Konstruktion und methodenbewusste Anwendung von Persönlichkeitsfragebogen

Jochen Fahrenberg, Freiburg i.Br.

Vorbemerkung
Der folgende Text ist die gekürzte und modifizierte Fassung eines der neuen Kapitel in der 8. Auflage des Freiburger Persönlichkeitsinventars FPI-R (2009).

1 Konstruktion und Evaluation von Persönlichkeitsinventaren – kritische Differenzierungen

Zeitweilig geschah eine lebhafte Entwicklung neuer Ideen für die Konstruktion und Evaluation von psychologischen Tests: Multitrait-Multimethod-Analyse, Faktorenanalyse, Clusteranalyse, multidimensionale Skalierung, Rasch-Skalierung und andere Item-Response Modelle, Latent Class Modelle, Entscheidungsnutzen und andere Prinzipien der modernen Assessmenttheorie, Multimodale Diagnostik, Aggregationstechniken, Multi-Level-Analyse, Generalisierbarkeitstheorie und Symmetrieprinzipien, Multivariate Reliabilitätstheorie, Prinzipien der Qualitätssicherung.

Wie steht es gegenwärtig mit methodologischen Innovationen in diesem Überlappungsbereich der Differentiellen Psychologie, der Konstruktion und Evaluation von Tests sowie der grundlegenden Assessmenttheorie? Der Blick richtet sich natürlich auf die neueren Lehr- und Handbücher, die das Grundwissen über die Konstruktion und Evaluation von Tests und Fragebogen vermitteln wollen (u.a. Amelang & Schmidt-Atzert, 2006; Borg & Staufenbiel, 2007; Bühner, 2006; Jäger & Petermann, 1995; Kubinger & Jäger, 2003; Moosbrugger & Kevala, 2007; Petermann & Eid, 2006; Rost, 2004). – Die Weiterentwicklung formaler Messmodelle bzw. Modellierungen ist nicht zu übersehen, doch sind diese für *Persönlichkeitsfragebogen* nur sehr bedingt geeignet. Im übrigen fällt die überzogene Auseinandersetzung über die „richtige“ Anzahl von Faktoren in Persönlichkeitsfragebogen auf. Demgegenüber scheinen andere Forschungsansätze, die für die Konstruktion und praktische Evaluation von Fragebogen und Tests ungleich wichtiger wären, fast zu stagnieren. In einigen der Lehrbücher werden interessante Themen und Kontroversen der neueren Testmethodik kaum erwähnt. Die folgenden Kommentare sind z.T. thesenartig formuliert und können an dieser Stelle nur auf ausgewählte Probleme und Entwicklungen aufmerksam machen.

2 Skalierung: Kontroversen über Messtheorie

Zwischen den Lehrbüchern der Testtheorie und Testkonstruktion scheint eine große Übereinstimmung zu bestehen: Die Daten von Persönlichkeitsfragebogen (und Stimmungsskalen) werden als Intervalldaten angesehen. Einige scheinen hier nur eine pragmatisch-bequeme und harmlose Konvention im Sinne „des Üblichen“ zu sehen. Andere übertragen ohne offensichtliche Bedenken ihre messtheoretischen Überzeugungen aus dem Bereich der objektiven Intelligenz- und Leistungstests auf die introspektiven Selbstbeurteilungen. Die Voraussetzungen und die Konsequenzen dieses Postulats werden nur sehr selten diskutiert. Zwar gibt es kompetente Darstellungen der Messtheorie hinsichtlich Repräsentation, homomorpher Abbildung, Eindeutigkeit, Deutbarkeit, Skalentheorie, doch bleiben diese Konzepte abstrakt (u.a. Borg & Staufenbiel, 2007; Orth, 1983; 1995; Yousfi & Steyer, 2006). Aber was bedeutet die Forderung, die resultierenden Testwerte sollten die empirischen Merkmalsrelationen adäquat abbilden, d.h. in adäquate Zahlenrelationen transformieren? Ein Bezug zu den unterschiedlichen psychologischen Datenquellen und ihren speziellen Verhältnissen wird kaum hergestellt. – Könnte es sich bei der „Messung“ von Introspektionen und Selbstbeurteilungen um „Messung durch willkürliche Festlegung“, nur um ein „numerisches Etikettieren“ handeln? Welche Konsequenzen ziehen solche Postulate über die Passung von Selbstbeurteilungen und Mess-Struktur nach sich? Können solche Aussagen psychologisch adäquat sein? – Bemerkenswert ist diese Zurückhaltung der Testtheoretiker schon, denn es entspräche ja der Position des kritischen Rationalismus im Sinne Stegmüllers (1973, S. 44), solche fundamentalen Voraussetzungen auf der Metaebene ebenfalls zum Thema einer rationalen Rechtfertigungsdebatte der Fachwissenschaftler zu machen.

Skeptische Stimmen wie von Michel und Conrad (1982) und die Zweifel, ob Interpretationen angemessen sind, die über das Ordinalskalenniveau hinausgehen, sind heute selten. Auch Krauth (1995) äußert sich nur indirekt. Er definiert Items als Reize, auf die Reaktionen erfolgen, stellt jedoch später fest: „Items auf Intervallskalenniveau werden in der Psychologie so gut wie nie verwendet“ (S. 32). Da die zugrundliegenden (latenten) Eigenschaftsdispositionen nie direkt beobachtbar sind, sei es nicht sinnvoll für solche Variablen überhaupt ein Skalenniveau zu definieren, doch sei genau zu überlegen, ob die latenten und die manifesten Variablen in einem Modell verknüpft werden sollten, wenn keine eindeutigen Beziehungen angenommen werden können. In allgemeiner Weise distanziert sich Krauth von jenen Autoren, „die leugnen, dass man bei Anwendung statistischer Verfahren auf das Skalenniveau Rücksicht nehmen müsse“ (S. 34). Er folgt den Ansätzen, die Items auf Ordinalskalenniveau definieren, ohne jedoch zwischen Intelligenz- und Leistungstests

und Persönlichkeits- und Stimmungsskalen abzugrenzen. Dass diese messtheoretischen Entscheidungen beliebig wären, ist auch dann nicht anzunehmen, wenn vorsichtig formuliert wird: "Die Skalenqualität einer Messung ist also letztlich von theoretischen Entscheidungen, d.h. von Interpretationen abhängig" (Bortz, Lienert & Boehnke, 2000, S. 66).

Borg und Staufenbiel (2007) meinen, dass das Skalenniveau der Ausgangsdaten nicht *vorab* empirisch oder argumentativ begründet werden müsse, es käme nicht darauf an, ob das Skalenniveau „wahr" sei, sondern ob das Messmodell nützlich ist. Das Skalennniveau wird also zugewiesen aufgrund von Hypothesen, wie die erhaltenen Werte mit anderen Beobachtungen zusammenhängen (S. 7). Andererseits sei die Frage der Darstellbarkeit von Daten mit besonderen Eigenschaften nicht trivial. Aufgrund der Strukturgleichheit zwischen dem empirischem und dem numerischem Relativ besteht die Aussicht, dass sich die Ergebnisse der Berechnungen zuverlässig auf die Empirie rückübersetzen lassen (S. 392). Eine latente Variable, welche die Beschreibung eines komplexen Sachverhalts auf eine formal bzw. mathematisch relativ einfache Weise beschreibt, gilt hier als „Erklärung". Offensichtlich ist *nicht* die u.a. von Dawes (1977, siehe auch Dawes, Faust & Meehl, 1989) vertretene *pragmatische* Auffassung gemeint, in solchen Zahlenzuweisungen nur Indizes zu erkennen, die mehr oder minder nützlich sein können (index measurement).

Rost (2004) begründet Messmodelle allgemein, indem er sich auf Verhaltensaussagen des Typs „A ist intelligenter als B" bezieht und fragt, wie sich diese theoretische Aussage interpretieren lässt. „Man benötigt hierfür ein *formales Modell*, das in Form einer mathematischen Gleichung den angenommenen Zusammenhang zwischen der Wahrscheinlichkeit des Auftretens der Verhaltensweisen ..(…).. und der Personeneigenschaft .. (…) .. sowie den Situationsmerkmalen .. (…) .. beschreibt. Ein formales Modell ist notwendig, weil sonst nicht über die Gültigkeit der Theorie und somit die Wissenschaftlichkeit der Aussagen entschieden werden kann" (S. 24-25). Ein Modell wird erst mit der Schätzung der Modellparameter zu einer Theorie für den betreffenden Inhalt. Welches Modell soll ausgewählt werden? „Natürlich dasjenige, welches die *Annahmen der jeweiligen Theorie* am besten widerspiegelt und welches diejenigen Aspekte der Wirklichkeit abzubilden vermag, die mit dem Test erfasst werden sollen" (S. 28). – Ist damit gemeint, dass erst mit der Prüfung metrischer Strukturhypothesen Wissenschaft entsteht? Damit würde völlig übersehen, dass sogar in der Biologie u.a. Naturwissenschaften sowie in der Medizin bestimmte Teildisziplinen ohne metrische Messmodelle auskommen und dennoch grundlegende Erkenntnisfortschritte leisteten. Auf der anderen Seite warnt Rost vor der Anpassung von Modellen, welche die gewünschten Aussagen gar nicht abbilden können. Dieser Merksatz ist allerding hier im aktuellen Kontext doppeldeutig: „Insofern kann sich eine richtig verstandene Testtheorie als größter Kritiker der Testpraxis erweisen" (S. 29).

Die zitierten Autoren begreifen die Messung – im Sinne von Louis Guttman – als Prüfen von Strukturhypothesen, stellen also eine enge Beziehung zwischen Messen und Theorie her. Sie lassen jedoch im Dunklen, was dies speziell für die subjektiven Auskünfte, Selbstbeurteilungen, Befindensweisen und Erlebnisse, d.h. für die Inhalte der am häufigsten verwendeten psychologischen Tests, Persönlichkeitsfragebogen, Stimmungsskalen und Klinische Skalen, bedeuten kann. Die Beispiele werden fast regelmäßig aus anderen Bereichen gewählt. Was für *beobachtbare* Verhaltensmerkmale überzeugen kann, wird hier kommentarlos auf *subjektive* Auskünfte generalisiert, ohne dieses Dilemma der psychologischen Diagnostik aufgrund von Selbstbeurteilungen deutlich zu machen. Auch ein Querverweis auf die Skalierungen der Psychophysik würde nicht viel klären, denn dort ist der Messvorgang durch die physikalische Variation der Stimuli auf besondere Weise strukturiert. „Ohne dass ein Verfahren das Gütekriterium *Skalierung* erfüllt, sind Betrachtungen über Validität (Gültigkeit), Messgenauigkeit und Objektivität eigentlich müßig" postulieren Westhoff et al. im *Grundwissen* (Westhoff et al., 2004, S. 181). Unausgesprochen bleibt der Bezug auf das Ideal der Rasch-Skalierung und die Physik. „… erfüllt ist das Kriterium, wenn die laut Verrechnungsvorschrift resultierenden Testwerte die empirischen Verhaltensrelationen adäquat abbilden". Wie könnte dieses Postulat für Introspektionen, Selbstbeurteilungen und Selbstbeobachtungen des Verhaltens mit Sinn gefüllt werden (vgl. Wittgensteins Argumentation zur Protokollierung von Erlebnissen)? Bemerkenswert ist ein neuer Begriff: Es gibt „nicht-skalierbare" Personen, die den Nachteil haben, mit ihrem Antwortmuster nicht zu den Messmodellen zu passen. Zu diesem Eindruck der Einseitigkeit passt auch, dass regelmäßig die Axiome der Klassischen Testtheorie nachhaltig kritisiert werden, jedoch eine entsprechende Kritik der Item-Response-Modelle sowie Hinweise auf prinzipielle Vorbehalte oft sehr dezent bleiben (siehe z.B. Kubinger, 2002a; Kubinger & Jäger, 2003; Rost, 2002, 2004, 2006, vgl. auch Amelang & Schmidt-Atzert, 2006).

Auf der anderen Seite steht eine nicht geringe Zahl von Psychologen, die psychologische bzw. erkenntnistheoretische Kritik an einer ihres Erachtens unreflektierten Mess- und Testtheorie und einer „pseudo-naturwissenschaftlich" ausgerichteten Psychologie üben. Die uneingeschränkten messtheoretischen Postulate können dogmatisch wirken und provozieren Widerspruch. Diese grundsätzliche Kritik führte zu der deutlich zunehmenden Strömung der „qualitativen" Methodik (vgl. Flick, von Kardorff & Steinke, 2000). Aber der Begriff „qualitativ" ist unglücklich gewählt, weil er vieldeutig und missverständlich ist (Fahrenberg, 2002, 2008b). „Qualitativ" dient vielfach als Etikett einer Auffassung, die sich von der akademischen Mess- und Testtheorie distanziert, Defizite der Argumentation aufzeigt und zugleich eine größere Praxisnähe behauptet. Die psychologische Berufspraxis ist ja zweifellos viel stärker am interpretativen Paradigma ausgerichtet als an dem experimentell-statistischen Paradigma. Besser ist es, von interpretierenden Verfahren im

Unterschied zu metrischen Methoden, Tests und Skalen in der Psychologie zu sprechen. Letztlich müssen natürlich auch experimentelle Befunde, Messmodelle und Skalierungen in einem primär psychologischen Kontext inhaltlich interpretiert, d.h. mit anderem psychologischen Wissen in Beziehung gesetzt werden.

Zu dieser Kontroverse gehören auch philosophisch-erkenntnistheoretische Argumente, dass hier sowohl subjektiv-mentale Phänomene als auch psychologische Eigenschaftskonzepte reduziert werden, ohne die Defizite klar zu legen (vgl. Jüttemann, 1991, 2004). Die fundamentale Kritik an der „Vermessung des Menschen" kann, wenn zur Reduktionismuskritik auch ideologiekritische bzw. gesellschaftskritische Argumente hinzukommen, zu einer weiteren Distanzierung vom sog. Mainstream führen (vgl. Walter, 1999).

Wenn die Kontroverse über die Intervallmessung subjektiver Auskünfte in den meisten Lehrbüchern eine so geringe Rolle spielt, scheint das nicht zu der Betonung der „Skalierung" als wichtiges Gütemerkmal zu passen (u.a. Kubinger, 2003b, Westhoff et al., 2004). Becker (2003b, S. 355) unterstreicht das Problem: „Bei Fragebogen, die nach der klassischen Testtheorie konstruiert wurden und ausgewertet werden, besteht die Gefahr, dass das unterstellte Intervallskalen-Niveau nicht gegeben ist, woraus eine mangelnde Verrechnungs-Fairness und Verzerrungen im Extrembereich der Scoreverteilung resultieren" (siehe auch Borkenau, 2006). – Aber erhält das empirische Relativ der Selbstbeurteilungen schließlich das gewünschte metrische Relativ, indem jetzt nur Messmodelle benutzt werden, die eben dies postulieren müssen? – Die nachhaltige Überzeugtheit der Vertreter und der Kritiker der messtheoretischen Postulate verweist auf Vorentscheidungen, die außerhalb des Messmodells liegen. Empfiehlt sich hier nicht eine mittlere Position, pragmatisch an Heuristik und Nutzen interessiert oder aus einer Einsicht in die Komplementarität solcher kategorial verschiedenen Auffassungen? Hat das von Wilhelm Wundt ausdrücklich geforderte perspektivische Denken (vgl. Fahrenberg, 2008a; Jüttemann, 2006) in der Psychologie immer noch zu wenig Einfluss?

3 Fragwürdige Übertragung messtheoretischer Axiome auf Introspektionen und Selbstbeurteilungen

Grundsätzlicher als bei Intelligenz- und Leistungstests stellt sich das Skalierungsproblem. Typische Ordinaldaten werden gewonnen, wenn ein einzelner kompetenter Beobachter oder eine Gruppe trainierter Beobachter die Ausprägung von Merkmalen beurteilen und ihre Einschätzungen in Rangordnungen von Größer-Kleiner-Beziehungen ausdrücken. Demgegenüber beruhen Persönlichkeitsfragebogen primär auf Selbstbeurteilungen. Weder ist ein direkter Vergleich mit dem Selbstbild und der Befindlichkeit anderer Menschen möglich, noch besteht in der Regel ein methodisches Training. Ob die Einstufungen *faktisch* wiederholbar sind oder ob eine Beurteiler-Übereinstimmung besteht, kann grundsätzlich nicht geprüft werden. Wohl alle Items verlangen implizit eine retrospektive Auskunft und eine Aggregation des gemeinten Persönlichkeitsmerkmals (Gefühle, Befinden, Verhaltensweisen) über nicht näher definierte Zeiträume der Lebensspanne, über Klassen von (auch hypothetischen, vielleicht nie erlebten) Situationen und über Klassen von Detailaspekten, wobei die individuellen Gewichtungen unbekannt bleiben. Selbstbeurteilungen liefern also Nominaldaten (oder intraindividuelle Ordinaldaten besonderer, „ipsativer" Art). Es sind subjektive Schätzverfahren hinsichtlich nicht direkt messbarer Merkmale mentaler Repräsentationen mit unbekanntem numerischem Relativ, in eigentümlichen, vermutlich von Individuum zu Individuum unterschiedlichen, pseudo-numerischen Bezugssystemen, die eventuell auch von Deskriptor zu Deskriptor variieren werden – subjektive Aggregationen und subjektive Metriken.

Wer die Definitionen einer Intervallskala kennt, wird grundsätzlich zweifeln, wenn den mehrstufigen Itemantworten sowie den addierten Testwerten Intervallskalen unterstellt werden. Die Gleichheit der Skalenintervalle ist nicht gegeben und folglich sind die Verhältnisse der Intervalle nicht definiert. Deshalb sind lineare Transformationen und die entsprechenden Rechenoperationen bzw. die Annahmen über die Wahrscheinlichkeitsverteilung der untersuchten Variablen definitionsgemäß unzulässig; auch die simple Addition einzelner**,** heterogener Itemwerte zu einem Skalenwert verletzt die Grundannahme. Über die Konsequenzen dieses Sachverhalts existieren allerdings in der Fachliteratur große Meinungsunterschiede. In der psychologischen Testmethodik und Forschung ist es eine weit verbreitete Gewohnheit, auch diesen – nur als numerisch *erscheinenden* – Selbstbeurteilungen die Qualität von Intervallskalen zuzubilligen, wie es in anderen Bereichen, z.B. bei Intelligenz- und Leistungstests geschieht. Im Sinne eines einheitlichen Messmodells (vgl. die Argumente von Guttman, Rasch und Nachfolgern) ist diese Entscheidung verständlich, zumal sie neben den Strukturhypothesen große Vorteile für die statistischen Analysen mit sich bringen in der Hoffnung auf eine bessere "Informationsausschöpfung". Dennoch bleibt es erstaunlich, wenn sehr anspruchsvolle statistische Strukturanalysen und Modellierungen gerade anhand der metrisch sehr zweifelhaften Selbstbeurteilungen in Fragebogen unternommen werden. Bei dieser "liberalen" Einstellung gehen so viele messtheoretische und psychologische Vorentscheidungen zur Repräsentation von Eigenschaften ein, dass die Argumentation unübersichtlich wird oder ganz unterbleibt. Erst die neuesten Latent Class Modelle enthalten hier einige Fortschritte.

4 Eigenart von Persönlichkeitsfragebogen und Fragebogenmethodik

Einige Lehrbücher zählen Regeln auf, wie Fragebogenitems formuliert werden sollten. Zweifellos kann kaum genug getan werden, die Verständlichkeit zu optimieren – so weit es eben geht. Auffällig ist dagegen, dass der keineswegs minder wichtigen Operationalisierung des gemeinten Konstrukts oft sehr viel weniger Aufmerksamkeit gewidmet wird. Wenn hauptsächlich die Formulierung und äußere Gestaltung der Items erörtert werden, kann das an dem hohen Schwierigkeitsgrad einer Debatte über adäquate Operationalisierungen von Eigenschaftskonstrukten liegen. Sie müsste jedoch exemplarisch geführt werden. Am ehesten geschah dies wohl für Eysencks Sekundärfaktoren E und N.

Persönlichkeitsfragebogen (Persönlichkeitsinventare) sollen Informationen über ausgewählte Bereiche der Persönlichkeit geben: aufgrund von Introspektionen, Selbstberichten, Selbstbeurteilungen, retrospektiven und aktuellen Verhaltensberichten, Selbstbeobachtungen von Verhaltensweisen bzw. biographische Fakten, die grundsätzlich auch anderen Beobachtern zugänglich wären. Die grundsätzlichen Unterschiede zu Intelligenz- und Leistungstests bzw. Verhaltensmessungen dürfen nicht übersehen werden. Diese Unterschiede sind kategorial, denn Selbstberichte und Selbstbeurteilungen, können zwar von Dritten bezweifelt oder im Detail als unzutreffend erkannt werden, bleiben aber in zentralen Bereichen subjektiv und unwiderlegbar. Der Anteil potentiell objektivierbarer Aspekte eines typischen Persönlichkeitsfragebogens ist gering.

Der Begriff „Verhalten“ ist irreführend, wenn nicht kategorial zwischen dem intersubjektiv beobachtbaren, manifesten, aktuellen Verhalten und den sprachlichen Mitteilungen und Selbstbeurteilungen des individuellen Verhaltens unterschieden wird. Auch die verbreitete Unterscheidung von Selbst-Einstufungen und Fremd-Einstufungen kann sehr irreführend sein, wenn nicht geklärt ist, ob diese Fremdbeurteilungen, z.B. von Bekannten oder in einem Interview, sich primär auf die Selbstauskünfte der betreffenden Person in alltäglichen Mitteilungen stützen. Noch schwerer einzuschätzen ist die gemeinsame Varianz solcher Selbst- und Fremdeinstufungen aufgrund alltagspsychologischer Schemata auf beiden Seiten (Asendorpf, 2007; Baumann & Stieglitz, 2008, Kenrick & Funder, 1988). Schon R.B. Cattell hatte darauf gedrängt, zwischen den Datenquellen Behavior *Rating* und Behavior *Measurement* zu unterscheiden.

Fragebogen sind „subjektive Verfahren“. Oft werden eine Kompetenz und eine Bereitschaft zur Selbstbeschreibung, Wissen und sprachliche Fähigkeiten, als Voraussetzungen solcher Persönlichkeitsfragebogen genannt. Eine Voraussetzung dieser Technik, so schreiben Amelang und Schmidt-Atzert (2006, S.241) besteht darin, dass „die Betreffenden sich selbst überhaupt kennen und zu beobachten imstande sind.“ Rost (2004) nennt drei Voraussetzungen: die Einsicht in eigene kognitive Prozesse, die Bereitschaft, das reale Selbstbild zu offenbaren und das Vorhandensein von geeigneten Beurteilungsmaßstäben aufgrund sozialer Vergleichsprozesse. Wie diese individuellen Fähigkeiten empirisch festzustellen sind, bleibt offen. Auch die Unterscheidungen von Selbsttäuschung und Fremdtäuschung oder von offenbartem und privatem Selbstbild mögen plausibel klingen, helfen jedoch ohne diagnostische Mittel, die Anteile überzeugend zu trennen, kaum weiter. Auf Defizite der Selbstbeobachtung und Selbsterkenntnis hinzuweisen (Lösel, 1995) oder den „privilegierter Zugriff“ durch Einschätzungen weiterer Beurteiler ergänzen zu wollen (Borkenau, 2006) befreit nicht von der *strukturellen Subjektivität* solcher Selbstbeurteilungen und Verhaltensberichte – und der Frage, wie die Hinweise auf Antworttendenzen adäquat zu interpretieren wären (siehe die ausführliche Diskussion dieses Themas im Manual des FPI-R.

Die Selbstauskünfte verlangen vielschichtige Urteilsprozesse: Erinnerungen an eigene Gewohnheiten, globale Einschätzungen, wie man sich im Allgemeinen verhalte, einen direkten oder indirekten Vergleich mit Anderen, eine Selbstbeurteilung und Selbstdarstellung. Die erhaltenen Testwerte repräsentieren im Unterschied zu typischen Intelligenz- und Leistungstests komplizierte subjektive und multi-referentielle Rekonstruktionen. Sie sind durch kognitive Schemata und soziale Stereotype, alltagspsychologische Vorstellungen, formale Antwort-Tendenzen und Erwägungen der sozialen Erwünschtheit, Retrospektionseffekte, Urteilsheuristiken u.a. Bedingungen beeinflusst. – Diese Kennzeichnung der Persönlichkeitsfragebogen scheint ihrer weiten Verbreitung zu widersprechen. Doch Selbstbeurteilungen sind am leichtesten zugänglich, einfach, ökonomisch, standardisiert, sie haben eine Augenscheinvalidität. Wer auf diese Selbstbeurteilungen verzichtet, verliert viele – auch durch ein langes Interview – nur bedingt zu ersetzende Informationen.

Persönlichkeitsfragebogen werden in einem *psychometrisch unterlegten Prozess psychologischer Interpretation* konstruiert. Folglich erfordern auch die erhaltenen Testwerte eine Interpretation, und zwar im Kontext aller dienlichen Informationen. Überhaupt werden die Strategien und die traditionellen Regeln der psychologischen Interpretation in den Lehrbüchern und in der akademischen Ausbildung weithin vernachlässigt oder werden als sog. qualitative Verfahren abgespalten, statt einzuräumen, das jedes experimentelle Ergebnis und jeder psychologische Testwert einer *psychologischen Interpretation* bedarf. Deshalb wurde die kurze Interpretation eines Fragebogenprofils am Beispiel des FPI-R als Abschnitt in ein Buch über *Psychologische Interpretation* aufgenommen (Fahrenberg, 2002).

Historische Hinweise

Als psychologie-historische Fußnote ist anzumerken, dass bereits Immanuel Kant in den Befragungen der Menschen einen der Gründe sah, der Psychologie nur den Rang einer empirischen, aber nicht einer exakten Wissenschaft einzuräumen. Zutiefst skeptisch war auch Wilhelm Wundt (1907, 1908) gegenüber Befragungen, während er andererseits die erste methodisch anspruchsvolle Interpretationslehre für psychologische Befunde schuf (Quellenangaben siehe Fahrenberg, 2008b). Wundt lehnte die gerade auftauchende Fragebogenmethodik ab, da den sorgfältigsten und den unzuverlässigen Aussagen gleiches Gewicht beigelegt werde. „Man versendet Bogen mit einer Anzahl Fragen … (…) … an eine möglichst große Anzahl von Personen, sammelt die Antworten und sucht sie statistisch zu verarbeiten. Dass diese Methode lediglich die Mängel der gewöhnlichen, nicht experimentell controllirten Selbstbeobachtung durch die bei ihr unvermeidlichen Missverständnisse, die unterschiedslose Behandlung guter und schlechter, zuverlässiger und unzuverlässiger Beobachter ins Unberechenbare vergrößert, ist an und für sich einleuchtend. Darum sollte man wenigstens die Anwendung derselben auf solche äußere Fragen beschränken, zu deren Beantwortung überhaupt keine psychologischen Beobachtungen erforderlich sind“ (Wundt, 1902-1903, II, S. 275).

Dagegen meinte Oswald Külpe (1920): „Das Vorurteil gegen den Fragebogen beruht auf unzweckmäßiger Anwendung desselben. Er wurde nämlich vielfach an sehr ungleichwertige Personen versandt, enthielt oft Fragen, die sich gar nicht ohne weiteres beantworten ließen und war so reich an Fragen, daß sich die meisten nicht die Mühe nahmen, sie sorgfältig zu beantworten oder ganz darauf verzichteten. Wo aber solche Fehler vermieden werden, kann der Fragebogen recht brauchbare Ergebnisse liefern“ (S. 56-57). Als Beispiel guter Fragebogen nannte Külpe u.a. einen Fragebogen zum Thema Vererbung psychischer Dispositionen von Heymans und Wiersma (1906), in dem Antworthäufigkeiten prozentual ausgewertet wurden. – Von einem konstruierten Persönlichkeitsfragebogen im engeren Sinn ist wohl nur dann zu sprechen, wenn im Unterschied zu ad hoc Fragenlisten außer einer äußeren Standardisierung auch die Auswahl, die Gewichtung und der innere Zusammenhang der einzelnen Fragen beschrieben werden, wie es erst mit der Einführung der Korrelationsrechnung möglich wurde. Als erster Persönlichkeitsfragebogen gilt deswegen Lankes (1915) Publikation des *Interrogatory on Perseveration Tendency*, noch vor dem bekannteren *Personal Data Sheet* von Woodworth (1918), d.h. einem Fragebogen, der ein psychiatrisch orientiertes Interview von Rekruten ersetzen sollte. Lankes bildete Gruppen von Items und verwendete die Korrelationen zwischen diesen und einer Zusammenstellung anderer Perseverationstests zur empirischen Itemselektion.

5 Itementwicklung als Operationalisierung von Persönlichkeitseigenschaften

Die Itementwicklung für Persönlichkeitsfragebogen muss im Unterschied zu Intelligenz- und Leistungstests eigenständigen Prinzipien folgen. Hier sind geeignete Kompromisse zwischen der Konstrukt- und Kriterien-Validität, den Gütekriterien der inneren Konsistenz bzw. Homogenität (Reliabilität), der Testökonomie und Standardisierung zu finden. Die theoretischen Konstrukte der Persönlichkeitsforschung (Eigenschafts- und Zustands-Konzepte) stammen primär aus Selbstberichten und Selbstbeurteilungen. Dies ist eine fundamental andere Datenquelle als das objektiv aufgezeichnete Verhalten in einem Intelligenztest. Persönlichkeitseigenschaften werden allgemein als *sehr facettenreiche* Dispositionen verstanden, die sich zwar zeit- und situationsabhängig unterschiedlich, jedoch relativ überdauernd im individuellen Erleben und Verhalten manifestieren. Nur perspektivische Unterschiede bestehen zwischen den Eigenschaftskonstrukten und den Konstrukten im Bereich der Zustandsänderungen (Befindlichkeit, Stimmungsskalen u.a.), wobei deren höhere intra-individuelle Variabilität natürlich nicht einfach ein Messfehler ist.

Für die Itementwicklung gilt, dass die psychologisch wichtigsten theoretischen Konstruktkomponenten durch ein Spektrum inhaltlich ähnlicher Items indiziert werden müssen. Die Homogenität und Eindeutigkeit im statistischen Sinne sind nachrangig gegenüber der Auswahl inhaltlich zutreffender Bedeutungsgehalte. Ob der Itempool als adäquat (die Konstruktbedeutung intensional erschöpfend) gelten kann, muss der Autor und müssen die anschließenden Evaluationen zeigen, wobei der pragmatische Nutzen bei Assessmentaufgaben wohl die größte Bedeutung hat. Typische Operationalisierungsfehler entstehen aus verschiedenen Gründen: Ein Konstrukt hat mehr Bedeutungskomponenten als durch die verwendeten Variablen erfasst sind, die verwendete Variable gehört nicht zu dem gemeinten Konstrukt, die verwendete Variable enthält auch Aspekte anderer Konstrukte, die dann fälschlicherweise dem Zielkonstrukt zugeschrieben werden. Außerdem gibt es konstrukt-irrelevante Varianz, z.B. wegen extremer Itemschwierigkeiten. Die Itementwicklung für Persönlichkeitsfragebogen stellt spezielle und schwierigere Anforderungen, damit Operationalsierungsfehler und formale Mängel vermieden werden. Dies verlangt ein besonderes Vorgehen:

- Die Itementwicklung setzt voraus, dass das oft schwer abzugrenzende und nicht einfach zu explizierende Eigenschaftskonstrukt aufgrund zentraler persönlichkeitspsychologischer Arbeiten und eigener Forschungserfahrung in diesem Bereich gut bekannt ist. Bewährt hat sich die Diskussion von Itementwürfen mit Fachkollegen in einem mehrstufigen Verfahren.
- Das Eigenschaftskonstrukt muss semantisch in alltagssprachliche, zumutbare und relativ einfach beantwortbare Fragen oder Aussagen umgesetzt werden, wobei die Differenzierungsleistung zwischen Personen, aber z.T.

auch zwischen Situationen sowie mögliche Einflüsse von Geschlecht, Alter u.a. soziodemographischen Bedingungen zu bedenken sind.
- Items beziehen sich u.a. auf Erlebnisse, Gewohnheiten, Tätigkeiten, die diskontinuierlich auftreten: Folglich ist oft ein Bezug zu Situationen und Zeiträumen herzustellen. Je genauer dies geschieht, desto weniger werden sich u.U. die pauschalen Antworttendenzen, subjektiven Aggregationsstile und Retrospektionseffeke auswirken und desto länger und schwieriger wird die Formulierung.

Die Itementwicklung verlangt Kompromisse zwischen gegensätzlichen Prinzipien und unterscheidet sich in mehrfacher Hinsicht von der Entwicklung eines Intelligenztests mit dessen vergleichsweise einfachen Konstrukten. Wenn es in einem Untertest z.B. darum geht, Symbole zuzuordnen, schwieriger werdende Rechenaufgaben zu lösen oder sich Zahlenreihen zu merken, kann für die vielen einzelnen Operationen gewiss ein hohe Homogenität hinsichtlich dieser elementaren Intelligenzfunktionen behauptet und dementsprechend skaliert werden. Demgegenüber haben die persönlichkeitspsychologisch wohl am besten bewährten Konstrukte der Sekundärfaktoren Extraversion und Emotionalität im Sinne Eysencks so viele wichtige Facetten, dass auch einige verhältnismäßig heterogen erscheinende Aussagen kombiniert werden müssen, um dieses Eigenschaftskonstrukt zu repräsentieren. Die typischen Extraversions-Items zu impulsivem und zu geselligem Verhalten betreffen keine sehr homogenen, sondern unterschiedliche, nicht notwendig hochkorrelierte Facetten eines theoretischen Konstrukts. Die Itemselektion, die ausschließlich nach dem Prinzip der Rasch-Skalierung oder nur aufgrund der Faktorladung durchgeführt würde, müsste zu extrem homogenen Skalen führen, wenn das Dilemma innere Konsistenz und adäquate Konstruktvalidität nicht erkannt wird. Auch die sog. Guttman-Skala fordert im Prinzip, dass jede Person alle Items positiv beantwortet, deren Schwierigkeitsgrad unter ihrer habituellen Fähigkeit (Eigenschaft) liegt, und keine „löst“, deren Schwierigkeit darüber liegt. Was würde dies z.B. für die Dimension der Klagsamkeit über diverse Beschwerden, eine der varianzstärksten und stabilsten Persönlichkeitseigenschaften überhaupt, anschaulich bedeuten? Sollte jemand der die selteneren („schwierigeren“) Magenschmerzen hat, aus Gründen der Eindeutigkeit auch die häufigeren („leichten“) Kopfschmerzen haben, aber keinesfalls die sehr seltenen Herzschmerzen nennen? Cattell und Warburton (1967) vertraten die entschiedene Auffassung, dass Persönlichkeitskonstrukte mit bedeutender empirischer Validität nie faktoriell rein sind, sondern mehrere Komponenten enthalten.

Das zu wenig erläuterte Risiko lautet, dass während der Itemselektion nicht nur Daten an Modelle angepasst werden, sondern tendenziell die psychologischen Eigenschaftskonstrukte manipuliert werden. Die Schwierigkeiten bei der Itemselektion für Persönlichkeitsfragebogen liegen darin, die statistisch-formalen Eigenschaften eines Items *und* die inhaltlichen Bedeutungen zu berücksichtigen. Die erforderliche multivariate Sichtweise des Operationalisierungsprozesses unterscheidet sich hier von den Axiomen der Eindeutigkeit und Sparsamkeit. Konfirmatorische Analysen können, abgesehen von ihren internen Schwierigkeiten mit den Schätzern und mit der Variablenzahl, wichtige Absichten der Itemauswahl nicht abbilden. Beispielsweise wurden während der Konstruktion der FPI-Skalen Extraversion und Emotionalität einige Items mit relativ niedriger Trennschärfe bzw. Ladung, aber *gegensätzlichem Vorzeichen* als Suppressor-Items ausgewählt, um die Null-Korrelation beider FPI-Skalen – erfolgreich – zu garantieren. Dies geschah in der Absicht, Eysencks Konzeption nachzubilden.

Aus solchen Gründen ist die schematische Itemanalyse zur Maximierung der Konsistenz von Persönlichkeitsfragebogen unangebracht. Die Operationalisierung ist eine psychologische Aufgabe und braucht das theoretische Annahmengefüge über das Konstrukt, kann also durch den Formalismus einer Item- oder Faktorenanalyse nicht ersetzt, sondern nur unterlegt werden.

6 Itemkritik und Aggregationen

Fragebogenitems haben typische Inhalte (Angleitner, 1976; Angleitner & Wiggins, 1986; Lösel, 1995): Beschreibungen eigener Reaktionen und Reaktionen anderer, eigene Zuschreibungen von Eigenschaften, Wünschen und Interessen, Einstellungen und Überzeugungen, außerdem biographische Fakten u.a. Die Itemformulierung und unscharfe Quantoren verlangen bei den Befragten unterschiedlich komplexe kognitive Prozesse. Deshalb meint Lösel: „Da die Instruktionen der meisten Fragebogen zudem auffordern, möglichst spontan zu reagieren, ergibt sich eine Beantwortungssituation, die derjenigen projektiver Tests nicht unähnlich ist.“ (S. 365). Damit ist zwar das Testprinzip projektiver Verfahren, latente oder unbewusste Tendenzen zu provozieren, kaum erschöpft, doch bestehen zweifellos erhebliche Interpretationsspielräume für die Selbstbeurteilung und Selbstdarstellung.

Bereits die Antwort auf ein typisches Fragebogen-Item (z. B. “Ich bin häufig angespannt”) liefert ein kompliziertes Aggregat, denn eigentlich muss nun über die erlebten Facetten der Anspannung (mental, emotional, körperlich), über Situationen und Häufigkeiten nachgedacht werden. Diese *subjektive* Aggregation findet „irgendwie“ bereits bei jedem Item statt, während bei einem Intelligenztest gewöhnlich erst der Untersucher über seriell wiederholte Aufgaben und Aufgabengruppen aggregiert. Unter Aggregation wird in der Regel nur diese vom Untersucher vorgenommene, in der Regel additive Zusammenfassung von Elementen verstanden (vgl. Amelang & Schmidt-Atzert, 2006; Schweizer, 1990).

Auf Aggregation basiert auch das Spearman-Brown-Prinzip der Reliabilitäts-Steigerung durch Verlängerung des Tests, d. h. Hinzufügen relativ homogener Items. Die neuere Diskussion über Aggregation wurde u. a. durch „multiple act criteria“ im Sinne von Fishbein und Ajzen (1974) und durch die Mischel-Epstein-Kontroverse über die angebliche Validitätsgrenze bei r = 0.30 angeregt, denn diese Kontroverse muss unter dem Gesichtspunkt der speziellen Datenaggregation differenziert werden.

Das Linsenmodell von Brunswik (1956) bezieht sich auf die repräsentative Auswahl von Variablen. Wenn es z. B. um statistische Vorhersagen des Verhaltens aus bestimmten Testbefunden geht, dann sollte zwischen dem Satz der Prädiktorvariablen und dem Satz der Kriterienvariablen eine symmetrische Beziehung (Linsendarstellung) bestehen, d.h. die Breite und Güte der Prädiktoren und der Kriterien sollten sich entsprechen. Wittmann (1987, 1988, 2002; siehe Beauducel, Biehl, Bosnjak, Conrad, Schönberger & Wagener, 2005) hat das Konzept von vier Datenboxen (Prädiktoren, experimentelles Treatment, nicht-experimentelles Treatment und Kriterien) in Anlehnung an Brunswik und Cattell entwickelt, um die notwendigen Präzisierungen von Assessmentstrategien und Validitäts- und Reliabilitätsaspekten zu erreichen. Aggregiert werden kann über Zeitpunkte (Messwiederholungen), über Situationen (Settings, Untersuchungsbedingungen), über Items (Konstrukt-Facetten, Verhaltensweisen) und andere Dimensionen der Datenbox, so dass ein mehrdimensionales Aggregat entsteht. Ein asymmetrisches Aggregationsniveau läge dann vor, wenn z. B. der Testwert eines Persönlichkeitsfragebogens für “Extraversion” als Prädiktor herangezogen wird, um die an einem bestimmten Tag beobachtbare Geselligkeit und Unternehmungslust vorherzusagen. Der Testwert E als Index einer überdauernden Persönlichkeitseigenschaft entsteht durch zeitliche und inhaltliche Aggregation vieler Erfahrungen des Individuums in jeder Itemantwort und durch testmethodische rechnerische Aggregation über viele Items. Dagegen bezieht sich die Verhaltensbeobachtung des Kriteriums nur auf einen kurzen Zeitraum, so dass hier erweiterte, symmetrische Aggregationen notwendig sind. Das Verfahren kann pragmatisch kriteriums-orientiert (Indexmessung) oder theoretisch konstrukt-orientiert sein. Wittmann fordert, der Planung adäquater Validierungsuntersuchungen im Vergleich zu den oft überwertig diskutierten Reliabilitätsberechnungen mehr Gewicht zu geben.

7 Konstruktionsprinzipen und Plädoyer für eine multistrategische Heuristik

Einige Autoren geben eine Typologie von Konstruktionsstrategien und versuchen, zwischen induktiver versus deduktiver Strategie, empirischer versus rationaler, theoriegeleiteter versus a-theoretischer zu unterscheiden oder sprechen von konstrukt-orientiertem Verfahren, intuitiver Konstruktion nach vermuteter Inhaltsvalidität oder von Skalenbildung a priori. Diese Kennzeichnungen sind sehr missverständlich. Vor allem wird zu leicht übersehen, dass jedes Vorhaben, jede Itementwicklung und jeder Itempool bereits Vorentscheidungen enthält. Wenn heute ein Persönlichkeitsinventar ohne Rücksicht auf jegliches theoretische Wissen oder bisherige Konstruktionen geschaffen werden könnte, würde das einen katastrophalen Eindruck vom Stand der wissenschaftlichen Psychologie vermitteln. Unvermeidlich sind theoretische Vorentscheidungen u.a. zum psychologischen Gültigkeitsbereich (Auswahl der interessierenden Konstrukte oder lexikalischen Domänen), zum Geltungsbereich hinsichtlich der angezielten Populationen (Alter, Status beispielsweise als Bewerber oder als Patient usw.), Anzahl der Konstrukte und Bevorzugung einer größeren oder geringeren, heuristischen oder sparsamen Varianzausschöpfung, d.h. mit reichhaltigen, z.T. überlappenden Skalen oder möglichst sparsamer Auswahl, d.h. starker Reduktion. Viele weitere Entscheidungen sind nötig: über das Spektrum der zu erfassenden Facetten, Itemtyp, Antwortformat, Itempool, testtheoretisches Modell, Qualität der Normierung, Überprüfung der Gütemerkmale, Revision usw.

Die Persönlichkeitsinventare wurden sämtlich nach der allgemein bewährten, wiederholten, induktiv-hypothetisch-deduktiven Strategie entwickelt und unterscheiden sich nur graduell, welches Gewicht den inhaltlichen oder den teststatistischen Konzepten gegeben wird, wie elaboriert, repliziert oder multistrategisch dieser Konstruktionsprozess ist. Alle neueren Persönlichkeitsinventare sind primär intern aufgrund eines mehr oder minder bevölkerungsrepräsentativen Datensatzes konstruiert worden, d.h. in Kenntnis empirischer Item- und Skalenkennwerte, im ersten Schritt jedoch in der Regel ohne aktuelle Kriterieninformation. Der Konstruktion aufgrund externer Itemvaliditäten wie beim MMPI wurde nicht gefolgt, wobei – vom erheblich größeren Aufwand abgesehen – auch die zu erwartende größere Skalen-Heterogenität verantwortlich sein wird (ein Hinweis auf das Reliabilitäts-Validitäts-Dilemma).

Multistrategische Konzeption des FPI

Das FPI hat mit seinem multistrategischen Konstruktionsverfahren einigen Rezensenten Schwierigkeiten bereitet. Nach den oben genannten Typisierungen wurde es mal als *induktiv* wegen der Faktorenanalyse, als *deduktiv* wegen der entschiedenen Auswahl von Persönlichkeitseigenschaften für bestimmte Forschungs- und Praxisbereiche, andererseits als *a-theoretisch* bezeichnet (Kubinger, 1995). Tatsächlich wurde keine „neue Persönlichkeitstheorie“ vorgelegt, was heute in umfassender Weise kaum noch möglich zu sein scheint. Die Autoren wählten Persönlichkeitseigenschaften aus, die für ihre Arbeitsgebiete in der Psychologie wichtig waren. Offensichtlich hat außerdem das psychologische Abwägen von itemmetrischen und faktorenanalytischen Parametern irritiert, so dass die Konstruktion schwer einzuordnen schien. Es kam hinzu, dass die statistischen Kennwerte nur als *heuristische* Hinweise gewertet wurden, die ausdrücklich der

psychologischen Konstruktoperationalisierung untergeordnet sind. Auch die Durchführung sehr großer und tatsächlich repräsentativ erhobener und nicht nachträglich zusammengesetzter Normierungsstichproben ist auch heute noch unüblich. – Das FPI pauschal als ein „faktorenanalytisch" konstruiertes Inventar zu bezeichnen, trifft also nicht das Besondere der Konstruktionsweise. Das FPI scheint so wenig in das Bild zu passen, dass die multistrategische Prozedur zur Konstruktion und Rekonstruktion von anderen Testautoren kaum diskutiert und von Rezensenten kaum angesprochen wurde.

Anfänglich spielte unter dem Einfluss Eysencks u.a. auch das Konzept der behavioralen (aktuarisch-statistischen) Interpretation der Itemantworten eine Rolle, d.h. die empirische Diskriminationsleistung zwischen Statusgruppen, beispielsweise Patienten mit psychosomatischen Störungen im Vergleich zu Gesunden (vgl. die Versuche mit nicht-verbalen Skalenitems, Amelang et al., 2002; Amelang & Schmidt-Atzert, 2006). Die Fragebogenantworten konsequent als Selbstbeurteilungen anzusehen, ändert nichts an der Möglichkeit, solche Aussagen sinnvoll mit den Daten von anderen Bezugsgruppen und Normen zu vergleichen, verlangt jedoch eine andere Interpretationsweise, die möglichst viele Kontextinformationen einzubeziehen versucht. Parallel zu dieser Klarstellung wurde versucht, Alternativen zu den metrisch voraussetzungsreichen Techniken der Itemanalyse und Faktorenanalyse zu finden, d.h. für kategoriale Daten eventuell eher geeignete Analyseverfahren.

Bereits 1973, in der 2. Auflage, wurde über Clusteranalysen, d.h. die nicht-metrische Guttman-Lingoes Smallest Space Analysis, und hierarchische Clusteranalysen berichtet. Daneben wurden heuristisch auch multidimensionale Skalierungen, eine Skalogramm-Analyse von Guttman (R. Hampel), Analysen der Reproduzierbarkeit der Strukturen in verschiedenen Modellen der Faktorenanalyse (W.W. Wittmann & R. Hampel) und mit einem der ersten funktionsfähigen Computerprogramme auch Rasch-Skalierungen mit Modell-Tests nach Fischer durchgeführt (J. Fahrenberg). Diese Analysen sollten die Invarianz der Itemgruppierungen jeder Skala gegenüber verschiedenen Analysemethoden prüfen. Je nach Verfahren zeigten sich in einer Anzahl von Items Abweichungen. So ergab u.a. auch die Rasch-Skalierung Hinweise auf viele modellunverträgliche Items, mit dem seltsamerweise besten Resultat für die Skala Offenheit. Insgesamt konnten in den Ergebnissen keine praktisch verwertbaren Hinweise auf gemeinsame sprachliche oder formale Mängel gefunden werden, um Items deswegen zu eliminieren. Aus dieser Sachlage wurde der Schluss gezogen, dass das voraussetzungsvolle Verfahren der Rasch-Skalierung sich für facettenreiche Persönlichkeitskonstrukte nicht eignet, es sei denn auf Kosten einer massiven Skalenkürzung oder einer Homogenisierung durch Zufügen inhaltlich redundanter Items. Die Versuche zur Exploration der Ergebnisse nach damaligem Methodenstand waren unergiebig, da weder formale noch inhaltliche Gemeinsamkeiten der kritischen Items zu erkennen waren, eine Nutzanwendung im Vergleich zu den relativ robusten Item- und Faktorenanalysen nicht einleuchten konnte.

Es trifft zu, dass der primäre Konstruktionsprozess des FPI durch die Berechnung von Itemparametern, durch Item- und Faktorenanalysen und mit den normierten Testwerten eine Intervallskalierung unterstellt hat. Die Addition der Itemantworten zu individuellen Testwerten könnte als das einfache Zählen kategorialer Daten aufgefasst werden. Dabei würde jedoch übersehen, dass den Items für den interindividuellen Vergleich näherungsweise nicht nur gleiches numerisches Gewicht (sehr ähnliche Itemparameter), sondern auch ein psychologisch äquivalentes Bedeutungsgewicht unterstellt wird.

In späteren Auflagen und Revisionen des FPI wurden neben der Faktorenanalyse zur Ordnung der Items – *nicht* zur metrischen Dimensionierung des Pools – nur noch Clusteranalysen nach Wards Verfahren durchgeführt. Die Gruppierung der Items entsprach weitgehend der Faktorenanalyse bzw. der Item-Skalen-Zuordnung (mit einzelnen Abweichungen, siehe Manual des FPI-R). Clusteranalysen verlangen Entscheidungen: welcher Ähnlichkeitskoeffizient und welcher Algorithmus der Clusterung verwendet und bis zu welcher Clusterzahl zusammengefasst wird. Die Ergebnisse solcher Clusteranalysen gelten je nach Verfahren und (oft zu kleinen) Personenstichproben als nicht sehr robust, jedoch stellt sich die begründete Frage an künftige Fragebogenkonstruktionen, weshalb solche und ähnliche Verfahren, die messtheoretisch voraussetzungsärmer sind, nicht ausreichen könnten.

8 Gibt es für die Konstruktion von Persönlichkeitsfragebogen ein überzeugendes Messmodell?

Im Laufe der Zeit wurde verschiedentlich vorgeschlagen, ein statistisch begründetes Konzept oder Messmodell als die Standardmethode der Testkonstruktion zu akzeptieren: die Itemselektion nach interner Trennschärfe und externer Itemvalidität, die Faktorenanalyse, das Rasch-Modell bzw. neuere Item-Response Modellierungen. Offensichtlich hat jedes Konzept Vorzüge und Nachteile, die in den jeweiligen Voraussetzungen und spezifischen Anwendungsschwierigkeiten liegen. Bei allen formalen Vorzügen der neueren Item-Response-Modelle, wäre es jedoch inadäquat, sie unterschiedslos für alle psychologischen Testkonstruktionen als die „Methode der Wahl" anzusehen. Der Behauptung der prinzipiellen Überlegenheit dieser Modellierungen steht zumindest zweierlei entgegen: Diese Modelle machen sehr einschränkende Voraussetzungen, die aber in ihren z.T. unerwünschten und sogar negativen Konsequenzen für die psychologischen

Eigenschaftskonstrukte nicht hinreichend diskutiert werden. Es besteht ein besonderes Missverhältnis zwischen den strikten formalen Voraussetzungen und dem Rechenaufwand einerseits und der mangelnden Rechtfertigung, den Selbstbeurteilungen metrische Intervallskalen zu unterstellen. Zutiefst fragwürdig ist die Botschaft, dass diese Modelle unterschiedslos für elementare Intelligenz- und Leistungstests mit extrem homogenen, auf eine Dimension der Lösungsschwierigkeit ausgewählten Items, und ebenso auf Persönlichkeitsfragebogen angewendet werden sollten. Es mangelt an einer kritischen Darstellung der immanenten Voraussetzungen. Es fehlt weiterhin an expliziten Methodenstudien, die an einem geeigneten Datensatz die konvergenten und divergenten Resultate der verschiedenen Konstruktionsweisen, jeweils lege artis durchgeführt, aufzeigen und kommentieren.

Das Messmodell wird aus der sog. Item-Charakteristik-Funktion deutlich. Sie bildet die Beziehung zwischen dem Antwortverhalten, d.h. der Lösungswahrscheinlichkeit für ein Item, und den postulierten Modellparametern, d.h. der Itemschwierigkeit und der zur „Lösung" benötigten Fähigkeit ab. Aus der Kenntnis der Fähigkeit einer Person und der Itemschwierigkeit kann vorhergesagt werden, ob die Person das Item lösen kann. Diese Messkonzeption entspricht weitgehend dem Ideal einer physikalischen Messung, ist jedoch ausdrücklich probabilistisch (stochastisch) hinsichtlich der Parameterschätzung gemeint und nicht deterministisch. Das bekannteste probabilistische Modell ist die Rasch-Skalierung für dichotome Itemantworten. Instruktiv ist ein Aufsatz von Kubinger und Draxler (2007), die einige Probleme der Testkonstruktion nach dem Rasch-Modell untersuchen und dabei neuere Modellentwicklungen beschreiben. Der Entscheidungsspielraum der Untersucher wird deutlich, die Abhängigkeit der Resultate von den gewählten Risikoschätzungen und von der Personenzahl (mit problematischen Auswirkungen *hoher* Personenzahlen für die statistischen Schätzungen) sowie von der Art der verfügbaren Computerprogramme. Ein anderes Beispiel: Die lokale stochastische Unabhängigkeit der Itemantworten zu behaupten, wie es in einer Fußnote geschieht, ist für Persönlichkeitsfragebogen sachlich falsch, denn es gibt Untersuchungsergebnisse für diesen, psychologisch recht trivialen Sachverhalt, dass der Kontext einen – wenn auch geringen – Effekt haben kann (z.B. Krampen et al., 1992). Was auf elementare Aufgabenserien, beispielsweise das Zuordnen von Symbolen u.a. Aufgaben zutreffen mag, wurde übergeneralisiert.

Die Methodik der Latent Class Analyse, Zusammenhänge zwischen manifesten, *kategorialen* Variablen durch latente *kategoriale* Variablen aufzuklären, scheint für die Konstruktion von Persönlichkeitsfragebogen eher geeignet zu sein als die Latent Trait Modelle mit der unterstellten Intervallskala. Die von Rost (2004) dargestellten Modelle enthalten, als Pendant zu den *Itemfunktionen* bei den quantifizierenden Testmodellen, *Itemprofile* mit hier uneingeschränkter Variierbarkeit der Lösungswahrscheinlichkeiten, so dass unterschiedliche Repräsentationsformen bzw. diskrete Itemfunktionen zugelassen sind. Die Latent Class Analyse kann zur Clusteranalyse dienen (paarweise Ähnlichkeiten oder Distanzen aller zu klassifizierenden Objekte) und zur Ähnlichkeitsbeurteilung von Personen, von denen kategoriale oder ordinale Daten vorliegen. Die latente Klassenanalyse verlangt keine Auswahl von Ähnlichkeitsmaßen, hat jedoch ebenfalls Interpretationspielräume, u.a. bei der Bestimmung der Maxima. Daneben sind Analysen von Konsistenzen, Separierung von Varianzquellen, Situationsspezifität, Identifikation auffälliger Personen bzw. Items möglich. Außerdem wurden Varianten zur Untersuchung multivariater Assoziationen entwickelt (vgl. u.a. Forman, 2002; Rost, 2004, 2006; Schweizer, 1999). – Die Meinungsunterschiede, insbesondere bei Persönlichkeitsfragebogen, werden bleiben: Sind auch hier bestimmte Messmodelle bzw. Skalierungen grundsätzlich überlegen? Oder handelt es sich um heuristische Verfahren, um innerhalb eines persönlichkeitspsychologischen Annahmengefüges statistische Hinweise zur empirisch geleiteten Ordnung von Konstruktfacetten zu gewinnen und empirisch nützliche Indizes abzuleiten?

Weshalb einzelne Items nicht zu einer Skala zu passen scheinen, ist eine wichtige Frage. Falls es gelingt, auch Antworttendenzen und mögliche Beantwortungsfehler bzw. *nicht-skalierbare Personen* zu identifizieren (z.B. Austin, Deary, Gibson, McGregor & Dent, 2006; Ponocny & Klauer, 2002) oder Antworttendenzen zuverlässig von den individuellen Eigenschaftswerten zu separieren, wären das interessante Befunde. Wie solche Muster mit persönlichkeitspsychologischen Unterschieden konfundiert sind, wird jedoch auf diese Weise kaum zu klären sein. Gegenwärtig scheint der testkonstruktive Gebrauchswert erst schwach ausgebildet zu sein.

Problematisch ist auch die Empfehlung mehrstufiger statt dichotomer Itemantworten (siehe Kubinger, 2002a; Moosbrugger & Kelava, 2007). Dieser Vorschlag ist aus dem Ideal der Intervallskalierung verständlich, bringt jedoch zusätzliche semantische und statistische Schwierigkeiten. Wie schon frühere Autoren beschrieben haben, und Rohrmann (1978) aufgrund einer breiten Umfrage zeigte, ist das populäre Verständnis solcher Graduierungen und Quantoren für Intensitäts-, Wahrscheinlichkeits- und Bewertungs- (Zustimmungs-) Skalen sehr divergent. Automatisch stellt sich die Frage nach einem Skalenmittelpunkt und dessen normativer Funktion (u.a. Schwarz, 1990, 2007; Schwarz & Scheuring, 1992). Außerdem müssen natürlich auch hier die Verteilungsanomalien berücksichtigt werden. Insofern sollten Postulate hinsichtlich der mehrstufigen Antwortformate überdacht werden (siehe die u.a. von Rost, 2004, 2006 untersuchten Probleme mehrstufiger Antwortformate).

Bemerkenswert ist, dass seit Aufkommen der Rasch-Skalierung in Deutschland in den 1970er Jahren, zwar einige Leistungstests, kaum jedoch Persönlichkeitsfragebogen auf diese Weise konstruiert wurden. Das erste mehrdimensionale

Persönlichkeitsinventar dieser Art (Hehl & Hehl, 1975) scheint fast vergessen zu sein. Es muss sich noch zeigen, ob originelle und im empirischen Gebrauchsnutzen überlegene Inventare publiziert werden. Bisher handelte es sich primär um Re-Analysen und fragwürdige Demonstrationen.

9 Problematischer Gebrauch der konventionellen Itemanalyse, Faktorenanalyse und IR-Modellierungen

Wie die Itemanalyse so ist auch die Methodik der Faktorenanalyse primär für die Intelligenzforschung entwickelt und später auf die Konstruktion von Persönlichkeits-Fragebogen und Stimmungsskalen übertragen worden. Bei *schematischer* Anwendung begünstigen die konventionelle Itemanalyse und die faktorenanalytische Methodik die Entstehung sehr homogener Skalen, d.h. die formale Maximierung der inneren Konsistenz ohne Rücksicht auf die empirische Validität. Das geschieht, indem anfänglich enthaltene, d.h. vorgegebene oder noch unzureichend erkannte Dubletten (Tripletts usw.) weitgehend redundanter Items aufgrund ihrer sehr hohen Trennschärfeindizes oder hohen Kommunalität die Ladungsmuster bzw. die Rotation dominieren, mit Folgeschäden bei der Beurteilung der relativen Varianzanteile und hinsichtlich anderer Eigenschaften. Generell besteht also ein hohes Risiko, dass inhaltlich sehr ähnliche Items, also kleine sprachliche und inhaltliche Varianten, technisch aufgrund ihrer höheren gemeinsamen Varianz begünstigt werden. Sehr hohe Trennschärfe-Koeffizienten und Faktorladungen sollten gerade ein Anlass sein, solche Items zu eliminieren, weil sie – bei praktisch begrenzter Itemzahl – weder testökonomisch noch vielversprechend für die externe Validität sein können. Deshalb muss das Gütekriterium der inneren Konsistenz relativiert werden. Einige Publikationen tragen zu einer unausgewogenen Bewertung bei, wenn sie vorrangig über die Reliabilitäten von Skalen berichten und ein „je höher, desto besser“ suggerieren. Der *zu geringe* Reliabilitätskoeffizient eines Tests (Anteil wahrer Varianz/Fehlervarianz) limitiert zwar die maximal erreichbaren Validitätskoeffizienten (vorhersagbare Kriterienvarianz). Aber eine *hohe* innere Konsistenz eines Persönlichkeitsfragebogens (extreme Item-Homogenität nach Rasch-Modell) bedeutet – anders betrachtet – Redundanz, geringere Testökonomie und potentiell einen Verlust an u.U. wesentlichen Facetten des gemeinten theoretischen Konstrukts und damit potentiell einen Verlust externer Validität. Lienert (1961) stellte fest:

Es liegt eine gewisse Kunst darin, einen Test sowohl möglichst reliabel wie auch zugleich möglichst valide zu gestalten; die Reliabilität scheint eher durch homogene Aufgaben, die empirische Validität dagegen durch heterogene Aufgaben gewährleistet zu sein. Man spricht in diesem Zusammenhang von einer *partiellen Inkompatibilität* der beiden Kardinalkriterien, indem man das eine anstrebt, gefährdet man das andere (S. 294-295).

Bemerkenswert ist, dass dieses Dilemma, trotz Lienerts Erklärung, keineswegs in allen Lehrbüchern der Testkonstruktion erwähnt oder hinreichend erklärt wird. Demgegenüber geht Rost (2004) zwar auf die statistische Formulierung des Problems ein, erläutert auch die Gefahren bei schematischer Itemselektion, hält die Frage jedoch für kein Problem der *Testtheorie.* Er meint, dass die interne Konstruktion von Messwerten von der Frage der externen Validität getrennt werden sollte. Die notwendige Heterogenität von Prädiktoren zur Vorhersage von komplexen Kriterien solle durch die Kombination eines Testwertes mit anderen Testwerten hergestellt und nicht innerhalb eines Tests angesiedelt werden (S. 394). – Ist auch diese Auffassung primär aus Sicht der Intelligenz- und Leistungstests geprägt?

Aus den skizzierten Gründen kann eine *schematische* Anwendung faktoren- und itemanalytischer Strategien zu inadäquaten Skalenkonstruktionen führen. Eine ausschließlich nach Rasch-Homogenität und verwandten Konzepten, d.h. ohne multistrategische Kontrollen durchgeführte Konstruktion von *Persönlichkeitsfragebogen* kann zu schwerwiegenden Operationalisierungsfehlern und einer Einschränkung der empirischen Kriterienvalidität führen. Die in der Konstruktion von Persönlichkeitsfragebogen zu treffenden Kompromisse unterscheiden sich grundsätzlich vom Bereich der Intelligenz- und Leistungstests, wie natürlich auch jener Bereich sich in einigen Prinzipien von der Sensorischen Psychophysik oder der psychophysiologischen Diagnostik mit ihrer physiologischen und physikalischen Messmethodik unterscheidet**.**

Innere Konsistenz, Homogenität und deren Bedeutung

Das Konzept der inneren Konsistenz in der traditionellen Testtheorie hat Grenzen und Mängel. Abgesehen von der Stichprobe hängt der Koeffizient u.a. von der Anzahl der Items ab, wobei auch relativ heterogene Items rechnerisch zur Erhöhung beitragen können. Eine hohe Reliabilität ist eine wichtige Voraussetzung, dass eine Skala nicht nur für Screeningzwecke bzw. Gruppenuntersuchungen, sondern auch für Diagnostik im Einzelfall nützlich ist. Die Konsistenzkoeffizienten beschreiben den inneren Zusammenhang der Items, Stabilitätskoeffizienten weisen auf die für Prognosen wichtige Erwartung künftiger Testwerte hin. Die möglichen Einschränkungen und die Techniken zur Erhöhung der Reliabilität nehmen einen großen Raum in den Lehrbüchern ein. Der hauptsächliche praktische Grund, abgesehen von messmethodischen Idealvorstellungen, ist die Berechnung von Vertrauensintervallen der individuellen Testwerte (Konfidenzintervalle anhand des Standardmessfehlers bzw. Standardschätzfehlers). Die erwünschten, kleinen Vertrauens-

intervalle setzen hohe Reliabilitätskoeffizienten voraus und die Differenzierung gelingt besser, wenn die Skalen relativ viele Items und eine große Varianz haben. Im Falle des FPI wurden 1970 in Anlehnung an R.B. Cattell die Stanine-Grobnormen eingeführt, um auf diese Unsicherheiten hinzuweisen, zugleich wurden große Normierungsstichproben verwendet, um in den Tabellen die beträchtlichen Effekte von Geschlechtszugehörigkeit und Altersgruppe abbilden zu können. Die Vertrauensintervalle der individuellen Testwerte aus typischen Persönlichkeitsfragebogen sind relativ groß; sie lassen sich, wenn nötig, durch mehr Items auf Kosten der Zumutbarkeit und der Testökonomie verringern. Das revidierte FPI-G enthielt ursprünglich 210 Items für 12 Skalen, das FPI-R, vor allem aus Gründen der Zumutbarkeit, nur noch 136 (+2) Items für die 10 Standardskalen sowie E und N.

Jede Beurteilung von Risiken wird sich an der gewünschten Entscheidungssicherheit und an den möglichen nachteiligen Folgen orientieren (siehe Amelang & Schmidt-Atzert, 2006). In der Praxis muss sich die Signifikanzbeurteilung von Testwerten nach der Fragestellung richten, d.h. nach Kosten-Nutzen-Abwägungen, Fehler der ersten und der zweiten Art, nach der Beurteilung, ob ein relativ homogenes oder heterogenes Merkmal erfasst wird, und nach den Alternativen, falls auf den betreffenden Test verzichtet wird. Es darf auch nicht übersehen werden: Viele Entscheidungen der diagnostischen Praxis fallen nach anderen Risikoschätzungen als dem statistischen 5 %-Niveau. Lienert (1961) schilderte aufgrund seiner breiten Forschungserfahrung in Medizin und Psychologie einige Beispiele, wo ein sehr viel höheres Risiko akzeptiert wird, weil die Alternativen sehr unerwünscht sind. Ein Verfahren mit niedriger Reliabilität kann ggf. immer noch wichtige Hinweise geben. Allgemeingültige Richtwerte zur Höhe der Reliabilität können folglich nicht gegeben werden, denn es sind zu viele Bedingungen zu berücksichtigen (vgl. u.a. Amelang & Schmidt-Atzert, 2006; Moosbrugger & Kelava, 2007).

Homogenität im Sinne der IR-Modellierungen ist nicht identisch mit Homogenität im Sinne von Item-Korrelationen oder mit Faktor-Reinheit. Aber allgemein gilt: Geringe Konsistenz (Homogenität) hat nicht notwendig geringere Kriterienvalidität zur Folge, umgekehrt ist hohe Konsistenz (Homogenität) keine Gewähr für Validität, so betonten Michel und Conrad (1982):

Aus der Beziehung zwischen Reliabilität und Validität ergibt sich im übrigen, dass die in der Literatur fast durchweg gestellten Anforderungen an die Reliabilität von Tests überspitzt sind. Die …(…)… immer wieder gestellte Forderung, dass sich Reliabilitätskoeffizienten um oder über .90 bewegen sollten, steht in einem krassen Missverhältnis zu den praktisch erreichten Validitätskoeffizienten, die nur selten über .60 liegen. Es muss deshalb mit Nachdruck wiederholt werden, was Guilford bereits 1946 ausführte: ‚Relativ zu viel Aufmerksamkeit wird der Reliabilität und zu wenig der Validität geschenkt… Eine hohe Reliabilität sollte nie als selbständiges Ziel angestrebt werden. Sie ist nur in soweit wichtig, als sie zur Validität beiträgt' (S. 432). (S. 53-54)

Die konventionellen Reliabilitätsschätzungen sind wie die internen Item- und Faktorenanalysen, auf einfachste Weise möglich, der Nachweis einer neuen (externen) Kriterienkorrelation (nicht bloß mit ähnlichen Fragebogen) oder sogar eines inkrementellen Nutzens für reale Assessment-Entscheidungen sind unvergleichlich viel schwieriger und aufwändiger. – Die Wechselbeziehungen zwischen Gütekriterien sind in den Lehrbüchern oft nur ein nachgeordneter Aspekt. Die Problematik der Itemselektion mit dem Abwägen von Homogenität und Konstruktpräzisierung wird fast nie an einem realistischen Beispiel mit den speziellen Konsequenzen erläutert. Andererseits ist deutlich, dass Guilford, Michel und Conrad die Testwerte als mehr oder minder valide Prädiktoren von Kriterien begriffen und keine Mess-Struktur-Theorie einer Domäne anstrebten.

Sparsame Beschreibung mittels unkorrelierter Dimensionen und Skalen?

Das Prinzip der sparsamen Beschreibung ist dem naturwissenschaftlichen Denken entlehnt und hat dort große Überzeugungskraft (im Unterschied zur ursprünglichen Verwendung in der Philosophie als „Rasiermesser" im Sinne von Wilhelm von Ockham). Die kommentarlose Übernahme in die psychologische Methodenlehre, z.B. die Konstruktion von Persönlichkeitsfragebogen, wird grundsätzliche wissenschaftsmethodische Kritik auslösen. Sollte es ein generelles Prinzip sein, auch einen Motivationskonflikt oder einen Entwicklungsverlauf immer anhand eines minimalen Systems orthogonaler Faktoren auf die psychologisch sparsamste und einfachste Weise diagnostisch zu beschreiben? Zu den möglichen Fehlbewertungen faktorenanalytischer Ergebnisse gehört, die oft nützlichen Dimensionierungen als überlegene Strukturaussagen zu deuten, obwohl es sich nur um *eine* vielleicht elegantere Ordnung unter vielen anderen Möglichkeiten handelt. Hier schließen sich zu leicht weitere Vorentscheidungen an. Nicht allein die Faktoren, sondern auch die Skalenwerte sollen möglichst niedrig korrelieren, d.h. die Skalen werden so abgeleitet, dass sie möglichst viel der Varianz ausschöpfen. Solche Postulate können leicht zu Einseitigkeiten oder Fehlbewertungen der faktorenanalytischen Konstruktion von Persönlichkeitsfragebogen führen. Ungleich wichtiger bleibt ja die inhaltliche Rechtfertigung des Eigenschaftskonstrukts und der psychologische Kontext von Forschung und Diagnostik eines wichtigen und facettenreichen Eigenschaftkonstrukts – unabhängig von teilweisen Überlappungen der Varianz.

Kontroversen über die „richtige“ Anzahl von Persönlichkeitsfaktoren

In der faktorenanalytisch ausgerichteten Testkonstruktion taucht immer wieder ein Postulat auf, das schon zuvor nicht überzeugen konnte: die Behauptung, *die* basalen Faktoren der Persönlichkeit (oder der Stimmungen, Emotionen usw.) gefunden zu haben. Als ob es eine real existierende Entität zu erfassen gebe – statt nur ein mehr oder minder nützliches Beschreibungssystem geschaffen zu haben. Schon Guilford (1959), einer der Pioniere faktorenanalytischer Intelligenz- und Persönlichkeitsforschung, hat vor 50 Jahren darauf hingewiesen, dass höchstens dann über die basale und erschöpfende Anzahl der relevanten Faktoren diskutiert werden kann, wenn das *Universum* der Items repräsentiert ist. Eine Random-Stichprobenziehung ist jedoch, trotz aller lexikalischen Sammelversuche, nicht praktikabel. Denn was zur *Persönlichkeits-Sphäre* (wie es R.B. Cattell, 1957, nannte) gehören soll oder nicht, bleibt eine beliebige Abgrenzung (vgl. die Themenliste des Handbuchs der Persönlichkeitspsychologie, Weber & Rammsayer, 2005). So steht es frei, außer den Temperamentseigenschaften andere Bereiche hinzuzunehmen: Selbstkonzepte und Kontrollüberzeugungen, überdauernde Einstellungen und Interessen, aber auch Grundstimmungen, wiederkehrende körperliche Befindensweisen und alltägliche Beschwerdentendenzen (Klagsamkeit). Weshalb fehlen in den Lehrbüchern der Persönlichkeitspsychologie weitgehend die biographischen Grundzüge und typische Lebensthemen, die individuellen Wertorientierungen und Ziele, die persönlichkeitsbestimmenden weltanschaulichen und religiösen Überzeugungen? Wer könnte überzeugende Grenzen ziehen zu den prägenden Motivationen, Handlungsbereitschaften, Fähigkeiten, Kreativität, kognitiven Stilen usw.? Selbstverständlich würde auch die Erweiterung des Itempools in den Grenzbereich der Psychopathologie und des nicht gerade seltenen devianten Verhaltens die Anzahl der resultierenden Faktoren und möglichen Skalenbildungen wesentlich erhöhen und inhaltlich modifizieren. – Eine definitive Anzahl von Grund-Dimensionen festlegen zu wollen, ist noch weniger sinnvoll als entsprechende Behauptungen im Bereich der Intelligenzforschung über eine fixe Anzahl von Faktoren. Wäre es nicht überzeugender, im Gegensatz zu diesen Kontroversen über 3, 4, 5, 6, 7 oder mehr Faktoren, die Auswahl der Persönlichkeitskonstrukte inhaltlich zu rechtfertigen: im Hinblick auf die Fragestellungen und auf die praktischen Aufgaben des Assessment, z.B. in typischen Anwendungsfeldern?

Die Frage nach der minimalen bzw. weithin varianzausschöpfenden, insofern „richtigen“ Anzahl von Faktoren zur Beschreibung der Persönlichkeit hat eine lange Vorgeschichte, u.a. in den von Cattell und Guilford ausgearbeiteten sehr umfangreichen Inventarisierungen. Diese Diskussion über die zutreffende Anzahl hauptsächlicher Grundeigenschaften wurde wieder angefacht durch die von Costa und McCrae intensiv publizierten und breit exportierten Behauptungen über die sog. Big Five im Persönlichkeitsinventar NEO-FFI bzw. NEO-PI-R (Borkenau und Ostendorf, 1993, 2004). Dies geschah ohne eine hinreichende Aufklärung über die zweifelhaften Voraussetzungen, ohne Rücksicht auf die zirkulären Vorentscheidungen zum Itempool und ohne tiefergehende Rechtfertigung des Postulats, die möglichst „sparsamen“ Beschreibungen als Ideal hinzustellen. Eysencks fundamentale Faktoren E und N wurden zwar rekonstruiert, darüber hinaus scheinen die psychologischen Interpretationen deutlich mühsamer auszufallen. Mit den seltsamer Weise erst im zweiten Schritt abgeleiteten Facetten sollten dann nachträglich mehr psychologische Differenzierungen erreicht werden (vgl. Paunonen & Ashton, 2001). Dabei darf nicht übersehen werden, dass es sich bei der deutschen Version des NEO-Inventars um eine einfache Übersetzung ohne Rekonstruktion auf der Basis einer hinreichend großen bevölkerungsrepräsentativen Stichprobe handelte, ohne die rationale Konstruktion und den psychologischen Reduktionismus in den Voraussetzungen genauer darzulegen. Informationen, welche der Items bloß übersetzt und welche sinngemäß an die deutsche Mentalität angepasst wurden, wären kulturpsychologisch interessant. Grundlegende Überprüfungen stehen hier tatsächlich noch aus (siehe Andresen & Beauducel, 2008; Becker, 1996, 2000). Die direkteste Kritik an dem überwertigen Anspruch der NEO-FFI-Konzeption (NEO-PI-R) liefern ja – mit guten Argumenten und empirischen Befunden – die Autoren der Inventare mit 2, 3, 4, 6, 7 usw. Persönlichkeitsfaktoren, damit den Anspruch der NEOFF-Autoren relativierend oder auch falsifizierend. Die Zukunft wird zeigen, wie schnell die Proliferation fortschreitet. Noch zu wenig aufgeklärt ist jedoch die zirkuläre Beziehung zwischen der Präselektion des Itempools als Teil der Persönlichkeitssphäre und der Anzahl und den Inhalten der erhaltenen Faktoren.

Ein zu wenig behandeltes Teilproblem sind die kulturpsychologischen Vorurteile solcher Fragebogen. Eysenck erwähnte einmal in einem Gespräch über EPI und FPI, dass nach seiner Ansicht die Mentalität der Menschen in England und in den USA (damals) sich in psychologisch wesentlichem Ausmaß unterscheiden, sogar die englischen und deutschen Auffassungen über Itembedeutungen, z.B. im Bereich der Extraversion-Introversion, ähnlicher seien als englische und amerikanische. Doch was folgt aus solchen Annahmen für die Übersetzung von Items? Der *universalistische* Anspruch (McCrae & Costa, 1997) erinnert an frühere, gelegentlich fast imperialistisch anmutende Postulate amerikanischer Autoren in der kulturanthropologischen Feldforschung mit ihren Behauptungen, die USA und die asiatischen Kulturen vor allem auf einer grundlegenden Dimension „Individualismus-Kollektivismus“ unterscheiden zu können. Die Einseitigkeit solcher Forschung ist erst allmählich und z.T. erst unter dem Einfluss chinesischer Kollegen erkannt worden (vgl. Hofstede , 2006; Marsella et al., 2000; sowie zu den kulturellen Unterschieden der Social Axioms, Leung et al., 2002). Da die statistischen Vergleiche zumeist anhand von Gelegenheitsstichproben, d.h. „College Students“ mit einem starken Selektionsbias, unternommen wurden, ergeben sich fundamentale Zweifel an zu oberflächlichen inter-kulturellen Projekten (McCrae, Terracciano et al., 2005).

Bereits zwischen den europäischen Ländern bzw. Kulturen scheinen beträchtliche Unterschiede in der Bedeutung von Fragebogenitems zu bestehen. Aus diesen Erfahrungen ist die Kritik an den simplifizierenden inter-kulturellen Übersetzungen von Persönlichkeitsfragebogen, Stimmungsskalen usw. entstanden. Der ersten Phase mit schlichten Übersetzungen ohne jeden Versuch, die interkulturelle Äquivalenz zu kontrollieren, folgte eine zweite Phase mit „nicht wörtlichen, sondern sinngemäßen" Übersetzungen, ohne jedoch psychologisch zu erläutern, empirisch zu begründen und zu dokumentieren. Diese Schwierigkeiten sind allgemein unterschätzt worden. Dem sollen künftig Richtlinien vorbeugen.

Fragwürdige Übersetzungen und Adaptationen ohne systematische Rekonstruktion

Da sehr viele der deutschen Fragebogen und Skalen Adaptationen angloamerikanischer Verfahren sind, stellt ich die schwierige Frage der angemessenen Übersetzung. Wird eine möglichst treffende lexikalische Übersetzung gesucht oder vielmehr eine psychologisch äquivalente Formulierung, die u.U. einen anderen Inhalt für ein Item verlangt? Viele Testinhalte könnten stark sprachabhängig, auch idiomatisch, und kulturabhängig sein. Die Schwierigkeiten beginnen bereits, geeignete Übersetzungen für die Namen der Persönlichkeitsfragebogen zu finden.

Es mangelt auf dem Gebiet persönlichkeitspsychologischer Deskriptoren an genauen Berichten über die sprachlichen Kompromisse bei der Übersetzung oder der absichtlichen psychologischen Transposition von Iteminhalten. Die neuere Testpsychologie kann sich nicht mehr mit der früher oft praktizierten, schlichten Übernahme englischer Fragebogen begnügen. Gewiss reicht es nicht mehr aus, Items einfach zu übersetzen oder – in etwas fortgeschrittener Weise– eine unabhängige Rückübersetzung zu lesen. Es muss auch geprüft werden, ob noch das gemeinte Konstrukt erfasst wird. Nur sehr selten wird der naheliegenden methodischen Forderung entsprochen, bei der Adaptation amerikanischer Tests auch die gesamte Konstruktion mit einem zunächst weiter gefassten deutschen Itempool zu wiederholen und sich nicht allein mit der Erhebung von Testdaten bei deutschen Gelegenheits-Stichproben zu begnügen. – Es sei denn, man wollte die vollständige Angleichung der psychologischen Profile der deutschen an die amerikanische Bevölkerung behaupten.

Angesichts dieser grundlegenden und weitgehend ungelösten Probleme ist es auffällig, wenn die Lehrbücher der Persönlichkeitspsychologie den Postulaten und Kontroversen über „die" Persönlichkeitsfaktoren so viel Raum geben (z.B. Amelang et al., 2006; Asendorpf, 2007). Das Missverhältnis ist noch auffälliger, wenn sehr viel wichtigere Themen der Konstruktion und Evaluation von Fragebogen sowie der psychologischen Diagnostik nur kurz oder nur abstrakt, aber kaum im Hinblick auf die Konsequenzen für neue Assessmentstrategien behandelt werden: u.a. das erwähnte Konsistenz-Validitäts-Dilemma, das innovative Ambulante Assessment, die aus den Symmetrieprinzipien (Brunswik-Linse) entwickelten Konzepte, die Generalisierbarkeitstheorie, die multivariate Reliabilitätstheorie oder die multimodale Diagnostik – und die Konsequenzen dieser Entwicklungen für Forschung und Praxis.

Funder (2006) sieht heute in seinem Rückblick auf die in den 1970er Jahren aufblühende Person-Situation-Debatte eine falsche Dichotomie zwischen persönlichen und situativen Determinanten des Verhaltens, statt gründlicher die funktionellen Zusammenhänge zu erforschen. Einen Grund für die Fortdauer dieser Debatte vermutet er in den zugrundliegenden und undiskutierten philosophischen Überzeugungen hinsichtlich Individualität versus Gleichheit, Willensfreiheit versus Determinismus, Konsistenz versus Flexibilität. – Wird es vielleicht einmal einen ähnlichen wissenschaftspsychologischen Rückblick geben auf die essentialistisch anmutende Kontroverse über die richtige Anzahl von Persönlichkeitsfaktoren in Fragebogen? Verbunden mit der Einsicht, dass durch Faktorenanalysen oder Messmodelle weder die psychologische Auswahl geeigneter Persönlichkeitskonstrukte entschieden, noch die strukturelle Subjektivität der Selbstbeurteilungen aufgehoben werden kann?

10 Multitrait-Multimethod

Angesichts der strukturellen Subjektivität der Selbstbeurteilungen und ihrer fundamentalen Unsicherheit muss nach Bestätigungen durch andere Informationsquellen gesucht werden. Diese Forderung ist in dem Prinzip der *multiplen Operationalisierung* enthalten und führte Campbell und Fiske (1959) zur Idee der Multitrait-Multimethod-Matrix, um die Behauptung der *konvergenten* Validität verschiedener Methoden für *ein* Eigenschaftskonstrukt und zugleich ihre *diskriminante* Validität hinsichtlich *verschiedener* Eigenschaftskonstrukte zu prüfen. Adäquate MTMM-Untersuchungen mit *verschiedenen Datenebenen* wurden nur sehr selten durchgeführt, meist mit frustrierenden Ergebnissen angesichts der Diskrepanz zwischen psychologischen Methoden, die sich angeblich auf dasselbe Eigenschaftskonstrukt beziehen. Sehr häufig zeigte sich, dass als einheitlich angenommene Konstrukte eher als Anordnung von relativ unabhängigen Sub-Konstrukten aufzufassen sind. Die befriedigende Konvergenz multipler Indikatoren ist eher die Ausnahme, Divergenzen bzw. unerwartet niedrige Korrelationen sind häufig. Deshalb hat u.a. Fiske (1971, 1978, 1987) für die Persönlichkeitsforschung sehr viel genauere operationale Definitionen von Subkonstrukten und speziellen „construct-operation-units" verlangt (siehe auch *Themenheft Multimodale Diagnostik*, Fahrenberg, 1987b; Wittmann, 1987, 1988, 2002; vgl. Eid & Diener, 2005).

Definitionen von MTMM-Plänen und deren Auswertung mit Korrelationskoeffizienten oder konfirmatorischen Faktorenanalysen schildern Schermelleh und Schweizer (2006). Als Beispiel dient hier die Beurteilung von drei Persönlichkeitseigenschaften nach den drei Methoden: Selbsteinstufung, Fremdeinstufung, elterliche Einstufung. Die Koeffizienten der konvergenten Validität sind zwar signifikant, aber so niedrig, dass keine der Methoden die andere ersetzen könnte. Was bedeutet diese geringe Effektstärke, sogar bei drei sehr ähnlichen und außerdem empirisch konfundierten Datenquellen, für den allgemeinen Gebrauch von Fragebogen bzw. Selbst- und Fremdbeurteilungen in der psychologischen Diagnostik? Zu den MTMM-Matrizen gab es zwar zahlreiche statistische und konzeptuelle Beiträge (u.a. Eid, Nussbeck & Lischetzke, 2006; Ostendorf, Angleitner & Ruch, 1986), aber es mangelt in der Persönlichkeitsdiagnostik immer noch an überzeugenden empirischen Studien. Wie würden die MTMM-Befunde erst aussehen bei einem eigentlich multimodalen Ansatz, der auch *unabhängige* Datenquellen, Verhaltensmessungen, nicht-reaktive Maße u.a. einschlösse?

Die Idee der multiplen Operationalisierung ist gewiss einleuchtend. Sie wurde jüngst sogar in den Bereich der qualitativen Methodik, d.h. in die Methodenlehre des interpretativen Paradigma, verbreitet. Seltsam ist nur der dort gewählte Begriff „Triangulation“, weil damit *ursprünglich* gerade die genaue *quantitative* geometrische Ortsmessung von verschiedenen Standpunkten aus gemeint ist (vgl. Fahrenberg, 2002; Flick, 2008; Flick et al., 2000). Kompetente Psychologen/innen werden in vielen Fällen multiple Operationalisierungen anstreben, d.h. eine Methodenkombination auswählen, vor allem wenn es um verhältnismäßig breite theoretische Begriffe (Angst, Emotionalität, Aggressivität, Intelligenz u.a.) geht oder wenn es auf riskante, folgenreiche Entscheidungen ankommt. Mängel bei der Operationalisierung psychologischer Konstrukte können Validierungsstudien entscheidend beeinträchtigen, u.a. wenn Prädiktor und Kriterium unsymmetrisch geplant sind.

11 Multimodale Diagnostik

Eine methodisch fortgeschrittene Strategie der multiplen Operationalisierung wichtiger Konstrukte ist die Multimodale Diagnostik, die ausdrücklich *kategorial verschiedene Datenebenen* berücksichtigt. Außer den Selbsteinstufungen und den vielfach mit solchen Selbstaussagen konfundierten Fremdeinstufungen werden unabhängige Verhaltensdaten, objektive Tests und Messungen sowie hinsichtlich einiger Konstrukte auch physiologische Parameter aufgenommen. Hauptsächlich R. B. Cattell hat ein sehr umfangreiches Forschungsprogramm zur Inventarisierung von Faktoren unternommen, und ein wichtiges Prinzip war dabei der *multimodale* Ansatz: Lebenslaufdaten, Selbstberichte (Einstufungen, standardisierte Fragebogen), Verhaltensbeurteilungen, Verhaltensbeobachtungen, echte Verhaltensmessungen, objektive Tests, physiologische Messwerte, sollten aufgrund ihrer konvergenten Validität zur wissenschaftlichen Beschreibung universeller Eigenschaften der Persönlichkeit, der Fähigkeiten, der Zustandsänderungen, Motivationen, Einstellungen usw. führen. Dieses anspruchsvolle und sehr aufwändige Forschungsprogramm konnte wegen der oft nur minimalen gemeinsamen Varianz der hypothetischen Indikatoren „derselben“ Persönlichkeitseigenschaft nicht überzeugen und fand auch keine Fortsetzung. Dagegen hat Eysenck keine systematischen multimodalen Analysen durchgeführt, trotz der für seine Theorie wesentlichen physiologischen Korrelate der Persönlichkeitsfaktoren.

In einer wichtigen Übersicht hatten Seidenstücker und Baumann (1978) innerhalb der klinischen Psychologie einen Trend zur multimodalen Diagnostik gesehen und später (Seidenstücker & Baumann, 1987) sogar von einem *Standard* gesprochen. Auch in anderen Bereichen der Diagnostik wurde diese Idee aufgenommen (siehe Themenheft der *Diagnostica*, Fahrenberg, 1987). Kürzlich veröffentlichten Baumann und Stieglitz (2008) eine Bilanzierung: Multimodale Diagnostik – 30 Jahre später. Sie gehen davon aus, dass die verschiedenen Datenebenen gleichrangig, ohne Vorurteil angesehen werden sollten, wobei erst praktisch nach Aufgabenbereichen, Konstrukten, Funktionsbereichen und spezieller Eignung, beispielsweise Änderungssensitivität, zu differenzieren wäre. Zusammenfassend ergibt sich, dass Selbst- und Fremdbeurteilungen in vielen Bereichen der Klinischen Psychologie oft nur in mittlerer Höhe korrelieren, d.h. eine beträchtliche Anzahl von Einzelfällen verschieden (falsch?) klassifiziert würde, teils als Überschätzung, teils als Unterschätzung der psychischen Störungen. Ein Teil des Problems ist die extreme Anzahl psychologischer Verfahren. Nach Baumann und Stieglitz existieren mehr als 100 Skalen zur Diagnostik der Depressivität und etwa eine gleiche Anzahl zur Angstdiagnostik. Die Verhaltensmessungen und psychophysiologischen Methoden, die nur in einigen Teilbereichen eingesetzt werden können, wurden im Review von Baumann und Stieglitz ausgeklammert. Bei der Würdigung der Gesamtbilanz ist zu bedenken, dass sich viele der sogenannten Fremdbeurteilungen in mehr oder minder hohem Ausmaß auf die Selbstberichte der Patienten stützen, also methodisch konfundiert sind. Auch das AMDP-System zur Dokumentation psychiatrischer Befunde enthält eine große Zahl solcher kategorial unklaren Ratings: Von 100 Items beruhen 50 auf Selbstbeurteilungen, 20 auf Beurteilungen des Einstufers oder verlässlicher Auskunft Dritter und 30 auf beiden Informationsquellen (siehe Stieglitz, 2000).

Wenn die häufigen Divergenzen vor allem in der Klinischen Psychologie irritieren, könnte das mitbedingt sein durch die offensichtlichen Konsequenzen der diagnostischen Urteilsbildung; außerdem gibt es häufig divergente katamnesti-

sche Informationen. Hier kann aus einem fachlichen Dilemma auch ein berufsethisches Problem werden, denn es mangelt an Konventionen, wie mit den häufig zu erwartenden Diskrepanzen umzugehen ist. Baumann und Stieglitz ziehen ihr Fazit: „Auch wenn theoretisch begründbar, inhaltlich notwendig und methodisch nachweisbar eine multimodale Diagnostik notwendig ist, erweist sich deren Umsetzung bis zum heutigen Tag oft als schwierig …“ (2008, S. 199). Einen Grund, weshalb dieser Ansatz nicht die nötige Verbreitung findet, sehen sie darin, dass es für die Diagnostik – im Gegensatz zur Therapie – keine verbindlichen Leitlinien gebe.

Angstforschung

Eine besonders anspruchsvolle Strategie der Operationalisierung ist das *Drei-Ebenen-Konzept* des Assessment: introspektiv-verbale, behaviorale und physiologische Daten sind zu kombinieren. Die empirisch häufig auftretenden Divergenzen wurden auch als "response fractionation" bezeichnet. Bei unimodaler Diagnostik kann es u.U. zu schwerwiegenden Fehleinschätzungen kommen. Die gelegentlich auch als *Drei-Systeme-Konzept* bezeichnete Vorstellung scheint in diesem Bereich lange den Blick für die notwendige Analyse von *multiplen* Systemen und Reaktionsmustern verstellt zu haben. Heute sollte z.B. in der psychophysiologischen Angstforschung die breite Messung peripher-physiologischer Parameter selbstverständlich sein. In der Praxis der Diagnostik und Therapiekontrolle mangelt es jedoch an Regeln und Konventionen, wie die häufigen Diskrepanzen in der diagnostischen Urteilsbildung und Therapiekontrolle zu bewerten sind – sofern die Untersucher nicht der Einfachheit halber von vornherein auf physiologische Befunde verzichtet haben. Nicht einmal in der Terminologie hat es sich durchgesetzt, konsequent zwischen Angstgefühl, Angstverhalten und vegetativ-endokriner bzw. motorischer Angstphysiologie zu unterscheiden.

Lawyer und Smitherman (2004) analysierten Fachzeitschriften und stellten fest, dass die multimodale Diagnostik der Angst während der letzten Jahrzehnte abnahm und sich relativ mehr Autoren mit den Selbstberichten begnügten. Die Diagnostik von Angststörungen und Phobien ist ein anschauliches Beispiel, denn auf diesem Gebiet wurden die häufigen Diskrepanzen verschiedener Beschreibungsebenen seit Jahrzehnten untersucht, als wichtig bezeichnet, aber sehr häufig wieder ausgeklammert, weil es keine einfachen Lösungswege gibt. Andere Autoren gehen auf dieses Problem überhaupt nicht ein (Hoyer, Beauducel & Franke, 2002; Hoyer & Helbig, 2005). – Für die Theoretiker und für die Verhaltenstherapeuten bedeutet es gleichermaßen eine schwierige Herausforderung, wenn z. B. bei Patienten mit akuten Angststörungen und Phobien das Angstgefühl (subjektiv-verbale Ebene), das ängstliche Vermeidungsverhalten (behaviorale Ebene) und die vegetativ-endokrine Angsterregung (physiologische Ebene) weder zu Beginn, noch im Prozess oder am Ende einer Therapie konvergent sind. Der globale Begriff „Angst” könnte sehr irreführend sein. Statt die diskrepanten Informationen zu übergehen, ist vielfach eine gründlichere multimodale Untersuchungen angebracht. Erst solche Prozessanalysen könnten die offenen Fragen der differentiellen Indikation und Therapieevaluation beantworten. Therapieverläufe mit zunehmender bzw. hoher Kopplung (Konkordanz) von Funktionssystemen könnten im Vergleich zu diskordanten Prozessen effektiver und nachhaltiger sein (Fahrenberg & Wilhelm, 2009).

Die genannten Lehrbücher der Testmethodik vermitteln den Eindruck, dass im Konzept der multimodalen Diagnostik eher ein abstraktes Problem gesehen wird statt auch die praktisch-diagnostischen Konsequenzen aufgrund der Reviews von Baumann und Mitarbeitern zu erörtern — trotz der großen Tragweite dieser Ergebnisse und der anschließenden Kritik bzw. den Vorschlägen, an einem Standard zu arbeiten. Mühlig und Petermann (2006) skizzieren nur den Ansatz multimodaler Diagnostik, indem sie mögliche Datenquellen bzw. Methodentypen aufzeigen, ohne Schlussfolgerung zur Kombination, zu möglichen Standards, ohne Diskussion der notorischen Enttäuschungen und Widersprüche (vgl. jedoch Lösels, 1995, dringende Forderung, diese Ansätze zu verbessern). Auch im Grundwissen zur berufsbezogenen Eignungsbeurteilung nach DIN 33430 Westhoff et al., 2004) werden Abweichungen zwischen den verschiedenen Datenquellen (vgl. Schuler & Schmitt, 1987) zwar erwähnt, jedoch nicht genauer behandelt. Die optimistische Einschätzung der Konvergenzen in den Assessment Centern und Beobachterkonferenzen steht in starkem Kontrast zur Einschätzung der klinischen Diagnostik durch Baumann und Stieglitz.

Amelang und Schmidt-Atzert (2006) haben den Eindruck, dass institutionalisierte Diagnostik meist unimodal und individuelle Diagnostik meist multimodal ist. Bei mäßiger Konkordanz von Daten aus verschiedenen Quellen gebe es Möglichkeiten der Verbesserung: Aggregation über Messzeitpunkte, über Kriterumsbereiche und regressionsanalytische Kombination. „Als Leitsatz hat hierbei nach allgemeiner Auffassung zu gelten, dass ein Befund erst dann als gesichert anzusehen ist, wenn er durch mindestens 2 verschiedene Methoden möglichst unterschiedlicher Art bestätigt wird“ (S. 372). Bei divergierenden Befunden hat der Diagnostiker, zumindest in den Individualuntersuchungen, die „Möglichkeit, den Ursachen von Diskrepanzen durch Gespräche mit den Untersuchten, durch Analyse der verwendeten Methoden und beobachteten Prozesse oder Hinzuziehung weiterer Informationen nachzugehen“ (2006, S. 372). – Bei der Tragweite von Divergenzen bzw. Fehlentscheidungen wären hier möglichst genaue Prinzipien und an Gutachtenbeispielen erläuterte Regeln interessant.

12 Generalisierbarkeitstheorie

Die Generalisierbarkeitstheorie von Cronbach, Gleser, Nanda und Rajaratnam (1972) erweiterte die – abgesehen von Retest-Korrelationen – nur auf interne Reliabilitätsprüfung angelegte Testtheorie. Praktisch wichtiger ist die Zuverlässigkeit eines Tests in den Anwendungsbereichen. Mit welchem Risiko kann der individuelle Testwert auf andere Gelegenheiten, d.h. andere Zeitpunkte, Untersucher, Untersuchungsbedingungen, ähnliche Tests, Testmaterialien usw. verallgemeinert werden? Die verschiedenen Varianzquellen, die erwünschten und – je nach Perspektive – unerwünschten Varianz- (Fehler-) Quellen werden durch multifaktorielle Varianzanalysen geschätzt, um Entscheidungen zu erreichen. Die Generaliserbarkeitstheorie trifft sich hier mit der Frage nach ökologischer Validität.

Ein Seitenblick auf die Messung des Blutdrucks kann verdeutlichen, welche große praktische Bedeutung der Generalisierbarkeitstheorie im Sinne von Cronbach et al. zukommt. Für eine medizinisch notwendige, repräsentative Blutdruckmessung müssen berücksichtigt werden: verschiedene Geräte und Untersucher, verschiedene Gelegenheiten (Settings, Tätigkeiten, Bedingungen) und verschiedene Tageszeiten (siehe Fahrenberg, 2005; Gerin et al., 1998). Welches Minimum an solchen Bedingungen sichert eine akzeptable Generalisierbarkeit der Blutdruckbestimmung? Demgegenüber mangelt es in der Testpsychologie an solchen Generalisierbarkeitsstudien. Ebenso mangelt es an geeigneten messmethodischen Konventionen, beispielsweise den bekannten Bland-Altman-Diagrammen (1986) zur Kennzeichnung der Genauigkeit und der Präzision einer *physiologischen* Messung, und es fehlen systematische Vergleiche, d.h. institutionalisierte Verfahren zur Kontrolle der Standards – wie sie in der Medizin unentbehrlich sind.

Die Generalisierbarkeitstheorie und „multivariates Denken“ führten Wittmann zu den Prinzipien seiner an Brunswiks Prinzipien von Repräsentativität und Symmetrie orientierten „multivariaten Reliabilitätstheorie“. Diese innovative Konzeption ist für die Validierung psychologischer Tests (siehe oben zur Aggregation) und generell in der Evaluationsforschung wichtig. Während die klassische Reliabilitätstheorie in der Regel sehr ausführlich behandelt wird, fehlt meist die Generalisierbarkeitstheorie, und von der multivariaten Reliabilitätstheorie im Sinne Wittmanns wird höchstens eine der ersten Arbeiten, aber nicht die neuere Entwicklung zitiert (Wittmann & Schmidt, 1983; Wittmann, 1987, 1988, 2002; Wittmann, & Klumb, 2006; Beauducel et al., 2005).

13 Qualitätssicherung

Die Konventionen zur Qualitätssicherung bedeuten über die bisherigen Richtlinien des Testkuratoriums (2007) hinaus einen großen Fortschritt für die Evaluation psychologischer Untersuchungsverfahren (Kersting, 2006; Moosbrugger, Stemmler & Kersting, 2008). Auch die in vielen Abschnitten relativ ausführliche Darstellung des *Grundwissens für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (Westhoff et al., 2004) enthält wichtige Informationen und Forderungen. Allerdings führt hier die Ausrichtung auf Personalwesen und Eignungsbeurteilung zu einigen Übergeneralisierungen, die z.B. für Persönlichkeitsfragebogen und für andere Anwendungsfelder psychologischer Diagnostik unzutreffend sind. Auch deshalb ist es sinnvoll, auf die Unterschiede und Widersprüche aufmerksam zu machen.

Während in einem der Kapitel die Intervallskalierung zutreffend definiert und auf ihre inkonsequente, nur als Gewohnheit zu verstehende Verwendung hingewiesen wird, steht in einem anderen Kapitel, wie wichtig das Gütekriterium der „Skalierung“ sei (siehe oben). Gerade hier fehlt die grundsätzliche Unterscheidung zwischen Intelligenz- und Leistungstests bzw. Verhaltensbeobachtungen einerseits und Persönlichkeitsfragebogen bzw. ähnlichen Selbstbeurteilungsskalen andererseits. Die Folgen zeigen sich in den Abschnitten, in denen kommentarlos metrische Analysemethoden, sogar Rasch- und andere IR-Modelle unterschiedslos für alle psychologischen Testkonstruktionen empfohlen werden. Da auch bei mehreren anderen Themen die Gegenargumente aus den fachlichen Kontroversen fehlen, entsteht der Eindruck einer Harmonisierung. So ist auch die Beschreibung der multimodalen Diagnostik unzureichend. Die behaupteten Konvergenzen der Assessment-Center-Forschung sind schwächer als angedeutet und besagen auch weniger als gemeint, denn der Entscheidungsnutzen könnte nur zugleich mit der Schadensfunktion wirklich evaluiert werden: von den abgewiesenen, vielleicht viel höher qualifizierten Bewerbern gibt es in der Regel kein follow-up. Die optimistische Schilderung des Forschungsstandes darf, wie zuvor erläutert, keinesfalls auf andere Gebiete der psychologischen Diagnostik, z.B. auf die klinisch-psychologischen Diagnostik verallgemeinert werden.

Das Problem der Antworttendenzen sowie die situtationsbedingten und reaktiven Effekte in allen psychologischen Tests werden zu kurz behandelt. Im Hinblick auf das praktisch gegebene Interpretationsproblem divergenter Befunde in einem diagnostischen Urteil werden keine Prinzipien oder praktischen Konventionen erläutert. – Vielleicht kann von diesem als Einführung gedachten *Grundwissen* nicht erwartet werden, dass komplizierte Themen behandelt werden: u.a. das Konsistenz-Validitäts-Dilemma, die multivariate Perspektive auf das Assessment individueller Differenzen (im Raum von Personen, Situationen, Variablen, Wiederholungen, Variabilitäten und Konsistenzen) und die zugehörigen Assessmentstrategien für speziell gemeinte psychologische Konstrukte (siehe Stemmler, 1996, 2001), und darüber hi-

naus Wittmanns generalisierte Reliabilitätstheorie. Es mangelt jedoch bei einigen Themen an den notwenigen Vorbehalten und auch an Literaturhinweisen.

Gewiss sind die Beurteilungen nach dem neuen TBS-TK System notwendig, auch wenn deswegen – in paradoxer Weise – seriös publizierte Tests als problematischer erscheinen mögen als die im Personalwesen extrem verbreiteten, aber in ihrer Qualität wegen der Geheimhaltung der Gütekriterien nicht zu bewertenden, nicht-prüffähigen „freien" Tests zur Persönlichkeitstypisierung (Kersting, 2006). Der Prozess der Konsensbildung dauert an, denn es bestehen noch Einseitigkeiten und Widersprüche. So ist die Kritik an den Maßstäben der formalen Bewertung, u.a. im holländischen COTAN-System (vgl. Kersting, 2006) gerechtfertigt. Zumindest für die Normierung von Persönlichkeitsfragebogen ist der als gut bezeichnete Umfang von Normstichproben von N = 400 völlig unzureichend und ohne Bewertung der sozialwissenschaftlichen Qualität solcher Erhebungen hinsichtlich Quotierung und Repräsentativität wenig aussagekräftig, zumal es sich sehr häufig nur um Studierende oder um post hoc zusammengesetzte Gelegenheitsstichproben handelt. Erst ein Vielfaches (N > 2000, repräsentativ quotiert) dürfte für die deutschen Verhältnisse angemessen sein.

Die in DIN 33430 (vgl. Westhoff et al., 2004, S. 191) aufgestellte Forderung, der empirische Nachweis der Gültigkeit der Eichwerte nach 8 Jahren sei „absolut notwendig", ist eine fragwürdige Verallgemeinerung, zumindest für Persönlichkeitsfragebogen empirisch nicht belegt und deswegen willkürlich. Das FPI-R ist wohl das einzige deutsche Persönlichkeitsinventar mit einer Wiederholung der repräsentativen Normierung: im Abstand von 16 Jahren und ohne markante Unterschiede. Auch dieser Befund darf natürlich nicht verallgemeinerst werden. Die Bewertung der Reliabilitätskoeffizienten (> .90 als „gut") wird von Kersting zutreffend kritisiert, ohne jedoch methodisch u.a. auf das zugrundeliegende Konsistenz-Validitäts-Dilemma einzugehen und darzulegen, weshalb bei Persönlichkeitsfragebogen eine sehr hohe Konsistenz der Skalen unerwünscht sein kann.

Qualitätssicherungen durch unabhängige Test-Beurteilungen sind notwendig, weil sie oft für Entscheidungen von hoher Tragweite verwendet werden. Die Richtlinien fordern, dass die Qualitätssicherung den gesamten diagnostischen Prozess umfassen soll, d.h. nicht nur die eingesetzten Verfahren und deren Gütekriterien, sondern auch die Kompetenz der Anwender, die Qualifikation der beteiligten Personen und den Prozess der diagnostischen Urteilsbildung. Genannt werden u.a.: akkurate Testauswertung und Analyse der Testergebnisse, angemessene Interpretation der Testergebnisse, klare und exakte Weitergabe der Testergebnisse, Überprüfung der Angemessenheit eines Tests und seiner Anwendung. Hier wäre ein Codebook bzw. eine Sammlung geeigneter Beispiele notwendig, um durch Veröffentlichung und öffentliche Diskussion problematischer Geschehnisse zu lernen und Regeln zu entwickeln. Zu den hochproblematischen Anwendungen gehören Persönlichkeitstests oder klinisch-psychopathologische Skalen, z.B. Angstskalen, im Internet, aber auch die automatisierte Erstellung von Gutachten mittels Computerprogrammen ohne Beteiligung testmethodisch und diagnostisch kompetenter Fachpsychologen (vgl. International Guidelines on Computer-Based and Internet Delivered Testing).

Lässt es sich vorstellen, dass die abstrakten Prinzipien der Qualitätskontrolle erweitert werden und auch in der Psychologie eine öffentliche fachliche Diskussion über diagnostische Fehler wie in der Medizin stattfindet? Solange Menschen aktiv sind, wird es auch professionelle Fehler geben, aus denen gelernt werden kann, ohne dass gleich die Ethik-Kommissionen bemüht werden müssen. Zumindest für größere Institutionen ist das in Kliniken bereits eingeführte „Fehler-Reportingsystem" ein Vorbild. Das System ermöglicht es, Vorfälle, Probleme sowie Warnungen vor möglichen Risiken zu melden, innerhalb der Institution auch annoym.

14 Diagnostische Urteilsbildung: Datenintegration statt Interpretation?

Das diagnostische Gutachten (Amelang & Schmidt-Atzert, 2006; Fisseni, 2004) verknüpft die einzelnen Befunde zu einem diagnostischen Urteil. Hier muss erneut grundsätzlich zwischen Intelligenz- und Leistungstests und Persönlichkeitsfragebogen unterschieden werden. Von den einzelnen Validitätshinweisen ist die *empirische Validität der Testinterpretation* zu unterscheiden. Bei Persönlichkeitsfragebogen sind, noch kritischer als z.B. bei Intelligenz- und Leistungstests, der Kontext der Anwendung und die Möglichkeit von Antworttendenzen zu berücksichtigen. Im Interpretationsprozess der diagnostischen Urteilsbildung muss grundsätzlich zwischen der Selbstbeurteilung des Verhaltens und dem manifesten Verhalten unterschieden werden. Das Wissen über die methodischen Schwierigkeiten der Selbstbeurteilungen in Fragebogen (und Interviews) und die bekannten Methodenprobleme von Persönlichkeitsfragebogen machen deren fachlich adäquate Auswertung und Anwendung zu einer herausfordernden Aufgabe. Darüber hinaus existieren die kritischen Einsichten aus der multimodalen Diagnostik, die viele wichtige Praxisbereiche betreffen. Nach welchen Regeln sollen diese Daten verknüpft und für die Urteilsbildung interpretiert werden?

In seinen *Essentials of psychological testing* hatte Cronbach (1970) drei typische Strategien der Interpretation von Fragebogenergebnissen unterschieden: Der Persönlichkeitsfragebogen wird als Selbstbeschreibung angesehen, die Inhalte dienen einer psychologisch bzw. psychoanalytisch orientierten, inhaltlichen Interpretation oder werden aktuarisch, d.h.

aufgrund von Kriterienkorrelationen, verwendet. Als Anwendungsmöglichkeiten beschrieb er hautsächlich die Unterscheidung zwischen Patienten und Gesunden, die Suche nach auffälligen Personen für eine genauere Untersuchung, die Klassifikation von Patienten. Außerhalb des klinischen Bereichs dominieren die Vorhersage des Berufserfolgs bzw. des akademischen Erfolgs und die psychologische Begutachtung für institutionelle Entscheidungen. – Die Aufgaben der Klassifikation, die Selektion bzw. das Screening, die Vorhersage und Begutachtung werden auch heute unterschieden. Doch in der differentiellen Psychologie wurden seit Cattell zahlreiche typische Assessmentstrategien entwickelt, um die inter- und intraindividuelle Variabilität der multivariaten Datenbox perspektivisch zu differenzieren und die speziellen Konstrukte, Aggregate und Indizes zu unterscheiden. In den meisten Lehrbuchtexten zur psychologischen Diagnostik fehlen diese fortgeschrittenen Konzepte (und der Begriff Assessmentstrategie).

Praktische Regeln der Kombinatorik, die in der scheinbar als veraltet geltenden diagnostischen Psychologie (Heiß, 1964; vgl. Fahrenberg, 2002) intensiv trainiert wurden, werden heute kaum noch erwähnt. Nur als Hinweis sind zu nennen u.a. *Aspekte* wie Interpretationsebenen und Interpretationstiefe, Interpretationsdivergenz (Widerspruchsanalyse), Vermeidung von Vorurteilen und Reflexion der Abhängigkeiten, Kontrolle durch eine Interpretationsgemeinschaft, *Prinzipien* der Folgerichtigkeit, Ebenen des Kontextbezugs, Einpassung in Muster, Bedeutung von sog. Dominanten, *technische Regeln* über typische Marker für individuelle Auffälligkeiten, über Gewichtung und Kombinatorik usw. Demgegenüber werden heute oft nur die logischen Regeln, d.h. die konjunktive, additive und disjunktive Verknüpfung erläutert, oder allgemeine Unterschiede zwischen hypothesengeleiteter und explorativer Diagnostik. Oft wird der Informationsverarbeitungs-Prozess mit ausgedehnten Flussdiagrammen illustriert (z.B. Westhoff, Hagemeister & Strobel, 2006; Schmitt & Gschwendner, 2006). Viele Lehrbuchbeiträge über die Methodik psychologischer Gutachten bleiben sehr allgemein oder schildern primär die äußere Gestaltung und Kommunikation der Ergebnisse. In der Literatur gäbe es „verstreute empirisch gesicherte Regeln“ (Westhoff et al., 2006, S. 398) und in der entscheidungsorientierten Diagnostik wären diese Regeln zusammengetragen und stünden in Form von Checklisten zu Verfügung (Kubinger, 2003a; Westhoff & Kluck, 2003). Diese Checklisten enthalten eine große Anzahl von Gesichtspunkten mit sehr knappen Erläuterungen ohne übergreifende strategische Konzeption. Nur sehr selten werden, wie von Fisseni (2004), mehr Hinweise auf die notwendige Kombinatorik gegeben. Demgegenüber schreibt Westhoff (2004) von der notwendigen Komplexitätsreduktion, die vorab geplant werden und nachvollziehbar sein müsse, und weist allgemein darauf hin, dass Aussagen zu kombinieren sind: „Sollten sich Informationen widersprechen, so ist diese Tatsache zu berichten …“ – Aber sollen die Experten ggf. die Deutung der Widersprüche den Auftraggebern überlassen? Aus dieser Sicht scheinen psychologische Widersprüche zwischen Befunden nur Ausnahmen zu sein, für deren Interpretation keine methodischen Regeln oder Standards entwickelt werden müssen.

Das bekannte Schema der richtigen und falschen Diagnosen ist oft dargestellt, jedoch wird selten erläutert, weshalb für Cronbach neben der Nutzenfunktion auch die *Schadensfunktion* solcher professionellen Entscheidungen wichtig war. Die rationale Bewertung des Schadens verlangt natürlich das zu tun, was meistens fehlt, d.h. ein unverzerrtes follow-up auch der abgewiesenen, letztlich vielleicht viel geeigneteren Bewerber bzw. der falsch diagnostizierten oder falsch behandelten Patienten.

Zwar sind, hauptsächlich in der älteren Literatur, einige Hinweise auf die Kompetenzen des Diagnostikers zu lesen, doch bleiben diese sehr allgemein: „Der kompetente Psychodiagnostiker ist sich der verschiedenen diagnostischen Perspektiven und ihrer konzeptionellen und methodologischen Herausforderung bewusst: Neben dem bislang diskutierten ‚Datenverarbeitungsmodell‘ und der ‚psycho-sozio-ökologischen Perspektive‘ des diagnostischen Prozesses gibt es noch andere diagnostische Modell-Perspektiven …“ (Booth, 1995, S. 144). Fiedler (1984) schreibt über den Diagnostiker: „Anstatt mit formalen Methoden zu konkurrieren, sollte er seine Kräfte und seine Arbeitszeit für jene Probleme reservieren, die statt formaler Methoden den menschlichen Verstand benötigen, d.h. für die es weniger auf Präzision und absolute Reliabilität ankommt als auf Kreativität, Flexibilität, soziale Intelligenz, Improvisation, komplexe Mustererkennung und nicht zuletzt ‚Sprachgefühl‘“ (S. 309).

Die Kontroverse zwischen statistischer und „klinischer“ Urteilsbildung (Meehl, 1954; Wiggins, 1973) und die mögliche *Kombination* beider Strategien haben zeitweilig großes Interesse gefunden. Ein typischer Untersuchungsansatz war damals die statistische Auswertung bzw. die klinisch-diagnostische Interpretation von MMPI-Profilen im Vergleich zur „richtigen“ psychiatrischen Diagnose – ein Evaluationsverfahren, das heute auf *beiden* Seiten überholt ist. Diese Auseinandersetzung verlangt Differenzierungen, u.a. zwischen den statistisch evaluierbaren Prognosen und den diagnostischen Urteilen im Einzelfall, und vor allem eine gründlichere Evaluation der Kriterienvalidität, der Konstruktoperationalisierung und der Datenaggregationen. Die Kontroverse ist keineswegs befriedigend geklärt. Dennoch taucht dieses Thema in den Lehrbüchern nur noch am Rande auf, entweder als ältere Kontroverse in einem historischen Rückblick oder pauschal als der Gegensatz nomothetischer und idiographischer Auffassungen (Fisseni, 2004; Heil, 1995; Petermann, 1995) ‚statt die Chancen der strategischen Kombination darzustellen. Diese Debatte müsste auf dem heutigen Stand fortgesetzt und an realistischen Lehrbeispielen vertieft werden (z.B. Dahle, 2005).

Deutet heute vielleicht der technisch klingende Begriff *Datenintegration* darauf hin, dass von vornherein Homogenes, sehr Ähnliches zu kombinieren ist wie bei einer mathematischen Funktion? Kann es auch auf die psychologische Interpretation tiefreichender Widersprüche von Befunden ankommen und um den psychologischen Kontext gehen? Konkret bleibt die heuristisch-beziehungsstiftende Aufgabe der Interpretation oft unerwähnt. Dementsprechend wird der Begriff Interpretation (auch in den Sachregistern erscheint höchstens die Interpretationsobjektivität) vermieden und damit das umfangreiche System von traditionellen Strategien und die Regeln der psychologischen Interpretation ausgeklammert, wenn nicht vergessen. Die Flussdiagramme und langen Checklisten deuten an, dass auch nach einer Debatte von ca. 40 Jahren die Hoffnung auf eine weitgehende Algorithmisierung und ein intelligentes Computerprogramm zur diagnostischen Urteilsbildung fortbesteht. Auch das Aufzählen von möglichen Fehlern und Verzerrungen im Prozess der diagnostischen Urteilsbildung (u.a. Westhoff & Kluck, 2003) demonstriert, wie wichtig methodenkritische Reflexionen sind.

Aus dem Nachweis von möglichen Denkfehlern der Urteilsbildung oder der „Entmystifizierung" des Diagnostikers, folgt natürlich nicht, dass die traditionellen Strategien der Interpretation insgesamt falsch oder irrational waren. So hielten es die Herausgeber des neuen *Handbuchs der Psychologischen Methoden und Evaluation* nicht für wichtig, ein Kapitel über die Prinzipien Psychologischer Interpretation aufnehmen. Damit vertieft sich der Eindruck, dass ein Teil der Methodik aufgegeben wird, die bereits für Wilhelm Wundt und dessen psychologische Interpretationslehre wesentlich war.

Die einseitig kognitionspsychologisch-formale Perspektive der Informationsverarbeitung berücksichtigt zu wenig, dass ein persönlichkeitstheoretisches Bezugssystem vorhanden sein muss. Die diagnostische Urteilsbildung kann ja nicht allein als *theoriefreie* Datenverarbeitung für eine bestimmte Aufgabenstellung ablaufen. Die Urteilsbildung des Diagnostikers findet unvermeidlich – ausgesprochen oder unausgesprochen – zugleich in einem persönlichkeitstheoretischen Bezugssystem statt. Dieser Interpretationsrahmen wird zu selten diskutiert als ob die diagnostischen Fragestellungen psychologisch isoliert zu beantworten wären. Wie könnten die die berufsbezogenen und die pädagogischen Beurteilungen oder die ätiologischen und die therapeutischen Konzeptionen *ohne theoretischen Bezug* auf zugrundlegende Annahmengefüge oder „Funktionsmodelle" von Einstellungen, Motiven und Persönlichkeitseigenschaften getroffen werden? Die Implikationen solcher Vorentscheidungen für die psychologische Diagnostik sind noch kein Thema der Lehrbücher und wohl auch kaum der akademischen Lehre.

Die skizzierten Tendenzen wären dann als Fehlentwicklung anzusehen, wenn sie – in einer ohnehin reduzierten fachlichen Ausbildung – generell zu einer Unterbewertung der grundlegenden Prinzipien der differentiellen Psychologie und Persönlichkeitstheorie führen und zu einer Kürzung des gründlichen Trainings in psychologischer Diagnostik.

15 Strategische Konsequenzen – Ausblick von einer mittleren Position

Selbstberichte in Fragebogen und Interviews sind zweifellos geeignet – und unersetzlich – wenn die subjektive (mentale) Repräsentation des Erlebens, der Einstellungen und des Verhaltens erfasst werden sollen. Solche Selbstbeurteilungen sind am leichtesten zu erhalten, standardisiert und testökonomisch, sie haben eine Augenscheinvalidität. Wer auf Persönlichkeitsfragebogen verzichtet, verliert viele – auch durch ein langes Interview nur bedingt zu ersetzende – Informationen. Aber diese Selbstbeurteilungen können die Untersuchung des aktuellen, manifesten Verhaltens im Alltag nicht ersetzen, sondern die Verhaltensunterschiede nur zu erläutern helfen. Grundsätzlich kann zwischen verschiedenen Absichten und Strategien unterschieden werden.

Strategien und Evaluationen

Cronbach (1970, S. 555f.) argumentiert, dass ein Persönlichkeitsfragebogen deskriptiv – wie ein Spiegel – dem Untersuchten hilft, von sich selbst ein Bild zu machen, auch im Vergleich zu anderen Menschen. Ein psychologischer Berater kann die Testwerte als Ausgangsinformationen für das weitere Kennenlernen verwenden und dabei Irrtümer der Testinterpretationen verringern. In wie weit die Beschreibung aufgrund der Testwerte mit dem Selbstbild übereinstimmt, kann mittels anderer Informationen oder eventuell durch weitere Fragen, wie jemand eigentlich sein möchte, vertieft werden. Die Eigenschaftsbezeichnungen der Skalen zu verwenden oder das Testprofil zu zeigen, sei nicht ratsam.

Diese *explorative Strategie* eignet sich für einen biographisch-diagnostischen Ansatz in der psychologischen Beratung und Fallarbeit. Auch die gegenüber Messmodellen und statistischen Auswertungen skeptischen Psychologen werden einen Persönlichkeitsfragebogen als einen besonders strukturierten Teil einer breiteren psychologischen Untersuchung akzeptieren können. In diesem Zusammenhang könnten am ehesten auch die möglichen Antworttendenzen erkannt und in das Gesamtbild einbezogen werden.

Die Vorzüge der Persönlichkeitsfragebogen werden jedoch *strategisch* erst durch den *Vergleich* der individuellen Testprofile mit den Normwerten genutzt. Entsprechen die individuellen Eigenschaftsschilderungen der Durchschnittsbevölkerung oder weichen sie deutlich ab? Deshalb sind sehr große bevölkerungsrepräsentative Stichproben gerade für die Konstruktion und die differenzierte Normierung von Persönlichkeitsfragebogen unverzichtbar. (Außerdem ermöglichen große Stichproben die zur Absicherung von Forschungsarbeiten notwendige, simultane statistische Kontrolle einer Anzahl soziodemographischer Merkmale, wie es mit der Methode statistischer Zwillinge geschieht.) Die Wiederholung der repräsentativen Normierung des FPI-R hatte das überraschende Ergebnis, dass die Itemmuster und die Testwert-Verteilungen (Normen) nach 17 Jahren nur geringfügig voneinander abwichen. Die FPI-Skalen repräsentieren also psychologische Konstrukte, die offensichtlich in den Selbstbeschreibungen der Durchschnittsbevölkerung einen herausragenden und überdauernden Platz haben. Diese Konzepte überlappen sich, wie sich zeigen liess, deutlich mit den Konzepten von direkten Selbsteinstufungen und Fremdbeurteilungen. Es handelt sich um robuste Dimensionen eines differentiell-psychologischen Beschreibungssystems. Aus der Einsicht in die *strukturelle Subjektivität* dieser vielschichtigen Selbstbeurteilungen und Selbstberichte folgt noch nicht, dass alle Antworten wirklichkeitsfern sind, d.h. die gesamte oder ein überwiegender Teil der Varianz ohne Bezug zum manifesten und künftigen Verhalten ist. Die extremen sozialpsychologisch-konstruktivistischen Erklärungsversuche, dass es sich grundsätzlich *nur* um Stereotypien der Alltagspsychologie handele, reichen hier nicht aus. Es gibt systematische empirische Zusammenhänge zwischen den Testwerten von Persönlichkeitsinventaren und Statusmerkmalen soziodemographischer, klinischer und beruflicher Art, zu anderen selbstberichteten Inhalten, zu Fremdeinstufungen und zu vielen empirisch prüfbaren Daten und Verhaltensweisen (Kapitel 6 und 7).

Ein standardisierter Fragebogen ist unentbehrlich, um solche Fragen zu beantworten, wie typisch in der Bevölkerung bestimmte Selbsteinschätzungen sind, z.B. der Lebenszufriedenheit, der Beanspruchung (des Stresserlebens), der körperlichen Beschwerden und Gesundheitssorgen, der Lebenszufriedenheit oder der Extraversion-Introversion, und wie ausgeprägt dieses Bild vergleichsweise bei einer einzelnen Person oder in Personengruppen ist. Zur psychologischen Diskrimination von klinischen und anderen Gruppen und zur Unterscheidung von Gesunden liegen wohl die meisten Untersuchungen vor. Die *Strategie des Screening* ist eine typische Anwendung eines Persönlichkeitsfragebogens, z.B. um in klinischen Einrichtungen auf testökonomische, und deshalb organisatorisch einfache Weise auf solche Personen aufmerksam zu werden, die in einem anschließenden Schritt genauer untersucht werden sollten. Ein solches Screening in einer mehrstufigen Strategie kann in vielen Institutionen nützlich sein, beispielsweise in der Psychosomatischen Klinik, in der Neurologischen Reha-Klinik oder in einer allgemeinen Rehabilitationseinrichtung für chronisch Kranke.

Die Grenzen der Fragebogenmethodik werden deutlicher, wenn nach den direkten Verhaltenskorrelationen gefragt wird. Wer auf die psychologischen Informationen aus Persönlichkeitsfragebogen nicht verzichten will, muss multimodal vorgehen und dabei die strukturelle Subjektivität von Selbstberichten und Selbstbeurteilungen akzeptieren. Selbstbeurteilungen und verbale Auskünfte über eigenes Verhalten sind eben keine objektiven Verhaltensdaten. Die deutlich begrenzten und z.T. widersprüchlichen Resultate im Hinblick auf objektivierbare Kriterien sind zwar seit langer Zeit bekannt, scheinen allerdings in manchen Anwendungssituationen unterschätzt oder ausgeklammert zu werden. Die monomethodisch angewendeten Fragebogen sind für einige psychologische Fragestellungen zweckmäßig, sollten aber sonst nur ein Bestandteil einer multimodalen Diagnostik sein, um krasse Fehlbeurteilungen zu vermeiden. – Diese Beschränkungen bewusst zu halten, bleibt wichtig, um zur vorsichtigen Interpretation zu motivieren. Weder eine unkritische behaviorale Interpretation noch die ausschließlich selbsttheoretische oder alltagspsychologische Deutung treffen die Eigenart von Persönlichkeitsfragebogen.

Strategische Schwierigkeiten grundsätzlicher Art bestehen in der Persönlichkeitsforschung, die sich ausschließlich auf Fragebogen stützt wie es noch häufig der Fall ist. Offensichtlich steht die Persönlichkeitsforschung weiterhin vor der von Fiske erkannten konzeptuellen und methodischen Herausforderung, anstelle der globalen Eigenschaftskonstrukte wesentlich kleinere Beschreibungseinheiten zu entwickeln. Auch die neue Methodik des *ambulanten Assessment* führt zu der Erfahrung, wie häufig Diskrepanzen zwischen den Beschreibungsebenen von Emotionen, körperlicher Aktivität, Symptomen und Verhaltensstörungen sind. Diese Befunde sind jedoch keineswegs spezifisch für die Persönlichkeitsfragebogen oder die Einstufungen des Befindens und Erlebens. Sie sind typisch für eine psychologische Diagnostik, die subjektiv-verbale, behaviorale und physiologische Methoden kombiniert. Die notorischen Diskrepanzen sind aus der behavioralen und psychophysiologischen Diagnostik chronischer Angststörungen und Phobien seit langem bekannt. Gelegentlich wird vorschlagen, in der psychologischen Diagnostik nicht uni-methodisch vorzugehen, sondern multimethodisch verschiedene Verfahren zur Absicherung zu verwenden. Diese Empfehlung müsste noch entschiedener lauten: nicht *multi-methodisch* (z.B. mit einer Sammlung verschiedenster Fragebogen), sondern *multi-modal* auf kategorial verschiedenen Datenebenen. Die *multimodale Strategie* ist jedoch gegenwärtig weder testtheoretisch noch in hinsichtlich praktisch-diagnostischer Standards hinreichend ausgearbeitet.

Die Fragebogenantworten konsequent als Selbstbeurteilungen anzusehen, ändert nichts an der Möglichkeit, solche Aussagen sinnvoll mit den Daten von Bezugsgruppen und Normen sowie Verhaltensinformationen zu vergleichen, verlangt jedoch eine andere Interpretationsweise und – wo immer möglich – eine *multimodale Strategie*, die möglichst viele Kontextinformationen einzubeziehen versucht. Eine mittlere Position einzunehmen, bedeutet auch, bei pauschaler Kritik an dieser Methodik nach der wissenschaftlichen Qualität der Argumente zu fragen und vor allem danach: Welche Alternative für die Persönlichkeitsforschung und Persönlichkeitsdiagnostik schlagen die Kritiker beim Verzicht auf die Fragebogenmethodik vor?

Zur Testkonstruktion

Es bestehen begründete Zweifel, ob die propagierten Latent Trait-Modelle für die facettenreichen und deshalb nicht perfekt „homogenisierbaren" Persönlichkeitskonstrukte gerechtfertigt sind. Aus einer mittleren Position, welche die Argumente der Kritiker und der Befürworter aufzunehmen sucht, werden solche Konzepte nicht grundsätzlich abgelehnt, sondern als *heuristische Strategien der Testanalyse* angesehen. Künftig werden vielleicht geeignete Latent Class Modelle und – über die konventionelle Clusteranalyse hinaus – andere fortgeschrittene Verfahren der Musteranalyse für nominale Daten entwickelt und auch erprobt sein. Diese voraussetzungsärmeren Analysen könnten u.U. zu überzeugenderen Strukturaussagen und entsprechenden Item-Aggregaten in den Persönlichkeitsfragebogen führen.

Die allgemeine Kritik an Fragebogen dauert, seit Wilhelm Wundt, nun mehr als 100 Jahre an. Dabei rückt das Thema der Antworttendenzen in Persönlichkeitsfragebogen oft so sehr in den Vordergrund der Methodenkritik, dass andere Perspektiven zu kurz kommen. Die möglichen Effekte formaler Antworttendenzen oder der sozialen Erwünschtheit dürfen nicht verharmlost werden. Es ist jedoch immer deutlicher geworden, dass die operationale Definition und Differenzierung solcher Tendenzen oder die Abgrenzung von den typischen Persönlichkeitsmerkmalen nicht erreicht wurde. Abgesehen von der Anwendung von Persönlichkeitsfragebogen zum Zweck eines ersten Screenings, gehört deshalb zur Interpretation eines Testprofils der psychologische Kontext und eine *zu trainierende Kompetenz in der psychologischen Interpretation* von konvergenten und von divergenten diagnostischen Informationen.

Ausblick

Strategisch folgt aus diesen Überlegungen, primär an der Umsetzung der methodischen Fortschritte in dieser Richtung zu arbeiten, u.a. unter den Stichworten: Generalisierbarkeit und ökologische Validität, repräsentative, symmetrische Validierungsforschung und multimodale Diagnostik. Wichtiger als konventionelle Gruppenvergleiche, Inter-Test-Korrelationen mit anderen Persönlichkeitsfragebogen oder Korrelationen mit anderen Informationen sind *anspruchsvollere Assessmentstrategien* und die Prüfung des Entscheidungsnutzens solcher Persönlichkeitsdaten im Hinblick auf praktisch relevante Kriterien. Dazu sollten auch überzeugende Replikationen wichtiger Befunde mit relativ großer Personenzahl zunächst innerhalb und dann zwischen den Arbeitsgruppen, gehören. Da solche systematischen Replikationen in der Psychologie (im Unterschied zu den Naturwissenschaften) weithin fehlen bzw. als nicht besonders wichtig oder als nicht kreativ zu gelten scheinen, ist ein Seitenblick auf die Medizin angebracht. Hier führen die international und multizentrisch durchgeführten Studien, z.B. über bereits eingeführte Medikamente, oft zu überraschenden Ergebnissen und gelegentlich sogar zur dramatischen Beendigung solcher Studien. Weshalb sollte die Evaluation psychologischer Prädiktoren und Interventionen einfacher sein?

Überzeugende Validierungsstudien haben eine höhere Priorität als die Fokussierung auf Messmodelle oder die Person-Situation-Debatte anhand dafür ungeeigneter Fragebogendaten oder die essentialistischen Behauptungen über die richtige Anzahl von hauptsächlichen Persönlichkeitsfaktoren. Die eigenen Bemühungen und problematischen Erfahrungen führten zu der Schlussfolgerung und dem wiederholten Plädoyer, dass inhaltlich und statistisch überzeugende Evaluationen nur *innerhalb* großer Institutionen zweckmäßig und aussichtreich sind. Nur dort sind auch die notwendigen follow-up Informationen bzw. Katamnesen zu gewinnen, die zur Beurteilung des Entscheidungsnutzens notwendig sind. Die bisherigen Untersuchungen sind in ihrer Überzeugungskraft beeinträchtigt, da solche Informationen in der Regel fehlen und deswegen neben der Nutzenfunktion die zugehörige Schadensfunktion der diagnostischen Entscheidungen bzw. des Screenings unbekannt bleibt. Die Absicht der Qualitätssicherung stößt hier an Grenzen, die gewöhnlich nicht erörtert werden. Falls einige Institutionen, Kliniken, Rehabilitationseinrichtungen, Organisationen, Betriebe, staatliche Verwaltungen tatsächlich Persönlichkeitsdaten und Persönlichkeitsfragebogen *intern* evaluieren, was anzunehmen ist, bleiben die Ergebnisse in der Regel unzugänglich. Dies kann mit dem Datenschutz zusammenhängen, obwohl es geeignete Wege der Anonymisierung gibt, oder mit der geringen Bereitschaft, solche Erfahrungen weiterzugeben. – Ob das zunehmende Bewusstsein von Qualitätssicherung diese Grenzen im gemeinsamen Interesse überwinden wird?

Literaturverzeichnis

weitere Hinweise siehe Manual

Fahrenberg, J., Hampel, R. & Selg, H. (2009, im Druck). Das Freiburger Persönlichkeitsinventar (revidierte Fassung FPI-R und teilweise geänderte Fassung FPI-A1 (8. Aufl.). Göttingen: Hogrefe.

Amelang, M., Bartussek, D., Stemmler, G. & Hagemann, D. (2006). Differentielle Psychologie und Persönlichkeitsforschung (6. Aufl.). Stuttgart: Kohlhammer.

Amelang, M., Schäfer, A. & Yousfi, S. (2002). Comparing verbal and non-verbal personality scales: Investigating the reliability and validity, the influence of social desirability, and the effect of fake good instructions. Psychologische Beiträge, 44, 24-41.

Amelang, M. & Schmidt-Atzert, L. (2006). Psychologische Diagnostik und Intervention (4. Aufl.). Heidelberg: Springer.

Andresen, B. & Beauducel, A. (2008). TBS-TK Rezension: NEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung (NEO-PI-R). Report Psychologie, 11/12, 543-544.

Angleitner, A. (1976). Methodische und theoretische Probleme bei Persönlichkeitsfragebogen unter besonderer Berücksichtigung neuerer deutschsprachiger Fragebogen. Unveröff. Habilitationsschrift. Universität Bonn.

Angleitner, A. & Wiggins, J. S. (Eds.). (1986). Personality assessment via questionnaire. Current issues in theory and measurement. Berlin: Springer.

Asendorpf, J.B. (2007). Psychologie der Persönlichkeit (5.Aufl.). Berlin: Springer.

Assmann, B., Dähne, A., Hinz, A. & Ettrich, C. (2003). Essstörungsspezifische Psychodiagnostik bei Anorexia nervosa und Bulimia nervosa. Suchtmedizin in Forschung und Praxis, 5(3), 185-191.

Austin, E.J., Deary, I.J., Gibson, G.J., McGregor, M.J. & Dent, J.B. (2006). Individual differences in response to scale use: Mixed Rasch modelling of response to NEO-FFI items. Personality and Individual Differences, 40, 1235-1245.

Internet-Dokument - (Zugriff am 10.3.2009) http://www.soz.jku.at/Portale/Institute/SOWI_Institute/Soziologie/aes/content/e50/e1512/e2128/StatisticalMatching_ger.pdf

Baumann, U. & Stieglitz, R.D. (2008). Multimodale Diagnostik – 30 Jahre später. Zeitschrift für Psychiatrie, Psychologie und Psychotherapie, 56, 191-202.

Beauducel, A., Biehl, B., Bosnjak, M., Conrad, W., Schönberger, G. & Wagener, D. (Hrsg.). (2005). Multivariate research strategies. Festschrift in Honor of Werner W. Wittmann. Aachen: Shaker.

Becker, P. (1996). Wie big sind die Big Five? Zeitschrift für Differentielle und Diagnostische Psychologie, 17, 209–221.

Becker, P. (2000). Die „Big Two“. Seelische Gesundheit und Verhaltenskontrolle: zwei orthogonale Superfaktoren höherer Ordnung? Zeitschrift für Differentielle und Diagnostische Psychologie, 21, 113–124. Becker, P. (2003a). Trierer Integriertes Persönlichkeitsinventar TIPI. Göttingen: Hogrefe.

Becker, P. (2003b). Persönlichkeitsfragebogen. In: K.D. Kubinger & S. Jäger (Hrsg.). Schlüsselbegriffe der psychologischen Diagnostik (S. 332-337). Weinheim: Beltz.

Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 8 (1), 307-310.

Booth, J.F. (1995). Kompetenz. In S. Jäger & F. Petermann (Hrsg.). Psychologische Diagnostik. (3. Aufl.). (S. 138-146). Weinheim: Beltz, Psychologie-Verlags-Union.

Borg, I. & Staufenbiel, T. (2007). Lehrbuch Theorien und Methoden der Skalierung (4. Aufl.). Bern: Huber.

Borkenau, P. (2006). Selbstbericht. In: F. Petermann & M. Eid, M. (Hrsg.). Handbuch der psychologischen Diagnostik (S. 135-143). Göttingen: Hogrefe.

Borkenau, P. & Ostendorf, F. (1989). Untersuchungen zum Fünf-Faktoren-Modell der Persönlichkeit und seiner diagnostischen Erfassung. Zeitschrift für Differentielle und Diagnostische Psychologie, 10, 239–251.

Borkenau, P. & Ostendorf, F. (1993). NEO-Fünf-Faktoren Inventar nach Costa und McCrae. Handanweisung. Göttingen: Hogrefe.

Bortz, J., Lienert, G.A. & Boehncke, K. (2000). Verteilungsfreie Methoden in der Biostatistik (2. Aufl.). Berlin: Springer.

Brunswik, E. (1956). Representative design and probabilistic theory in a functional psychology. Psychological Review, 62, 193-217.

Bühner, M. (2006). Einführung in die Test- und Fragebogenkonstruktion (2. Aufl.). München: Pearson Studium.

Campbell, D.T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cattell, R.B. (1957). Personality and motivation. Structure and measurement. New York: World Book.

Cattell, R. B. & Warburton, F. W. (1967). Objective personality and motivation tests: a theoretical introduction and practical compendium. Urbana: University of Illinois Press.

Cronbach, L.J. (1970). Essentials of psychological testing (3rd ed.). New York: Harper and Row.
Cronbach, L.J., Gleser, G. C., Nanda, H., Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.
Dahle, K.P. (2005). Psychologische Kriminalprognose. Wege zu einer integrativen Methodik für die Beurteilung der Rückfallwahrscheinlichkeit bei Strafgefangenen. Herbolzheim: Centaurus.
Dawes, R. M. (1977). Grundlagen der Einstellungsmessung. Weinheim: Beltz.
Dawes, R. M., Faust, D. & Meehl, P. E. (1989). Clinical versus actuarial judgment. Science, 243, 1668-1674.
Eid, M. & E. Diener, E. (Eds.). (2005). Handbook of multimethod measurement in psychology. Washington, D.C.: American Psychological Association.
Eid, M., Nussbeck, F.W. & Lischetzke, T. (2006). Multitrait-Mulimethod-Analyse. In: F. Petermann & M. Eid, M. (Hrsg.). Handbuch der psychologischen Diagnostik (S. 332-345). Göttingen: Hogrefe.
Fahrenberg, J. (Hrsg.). (1987). Multimodale Diagnostik. [Themenheft]. Diagnostica, 33 (3).
Fahrenberg, J. (2002). Psychologische Interpretation. Biographien - Texte - Tests. Bern: Huber.
Fahrenberg, J. (2005). Representative design and the laboratory field-issue. In: A. Beauducel, B. Biehl, M. Bosniak, W. Conrad, G. Schönberger & D. Wagener (Eds.). Multivariate research strategies. Festschrift in Honor of Werner W. Wittmann (S. 237-260). Aachen: Shaker.
Fahrenberg, J. (2008a). Die Wissenschaftskonzeption der Psychologie bei Kant und Wundt. In: e-Journal Philosophie der Psychologie -Online10, 1-22. Zugriff am 1.3.2009. http://www.jp.philo.at/texte/FahrenbergJ2.pdf [35 Seiten, 115 Literaturhinweise, 199 KB].
Fahrenberg, J. (2008b). Wilhelm WUNDTs Interpretationslehre [43 Absätze]. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, Online 9(3), Art. 29. Zugriff am 1-3-2009. http://nbnresolving.de/urn:nbn:de:0114-fqs0803291.
Fahrenberg, J. & Wilhelm, F. H. (2009). Psychophysiologie und Verhaltenstherapie. In: J. Margraf & S. Schneider (Hrsg.). Lehrbuch der Verhaltenstherapie (3. Aufl.). (S. 163-179). Berlin: Springer.
Fiedler, K. (1984). Diagnostische Fähigkeiten und diagnostische Erfahrung. In: R.S. Jäger, A. Mattenklott & R.D. Schröder (Hrsg.). Diagnostische Urteilsbildung in der Psychologie. Grundlagen und Anwendungen (S. 303-327). Göttingen: Hogrefe.
Fishbein, M. & Ajzen, I. (1974). Attitudes towards objects as predictors of single and multiple behavioral criteria. Psychological Review, 81, 59-74.
Fiske, D.W. (1978). Strategies for personality research. San Francisco: Jossey-Bass.
Fiske, D.W. (1987). On understanding our methods and their effects. Diagnostica, 33, 188-194.
Fiske, D.W. & Pearson, P. H. (1970). Theory and techniques of personality measurement. Annual Review of Psychology, 21, 49-86.
Fisseni, H.-J. (2004). Lehrbuch der Psychologischen Diagnostik (3. Aufl.). Göttingen: Hogrefe.
Flick, U., von Kardorff, E. & Steinke, I. (Hrsg.). (2000). Qualitative Forschung – Ein Handbuch. Hamburg: Rowohlt.
Forman, A.K. (2002). Identifying types, response errors, and unscalable respondents from personality questionnaires. Psychologische Beiträge, 44, 78-93.
Funder, D.C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. Journal of Research in Personality, 40, 21-34.
Gerin, W., Christenfeld, N., Pieper, C., Su, O., Stroessner, S. J., Deich, J. & Pickering, T. G. (1998). The generalizability of cardiovascular responses across settings. Journal of Psychosomatic Research, 44, 209-218.
Gorin, A.A. & Stone, A.A. (2001). Recall biases and cognitive errors in retrospective self-reports: A call for momentary assessments. In: A. Baum, T. A. Revenson & J. E. Singer (Eds.). Handbook of health psychology (pp. 405-413). New Jersey: Erlbaum.
Guilford, J.P. (1959). Personality. New York: McGraw Hill.
Hehl, F.J. & Hehl, R. (1975). Persönlichkeits-Skalen System PSS 25. Weinheim: Beltz.
Heil, F.H. (1995). Klinische versus statistische Urteilbildung. In: S. Jäger & F. Petermann (Hrsg.). Psychologische Diagnostik (3. Aufl.). (S. 39-42). Weinheim: Beltz, Psychologie-Verlags-Union.
Heiß, R. (1964). Technik, Methodik und Problematik des psychologischen Gutachtens. In: R. Heiß (Hrsg.). Handbuch der Psychologie. Band 6. Psychologische Diagnostik (S. 975-995). Göttingen: Hogrefe.
Heymans, G. & Wiersma, E. (1906). Beiträge zu einer speziellen Psychologie auf Grund einer Massenuntersuchung. Zeitschrift für Psychologie und Physiologie des Sinnesorgane, 42, 81-127.
Hofstede, G.H. (2006). Culture's consequences: comparing values, behaviors, institutions, and organizations across nations (2nd ed.). Thousand Oaks, Calif.: Sage.
Hoyer, J., Beauducel, A. & Franke, G.H. (2002). Kriterien der Angstdiagnostik – Woran der Anwender auch denken muss. In: Hoyer, J. & Margraf, J. (Hrsg.). Angstdiagnostik – Grundlagen und Testverfahren (S. 77-90). Berlin: Springer.
Hoyer, J. & Helbig, S. (2005). Diagnostische Verfahren bei Angststörungen. Psychotherapie im Dialog, 6(4), 425-430.
Ille, R., Lahousen, T., Rous, F., Hofmann, P. & Kapfhammer, H. P. (2005). Persönlichkeitsprofile und psychische Abweichungen bei psychiatrisch-forensisch begutachteten Straftätern. Der Nervenarzt, 76(1), 52-60.

Jäger, S. & Petermann, F. (Hrsg.). (1995). Psychologische Diagnostik. (3. Aufl.). Weinheim: Beltz, Psychologie-Verlags-Union.

Jüttemann, G. (1991). Systemimmanenz als Ursache der Dauerkrise „wissenschaftlicher" Psychologie. In: G. Jüttemann, M. Sonntag & C. Wulf (Hrsg.). Die Seele. Ihre Geschichte im Abendland (S. 340-363). Weinheim: Psychologie Verlags Union.

Jüttemann, G. (2004). Psychologie als Humanwissenschaft. Ein Handbuch. Göttingen: Vandenhoeck & Ruprecht.

Jüttemann, G. (Hrsg.). (2006). Wilhelm Wundts anderes Erbe. Ein Missverständnis löst sich auf (S. 13-30). Göttingen: Vandenhoeck & Ruprecht.

Kenrick, D.T. & Funder, D.C. (1988). Profiting from controversy. Lessons from the person-situation debate. American Psychologist, 43, 23-34.

Kersting, M. (2006). Zur Beurteilung der Qualität von Tests: Resümee und Neubeginn. Psychologische Rundschau, 57, 243-253.

Krauth, J. (1995). Testkonstruktion und Testtheorie. Weinheim: Beltz, Psychologie-Verlags-Union.

Kubinger, K.D. (Ed.). (2002a). Special issue: Personality questionnaires: Some critical points of view. Psychologische Beiträge, 44, 3-158.

Kubinger, K.D. (2003a). Gutachten, psychologisches. In: K.D. Kubinger & S. Jäger (Hrsg.). Schlüsselbegriffe der psychologischen Diagnostik (S. 187-194). Weinheim: Beltz.

Kubinger, K.D. (2003b). Gütekriterien. In: K.D. Kubinger & S. Jäger (Hrsg.). Schlüsselbegriffe der psychologischen Diagnostik (S. 195-201). Weinheim: Beltz.

Kubinger, K.D. & Draxler, C. (2007). Probleme bei der Testkonstruktion nach dem Rasch-Modell. Diagnostica, 53, 131-143.

Kubinger, K.D. & Jäger, S. (Hrsg.). (2003). Schlüsselbegriffe der psychologischen Diagnostik. Weinheim: Beltz.

Külpe, O. (1920). Vorlesungen über Psychologie (hrsg. von Karl Bühler). Leipzig: Hirzel.

Lankes, W. (1915). Perseveration. British Journal of Psychology, 7, 387-419.

Lawyer, S.R. & Smitherman T.A. (2004). Trends in anxiety assessment. Journal of Psychopathology and Behavioral Assessment, 26, 101-106.

Leung, K., Bond, M. H., de Carrasquel, S. R., Munoz, C., Hernandez, M., Murakami, F., Yamaguchi, S., Bierbrauer, G. & Singelis, T. M. (2002). Social Axioms. The search for universal dimensions of general beliefs about how the world functions. Journal of Cross-Cultural Psychology, 33, 286-302.

Lienert, G.A. (1961). Testaufbau und Testanalyse (2. Aufl.). Weinheim: Beltz.

Lienert, G.A. & Raatz, U. (1994). Testaufbau und Testanalyse (5. Aufl.). Weinheim: Beltz, Psychologie-Verlags-Union.

Lösel, F. (1995). Persönlichkeitsdaten (Tests). In: S. Jäger & F. Petermann (Hrsg.). Psychologische Diagnostik (3. Aufl.). (S. 362-380). Weinheim: Beltz, Psychologie-Verlags-Union.

Marsella, A.J., Dubanoski, J., Hamada, W.C. & Morse, H. (2000). The measurement of personality across cultures. American Behavioral Scientist, 44, 41-62.

McCrae, R.R. & Costa, P.T. (1997). Personality trait structure as a human universal. American Psychologist, 52, 509-516.

McCrae, R.R. & Terracciano, A. (2005). Personality profiles of cultures: Aggregate personality traits. Journal of Personality and Social Psychology, 89, 407-425.

Meehl, P.E. (1954). Clinical versus statistical prediction. A theoretical analysis and a review of evidence. Minneapolis: University of Minnesota Press.

Michel, L. & Conrad, W. (1982). In: K.J. Groffmann & L. Michel (Hrsg.). Persönlichkeitsdiagnostik. Enzyklopädie der Psychologie. Psychologische Diagnostik Band 3 (S. 1-129). Göttingen: Hogrefe.

Mittenecker, E. (1982). Subjektive Tests zur Messung der Persönlichkeit. In K. J. Groffmann & L. Michel (Hrsg.). Persönlichkeitsdiagnostik. Enzyklopädie der Psychologie. Psychologische Diagnostik. Band 3 Persönlichkeitsdiagnostik (S. 57-131). Göttingen: Hogrefe.

Moosbrugger, H. Kelava, A. (Hrsg.). (2007). Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer.

Orth, B. (1983). Grundlagen des Messens. In H. Feger & J. Bredenkamp (Hrsg.). Enzyklopädie der Psychologie. Themenbereich B. Methodologie und Methoden. Serie I Forschungsmethoden der Psychologie. Band 3 Messen und Testen (S. 136-180). Göttingen: Hogrefe.

Orth, B. (1995). Meßtheoretische Grundlagen der Diagnostik. In: S. Jäger & F. Petermann (Hrsg.). Psychologische Diagnostik (3. Aufl.). (S. 286-295). Weinheim: Beltz, Psychologie-Verlags-Union.

Ostendorf, F., Angleitner, A. & Ruch, W. (1986). Die Multitrait-Multimethod Analyse: Konvergente und diskriminante Validität der Personality Research Form. Göttingen: Hogrefe.

Paunonen, S.V. & Ashton, M.C. (2001). Big Five factors and facets and the prediction of behavior. Journal of Personality and Social Psychology, 81, 524-539.

Petermann, F. (1995). Historische Kontroversen. In: S. Jäger & F. Petermann (Hrsg.). Psychologische Diagnostik (3. Aufl.). (S. 36-39). Weinheim: Beltz, Psychologie-Verlags-Union.

Petermann, F. & Eid, M. (2006). Handbuch der psychologischen Diagnostik. Göttingen: Hogrefe.

Ponocny, I. & Klauer, K.C. (2002). Towards identification of unscalable personality questionnaire respondents: The use of person fit indices. Psychologische Beiträge, 44, 94-107.

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. Zeitschrift für Sozialpsychologie, 9, 222-245.

Rost, J. (2002). When personality questionnaires fail to be unidimensional. Psychologische Beiträge, 44, 108-125.

Rost, J. (2004). Lehrbuch Testtheorie – Testkonstruktion (2. Aufl.). Bern: Huber.

Rost, J. (2006). Latent-Class-Analyse. In: F. Petermann & M. Eid, M. (Hrsg.). Handbuch der psychologischen Diagnostik (S. 275-287). Göttingen: Hogrefe.

Schermelleh-Engel, K. & Schweizer, K. (2006). Multitrait-Multimethod-Analyse. In H. Moosbrugger & A. Kelava (Hrsg.), Testtheorie: Test- und Fragebogenkonstruktion (3. Aufl.). (S. 325-341). Heidelberg: Springer.

Schmitt, T. & Gschwendner, T. (2006). Regeln der Datenintegration. In: F. Petermann & M. Eid, M. (Hrsg.). Handbuch der psychologischen Diagnostik (S. 383-395). Göttingen: Hogrefe.

Schuler, H. & Schmitt, N. (1987). Multimodale Messung in der Personalpsychologie. Diagnostica, 33, 259-271.

Schwarz, N. (1990). Assessing frequency reports of mundane behaviors: Contributions of cognitive psychology to questionnaire construction. In: C. Hendrick & M. S. Clark (Eds.). Research methods in personality and social psychology (pp. 98-119). Newbury Park, CA: Sage.

Schwarz, N. (2007). Retrospective and concurrent self reports. The rationale for real-time data capture. In: A.A. Stone, S. Shiffman, A.A. Atienza & L. Nebeling (Eds.). The science of real time data capture. Self reports in health research (pp. 11-26). New York: Oxford University Press.

Schwarz, N. & Scheuring, B. (1992). Selbstberichtete Verhaltens- und Symptomhäufigkeiten: Was Befragte aus Antwortvorgaben des Fragebogens lernen. Zeitschrift für Klinische Psychologie, 21, 197-208.

Schweizer, K. (1989). Eine Analyse der Konzepte, Bedingungen und Zielsetzungen von Replikationen. Archiv für Psychologie, 141, 85-97.

Schweizer, K. (1990). Der Aggregationseffekt. Konsequenzen der Aggregation und der Disaggregation von Daten. Frankfurt a. M.: Verlag Peter Lang.

Schweizer, K. (Hrsg.). (1999). Methoden für die Analyse von Fragebogendaten. Göttingen: Hogrefe.

Seidenstücker, G. & Baumann, U. (1978). Multimodale Diagnostik. In: U. Baumann, H. Berbalk & G. Seidenstücker (Hrsg.). Klinische Psychologie: Trends in Forschung und Praxis. Band 1 (S. 134-182). Bern: Huber.

Seidenstücker, G. & Baumann, U. (1987). Multimodale Diagnostik als Standard in der Klinischen Psychologie. Diagnostica, 33, 243-258.

Stegmüller, W. (1973). Aufgaben und Ziele der Wissenschaftstheorie. Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie. Band IV. Berlin: Springer.

Stemmler, G. (1996). Strategies and designs in ambulatory assessment. In: J. Fahrenberg & M. Myrtek (Eds.). Ambulatory Assessment: computer-assisted psychological and psychophysiological methods in monitoring and field studies (pp. 257-268). Seattle, WA: Hogrefe & Huber.

Stieglitz, R.D. (2000). Diagnostik und Klassifikation psychischer Störungen. Göttingen: Hogrefe.

Walter, P. (1999). Die „Vermessung des Menschen“: Meßtheoretische und methodologische Grundlagen psychologischen Testens. In: S. Grubitzsch (Hrsg.). Testtheorie - Testpraxis: psychologische Tests und Prüfverfahren im kritischen Überblick (2. Aufl.). (S. 98-127). Eschborn: Klotz.

Weber, H. & Rammsayer , T.(Hrsg.). (2005). Handbuch der Persönlichkeitspsychologie und Differentiellen Psychologie. Göttingen: Hogrefe.

Westhoff, K. (2004). Die Eignungsbeurteilung. In: K. Westhoff et al. (Hrsg.). Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430 (S. 201-206). Lengerich: Pabst.

Westhoff, K., Hagemeister, C. & Strobel, A. (2006). Psychologische Begutachtung. In: F. Petermann & M. Eid, M. (Hrsg.). Handbuch der psychologischen Diagnostik (S. 396-406). Göttingen: Hogrefe.

Westhoff, K., Hellfritsch, L.J., Hornke, L.F., Kubinger, K.D., Lang, F., Moosbrugger, H., Püschel, A. & Reimann, G. (Hrsg.). (2004). Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430. Lengerich: Pabst.

Westhoff, K. & Kluck, M.L. (2003). Psychologische Gutachten schreiben und beurteilen (4. Aufl.). Berlin: Springer.

Wiggins, J.S. (1973). Personality and prediction: Principles of personality assessment. Reading, Mass.: Addison-Wesley.

Wittmann, W.W. (1987). Grundlagen erfolgreicher Forschung in der Psychologie: Multimodale Diagnostik, Multiplismus, multivariate Reliabilitäts- und Validitätstheorie. Diagnostica, 33, 209-226.

Wittmann, W.W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), Handbook of multivariate experimental psychology (2nd ed.). (pp. 505-560). New York: Plenum Press.

Wittmann, W.W. (2002). Brunswik-Symmetrie: Ein Schlüsselkonzept für erfolgreiche psychologische Forschung. In M. Myrtek (Hrsg.). Die Person im biologischen und sozialen Kontext (S. 163-186). Göttingen: Hogrefe.

Wittmann, W.W. & Klumb, P.L. (2006). How to fool yourself with experiments in testing theories in psychological research. In R.R. Bootzin & P.E. McKnight (Eds.), Strengthening research methodology: Psychological measurement and evaluation (pp. 185-211). Washington, DC: American Psychological Association.

Wittmann, W. W. & Schmidt, J. (1983). Die Vorhersagbarkeit des Verhaltens aus Trait-lnventaren. Forschungsberichte des Psychologischen Instituts der Universität Freiburg, Nr. 10.

Wundt, W. (1902-1903). Grundzüge der Physiologischen Psychologie. Band 1-3 (5. Aufl.). Leipzig: Engelmann.

Wundt, W. (1907). Über Ausfrageexperimente und über die Methoden zur Psychologie des Denkens. Psychologische Studien, 3, 301-360.

Wundt, W. (1908). Kritische Nachlese zur Ausfragemethode: Archiv für die gesamte Psychologie, 11, 445-459.

Wundt, W. (1921). Logik. Eine Untersuchung der Prinzipien der Erkenntnis und der Methoden Wissenschaftlicher Forschung. Band 3. Logik der Geisteswissenschaften (4. Aufl.). Stuttgart: Ferdinand Enke.

Yousfi, S. & Steyer, R. (2006). Messtheoretische Grundlagen der Psychologischen Diagnostik. In: F. Petermann & M. Eid, M. (Hrsg.). Handbuch der psychologischen Diagnostik (S. 46-56). Göttingen: Hogrefe.