-Brown* auf.

Für das DI liegen dabei die folgenden Informationen zu diesem Aspekt der Reliabilität vor (vgl. Tab. 5). Wie man sieht, ist der Wert, den *Beck* angibt, befriedigend hoch, auch der Wert für die Gesamtstichprobe unserer Unter-

Tabelle 5: *Halbierungskoeffizienten aus verschiedenen Untersuchungen, aufgewertet nach Spearman-Brown*

	Beck et al. (1961) N = 200	Gruppen 1, 2, 3 N = 102	Gruppe 1 N = 30	Gruppe 2 N = 30	Gruppe 3 N = 42
r_{tt}	.86	.85	.47	.19	.49
r_{tt}, korr.	.93	.92	.65	.32	.64

suchung ergibt einen zufriedenstellenden Koeffizienten. Die Halbierungskoeffizienten für die einzelnen Gruppen sind aber unzureichend, denn bekanntlich wird von Tests gefordert, daß sie Reliabilitätskoeffizienten von 0,85 bis 0,95 aufweisen sollten (*Ekmann*, 1955, p. 35; *Hofstätter*, 1962, p. 292). Auch dieses Ergebnis kommt dadurch zustande, daß die Testwerte vorwiegend die Gruppenzugehörigkeit wiedergeben und nicht eine differenzierte Ordnung nach dem Kriterium der Tiefe der Depression innerhalb der drei Gruppen.

2.3.4. Konsistenzschätzungen. Unter Konsistenz eines Tests wird das Ausmaß verstanden, „in welchem alle Items eines Tests dasselbe messen" (*Horst*, 1971, p. 310). Wie *Horst* kritisiert, ist es allerdings nicht so, daß die Konsistenz oder Homogenität eines Tests mit seiner Reliabilität zusammenfallen muß. In der Literatur wird aber auch jene als ein Aspekt der Reliabilität angesehen.

Als Methoden der Konsistenzschätzung dienen die Verfahren von *Kuder-Richardson* (*Lienert*, 1967, p. 225 ff.) und von *Cronbach* (1951). Numerische Angaben dazu sind in der Studie von *Weckowitz* et al. (1967, p. 26) enthalten und in unserer eigenen Untersuchung (vgl. Tab. 6).

Tabelle 6: *Konsistenzschätzungen für das DI*

	Weckowitz (1967) N = 254	N = 391		Gruppen 1, 2, 3 N = 102	Gr. 1 N = 30	Gr. 2 N = 30	Gr. 3 N = 42
r_{tt} nach *Kuder-Richardson*	.53	.78	alpha nach *Cronbach*	.92	.66	.54	.60

Wie bereits *Weckowitz* schreibt, reflektiert der Wert aus seiner ersten Stichprobe die Tatsache, daß die Iteminterkorrelationen des Tests relativ gering sind. Dies hat u. a. die Ursache darin, daß in diese Stichprobe nur Patienten aufgenommen wurden, die einen Gesamtwert von über 17 Punkten besaßen. Die Ergebnisse für die Gesamtstichprobe sind etwas besser, da hier die volle Variationsbreite des Tests ausgenützt wird. Ähnliches zeigte sich auch in unserer Untersuchung: der Konsistenzkoeffizient für die Gesamtstichprobe ist befriedigend hoch, und man würde meinen, man hat es hier mit einem idealen Test hinsichtlich der Homogenität zu tun. Leider sinken die Koeffizienten für die einzelnen Gruppen wieder beträchtlich ab, d. h. man muß genauso wie vorher feststellen, daß das Kriterium der Homogenität für das vorliegende Verfahren nicht erfüllt ist. Dies hat aber auch noch weitere Konsequenzen: ein hoher Konsistenzkoeffizient ist eine Voraussetzung, um die Werte der einzelnen Testitems

zu einem Gesamttestwert addieren zu können. Diese Berechtigung scheint hier in Frage gestellt zu sein.

2.4. Gültigkeit (Validität)

Vorläufig sei unter Validität eines Tests seine „psychologische Bedeutung" (*Drenth*, 1969, p. 180) verstanden, eigentlich ist aber die theoretische Bedeutung gemeint, die er besitzt (*Lukesch*, 1973 a).

2.4.1. Logische bzw. inhaltliche Validität. Die einzelnen Fragen, die das Depressionsinventar enthält, beziehen sich auf Symptome, die für Depressive kennzeichnend sind. Ein Maß für diese Art der Gültigkeit gibt es nicht, außer man wollte Fachleute beurteilen lassen, ob die aufgeführten Fragen auch tatsächlich für diese Art der Erkrankung zutreffen. Nach den Angaben des Testautors (*Beck*, 1967, p. 189) sind aber die Fragen charakteristisch für die von Patienten geschilderten Symptome, was bedeuten würde, daß dem Verfahren logische und auch inhaltliche Validität zukommt.

2.4.2. Empirische Validität. Unter diesem Begriff kann man alle Untersuchungen subsumieren, in denen über empirisch feststellbare Korrelationen zu Außenkriterien (vgl. Abb. 1) berichtet wird (z. B. psychiatrische Beurteilungen, Kor-

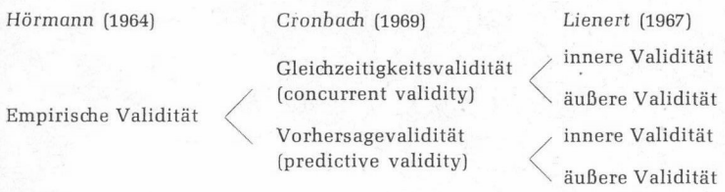

Abbildung 1: *Gliederungsmöglichkeiten der empirischen Validität*

relationen zu anderen Tests, Beziehung zu bestimmten vorher definierten Gruppen usw.).

2.4.2.1. Gleichzeitigkeitsvalidität. Werden Test und Kriterium zum selben Zeitpunkt erhoben, so spricht man von Gleichzeitigkeitsvalidität (*Cronbach*, 1969, p. 103). Hier kann man wieder unterscheiden, ob es sich bei dem Kriterium, das verwendet wird, um einen anderen Test handelt oder um ein Kriterium, das selbst kein Test ist (innere und äußere Validität, *Lienert*, 1967).

2.4.2.1.1. Innere Validität. Bedeutsame Korrelationen des DI zu anderen Verfahren, die ebenfalls vorgeben, Depressivität zu messen, konnten gefunden werden (vgl. Tab. 7).

Es zeigt sich jedoch dabei, daß in der Literatur nicht allzu häufig über solche Koeffizienten berichtet wird. Die Korrelationen mit Verfahren, die vorgeben, dasselbe zu messen, sind bis auf eine Ausnahme alle sehr signifikant; es macht dabei keinen Unterschied, ob die Verfahren auf Selbst- oder auf Fremdbeurteilungen beruhen. Allerdings weist ihre Höhe auch darauf hin, daß die Skalen nicht identische Merkmale erfassen; dies ist eine Tatsache, auf die schon *Fahy*

Tabelle 7: *Korrelationen des DI mit anderen Testverfahren*

	Fahy (1969) N = 51	*Schwab* et al. (1967 b) N = 153	*Nußbaum* et al. (1963) N = 19	*Pöldinger* et al. (1969 a) N = 24	*Vinar* (1966) N = 41	*Lubin* (1965) N = 62
Self-rating-depression-scale (*Zung*, 1965)	.57***					
Hamilton-rating-scale (*Hamilton*, 1967)	.25*	.75***			.68***	
Depression-adjective-checklist (*Lubin*, 1965)	.42***				.40 – .66	
Mental-status-schedule (*Spitzer* et al., 1967)	.45***					
MMPI-D-Skala			.75***			
Depression-rating-scale (*Wechsler* et al., 1963)				.79***		

*** sign. 1%, ** sign. 5%, * sign. 10%

(1969, p. 314) hingewiesen hat: „ . . . depressive rating scales seldom duplicate each other. Depending on their design they reflect qualitatively different patterns of depression."

2.4.2.1.2. Äußere Validität. Interessant sind hier vor allem die Beziehungen, die man zwischen dem DI und psychiatrischen Beurteilungen gefunden hat (hier sollen solche psychiatrische Beurteilungen referiert werden, die nicht auf standardisierten Interviews beruhen, sondern eher auf dem klinischen Eindruck). Man findet in den Untersuchungen (vgl. Tab. 8) durchwegs sehr signifikante

Tabelle 8: *Korrelationen des DI mit psychiatrischen Beurteilungen*

Autor	N	Art des Ratings	DI
Nußbaum et al. (1963)	19	pre-treatment-rating post-treatment-rating	.66*** .73***
Metcalfe et al. (1965)	120	globales psychiatrisches Rating	.62***
Beck (1967)	226 183	alternative Beurteilung	.65*** .67***
Fahy (1969)	51	globales psychiatrisches Rating	.37***
Pichot (1969)	41	globales psychiatrisches Rating	.41***
Lukesch (1973)	60	depressiv / nicht depressiv	.898***

*** sign. 1%

Beziehungen berichtet. Die Höhe der Koeffizienten weist aber wieder darauf hin, daß es sich nicht um identische Maße handelt. Dies ist auch weiter nicht verwunderlich, da das DI die subjektiv erlebte Schwere der Depression erfassen soll, während die Beurteilungen von Klinikern sich an äußerlich feststellbaren Merkmalen orientieren. Der sehr hohe Wert, der sich bei unserer eigenen Untersuchung ergeben hat, ist darauf zurückzuführen, daß die Verteilungen für das Gesamtresultat sich für die Gruppen der „Gesunden" und der „Kranken" nicht überlappen. Es kann aber auch bei diesem positiven Resultat nur festgestellt werden, daß der DI die Gruppenzugehörigkeit in valider Weise wiedergibt; nicht kann aber gefolgert werden, daß durch dieses Ergebnis auch belegt sei, daß das DI auch eine zuverlässige Reihung der Pbn nach dem Merkmal „Depression" ermöglicht, die auch innerhalb der verschiedenen Gruppen Gültigkeit besitzt.

2.4.2.2. Vorhersagevalidität. Informationen, in welchen Beziehungen der DI, der zu einem bestimmten Zeitpunkt erhoben wird, zu Kriterien steht, die in der Zukunft erst erfaßt werden können, liegen nicht vor.

2.4.3. *Konstruktvalidierung.* Dieser Begriff bezieht sich auf die Forderung, daß ein Testverhalten durch eine dahinterstehende Theorie erklärbei sei, m. a. W. der Test soll eine optimale Operationalisierung von theoretisch bedeutsamen Parametern darstellen. Rückschlüsse darauf kann man für das DI einmal in der Weise ziehen, daß man die Argumente, die für seine inhaltliche Validität zählen können, wiederholt. Andere Möglichkeiten bestehen ebenfalls: wie *Hörmann* (1964, p. 36) angibt, kann ein Hinweis auf Konstruktvalidität gefunden werden, indem man die Beziehungen (das nomologische Netzwerk, d. h. das Netzwerk gesetzesmäßig miteinander verbundener Aussagen) zu anderen Kriterien aufweist, die theoretisch vorhergesagt oder in anderen Untersuchungen schon empirisch bewährt sind; letztlich sei auch noch auf Versuche hingewiesen, die Konstruktvalidität durch faktoranalytische Untersuchungen zu belegen (faktorielle Validität).

2.4.3.1. Beziehungen zu weiteren Kriterien. Hier sind die Untersuchungen von *Beck* (1967, p. 203) und seinen Mitarbeitern zu erwähnen, die feststellen konnten:

a) Patienten mit höheren DI-Werten berichten signifikant häufiger Träume mit masochistischem Inhalt (*Beck & Ward*, 1961) und die Korrelation mit einem Masochismus-Fragebogen ist statistisch sehr signifikant (N = 109, r = .51) von Null unterschieden (*Beck*, 1961).

b) Hohe Werte in dem DI sind weiters mit einem negativen Selbst-Konzept verbunden (*Beck & Stein*, 1960).

c) Patienten mit hohen Werten tendieren weiters dazu, nach einem schlechten Abschneiden sich selbst extrem wenig zuzutrauen (*Loeb* et al., 1964) und

d) ihre tatsächliche Leistung zu unterschätzen (*Loeb* et al., 1964; 1967; 1971).

e) Das DI korreliert nur gering mit psychiatrischen Beurteilungen (N = 606, r = .14) über die Ängstlichkeit der Patienten (*Beck*, 1967, p. 201).

f) Es besteht eine positive Beziehung zur Feindseligkeit gegen sich selbst (r = .47), die ebenfalls signifikant ist (*Gottschalk* et al., 1963), und eine negative Korrelation zu nach außen gerichteter Feindseligkeit.

g) Schließlich zeigte sich auch eine bedeutsame negative Beziehung zwischen psychiatrischen Ratings, aufgrund des DI und des MMPI, und den Ergebnissen eines Tests, der den Sinn für Humor erfassen sollte (N = 18, r = .42–.73) (*Nußbaum & Michaux*, 1963).

Diese Ergebnisse sind eigentlich zu erwarten gewesen. Offen bleibt aber die Frage, ob diese auch für die übersetzten Versionen des DI Geltung haben. Durch die viel geringere Reliabilität der deutschen Version ist zumindest zu befürchten, daß sich diese Beziehungen nicht in gleicher Höhe bestätigen lassen werden.

h) *Blaser* et al. (1968 b, p. 36) fanden mit dem Fragebogen von *Beck-Pichot* einen auf dem Faktor „traurige Verstimmung" signifikant höheren Wert bei endogenen Depressiven (N = 80, p = 0,05); die Werte bei den anderen Dimensionen sind zwar ebenfalls bei den endogenen Depressiven erhöht, aber statistisch nicht bedeutsam.

i) *Pöldinger* et al. (1969 b, p. 294) fanden ebenfalls bei endogenen Depressiven (N = 73) tatsächlich höhere Werte im DI als bei psychogen Depressiven (N = 69) (p = 0,05).

j) Sie fanden außerdem, daß somatische Beschwerden (Einschlaf-, Durchschlafstörungen, frühes Erwachen, Kopfschmerzen, Schwindel und Herzbeschwerden) mit der Depressionstiefe signifikant zunehmen, sowie auch die Neigung zu Hypochondrie.

k) Der Therapieerfolg spiegelt sich auch in einem Absinken der Werte in dem DI wider (vgl. Tab. 9). *Beck* (1967, p. 199) fand, daß die Änderungen in der

Tabelle 9: *Therapieerfolg und DI*

	Fahy (1969) N = 50		*Blaser* et al. (1968) N = 41	
	Eintritt	Entlassung	Eintritt	Entlassung
$\bar{X}$	26,6	12,9	23,3	8,4
s	1,51	1,63		

Depressionstiefe, gemessen mittels DI und psychiatrischen Urteils, in 85 % der Fälle übereinstimmen.

2.4.3.2. Faktorielle Validität. Der Einsatz von Faktorenanalysen zur Konstruktvalidierung ist nur in dem Ausmaß gerechtfertigt, in welchem es Theorien gibt, die durch das Verfahren der Faktorenanalyse bewährt werden können. Zumeist ist das Vorgehen aber umgekehrt, d. h. man führt eine Faktorenanalyse durch und versucht dann die Ergebnisse in Übereinstimmung mit implizit gehegten Hypothesen zu bringen, d. h. die Faktorenanalyse wird in ihrer theoretischen Tragweite überschätzt, indem sie sozusagen die Theorie liefern soll, die einem Testverfahren zugrunde liegt. Abgesehen davon, ist die Anwendung der Faktorenanalyse zumeist mit schwerwiegenden Mängeln behaftet, deren Aufzählung im einzelnen aber zu weit führen würde (*Lukesch*, 1973 b).

Zu dem DI liegen einige faktorenanalytische Untersuchungen vor (*Fahy*, 1969; *Pichot & Lemperiere*, 1964; *Weckowitz* et al., 1967) und auch eine eigene. Es ist dabei aber so, daß die durchgeführten Analysen nicht der Bewährung einer vorher konzipierten Theorie dienen, sondern einfach Versuche sind, dieses Verfahren einmal anzuwenden und zu sehen, ob sich interpretierbare Ergebnisse zeigen.

Pichot (1969 a) führte mit der erweiterten Version des DI mit 33 Items Faktorenanalysen (N = 133) nach mehreren verschiedenen Methoden durch, die jeweils zu einer verschieden großen Zahl von interpretierbaren Faktoren führten (3–6). Er glaubt aber doch aus dem Vergleich der verschiedenen Lösungen fünf gemeinsame interpretierbare „Kernfaktoren" zu finden, nämlich a) vitale Verlangsamung, b) Strafbedürfnis, c) depressive Stimmung, d) somatische Beschwerden, e) quälenden „Monoideismus". Nach der von *Beck* (1967, p. 204) wiedergegebenen Faktorenanalyse von *Pichot* zu schließen, erklärt der Faktor „vitale Depression" (gekennzeichnet vor allem durch somatische Symptome) den größeren Anteil der Varianz. Das würde auch mit der Feststellung von *Fahy* (1969, p. 313) übereinstimmen, daß das DI eher „endogene" Komponenten erfaßt.

Weckowicz et al. (1967) extrahierten bei ihrer Analyse (N = 254) sechs Faktoren, die allerdings insgesamt nur 34,4 % der Varianz in den Daten wiedergeben können, und schließen, daß drei davon interpretierbar sind. Als Benennungen wählen sie a) guilty depression, b) retarded depression, c) somatic disturbance. Sie glauben aber auch, daß ihre Analyse nur heuristischen Wert besitzt und durch andere Untersuchungen (mit physiologischen Maßen und Daten aus der Verhaltensbeobachtung) ergänzt werden müßte.

Von uns wurden mehrere Faktorenanalysen[3]) gerechnet. Als Ausgangsdaten dienten die Items aus dem DI und einige daraus abgeleitete Maße (Gesamtwert; affektive, soziale, somatische Beschwerden nach *Blaser* et al., 1968 b). Die durchgeführten Analysen können sehr stark kritisiert werden, da die Versuchspersonenanzahl bei den Einzelgruppen nicht in der erwünschten Relation zur Zahl der Variablen steht (*Pawlik*, 1968, p. 278). Aus diesem Grunde soll nur eine grobe Sichtung der Resultate vorgenommen werden (vgl. Tab. 10).

Tabelle 10: *Ergebnisse von Faktorenanalysen des DI an mehreren Stichproben*

	Gruppen 1, 2, 3	Gruppen 1, 2	Gruppe 1	Gruppe 3	Gruppe 2
Anzahl der extrahierten Faktoren	2	2	4	5	5
Varianzaufklärung (in % der Ges.-Var.)	51,91	55,14	46,43	46,28	47,45

Die Analysen zeigen vor allem, daß die Interkorrelationen bei den geteilten Gruppen der Depressiven und Nichtdepressiven stark absinken; aus diesem

[3]) Kommunalitätenschätzung wurde iterativ vorgenommen; extrahiert wurde nach der Hauptachsenmethode; die Anzahl der Faktoren wurde durch das Kriterium „Eigenwerte größer Eins" festgelegt; anschließend wurde eine Varimax-Rotation durchgeführt.

Grunde nimmt die Faktorenanzahl bei den Unterstichproben zu. Trotz mehrerer Faktoren, die das Extraktionskriterium erfüllen, ist der Anteil der aufgeklärten Varianz geringer. Aus den Matrizen der Faktorenladungen geht hervor, daß bei den beiden ersten Analysen nur unipolare Faktoren vorkommen, während die Faktoren bei den einzelnen Gruppen auch bipolar ausgeprägt sind. Die Kommunalitäten der vier letzten Variablen sind jeweils sehr hoch. Dies ist aber leicht zu erklären, da es sich hierbei um abgeleitete Maße handelt. Die Faktorendifferenzierung wurde durch eine Homogenisierung der Stichproben erreicht. Es ist u. E. daher müßig, eine Interpretation oder Benennung der einzelnen Faktoren vorzunehmen. Letztlich würde dies nur zu dem Ergebnis führen, daß man sich auf eine Interpretation einigt, die am plausibelsten erscheint (d. h. eine Gruppierung der Items nach a priori Gesichtspunkten) und die nicht völlig mit den extrahierten Faktoren in Deckung gebracht werden kann (vgl. *Pichot*, 1969 a).

Abschließend kann man zu dem Punkt der Konstruktvalidität sagen, daß eine Reihe von Beziehungen mit dem DI von *Beck* nachweisbar sind, die durch die Depressionsforschung im allgemeinen vorhergesagt werden konnten. Faktorielle Validierungsversuche schlagen aber fehl, was nicht als Mangel des Tests gedeutet werden soll, sondern eher durch die angewandte Methode bedingt ist.

2.5. Normierung

Um aus einem Test diagnostisch brauchbare Schlüsse ziehen zu können, ist es vorteilhaft, wenn er normiert ist. Normierung nach verschiedenen Aspekten ist dann vor allem notwendig, wenn sich zeigen sollte, daß seine Ergebnisse von „äußeren" Variablen (wie Alter, Geschlecht, sozio-ökonomischem Status) abhängig sind. (Von den Möglichkeiten der Testkonstruktion von stichprobenunabhängigen Tests nach dem Rasch-Modell sei in diesem Zusammenhang abgesehen, *Fischer*, 1967.)

2.5.1. Beziehung zum Alter. Wiederholt wurde die Hypothese aufgestellt, daß ältere Patienten einen höheren Wert in Depressionsfragebögen erreichen müßten. *Blaser* et al. (1968 b, p. 305) fanden bei älteren Patienten zwar tendenziell höhere Werte, die aber nicht statistisch signifikant waren. Ein Hinweis auf tendenziell höhere Werte bei Älteren ist auch bei *Pöldinger* et al. (1969 b, p. 295) zu finden; Angaben über statistische Signifikanz jedoch fehlen. Korrelative Beziehungen zum Alter konnten ebenfalls nicht nachgewiesen werden (vgl. Tab. 11). Auch bei der Faktorenanalyse, die *Fahy* (1969, p. 310) durchführte,

Tabelle 11: *Korrelationen zwischen DI und Alter*

Metcalfe et al. (1965)	*Beck* (1967)	Gruppen 1, 2	Gruppe 1	Gruppe 2
N = 120	N = 606	N = 60	N = 30	N = 30
.12	.025	.009	.018	—.081

zeigte sich, daß zu dem Faktor, auf dem das DI hoch lädt (.824), die Altersvariable in keiner Beziehung steht (.091), während für den Faktor, der die Altersvariable hauptsächlich enthält (.749), gerade das Umgekehrte gilt (DI = .070).

Aus diesen vorläufigen Ergebnissen kann man schließen, daß eine Normierung des Verfahrens nach Altersgruppen trotz anderer Erwartungen nicht notwendig ist.

2.5.2. Beziehungen zum Geschlecht. Hier liegen die Verhältnisse etwas anders, das DI scheint anfällig zu sein für geschlechtsspezifische Differenzen. Bei *Metcalfe* et al. (1965, p. 242) liegen zwar negative Befunde vor, aber *Pöldinger* et al. (1969 b, p. 295) konnten bei Frauen (N = 73) im Vergleich zu Männern (N = 54) signifikant höhere Werte nachweisen und damit eine Vermutung von *Blaser* et al. (1968 b, p. 305) stützen. Die Größe dieses Effekts darf aber nicht überschätzt werden. Dies sieht man aus den Korrelationen zwischen Geschlecht und DI (vgl. Tab. 12). Als einziger Koeffizient ist der von *Beck* angeführte signifikant, erklärt aber auch nur einen eigentlich zu vernachlässigenden Anteil an der Variabilität der Daten (etwa 4 %). Vorsichtigerweise müßte man aber bei einer Normierung des Verfahrens solche Unterschiede berücksichtigen.

Tabelle 12: *Korrelationen zwischen DI und Geschlecht*

	Beck (1967)	Gruppen 1, 2	Gruppe 1	Gruppe 2
	N = 606	N = 60	N = 30	N = 30
	—.189***	—.030	.014	—.025

*** sign. 1%

2.5.3. Beziehungen zum sozio-ökonomischen Status. Die empirischen Befunde zu diesem Punkt sind sehr spärlich. *Beck* (1967) führt einen negativen Korrelationskoeffizienten zwischen DI und der Höhe der Schulausbildung (r = —.163) und auch *Schwab* et al. (1967 b) geben an, daß einige Symptome, die kennzeichnend für Depression sind, mit niedrigem sozio-ökonomischem Status einhergehen.

2.5.4. Differenzierung zwischen „Normalen" und „Depressiven". Wendet man das DI bei nicht-depressiven Probanden an, so zeigt sich, daß die von den Patienten abgegebenen Selbstratings klar zwischen Kranken und Nicht-Kranken unterscheiden (Tab. 13).

Tabelle 13: *Mittelwerte und Streuungen der Gesamtwerte aus dem DI*

	Metcalfe et al. (1965)	*Beck* (1967)	Gruppe 1	Gruppe 2	Gruppe 3
	N = 32	N = 115	N = 30	N = 30	N = 42
$\bar{X}$	5,4	10,9	24,33	4,00	4,79
s	5,8	8,1	6,63	2,75	3,39

Die Verteilung der Gesamtwerte aus dem DI der Gruppe der psychisch Gesunden überlappt sich dabei nicht mit der Gruppe der Depressiven. Aus diesem Grunde ist auch die Korrelation zwischen der Zugehörigkeit zur Gruppe der Gesunden bzw. Kranken mit r = —.898 erwartungsgemäß sehr hoch. Die Werte

der Gruppen 2 und 3 stimmen aber eher mit den Angaben von *Metcalfe & Goldman* überein als mit denen von *Beck*.

2.5.5. Tiefe der Depression. Es ist eine bisweilen strittige Frage, ob mit einem bloß quantitativen Konzept alle Depressionsformen erfaßbar seien oder ob unterschiedliche nosologische Zuordnungen nur auch in verschiedenen Werten im DI zum Ausdruck kommen. Dieser wichtigen Frage kann in diesem Zusammenhang nicht weiter nachgegangen werden. Es sollen nur die Werte aus verschiedenen Untersuchungen wiedergegeben werden (vgl. Tab. 14). Diese können

Tabelle 14: *Die Tiefe der Depression im DI*

	Metcalfe et al. (1965)			*Beck* (1967)			*Blaser* et al. (1968 b)	
	leichte	mittlere	schwere	leichte	mittlere	schwere	psychogen	endogen
	N = 44	N = 20	N = 24	N = 127	N = 134	N = 33		
$\bar{X}$	14,3	24,2	29,5	18,7	25,4	30,0	18,3	25,4
s	8,3	10,8	6,5	10,2	9,6	10,4		

als Interpretationshilfen gelten. Die Angaben von *Blaser* deuten auch eine mögliche Beziehung zwischen den Werten im DI und nosologischen Einheiten an.

Zusammenfassung

Es wurde versucht, mittels der in der Literatur angegebenen und eigener Ergebnisse einen Überblick über die Gütekriterien des Depressionsinventars von *Beck* et al. zu geben. Dazu kann man im einzelnen feststellen:

a) Die Standardisierung des Verfahrens ist noch nicht so weit gediehen, wie es für einen Test der Fall sein sollte.

b) Für die deutsche Übersetzung haben sich aufgrund eigener Ergebnisse schwerwiegende Einwände bezüglich der Itemcharakteristiken erheben lassen. Sollten sich die Ergebnisse an anderen und größeren Stichproben replizieren lassen, so wäre eine Revidierung des DI angezeigt.

c) Auch die Angaben zur Zuverlässigkeit des Verfahrens sind nicht optimal.

d) Die Untersuchungen zur Gültigkeit des Verfahrens haben in sehr vielen Fällen gezeigt, daß die Ergebnisse mit begründeten Erwartungen aus der allgemeinen Depressionsforschung übereinstimmen. Diese Untersuchungen wurden aber zumeist mit der Originalform des DI vorgenommen; es kann also nicht als sicher angenommen werden, daß sich diese Ergebnisse bestätigen lassen.

e) Das DI erlaubt eine eindeutige Trennung von Depressiven und Nicht-Depressiven. Aufgrund der dargestellten Ergebnisse ist es aber unwahrscheinlich, daß die Werte aus dem DI eine graduelle Reihung der Probanden nach dem Merkmal „Depression" erlauben.

f) Es erscheint möglich, daß bei Gruppen, die durch grobe Klassifizierungen (mehr kann aufgrund der Reliabilitätsmängel nicht erreicht werden) mittels des DI zusammengefaßt werden, Unterschiede nachgewiesen werden können (z. B.

bei physiologischen Messungen), deren Erklärung auch theoretisch von Interesse sein könnte. Insofern ist diesem ad-hoc konstruierten Verfahren ein heuristischer Wert nicht abzusprechen.

LITERATUR

APA, *Technical recommendations for psychological tests and diagnostic techniques*, *Psychological Bulletin*, 1954, *51*, 201—238.

ASH, P., The reliability of psychiatric diagnosis, *Journal of Abnormal and Social Psychology*, 1949, *44*, 272—276.

BECK, A. T., A systematic investigation of depression, *Comprehensive Psychiatry*, 1961, *2*, 162—170.

BECK, A. T., *Depression: clinical, experimental and theoretical aspects*, New York, Harper & Row, 1967.

BECK, A. T., FESHBACH, S., & LEGG, D., The clinical utility of the digit symbol test, *Journal of Consulting Psychology*, 1962, *26*, 263—268.

BECK, A. T., & STEIN, D., The self-concept. Manuscript, 1960 (zit. nach A. T. BECK, 1967).

BECK, A. T., & WARD, C. H., Dreams of depressed patients: characteristic themes in manifest content, *Archives of General Psychiatry*, 1961, *5*, 462—467.

BECK, A. T., WARD, C. H., MENDELSON, M., MOCK, J., & ERBAUGH, J., An inventory for measuring depression, *Archives of General Psychiatry*, 1961, *4*, 561—571.

BLASER, P., KÖNIG, U., & PÖLDINGER, W., Das Problem der Quantifizierung in der Depressionsforschung, in: F. LABHARDT (Ed.), *Depressionen und ihre Behandlung*, Basel, Karger, 1968, p. 33–42. (a)

BLASER, P., LÖW, P., & SCHÄUBLEIN, A., Die Messung der Depressionstiefe mit einem Fragebogen, *Psychiatria Clinica*, 1968, *1*, 299–319. (b)

CARL, W., Eine Untersuchung zur Faktorenstruktur von Antworttendenzen ("response sets") bei Antwortskalen verschiedener Stufenzahl, *Zeitschrift für experimentelle und angewandte Psychologie*, 1968, *15*, 419—434.

CATTELL, R. B., *Handbook of multivariate experimental psychology*, Chicago, Rand McNally, 1966.

CRONBACH, L. J., *Studies of acquiescence as a factor in the true-false test*, *Journal of Educational Psychology*, 1942, *33*, 401—415.

CRONBACH, L. J., Coefficient alpha and the internal structure of tests, *Psychometrika*, 1951, *16*, 297—334.

CRONBACH, L. J., *Essentials of psychological testing*, New York, Harper & Row, 1969.

DELAY, J., PICHOT, P., LEMPERIERE, T., & MIROUZE, R., La nosologie des états dépressifs: rapports entre l'étiologie et la semilogie. 2. Résultats due Questionaire de Beck, *Encéphale*, 1963, *52*, 497—505.

DOERING, C. R., & RAYMOND, A. F., Reliability of observation in psychiatric and related characteristics, *American Journal of Orthopsychiatry*, 1934, *4*, 249—257.

DRENTH, P. J., *Der psychologische Test*, München, Barth, 1969.

EDWARDS, A. E., The relationship between the judged desirability of a trait and the probability that the trait will be endorsed, *Journal of Applied Psychology*, 1953, *37*, 90—93.

EKMAN, G., Konstruktion und Standardisierung von Tests, *Diagnostica* 1955, *1*, 15—19, 32—36.

FAHY, T., Some problems in the assessment of current mental status of depressed patients, in: H. HIPPIUS & H. SELBACH (Ed.), 1969, p. 305—316.

FISCHER, G. (Ed.), *Testtheorie*, Bern, Huber, 1968.

FOGEL, M. L., CURTIS, G. C., KORDASZ, F., & SMITH, W. G., Judges' ratings, self-ratings, and checklist report of affects, *Psychological Reports*, 1966, *19*, 299—307.

GOTTSCHALK, L., GLESER, G., SPRINGER, K., Three hostility scales applicable to verbal samples, *Archives of General Psychiatry*, 1963, *9*, 254—279.

GULLIKSEN, H., *Theory of mental tsets*, New York, Wiley, 1950.

HAMILTON, M., Development of a rating scale for primary depressive illness, *British Journal of Social and Clinical Psychology*, 1967, *6*, 278—296.

HIPPIUS, H., & SELBACH, H. (Ed.), *Das depressive Syndrom*, München, Urban & Schwarzenberg, 1969.

HOETH, F., BÜTTEL, R., & FEYERABEND, H., Experimentelle Untersuchungen zur Validität von Persönlichkeitsfragebögen, *Psychologische Rundschau*, 1967, *18*, 169—184.

HÖRMANN, H., *Aussagemöglichkeiten psychologischer Diagnostik*, Göttingen, Hogrefe, 1964.

HOFSTÄTTER, P. R., *Psychologie*, Frankfurt/Main, Fischer, 1962.

HORST, P., *Messung und Vorhersage*, Weinheim, Beltz, 1971.

LIENERT, G. A., *Testaufbau und Testanalyse*, Weinheim, Beltz, 1967.

LOEB, A., BECK, A. T., & DIGGORY, J., Differential effects of success and failure on depressed and nondepressed patients, *Journal of Nervous and Mental Disease*, 1971, *152*, 106—114.

LOEB, A., BECK, A. T., DIGGORY, J. C., & TUTHILL, R., Expectancy, level of aspiration, performance, and self-evaluation in depression, *Proceedings of the 75th Annual Convention of the American Psychological Association*, 1967, *2*, 193—194.

LOEB, A., FESHBACH, S., BECK, A. T., & WOLF, A., Some effects of reward upon the social perception and motivation of psychiatric patients varying in depression, *Journal of Abnormal and Social Psychology*, 1964, *68*, 609—616.

LUBIN, B., Adjective checklists for measurement of depression, *Archives of General Psychiatry*, 1965, *12*, 57—62.

LUKESCH, H., Zur Validitätsfrage in der psychologischen Diagnostik: Die Reformulierung eines Problems, *Zeitschrift für Klinische Psychologie und Psychotherapie*, 1973 (im Druck). (a)

LUKESCH, H., Die Anwendung der Faktorenanalyse. Kritik der Praxis einer Methode, Vortrag, gehalten am Psychologischen Institut der Freien Universität Berlin, Jänner 1973. (b)

LUKESCH, H., Depression und Intelligenz, *Zeitschrift für Klinische Psychologie und Psychotherapie*, 1974 (im Druck).

MAGNUSSON, D., *Testtheorie*, Wien, Deuticke, 1969.

METCALFE, M., & GOLDMAN, E., Validation of an inventory for measuring depression, *British Journal of Psychiatry*, 1965, *111*, 240—242.

MICHEL, L., Allgemeine Grundlagen psychometrischer Tests, in: R. HEISS (Ed.), *Handbuch der Psychologie, Bd. 6, Psychologische Diagnostik*, Göttingen, Hogrefe, 1964, p. 19—70.

MITTENECKER, E., Subjektive Tests zur Messung der Persönlichkeit, in: R. HEISS (Ed.), *Handbuch der Psychologie, Bd. 6, Psychologische Diagnostik*, Göttingen, Hogrefe, 1964, p. 461—487.

NUSSBAUM, K., & MICHAUX, W. W., Response to humor in depression: a prediction and evaluation of patient change? *Psychiatric Quaterly*, 1963, *37*, 527—539.

NUSSBAUM, K., WITTIG, B. A., HANLON, T. E., & KURLAND, A. A., Intravenous nialamide in the treatment of depressed female patients, *Comprehensive Psychiatry*, 1963, *4*, 105—116.

PICHOT, P., Les aspects symptomatique des états dépressifs, *Schweizer Archiv für Neurologie, Neurochirurgie und Psychiatrie*, 1964, *94*, 392—410.

PICHOT, P., Überlegungen zur Faktorenanalyse des depressiven Syndroms, in: H. HIPPIUS & H. SELBACH (Ed.), 1969, p. 269–277. (a)

PICHOT, P., Überlegungen zur Quantifizierung depressiver Symptome, in: W. SCHULTE & W. MENDE (Ed.), *Melancholie in Forschung, Klinik und Behandlung*, Stuttgart, Thieme, 1969, p. 76–79. (b)

PICHOT, P., & LEMPERIERE, T., Analyse factorielle d'un questionaire d'autoévaluation des symptomes dépressifs, *Revue de Psychologie Applique*, 1964, *14*, 15—29.

PICHOT, P., PIROT, J., & CLYDE, D. J., Analyse de la symptomatologie depressive subjective, *Revue de Psychologie Applique*, 1966, *16*, 105—115.

PÖLDINGER, W., & BLASER, P., Experimentalpsychologische Untersuchungen bei psychisch Kranken, *Kurse ärztlicher Fortbildung*, 1968, *18*, 185—192.

PÖLDINGER, W., BLASER, P., & GEHRING, A., Anforderungen an quantifizierende Methoden in der Depressionsforschung, in: W. SCHULTE & W. MENDE (Ed.), *Melancholie in Forschung, Klinik und Behandlung*, Stuttgart, Thieme, 1969, p. 80–87. (a)

PÖLDINGER, W., BLASER, P., & GEHRING, A., Zur Quantifizierung psychopathologischer und somatischer Symptome bei depressiven Verstimmungszuständen, in: H. HIPPIUS & H. SELBACH (Ed.), 1969, p. 291–304. (b)

PÖLDINGER, W., & GEHRING, A., Vegetative Untersuchungen im Rahmen der Depressionsdiagnostik, *Kurse ärztlicher Fortbildung*, 1968, *18*, 190—192.
REITER, I., Eine Untersuchung der Gedächtnisfunktion bei Depressiven, unveröff. Diss., Salzburg, 1972.
SCHMIDT, H. O., & FONDA, C. P., The reliability of psychiatric diagnosis: a new look, *Journal of Abnormal and Social Psychology*, 1956, *52*, 262—267.
SCHWAB, J. J., BIALOW, M. R., CLEMMONS, R. S., & HOLZER, C. E., Hamilton rating scale for depression with medical in-patients, *British Journal of Psychiatry*, 1967, *113*, 83–88. (a)
SCHWAB, J. J., BIALOW, M. R., & HOLZER, C. E., A comparison of two rating scales for depression, *Journal of Clinical Psychology*, 1967, *23*, 94–96. (b)
SELG, H., & BAUER, W., *Forschungsmethoden der Psychologie*, Stuttgart, Kohlhammer, 1971.
SPITZER, R. L., FLEISS, J. L., ENDICOTT, J., & COHEN, J., The Mental Status Schedule, *Archives of General Psychiatry*, 1967, *16*, 479—493.
VINAR, O., Auswertung der seinen Zustand betreffenden Angaben des Kranken im Verlauf der Psychopharmakotherapie, *Arzneimittel-Forschung*, 1966, *16*, 285—287.
VINAR, O., & GROF, P., Die depressive Symptomatologie im Lichte des Beck'schen Fragebogens, in: H. HIPPIUS & H. SELBACH (Ed.), 1969, p. 318.
WARD, C. H., BECK, A. T., MENDELSON, M., MOCK, J. E., & ERBAUGH, J. K., The psychiatric nomenclature: reasons for diagnostic disagreement, *Archives of General Psychiatry*, 1962, *7*, 198—205.
WECHSLER, H., GROSSER, G. H., & BUSFIELD, B. L., The depression rating scale, *Archives of General Psychiatry*, 1963, *9*, 334—343.
WECKOWICZ, T. E., MUIR, W., CROPLEY, A. J., A factor analysis of the Beck inventory of depression, *Journal of Consulting Psychology*, 1967, *31*, 23—28.
ZUNG, W. W. K., RICHARDS, C. B., & SHORT, M. J., Self-rating depression scale in an outpatient clinic, *Archives of General Psychiatry*, 1965, *13*, 508.

(Eingegangen am 22. August 1973)

Anschrift des Verfassers: Dr. H. Lukesch, Universität Konstanz, Fachbereich Erziehungswissenschaft, 775 Konstanz, Postfach 733

R. Scheller und H. Sittauer

Analytische Diskrimination dreier hirnorganischer Gruppen anhand von HAWIE-Daten[1])

(Aus dem Fachbereich I – Abteilung Psychologie – der Universität Trier-Kaiserslautern in Trier)

1. *Problemstellung*

Um die Brauchbarkeit des HAWIE als differentialdiagnostisches Instrument im Bereich hirnorganischer Erkrankungen zu überprüfen, varianzanalysierte *Scheller* (1973) die HAWIE-Ergebnisse von 80 Hirnarteriosklerotikern, 80 Hirnatrophikern und 80 Hirntraumatikern. Die ermutigenden Ergebnisse dieser Untersuchung werden in der vorliegenden Arbeit durch eine diskriminanzanalytische Bearbeitung der HAWIE-Daten mit dem Ziel ergänzt, ein zusätzliches

[1]) Die Verfasser danken Herrn Dipl.-Psych. F. Heil für die kritische Durchsicht des Manuskripts.

diagnostisches Hilfsmittel für die in der Praxis so bedeutsame Einzelfallaussage bereitzustellen.

Obwohl *Dahl* (1968) die Verwendung der Driskriminanzanalyse als Auswertungsverfahren für klinisches Testmaterial mit Skepsis betrachtet, belegen neuere Untersuchungen die Effizienz dieser Methode (vgl. z. B. *Jungebloed*, 1973; *Kornmann*, 1971; *Michaelis*, 1972 a, 1972 b; *Trappl* et al., 1971). Sie verdeutlichen ebenso ihre zunehmende Relevanz für medizinisch-psychologische Diagnostik und Prognostik.

2. *Methodologische Vorbemerkungen*

Michaelis (1972, p. 1) bezeichnet als Diskriminanzanalyse „eine Gruppe multivariater statistischer Verfahren, die es gestatten, ein Individuum aufgrund einer Reihe von beobachteten Merkmalen einer bestimmten Gruppe gleichartiger Individuen zuzuordnen. Beispiele hierfür sind die . . . Zuordnung eines Patienten zu einer Gruppe von Patienten mit einer bestimmten Krankheit (Diagnostik) oder innerhalb derselben Krankheit zu einer bestimmten Verlaufsform (Prognostik)". – Eine auf diskriminanzanalytischen Überlegungen basierende Einzelfallaussage bedarf der Formulierung einer geeigneten Zuordnungsvorschrift. Sie wurde in der vorliegenden Untersuchung mit der von *Faber & Nollau* (1969) erweiterten linearen Diskriminanzanalyse (*Fisher*, 1936; vgl. hierzu auch *Heil*, 1968; *Hope*, 1968; *Kendall & Stuart*, 1966) ermittelt[2]). Die Bestimmung der effizientesten Untermenge von Subtests erfolgte mit einem gleitenden Variablensatz, wobei sukzessiv eine bis maximal drei Subtestvariablen aus der Datenmatrix X entfernt wurden. Als Beurteilungskriterien für die hierbei erzielten Ergebnisse dienten die Güte der statistischen Sicherung und die Anzahl der richtigen Zuordnungen in die jeweilige Kriteriumsgruppe. Die Überprüfung der Distanzen im gesamten Diskriminanzraum und der Unterschiede zwischen jeweils zwei Hirnorganikergruppen auf Signifikanz wurde mit der von *Rao* (1952) entwickelten Approximation von „*Wilks* Lambda" an die F-Prüfverteilung vorgenommen; sie gilt der von *Bartlett* (1937) angegebenen Transformation in die Chi-Quadrat-Verteilung als überlegen (vgl. *Cooley & Lohnes*, 1962).

3. *Erläuterung von Diskriminanzfunktion und Zuordnungsvorschrift*

Aus der Matrix $[x_{ij}]$ der $i = 1, 2, \ldots, m = 10$ Subtestvariablen können bei $I = 3$ Gruppen zwei Diskriminanzfunktionen $d_{1.}$ und $d_{2.}$ errechnet werden. Diese beiden Funktionen $d_{k.}$ bestehen als Linearkombinationen aus je einem Vektor $\beta_{k.}$, dessen m Elemente β_{ki} als Gewichte der Subtestvariablen $x_{i.}$ in der k-ten Diskriminanzfunktion interpretiert werden können und die der Bedingung der optimalen Diskrimination genügen. Durch die Anwendung der Gewichte β_{ki} auf die Wertpunktmatrix x_{ij} erhält man für den j-ten Probanden die beiden Diskriminanzwerte d_{1j} und d_{2j} nach der Formel

$$(1) \quad d_{kj} = \sum_{i=1}^{m} \beta_{ki} \cdot x_{ij}$$

Die beiden Diskriminanzwertevektoren $d_{1.}$ und $d_{2.}$ sind unkorreliert und dienen der von *Faber & Nollau* (1969) entwickelten und bewiesenen Zuordnungsvorschrift als Basisinformation. Hierbei wird der Abstand s_{gj} vom Zentroid Z_g mit dem Abstand s_{lj} von allen übrigen Gruppenzentroiden für jeweils einen Probanden paarweise verglichen. Die Versuchsperson j gehört der Gruppe g an, wenn ihr Abstand s_{gj} vom Zentroid Z_g kleiner ist als die Abstände s_{lj} von den anderen Zentroiden Z_l, bzw., wenn alle Abstandsdifferenzen ${}_{gl}a_j$ größer Null sind. Der Proband j wird also der Gruppe g zugeordnet, wenn

$$(2) \quad {}_{gl}a_j = s_{lj} - s_{gj} > 0 \quad \text{für alle } l \neq g$$

[2]) Sämtliche Berechnungen wurden auf der TR 440 im Rechenzentrum der Universität Trier-Kaiserslautern durchgeführt.