Zur Validitätsfrage in der psychologischen Diagnostik: Die Reformulierung eines Problems

Von Helmut Lukesch, Konstanz

1. Die Konzeption diagnostischer Verfahren

Nachdem es auch für die Psychologie als akademische Wissenschaft eine unbezweifelbare Tatsache geworden war, daß zwischen den einzelnen Menschen Unterschiede bestehen (J. McKeen Cattell, 1890) wurde es als ein legitimes Anliegen und Ziel psychologischer Forschung angesehen, diese Unterschiede zu ergründen und möglichst genau[1] zu erfassen (W. Stern, 1921). Im allgemeinen geschah dies, indem man Testverfahren konzipierte, d. h. genormte Situationen, in denen auf ein vorgegebenes Stimulusmaterial reagiert werden muß[2], und zwar so, daß eine Varianz in den individuellen Reaktionen erhalten bleibt[3].

Man kann sich nun leicht vorstellen, daß es eine Unzahl von solchen Situationen gibt, in denen verschiedene Menschen verschieden reagierten. Ein Problem, das sich neben anderen Fragen[4] dabei immer wieder stellen wird, ist: welche Bedeutung kommt dem gezeigten Verhalten zu oder was erfaßt ein Test „wirklich", also ganz allgemein die Frage nach der Validität oder Gültigkeit diagnostischer Verfahren. Dabei wird vorerst unter Validität – in Anlehnung an P. J. Drenth (1969, S. 180) die „psychologische Bedeutung" eines Tests verstanden.

Versucht man, diese Frage zu beantworten, so sollte man sich das Vorgehen bei dem Entwurf eines „Tests" vor Augen halten: Besonders in der älteren Testpraxis war es üblich, daß der Experte (d. i. der Testautor) sich irgendeine Situation ausdachte, auf welche auf Grund meist ungeklärter Bedingungen mit einer

[1] Im Anschluß an die Untersuchungen von J. McKeen Cattell kam es sehr bald zu einer ausgearbeiteten Theorie des Testwertes. Diese Überlegungen, die in zusammengefaßter Form in den verschiedenen Lehrbüchern der Testtheorie vorliegen (H. Gulliksen, 1950; F. M. Lord und M. R. Novick, 1967), sollen hier nicht behandelt werden, da sie größtenteils Reliabilitätsprobleme betreffen. Erst in den neueren Lehrbüchern der Testtheorie sind wieder Überlegungen zur Validität zu finden (G. Fischer, 1968).

[2] Diese Definition ist verhältnismäßig weit. Weitere Kriterien, denen ein Verfahren, das die Bezeichnung „Test" verdient, genügen muß, sind bei G.E kman (1955), G. A. Lienert (1967), P. J. Drenth (1969) zu finden.

[3] Sicherlich sind auch solche Situationen denkbar, in welchen die Varianz individuellen Verhaltens auf ein Minimum eingeschränkt ist (z. B. durch eine entsprechende Instruktion). Nur sind solche Untersuchungsbedingungen für diagnostische Untersuchungen nicht sinnvoll.

[4] Andere Probleme betreffen die Meßgenauigkeit oder Reliabilität der Verfahren, ihre Situationskonstanz u. dgl. mehr.

hinreichend großen Variationsbreite der Reaktionen von seiten der Probanden zu rechnen ist. Man denke hier an den Formdeuteversuch von H. Rorschach, an Apperzeptionstests, an Farbwahltests, aber auch an die Verfahren und Aufgaben, welche der Feststellung der „Intelligenz"[5] dienen sollen. Die Überlegungen der Testautoren gingen dabei in den seltensten Fällen von klar formulierten Hypothesen über den Zusammenhang zwischen dem Testverhalten und dessen möglicher Bedeutung aus, sondern der Vorgang war umgekehrt: man hatte Verhaltensweisen provoziert und versuchte nun, deren Bedeutung abzuklären.

1.1. Logische Validität

Es ist bei diesem Vorgang der Erklärungssuche für das Testverhalten zu konstatieren, daß dem Testverhalten diese oder jene Interpretation zugemessen wird, weil der Testautor dies für plausibel hält (so will u. a. R. Brickenkamp logische Validität für den Test d 2 aus Selbstbeobachtungen herleiten [1962, S. 21][6]).

Gefährlich wird diese Art der Begründung, weil der Anspruch, der mit dem Wort „logisch" erhoben wird, zumeist nicht gerechtfertigt ist. Wollte man versuchen zu erklären, was damit gemeint ist, so könnte man sagen, daß es eine Anzahl empirisch bewährter und in sich widerspruchsfreier Prämissen gibt, aus denen mit einer logisch gültigen Schlußform die Validität des Verfahrens abgeleitet werden kann[7]. Nun ist es aber bei der Berufung auf logische Validität gerade so, daß diese Prämissen nicht explizit formuliert werden und daß sie auch nicht hinreichend bewährt sind, sondern einzelne Hypothesen oder vielmehr bloße Vermutungen darstellen[8], für welche empirische Belege noch fehlen.

L. Michel (1964) plädiert aus ähnlichen Überlegungen dafür, den Terminus nicht mehr zu verwenden. Vermutungen oder Erklärungsansätze, die auf solchen Grundlagen basieren, bedürfen der empirischen Untermauerung, um auch für die Diagnostik brauchbar zu sein. Der Hinweis auf logische Gültigkeit kann dafür kein Ersatz sein, sondern nur Anstoß zu fundierterer Theorienbildung oder empirischer Forschung.

1.2. Augenscheinlichkeits-(face-)Validität

Auch diese Bezeichnung sagt über die Bedeutung eines Testverhaltens nur aus, daß es sich um noch unbestätigte Vermutungen handelt. Wird logische Validität einem Verfahren eher unter Berufung auf Experten und Autoritäten zugeschrieben, so bezeichnet Augenscheinlichkeitsvalidität die Tatsache, daß auch die Probanden Vermutungen über die Bedeutung eines Testverhaltens haben können. Diese Tatsache rechtfertigt zwar nicht den Gebrauch des Wortes „Validität", kann aber nach

[5] Daß es sich auch bei diesem doch so geläufigen Begriff eigentlich um ein noch unzureichend erforschtes Konstrukt handelt, darauf weisen neuere Untersuchungen hin (vgl. E. Roth, W. D. Oswald und K. Daumenlang, 1972).

[6] Dieser Test ist hier nur als Illustration angeführt. Ein abwertendes Urteil über den Test, der sich auf viele empirische Bewährungsuntersuchungen stützen kann, ist damit nicht impliziert.

[7] Diese Interpretation kommt dem Vorgang der Begriffsvalidierung, der später behandelt wird, nahe und wurde in dieser Bedeutung auch von L. Cronbach (1949) gebraucht.

[8] Der Ausdruck „Hypothese" wird in der Literatur sehr willkürlich verwendet. Kriterien, nach deren Erfüllung ein Satz als Hypothese gelten kann, sind bei P. Weingartner (1971, S. 57ff.) oder M. Bunge (1967, Bd. I, S. 222ff.) zu finden.

L. Michel (1964, S. 52) für die Diagnostik insofern von Bedeutung sein, als die Probanden u. U. besser motiviert sein können, wenn sie den Sinn eines Verfahrens einzusehen glauben (Vgl. Guilford, 1954). Sollte man aber der Meinung sein, daß Augenscheinlichkeitsvalidität für die Begründung eines Verfahrens allein genügt, so müßte mit P. J. Drenth (1969, S. 193) gesagt werden, daß dieser Begriff nicht weit entfernt ist von „faith-validity" oder mit R. B. Cattell (1957, S. 338), daß er vermutlich der „false-validity" entspricht.

Stellt man aber einen Test zur Erfassung eines willkürlich gewählten, alltagssprachlichen Konstrukts zusammen, so wird man allerdings nach logischer und Augenscheinlichkeitsvalidität eine erste Vorauswahl der Testfragen treffen. Wie von L. R. Goodstein und P. Slovoc (1971, S. 255 f.) ausgeführt, ist die Chance bei Testitems mit hoher face-validity Beziehungen zu empirischen Kriterien zu finden höher, als wenn man den anfänglichen Item-Pool zufällig zusammengestellt hätte. Allerdings muß man damit auch in Kauf nehmen, daß der Test von den Probanden leichter durchschaut und somit beeinflußt werden kann.

Trügerisch wäre allerdings die Hoffnung, daß Augenscheinlichkeitsvalidität und empirische Validität isomorph seien oder sich gar decken. Die Berufung auf Augenscheinlichkeitsvalidität kann ebenfalls nur andeuten, daß hier vorerst nur vage begründete Vermutungen über die Erklärung eines Testverhaltens vorliegen und nicht mehr.

1.3. Ganzheitliche Auswertung

In dem Zusammenhang mit der Überprüfung der Validität von Testverfahren findet man in der älteren deutschen Literatur das holistische Argument, das bisweilen auch in anderen Erklärungszusammenhängen für relevant gehalten wird[9]. Daß sich im Testverhalten etwas, was zumeist die „gesamte Persönlichkeit" des Probanden genannt wird, in ganzheitlicher Weise ausdrückt und daher auch ganzheitlich interpretiert werden muß, wird bisweilen als besonderer Vorzug von diagnostischen Verfahren geschildert[10] (H. Lossen, 1955; A. Wellek, 1958).

Dagegen lassen sich einmal empirische Untersuchungen anführen, z. B. diejenigen von H. J. Eysenck (1954), in welchen nachgewiesen werden konnte, daß die „atomistische" Auswertung des Rorschach-Tests den ganzheitlichen Ratings von Experten überlegen war. Aber das Problem muß auch noch von einer anderen Seite her gesehen werden. Es besteht hohe empirische Evidenz, daß auf intuitivem oder ganzheitlichem Weg aus einem Test wie z. B. dem TAT (W. J. Revers und K. Taeuber, 1969) zutreffend biographische Situationen in ihrer Erlebnisbedeutung diagnostiziert werden können und daß im Moment keine Regeln angebbar sind, nach denen ein erfahrener TAT-Diagnostiker seine Schlüsse vornimmt. Eine andere Sache ist es aber zu behaupten, daß es prinzipiell unmöglich sei, für ein solches Vorgehen Regeln zu finden. Intuition kann für einen Diagnostiker sehr hilfreich sein, für den Wissenschaftler aber ist es der Hinweis darauf, daß ein Verfahren noch nicht hinreichend kontrolliert wird[11].

[9] Dieses Argument, das eine der Doktrinen des sog. Historizismus darstellt, wird von K. R. Popper (1971, S. 61 ff.) einer Analyse unterzogen.

[10] So werden z. B. die Vorzüge des „Mehrdimensionalen Zeichentests" von R. Bloch (1971) geschildert!

[11] Das Argument des „Verstehens" im Begründungszusammenhang, das bei solchen Diskussionen bisweilen zu finden ist, soll hier nicht weiter zur Sprache kommen. Zusammenfassungen darüber sind zu finden bei H. Rohracher (1971, S. 97 f.) oder W. Stegmüller (1969).

für bestimmte Entscheidungen liefern, für andere Vorhersagezwecke aber ungeeignet sein, als Beispiel ist ein Intelligenztest denkbar, der mit erfolgreichem Abschluß des Hochschulstudiums insgesamt in mittlerem Ausmaß korreliert. Bei der Entscheidung, ob jemand überhaupt ein Hochschulstudium beginnen soll, könnte ein solcher Test einen Beitrag leisten. Für die Beantwortung der Frage, welche Studienrichtung der Bewerber aber wählen soll, wird ein solcher Test nur wenig leisten können, da nur ein allgemeiner Zusammenhang zwischen Testresultat und Hochschulabschluß überhaupt besteht.

Solche Überlegungen führten dazu, daß von *differentieller Validität* gesprochen wird (L. Cronbach, 1969, S. 356). Damit ein Test ein ideales Klassifikationsinstrument darstellt, sollte es so sein, daß er mit einem bestimmten Kriterium hoch, mit allen anderen aber gar nicht korreliert. Je nach Fragestellung kann man also einem Test differentielle Validität zubilligen oder nicht. Dieses Prinzip kann auch auf verschiedene Validitätsbestimmungen bei unterschiedlichen Alters- oder Berufsgruppen übertragen (z. B. H. Schlange et al., 1972, S. 22) werden. Ebenso ist es denkbar, daß die Gültigkeitsbestimmung für einen Test in der Form der korrelativen Beziehung zwischen Test und Kriterium je nach situativen Randbedingungen anders ausfallen kann und daher auch gesondert erhoben werden sollte.

Nach dieser utilitaristisch zu nennenden Anschauung, derzufolge Validität mit Brauchbarkeit gleichgesetzt wird, besitzt ein diagnostisches Verfahren so viele „Validitäten" wie es Situationen gibt, die auf Grund des Testverhaltens erfolgreich vorhergesagt werden können. Frägt man aber weiter, so müßte auch hier eine Analyse der Bedingungen, aus denen Test- und Kriteriumsverhalten folgen, Gemeinsamkeiten aufweisen können.

2.3. Praktische Schwierigkeiten bei der Kriteriumsvalidierung

1. Wie G. A. Lienert (1967, 17f.) schreibt, ist der erhaltene Korrelationskoeffizient zwischen Test und Kriterium neben dem Grad der „Gemeinsamkeit", den beide Verhaltensreihen besitzen mögen, auch von der Reliabilität des Tests und des Kriteriums abhängig. Durch Meßungenauigkeiten auf beiden Seiten bei der empirischen Erhebung oder wegen der zeitlichen Fluktuation der Merkmale wird also der erhaltene Validitätskoeffizient sinken (nachträgliche Korrekturen sind zwar auf statistischem Wege möglich (H. Gulliksen, 1950, S. 101 ff.), aber auch mit bestimmten Vorannahmen belastet).

2. Bei der Erhebung von Kriteriumswerten zu einem späteren Zeitpunkt als die eigentliche Testdurchführung wird sich oft ergeben, daß die ursprüngliche Stichprobe nicht mehr vollständig zur Verfügung steht. Diese Varianzeinbuße ist z. B. dadurch bedingt, daß nur die Besten in einem Test für einen Betrieb aufgenommen werden und später noch einmal getestet werden können. Durch solche systematischen Selektionen wird der erhaltene Validitätskoeffizient beeinträchtigt (D. Magnusson, 1969, S. 155).

3. Erhebt man Kriteriums- und Testverhalten zur gleichen Zeit und verwendet die erhaltenen Korrelationskoeffizienten als Grundlage für Entscheidungen, so besteht auch hier die Möglichkeit von Fehlschlüssen. Von einem Generaldirektor wird man z. B. kaum erwarten können, daß er sich einem Pauli-Test unterzieht und selbst wenn er dies tun sollte, so ergeben die Resultate noch keinen Hinweis darauf, welche Qualitäten ein Bewerber, der am Anfang der Laufbahn steht, auch tatsächlich haben muß. Auch ist der Fall denkbar, daß die Eigenschaften, die z. B. für einen erfolgreichen Manager charakteristisch sind, für einen Anfänger eher hin-

derlich sind (eine oder mehrere Herzattacken sind nach dem Image eines Managers für diesen zwar charakteristisch, dennoch wird es niemanden einfallen, einen Bewerber, dessen Endziel eine Managerposition sein könnte, nach der Anzahl der Herzattacken auszusuchen).

4. Es sollte auch in Beracht gezogen werden, daß bei der Zuordnung von Zahlen zu einem Kriteriumsverhalten (d. i. dessen Messung) ein Informationsverlust einhergehen kann (z. B. kann es dazu kommen, daß das Kriteriumsverhalten nur grob klassifiziert werden kann, da die Analyse des Kriteriums nur ungenügend möglich war) und eine Erhöhung der Zufallsvarianz in Kauf genommen werden muß (z. B. indem die Rater bei der Beurteilung des Kriteriumsverhaltens nicht genau übereinstimmen). Beide Möglichkeiten wirken sich negativ auf die Höhe des Validitätskoeffizienten aus.

5. Die erhaltenen Korrelationskoeffizienten sind bisweilen nicht sehr stabil, sondern wechseln von Stichprobe zu Stichprobe. Es ist daher zu fordern, daß man die erhaltenen Validitätskoeffizienten mittels *Kreuz-Validierung* auf ihre Stabilität hin überprüft. Wie J. Guilford ausführt (1954, S. 406), kann dies dadurch geschehen, daß man an einer ersten Stichprobe Validitätskoeffizienten erhebt und diese zur Schätzung der Güte der Vorhersage an einer zweiten Stichprobe verwendet. Hat man bereits zwei Stichproben, so kann man auch eine doppelte Kreuz-Validierung vornehmen. Ergeben sich bei diesem Verfahren signifikante Änderungen in der Güte einer Vorhersage, so sind die entsprechenden Regressionsgleichungen zu ändern. Sollte aus ökonomischen Gründen keine zweite Stichprobe getestet werden können, so ist es auch denkbar, daß man die erste Stichprobe nach einem Zufallsverfahren teilt und mit diesen beiden Gruppen eine Kreuzvalidierung durchführt. Alles in allem wird man bei diesem Vorgehen keine sehr hohen Maßzahlen für die Validität eines Verfahrens bekommen. Daß einzelne Korrelationskoeffizienten sicherlich signifikant werden, ist nur ein schwacher Trost, denn statistische Signifikanz kann allenfalls nur als eine Vorbedingung für sachgerechte Interpretation angesehen werden (Vgl. G. Kleiter [1969])[14].

Durch die Anwendung entsprechend verfeinerter statistischer Verfahren können die Validitätskoeffizienten aber auch erhöht werden. Es müssen hier vor allem die multiplen Ansätze herangezogen werden, bei denen mehrere Prädiktoren oder mehrere Kriteriumsmaße oder beides gleichzeitig analysiert werden (W. Cooley und P. Lohnes, 1971). In diesem Zusammenhang ist aber auch die Verwendung der Konfigurationsfrequenzanalyse zu nennen (G. A. Lienert, 1971 u. a. m.) für den Fall von Daten auf nominalem Skalenniveau und die Anwendung von Moderatorentechniken (D. Bartussek, 1970) bei Intervalldaten. Auf diese Möglichkeit weist auch E. E. Ghiselli (1971) nachdrücklich hin. Durch den Gebrauch kurvilinearer Regressionsansätze müßte bei entsprechenden Daten ebenfalls eine Erhöhung der Validitätsmaße möglich sein (N. R. Draper und H. Smith, 1966; J. Adam, J.-H. Scharf und H. Enke, 1971).

2.4. Die Validität des Kriteriums

Bei der kriteriumsbezogenen Validierung wird implizit vorausgesetzt, daß es Kriterien gibt, welche ein besseres oder gül-

[14] Wie G. Kleiter ausführt, wird praktisch jeder Korrelationskoeffizient statistisch signifikant, wenn die Stichprobe, an der er erhoben wird, entsprechend groß ist. Ob man solche Korrelationskoeffizienten aber noch sachlich interpretieren kann, ist damit noch nicht geklärt.

tigeres Maß für das anstehende Problem sind als das diagnostische Instrument. Warum man nicht gleich das Kriteriumsverhalten als Grundlage für weitere Entscheidungen nimmt, liegt darin, daß das Kriteriumsverhalten zum gegebenen Zeitpunkt noch nicht vorliegt oder aber auch darin, daß Kriteriumsmaße teuer und umständlich zu erheben sind.

Wie bereits angedeutet, wird es bei unreliablen Kriteriumsmaßen, wie es z. B. psychiatrische Diagnosen darstellen, nur zu geringen Übereinstimmungen mit Testergebnissen kommen können. Dies konnte erneut von P. Matussek (1971) nachgewiesen werden; aus dessen Untersuchung über die Folgen der KZ-Haft kann man ersehen, daß die Diagnose, die von Psychiatern abgegeben wird, durchaus von anderen Faktoren bedingt ist als durch Krankheitssymptome und -syndrome selbst.

Bei der Suche nach valideren Kriterien verwendet man auch solche Tests, von denen man annimmt, daß sie sich bereits bewährt haben. Ist aber die Übereinstimmung sehr hoch, so bedeutet dies, daß dasselbe erfaßt wird (Vgl. I. Stelzl, 1970) – wobei noch die Frage ungeklärt bleibt, was erfaßt wird. Dies ist aber nur dann wünschenswert, wenn man ein zweites Verfahren für denselben Anwendungsbereich (oder für dieselbe theoretische Frage) entwerfen will. Ist die Übereinstimmung aber gering, so muß dies nicht unbedingt gegen das Verfahren sprechen, denn damit hat man gezeigt, daß man einen anderen Aspekt im Verhalten der Probanden erfaßt hat – wobei wieder offen bleibt, was man eigentlich erfaßt –, z. B. wird bei den sogenannten Leistungstests immer wieder betont, daß die Testparameter unabhängig von Intelligenzmaßen sind, d. h. daß sie etwas anderes erfassen als Intelligenztests (Vgl. H. Bartenwerfer, 1964)[15].

Die Schwierigkeit bei der Erfassung von Kriterien besteht eigentlich darin, daß für die Maße zur Erfassung des Kriteriumsverhaltens genau so viel Mühe und Arbeit verwendet werden müßte wie für die Konstruktion eines Diagnostikums selbst. Damit sind aber Testautoren endgültig überfordert, und das Resultat ist, wie R. Ebel schreibt (1961, S. 643), folgend: „What happens ... is that we accept highly questionable criteria, obtain discouragingly low correlations, and finally give the whole thing up as a bad job."

2.5. Diskussion der Kriteriumsvalidierung

Den Versuchen der bloßen Kriteriumsvalidierung unter Absehen von der Testbedeutung unterliegt nach H. Hörmann (1964) das diagnostische Modell des Behaviorismus. Damit ist gemeint, daß der Validierungsprozeß sich in dem Finden korrelativer Beziehungen erschöpft (vgl. Abb. 1).

Bei der kriteriumsbezogenen Validierung wird der Zusammenhang zwischen einem Test und einem Priterium (a), einem Test mit mehreren Kriterien (b) oder mehreren Tests mit mehreren Kriterien (c) zu erfassen versucht. Die Funktion (f), die zwischen Test und Kriterien besteht, kann als Korrelations- oder Regressionsansatz interpretiert werden.

Wollte man dieses Modell mit allen seinen Konsequenzen beibehalten, so dürfte man aus dem Testverhalten keine Schlüsse ziehen (d. h. die Anwendung des Tests genereller gestalten), die über die eng umschriebene Situation, in welcher Test und Kriterium erhoben wurden, hinausgehen.

[15] Auch D. Campbell (1960) verlangt, daß publizierte Tests systematisch auf ihren Zusammenhang mit Intelligenzmaßen, sozialer Erwünschtheit, Zustimmungstendenz und anderen Antworttendenzen und Selbstbeurteilungen geprüft werden.

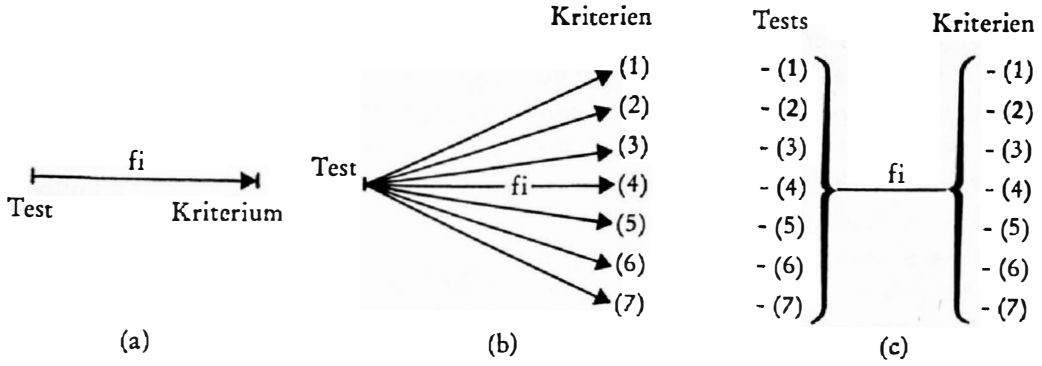


Abb. 1:
Veranschaulichung der kriteriumsbezogenen Validierungsversuche

Für bestimmte Zwecke mag dieser Ansatz genügen (z. B. für Personalauslese). Eigentlich setzt man sich aber dabei Scheuklappen auf, indem man einen Rekurs auf gemeinsame Bedingungen, aus denen vielleicht sowohl Test- wie Kriteriumsverhalten folgen, ausschließen will.

Auch werden sich hierbei in der Praxis Schwierigkeiten ergeben, denn die Mitteilbarkeit der Ergebnisse psychologischer Untersuchungen, die auch eines der Hauptprobleme in der psychologischen Praxis darstellt (K. Holzkamp, A. O. Jäger und F. Merz, 1966; H. Hartmann, 1970; G. Jüttemann 1972, S. 95), wird durch ein solches Modell beträchtlich erschwert. Außerdem erfolgt ja auch eine Kriteriumsvalidierung nicht „blind", sondern man geht zumindest implizit von Hypothesen über den Zusammenhang von Test und Kriterien aus und beginnt nicht mit bloßem trial and error.

Und schließlich hat dieses Vorgehen trotz allen methodischen Aufwandes noch nicht die Ergebnisse gezeigt, die man sich erhoffen könnte. So kann die Psychologie die Frage der Verkehrstauglichkeit noch nicht mit genügender Sicherheit beantworten (vgl. D. Oswald, 1971), auch ist es der psychologischen Forschung bis heute noch nicht gelungen, solche Testverfahren zu entwerfen und empirisch zu validieren, mit deren Hilfe eine zufriedenstellende Selektion der Studienanfänger im Fach Psychologie vorgenommen werden könnte (vgl. B. Gasch, 1970)[16].

Diese kritischen Punkte am Konzept der Kriteriumsvalidierung sollen aber nicht so verstanden werden, daß die Feststellung von Beziehungen auf empirischer Ebene nutzlos ist; empirische Überprüfungen werden in einer Wissenschaft wie der Psychologie immer notwendig und entscheidend sein.

2.6. Zur Logik der Entscheidung bei kriteriumsvalidierten Verfahren

Das Schema, dem Aussagen oder Entscheidungen folgen, die man auf Grund von Daten aus kriteriumsorientierten Untersuchungen von psychodignostischen Verfahren trifft, kann wie folgt charakterisiert werden: Man hat zuerst festgestellt, daß ein Test A mit einem erfolgreichen Verhalten in einer Berufssituation korreliert (vielleicht wurde der Zusammenhang sogar an einer repräsentativen Stichprobe erfaßt). Sollte sich ein hoher Validitäts-

[16] Dies ist sicher eine simplifizierende Behauptung. Selbst wenn es gelänge, valide Auswahlverfahren zur Verfügung zu stellen, so ist damit noch nicht gesagt, daß diese Verfahren auch zur Anwendung kämen.

koeffizient ergeben, so wird man den individuellen Testwert X eines Probanden in die Regressionsgleichung einsetzen und dann eine Aussage – zumindest auf nominalem oder ordinalem Niveau – über die Eignung dieses Probanden in dem betreffenden Beruf machen.

Man kann nun zu der Ansicht neigen, daß dieses Vorgehen dem deduktiv-nomologischen Verfahren zur Erklärung eines einzelnen Ereignisses in empirischen Wissenschaften entspricht, daß nämlich hier aus einem (oder mehreren) Gesetz(en) und der Feststellung von Anfangsbedingungen mittels eines logisch gültigen Schlusses eine Erklärung oder Vorhersage eines singulären Ereignisses folgt[17].

G: „Wenn jemand einen bestimmten Testwert X in dem Test A erhält, dann wird er sich in dem Beruf erfolgreich verhalten."
A: „Herr Maier hat den Testwert X im Test A erreicht."

E: „Herr Maier wird sich in dem Beruf bewähren."

Dieses Schlußschema kann aber auf den vorliegenden Fall nicht angewandt werden, denn es setzt voraus, daß die gesetzesartigen Aussagen im Explanans deterministische Aussagen sind. Kriteriumsorientierte Validierungsstudien von diagnostischen Verfahren führen aber nur zu Wahrscheinlichkeitsaussagen, und zwar zu sog. „induktiv-statistischen Systematisierungen" (W. Stegmüller, 1969, S. 627). Übertragen auf das vorliegende Beispiel muß die Formulierung korrekter lauten:

„Die Wahrscheinlichkeit, daß jemand, der den Testwert X im Test A besitzt, zur Klasse V (der Menschen, die in dem Beruf ein erfolgreiches Verhalten zeigen) gehört, beträgt r."
„Herr Maier erreicht in dem Test A den Testwert X."

„Herr Maier wird sich in dem Beruf bewähren."

Dieser Folgerung kommt nun eine gewisse induktive Wahrscheinlichkeit zu. Es ist dabei zu beachten, daß hier kein logischer Schluß vorliegt, sondern eine induktive Aussage, welche dem letzten Satz relativ zum Ausgangsdatum und der statistischen Verallgemeinerung einen bestimmten Grad an Bestätigung zuschreibt. Wie W. Stegmüller zeigt (1969, S. 631 f.) kommt nur der Folgerung, nicht aber dem Ereignis oder der Vorhersage eine bestimmte Wahrscheinlichkeit zu (denn das Ereignis tritt ein oder tritt nicht ein).

Diese Art der induktiv-statistischen Systematisierung führt noch zu einer Reihe von Problemen wissenschaftstheoretischer Art, die teilweise noch ungelöst sind. Für das pragmatische Vorgehen in der Diagnostik lassen sich in Anschluß an W. Stegmüller aber einige Regeln formulieren, welche eine korrekte Anwendung solcher Systematisierungen erlauben sollen.

1. Bei einer Entscheidung über Sachverhalte, über welche statistische Hypothesen vorliegen, muß man sich auf das ganze im Moment verfügbare Wissen stützen, d. h. eine Entscheidung ist immer nur relativ

[17] Dieses Schema für deduktiv-nomologische Erklärungen wird ausführlich bei W. Stegmüller (1969, S. 86 f.) besprochen. Es kann allgemein wie folgt angeschrieben werden:

Explanans	$A_1, \ldots A_n$	(Sätze, welche die Antecedensbedingungen beschreiben)
	$G_1, \ldots G_n$	(allgemeine Gesetzmäßigkeiten)
Explanandum	E	(Beschreibung des zu erklärenden Ereignisses)

auf das zu einem Zeitpunkt verfügbare Wissen rational[18].

Beispiel: Weiß man, daß jemand eine bestimmte phobische Vorstellung hat und daß diese durch bestimmte Arten der Therapie geheilt werden kann, so wird man die Folgerung für wahrscheinlich halten, daß jemand, der diese phobische Vorstellung hat und behandelt wird, geheilt wird. Bekommt man aber nun die Zusatzinformation, daß der Betreffende stark extrovertiert ist, und liegt die statistische Aussage vor, daß Extravertierte therapieresistent sind, so wird man die Folgerung, daß dieser Patient geheilt wird, nicht mehr mit der gleichen Wahrscheinlichkeit annehmen.

2. Werden probabilistische Hypothesen zur Erklärung oder Vorhersage verwendet, so ist die schärfere Information der schwächeren vorzuziehen.

Beispiel: Die Wahrscheinlichkeit, daß jemand in einem Test A dem Testwert X erreicht und sich erfolgreich in einem Beruf verhalten wird, beträgt r. Die Wahrscheinlichkeit, daß jemand in einem Test B den Wert Y erreicht und sich in dem gleichen Beruf nicht erfolgreich verhalten wird, beträgt s. Weiß man nun, daß jemand, der im Test A den Wert X und im Test B den Wert Y erhält, sich mit einer Wahrscheinlichkeit von t in dem Beruf erfolgreich verhalten wird, so ist die in dieser statistischen Hypothese enthaltene Information den beiden anderen vorzuziehen.

3. Weiß man nicht, welche Information die stärkere ist, so ist eine Zuordnung im Moment nicht möglich, d. h. man weiß nicht, welche Art der statistischen Systematisierung korrekt ist.

Beispiel: Gegeben sei der obige Fall, nur liege keine Information vor über den Einfluß der Zugehörigkeit zu der Klasse derjenigen, die den Wert X im Test A und den Wert Y im Test B erhalten haben, mit dem Verhalten im Beruf vor. In diesem Fall kann die Frage, ob sich ein bestimmter Mensch, der sowohl den Wert X im Test A und den Wert Y im Test B erhalten hat, in dem betreffenden Beruf erfolgreich verhalten wird, nicht entschieden werden. Praktisch bedeutet dies, daß man weitere Informationen suchen muß[19].

Diese Regeln sehen davon ab, daß auch die Feststellung der Antecedensbedingungen nur mit bestimmten Wahrscheinlichkeiten vor sich geht. Zusammenfassend ist zu betonen, daß die Folgerungen nur mit der durch die Antecedensbedingungen und statistischen Hypothesen spezifizierten bedingten Wahrscheinlichkeiten gelten.

3. Das Ziel: Die Suche nach „Konstrukten"

Gegen alle bisher aufgezeigten Validierungsvorschläge, wenn damit ein endgültiger Abschluß der Suche nach der Testbedeutung gemeint war, haben sich schwerwiegende Bedenken formulieren lassen: für logische, face- und inhaltliche Validität haben sich keine zureichenden empirischen Verankerungen finden lassen; bloße empirische Validitätsstudien erwiesen sich als unbefriedigend, wenn damit ein völliges Absehen von der Suche nach der theoretischen Bedeutung eines Tests gemeint war, denn damit sind sehr beschränkte Anwendungs-, Interpretations- und Mitteilungsmöglichkeiten verbunden. Mit den bisher genannten Spielarten des Validitätsbegriffes sind aber noch nicht alle Möglichkeiten erschöpft. Eine weitere Gruppe von Überlegungen rankt sich um die Ausdrücke der „Konstrukt-" oder „Begriffs-"Validität, und bei Durchsicht

[18] „Eine gültige induktive Relation ist jedoch nur dann auf eine konkrete Situation anwendbar, wenn gewährleistet ist, daß im Datum alles für die Hypothese relevante Erfahrungswissen aufscheint" (W. Stegmüller, 1969, S. 661).

[19] Diese Regeln wurden in Anschluß an Hempels Prinzip der maximalen Bestimmtheit formuliert (vgl. W. Stegmüller, 1969, S. 664 f.).

der Literatur findet man, daß diese Validitätsart das erklärte Ziel aller Bemühungen der Bedeutungserhellung eines diagnostischen Verfahrens darstellt.

3.1. Bedeutung von „Konstrukt-" oder „Begriffs"-Validität

Die im deutschen Sprachraum bekanntesten Überlegungen zu dieser Methode der Gültigkeitsbestimmung für ein diagnostisches Verfahren stammen von H. Hörmann (1961, 1964) und werden auch oft zitiert (L. Michel, 1964; P. J. Drenth, 1969; Th. Herrmann, 1969; u. a. m.). Der Sprachgebrauch läßt dabei vermuten, daß nahegelegt wird, daß „Konstrukte" gleich welcher Art als etwas Handfestes angesehen werden, etwas, was tatsächlich in der Realität existiert und greifbar ist. Die dadurch erfolgte Reifikation von Persönlichkeitseigenschaften u. dgl. wird auch von D. Campbell (1960) in Anschluß an H. Bechtoldt (1959) kritisiert.

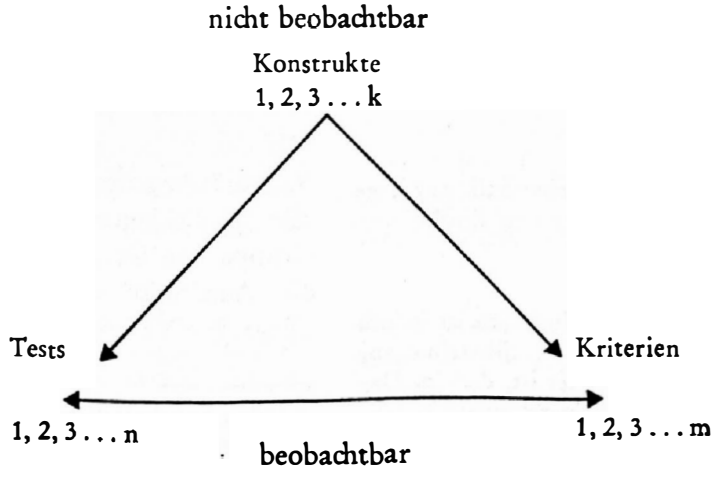
Bei der Besprechung von „Konstruktvalidität" wird davon ausgegangen, daß eine Testleistung bedingt ist durch dahinterstehende Eigenschaften, Faktoren oder im weitesten Sprachgebrauch von „analytischen Einheiten" im Sinne E. Roths (1969). Korrelationen und Zusammenhänge zwischen Test- und Kriteriumsverhalten werden ebenfalls zu erklären versucht, indem man auf gemeinsame Bedingungen hinweist (vgl. Abb. 2). (Beispiele für diese Art der Bedeutungserhellung findet man bei den angegebenen Stellen bei H. Hörmann.)

> Bei der Konstruktvalidierung wird versucht, die (korrelativen) Beziehungen, die auf empirischer Ebene zu beobachten sind, durch dahinterstehende Konstrukte zu erklären.

Im Gegensatz zu den anderen Validierungsversuchen wird hierbei ausdrücklich die Forderung aufgestellt, daß die Bedeutungserhellung eines Tests immer verbunden ist mit seiner Verankerung in einer „Theorie"[20]. Diese stellt das Bezugssystem

[20] R. Meili ist dabei aber der Meinung, daß es im Bereich der Persönlichkeitsforschung noch keine hinreichend ausgearbeiteten Theorien gibt und daß daher die Diskussion der Konstruktvalidierung nicht vordringlich ist (1965, S. 315).

Abb. 2:
Zusammenhang zwischen Test- und Kriteriumsverhaltensweisen mit dahinterstehenden Konstrukten



dar, aus welchem empirisch überprüfbare Hypothesen ableitbar sind, deren empirische Überprüfung mittels Testverfahren geschehen kann und schließlich soll die Bewährung[21] oder Falsifizierung dieser Hypothesen eine Theorie stützen oder dazu führen, daß diese modifiziert wird. Damit verbunden ist aber auch, daß die Testverfahren, die eine Überprüfung einer Hypothese möglich machen, ebenfalls bewährt, sprich: validiert, werden. Dieser Vorgang wird auch als „Prozeß der sukzessiven Approximation" beschrieben (H. Hörmann, 1961, S. 48), womit gemeint ist, daß schrittweise eine bessere Annäherung an die Realität erreicht werden kann. (Auf die gegenseitige Anregung von Grundlagenforschung und Praxis weist auch W. Witte [1966] hin.) Genauso wird dabei betont, daß Verhalten in der diagnostischen Situation und Persönlichkeitskonstrukt nicht kongruent sind, sondern ein Mehr an Bedeutung besitzen.

3.2. Die Unterscheidung zwischen intervenierender Variable und hypothetischem Konstrukt

Der eigentliche Grund, warum der „Konstruktvalidität" solche Bedeutung zugemessen wird, dürfte allgemeiner sein und mit der Theorienbildung in empirischen Wissenschaften überhaupt verbunden sein. Es handelt sich dabei u. a. um die Frage, ob eine Theorie einer empirischen Wissenschaft allein aus Beobachtungsbegriffen aufgebaut sein kann[22]. Würde man nur im Bereich des direkt Beobachtbaren bleiben, so würde man erst gar nicht zu einer Theorie gelangen, sondern höchstens zu einer Sammlung von Fakten und Phänomenen und eventuell zu einer Beschreibung und Systematisierung im Vorfeld einer Theorie. Mittels nicht direkt beobachtbarer Konstrukte (explikative Konstrukte" im Sinne Th. Herrmanns [1969, S. 61] oder „analytische Einheiten" im Sinne E. Roths [1969, S. 37 ff.]) aber wird versucht, einen weiten Bereich von Beobachtungskonzepten und Beobachtungstatsachen (im vorliegenden Fall: verschiedene Test- und Kriteriumsverhaltensweisen) zu systematisieren, diese zu erklären und auch vorhersagbar oder manipulierbar zu machen.

Diese Begriffe, die nicht vollständig auf Beobachtung zurückgeführt werden können, werden nach einem Vorschlag von K. MacCorquodale und P. Meehl (1948) in sogenannte „intervenierende Variable" und „hypothetische Konstrukte" eingeteilt. Folgt man der Bedeutung, welche die beiden Autoren diesen Begriffen beilegen, so beziehen sich zwar beide auf eine Reihe von Beobachtungstatsachen, welche sie systematisieren und auch erklären sollen, während aber hypothetische Konstrukte die Existenz von Entitäten voraussetzen, sind intervenierende Variable nur kurze Zusammenfassungen von Beobachtungen (S. 105 f.). Diese mit Voraussetzungen (z. B. ontologischer Art) vorbelastete Unterscheidung wird von M. Bunge (1967 a, S. 93 f.) dahingehend modifiziert, daß hypo-

[21] Man findet in der älteren Literatur des öfteren die Meinung vertreten, daß universelle Hypothesen einer faktischen Wissenschaft verifiziert werden könnten. Daß dies im strengen Sinn nicht möglich ist, wird durch die Überlegungen von K. R. Popper (1966) nahegelegt.

[22] Nach M. Bunge (1967 a, S. 92) beziehen sich Beobachtungsbegriffe auf direkt beobachtbare Objekte, wie z. B. „Körper" oder „Stimulus". Begriffe, die nicht direkt auf Beobachtung zurückgeführt werden können, sind z. B. in den Axiomen einer Theorie zu finden, wie die Begriffe „Masse" oder „Kraft" in der klassischen Physik. Aber auch, wenn man „Stimulus" als „Informationsträger" interpretiert (M. Toda, 1972), so ist damit mehr gemeint als die unmittelbar beobachtbare Reizgegebenheit.

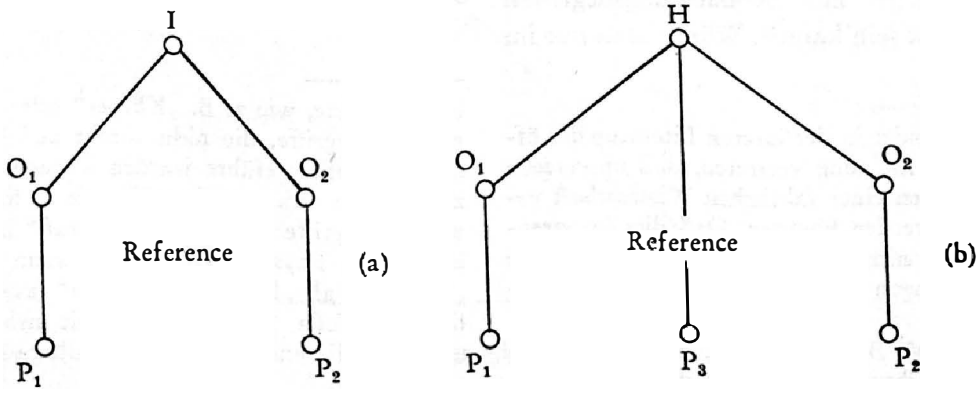
thetische Konstrukte im Unterschied zu den intervenierenden Variablen auch unabhängig von den Beobachtungstatbeständen, zu deren Erklärung sie angenommen werden, gemessen werden können (vgl. Abb. 3). Ob ein Konzept nun als intervenierende Variable oder als hypothetisches Konstrukt anzusehen ist, hängt vom Stand der Theorienbildung in dem betreffenden Bereich ab[23]; durch weitere Forschung oder Interpretation von Sachverhalten können bestimmte Konzepte vom Status einer intervenierenden Variable in den eines hypothetischen Konstrukts übergehen. Weiters ist anzumerken, daß dieselbe Wortmarke (z. B. „Trieb" oder „habit") je nach theoretischem Kontext einmal ein hypothetisches Konstrukt und einmal eine intervenierende Variable bezeichnet.

(a) Die intervenierende Variable I vermittelt zwischen den Beobachtungsbegriffen O_1 und O_2, welche durch die Eigenschaften P_1 und P_2 gekennzeichnet sind, wobei I selbst keine solche Eigenschaft besetzt. (b) Das hypothetische Konstrukt H vermittelt zwischen den Beobachtungsbegriffen O_1 und O_2 und hat darüber hinaus noch einen objektiven Referenten, nämlich die Eigenschaft P_3 (M. Bunge, 1967 a, S. 93).

Zur Illustration sei ein Beispiel angeführt: Die sehr globale Bezeichnung „Intelligenz" wird verwendet, um Verhaltensweisen, die z. B. im Alltag oder in Intelligenztests vorzufinden sind, zu umreißen und mit einer einheitlichen Wortmarke zu belegen (auch Faktorenanalysen in diesem Bereich führen den Intelligenzbegriff nicht über den Status einer intervenierenden Variable hinaus, sie führen nur zu einer mit gewissen Annahmen belasteten Systematisierung). Durch weitere Forschung kann aber „Intelligenz" genauer analysiert werden, indem z. B. Annahmen über physiologische Vorgänge systematisch in Betracht gezogen werden; so kann man zu der theoretisch begründeten und empirisch überprüfbaren Hypothese gelangen, daß „Intelligenz" mit „Informationsverarbeitungsgeschwindigkeit" (also bestimmten Parametern des physiologischen Substrates) zu tun hat. Da es nun möglich ist, diese auch unabhängig von Intelligenztests zu messen (z. B. durch evozierte Potentiale [J. Ertl, 1966]), bekommt „Intelligenz",

[23] Die Unterscheidung zwischen deskriptiven und explikativen Persönlichkeitskonstrukten, wie sie bei Th. Herrmann (1969) oder R. Meili (1971) zu finden ist, stimmt nur bedingt mit der hier versuchten Definition überein.

Abb. 3:
Die Unterscheidung zwischen hypothetischem Konstrukt und intervenierender Variabler



interpretiert als Informationsverarbeitungsgeschwindigkeit, den Status eines hypothetischen Konstrukts. Dieses Vorgehen läßt sich nun ausbauen, wenn man über die Funktionen, die für den Ablauf intelligenten Verhaltens verantwortlich gemacht werden können, Modelle bildet und deren theoretisch bedeutsamen Parameter meßbar macht.

Im gegenwärtigen Schrifttum ist trotz Bemühungen um eine einheitliche Sprachverwendung ein wechselweiser Gebrauch beider Bezeichnungen festzustellen, bzw. ist ein Trend dahingehend ersichtlich, daß nur mehr von hypothetischen Konstrukten gesprochen wird und damit ein höherer Wissensstatus vorgegeben wird, als er tatsächlich auf Grund der bisherigen Theorienbildung angenommen werden kann.

Dieser Trend, in der Psychologie von „Konstrukten" zu reden, hat auch seinen Hintergrund in der vagen und theoretisch unbedeutenden Hypothesenformulierung (P. E. Meehl, 1967) und Hypothesenüberprüfung. Besonders in dem Genügen mit dem Finden von korrelativen Beziehungen zeigt sich die Tendenz, fundierter Theorienbildung aus dem Wege zu gehen (vgl. G. Jüttemann, 1972).

3.3. Operationale Definition

H. Hörmann betont (1961, S. 49), daß es als Vorteil der Konstruktvalidierung angesehen werden könne, daß die Konstrukte nicht auf Beobachtbares reduziert werden, sondern daß ihre Bedeutung ermittelt wird, indem sie in ein Netz gesetzesartiger Aussagen eingebettet werden. Reduktion auf bloß Beobachtbares aber werde betrieben, indem man operational definiere[24]. Da dem Operationalismus als wissenschaftstheoretisches Konzept in einer Wissenschaft wie der Psychologie, in der methodische Fragen viel diskutiert werden, einige Bedeutung zukommt, seien einige Anmerkungen dazu erlaubt.

Zuerst erhebt sich die Frage, ob durch die Zuordnung von Meßvorgängen zu Konstrukten diese auch tatsächlich definiert werden. Unter dem Begriff Definition wird üblicherweise eine Entsprechung auf der Zeichenebene, also zwischen zwei oder mehreren Zeichen verstanden[25]. Mit diesen Zeichen werden Konstrukte, wissenschaftliche Ideen, Theorien usw. bezeichnet. In einer faktischen oder empirischen Wissenschaft kommt solchen Begriffen eine besondere Bedeutung zu, die eine Entsprechung mit der Wirklichkeit aufweisen können[26]. Die Beziehung, die zwischen Zeichen, Bezeichnetem und dem dem Bezeichnetem „physikalisch" Entsprechenden kann man nach einem Vorschlag von M. Bunge (1967 a, S. 58) als Denotation bezeichnen (vgl. Abb. 4). Eine mögliche Art, das mit dem Zeichen Gemeinte und die Realität in Beziehung zu bringen, kann durch sog. operationale Definition oder vielmehr: operationale Referition erfolgen. Durch die Zuordnung von Meßvorgängen zu einem theoretischen Konzept

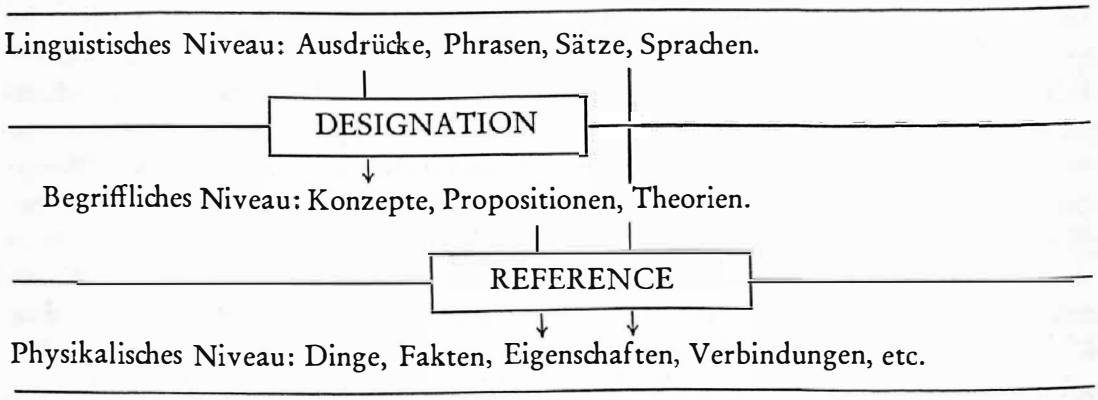
[24] Über die Thesen des Operationalismus (vgl. P. Bridgeman, 1927) und seinen Einfluß auf verschiedene Wissenschaften gibt es eine ausführliche Literatur (J. Klüver, 1971); die Ausführungen hier sind an das Konzept von M. Bunge (1967, 1972) angelehnt.

[25] Definitionen sollen dabei einer Reihe von Kriterien genügen, deren wichtigste in der Forderung nach Nicht-Kreativität zu sehen ist.

[26] Die Einführung von dem Konzept von Geistern als intervenierende Variable (zur Erklärung bestimmter Beobachtungstatsachen) ist z. B. für eine faktische Wissenschaft bedeutungslos; damit ist aber nicht gesagt, daß die Aussage, daß es Menschen gibt, die an Geister glauben, bedeutungslos ist. Es könnte sogar so sein, daß die Suche nach einer Erklärung für diese Aussage ein wichtiges Anliegen einer empirischen Disziplin, z. B. der Psychopathologie, sein kann.

wird aber im strengen Sinn nicht definiert, sondern dieses Konzept bekommt eine mögliche empirische Interpretation (die ihrerseits auch eine empirische Überprüfung von Hypothesen ermöglicht).

Abb. 4:
Die Beziehung zwischen linguistischem, begrifflichem und physikalischem Niveau (DENOTATION)



Mit einem Ausdruck (z. B. Tisch, table etc.) wird ein Begriff bezeichnet (z. B. die Idee eines Tisches) (=Designation). Ausdrücke und Begriffe können sich wieder Entsprechungen in der Wirklichkeit aufweisen (z. B. ein tatsächlich vorhandener Tisch ist ein „Referent" sowohl für das Zeichen als auch den Begriff Tisch) (=Reference).

Die hier vertretene Ansicht über den Operationalismus steht damit in Widerspruch zu der ursprünglichen von P. Bridgeman, der ja die extreme Meinung vertreten hatte: „If we had more than one set of operations we have more than one concept, and strictly there should be a seperate name to correspond to each different set of operations." (1927, S. 10). Wollte man diesen Standpunkt wirklich teilen – selbst P. Bridgeman hat ihn in dieser extremen Form nicht beibehalten –, so hätte man, wenn man die Länge eines Gegenstandes durch verschiedene Meßmethoden erfaßt (einmal mittels Maßband, einmal durch Winkelbestimmung etc.) immer verschiedene Begriffe definiert[27]. Es scheint aber doch so zu sein, daß unabhängig von den verschiedenen Meßmethoden diese demselben Konzept, nämlich dem Längenkonzept, zugeordnet sind, und daß der verwendete Längenbegriff der gleiche bleibt[28].

3.4. Empirische Erfassung von Konstrukten

Geht man nicht so weit, wie die Philosophie des Operationalismus meint, so bleibt

[27] Auch die Tatsache, daß die Resultate von Meßvorgängen, die operational nicht identisch sind, die gleichen sind (wie z. B. die Messung der Temperatur mit einem Quecksilber- und einem Alkoholthermometer) würde von einem strengen Operationalisten – wie von C. G. Hempel (1966, S. 92) kritisiert wird – nicht als Grund dafür anerkannt werden, daß dasselbe Konzept erfaßt wird.

[28] Weitere Punkte der Kritik an dem wissenschaftstheoretischen Konzept des Operationalismus sind bei M. Bunge zu finden (1967 a, S. 148). Positiv hervorzuheben ist aber der begründete Empirismus, der durch operationales Vorgehen gesichert werden soll, u. zw. durch (1) Vermeidung verbaler Definitionen, (2) empirische Interpretierbarkeit (allerdings nicht aller) wissenschaftlicher Termini und (3) der empirischen Prüfbarkeit (der meisten) wissenschaftlichen Hypothesen.

immer die Frage offen, ob ein Konzept durch adäquate Meßmethoden erfaßt wird oder nicht. Man kann sich fragen, ob tatsächlich die Länge eines Gegenstandes erfaßt wird, wenn man diesen abwiegt. Aber auf ein solches Verfahren wird niemand verfallen, denn bei dem Beispiel handelt es sich um einen Beobachtungsbegriff, dessen Messung auf der Hand liegt.

Schwierigkeiten ergeben sich dann, wenn es sich um nicht direkt Beobachtbares handelt (Trieb, Information, Bedeutung, Intelligenz). Nicht direkt Beobachtbares kann dabei nur erfaßt werden, wenn Hypothesen über den Zusammenhang zwischen Beobachtungstatbeständen und den nicht direkt beobachtbaren Konzepten vorliegen. Dies setzt aber wieder ausgearbeitete Theorien voraus, welche eine sachgemäße Ableitung von einer Skala, der Skaleneinheit, dem Index für das nicht direkt beobachtbare Konzept, von numerischer Quantifizierung und schließlich der eigentlichen Messung erlaubt (M. Bunge, 1972). Die mathematische Meßtheorie allein kann diese Aufgabe nicht leisten und erhebt diesen Anspruch auch nicht, da die genannten Auswahlen durch die Sache begründet sein müssen.

Die Problematik diagnostischer Verfahren besteht aber nun gerade darin, daß Meßverfahren vorliegen, für die eine theoretische Erklärung gesucht wird; man weiß nicht, welchen Konstrukten diese Meßverfahren zugeordnet werden sollen. Daher wird versucht, die Theorienkonstruktion vom anderen Ende her aufzubauen, d. h. man beginnt zu messen, ohne genügend explizite Hypothesen über das Ziel der Messung zu haben.

Prinzipiell unterscheiden sich Messungen in der Psychologie aber nicht von solchen in entwickelteren Wissenschaften wie z. B. der Physik (wenn auch eine Reihe spezifischer technischer Probleme damit verbunden sind). Wird operationales Vorgehen nur in dem eingeschränkten Sinn verstanden, daß wissenschaftliche Begriffe und Hypothesen mögliche empirische Interpretationen erfahren und dadurch empirisch überprüfbar werden, so kommt dem auch in der psychologischen Diagnostik überragende Bedeutung im Sinne der Konstruktvalidierung zu. Diagnostische Verfahren sind – so verstanden – Operationalisierungsversuche von Theorien oder Teilen von Theorien; also theoretischen Begriffen zugeordnete Meßverfahren, die mehr oder minder genau sein können und besser oder schlechter ein theoretisches Konzept treffen können. Durch diese Meßverfahren bekommen diese theoretischen Begriffe empirische Bedeutung[29], sie werden dadurch schärfer formuliert und ihr empirischer Gehalt vergrößert.

Die Frage der Validität diagnostischer Verfahren hat aber ihre Grenze in der Vagheit der formulierten Hypothesen, der Konsistenz und dem empirischen Gehalt von Theorien und der den Theorien entsprechenden Modellvorstellungen. Solange solche explizit formulierten Theorien umfassender Art in der Persönlichkeitsforschung aber nicht vorliegen, hat es wenig Sinn, von den Möglichkeiten einer Konstruktvalidierung zu sprechen. Für solche Theorien aber mit beschränktem Aussagebereich, die bisher klar formuliert wurden, ist es relativ leicht, Meßmethoden zu entwickeln, die den Kriterien der Konstruktvalidität entsprechen. Z. B. müßte es relativ leicht sein, Verfahren zu konstruieren, welche Aussagen über Zuflußkapazität und Zuflußgeschwindigkeit für bestimmte Modalitäten erlauben, da es sich hier um „geklärte" Begriffe handelt, die einen festen Platz in der Informationstheorie

[29] Weitere Formen der Bedeutungsfindung sind u. a. damit gegeben, daß diese Begriffe untereinander verbunden sind, indem sie in Form von Gesetzesaussagen aufeinander bezogen sind.

besitzen. Ungleich schwieriger ist es aber, ein Verfahren ad hoc zu entwerfen, um einen Begriff wie „Angst" oder „Aggressivität" meßbar zu machen. Bei den letzteren handelt es sich um alltagssprachliche Begriffe, die theoretisch wenig abgeklärt sind, über die es zwar sehr viele verbal formulierte Behauptungen gibt, aber nicht ein: einheitliche Theorie, in welcher klar unterschieden ist zwischen Grundbegriffen, Axiomen und daraus ableitbaren Folgerungen. Die Verfahren, die aber dazu vorliegen, sind nur vorläufige Vorschläge, die der Theorienbildung weiterhelfen sollen, wobei es immer schwierig sein wird, die Frage zu beantworten, ob man zuerst mit empirischen Untersuchungen mit mehr oder minder ungeklärten Voraussetzungen beginnen soll oder mit profunder Schreibtischarbeit.

3.5. Die Verwendung der Faktorenanalyse zur Konstruktvalidierung

Die Methode der Faktorenanalyse wird von manchen Autoren als bevorzugtes Mittel der Konstruktvalidierung angesehen (G. A. Lienert, 1967; L. Michel, 1964; J. P. Guilford, 1954); auch dabei wäre zu untersuchen, was damit gemeint sein könnte.

Man kann zwei Fälle unterscheiden: *Faktorielle Validität* kann einem diagnostischen Verfahren zugesprochen werden, das an eine Theorie angelehnt ist, welche durch Faktorenanalysen zu bewähren versucht wurde. Die Tests müßten dabei den Modellannahmen der Theorien entsprechen, von denen sie Operationalisierungen sind[30]. Konstruktvalidität kommt diesen Verfahren also in dem Ausmaß zu, in welchem sie geeignet sind, diese Modellannahmen zu erfüllen bzw. in dem Ausmaß, in welchem die diesen Tests zugrunde liegenden Persönlichkeitstheorien adäquate Abbildungen der Wirklichkeit sind. Von faktorieller Validität könnte man aber auch dann sprechen, wenn nur bekannt ist, daß von den Testitems Faktorenanalysen gerechnet wurden. In diesem eingeschränkten Fall könnte man wieder unterscheiden, ob die Faktorenanalyse verwendet wurde, um eine Itemselektion zu treffen und um homogene Subskalen zu formulieren oder ob man damit eine reine Pflichtübung absolvierte, welche für die Testkonstruktion keine weiteren Folgen mehr hatte. Der letztere Fall disqualifiziert sich selbst. Aber auch für die beiden anderen müßten eigentlich alle die Gründe genannt werden, welche die Methode der Faktorenanalyse als suspekt erscheinen lassen, da sehr oft nicht alle methodischen Voraussetzungen vorhanden waren, wie sie z. B. bei K. Überla zusammengestellt sind, oder weil die Modellannahmen der Faktorenanalyse von ihr selbst nicht erfüllt werden können (H. Kallina, 1967; G. Fischer, 1967; K. Th. Kalveram, 1970).

4. Zusammenfassung

Die Frage, was psychologische Verfahren eigentlich erfassen, hat zu verschiedenen Möglichkeiten der Suche nach der Bedeutung von Testverfahren geführt. Im wesentlichen kann man dabei drei Gruppen von Begriffen unterscheiden, welche diese Bemühungen kennzeichnen können, u. zw. Analyse der Gültigkeit eines diagnostischen Verfahrens durch Feststellen von (1)

[30] Es soll dabei nicht außer acht gelassen werden, daß die tatsächlich konstruierten Verfahren nur eine Möglichkeit der Operationalisierung sind, m. a. W. die Tests sind nicht mit der Persönlichkeitstheorie gleichzusetzen („... the usual operational referent of a concept may be only one of several possible" [R. B. Cattell, 1957, S. 334]).

inhaltlicher, logischer und face-Validität, durch (2) empirische Bewährungsuntersuchungen und durch (3) Begriffs- oder Konstruktvalidität. Diese drei Gruppen von Validitätsbegriffen spiegeln dabei die Entwicklung der Theorienbildung in einer empirischen Wissenschaft wider.

Für die Psychologie ist es momentan charakteristisch, daß sich das Hauptgewicht der Validierungsbemühungen auf Feststellungen im Bereich der empirischen Validität beziehen. Die Psychologie versucht damit den Ansprüchen, welche von den verschiedenen „Konsumenten" psychologischen Wissens gestellt werden, nachzukommen, u. zw. indem mittels statistischem Methodeninventar möglichst genaue Vorhersagen geleistet werden.

Diese Untersuchungen haben allerdings nur beschränkte theoretische Bedeutung. Durch vermehrte Theorienbildung, d. h. Systematisierung des vorliegenden Beobachtungswissens und Ordnung von diesem in Grundbegriffe, deren Verbindung in empirischen Gesetzmäßigkeiten und Einordnung in übergreifende Hypothesensysteme (das ist Anliegen der Konstruktvalidierung), könnte die Grundlage für besser fundierte Verfahren gerade im Bereich der Persönlichkeitspsychologie geschaffen werden. Solange aber scharfe Hypothesen und relevante Grundbegriffe weitgehend fehlen oder aber solange man glaubt, daß die Verfügung über bestimmte Analysetechniken das Nachdenken ersetzen kann, wird eine Validitätsbestimmung im Sinne der Konstruktvalidität illusorisch sein.

Ob die aus der Grundlagenforschung ableitbaren diagnostischen Verfahren üblichen Papier- und Bleistifttests entsprechen werden, kann man erst ersehen, wenn man an die Überprüfung einzelner durch Theorienbildung als relevant erachteter Merkmale und Begriffe geht.

Derzeit muß man für die praktische Anwendung zu den Mitteln greifen, die durch empirische Bewährungsuntersuchungen bereitgestellt werden, d. h. der Information über Gültigkeitskoeffizienten und das Wissen, welches – in meist unsystematischer Form – über die Anwendung und Interpretationsweite von diagnostischen Verfahren vorliegt. Auch hier bestünde ein Ansatz für weitere Formulierung von einschränkenden Bedingungen, unter denen der Einsatz diagnostischer Verfahren sinnvoll erscheinen kann. Solche Regeln könnten z. B. auch gefunden werden, indem überprüfte Hypothesen aus der Sozialpsychologie in die psychologische Diagnostik systematisch eingebaut werden. (Die Formulierung solcher Regeln könnte auch zu einer realistischeren Beurteilung der Ansprüche, die durch die Praxis an die Psychologie gestellt werden, führen.)

Abschließend kann man auch konstatieren, daß die Diskussion von Validitätsfragen für solche Wissenschaften kennzeichnend ist, deren theoretisches Standardwissen noch nicht sehr gesichert ist.

LITERATUR

Adam, J., Scharf, J.-H., Enke, Methoden der statistischen Analyse in Medizin und Biologie. Stuttgart 1971.

Anastasi, A., Some current developments in the measurement and interpretation of test validity, in: Gronlund, N. E. (Hrsg.), Readings in measurement and evaluation. New York–London 1968 a, S. 182–192.

–, Psychological testing. New York 1968 b.

APA: Technical recommendations for psychological tests and diagnostic techniques, in: Psychol. Bull.; Suppl. 51 (1954) S. 201–238.

Bartenwerfer, H.: Allgemeine Leistungstests, in: Heiß, R. (Hrsg.), Psychologische Diagnostik, Handbuch der Psychologie, Bd. 6, Göttingen 1964.

Bartussek, D., Eine Methode zur Bestimmung von Moderatoreffekten, in: Diagnostica 16 (1970), S. 57–76.

Bechtoldt, H., Construct validity: A critique, in: American Psychologist 14 (1959), S. 619–629.

Bloch, R. (Hrsg.), Bild und Persönlichkeit. Der mehrdimensionale Zeichentest (MDZT). Bern–Stuttgart 1971.

Brickenkamp, R., Test d 2 Aufmerksamkeitsbelastungstest. Göttingen 1962.

Bridgeman, P., The logic of modern physics. New York 1927.

Bunge, M., Scientific Research, I (a), II (b). Berlin–Heidelberg–New York 1967.

–, On confusing 'measure' with 'measurement' in the methodology of behavioral science. Unpublished paper. McGill University, Montreal 1972.

Campbell, D. T., Recommendations for APA test standards regarding construct, trait, or discriminant validity, in: American Psychologist 15 (1960), S. 546–553.

Cattell, J. McKeen, Mental tests and measurement, in: Mind 15 (1899), S. 373–381.

Cattell, R. B., Personality and motivation, structure and measurement. New York 1957.

Cooley, W. W., Lohnes, P. R., Multivariate data analysis. New York–London–Sydney–Toronto 1971.

MacCorquodale, K., Meehl, P. E., On a distinction between hypothetical constructs and intervening variables, in: Psychological Review 55 (1948), S. 95–107.

Cronbach, L. J., Essentials of psychological testing. New Lork–London–Tokyo 1969 (1949).

Cronbach, L. J., Gleser, G. C., Psychological tests and personnel decisions. Chicago–London 1965.

Draper, N. R., Smith, H., Applied regression analysis. New York–London–Sydney 1966.

Drenth, P. J. D., Der psychologische Test. München 1969.

Ebel, R., Must all tests be valid?, in: American Psychologist 16 (1961), S. 640–647.

Ekman, G., Konstruktion and Standardisierung von Tests, in: Diagnostica 1 (1955), S. 15–19.

Ertl, J., Evoked potentials and intelligence, in: Revue de l'Université d'Ottawa 36 (1966), S. 599–607.

Eysenck, H. J., Probleme der diagnostischen Untersuchung und Demonstration des Charakter-Interpretationstests. Göttingen 1954.

Fischer, G. (Hrsg.), Psychologische Testtheorie. Bern–Stuttgart 1968.

–, Zum Problem der Interpretation faktorenanalytischer Ergebnisse, in: Psychologische Beiträge 10 (1967), S. 122–135.

Gasch, B., Erfolg im Psychologie-Studium. Erlangen 1970 (unveröff. Diss.).

Ghiselli, E. E., Moderating effects and differential reliability, in: Goodstein, L. D., Lanton, R. I. (Hrsg.), Readings in personality assessment. New York–London–Sydney–Toronto 1971, S. 700–707.

Goldman, L., Using tests in counseling. New York 1961.

Goodstein, L. R., Slovic, P., Importance of test item content: an analysis of a corollary of the deviation hypothesis, in: Goodstein, L. R., Lanton, R. I. (Hrsg.), Readings in personality assessment. New York–London–Sydney–Toronto 1971, S. 253–267.

Guilford, J. P., Psychometric methods. New York–Toronto–London 1954.

Gulliksen, H., Theory of mental tests. New York 1950.

Hartmann, H., Psychologische Diagnostik. Stuttgart–Berlin–Köln–Mainz 1970.

Hempel, C. G., Philosophy of natural science. London–Sydney–Toronto–New Delhi–Tokyo 1966.

Herrmann, Th., Lehrbuch der empirischen Persönlichkeitsforschung. Göttingen 1969.

Holzkamp, K., Jäger, A. O., Merz, F., Prognose und Bewährung in der psychologischen Diagnostik. Göttingen 1966.

Hörmann, H., Aussagemöglichkeiten psychologischer Diagnostik. Göttingen 1964.

–, Zur Validierung von Persönlichkeitstests, insbesondere von projektiven Verfahren, in: Psychologische Rundschau 12 (1961), S. 44–49.

Horst, P., Messung und Vorhersage. Weinheim–Berlin–Basel 1971.

Jenkins, J. G., Validity for what? In: Barnette, W. L. (Hrsg.), Readings in psychological tests and measurement. Homewood 1964, S. 155–161.

Jüttemann, G., Was nützen uns Eigenschaftskonstrukte?, in: Psychologische Rundschau 23 (1972), S. 91–114.

Kallina, H., Das Unbehagen in der Faktorenanalyse, in: Psychologische Beiträge 10 (1967), S. 81–87.

Kalveram, K. Th., Über Faktorenanalyse, in: Archiv für die gesamte Psychologie 122 (1970), S. 92–118.

Kleiter, G., Krise der Signifikanztests in der Psychologie, in: Jahrbuch für Psychologie, Psychotherapie und medizinische Anthropologie 17 (1969), S. 144–163.

Klüver, J., Operationalismus. Kritik und Geschichte einer Philosophie der exakten Wissenschaften. Stuttgart–Bad Cannstatt 1971.

Lienert, G. A., Die Konfigurationsfrequenz-

analyse, in: Zeitschrift f. Klinische Psychologie u. Psychotherapie 19 (1971) S. 99–115.
–, Testaufbau und Testanalyse. Weinheim–Berlin 1967.
–, Mechanisch-technischer Verständnistest. Göttingen 1958.
–, Form-Lege-Test. Göttingen 1964.
–, Drahtbiegeprobe als standardisierter Test. Göttingen 1961.
Lord, F. M., Novick, M. R., Statistical theories of mental test scores. Reading Menlo Park–London–Don Mills 1968.
Lossen, H., Einführung in die diagnostische Psychologie. Stuttgart–Bad Cannstatt 1955.
Magnusson, D., Testtheorie. Wien 1969.
Matussek, P., et al., Die Konzentrationslagerhaft und ihre Folgen. Berlin–Heidelberg–New York 1971.
Meehl, P. E., Theory-testing in psychology and physics: A methodological paradox, in: Philosophy of science 34 (1967), S. 103–115.
Meili, R., Lehrbuch der psychologischen Diagnostik. Bern–Stuttgart 1965.
–, Deskriptive und erklärende Persönlichkeitseigenschaften, in: Zeitschrift f. exp. u. angew. Psychol. 18 (1971), S. 621–628.
Michel, L., Allgemeine Grundlagen psychometrischer Tests, in: Heiss, R. (Hrsg.): Psychologische Diagnostik. Handbuch d. Psychologie, Bd. 6, Göttingen 1964.
Oswald, W. D., Persönlichkeit und Kraftfahreignung. Stuttgart 1971.
Popper, K. R., Das Elend des Historizismus. Tübingen 1971.
–, Logik der Forschung. Tübingen 1960.
Revers, W. J., Taeuber, K., Der thematische Apperzeptionstest (TAT). Bern–Stuttgart 1969.
Rohracher, H., Einführung in die Psychologie. Wien 1971.
Roth, E., Persönlichkeitspsychologie. Stuttgart–Berlin–Köln–Mainz 1969.
–, Oswald, W. D., Daumenlang, K., Intelligenz. Stuttgart–Berlin–Köln–Mainz 1972.
Sader, M., Möglichkeiten und Grenzen psychologischer Testverfahren. Bern–Stuttgart 1961.
Schlange, H., Stein B., Boettcher, I. v., Taneli, S., Der Göttinger Formreproduktions-Test. Beiheft zur Handanweisung. Göttingen 1972.
Stegmüller, W., Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Bd. I, Wissenschaftliche Erklärung und Begründung. Berlin–Heidelberg–New York 1969.
Stelzl, I., Definitionsmöglichkeiten von Persönlichkeitseigenschaften, in: conceptus 4 (1970) S. 49–52.
Stern, W., Die differentielle Psychologie und ihre methodischen Grundlagen. Leipzig 1921.
Toda, M., Dynamic decision theory: a study of human planned control activities. In: Abstract Guide XXth International Congress of Psychology. Tokyo 1972.
Überla, K., Faktorenanalyse. Berlin–Heidelberg–New York 1971.
Weingartner, P., Wissenschaftstheorie I. Einführung in die Hauptprobleme. Stuttgart–Bad Cannstatt 1971.
Wellek, A., Exploration und ganzheitliches Verfahren, in: Psychologische Rundschau 9 (1958), S. 24–28.
Witte, W., Zu den Beziehungen zwischen praktischer Psychologie und psychologischer Grundlagenforschung, in: Psychol. Beiträge (1966), S. 368–377.

Dr. H. Lukesch
Universität Konstanz
Fachbereich Erziehungswissenschaften
D–7750 Konstanz
Jacob-Burkhardt-Straße