Moderatoren, differentielle Diagnostizierbarkeit und Vorhersage

Reinhold Jäger

Forschungsbricht Nr. 1

Mai, 1977

Anschrift des Verfassers: Dr. Reinhold Jäger,
Otto-Selz-Institut für Psychologie und
Erziehungswissenschaft
Schloß, EO, 6800 Mannheim 1

I Begriffsbestimmung

Die entscheidende Frage der differentiellen Psychologie ist "Wo bestehen Unterschiede zwischen Probanden bzw. Probandengruppen und wie sind diese zu interpretieren?".

Die empirischen Vorgehensweisen sind deshalb darauf ausgerichtet, mit Hilfe von Prädiktoren Unterschiede zwischen diversen Probanden bzw. Probandengruppen zu identifizieren.

So wird u.a. in den verschiedensten Lehrbüchern der differentiellen - und Persönlichkeitspsychologie berichtet, daß psychologische Tests in verschiedenen Teilpopulationen verschiedene Validitäten besitzen (s. ANASTASI, 1968; HERMANN, 1969).

Unabhängig davon, welche Art der Validität in Frage steht und wie diese bestimmt wird, ist festzustellen, daß diese unter Einbeziehung von Drittvariablen, sog. Moderatoren, nicht konstant bleibt. Dabei existiert bezüglich des Begriffes "Moderator" ein Konsensus; "unter Moderatorvariablen verstehen wir jene Variablen oder Variablensysteme, die bestehende Abhängigkeiten zwischen wiederum Variablen oder Variablensystemen in der Größe und/oder Richtung verändern" (JÄGER, 1974 a, S. 1oo).

Übertragen auf einen korrelationsstatistischen Ansatz bedeutet dies, daß bei der Aufteilung einer Gesamtstichprobe in verschiedene Teilstichproben korrelative Änderungen eintreten, wie sie beispielsweise in Abb. 1 dargestellt werden. Die Aufteilung der Gesamtstichprobe in Teilstichproben erfolgt hierbei durch Verwendung von Moderatoren, so z.B. Geschlecht, Alter, sozio-ökonomischer Status etc.

Eine eindeutige Aussage in Richtung auf eine erhöhte Korrelation, wie sie beispielsweise von WESTMEYER (1972) unterstellt wird, kann allerdings nicht gemacht werden, noch wird eindeutig in Richtung auf eine kausale Interpretation Bezug genommen.

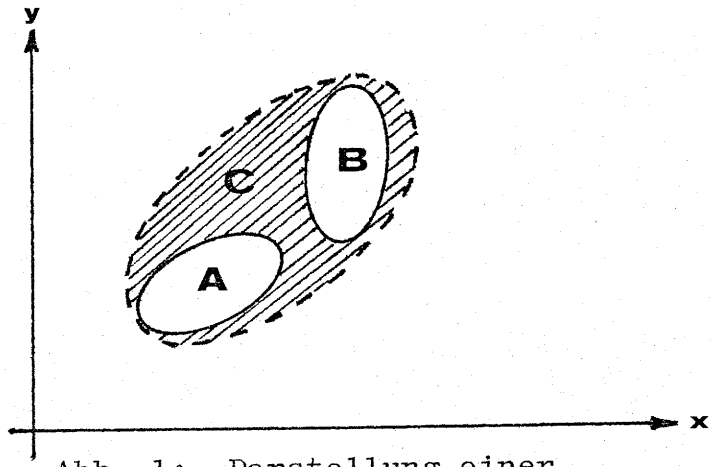

Abb. 1: Darstellung einer Moderatorwirkung (A+B≠C)

MICHEL & ISELER (1968) schlagen als Oberbegriff zu den von anderen Autoren verwendeten Begriffen, wie differentielle Validität, differentielle Vorhersagbarkeit etc. (s. BRODGEN, 1951; CRONBACH, 1960; ANASTASI, 1961; GHISELLI, 1963; JANKE, 1964; MICHEL, 1964; HÖRMANN, 1964; LIENERT, 1969) den Begriff differentielle Diagnostizierbarkeit vor.

Der gemeinsame Bezug auf das Gütekriterium der Validität erscheint dabei sehr einseitig, da die Güte eines Testverfahrens oder einer anderen Datenerhebungsmethode nur durch eine Vielzahl von verschiedenen Kriterien abgeschätzt werden kann. So spricht GHISELLI (1973) auch von differentiellen Effekten im Zusammenhang mit der Reliabilität. Auch eine Eingrenzung der Validität auf den Bereich der (Personen-) Population ist wenig sinnvoll.

Im Rahmen einer Präzisierung des generellen Validitätsbegriffes hat WESTMEYER (1972) festgestellt: "Validität läßt sich darstellen als sechsstelliger Faktor » val « , der mit den Subjektvariablen t,c,p,o,v,z den zusammengesetzten Term » val « (t,c,p,u,v,z) wiedergibt (S. 53).

Der hier genannte Validitätsterm ist dann so zu interpretieren, daß der dem Term zugeordnete quantitative Wert W die Höhe der Validität eines "Prädiktors t in Bezug auf das Kriterium c bei Anwendung auf die Personenklasse p und der Umgebungsbedingung u durch den Versuchsleitertyp v während eines Zeitbereiches z" darstellt (WESTMEYER, 1972, S. 53).

Auch diese Präzisierung ist nur beschränkt verallgemeinerbar, da im Prinzip davon ausgegangen werden muß, daß die im Test verwendeten Items eine repräsentative Stichprobe darstellen, von der man auf das entsprechende Verhalten außerhalb der Testsituation bzw. das ganze Spektrum von Verhaltensweisen der gleichen Art (Verhaltensuniversum) schließen kann.

Unterstellt man einen solchen Repräsentationsschluß, so sind die Subjektvariablen t, c, p, u, v, z ihrerseits entweder als repräsentative Stichprobe aus der Population von t, c ..., v, z zu verstehen oder aber als konkrete Realisierung, d.h. daß z.B. t ein Elementarereignis aus der Gesamtmenge aller Tests (T) darstellt, daß ebenso c eine konkrete Realisierung aus der Menge C ist, etc.

Ein Repräsentationsschluß bietet sich in diesem Kontext weniger an, da er zu sinnlosen diagnostischen Schlußfolgerungen führen würde, wogegen eine konkrete Realisierung, wie sie oben angesprochen wird, entsprechende praktische Relevanz besitzt.

Für die Formulierung eines Reliabiltitätsterms läßt sich analog zu dem o.a. Validitätsterm eine Aussage machen. Reliabilität läßt sich in Anlehnung an die von WESTMEYER (1972) gewählte Terminologie folgendermaßen auffassen: "Reliabilität läßt sich darstellen als sechsstelliger Funktor »rel« , der mit den Subjektvariablen t,p,u,v,z,q den zusammengesetzten Term »rel« (t,p,u,v,z,q) bildet".

In Ergänzung zu den vorher erläuterten Subjektvariablen t, p,u,v,z stellt nun q diejenige Zeitspanne dar, innerhalb der t (im Falle der Retestreliabilität) wiederholt wird. Die im Validitätsterm angeführte Subjektvariable c entfällt, da im Falle der (Retest-) Reliabilität c zu t wird (die Reliabilität ist in diesem Zusammenhang die Validität von t mit sich selbst).

Unter Heranziehung der vorangegangenen erweiterten Darstellung des Validitäts- und Reliabilitätsbegriffes läßt sich "differentielle Diagnostizierbarkeit" als zusammenfassender Terminus all jener Gütekriterien auffassen, die bezüglich der Subjektvariablen t,c,p,u,v, q,z keine Konstanz besitzen, d.h. also, daß die Validität von t auf dem Hintergrund einer spezifischen Personenklasse p_1 anders ausfallen kann als bei p_2, daß die Umgebungsbedingung u_1 kausal für die in quantitativer Hinsicht veränderte Reliabilität gegenüber u_2 verantwortlich gemacht werden kann etc.

II. Probleme der Vorhersage

Im Raume der Sozialwissenschaften ist es das erklärte Ziel, nicht nur eine Diagnose über den jetzigen Zustand von Personen bzw. Personengruppen durchzuführen, sondern womöglich Anhaltspunkte dafür zu gewinnen, ob künftiges Verhalten dieser Personen bzw. der Personengruppen prognostiziert werden kann. Schon in den frühen Jahren der psychologischen Forschung, in denen eine Auseinandersetzung mit lernpsychologischen Gesetzmäßigkeiten erfolgte, versuchte THORNDIKE (1913) mit dem sog. Effektengesetz auf der Basis eines in der Ist-Situation gegebenen Endzustandes das Auftreten des gleichen Verhaltens in der Zukunft vorherzusagen.

Auch die momentanen Bestrebungen in der Hochschulforschung sind darauf ausgerichtet, Verhalten - hier Leistungsverhalten im Examen - zu prognostizieren, indem fachspezifische Hochschuleingangstests Verwendung finden,

mit deren Hilfe die Geeignetsten für ein Studienfach unter den bereits durch das Abitur Qualifizierten gefunden werden (MICHEL & JÄGER, 1976).

Unabhängig von der jeweiligen Teildisziplin, bei der Probleme der Prognose auftreten, läßt sich in der jetzigen Situation der psychologischen Forschung feststellen, daß ein Dilemma darin existiert, daß die numerische Höhe derjenigen Koeffizienten, die eine quantitative Bestimmung der Güte der Prognose erlauben, bislang statistisch unbefriedigend ist.

Zur Lösung dieses Dilemmas bieten sich eine Reihe von methodischen Anhaltspunkten an, von denen drei in getrennter Diskussion hier kurz erörtert werden sollen (s.a. JÄGER, 1975).

1) Erhöhung der Anzahl der Prädiktoren

Ein sehr gebräuchliches Verfahren, um beispielsweise so vermutlich komplexe Kriterien wie allgemeine Schulleistung, Schulreife, Hochschulerfolg, Joberfolg, etc. statistisch zu prognostizieren, besteht darin, zusätzlich zu gängigen Prädiktoren weitere hinzuzufügen.

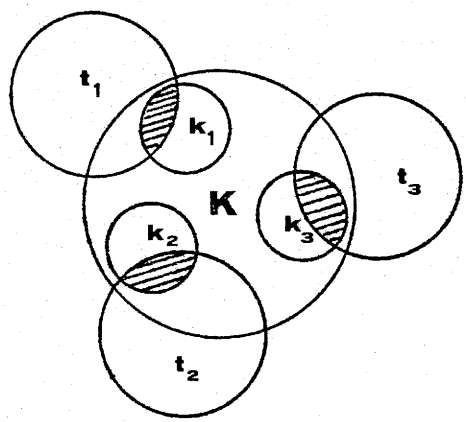

Abb. 2: Zusammenhang zwischen $t_1 \ldots, t_n$ und $k_1 \ldots\ldots, k_n$

Abb. 2 macht deutlich, daß die korrelative Übereinstimmung zwischen den Prädiktoren t_1 ....t_n und den Teilkriterien k_1 .....k_n umso höher ist, je eher ein Teilkriterium K_i durch einen Prädiktor T_i prognostiziert werden kann.

Sehr differenziert wurde dieser Sachverhalt in einer unlängst erschienenen Untersuchung von KRAPP (1975) im Rahmen einer Untersuchung über die multiple Prognose bei Schulanfängern demonstriert, wobei die dabei vorgefundene multiple Korrelation weit höher lag, als die bislang im gleichen Rahmen berichteten Koeffizienten.

Bei der Verwendung solcher statistischen Prognosemodelle wird in den meisten Fällen von linearen Beziehungen, in sehr wenigen Untersuchungen von nicht-linearen ausgegangen.

Ein Nachteil dieser Modelle tritt dabei zu oft in den Hintergrund, der aber hier einer Diskussion bedarf.

Es muß die Frage gestellt werden, ob die Voraussetzungen, die zur zufallskritischen Absicherung der Koeffizienten notwendig sind, überhaupt gegeben sind. Besteht schon bei einzelnen Prädiktoren die Schwierigkeit, eine Normalverteilung der Daten zu gewährleisten, so gelingt dies bei multiplen Ansätzen sowohl auf der Seite der Prädiktoren und Kriterien ungemein schwerer. In einer solchen mißlichen Situation muß deshalb ernsthaft die Frage diskutiert werden, ob eine sinnvolle Signifikanztestung unter Wahrung der Teststärke überhaupt möglich ist.

Überdies treten Probleme auf, die eine statistische Verbesserung der Zusammenhänge zwischen Prädiktor(en) und Zielvariable(n) zwar bewirken können, bei denen aber in Frage steht, ob die Verbesserungen nicht artifiziell bedingt sind.

Bei einer kritischen Analyse von multivariat angelegten Untersuchungen fällt auf, daß sehr oft Prädiktoren verwendet werden, die voneinander stochastisch abhängig sind. Solche stochastischen Abhängigkeiten (die dann zu entsprechenden hohen korrelativen Übereinstimmungen führen) liegen beispielsweise dann vor, wenn Variablen wie Alter und Entwicklungsstand, sozioökonomischer Status und Beruf der Eltern, Leistungen im Schulreifetest und Intelligenztest etc. verwendet werden. In diesen Fällen muß von vornherein der Anteil der gemeinsamen Varianz zwischen eben diesen Variablen groß sein, was graphisch durch eine entsprechende Überlappung der Varianzräume verdeutlicht wird (s. Abb. 3).

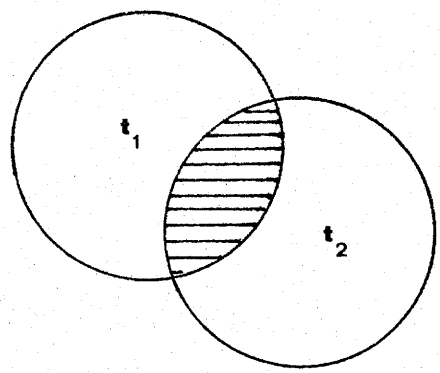

Abb. 3: Veranschaulichung der Determination als $t_1 \cap t_2$

Die Überschneidungsfläche $t_1 \cap t_2$ ist dann mit der durch $r^2_{t_1 t_2}$ ausgedrückten Determination identisch.

Ein Spezialfall im Rahmen dieser Überlegungen wird durch die SAUNDER'sche Regressionsgleichung, die sog. moderierte Regression, wiedergegeben (SAUNDERS, 1956). Bezeichnet man den kontinuierlich verlaufenden Moderator mit M, die ebenfalls zur Schätzung einer Variablen Y

verwendeten unabhängigen Variablen mit X, so resultiert aus der ursprünglich vorhandenen linearen Regression zwischen y und x

$$\hat{Y} = bX + a, \qquad (1)$$

die moderierte Regression:

$$\hat{Y} = c + eM + dX + fMX \qquad (2)$$

Das Kreuzprodukt MX wird dabei genauso wie X und M als Prädiktor aufgefasst. Ist aber ein Moderator M derart gestaltet, daß er stochastisch abhängig vom Prädiktor X ist, so wird im Prinzip eine Variable zweimal als Prädiktor verwendet (s. Gleichung (2)).

Stochastische Abhängigkeiten existieren auch in denjenigen Fällen, in denen Polynome höherer Ordnung verwendet werden, um eine bessere Anpassung an Regressionsverläufe zu erreichen. Dabei liegt ein Spezialfall der moderierten Regression vor.

Setzt man nämlich in Gleichung (2) anstelle des Moderators M die Zufallsvariable X, so resultiert daraus:

$$\hat{Y} = c + X(e + d) + fX^2 \qquad (3)$$

Wenn nun (e+d) durch h ersetzt wird, so wird aus (3):

$$\hat{Y} = c + hX + fX^2 \qquad (4)$$

also ein Polynom zweiter Ordnung. Damit ist die Beziehung zwischen einem Polynom zweiten Grades und der moderierten Regression hergestellt; man kann nun in diesen Zusammenhang sagen, daß ein Polynom zweiten Grades mit der an sich selbst moderierten Regression identisch ist, X ist der eigene Moderator. Faßt man

zunächst einmal die in (4) dargestellten Glieder X und X^2 als unabhängige Variablen auf, so erfüllt die Regression von X nach Y Minimumseigenschaften. Gleichzeitig existiert aber zwischen X und X^2 eine entsprechende stochastische Abhängigkeit, so wie sie auch in dem Kreuzprodukt der Gleichung (2) durch MX gegeben ist.

Bei der Überprüfung der korrelativen Beziehungen zwischen den stochastisch abhängigen Größen stellt sich heraus, daß sehr oft Korrelationen der Größenordnung $r_{MX} \geq .90$ oder $r_{X^2} \geq .95$ existieren. Diese Daten wurden anhand eigener empirischer Untersuchungen gewonnen. Von daher ist es einleuchtend, daß die mit Hilfe von multiplen Korrelationen (R^2) bestimmten Zusammenhänge zwischen Prädiktoren und Kriterium u.U. überschätzt sein müssen.

Die mathematische Ableitung wird bei JÄGER (1976) mitgeteilt.

Für den Fall eines Kriteriums Y und eines Prädiktors X und einer Variablen M ist die multiple Korrelation R zwischen Y und X und M:

$$R_{Y.XM} = \sqrt{\frac{r^2_{YX} + r^2_{YM} - 2r_{YX} \cdot r_{YM} \cdot r_{XM}}{1 - r^2_{XM}}} \qquad (5)$$

An anderer Stelle (CONGER & JACKSON, 1972) wird nachgewiesen, daß ein partialkorrelationsstatistischer Ansatz am besten geeignet ist, den Zusammenhang zwischen Variablen darzustellen, da er als unterste Grenze der Übereinstimmung zwischen Y und X angesehen werden kann (s. JÄGER, 1976).

Die Partialkorrelation zwischen Y und X unter Ausschaltung von M ist dann:

$$r_{Y.(XM)} = \frac{r_{YX} - r_{YM} \cdot r_{XM}}{\sqrt{(1 - r_{XM^2})(1 - r_{YM^2})}} \qquad (6)$$

Unter der Voraussetzung $r_{YM} = 0$ gilt:

$$R_{Y.XM} = r_{Y.(XM)}$$

d.h., es ist hier der ideale Suppressionsfall vorgegeben, bei allen anderen Fällen ($r_{YM} \neq 0$) existiert eine Differenz zwischen multipler Korrelation und Partialkorrelation. Umgekehrt läßt sich daraus ableiten, daß die Partialkorrelation für den Fall $r_{XM} = 0$, d.h. daß die Unkorreliertheit zwischen den als Prädiktoren verwendeten Variablen existiert, den wahren Zusammenhang unterschätzt, in diesem Zusammenhang ist

$$r_{Y.(XM)} < R_{Y.XM}$$

da

$$\frac{r_{YX}}{\sqrt{1-r^2_{YM}}} < \sqrt{r^2_{YX} + r_{YM}^2} \qquad (7)$$

Der Zusammenhang in diesem Rahmen wird durch eine entsprechende Graphik verdeutlicht (s. Abb. 4), bei der eine Konstellation von Daten so gewählt wurde, daß sich die Korrelationen r_{YX} und r_{YM} jeweils zu 1 ergänzen und von $r_{YX} = .1o$ bzw. $r_{YM} = .9o$ beginnend gegeneinander konvergieren.

Aus den Waagerechten und dem Schnittpunkt mit der gestrichelten Kurve in der Abbildung, beginnend mit der Geraden G, ist die jeweilige Unterschätzung abzulesen. Die Differenz zwischen

$$r_{Y.(XM)} \quad \text{und } R_{Y.XM}$$

ist dabei umso größer, je höher die Korrelation r_{YM} ausfällt; dieser Sachverhalt ist daraus zu erklären, daß das Korrekturglied (s.(7))

$$\sqrt{1 - r^2_{YM}}$$

die Höhe von r_{YX} entsprechend relativiert.

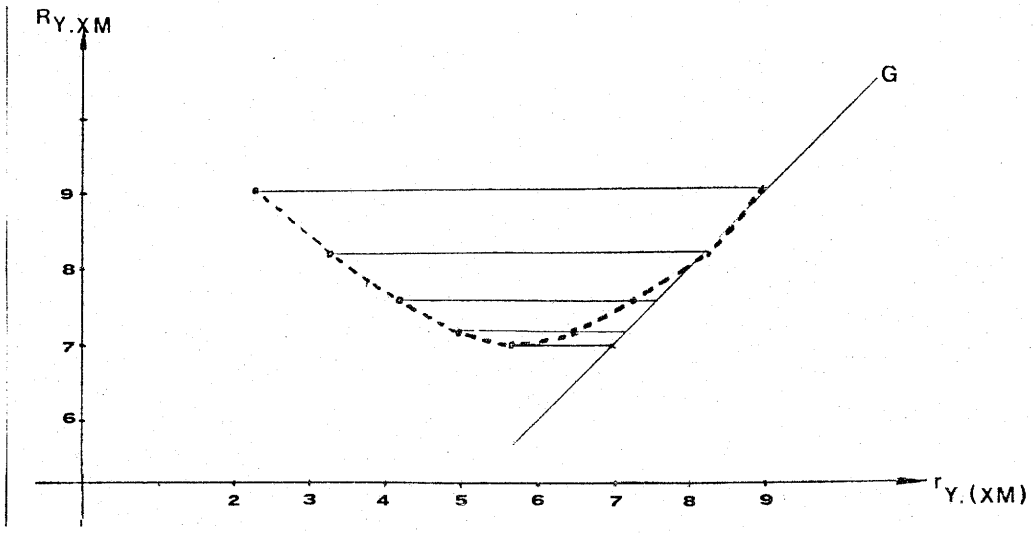

Abb. 4: Zusammenhang zwischen multipler- und Partialkorrelation bei $r_{XM} = 0$

Eine Entsprechung (wenn auch nicht eine numerisch gleiche) liegt im umgekehrten Falle vor ($r_{XM} \neq 0$). Dabei wird bei Korrelationen zwischen den als abhängig bezeichneten Variablen von sog. Multicollinearität gesprochen (s.a. JÄGER, 1976).

Stellt man auf dem Hintergrund der vorherigen Aussagen die Frage, ob nun zur Erhöhung der Validität nach neuen Variablen gesucht werden soll, so muß die Frage auf der Grundlage des bisher Gesagten differenzierter beantwortet werden:

1. Es müssen auf jeden Fall solche Variablen Verwendung finden, die mit dem in Frage stehenden Kriterium korrelieren.

2. Sie sollten möglichst gering mit den anderen Prädiktoren korrelieren und

3. zur Bestimmung des Zusammenhanges sollte im mehr als dreidimensionalen Rahmen ein partialkorrelationsstatistischer Ansatz verwendet werden; eventuell ein multipler Ansatz. Allerdings muß bei Zusammenhängen der vorher beschriebenen Art ($r_{XM} = 0$; $r_{YM} > .20$) zum Teil mit gravierenden Unterschätzungen gerechnet werden (s.o.).

4. Sofern auf der Basis von Moderatoren Teilpopulationen gebildet werden, ist besonders auf die Punkte 1 und 2 zu achten (s.o.), da vorab nie bestimmt werden kann, in welcher Weise Änderungen eintreten.

2. Verwendung von Suppressoren

Im Rahmen einer klassischen Betrachtung von sog. Suppressoren (S) geht man davon aus, daß es sich bei diesen (auch Suppressorvariablen bzw. Suppressortest genannt) um solche Variablen oder Tests handelt, die mit dem Kriterium niedrig und mit den anderen Variablen (oder Tests), die zur Prädiktion verwendet werden, hoch korrelieren (LIENERT, 1969). Graphisch läßt sich dieser Sachverhalt folgendermaßen veranschaulichen (s. Abb. 5):

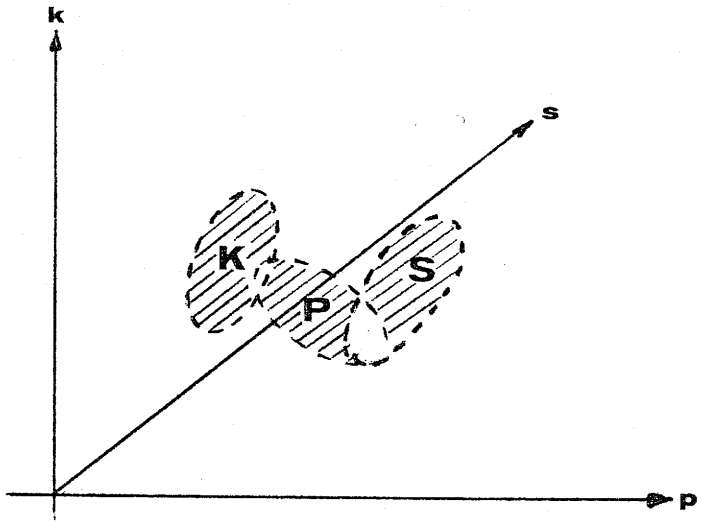

Abb. 5 : Veranschaulichung einer Suppressorwirkung

Da die Korrelation $r_{KS} = 0$, wie sie in Abb. 4 dargestellt wird, und gleichzeitig r_{PS} sehr hoch ist, muß es sich im Falle einer Prädiktion von P und S nach K um ein Artefakt der Methode handeln, wenn die Korrelation $R_{K.PS}$ höher ist als r_{KP}.

Offensichtlich liegt hier eine Verletzung des regressionsanalytischen Modelles vor, da nur solche Variablen zur Vorhersage verwendet werden sollen, die untereinander unabhängig sind; d.h. hier ist sog. Multicollinearität der Daten gegeben (s.o.), was dazu führt, daß eine Überschätzung der wahren Validität zwischen Kriterium und Prädiktor erfolgt. Im statistischen Sinne liegt hier eine inkonsistente Schätzung vor(s. JÄGER, 1976).

So gesehen muß die Verwendung von Suppressoren wesentlich kritischer betrachtet werden als dies in verschiedenen Lehrbüchern der Testtheorie dargestellt wird.

3. Verwendung von Moderatoren

Weiter vorne wurde die Wirkungsweise von Moderatoren dargestellt. Wenn nun auf der Grundlage von Moderatoren Subpopulationen gebildet werden, so entstehen hieraus eine Reihe von Problemen, die zum Teil an anderer Stelle von uns dargestellt wurden; deshalb soll hier besonderer Wert auf die angewandte psychologische Diagnostik gelegt werden.

Es gehört zu den Eigenarten der Testkonstrukteure, psychologische Testverfahren eher auf dem Hintergrund der zur Verfügung stehenden Gesamtstichprobe zu analysieren als auch in Teilpopulationen entsprechende Untersuchungen über die Gütekriterien von Items durchzuführen. Nur vereinzelt finden sich solche Hinweise (s. FAHRENBERG et al., 1973; JÄGER & JUNDT, 1973).

Besonders aus den Mitteilungen von FAHRENBERG et al. (1973, S. 27 f) geht hervor, daß es gerade im Zusammenhang mit Fragebogen sehr wichtig ist, Kreuzvalidierungen auch auf Itemebene vorzunehmen, will man sich nicht dem Vorwurf aussetzen, unter Umständen mit den gleichen Skalen in verschiedenen Teilpopulationen unterschiedliche Merkmale zu erfassen. Ein Weg dazu besteht in der Verwendung von Moderatoren, d.h. es werden die letztlich geplanten Normierungseinheiten dazu verwendet, Kreuzvalidierungen auf Itemebene durchzuführen.

Dieser Versuch wurde konsequent im Rahmen der Konstruktion eines neuen Fragebogens (BIV von JÄGER et al., 1976) durchgehalten. Konkret wurde die Frage gestellt, wie läßt sich eine Homogenisierung auf Skalenebene und eine populationsstabile Faktorenstruktur erreichen? Die Lösung wurde darin gesehen, nur solche Items beizubehalten, die über verschiedene Teilstichproben hinaus vergleichbare Trennschärfeindizes erbringen, wobei die Teilstichproben unter Zuhilfenahme von Moderatoren gebildet wurden.

Beispielhaft wird in Tab. 1 auf die Skala FAM des BIV Bezug genommen. Dabei zeigt sich, daß die Trennschärfen der Items als interne Validitätskriterien der Skala, und gleichfalls die Homogenitäten, die nach CRONBACH's α bestimmt wurde, über die verschiedenen Gruppen (Pat(ienten), No(rmale), A(lter)) hinweg wiederum homogen sind, was dafür spricht, daß die Skalen gleiches oder zumindest ein vergleichbares Merkmal erfassen.

Eine weitere Analyse der Daten zeigt, daß bei einer Faktorisierung (Hauptkomponentenmodell) auch über verschiedene Teilpopulationen hinweg die Faktorenstruktur auf Skalenebene erhalten bleibt. Tab. 2 gibt die entsprechenden Daten wieder. (s.a. JÄGER, 1977 b):

Item Nr.	Alle	Pat	NO	♂	♀	NO ♂	NO ♀	Pat.♂	Pat.♀	A ≤ 3o	A > 3o	Neuroti
8	.51	.54	.45	.49	.53	.42	.5o	.54	.54	.51	.51	.57
13	.44	.41	.49	.42	.47	.43	.57	.4o	.42	.48	.39	.63
15	.5o	.47	.54	.5o	.49	.57	.5o	.46	.48	.52	.47	.53
27	.42	.41	.44	.4o	.44	.45	.41	.37	.45	.53	.28	.43
29	.47	.43	.53	.44	.52	.56	.52	.35	.52	.45	.48	.56
31	.47	.47	.46	.43	.52	.4o	.57	.45	.49	.58	.39	.54
35	.61	.64	,56	.6o	.64	.53	.63	.64	.66	.62	.61	.7o
43	.52	.54	.51	.53	.53	.54	.46	.52	.56	.51	.51	.58
54	.61	.58	.66	.64	.57	.64	.66	.64	.53	.62	.59	.58
64	.62	.62	.62	.63	.61	.63	.63	.63	.61	.61	.6o	.62
7o	.57	.57	.56	.56	.59	.58	.53	.55	.61	.62	.5o	.64
77	.57	.56	.58	.56	.58	.6o	.55	.53	.59	.63	.52	.49
78	.55	.56	.53	.54	.56	.48	.61	.58	.54	.55	.53	.56
83	.46	.42	.51	.47	.44	.5o	.53	.45	.4o	.5o	.44	.41
88	.61	.59	.66	.64	.58	.61	.73	.65	.52	.65	.56	.64
α	.88	.87	.88	.87	.88	.88	.89	.87	.88	.89	.86	.89

Tab. 1 Trennschärfeindices und Homogenitätsindices der Skala FAM aus dem BIV von JÄGER et al. (1976)

	Patienten (N = 332)			Normale (N = 293)		
Skala	FA I	FA II	FA III	FA I	FA II	FA III
FAM	.22	.83	.01	.28	.73	.12
SOZAKT	.29	.20	.93	.45	.15	.66
PSYKON	.71	.15	-.02	.68	.29	.18
ICHSTIK	.73	.24	.20	.75	.16	.13
SOZLAG	.43	.39	.31	.50	.20	.31
ERZIEH	.22	.85	-.05	.24	.95	.14
N	.89	.20	.02	.88	.23	.00
E	.05	.09	-.51	-.01	-.07	-.66
% gemeinsame Varianz	62,1	21,9	16,0	69,5	16,5	14.00
% gesamte Varianz	44,2	18,3	14,0	49,0	14,7	13,3

Tab. 2 Faktorenanalyse auf der Basis der Skalen des BIV

Alle drei Faktoren bleiben soweit erhalten, daß Skalen, die substantielle Ladungen auf einem bestimmten Faktor der einen Population haben, ebenso substantiell auf dem gleichen Faktor der zugrundegelegten anderen Population laden. Dieses Ergebnis kann nur so interpretiert werden, daß zunächst einmal die faktorielle Validität des Inventars (BIV) über verschiedene Teilpopulationen erhalten bleibt. Damit ist ein entsprechendes Vorgehen im Zusammenhang mit einer Absicherung der Validität gerechtfertigt.

Gleichzeitig mit der Itemanalyse wurde die Homogenität der verschiedenen Skalen bestimmt. Hier erscheint eine Mindesthöhe angebracht, da auf dem Hintergrund der von LEHRL & KINZEL (1973) diskutierten Standardskala der kritischen Differenz ein Maß für die Brauchbarkeit von psychologischen Testverfahren abgeleitet werden kann.

Wir haben dies an anderer Stelle als "praktische Differenzierbarkeit" bezeichnet (JÄGER, 1974), womit ausgedrückt werden sollte, daß ein psychologischer Test wenigstens zwischen Probanden mit hoher und Probanden mit niedriger Merkmalsausprägung unterscheiden sollte, d.h. es sollten zumindest zwei kritische Differenzen existieren.
Diese sind dann gegeben, wenn die interne Zuverlässigkeit (r_{tt}) einer Skala oder eines Tests mindestens den Wert $\sqrt{.50}$ erreicht (s.a. JÄGER et al., 1976; JÄGER, 1977 a, 1977 b).

Diese Diskussion macht deutlich, daß:

1. Die Frage der differentiellen Diagnostizierbarkeit umgedreht werden kann, um entsprechende psychologische Skalen zu konstruieren. Damit ist eine wesentlich bessere Kontrolle der Validität möglich, als dies üblicherweise im Rahmen der Konstruktion von psychologischen Testverfahren geschieht;

2. ebenso ist die Zuverlässigkeit von Testverfahren, gemessen an Aspekten der Homogenität bzw. der internen Konsistenz, besser abzuschätzen.

Zusammengefasst könnte man sagen, daß es bei einer solchen Interpretation der differentiellen Diagnostizierbarkeit nun nicht mehr darum geht, zunächst eine höhere Prognose zu erhalten, sondern eher, eine entsprechende formale und metrische Kontrolle durchzuführen. Konsequenterweise wären nur solche Verfahren publizierbar und in der Praxis auch anwendbar, bei denen nachgewiesenermaßen bei allen Normierungseinheiten entsprechende Untersuchungen durchgeführt wurden.

Damit wird aber gleichzeitig folgendes ausgesagt: wenn durch die Bildung von Teilpopulationen, wie sie etwa im Rahmen verschiedener Normierungseinheiten vorliegen, korrelative Veränderungen zwischen einem Kriterium und

einem Prädiktor resultieren, so läßt sich dieser Sachverhalt zunächst einmal als unterschiedliche Validität eines Prädiktors interpretieren. Demnach besitzt der Prädiktor je nach Teilpopulation eine unterschiedliche Brauchbarkeit. Mit der Validität wird aber auch der inhaltliche Geltungsbereich eines Tests, eines Prädiktors, angesprochen, so daß aus unterschiedlichen Validitäten, die auf der Grundlage verschiedener Teilpopulationen zustandekommen, auch der Schluß gezogen werden muß, daß der Prädiktor ü.U. etwas Unterschiedliches erfasst. Gleichzeitig wird damit die Frage der Zuverlässigkeit der Messung angesprochen.

Ein absoluter Vergleich zwischen den verschiedenen Teilpopulationen ist damit nicht mehr gegeben. So ist vorstellbar, daß ein Test, von dem angenommen wird, er erfasse Merkmal A, in Wirklichkeit aber bei einer eingegrenzten Population das Merkmal B mißt. Entsprechende Beispiele sind hier aus dem Zusammenhang der verschiedenen Differenzierungshypothesen der Intelligenz ableitbar. Unabhängig von den methodischen Schwierigkeiten, die in diesem Rahmen bestehen, belegen die Differenzierungshypothesen, daß mit den bisherigen Konstruktionsansätzen keine gute Kontrolle der Validität einhergeht. Mit dem hier vorgeschlagenen Ansatz wird versucht, die entsprechende methodische Seite zu erbringen.

III Diskussion

In der vorangegangenen Darstellung wurde die Frage erörtert, welche Möglichkeiten existieren, um eine Erhöhung der Validität zwischen Kriterium und Prädiktor(en) zu erreichen. Dabei wurden drei mögliche Zugangsweisen erörtert.

Während bei zwei diskutierten Methoden (Erhöhung der Anzahl der Prädiktoren bzw. Verwendung von Suppressoren) im wesentlichen statistische Aspekte im Vordergrund standen, wobei

herauskristallisiert wurde, daß im Regelfalle auf Grund von Modellverletzungen die Validitäten überhöht ausfallen, ging es im Zusammenhang mit der Verwendung von Moderatoren um die Frage, wie die Validität und Reliabilität auch über verschiedene Teilstichproben gewährleistet werden kann.

Dieser letzte Aspekt ist vor allem auf dem Hintergrund der differentiellen Diagnostizierbarkeit zu sehen. Von der vorab gegebenen Information kann abgeleitet werden, daß kein Testergebnis unabhängig von einer Reihe von Funktoren gesehen werden kann. Deshalb ist es wichtig zu wissen, ob die Ergebnisse eines Tests unabhängig davon, ob beispielsweise der Test in einem Selektionszusammenhang bzw. in einer Beratungssituation angewendet wird, in gleicher inhaltlicher Hinsicht interpretiert werden können. Gleiche Überlegungen sind auf verschiedene Stimmungslagen, verschiedene Instruktionen etc. übertragbar. Von daher muß gefordert werden, daß ein Testverfahren nur bei solchen Populationen eingesetzt wird (und unter den entsprechenden Bedingungen), bei denen nachgewiesenermaßen Validität und Reliabilität statistisch gesichert sind. So muß festgestellt werden, daß bei einer internen Zuverlässigkeit eines Tests (im Sinne der Homogenität oder Interitemkonsistenz) von $r_{tt} < .7o$ keine diagnostisch sinnvolle Aussage mehr über interindividuelle Unterschiede gemacht werden kann. Anders ausgedrückt: die Fehlervarianz zu Lasten des Tests ist so groß, daß nicht einmal Extremwerte auf der entsprechenden Skala statistisch bedeutsam zu unterscheiden sind.

Was kann nun für die praktische psychologische Diagnostik aus der gesamten Erörterung herauskristallisiert werden?

1. Eine Erhöhung der Anzahl von Prädiktoren ist nur dann angezeigt, wenn die Korrelationen zwischen den Prädiktoren gering sind; möglicherweise läßt sich durch die Verrechnung der Daten zu Faktorenscores entsprechendes gewährleisten. Erste Hinweise hierzu finden sich in

der Literatur (AMELANG, 1975). Bei stochastisch abhängigen Variablen und relativ hohen Korrelationen der unabhängigen Variablen untereinander ist die Validität überhöht.

2. Von einer Verwendung von Suppressorvariablen bzw. Suppressortests ist abzuraten, da mit Hilfe solcher Variablen die durchgeführten Schätzungen inkonsistent werden. Bei Kreuzvalidierung ist mit entsprechend niedrigen Koeffizienten zu rechnen; die Ergebnisse sind nicht mehr wiederholbar, und die errechneten Validitäten sind überschätzt.

3. Bei einer Verwendung von Moderatoren ist mit erheblichen Auswirkungen auf die Gütekriterien zu rechnen. Die Validität kann nicht mehr entsprechend kontrolliert werden, und die Reliabilität sinkt u.U. derart ab, daß die Fehlervarianz zu Lasten des Testinstrumentes eine gesicherte Aussage über individuelle Merkmalsausprägungen nicht mehr zuläßt. Die hier vorgeschlagene methodische Restriktion erlaubt eine entsprechende Kontrolle, wenn auch die Anforderungen an den Itempool, bzw. die Bandbreite des diagnostischen Einsatzes eines Tests, erheblich eingeschränkt wird.

Deshalb muß zwangsläufig überprüft werden, ob die bisher publizierten Verfahren in dieser Hinsicht den entsprechenden statistischen Restriktionen gerecht werden, da nur dann auch nachfolgende diagnostische Schlußfolgerungen entsprechend gesichert sind.

Literatur

AMELANG, M.	1975	Validierung von Anforderungsprofilen für das Studium der Medizin, Zahnmedizin, Pharmazie und Psychologie. Vorläufiger Forschungsbericht (unveröffentlicht), Hamburg
ANASTASI, A.	1961[2]	Psychological Testing. N.Y.
ANASTASI, A.	1968[2]	Differential Psychology, N.Y.
BRODGEN, H.E.	1951	Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. Educ. Psych. Measmt. 11, 173-196
CONGER, A.J. & JACKSON, D.N.	1972	Suppressor Variables, Prediction and the Interpretation of Psychological Relationsship. Educ. Psycho. Measmnt. 12, 579-599
CRONBACH, L.J.	196o	Essentials of Psychological Testing. N.Y.
FAHRENBERG, J.; SELG, H. & HAMPEL, R.	1973[2]	Das Freiburger Persönlichkeitsinventar. FPI, Göttingen
GHISELLI, E.E.	1963	Moderating Effects and Differential Reliability and Validity. Jour. appl. Psychol., 41, 81-86
HERMANN, Th.	1969	Lehrbuch der empirischen Persönlichkeitsforschung. Göttingen
HÖRMANN, H.	1964	Theoretische Grundlagen projektiver Tests. In: HEISS, R. (Hg): Handbuch der Psychologie, Band 6, Göttingen
JANKE, W.	1964	Klassifikation. In: HEISS, R. (Hg): Psychologische Diagnostik. Hdb. der Psychologie, Band 6, Göttingen
JÄGER, R. & JUNDT, E.	1973	Mannheimer Rechtschreibtest (MRT) Göttingen und Braunschweig
JÄGER, R.	1974	Bemerkungen zu: Die Standardskala der kritischen Differenzen. Diagnostica, 2o, 165-168
JÄGER, R.	1974a	Moderatoransatz als vereinheitlichendes Prinzip. Möglichkeiten und Grenzen. Archiv für Psychologie, 126, 97-113
JÄGER, R.	1975	Probleme der Differentiellen Diagnostizierbarkeit. In: TACK, W.H.(Hg): Bericht über den 29.Kongreß der Deutschen Gesellschaft für Psychologie, Bd. 2, Göttingen

JÄGER, R. 1976 Ähnlichkeiten und Konsequenzen von Suppressorwirkungen und Multicollinearität. Psych. Beiträge, 18, 77-83

JÄGER, R.; LISCHER, S.; MÜNSTER, W. & RITZ, B. 1976 Biographisches Inventar zur Diagnose von Verhaltensstörungen (BIV). Göttingen

JÄGER, R. 1977a Quantifizierung von anamnestischen Daten. Ein Beitrag zur klinischen Diagnostik und zur differentiellen Diagnostizierbarkeit. in TACK, W.H. (Hg). Bericht über den 3o. Kongreß der Deutschen Gesellschaft für Psychologie. Göttingen

JÄGER, R. 1977b Differentielle Diagnostizierbarkeit. Theoretische und empirische Untersuchungen mit Moderatoren. Ein Beitrag zur psychologischen Diagnostik. Göttingen (im Druck).

KRAPP, A. 1975 Untersuchungen zur multiplen Prognose des Schulerfolgs bei Schulanfängern. Alternativen zur Schulreifediagnostik. Psychol. in Erz. u. Unterricht, 22, 78-87

LEHRL, S. & KINZEL, W. 1973 Die Standardskala der kritischen Differenz. Diagnostica, 19, 75-88

LIENERT, G.A. 1969[3] Testaufbau und Testanalyse. Weinheim

MICHEL, L. 1964 Allgemeine Grundlagen psychometrischer Tests. In: HEISS, R.(Hg): Psychologische Diagnostik Handbuch der Psychologie, Bd. 6, Göttingen

MICHEL, L. & ISELER, A. 1968 Beziehungen zwischen klinischen und psychometrischen Methoden der diagnostischen Urteilsbildung. In: GROFFMANN, K.-J. & WEWETZER, K.-H. (Hg): Person als Prozeß. Bern und Stuttgart.

MICHEL, L. & JÄGER, R. 1976 Entwicklung eines Hochschuleingangstests (HET) für Human- und Zahnmediziner. Allgemeine Probleme, Diskussion und Beschreibung des Projekts. (in Vorber.)

SAUNDERS, D.R. 1956 Moderator variables in prediction. Educ. Psychol. Measmnt. 16, 2o9-222

THORNDIKE, E.L. 1913 The psychology of learning. Educational Psychologa, Vol II, N.Y.

WESTMEYER, H. 1972 Logik der Diagnostik. Stuttgart.