Persönlichkeitsfragebogen – kritische und konstruktive Argumente

Jochen Fahrenberg, Institut für Psychologie, und
Rainer Hampel, Arnold-Bergstraesser-Institut für kulturwissenschaftliche Forschung
Universität Freiburg
Juli 2021

Zusammenfassung

Persönlichkeitsfragebogen werden weiterhin breit angewendet, doch mangelt es an einer systematischen und vertieften Diskussion der speziellen Konstruktionsprobleme. Die folgende Argumentation stützt sich auf drei vorausgegangene Arbeiten:

(1) Zur 9. Auflage des Freiburger Persönlichkeitsinventars (Fahrenberg, Hampel und Selg, 2020) wurde eine dritte bevölkerungsrepräsentative Erhebung (Institut für Demoskopie Allensbach) durchgeführt. So konnten die Skalenkonstruktion überprüft und die Testnormen teilweise, insbesondere für die Jüngeren, angepasst werden. Das Test-Manual enthält außerdem neuere Validierungshinweise zu den FPI-Skalen und das Kapitel „Methodenbewusste Anwendung von Persönlichkeitsfragebogen“ mit ausführlicher kritischer Diskussion der Testkonstruktion und Validierungsprobleme.

(2) In der zugehörigen Dokumentation (Hampel, 2020) werden weitere Analysen zur Testkonstruktion und Details der Validierung berichtet, u.a. zu Gesundheits- und Belastungsindikatoren bzw. dem Vergleich zwischen Bevölkerung und Patientengruppen 1999 und 2018, Eigenschaftsprofile aufgrund von Clusteranalysen („Persönlichkeitstypen“), Multitrait-Multimethod Hinweise.

(3) Unter dem Titel Was bedeutet ‚Testpflege‘? – Zur Qualitätssicherung von Persönlichkeitsfragebogen (Fahrenberg und Hampel, 2020) wurde der Prozess eines „kontinuierlichen Qualitätsmanagements“ diskutiert, insbesondere nach den Guidelines der International Test Commission (2005, 2013, 2015, 2017). Als ein möglicher Ansatz wurde hier die vergleichende Analyse der Datensätze aus den drei Normierungsstudien des FPI herangezogen, die jedoch nur eine statistische „Pseudo-Kohorte“ bilden (ein entsprechendes Panel wäre für einen Fragebogen mit mehreren Skalen nicht finanzierbar). Im weiteren Sinn umfasst Qualitätssicherung den kontinuierlichen Prozess eines Qualitätsmanagements, an dem die Testautoren und der Verlag sowie die Anwender und die Rezensenten Anteil haben. Jede neue Auflage eines Tests bietet die Gelegenheit, qualitätsverbessernde Befunde und Überlegungen aufzunehmen.

Summary

Personality questionnaires continue to be widely used, but there is a lack of systematic and in-depth discussion of the specific design problems. The following article is based on three previous works.

(1) For the 9th edition of the Freiburg Personality Inventory (Fahrenberg, Hampel and Selg, 2020), a third population-representative survey was carried out by the Institute for Demoscopy, Allensbach. In this way, the scale construction could be checked and the test norms partially adapted, especially for the younger cohort. The test manual also contains more recent validation evidence on the FPI scales, and a chapter ‘Method-conscious application of personality questionnaires’ with detailed critical discussion of test design and validation problems.

(2) In the associated FPI-documentation (Hampel, 2020), further analyses of test design and details of validation are reported, e.g., on health and stress indicators, comparison between population and patient groups in 1999 and 2018, trait profiles based on cluster analyses (‘personality types’), and multitrait-multimethod indications.

(3) Under the title ‘What does test maintenance mean? – quality assurance of personality questionnaires’ (Fahrenberg and Hampel, 2020), the process of ‘continuous quality management’ was discussed, with reference to the guidelines of the International Test Commission (2005, 2013, 2015, 2017). As a possible approach, the comparative analysis of the data-sets from the three standardization studies of the FPI was used, which, however, only form a statistical ‘pseudo-cohort’ (a corresponding panel would not be financially viable for a questionnaire with several scales). More broadly, quality assurance includes the continuous process of quality management in which the test authors and the publisher, as well as the users and the reviewers, participate. Each new edition of a personality test offers the opportunity of incorporating quality-enhancing findings and considerations.

1. Einleitung

In der psychologischen Praxis und in Forschungsvorhaben werden weithin Persönlichkeitsfragebogen verwendet. Im Kontrast zu diesem häufigen Gebrauch sind in Lehrbüchern der Testtheorie und Testkonstruktion ausführliche Diskussionen kaum zu finden, denn sie befassen sich, wenn es um „psychometrische" Prinzipien und entsprechende Konstruktionsmethoden geht, hauptsächlich mit objektiven Intelligenz- und Leistungstests. In Lehrbüchern der Differentiellen Psychologie und Persönlichkeitsforschung werden Persönlichkeitsfragebogen teils ausführlich referiert, ohne jedoch der vergleichenden Diskussion und den testkonstruktiven Problemen viel Raum geben zu können.

Eine Umfrage unter Mitgliedern des Berufsverbandes BDP hinsichtlich psychologischer Diagnostik in der Praxis (Roth, Schmitt & Herzberg, 2010) ergab außer einer Liste der am häufigsten verwendeten Testverfahren eine Reihe von Verbesserungsvorschlägen für das Studium. An erster Stelle wurde gefordert, dass „in der universitären Diagnostikausbildung eine kritische Diagnostik gelehrt wird. So soll auf die Grenzen der Testverfahren hingewiesen werden, und es soll für Fehlerquellen beim Testen eine Sensibilität erzeugt werden." (S. 126). Diese kritisch-methodenbewusste Einstellung kann durch die Weiterentwicklung eines Testmanuals unterstützt werden. Gerade bei den so einfach erscheinenden Persönlichkeitsfragebogen sind die Prinzipien der Testkonstruktion und der empirischen Validierung wichtig, um kritisch auswählen und interpretieren zu können. Persönlichkeitsfragebogen erfassen Selbstbeurteilungen und Selbstbeobachtungen und sie sollen facettenreiche Persönlichkeitseigenschaften repräsentieren. Folglich können die psychometrischen Postulate (Messmodelle) und die Konstruktionsweisen von objektiven Intelligenz- und Leistungstests nicht einfach auf Persönlichkeitsfragebogen übertragen werden, und die Nachweise externer Kriterienvalidität sind hier noch wichtiger.

Während der vergangenen Jahrzehnte wurden mehrere Persönlichkeitsfragebogen in Deutschland entwickelt und normiert. Nicht zu übersehen ist jedoch, dass es weitverbreitete Fragebogen gibt, bei denen es sich um einfache Übersetzungen amerikanischer Tests handelt. Selbst wenn die Qualität der sprachlichen Übersetzung durch Rückübersetzung durch qualifizierte bi-linguale Personen überprüft wird, fehlt in der Regel eine Analyse der „psychologischen Äquivalenz". Deshalb erfüllen solche deutschen Varianten nicht die Standards der International Test Commission ITC Guidelines on Test Use, Version 1.2 (International Test Commission, 2013) zu nennen: ITC Guidelines on Test Use (translations and description (https://www.intestcom.org/page/17), – Diese Richtlinien werden in der deutschen Fachliteratur kaum erwähnt. Der Grund ist unklar. Die Kommissionen der ITC hatten anscheinend kein deutsches Mitglied.

Hinsichtlich der sprachlichen Äquivalenz geben die Items „Ich gehe abends gerne aus" oder „Ich übe zwei Berufe aus" einfache Beispiele. Die sprachliche Übersetzung ist einfach, doch hat das abendliche Ausgehen in Italien einen anderen Stellenwert als in Norddeutschland, und es gibt in den USA und in einigen europäischen Ländern Gegenden, in denen die doppelte Berufstätigkeit sehr verbreitet, finanziell notwendig und fast selbstverständlich ist. Die kulturpsychologische Divergenzen gehen jedoch tiefer. Besonders interessant ist die grundsätzliche Kritik seitens einer Gruppe von Psychologen in Südafrika sowie einer Gruppe in China an dem amerikanischen Menschenbild, das in der großen Gruppe der Persönlichkeitsinventare mit vier bis sechs angeblich fundamentalen „großen" Persönlichkeitsfaktoren zu erkennen sei und universell gültig sein soll. Demgegenüber vermissen die Kritiker bestimmte Eigenschaftskonzepte, die von ihnen als fundmental angesehen werden. Gegen den universelle Anspruch jener amerikanischen Testautoren hat sogar die American Psychological Association indirekt Stellung genommen, indem die chinesische Psychologin Fanny M. Cheung den „Award for Distinguished Contributions to the International Advancement of Psychology" für ihre kulturkritischen Beurteilung und entsprechende Konstruktion des Chinese Personality Assessment Inventory erhielt (APA 2012).

Ein fundamentales Problem bleibt die bevölkerungsrepräsentative Normierung amerikanischer Tests in Deutschland, zumal die Forderung nach bevölkerungsrepräsentativer Normierung eines Fragebogens in der weitaus heterogeneren Bevölkerung der USA größte Anforderungen stellen würde. Der einfachere und weithin übliche Weg ist, vorhandene Datensätze aus verschiedenen Projekten zu kombinieren und dann nach der Bevölkerungsstatistik sekundär eine „repräsentative" Stichprobe zu konstruieren. Die unterschiedlichen Ziele und Organisationsweisen der so kombinierten „Gelegenheitsstichproben" werden gewöhnlich nicht mitgeteilt. Außerdem wächst die Tendenz, solche Erhebungen nur bei Personen mit Festnetz-Telefon (wie das ZDF-Politbarometer) zu unternehmen oder sich auf bereitwillige Teilnehmer einer u. U. vergüteten Internet-

Umfrageaktion zu beschränken. Auch bei der konventionellen und wesentlich aufwendigeren Erhebung durch erfahrene Mitarbeiter, die in der Regel direkt geschieht (Institut für Demoskopie Allensbach, ALLBUS u.a.), sind spezifische methodische Bedenken möglich. Wenn jedoch der Interviewer in der Regel beim Ausfüllen des Fragebogens anwesend ist, kann dies auch die Chance bieten, außer vielen zusätzlichen demoskopischen Daten, auch Informationen über berufliche und soziale Merkmale auch Charakteristika des Wohnens und explorativ sogar Verhaltenseinstufungen zu erhalten (vgl. entsprechende Befunde für das FPI).

Das alte Problem der sozialen Erwünschtheit bestimmter Antworten hat zu einem hohen Forschungsaufwand angeregt, ohne allgemein zu praktischen Schlussfolgerungen zu führen. Angesichts des altbekannten Problems der Divergenz von geäußerter Einstellung und tatsächlichen Verhalten, Reden und Tun, können die Erwartungen nicht allzu hoch sein. Die naheliegende Konsequenz lautet daher, gerade bei Fragebogenmethoden noch viel intensiver an der externen Validierung zu arbeiten, denn es kann hier keine, den objektiven Intelligenz- und Leistungstests vergleichbare, innere Validität behauptet werden.

Auf psychologisch relevante externe Kriterien bezogene Validierungen können heute wahrscheinlich nur noch aus den großen Institutionen der Praxisfelder (Kliniken, soziale Einrichtungen, Versicherungsträger, Betriebe, Verwaltungen u.a.) heraus initiiert und organisiert werden, denn nur dort sind die wichtigsten Indikatoren und Prädiktoren sowie die wesentlichen Informationen über den Entscheidungsnutzen (auch aus Gründen des Datenschutzes) zu gewinnen und Prinzipen der modernen Assessmenttheorie umzusetzen. Auch die testmethodisch wichtigen Multitrait-Multimethod (MTMM)-Studien stellen sehr hohe Anforderungen, wenn sie adäquat geplant und durchgeführt werden sollen. Umso wichtiger sind Nachweise externer Validität indem neuere Methoden eingesetzt werden, u.a. das digitale ambulante Assessment, um Befinden, soziale Situation, Tätigkeiten, Bewegungsaktivität, physiologische und ambiente (umweltbezogene) Parameter erfassen. – Generell ist der zeitliche und finanzielle Aufwand für Skalenkonstruktion, Normierung und Validierung hinsichtlich externer Kriterien so hoch, dass kooperative Projekte zur Qualitätskontrolle – und künftig bereits zur Entwicklung – mehrdimensionaler Persönlichkeitsfragebogen erforderlich sein werden.

Von den Autoren eines eingeführten Tests ist zu erwarten, dass bei jeder Neuauflage mindestens die Fortschreibung der Validitätshinweise erfolgt und in adäquaten Abständen eine neue Normierung, die mit der Kontrolle der Testkonstruktion verbunden wird. Die „Pflege“ ist umso mehr erforderlich, wenn ein Test relativ weit verbreitet ist. Die FPI-Autoren halten es außerdem für wissenschaftlich unverzichtbar, dass die bevölkerungsrepräsentativen Datensätze der Testkonstruktion in PsychData des ZPID archiviert sind und für wissenschaftliche Zwecke zur Verfügung gestellt werden: Reanalysen, vergleichende Untersuchungen, Unterrichtszwecke. Ein einfacher Datennehmer-Vertrag regelt, auch im Einverständnis mit dem Verlag, die Nutzungsmöglichkeiten. Auf diese Weise kann eventueller Besorgnis begegnet werden, dass sich Copyright-Verstöße ereignen und dieser Datensatz direkt für fremde Testpublikationen benutzt wird. Eine wesentliche Forderung der open-access-Bewegung, auch in der Auseinandersetzung über Replizierbarkeit, bleibt, dass Datensätze aus der Forschung, d.h. ebenso aus der Testkonstruktion, für die Fachwelt zugänglich gemacht werden – auch in der Psychologie.

Die skizzierten Argumente werden in den folgenden Abschnitten weiter erläutert und durch einige Literaturhinweise ergänzt:

- Konstruktion von Persönlichkeitsfragebogen – kritische Differenzierungen
- Fragwürdige Übertragung messtheoretischer Axiome auf Selbstbeurteilungen (Introspektion)
- Eigenständige Konstruktionsprinzipien der Persönlichkeitsfragebogen
- Kontroversen über die „größten und wichtigsten“ Persönlichkeitsfaktoren
- Bereitschaft zu offener Auskunft und Selbstschilderung
- Prinzipien der Assessmenttheorie
- Ambulantes Assessment
- Kritik an der Dominanz der Fragebogenmethodik
- Einstellungen zu Persönlichkeitsfragebogen, populäre Medienbeiträge und problematische Publikationen
- Qualitätssicherung

Aus der Einleitung des FPI-Manuals (2020)

Die Autoren des Freiburger Persönlichkeitsinventars sind überzeugt, dass Persönlichkeitsfragebogen für bestimmte, beschreibende und diagnostische Aufgaben wichtig sind und auch künftig unentbehrlich sein werden. Gerade Persönlichkeitsfragebogen verlangen jedoch eine methodenbewusst-kritische Anwendung und möglichst auch Strategien der multimodalen Diagnostik, d.h. Absicherungen durch andere Methoden und vorsichtige Interpretation. Wegen der äußerlich leichten Handhabung können vielleicht die Schwierigkeiten der Test-konstruktion übersehen werden. Diese kritische Diskussion der Vorzüge und der Probleme von Persönlichkeitsfragebogen gehören nach der Auffassung der Autoren auch in ein Test-Manual. Wie in den früheren Auflagen wird ausführlich über die Absichten der Konstruktion, über Validitätshinweise, über die Anwendung und auch über die Kritik an Persönlichkeitsfragebogen berichtet. Diese Informationen sind heute noch wichtiger, da im Vergleich zum früheren Diplom-Studiengang wahrscheinlich nur ein Teil der Absolventen eines Psychologie-Studiums (mit speziellem Vertiefungsfach) über eine hinreichende diagnostische Ausbildung, einschließlich Testkonstruktion und Gutachtenpraxis, verfügen wird.

Persönlichkeitsfragebogen und Klinische Skalen sind neben den Intelligenz- und Leistungstests die am häufigsten verwendeten psychologischen Tests. Die Durchsicht der aktuellen Lehrbücher der Psychologischen Diagnostik, Persönlichkeitspsychologie, Testkonstruktion und Statistik zeigt jedoch, dass diese – mit sehr wenigen Ausnahmen – keine gründlichen Kapitel über Persönlichkeitsfragebogen, über deren Voraussetzungen, unterschiedliche Konstruktion und problematische Normierung sowie die methodenbewusste Anwendung im Sinne der modernen Assessmenttheorie enthalten. – Inwieweit kann die primär für objektive Intelligenz- und Leistungstests entwickelte Methodik der Skalen- und Itemanalyse auf die Konstruktion von Persönlichkeitsfragebogen übertragen werden? Wird psychometrisch den introspektiven Selbstbeurteilungen im Persönlichkeitsfragebogen dasselbe Messmodell zugrunde gelegt wie den objektiv nach richtig oder falsch bewerteten Aufgabenlösungen? Bleibt das Konstruktionsziel möglichst homogener, nur in ihrer Schwierigkeit abgestufter, „paralleler" Messungen erhalten oder verlangen die facettenreichen Konstrukte der Persönlichkeitspsychologie eine andere Bewertung der inneren Konsistenz (Reliabilität)? Die Auswahl eines Persönlichkeitsfragebogens für eine spezielle Aufgabe in der Forschung oder in den Praxisfeldern der Psychologie bedeutet bereits, sich auf eine bestimmte Auswahl von Persönlichkeitseigenschaften festzulegen, und darüber hinaus eine spezielle Konzeption von „Persönlichkeit" und das jeweilige Konstruktionsprinzip zu akzeptieren. Angesichts der Vielfalt von Persönlichkeitstheorien, Fragebogen und Konstruktionsproblemen kann diese Entscheidung nicht einfach sein.

Wichtige Voraussetzungen für die psychologische Gültigkeit von Persönlichkeitsfragebogen

Ein methodisch gründlich konstruierter Persönlichkeitsfragebogen ermöglicht es, die individuelle Ausprägung bestimmter Persönlichkeitszüge zu vergleichen, d.h. auf andere Personen und auf die Normwerte der Bevölkerung zu beziehen. Dieses Beschreibungssystem hat jedoch zwei Voraussetzungen: (1) bereits die Testkonstruktion und nicht nur die Normierung müssen aufgrund einer hinreichend großen und bevölkerungsrepräsentativ erhobenen Stichprobe erfolgen und (2) die Testkonstruktion und Normierung sind in adäquaten Abständen zu kontrollieren. Andernfalls können die individuellen Fragebogenwerte zwar weiter berechnet werden, doch ohne zuverlässige Vergleichsmöglichkeiten. Grundsätzliche Vorbehalte bestehen hinsichtlich aller Testentwicklungen, die anstelle einer einheitlich erhobenen und bevölkerungsrepräsentativen Datenbasis nur einen aus diversen Untersuchungen nachträglich kompilierten Datensatz verwenden oder sogar methodisch ungeeignete Internet-„Umfragen" mit verzerrter Vorauswahl, die repräsentative Aussagen unmöglich machen.

Die regelmäßige Normierung und der Nachweis, dass das Beschreibungssystem prägnant und reproduzierbar ist, bilden die Grundlage und die formale Voraussetzung der psychologischen Gültigkeit. In welchem Umfang gibt es nun empirische Nachweise der externen psychologischen Gültigkeit im Hinblick auf deskriptive Zwecke und diagnostische Entscheidungen? Wie steht es mit reproduzierbaren Validitätshinweisen hinsichtlich psychologisch wichtiger Kriterien, beispielsweise zu den Bereichen Beruf und Belastung, Gesundheit und klinisch-psychologische Auffälligkeit? Welche Validitätshinweise gibt es, dass die Testwerte relevante psychologische Informationen geben – auch für das Alltagsverhalten und andere Aspekte der externen („ökologischen") Validität?

Doch Persönlichkeitsfragebogen erfassen Selbstbeurteilungen. Die Befragten sollen ausgewählte Aspekte ihres Erlebens und Verhaltens beschreiben, damit die relativ überdauernde Ausprägung bestimmter Persönlichkeitseigenschaften im Vergleich zu den repräsentativen Normwerten der Bevölkerung (oder bestimmter Personengruppen) psychologisch beurteilt werden

kann. Solche Fragebogen bestehen aus einer Anleitung und der Liste der Items, die häufig nicht in Frageform, sondern als Aussagen (Ich bin …) formuliert sind, um es den Befragten zu erleichtern, den eigenen Bezug zu dem psychologischen Inhalt herzustellen. Die Antworten sind überwiegend Selbstbeurteilungen. Als introspektive Aussagen können sie grundsätzlich nicht überprüft werden. Einige Items beziehen sich auf Selbstbeobachtungen von Verhaltensweisen oder sind Selbstauskünfte, die sich auf bestimmte Ereignisse beziehen. Solche Aussagen wären im Prinzip objektivierbar, beispielsweise indem festgestellt wird, ob jemand sich häufig mit Freunden trifft, tatsächlich für soziale Zwecke spendet oder durch aggressives Verhalten auffällig wurde.

Die Konzeption der Persönlichkeitsfragebogen ist einem standardisierten psychologischen Interview ähnlich, und historisch sind solche Fragebogen aus einfachen Formen solcher Interviews hervorgegangen. Doch es gibt zwei fundamentale Unterschiede: psychologisch ausgebildete Interviewer können ergänzende und klärende Fragen stellen und sie können ihrerseits die einzelnen Selbstbeurteilungen unmittelbar kritisch evaluieren: aus dem Kontext der Fragestellung, der aktuellen Situation und der bereits vorliegenden psychologischen Informationen. Die Fragebogenmethodik muss in der Regel ohne die ergänzende und kontrollierende Methodik des geschulten psychologischen Interviews und der zugehörigen Interpretationsmethodik auskommen. Insofern können die wünschenswerte breite Anwendbarkeit, die hohe Standardisierung und die „Testökonomie“ von Persönlichkeitsfragebogen auch als Defizite angesehen werden. Je nach Aufgabe bleibt abzuwägen, was adäquat und was praktisch möglich ist.

Selbstbeurteilungen und Selbstauskünfte sind aus mehreren Gründen problematisch. Auch wenn anzunehmen ist, dass die Mehrzahl der Befragten offen antwortet, könnten doch im Einzelfall eine mehr oder minder ausgeprägte „Selbstdarstellung“ zum Tragen kommen, um einen guten Eindruck zu machen. Testmethodische Bemühungen, diese Effekte mittels zusätzlicher Kontrollskalen hinsichtlich der Offenheit, der sozialen Erwünschtheit oder der internen Widerspruchsfreiheit der Antworten zu erfassen, haben nicht überzeugen können. In dieser Hinsicht bleibt die Testgültigkeit grundsätzlich eingeschränkt.

Grundsätzlich noch wichtiger ist die Konstruktionsstrategie von Persönlichkeitsfragebogen. Kann die primär für objektive Intelligenz- und Leistungstests entwickelte Methodik der Item- und Skalenanalyse auf die Konstruktion von Persönlichkeitsfragebogen übertragen werden? Sind objektive, nach richtig oder falsch bewertete Lösungen von Aufgaben und die primär auf Introspektion bzw. Selbstbeurteilung beruhenden Fragebogendaten im Prinzip gleichartig zu behandeln oder sollte die – unkommentierte – Anwendung solcher Messmodelle kritisiert und relativiert werden? Fragebogendaten sind sicherlich keine Messungen auf dem Niveau einer Intervallskala. Folglich sind die „psychometrischen“ Voraussetzungen der meisten statistischen Methoden, die für Intelligenz- und Leistungstests üblich sind, nicht gegeben. Es genügt jedoch nicht, wenn in Lehrbüchern die Voraussetzungen allgemein definiert werden, ohne die wesentlichen Konsequenzen für die Persönlichkeitsfragebogen zu erläutern und Strategien der Absicherung vorzuschlagen und zu überprüfen. Durch welche anderen, voraussetzungsärmeren statistischen Verfahren sind die eigentlich inadäquaten Item- und Faktorenanalysen des Itempools zu ergänzen und zu unterstützen, um die Skalenkonstruktion eher zu rechtfertigen?

In der Konstruktion der objektiven Tests werden inhaltlich relativ ähnliche, aber unterschiedlich schwierige Aufgaben aneinandergereiht, um parallele Messungen einer einzelnen Fähigkeit zu erreichen. Demgegenüber sind die theoretischen Konstrukte in der Persönlichkeitspsychologie sehr viel facettenreicher. Hier ebenfalls eine hohe innere Konsistenz (Homogenität) zu fordern, wäre inadäquat. Denn die höchste „Reliabilität“ nahe $r = .90$ wäre am einfachsten zu erreichen, wenn ein inhaltlich gleichbleibendes Item schematisch und nur in abgestuften Schwierigkeitsgraden wiederholt würde. Dagegen ist für die Skalen eines Persönlichkeitsfragebogen ein Kompromiss anzustreben: ein adäquater Facettenreichtum des theoretischen Konstrukts, z.B. „Lebenszufriedenheit“ in verschiedenen Bereichen wie Partnerschaft, Gesundheit, Beruf und Einkommen, bei hinreichender psychologischer Einheitlichkeit (Homogenität, Konsistenz) der beabsichtigen Skalenbildung. Dieses Beispiel zeigt einen weiteren grundsätzlichen Unterschied in der Konstruktion der Skalen von Intelligenztests und von Persönlichkeitsfragebogen.

2. Konstruktionsstrategien von Persönlichkeitsfragebogen

Zur Konstruktion von Persönlichkeitsfragebogen gibt es unterschiedliche Voraussetzungen und Strategien, deren Kenntnis für die adäquate Anwendung wichtiger ist als das Wissen über viele teststatistische Details. Es sind vier hauptsächliche Strategien zur Konstruktion von Persönlichkeitsfragebogen zu unterscheiden:
– die primär kriterienbezogene Strategie;
– die auf eine spezielle Persönlichkeitstheorie bezogene deduktive Strategie;
– die lexikalisch-induktive Strategie und
– die kombinierte Strategie, die hypothetisch-deduktive und empirisch-induktive Schritte verbindet und zu den Prinzipien der modernen Assessmenttheorie weiterführt.

Jede dieser Konstruktionsstrategien richtet sich auf (1) die interne und formale Zuverlässigkeit (Reliabilität), die für die differenzierte Beschreibung der individuellen Unterschiede notwendig ist, und (2) auf die interne und auf die externe Gültigkeit (Validität), die angibt, inwieweit die psychologisch gemeinte Persönlichkeitseigenschaft repräsentiert ist und mit welchen Kriterien empirisch belegte Zusammenhänge bestehen. Da alle statistischen Analysen spezielle Voraussetzungen machen, ist zu prüfen, ob diese psychometrischen Postulate adäquat sind oder nicht. Da für die individuellen Selbstbeschreibungen in einem Persönlichkeitsfragebogen offensichtlich keine Intervallskalen postuliert werden können, also kein „Messmodell“ im engeren Sinn, bestehen grundsätzliche Unterschiede zur Konstruktion von Intelligenz- und Leistungstests. Korrelieren die Skalenwerte des Fragebogens extern nicht nur mit den Ergebnissen ähnlicher Methoden, sondern vor allem auch mit psychologisch wichtigen Kriterien? Erst solche externen Validitätshinweise machen den möglichen Nutzen des Tests aus. Nach dem Entscheidungsnutzen zu fragen, enthält logisch auch die Alternative, an eine mögliche Schadensfunktion zu denken: wenn z.B. eine klinisch-psychologische Begutachtung zu einer unzutreffenden Diagnose führt oder ein besonders qualifizierter Bewerber fälschlich abgelehnt wird. Welche Kriterien wichtig sind, hängt von den speziellen Fragestellungen der Forschung und der Berufspraxis ab. Zu welchem deskriptiven, diagnostischen oder prognostischen Zweck werden die Testwerte erhoben?

Zwischen den genannten vier Konstruktionsstrategien gibt es markante Unterschiede: welche Ziele als vorrangig angesehen werden, die Anzahl der ausgewählten Eigenschaften, die internen psychometrisch-strukturellen Eigenschaften der Skalen (aufgrund von Itemanalyse, Faktorenanalyse, Clusteranalyse, Item-Response-Messmodell) oder primär die Kriterienvalidität aufgrund eines effektstarken Zusammenhangs mit psychologisch wichtigen Kriterien, auch als Chance von Vorhersagen (Prognosen).

Kriterienbezogene Konstruktionsstrategie

Hauptsächlich der kriterienbezogenen Konstruktionsstrategie folgte das erste große Persönlichkeitsinventar Minnesota Multiphasic Personality Inventory (MMPI) von Hathaway und McKinley (1943). Die Skalen wurden im Hinblick auf psychologisch-psychiatrische Diagnosegruppen konstruiert. Es gibt eine nachkonstruierte deutsche Version: MMPI-2 (Engel, 2000) bzw. MMPI-2-RF (Engel, 2019). Auch die zwei Persönlichkeitsskalen, die noch heute weit herausragen und konzeptuell die weiteste Verbreitung in der Psychologie fanden, wurden primär aufgrund klinisch-psychologischer Kriterien bestimmt und nicht durch problematische Messmodelle: Eysencks Skalen Emotionalität (ursprünglich: Neurotische Tendenz; Neurotizismus) und Extraversion-Introversion. In eine Fragebogen-Skala werden jene Items aufgenommen, die sich primär durch ihre Kriterienvalidität auszeichnen. Diese Strategie der Kriterienvalidierung wird gegenwärtig in der Persönlichkeitsforschung seltener befolgt, abgesehen von den wichtigen Klinischen Skalen, die bestimmte Dimensionen oder Syndrome psychopathologischer Art erfassen sollen, und anderen Skalen für spezielle Aufgabengebiete.

Deduktive Konstruktionsstrategie

Deduktiv entwickelte Fragebogen sollen die wesentlichen Eigenschaftskonzepte einer speziellen Persönlichkeitstheorie erfassen und in ihrem Zusammenhang nachbilden, d.h. ein „Persönlichkeitsmodell“ operationalisieren. Die psychologischen Inhalte sind weitgehend durch die betreffende Persönlichkeitstheorie vorgegeben und sind durch adäquate Items bzw. Skalen testkonstruktiv optimal nachzubilden. Ein bekanntes Beispiel ist die Personality Research Form, ein Persönlichkeitsfragebogen unter Bezug auf Murrays Theorie (deutsche Fassung Stumpf u.a., 1985); zu nennen ist auch der Gießen-Test GT-II (Beckmann, Brähler & Richter, 2012), der aus psychoanalytischer Orientierung entwickelt wurde. Diese deduktive Strategie

hat den Vorzug, dass die einzelnen Persönlichkeitseigenschaften in einem psychologischen und auch motivationalen (dynamischen) Zusammenhang konzipiert sind. Aus der Sicht anderer Testautoren kann dieser spezielle theoretische Bezug ein Nachteil sein, doch ist nicht zu übersehen, dass die anderen Konstruktionsstrategien stets nur ein Inventar von Persönlichkeitseigenschaften liefern, aber nichts über den Funktionszusammenhang dieser Dispositionen aussagen können. In vorwiegend deduktiver Strategie wurden einzelne Skalen auch aus anderen Forschungsansätzen entwickelt. Gerade in diesem Bereich ist ein auffälliger Interessenwandel zu erkennen. Vor hundert Jahren waren Eigenschaften wie Perseveration und Rigidität wichtig, in neuerer Zeit interessierten eher andere Konzepte: u.a. Repression-Desentization, Sensation Seeking, Kontrollüberzeugungen, Alexithymie, Resilienz, Embodiment oder Grit (Selbstkontrolle). Es sind teils neu konzipierte Eigenschaften, teils Varianten oder nur neue Namen für ähnliche frühere Konzepte.

Induktive und lexikalisch-induktive Konstruktionsstrategie

Induktiv vorzugehen, bedeutet hier, zunächst einen Bereich von Persönlichkeitsmerkmalen inhaltlich abzugrenzen und anschließend aus den einzelnen Merkmalen auf die zugrunde liegenden Faktoren (Dimensionen) zu schließen. Diese Datenreduktion wird meistens statistisch durch faktorenanalytische Verfahren vorgenommen. Empirisch wird es sich immer um eine problematische Auswahl handeln. Die Lehrbücher der Persönlichkeitspsychologie enthalten umfangreiche Zusammenstellungen: sog. Temperamentseigenschaften, Handlungseigenschaften, Bedürfnisse, Motive, Einstellungen, Interessen, Bewältigungsstile, Werthaltungen, vorwiegende Stimmungslagen usw. Auch die definitorische Abgrenzung zwischen Eigenschaft und Zustand ist in der Persönlichkeitsforschung (Robert Heiß „Person als Prozess“, 1948; R. B. Cattell „State-Trait“, 1957) als sehr fragwürdig erkannt worden. Die Definition jener Merkmale, die zentral zu „der Persönlichkeit“ gehören, bleibt eine willkürliche, d.h. psychologisch nicht überzeugend zu begründende, Festlegung. Vielleicht stammt die Annahme eines zentralen Persönlichkeitsbereichs noch aus der alten Temperamentslehre und Charakterkunde, in denen lebenslang ausgeprägte oder angeborene Wesens-Eigenschaften behauptet wurden? Jedenfalls existiert kein überzeugendes Kriterium, zwischen einem zentralen Bereich, der Domäne der „Persönlichkeit“, und den anderen Domänen von Persönlichkeitsmerkmalen zu unterscheiden. – Sind beispielsweise die soziale Orientierung einschließlich sozialer Verpflichtung, die Lebenszufriedenheit, die überdauernden Gesundheitssorgen persönlichkeitspsychologisch nicht mindestens so interessant wie Verträglichkeit oder Gewissenhaftigkeit? Gehört auch, wie in Cattells 16-PF, eine Skala zur allgemeinen Intelligenz in den „Persönlichkeits“-Fragebogen?

Die lexikalisch-induktiven Strategien bilden eine Gruppe von Verfahren, die vom alltagssprachlichen Vokabular ausgehen und postulieren, dass eine methodische Reduktion auf die „wichtigsten“ Persönlichkeitseigenschaften möglich ist. Selbstbeurteilungen und psychologische Beurteilungen anderer Menschen werden sprachlich mitgeteilt. Aus psycholinguistischer Sicht scheint es sehr nahe zu liegen, im ersten Schritt das psychologische Vokabular über „Persönlichkeit“ aufgrund der eingeschätzten semantischen Ähnlichkeiten auf eine praktikable Anzahl einzuschränken (wie u.a. von Allport & Odbert, 1936, in den USA vorgenommen). Im zweiten Schritt wird diese Auswahl empirisch und faktorenanalytisch reduziert, um die erhaltene Struktur im dritten Schritt als Persönlichkeitstest zu konstruieren und zu normieren. Solche lexikalisch begründeten Reduktionsversuche wurden in den USA von mehreren Autoren, u.a. von Cattell, Saunders and Stice (1957) für den Sixteen Personality Factor Test 16-PF vorgenommen mit den Dimensionen: Wärme, Logisches Schlussfolgern, Emotionale Stabilität, Dominanz, Lebhaftigkeit, Regelbewusstsein, Soziale Kompetenz, Empfindsamkeit, Wachsamkeit, Abgehobenheit, Privatheit, Besorgtheit, Offenheit für Veränderung, Selbstgenügsamkeit, Perfektionismus, Anspannung. Andere Autoren haben Inventare mit 4, 5, 6 oder mehr „wichtigsten“ Faktoren publiziert, darunter mehrere mit 5 Faktoren, den sog. Fünf-Faktor-Modellen (FFM). Von diesen ist die Variante von Costa und McCrae (1985) NEO Personality Inventory in Deutschland durch die Übersetzung von Borkenau und Ostendorf (1993, 2008) bekannt geworden.

Drei grundsätzliche Probleme sind zu erkennen: Enthält die Alltagssprache tatsächlich Ausdrücke für alle psychologisch wichtigen Persönlichkeitsmerkmale, so dass die gesamte wissenschaftliche Psychologie sprachlich nichts Neues hinzuzufügen braucht? Wie können Konstruktionen nach lexikalischer Methodik ohne sonstige persönlichkeitstheoretische Definitionen abgrenzen zwischen den Wörtern, die sich auf die Persönlichkeit beziehen und jenen, die andere Merkmale der Person meinen? Wie kann der Anspruch, auf diese Weise die wichtigsten Faktoren gefunden zu haben, überhaupt stichprobentechnisch begründet werden? Die ausgewählten Items bilden keine Zufallsstichprobe aus dem Universum der psychologischen Deskriptoren und die Konstruktion erfolgte primär anhand eines Datensatzes, der oft nur aus diversen anderen Datensätzen nachträglich zusammengesetzt wurde. Gerade für diese Konstruktionsstrategie ist es jedoch unerlässlich, dass die

ausgewählten Items und die untersuchten Personen gültige (repräsentative) Zufallsstichproben sind.

Wie überzeugend und universell können jene Fünf-Faktoren-Modelle (FFM oder sogenannte Big Five) sein, wenn mit induktiv-lexikalischer Methodik ebenso auch Systeme mit zwei, drei, vier, sechs oder 16 Faktoren begründet werden? Die zunehmende Methodenkritik am Mythos Big Five (Andresen, 2015) wird verstärkt durch die kulturpsychologisch begründete Kritik an diesem Menschenbild und an dem Anspruch amerikanischer Testautoren auf universelle Gültigkeit ihrer Auswahl von Persönlichkeitseigenschaften.

Die kombinierte deduktiv-induktive Konstruktionsstrategie

Diese Strategie hat das Ziel, für eine psychologisch begründete Auswahl von Persönlichkeitseigenschaften einen gültigen Fragebogen zu entwickeln. Es wird keine umfassende Theorie und kein übergeordneter Struktur- oder Funktionszusammenhang dieser Eigenschaften behauptet. Statt einen Fragebogen für „die Persönlichkeit im Allgemeinen" zu entwickeln, findet die Konstruktion auf der Ebene der einzelnen Persönlichkeitseigenschaften statt. Ausgewählt für die Konstruktion eines neuen Persönlichkeitsfragebogens werden jene Eigenschaften, die den Testautoren in den Bereichen ihrer Forschung und Praxis wichtig sind. Es liegt nahe, einen auf diese Weise entstandenen Fragebogen zu publizieren und dadurch allgemein zugänglich zu machen. Die persönlichkeitstheoretische Diskussion findet also hinsichtlich der einzelnen Eigenschaftskonstrukte statt, um die Iteminhalte so zu formulieren, dass die als wichtig angesehenen Facetten jedes Konstrukts repräsentiert sind. Diesem theoretisch-deduktiven Prozess folgt eine induktiv angelegte Überprüfung, um die Itemauswahl aufgrund statistischer Kennwerte der item-, faktoren- und clusteranalytischen Analysen vorzunehmen. Eventuell kann der primäre Itempool in mehreren Phasen inhaltlich erweitert oder reduziert werden, um die Prägnanz des gemeinten Konstrukts, die notwendige Itemzahl für eine Skalenbildung oder die sprachliche Formulierung zu optimieren.

Die skizzierte Strategie entspricht der auf vielen wissenschaftlichen Gebieten bewährten Kombination des deduktiven und des induktiven Verfahrens in einem hypothetischen Ansatz und einer wiederholten (rekursiven) empirischen Prozedur, die auch Revisionen einschließen kann. Ausgehend von den Arbeitsergebnissen der Testautoren und der Testanwender entsteht eine sich verbreiternde Erfahrungsbasis von Validitätshinweisen. Der heute unrealistische Anspruch auf ein gültiges „Persönlichkeitsmodell" wird ausdrücklich zurückgewiesen zugunsten einzelner Eigenschaftskonstrukte, die methodisch gleichartig bestimmt werden. Widersprochen wird auch dem Postulat, primär durch lexikalische (aufgrund englischer Wörterbücher) und statistische Reduktionen die Anzahl und Inhalte der psychologisch wichtigsten Persönlichkeitseigenschaften bestimmen zu können, selbst wenn die Konstruktion und die Normierung bevölkerungsrepräsentativ wären.

Auch die kombinierte Konstruktionsstrategie für Persönlichkeitsskalen hat Grenzen: Ein persönlichkeitspsychologisches Konzept deduktiv in einer sehr begrenzten Anzahl Items nachzubilden und diese Itemselektion induktiv zu kontrollieren, verlangt Kenntnis der Fachliteratur einschließlich bereits existierender Fragebogen und Forschungserfahrungen auf diesem Gebiet. Außerdem ist eine intensive kollegiale Diskussion sehr zweckmäßig, wenn nicht sogar notwendig, um die psychologischen Facetten des Konstrukts inhaltlich und sprachlich möglichst prägnant zu repräsentieren bzw. gegen andere Konstrukte abzugrenzen. Einer Fragebogenkonstruktion dieser Art sind Grenzen gesetzt. Falls auch Autoren aus anderen Bereichen der aktiven Persönlichkeitsforschung beteiligt wären, könnte durchaus ein größeres System gleichartig konstruierter Skalen entstehen. Vielleicht wird es in einer künftigen Entwicklungsphase aufgrund breiterer Kooperation ein System solcher Persönlichkeitsskalen geben, um je nach Fragestellung geeignete Module auswählen zu können.

Wie zur kriterienorientierte Strategie, so sind auch zur kombinierten Strategie herausragende Skalenkonstruktionen durch Eysenck mit den beiden wohl am besten reproduzierbaren und weithin als gültig angesehenen Dimensionen Emotionalität und Extraversion-Introversion zu nennen. Eysenck war in den 1940er Jahren als klinischer Psychologe tätig. Die Items der ursprünglichen Skala Neurotizismus wurden von ihm im Maudsley Medical Questionnaire MMQ (1958), dann mit Extraversion-Introversion im Maudsley Personality Inventory MPI (1959), in einem teils empirisch-induktiven teils hypothetisch-deduktiven Konstruktionsprozess entwickelt und mit kleineren Revisionen weitergeführt (siehe Eysenck Personality Inventory E-P-I, 1964; und die deutschen Versionen, Eggert, 1974; Ruch, 1999; vgl. auch Neyer & Asendorpf, 2018; Stemmler, Hagemann, Amelang & Spinath, 2016).

Assessmenttheorie

Die methodenbewusste kritischen Anwendung solcher Fragebogen erfordert einen Bezugsrahmen, der primär Persönlichkeitstheorie und Assessmenttheorie verbindet und sekundär die als adäquat angesehenen psychometrischen und teststatistischen Verfahren benutzt. Assessment in der Differentiellen Psychologie entspricht dem Begriff von Diagnostik, meint jedoch allgemeiner die Erfassung von psychologischen Merkmalen nach bestimmten methodischen Prinzipien zu einem praktischen Zweck, welcher eine rationale Entscheidung verlangt. Assessmentstrategien legen in einem Datenerhebungsplan fest, welches Konstrukt mit welchem Untersuchungs- und Auswertungs-Konzept erfasst werden soll. – Assessmenttheorie wird hier anstelle des herkömmlichen, oft klinisch (medizinisch) gemeinten Begriffs Diagnostik verwendet, um eine über Statusdiagnostik hinausgehende, allgemeinere theoretische Konzeption zu bezeichnen: allgemeine Prinzipien, Methoden und Methodenproblem, die sich bei der psychologischen Erfassung individueller Unterschiede als Grundlage systematischer Beschreibung und Entscheidung ergeben.

Die assessment-strategisch begründete Konstruktion von Skalen bzw. Item-Aggregaten führt diese kriterienbezogene Entwicklung als Kombination von induktiven und deduktiven Schritten weiter. Mit den Prinzipien der modernen Assessmenttheorie kann diese Strategie methodisch besser dargelegt und begründet werden. Diese Prinzipien der Assessmenttheorie haben sich seit längerem herausgebildet, es mangelt noch an der breiten und systematischen Umsetzung in praktikable Arbeitsmethoden für bestimmte Aufgabengebiete. Die Leitkonzepte werden an dieser Stelle nur genannt und später erläutert: Multitrait-Multimethod, multimodale Strategie, Generalisierbarkeitstheorie, Brunswiks Linsenmodell und Wittmanns multivariate Reliabilitätstheorie. Nach diesen Prinzipien kommt es darauf an, jeweils für bestimmte Aufgabengebiete Skalen zu konstruieren, die Prädiktoren für psychologisch interessierende Kriterien geben. – Ein unvermindert aktuelles Forschungsthema bleibt der Nachweis externer (ökologischer) Validität. Inwieweit ist von den Testwerten eines Persönlichkeitsfragebogens auf die tatsächlich im Alltag registrierten Indikatoren des individuellen Verhaltens und Befindens zu schließen? Erwähnenswert ist hier die Methodik des Ambulanten Assessment (Ecological Momentary Assessment, Experience Sampling Method), technisch unterstützt durch miniaturisierte, auch interaktive Systeme. Diese Begriffe deuten bereits ein hohes methodologisches Anspruchsniveau an, so dass von dieser innovativen Methodik kein direkter Ersatz der viel einfacheren Persönlichkeitsfragebogen erwartet werden kann. Die Methodik und Anwendung des ambulanten Assessment haben sich jedoch zu einer breiten Forschungsrichtung entwickelt. Die ersten großen und alltagnahen Untersuchungen wurden seit etwa 1980 hauptsächlich im deutschsprachigen Bereich unternommen. In Lehrbüchern der Differentiellen Psychologe und der psychologischen Diagnostik (ausgenommen Neyer & Asendorpf, 2018) ist diese innovative Forschungsrichtung auch nach 40 Jahren kaum rezipiert.

Wird es einen Strategiewechsel in der Konstruktion von Persönlichkeitsskalen geben? Von den fragwürdigen Messmodellen der internen Konstruktionsweise zu assessment-strategisch begründeten Item-Aggregaten statt der bisherigen Skalen? – Welche Fortschritte hinsichtlich der direkten Kriterienvalidierung möglich sind, muss sich empirisch zeigen. Ein allgemeiner Strategiewechsel der Fragebogenkonstruktion ist bis auf weiteres, auch wegen der höheren methodischen, auch technischen und organisatorischen Anforderungen im Vergleich zu den „Papier-und-Bleistift“ Verfahren, noch kaum zu erwarten, wohl aber der eine oder andere innovative Schritt.

3. Fragwürdige Übertragung messtheoretischer Axiome auf Selbstbeurteilungen

Persönlichkeitsfragebogen erfassen primär Selbstbeurteilungen. Weder ist ein direkter Vergleich mit dem Selbstbild und der Befindlichkeit anderer Menschen möglich, noch besteht in der Regel ein methodisches Training für diese introspektive Aufgabe, um die typischen Unsicherheiten und Urteilstendenzen zu verringern. Ob die Einstufungen *faktisch* wiederholbar sind oder wie die *Selbstbeurteilung* zustande kam, ist grundsätzlich nicht zu überprüfen, denn es gibt hier keine objektiven, unabhängig aufgezeichneten „Daten“. Bestimmte *Selbstauskünfte* könnten mit objektiven Informationen verglichen werden, doch verlangen die allermeisten Items eine retrospektive Beurteilung des gemeinten Persönlichkeitsmerkmals, d. h. eine Verhaltensweise, Reaktionsweise, Handlungstendenz, Einstellung, Gefühlslage, Befindensweise oder ein anderes Persönlichkeitsmerkmal. Es wird nicht nach aktuellen Zuständen oder Veränderungen gefragt, sondern allgemeiner, sodass es eigentlich auf eine zusammenfassende Einschätzung der wiederkehrenden bzw. überdauernden charakteristischen Merkmale

ankommt. Wer ein Item mit „stimmt“ oder „stimmt nicht“ beantworten will, wird das gemeinte Merkmal zunächst psychologisch interpretieren und dabei introspektiv beurteilen, inwieweit die Formulierung in ihren Details überwiegend zutrifft oder nicht. Am Beispiel „Ich habe häufig Kopfschmerzen“: Kopfschmerzen kommen mit sehr unterschiedlicher Lokalisation, Intensität und Verlaufsform vor und „häufig“ ist kaum zu definieren. – Dennoch werden wahrscheinlich wenige Menschen zögern, dieses Fragebogen-Item zu bejahen oder zu verneinen. Insofern wird nicht nur eine sprachlich-semantische Definition bzw. Interpretation verlangt, sondern indirekt auch eine Zusammenfassung über nicht näher definierte Zeiträume der Lebensspanne und teils auch über Klassen von (auch hypothetischen, vielleicht nie erlebten) Situationen erfragt. Solche semantischen Aspekte, die individuellen Aggregationen und spezielle Gewichtungen bleiben unbekannt.

Psychometrische Postulate

Selbstbeurteilungen aufgrund von Fragebogenitems sind also Daten im Sinne von „vorhanden“ oder „nicht vorhanden“. Falls statt dieses dichotomen Formats abgestufte Antworten vorgegeben werden, beispielsweise häufig-manchmal-selten-nie, werden Ordinaldaten gewonnen, die nur für den Vergleich innerhalb einer Person gelten (sog. *ipsative* Werte). Solche Abstufungen führen jedoch zusätzliche semantische Probleme ein, denn die verwendeten Quantoren sollten ihrerseits eindeutig definiert werden, und sie passen auf manche Frageninhalte nur schlecht; dies gilt entsprechend für Abstufungen der Intensität: sehr stark ausgeprägt … sehr schwach ausgeprägt. Mehrstufige Antwortformate haben sich nicht durchgesetzt, denn sie lösen das metrische Problem nur oberflächlich.

Aus diesen Argumenten folgt die fundamentale Unterscheidung zwischen objektiven Intelligenz- und Leistungstests einerseits und den Selbstbeurteilungen in Persönlichkeitsfragebogen andererseits. Intelligenz- und Leistungstests erfassen in standardisierter Weise Verhaltensstichproben objektiver Testleistungen. Selbstbeurteilungen in Persönlichkeitsfragebogen geben subjektive Schätzverfahren hinsichtlich nicht direkt messbarer Merkmale mentaler Repräsentationen mit unbekanntem numerischen 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 Definition einer Intervallskala kennt, wird grundsätzlich zweifeln, wenn eine Anzahl von Itemwerten zu einem Testwert auf der *Skala* für ein bestimmtes Persönlichkeitsmerkmal addiert wird, denn diese Itemwerte repräsentieren keine metrischen „Messwiederholungen“. In einem Intelligenztest sollen sich die Items in der Skala bzw. dem Untertest möglichst nur in ihrer Schwierigkeit unterscheiden, also homogen und konsistent sein, und deshalb kann die eindimensionale Skala aufgrund solcher Parallelmessungen sehr hohe interne Reliabilitätsindizes erreichen. – Die theoretischen Konstrukte der Persönlichkeitstheorien sind aus mehreren Gründen wesentlich komplizierter als ein Standard-Intelligenztest mit homogenen Untertests und folglich bleibt auch die Explikation dieser Persönlichkeitskonstrukte durch adäquate Indikatoren nur näherungsweise zu erreichen und grundsätzlich unabgeschlossen.

Die Antworten in einem Persönlichkeitsfragebogens sind nominale (kategoriale) Daten. Die einfache Addition der Itemwerte zu einem Testwert (Skalenwert) und deren interindividueller Vergleich setzen jedoch eine zugrundeliegende Skala mit gleichen Intervallen voraus. Dieser Einwand gegen die einfache Auswertung des Fragebogens trifft umso mehr die primäre Konstruktion der Skalen: traditionelle Itemanalyse, Faktorenanalyse, Item-Response (IR)-Modelle unterschiedlicher Art. – Lineare Transformationen und die entsprechenden Rechenoperationen bzw. die Annahmen über die Wahrscheinlichkeitsverteilung der untersuchten Variablen sind definitionsgemäß unzulässig. – Über die Konsequenzen dieser Feststellung existieren allerdings in der Fachliteratur große Meinungsunterschiede und überdauernde Kontroversen. In der psychologischen Testmethodik und Forschung ist es eine weit verbreitete Gewohnheit, auch den – nur als numerisch *erscheinenden* – Selbstbeurteilungen die Eigenschaften von Intervallskalen zuzubilligen, wie es in anderen Bereichen, z. B. bei objektiven Intelligenz- und Leistungstests, mit ihren sehr homogenen, parallelen Messungen 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“ und höhere Prägnanz. Dennoch bleibt es erstaunlich, wenn sehr anspruchsvolle statistische Strukturanalysen und Modellierungen gerade anhand der metrisch sehr zweifelhaften Selbstbeurteilungen in Fragebogen unternommen werden. Es gehen so viele messtheoretische und psychologische Vorentscheidungen zur Repräsentation von Eigenschaften ein, dass die Argumentation unübersichtlich wird.

Kontroversen über Messtheorie

Zwischen den Lehrbüchern der Testtheorie und Testkonstruktion scheint eine große Übereinstimmung zu bestehen: Die Daten von *Persönlichkeitsfragebogen* und ähnlichen *Klinischen Skalen* werden als *Intervalldaten* angesehen. Einige Autoren 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 Selbstbeurteilungen. Die Voraussetzungen und die Konsequenzen dieses Postulats werden nur sehr selten eingehend diskutiert. Zwar gibt es seit längerem 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), und es fehlen regelmäßig konkrete Stellungnahmen zur Konstruktion von Fragebogen im Unterschied zu Intelligenztests. Ein Bezug zu den unterschiedlichen psychologischen Datenquellen und ihren speziellen Verhältnissen wird kaum hergestellt. Was bedeutet die Forderung, die resultierenden Testwerte sollten die empirischen Merkmalsrelationen adäquat abbilden, d. h. in adäquate Zahlenrelationen transformieren? – Könnte es sich bei der „Messung" von Introspektionen und Selbstbeurteilungen um „Messung durch willkürliche Festlegung", nur um ein „numerisches Etikettieren" handeln? Welche Konsequenzen ziehen Postulate über die Passung von Selbstbeurteilungen und Mess-Struktur nach sich? Können solche Aussagen auch *psychologisch* adäquat sein?

Skeptische Stimmen wie von Michel und Conrad (1982) oder Michell (1999) und die Zweifel, ob Interpretationen angemessen sind, die über das Ordinalskalenniveau hinausgehen, sind 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" (1995, 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). Becker (2003, 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."

„Skalierung" ist ein wichtiges Gütemerkmal (u. a. Kubinger, 2002, 2003, Westhoff et al., 2004). 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). – Allerdings bleibt hier völlig offen, welches Modell der Persönlichkeitsdiagnostik gemeint ist und welche Vorhersagen des Verhaltens erreicht oder zumindest angestrebt werden. Auf der anderen Seite warnt Rost vor der Anpassung von Modellen, welche die gewünschten Aussagen gar nicht abbilden können. Dieser Merksatz ist allerdings im aktuellen Kontext doppeldeutig: „Insofern kann sich eine richtig verstandene Testtheorie als größter Kritiker der Testpraxis erweisen" (S. 29). 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 Skalenniveau 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 bereits als „Erklärung". Der tatsächliche Wert zur empirischen Vorhersage des manifesten individuellen Verhaltens bleibt auch hier ausgeklammert.

„Ohne dass ein Verfahren das Gütekriterium *Skalierung* erfüllt, sind Betrachtungen über Validität (Gültigkeit), Messgenauigkeit und Objektivität eigentlich müßig“ postulieren 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)? – Auch wenn die „lokale stochastische Unabhängigkeit“ der Itemantworten behauptet wird, mag das für elementare Aufgabenserien, beispielsweise das Zuordnen von Symbolen u. a. Aufgaben zutreffen, ist aber für Persönlichkeitsfragebogen falsch, denn es gibt Untersuchungsergebnisse für diesen, psychologisch recht trivialen Sachverhalt, dass der Kontext einen – wenn auch geringen – systematischen Effekt haben kann (z. B. Dörrer, 2011; Krampen, Hense & Schneider, 1992; Rothkirch, 2015; Schneider-Düker & Schneider, 1977). Die Empfehlung mehrstufiger statt dichotomer Itemantworten (siehe Kubinger, 2002; Moosbrugger & Kelava, 2007; Rost, 2006) kann für bestimmte Tests berechtigt sein, würde jedoch für Persönlichkeitsfragebogen zusätzliche semantische und statistische Schwierigkeiten bringen. 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. Zwangsläufig 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 heutige Empfehlungen hinsichtlich der mehrstufigen Antwortformate überdacht werden.

Die zitierten Postulate sind sehr weit entfernt von der u. a. durch Dawes (1977; Dawes, Faust & Meehl, 1989) vertretenen *pragmatischen* Auffassung, in solchen Zahlenzuweisungen nur Indizes zu erkennen, die mehr oder minder nützlich sein können, wenn Kriterien vorherzusagen sind („index measurement“). – Viele der zitierten Autoren begreifen die Messung – im Sinne von Guttman, Rasch und ihren Nachfolgern – 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, Befindens- und Verhaltensweisen und Erlebnisse, d. h. für die Inhalte der am häufigsten verwendeten psychologischen Tests, Persönlichkeitsfragebogen, Stimmungsskalen und Klinische Skalen, bedeuten kann. Die betreffenden Lehrbücher verzichten keineswegs auf Kapitel über Persönlichkeitsfragebogen, doch werden die konkreten Beispiele regelmäßig aus anderen Bereichen gewählt. Was für *beobachtbare* Verhaltensmerkmale oder Intelligenz- und Leistungstests überzeugen kann, scheint in den meisten Lehrbüchern kommentarlos auf *subjektive* Auskünfte generalisiert zu werden, 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. Bemerkenswert ist die Zurückhaltung der genannten 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 und die Alternativen persönlichkeitspsychologischer und klinischer Diagnostik darzulegen.

Auf der anderen Seite steht eine nicht geringe Zahl von Psychologen, die *psychologische* und *erkenntnistheoretische* Kritik an einer ihres Erachtens unreflektierten Mess- und Testtheorie und an einer „pseudo-naturwissenschaftlich“ an der Physik orientierten reduktionistischen Psychologie üben. Die uneingeschränkten messtheoretischen Postulate können dogmatisch wirken und provozieren Widerspruch. Diese grundsätzliche Kritik förderte wahrscheinlich die Strömung der „qualitativen“ Methodik (vgl. Mey & Mruck, 2010; Flick, von Kardorff & Steinke, 2000). Doch der Begriff „qualitativ“ ist unglücklich gewählt, weil er vieldeutig und missverständlich ist (Fahrenberg, 2002, 2015). „Qualitativ“ dient vielfach als Etikett einer Auffassung, die sich von der anscheinend dominierenden „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. Deshalb ist es adäquat, 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 Reduktionismus-Kritik auch

ideologiekritische bzw. gesellschaftskritische Argumente hinzukommen, zu einer weiteren Distanzierung oder sogar Abspaltung vom sog. Mainstream der Psychologie führen (vgl. u. a. Walter, 1999 oder die *Gesellschaft für Neue Psychologie*). – Die nachhaltige Überzeugtheit der Vertreter und der Kritiker der messtheoretischen Postulate verweist auf wissenschaftstheoretische Vorentscheidungen, die außerhalb des Messmodells liegen. Empfiehlt sich hier nicht eine mittlere Position, informiert über die fundamentalen wissenschaftstheoretischen Probleme und die statistischen Prozeduren, aber pragmatisch an Heuristik und psychologischem Nutzen interessiert?

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 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 einer Dimension der Lösungsschwierigkeit ausgewählten Items, und ebenso auf Persönlichkeitsfragebogen angewendet werden sollten. Es mangelt an einer kritisch vergleichenden Darstellung aller wesentlichen Voraussetzungen. Hinzu kommt, dass auf diesem Gebiet keine vergleichende Studie gefunden werden konnte, die an geeigneten Datensätzen und mit unabhängigen Untersuchern eine reproduzierbare Konvergenz der Ergebnisse von diversen metrischen IR-Strukturanalysen und ihrer zentralen Aussagen geliefert hat. In dieser Hinsicht scheint das Vorbild naturwissenschaftlicher Messmodelle noch nicht wirksam zu sein. Deswegen ist es angebracht, auf fragwürdige Vereinfachungen und überdauernde Schwierigkeiten aufmerksam zu machen.

Bemerkenswert ist, dass ein größerer Anteil von Personen nicht zu einem ausgewählten Messmodell zu passen scheint, sie wurden deshalb als „nicht-skalierbar“ bezeichnet (Austin, Deary, Gibson, McGregor & Dent, 2006; Forman, 2002; Ponocny & Klauer, 2002). – Ob der Begriff „nicht-skalierbare“ Person die Persönlichkeitspsychologie sinnvoll bereichert, ist eine andere Frage. Wird ein so etikettierter Mensch sich vielleicht diffamiert oder eher erleichtert fühlen, ein Individuum zu sein, bei dem ein Messmodell versagt? Erstaunlich ist, dass in diesem Zusammenhang nicht gründlich erörtert und repräsentativ untersucht wird, wie groß der Prozentanteil der „nicht-skalierbaren“ Personen ist, und welche psychologischen Hypothesen vorzubringen sind.

Die Schlussfolgerung aus diesen Überlegungen lautet, dass es gegenwärtig kein breit überzeugendes „Messmodell“ für die Konstruktion von Persönlichkeitsfragebogen gibt und wahrscheinlich weiterhin nicht geben wird. Die Meinungsunterschiede werden bleiben: Sind hier bestimmte Messmodelle bzw. Skalierungen grundsätzlich überlegen? Oder handelt es sich um mehrere heuristischen Verfahren, innerhalb eines persönlichkeitspsychologischen Annahmengefüges statistische Hinweise zur empirisch geleiteten Ordnung von Konstruktfacetten zu gewinnen und empirisch nützliche Indizes abzuleiten? Im Unterschied zu der breiten Entwicklung von Modellierungsmöglichkeiten mangelt es an einer prägnanten Darstellung des Geltungsbereichs bzw. der Indikation solcher Modellierungen. Welche Datenquellen verlangen welche Modellierungen? Von dieser Diskussion wäre allerdings zweierlei zu erwarten. Die Diskussion muss erstens auf die betreffende Domäne eingehen, d. h. sie muss sich inhaltlich auf die psychologische Operationalisierung dieser speziellen Klasse von theoretischen Konstrukten beziehen, und zweitens die Absichten der Testentwicklung kennen, d. h. die Assessmentstrategien und die möglichen empirischen Validierungskriterien berücksichtigen. Geht es vorrangig um die messtheoretischen Postulate von Eindimensionalität, spezifischer Objektivität und erschöpfender Statistiken der Eigenschaftsausprägung oder kommt es hauptsächlich auf den Entscheidungsnutzen eines empirisch validierten Testwertes an? – Rost (1999) argumentierte, dass das Rasch-Modell und die klassischen Prinzipien der Testkonstruktion nicht als konkurrierende, sondern als komplementäre Modellperspektiven angesehen werden können. Doch welche wechselseitigen Ergänzungsmöglichkeiten könnte es bei der

Konstruktion von Persönlichkeitsfragebogen geben? Falls die künftige Entwicklung von Persönlichkeitsfragebogen und ähnlicher Skalen sich primär an den Prinzipien der Assessmenttheorie, d. h. auch an den externen Kriterien und dem Entscheidungsnutzen orientieren würde, könnte sich ein dritter Weg ergeben.

4. Eigenständige Konstruktionsprinzipien der Persönlichkeitsfragebogen

Item-Entwicklung als Operationalisierung von Persönlichkeitseigenschaften

Persönlichkeitseigenschaften werden allgemein als *facettenreiche* Dispositionen verstanden, die sich zwar zeit- und situationsabhängig unterschiedlich, jedoch relativ überdauernd im individuellen Erleben und Verhalten manifestieren. Nur relative Unterschiede bestehen zwischen den Eigenschaftskonstrukten und den Konstrukten im Bereich der Zustandsänderungen der Befindlichkeit, u. a. der charakteristischen emotionalen und motivationalen Verläufe, wobei deren höhere intra-individuelle Variabilität natürlich nicht einfach ein Messfehler ist (siehe Heiß, 1948, „Person als Prozess“, Cattells, 1950, state-trait-Forschung). Die Konstruktionsweise von Persönlichkeitsfragebogen und ähnlichen Skalen unterscheidet sich hauptsächlich in zwei Bereichen von Intelligenz- und Leistungstests. Erstens in dem psychologischen Prozess der Itementwicklung und zweitens in der Zielsetzung der Skalenkonstruktion. Einerseits muss der psychologische Facettenreichtum einer Persönlichkeitseigenschaft nachgebildet werden, andererseits verlangt die metrische (formale) Prägnanz der Skala „parallele Messungen“ durch gleichartige Items. So besteht der Zielkonflikt: Metrisch homogen oder inhaltlich heterogen? Hohe interne Konsistenz (Reliabilität) der Skala oder optimale Gültigkeit hinsichtlich des Konstrukts und der externen Kriterien? Stemmler (1996) unterschied:

- den Bereich des Konstrukts: Ist ein Konstrukt durch Variation zwischen Personen, Settings/Situationen, Variablen oder Kombinationen hiervon definiert?
- die Operationalisierungen: In welchem Modus werden die Operationalisierungen vorgenommen (Person, Setting/Situation oder Variablen)?
- den Anwendungsbereich: Welches sind die Einheiten des Assessment, auf welche sich die Schlussfolgerungen beziehen; und in welchem Modus (Person, Setting/Situation oder Variablen) sind sie zu finden? (S. 261–262).

Von dieser Dreiteilung ausgehend gelangt er zu einer Taxonomie von neun „models of construct assessments“.

Für die Itementwicklung gilt, dass die psychologisch wichtigsten Komponenten des Konstrukts durch ein Spektrum inhaltlich ähnlicher Items indiziert werden müssen. Die Homogenität der Items im statistischen Sinne ist 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 Items erfasst sind, das verwendete Item gehört nicht zu dem gemeinten Konstrukt, das verwendete Item enthält gleichzeitig auch Aspekte eines anderen, 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 Operationalisierungsfehler und formale Mängel vermieden werden:

- Die Itementwicklung setzt voraus, dass häufig kaum abzugrenzende und nicht einfach zu explizierende Eigenschaftskonstrukte aufgrund zentraler persönlichkeitspsychologischer Arbeiten und auch eigener Forschungserfahrung in diesem Bereich gut bekannt sind. 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 sowie mögliche Einflüsse von Geschlecht, Alter u. a. soziodemografischen 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, aber desto länger und schwieriger wird die Formulierung.

Die Beantwortung eines Persönlichkeitsfragebogens verlangt – genau genommen – 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-referenzielle Rekonstruktionen solcher Selbstbeurteilungen als persönlichkeitspsychologische Feststellungen. Sie sind durch kognitive Schemata und soziale Stereotype, alltagspsychologische Vorstellungen, formale Antworttendenzen 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.

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 (u. a. 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.

Einige Lehrbücher nennen Regeln, wie Fragebogenitems formuliert werden sollten (u. a. Bühner, 2011). Die Verständlichkeit zu optimieren ist sehr erstrebenswert – soweit es eben geht. Dazu gehören natürlich die eingehenden fachlichen Erörterungen von theoretischem Konstrukt und adäquaten Indikatoren, die Variation der Formulierung und empirische Daten aus möglichst breit angelegten Voruntersuchungen. Ist auch an den Sprachgebrauch in den deutschsprachigen Nachbarländern zu denken und an die sich wandelnden Ausdrucksweisen und Tendenzen, wenn ein Fragebogen wie das FPI über 50 Jahre weitergeführt wird? Rückmeldungen über die Verständlichkeit, über die Anzahl der Fragen, über die Plausibilität und Akzeptanz des Tests sind am besten anlässlich der bevölkerungsrepräsentativen Normierung zu gewinnen.

Auffällig ist, dass den statistischen Operationen oft mehr Aufmerksamkeit gewidmet zu werden scheint als dem notwendigen und stufenweise-rekursiven Prozess, wie die gemeinte Persönlichkeitseigenschaft durch die Auswahl und Formulierung der Fragen erfasst wird, um das gemeinte Konstrukt psychologisch adäquat zu erfassen (Löhr & Angleitner, 1980). Wenn hauptsächlich die Formulierung und äußere Gestaltung der Items erörtert werden, kann das an dem viel höheren 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. Dennoch ist kein fachliches Protokoll einer solchen Explikation und Rekonstruktion eines persönlichkeitstheoretischen Konstrukts bekannt, um das Ineinander von deduktiven, itemmetrischen, induktiven und interpretativen Schritten im Hinblick auf die geplante Operationalisierung, die antizipierte Anwendung des Tests und dessen deskriptiven und prädiktiven Nutzen vermitteln zu können. Sollten solche kooperativen Versuche nicht in der fachlichen Ausbildung vorkommen? – 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.

Der Versuch der Operationalisierung einer Persönlichkeitseigenschaft mittels einer Fragebogenskala verlangt theoretische Vorentscheidungen u. a. zum psychologischen Gültigkeitsbereich (Auswahl der interessierenden Konstrukte), 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 die Breite der zu erfassenden Facetten, Itemtyp, Antwortformat, Itempool, testtheoretisches Modell, Qualität der bevölkerungsrepräsentativen Konstruktion und Normierung, Überprüfung der formalen Gütemerkmale und empirischer Aufwand für die wichtigsten Validitätshinweise.

Innere Konsistenz (Item-Homogenität) und deren Bedeutung

Wie die Itemanalyse so ist auch die Methodik der Faktorenanalyse primär für die Intelligenzforschung entwickelt und später auf die Konstruktion von Persönlichkeitsfragebogen und Stimmungsskalen übertragen worden. Bei *schematischer* Anwendung begünstigen beide Strategien, die konventionelle Itemanalyse und die faktorenanalytische Methodik, die Entstehung sehr homogener Skalen, d. h. die Maximierung der *inneren* Konsistenz ohne Rücksicht auf die *externe* empirische Validität. Das kann geschehen, wenn anfänglich enthaltene, d. h. vorgegebene oder noch unzureichend erkannte Dubletten oder Tripletts weitgehend redundanter Items aufgrund ihrer sehr hohen Kommunalität die Ladungsmuster bzw. die Rotation dominieren, mit Folgeschäden bei der Beurteilung der relativen Varianzanteile und 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.

Wenn es in der Skala eines Intelligenztests z. B. darauf ankommt, Symbole zuzuordnen, schwieriger werdende Rechenaufgaben zu lösen oder sich Zahlenreihen zu merken, kann jeweils für die vielen einzelnen Operationen gewiss eine hohe Homogenität hinsichtlich dieser elementaren Intelligenzfunktionen behauptet und dementsprechend skaliert werden. Demgegenüber haben auch die persönlichkeitspsychologisch wohl am besten bewährten Konstrukte der Sekundärfaktoren Extraversion und Emotionalität im Sinne Eysencks so viele wichtige Facetten, dass einige verhältnismäßig heterogen erscheinende Items kombiniert werden müssen, um dieses Eigenschaftskonstrukt zu repräsentieren. Die typischen Extraversions-Items zu impulsivem und zu geselligem Verhalten betreffen keine homogenen, sondern unterschiedliche, nicht notwendig hochkorrelierte Komponenten oder Facetten eines theoretischen Konstrukts. Die Itemselektion, die ausschließlich aufgrund der Höhe der Faktorladung oder der parallelen Messung bzw. Skalenverträglichkeit nach dem Prinzip der IR-Messmodelle vorgenommen wird, führt zu sehr homogenen Skalen. Folglich könnte sich die Chance einer externen empirischen Gültigkeit verringern. Auch die sog. Guttman-Skala forderte 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 Disposition zu *Körperlichen Beschwerden,* eine der varianzstärksten und stabilsten Persönlichkeitseigenschaften überhaupt und eng assoziiert mit der Dimension Emotionalität, anschaulich bedeuten? Sollte jemand, der die seltenen („schwierigen") Herzschmerzen hat, aus Gründen der Eindeutigkeit auch die häufigeren („leichten") Magenschmerzen, und jedenfalls die sehr häufigen Kopfschmerzen haben?

Sehr hohe Trennschärfe-Koeffizienten und Faktorladungen sollten bei einem Persönlichkeitsfragebogen eher ein Anlass sein, solche Items kritisch zu bewerten und eventuell 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. 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 einen Verlust an u. U. wesentlichen Facetten des gemeinten theoretischen Konstrukts und damit wahrscheinlich einen Verlust externer Validität. Generell sind Reliabilitätskoeffizienten als nur *formale* Aspekte eines Messmodells zweitranging gegenüber den Validitätsbelegen. Die Herausforderung bei der Itemselektion für Persönlichkeitsfragebogen liegt darin, die statistisch-formalen Eigenschaften eines Items *und* das bereits vorhandene persönlichkeitstheoretische Wissen und möglichst viel Kriterieninformation zu berücksichtigen. Aus diesen Gründen ist jede *schematische* Itemanalyse zur Maximierung der Konsistenz von Persönlichkeitsfragebogen unangebracht. Die Operationalisierung ist eine *psychologische* Aufgabe und benötigt das theoretische Annahmengefüge über das Konstrukt, kann also durch den Formalismus einer Item- oder Faktorenanalyse nicht ersetzt, sondern nur unterlegt 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. Homogenität im Sinne der IR-Modellierungen ist nicht identisch mit Homogenität im Sinne von Item-Korrelationen oder mit Faktor-Reinheit. Aber allgemein gilt: relativ geringere Konsistenz (Homogenität) hat nicht notwendig geringere Kriterienvalidität zur Folge, umgekehrt ist hohe Konsistenz (Homogenität) keine Gewähr für Validität, so betonten schon 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 insoweit wichtig, als sie zur Validität beiträgt' (S. 432)" (S. 53–54). Auch Lienert (1961) stellte diesen Zielkonflikt 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, siehe auch Lienert & Raatz, 1994). – Bemerkenswert ist, dass dieses Reliabilitäts-Validitäts-Dilemma, trotz dieser Erklärungen, keineswegs in allen Lehrbüchern der Testkonstruktion erwähnt oder hinreichend ausgeführt 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?

Natürlich sind die konventionellen Reliabilitätsschätzungen und die internen Item- und Faktorenanalysen, auf einfachste Weise möglich, dagegen ist der Nachweis einer neuen (externen) Kriterienkorrelation (nicht bloß mit ähnlichen Fragebogen) oder sogar eines inkrementellen Nutzens für reale Assessment-Entscheidungen sehr viel schwieriger und aufwändiger. 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 Vertrauensintervalle 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 nur die Stanine-Grobnormen eingeführt, um auf diese Unsicherheiten hinzuweisen, später wurden dann ungewöhnlich große Normierungsstichproben verwendet, um in den Tabellen die beträchtlichen Effekte von Geschlechtszugehörigkeit und differenzierten Altersgruppen 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 revidierte FPI-R, vor allem aus *Gründen der Länge und Zumutbarkeit*, nur noch 137 (+1) Items für die 10 Standardskalen sowie E und N. Eine relativ hohe bzw. hinreichende 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. Eine hochgradige Stabilität der individuellen Testwerte, etwa nach dem Vorbild von Intelligenztests, auch für Testwerte von Persönlichkeitsfragebogen zu fordern, wäre unangebracht. Die Idee einer sich dynamisch verändernden Persönlichkeit und das Konzept einer Person-Situation-Interaktion sind so alt wie die gegenteilige Vorstellung feststehender, „eingeritzter" Charaktereigenschaften. Es gibt durchaus eine kurz- und mittelfristige, funktionelle, motivations- und situationsabhängige intra-individuelle Variabilität. Methodisch ist es schwierig, differenzierte Aussagen über situative Effekte und relativ überdauernde Persönlichkeitsdispositionen zu machen. Fragebogendaten, die nur unter verschiedenen Instruktionsbedingungen gewonnen sind, werden hier nicht überzeugen (Deinzer, Steyer, Eid & Notz, 1995). Heute ist hier das ambulante Assessment die Methode der Wahl.

Explorative Verwendung der konventionellen Itemanalyse, Faktorenanalyse, Clusteranalyse und IR-Modellierungen

Im Jahr 1970 stützte sich der primäre Konstruktionsprozess des FPI auf Item- und Faktorenanalysen und unterstellte hier, wie auch bei den normierten Testwerten, eine Intervallskalierung. Die Zusammenfassung der Itemantworten zu individuellen Testwerten könnte zwar als das einfache Zählen nominaler Daten aufgefasst werden. Dabei wurde nicht übersehen, dass damit den einzelnen Items für den interindividuellen Vergleich psychometrisch ein gleiches Skalenintervall unterstellt wird. Bereits 1973, in der 2. Auflage des FPI, wurde über Clusteranalysen, d. h. die nicht-metrische Guttman-Lingoes Smallest Space Analysis, und hierarchische Clusteranalysen berichtet. Clusteranalysen nach Wards Verfahren dienten auch später neben der Faktorenanalyse als eines der Hilfsmittel zur empirischen Ordnung der psychologisch deduzierten Items des betreffenden Eigenschaftskonstrukts – *nicht* zur Dimensionierung des Pools. Die Gruppierung der Items entsprach weitgehend

der Faktorenanalyse bzw. der Item-Skalen-Zuordnung (mit einzelnen Abweichungen). Auch Clusteranalysen verlangen Entscheidungen: welcher Ähnlichkeitskoeffizient und welcher Algorithmus verwendet und bis zu welcher Clusterzahl zusammengefasst wird. Daneben wurden, wie im Manual der 2. Auflage berichtet, heuristisch auch multidimensionale Skalierungen, eine Skalogramm-Analyse von Guttman (R. Hampel), mit einem der ersten funktionsfähigen Computerprogramme (LGM des IPN in Kiel) auch Rasch-Skalierungen mit Modell-Tests nach Fischer durchgeführt (J. Fahrenberg) sowie Analysen der Reproduzierbarkeit der Strukturen in verschiedenen Modellen der Faktorenanalyse diskutiert (siehe auch Wittmann & Hampel, 1976). – 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 und seltsamerweise das beste 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. Diese explorativen Analysen der FPI-Daten mit LGM führten also je nach Blickwinkel zu der Schlussfolgerung, dass (1) der auf Maximierung der Homogenität zielende LGM-Algorithmus für die fachlich interessierenden Persönlichkeitseigenschaften ungeeignet ist oder dass (2) die durch von Item- und Faktorenanalysen unterstütze induktiv-deduktiv konstruierten Items bzw. Skalen keine suffizienten Messungen dieser Eigenschaften liefern. Aus den erhaltenen *LGM*-Ergebnissen wurde der Schluss gezogen, dass sich jenes voraussetzungsreiche Verfahren der Rasch-Skalierung 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.

Auch die 1999 unternommenen Analysen mit dem damals aktuellen Programm *WINMIRA*, beraten und unterstützt durch M. von Davier (2005), in Kiel und Washington, führten zu einem sehr ähnlichen Eindruck. Hinzu kam die irritierende Erfahrung, wie groß der Spielraum für den Untersucher ist, die Modelltest-Ergebnisse gerade bei der Analyse großer Datensätze durch die Variation von Kriterien und Signifikanzniveaus (hier gerade mit problematischen bzw. störenden Auswirkungen *hoher* Personenzahlen für die statistischen Schätzungen) in einem großen „Spielraum“ zu steuern. Es gibt wesentlich mehr Freiheitsgrade als beispielweise bei einer konventionellen Faktorenanalyse mit Varimax-Rotation. Erstaunlich ist, dass anscheinend auch weiterhin keine systematische Prüfung der *Reproduzierbarkeit* solcher Skalierungen, auch zur Konstruktion von Intelligenztests im Vergleich zu Persönlichkeitsfragebogen, zu existieren scheint. Wäre nicht zu erwarten, dass gerade bei der Anlehnung an naturwissenschaftliche Messmodelle auch kritisch-vergleichende Methodenstudien existieren bzw. in den Lehrbüchern nachdrücklich gefordert werden?

Sparsame Beschreibung mittels unkorrelierter Dimensionen und Skalen?

Das Prinzip der sparsamen Beschreibung ist dem naturwissenschaftlichen Denken entlehnt und hat dort große Überzeugungskraft. 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 anhand eines minimalen Systems orthogonaler Faktoren auf die psychologisch sparsamste und einfachste Weise zu beschreiben? Zu den möglichen Fehlbewertungen faktorenanalytischer Ergebnisse gehört, diese zur Datenübersicht 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 fragwürdige Vorentscheidungen an. Nicht allein die orthogonalen Faktoren bestimmen in der Regel das Beschreibungssystem, sondern auch die Skalenwerte sollen möglichst niedrig korrelieren. Solche Postulate können leicht zu Einseitigkeiten oder Fehlbewertungen der Konstruktion von Persönlichkeitsfragebogen führen. Ungleich wichtiger bleiben ja die inhaltliche theoretische Rechtfertigung und der psychologische Kontext von Forschung und Diagnostik eines wichtigen und facettenreichen Eigenschaftskonstrukts.

Testkonstruktive Alternativen und Kriterienvalidierung

Die Persönlichkeitsfragebogen unterscheiden sich graduell, welches Gewicht den inhaltlichen oder den teststatistischen Konzepten gegeben wird, wie elaboriert, repliziert oder multistrategisch dieser Konstruktionsprozess ist. Generell sind sie *intern* aufgrund eines mehr oder minder bevölkerungsrepräsentativen Datensatzes konstruiert, d. h. ohne maßgebliche externe

Kriterieninformation. Doch zuvor war gerade diese kriterienorientierte Konstruktion, primär auf psychiatrische Diagnosen bezogen, als großer Fortschritt angesehen worden. Das weithin bekannte *Minnesota Multiphasic Personality Inventory* MMPI (Hathaway & McKinley, 1943, deutsche Adaptation Engel, 2000) war auch für Eysenck ein Vorbild seiner herausragenden Testentwicklung hinsichtlich Emotionalität und Extraversion-Introversion. So ist zu überlegen, ob sein wissenschaftlicher Erfolg in diesem klinisch-psychologischen Bezug begründet ist. Diese Konstruktionsstrategie ist natürlich von der zumindest relativ überdauernden Gültigkeit und praktischen Relevanz jener Kriterien abhängig. Durch die Revisionen der Klassifikationssysteme in der Psychiatrie verlor das MMPI zeitweilig einige wichtige Kategorien seiner Kriterienbasis (Engel, 2000; 2019). Anzumerken ist, dass spezielle Fragebogen existieren, die sowohl kriterienorientiert als auch intern konstruiert sind, u. a. in der Klinischen Psychologie und in der Berufs- und Personalpsychologie. In diesem Sinne handelt es sich um lebenslaufbezogene Kriterien, die auch die wichtigsten biografischen Grundzüge und familiäre, berufliche und gesundheitliche Entwicklungen umfassen. Hauptsächlich sind jedoch klinisch-psychologische und psychiatrische Diagnosen, abweichendes Verhalten, Straffälligkeit, pädagogische, sozialtherapeutische, psychotherapeutische und rehabilitationspsychologische Verläufe zu nennen. Die sich über 30 Jahre erstreckende Kohortenstudie an der Züricher Psychiatrischen Klinik durch Jules Angst und Mitarbeiter ist ein bemerkenswertes Projekt. Es hätte über das verwendete FPI (bzw. dessen dreidimensionale Rekonstruktion) hinaus im Prinzip auch zu einer eigenständigen Testkonstruktion führen können. Die weit überzogene Auseinandersetzung über die „richtige“ Anzahl oder die „wichtigsten“ von einem Persönlichkeitsfragebogen zu erfassenden Eigenschaften fällt sogar in der deutschen Fachliteratur auf. Auf der testkonstruktiven Ebene bleiben solche Entscheidungen beliebige Grenzziehungen, die nicht mit einer *persönlichkeitspsychologischen Begründung* oder dem Nachweis *externer Validität* im Hinblick auf wichtige Praxisfelder der Psychologie verwechselt werden dürfen. Aus dieser Sicht ist zu fragen: Inwiefern haben IR-Modellierungen der Skalen von Persönlichkeitsfragebogen im Vergleich zu der traditionellen Konstruktionsmethodik überhaupt eine höhere oder gar überlegene Vorhersageleistung relevanter Kriterien der psychologischen Praxisfelder erreicht? Systematisch vergleichende Belege scheinen zu fehlen. So wäre beispielsweise zu prüfen, welche Kriterienvorhersagen durch Fragebogen der FFM-Gruppe (NEO-FFI u. a.) geleistet werden, wenn hierbei die von Eysenck inspirierten Skalen Extraversion und Emotionalität ausgeklammert würden. Ist nicht zu vermuten, dass in einer weiter fortgeschrittenen Entwicklungsphase erneut die externe, d. h. die auf relevanten Kriterien bezogene Skalenkonstruktion vorherrschen wird? Dann würden primär nicht die Messmodelle, sondern die Assessmentstrategien den Kern der empirisch begründeten Testtheorie bestimmen. Gegenwärtig scheinen solche Forschungsansätze, die für die Konstruktion und praktische Evaluation von Fragebogen ungleich wichtiger wären, relativ geringes Interesse zu finden und fast zu stagnieren. Gemeint ist insbesondere die kriterienbezogene prädiktive Validität (Vorhersageleistung) psychologisch relevanter Kriterien und individueller Entwicklungen. Die Konzeption eines Assessment-Centers im Bereich der Personalpsychologie gibt hier ein organisatorisches Beispiel, das jedoch in der Regel kein paralleles follow-up auch der abgelehnten Bewerber leisten kann. Der Entscheidungsnutzen kann jedoch nur dann schlüssig evaluiert werden, wenn auch der eventuelle Schaden solcher Auswahl-Entscheidungen kritisch erfasst ist. Die Schlussfolgerung lautet, dass solche Vorhaben nur noch aus den Institutionen heraus und nicht von externen Forschungsgruppen organisiert und getragen werden können. Es ist zu vermuten, dass in einigen solcher großen Institutionen Persönlichkeitstests eingesetzt und der Entscheidungsnutzen intern evaluiert wird, ohne jedoch die Ergebnisse zu publizieren. Selbst in jenen Institutionen würde die mögliche empirische Basis kaum für einen Persönlichkeitsfragebogen mit einer mittleren Anzahl von Skalen wie beim FPI-R ausreichen. Doch die zuvor geschilderte kombinierte Strategie der Konstruktion bedeutet, dass es nicht schwierig ist, weitere Skalen hinzufügen und gemeinsam zu normieren, ihre internen statistischen Beziehungen zu beschreiben und die einzelnen Skalen bei Bedarf bausteinartig einzubeziehen, ähnlich den Untertests eines mehrdimensionalen Intelligenztests. Gerade diese Konzeption wäre der Ausdruck Persönlichkeitsinventar adäquat. – Deutlich ist aber, dass solche Projekte die hochkoordinierte Zusammenarbeit von Psychologen und Institutionen voraussetzt. Einer fortgeschrittenen methodenbewussten Psychologie wäre dieses Vorgehen weitaus adäquater als die einzelnen, häufig wieder abgebrochenen, nicht revidierten oder von speziellen Interessen oder individuellen Maßstäben geprägten Testentwicklungen.

Zusammenfassend ergibt sich: Wegen der zugrundeliegenden Selbstbeschreibungen besteht ein fundamentaler Zweifel an dem Postulat einer interindividuell gültigen Metrik auf einer Intervallskala. Wesentliche methodische Besonderheiten sind auch die individuellen Urteilsprozesse, da die Befragten über viele Elemente der Erlebnis- und Verhaltensweisen, über mögliche Zeiträume, über verschiedenste Situationen (Gelegenheiten) aggregieren müssen. In diesem Bereich bestehen mehr semantische Probleme hinsichtlich der Iteminhalte und der Antwortmodi sowie ein generell stärkerer Einfluss von

Kommunikationsbedingungen und Antworttendenzen. Die auf dem Gebiet der Intelligenz- und Leistungstests gewonnen Strategien und Kriterien der Testkonstruktion sind nicht direkt im Sinne eines allgemeinen Messmodells auf Persönlichkeitstests zu übertragen. Persönlichkeitskonstrukte sind eben nicht item-homogen, sondern mit psychologischer Berechtigung inhaltlich sehr facettenreich-heterogen. Bei Reanalysen ist genau und kriterienorientiert zu prüfen, ob die Forderung nach metrisch hochkonsistenten Skalen nicht technisch zu einer Ansammlung von weitgehend homogen-redundanten, sprachlichen Varianten mit minimaler inkrementeller Kriterienkorrelation führt (Reliabilitäts-Validitäts-Dilemma). Können hier bestimmte Messmodelle als grundsätzlich überlegen gelten? Oder handelt es sich primär um *heuristische* Verfahren, die mehr oder minder fruchtbare *Hinweise* liefern, die Konstruktfacetten innerhalb eines persönlichkeitspsychologischen Annahmengefüges zu ordnen, empirisch nützliche Indizes abzuleiten und deren Kriterienvalidität zu belegen? Aus den skizzierten Gründen kann eine schematische Anwendung faktoren- und itemanalytischer Strategien oder spezieller Messmodelle, ohne multistrategische Kontrollen und ohne die – letztlich persönlichkeitstheoretische – Bewertung der Itemanalysen, zu psychologisch inadäquaten Skalenkonstruktionen führen.

5. Kontroversen über die „größten und wichtigsten“ Persönlichkeitsfaktoren

Durch die ältere Charakterkunde und die neuere Persönlichkeitsforschung zieht sich ein Programm, die wichtigsten Eigenschaften zu bestimmen. Umstritten ist nicht nur die Methodik, sondern auch die Abgrenzung. Die lexikalisch-induktive Methodik bezieht sich auf Lexika der Alltagssprache, um die Grundeigenschaften zu bestimmen. Aber enthalten diese Lexika bereits alle für die wissenschaftliche Psychologie wesentlichen Aspekte bzw. Wörter? Sind die Fachausdrücke aus der Psychologie und aus der bisherigen empirischen Persönlichkeitsforschung einfach entbehrlich? Welche psychologischen Merkmale bilden den Bereich (die Domäne) der Persönlichkeit und welche Merkmale der Individualität gehören nicht dazu und werden systematisch ausgeklammert? Wie wird diese Abgrenzung begründet und welche Konsequenzen ergeben sich für die Konstruktion von Persönlichkeitsfragebogen? Statistische Reduktionen dieser Art sind technisch möglich – sind sie, im Vergleich zu anderen Konstruktionsstrategien, auch psychologisch, theoretisch und hinsichtlich der deskriptiven und praktisch-diagnostischen Aufgaben adäquat?

In der faktorenanalytisch ausgerichteten Testkonstruktion taucht immer wieder die Behauptung auf, *die* basalen Faktoren der Persönlichkeit gefunden zu haben, als ob es um real existierende Entitäten ginge – statt nur um ein mehr oder minder nützliches Beschreibungssystem. Diese Behauptung, die wichtigsten oder die größten (und deshalb?) bedeutendsten Faktoren gefunden zu haben, ist typisch für die induktive Konstruktionsstrategie, die von einer möglichst umfassenden lexikalischen Basis psychologischer Deskriptoren von Persönlichkeit ausgeht. Vereinfacht gesagt, werden im ersten Schritt alle Wörter mit persönlichkeitspsychologischer Bedeutung gesammelt. Im zweiten Schritt werden diese nach sprachlich-semantischen Aspekten beurteilt und auf eine sehr viel kleinere Anzahl reduziert und zur Formulierung entsprechender Items verwendet, bevor im dritten Schritt eine größere Anzahl von Personen durch ihre Selbstbeurteilungen einen Datensatz liefert, dessen Faktorenanalyse eben jene großen Faktoren nachweisen soll.

Das Postulat, universell gültige Persönlichkeitsfaktoren erhalten zu haben, hat zwei wesentliche Voraussetzungen: die Domäne der hier gemeinten „Persönlichkeit“ muss prägnant von den anderen Merkmalen der Person (Individualität) abgegrenzt werden und die empirische Ableitung der Faktoren muss nachweisbar repräsentativ sein, d. h. nicht allein für den Raum der Items, sondern auch für den Raum der Personen gelten. Bereits Guilford (1959), einer der Pioniere faktorenanalytischer Intelligenz- und Persönlichkeitsforschung, hat vor 60 Jahren festgestellt, dass höchstens dann über die basale oder gar erschöpfende Anzahl der relevanten Faktoren diskutiert werden könnte, wenn das *Universum* der Items der betreffenden Domäne repräsentiert ist. Auch die Auswahl der Personen für die Testkonstruktion muss repräsentativ sein, denn die Selbstbeurteilungen in einem Fragebogen hängen in erheblichen Varianzanteilen mit Unterschieden hinsichtlich Alter, Schulabschluss u.a. sozioökonomischen Bedingungen zusammenhängen. Diese Forderung nach domän-repräsentativer Itemauswahl und bevölkerungsrepräsentativer Datenbasis ist deshalb so entscheidend, weil die lexikalisch-induktive Strategie über die statistische Prozedur der Faktorenanalyse hinaus *keine anderen Maßstäbe* enthält. Die Skalenkonstruktion richtet sich weder nach externen Kriterien noch geht es primär um das Ziel, bestimmte, aus der bisherigen Forschung bekannte *Persönlichkeitseigenschaften* durch gültige Items noch besser zu erfassen, d. h. theoretische Konstrukte systematisch und adäquat zu explizieren bzw. zu operationalisieren. Einerseits könnte in der lexikalisch-induktiven Methode ein Vorteil gegenüber den deduktiven, den kriterienorientierten und den kombinierten Strategien bestehen, andererseits stellt sich die Frage, inwieweit

solche Faktorisierungen der Alltagsprache – in der Konsequenz des psycholinguistischen Postulats – kognitions- und sozialpsychologisch betrachtet, noch stärker als die anderen Konstruktionsstrategien, Schemata und Stereotype repräsentieren.

Anspruch und Problematik der lexikalisch abgeleiteten Persönlichkeitsinventare

Von den USA ausgehend, insbesondere durch Goldberg (1981), Costa und McCrae (1985), und John (1990), in der Nachfolge früherer Autoren wie Cattell, Saunders and Stice (1957), ist eine größere Gruppe von lexikalisch angelegten und faktorenanalytisch reduzierten Persönlichkeitsfragebogen entstanden, außerdem zahlreiche Varianten aufgrund von Übersetzungen oder Nachkonstruktionen in anderen Sprachen. So existieren auch deutschsprachige Varianten NEO-FFI, NEO-PI-R, außerdem erst nachträglich konstruierte Unterskalen (30 bzw. 16 Facetten), Revisionen und Kurzformen. Gemeinsam ist dieser Gruppe von Tests, dass *genau 5* wichtige Persönlichkeitsfaktoren behauptet werden: die Big Five bzw. Fünf-Faktoren-Lösungen (auch Fünf-Faktoren-Modelle FFM). Diese Testautoren scheinen sich kaum mit den grundsätzlichen Einwänden gegen die Konstruktionsweise und gegen ihre Postulate auseinanderzusetzen. Eine aktuelle Übersicht über solche lexikalisch-induktiv entwickelten Fragebogen geben Neyer und Asendorpf (2018). Hier wird auch die neuere Entwicklung referiert, die zu einer erstaunlichen Diversifikation solcher „Modelle“ mit einer Spannweite zwischen 2 und 16 Hauptfaktoren führte. Eine frühere Übersichtstabelle enthielt bereits 14 solcher Konstruktionsversuche (Amelang & Bartussek, 1997, S. 366) und durch nachträglich Aufspaltung von Faktoren in Facetten sowie andere Revisionen und Kurzfassungen hat sich die Anzahl noch wesentlich erhöht. Angesichts dieser Diversifikation sollte heute auf den vagen Ausdruck *Big Five* verzichtet werden zugunsten der Typbezeichnung FFM.

Im deutschen Sprachgebiet bildet *Towards a taxonomy of personality descriptors in German: A psycho-lexical study* (Angleitner, Ostendorf & John, 1990) die umfangreichste psychologisch-lexikalische Studie, die vom Wahrig-Lexikon ausgehend zu einer sehr großen Liste von Deskriptoren führte. Mit einer nachträglichen Ergänzung durch weitere psychologisch wichtige Wörter aus der Fachliteratur führte die Analyse im nächsten Schritt zur Reduktion auf 5092 Adjektive und schließlich zu deren Zusammenfassung in 13 Kategorien, von denen nur 9 die als geeignet angesehenen Einheiten der Persönlichkeitsbeschreibung enthalten würden: Temperament and character traits; Abilities, talents, or their absence; Experiental states: emotions, moods, and cognitions; Physical and bodily states; Behavioral states: observable actitivities; Roles and relationships; Social effects: reactions of others; Pure evaluations; Attitudes and world views. Diese erste Reduktion wurde von acht Studierenden der Psychologie durchgeführt, jeweils zwei für 10 Prozent der Wörter, die folgende Reduktion auf die neun Kategorien durch zehn Beurteiler, u. a. auch graduierte Psychologen. Die Autoren berichten teilweise Übereinstimmungen, aber auch deutliche Diskrepanzen zu ähnlichen Projekten in anderen Sprachen. – Aus den wichtigsten Adjektiven dieser neun Kategorien oder Domänen müssten im nächsten Schritt die für Fragebogen geeigneten Items formuliert werden.

Zu erwarten war, dass die Entwicklung des deutschen *NEO-Fünf-Faktoren Inventar (NEO-FFI*) von Borkenau und Ostendorf (1993) auf jenen 9 Kategorien der geschilderten Untersuchung von Angleitner, Ostendorf und John aufbaut, doch der Untertitel „nach Costa und McCrae“ zeigt, dass es sich nur um eine Übersetzung bzw. Adaptation des amerikanischen Fragebogens handelt. Es wurde auch im Manual der revidierten Auflage NEO-PI-R kein Hinweis gefunden, ob viele oder alle jener 9 Kategorien der deutschen lexikalischen Untersuchung auch in dem amerikanischen NEO und dessen Varianten eine Entsprechung haben. Offensichtlich existieren große Lücken. Es gibt keine Aussagen oder persönlichkeitspsychologische Begründungen, welche dieser Domänen ausgewählt und welche weggelassen wurden, d. h. das Universum der Items (Deskriptoren) ist undefiniert. – In dem Manual zur revidierten Fassung des NEO-PI-R wird behauptet: „Das FFM geht davon aus, dass sich Persönlichkeitseigenschaften mittels Selbst- und/oder Bekanntenbeurteilung in fünf weitgehend unabhängige Dimension ordnen lassen“ (2004, S. 30). Außerdem wird eine genetische Grundlage dieser Beschreibungsweisen vermutet, statt zunächst der Frage nachzugehen, inwieweit solche Faktorisierungen der Alltagsprache aus kognitions- und sozialpsychologischer Sicht viele Schemata und Stereotype repräsentieren, statt uneingeschränkt essentialistisch gedeutet werden zu können. Borkenau und Ostendorf (1993) bevorzugen Begriff der *Sedimentationshypothese*, d. h. sie sind der Überzeugung, dass „alle Aspekte individueller Differenzen, welche bedeutsam, interessant oder nützlich sind oder waren, in die Sprache Eingang gefunden haben; je bedeutender eine solche individuelle Differenz, desto größer die Wahrscheinlichkeit, dass sie ein gesondertes Wort hervorbrachte. Die Sedimentationshypothese impliziert, dass Lexika das Universum aller bedeutenden individuellen Unterschiede abdecken.“ (Borkenau & Ostendorf, 1993, S. 5; ebenso Kubinger, 2009, S. 218–219). – Bereits Ludwig Klages (1948) prägte den schönen Begriff „Sprache als Quell der Seelenkunde“. Tatsächlich zitierten Allport und Odbert (1936), die wohl die erste methodisch zielstrebige lexikalische Analyse durchführten, außer Francis Galton und anderen Vorläufern auch Klages (The science of character, 1926/1932) und dessen Schätzung, dass in der deutschen Sprache

etwa 4000 Wörter sich auf innere Zustände beziehen. Zitiert wurde auch die davon beeinflusste empirische Analyse durch Franziska Baumgarten (1933, S. 81), die jedoch eine psychologisch unzureichende Abgrenzung praktiziert habe – offensichtlich heute noch ein zentraler Einwand. Aber kann wirklich behauptet werden, dass alle für die wissenschaftliche Psychologie wichtigen Deskriptoren von Persönlichkeit, auch aus sozial-, kulturund neuropsychologischer Sicht bereits als „Sedimente" vorhanden sind und nur noch richtig geordnet zu werden brauchen? Hätten auch Sigmund Freud für seine psychoanalytische Forschung, Henry A. Murray (1963) für seine persönlichkeitstheoretisch wichtige Motivationslehre oder Hans Thomae (1968) für seine anspruchsvolle biografisch-längsschnittliche Persönlichkeitsforschung die passenden Wörter bereits in der Alltagsprache finden können?

Welch relativ beschränkter Ausschnitt in dieser lexikalisch-faktorenanalytischen Fragebogenkonstruktion nur erfasst wird, zeigt die Benennung der Skalen an, selbst wenn die jeweiligen Facetten ein breiteres Spektrum bedeuten: Openness for Experience, Agreeableness und Conscientiousness sowie Neuroticism und Extraversion. Bereits die deutsche Liste der Deskriptoren in den genannten neun Kategorien bilden keine Zufallsstichrobe der Alltagssprache, noch sehr viel weniger gilt dies für die ausgewählten Domänen des amerikanischen bzw. des deutschen NEO-Itempool. Jedes Lehrbuch der Persönlichkeitspsychologie steht in einem massiven Widerspruch zu dem globalen Anspruch der FFM-Tests, denn bereits die Register bzw. die Hinweise auf bereits existierende Skalen reichen aus, um die Defizite der FFM-Fragebogen zu charakterisieren: beispielsweise Aggressivität, Autoritarismus, Dogmatismus, Frustrationstoleranz, Repression-Sensitization, Feldabhängigkeit, Kontrollüberzeugungen, Lebenszufriedenheit, Risikobereitschaft, Selbstkonzepte, Machtmotive, Belohnungsaufschub, Zwischenmenschliches Vertrauen. Persönlichkeitsprägende Einstellungen und Wertorientierungen scheinen fast völlig zu fehlen: die weltanschaulichen und religiösen Überzeugungen. Wer könnte überzeugende Grenzen ziehen zu den Handlungsbereitschaften, Kreativität, kognitiven Stilen usw.? Führt die nicht begründete Abgrenzung der FFM-Domäne auch dazu, dass beispielsweise die motivationsbezogenen Persönlichkeitskonzeptionen (*Personality Research Form* PRF; Stumpf et al., 1985) und psychoanalytisch orientierten Eigenschaftsbegriffe (*Der Gießen-Test II*, Beckmann, Brähler & Richter, 1977, 2012) völlig ausgeklammert werden? Bereits die Aufnahme weniger Items zu einigen der genannten Persönlichkeitsdomänen würde die FFM-Faktorenanalysen wesentlich modifizieren, umso mehr wenn noch weitere Domänen berücksichtigt würden: auch überdauernde habituelle Gesundheitssorgen oder die Neigung zu körperlichen und psychischen Beschwerden, auch einzelne Aspekte im Übergang zu psychischen Störungen (nur deskriptiv und erst im Grenzbereich zu einer möglichen Diagnose) oder Aspekte des nicht gerade seltenen devianten Verhaltens. Was meinen jene Autoren mit „Personality"?

Spekulation über Anzahl und psychologische Bedeutung von Persönlichkeitsfaktoren

Jede psychologische Abgrenzung einer Domäne *der Persönlichkeit* von anderen Bereichen der *Person* bleibt eine Fiktion. Andere Konstruktionsstrategien benötigen solche Postulate und Argumente nicht, denn sie beziehen sich entweder auf Kriterien oder deduktiv auf einzelne, bereits mehr oder minder gut bestimmte Persönlichkeitseigenschaften. In der Konstruktion dieser kritisierten Persönlichkeitsfragebogen sind weder das Universum der Deskriptoren von „Persönlichkeit" repräsentiert noch das Universum der Individuen, denn es wurden zumeist nur große, unter verschiedenen Bedingungen gesammelte Datensätze ohne einheitlich durchgeführte und quotierte bevölkerungsrepräsentative Erhebung analysiert. Folglich sind die methodischen Voraussetzungen nicht erfüllt, um repräsentative Ergebnisse postulieren zu können. Kann wirklich behauptet werden, dass die ersten fünf Faktoren solcher problematischen Itemsammlungen die theoretisch und praktisch wichtigsten Persönlichkeitseigenschaften ergeben? Die Entsprechung von „groß" und „wichtig" ist vielleicht in der Ökonomie oder in technischen Bereichen angebracht, aber in der differenziellen Psychologie und Persönlichkeitsforschung?

So bleibt der eigenartige Befund, dass die „Sprache als Quell der Seelenkunde" hier angeblich nicht mehr als fünf gemeinsame Faktoren liefert. Deren geringe Zahl könnte durch die Beschränkung der Domäne verursacht sein oder auf eine sehr hohe und dann psychologisch unzweckmäßige Homogenität der verwendeten in Items in wenigen Clustern hinweisen bzw. auf das Abbruchkriteriums Eigenwert = 1.0. Dieses Kriterium ist sehr fragwürdig und bietet nur eine Entscheidungshilfe unter anderen. Die Autoren des deutschen NEO-PI-R teilen nicht mit, welche psychologischen Inhalte durch die Faktoren 6, 7, 8 usw. angezeigt werden. Da es primär um die Duplikation der Skalen des amerikanischen Vorbilds geht, wird die erhaltene Faktorenstruktur nach der nur selten angewendeten *Prokrustes*-Methode rotiert (Ostendorf & Angleitner, 2004, S. 112), sodass hier zumindest der seltsame Name zu übersetzen ist: Der Riese Prokrustes ist eine Figur der griechischen Sagenwelt; er bot Reisenden ein Bett an, wenn sie jedoch zu groß waren, hackte er ihnen die Füße ab – redensartlich für den Zwang in ein eigentlich nicht passendes Schema.

Auf den kritischen Vergleich mit ähnlichen Publikationen anderer amerikanischer Autoren bzw. Fragebogen kann hier

verzichtet werden, denn die Publikationen über Big 3, Big 4 oder Big 6 bzw. Big 7, HEXACO (Ashton & Lee, 2009) usw. sind zugleich wechselseitig auch Widerlegungen ihres Anspruchs (vgl. Andresen, 2015). Eine definitive Anzahl von Grund-Dimensionen festlegen zu wollen, kann noch viel weniger überzeugen als ähnliche Postulate im Bereich der Intelligenzforschung. Die faktorenanalytische Methodik gibt höchstens Hinweise auf die Varianzanteile der extrahierten Eigenwerte, deren relative Größe von mehreren technischen Entscheidungen des Untersuchers abhängt, u. a. hinsichtlich Kommunalitäten, Extraktionsverfahren, orthogonaler oder schiefwinkliger Rotation (siehe bereits Wittmann 1977; Wittmann & Hampel, 1976), wobei im konkreten Fall jedoch die faktorenanalytischen Befunde fundamental von der Selektion der Items abhängen. Eine sehr hohe innere Konsistenz ist nur mit sehr ähnlichen Items zu erreichen, und diese Redundanz wird die gewünschte Kriterienvalidität entsprechend einschränken und den diagnostischen Nutzen der Skala verringern (Reliabilität-Validität-Zielkonflikt). Cluster sehr ähnlicher Items, bereits sogenannte Dubletten oder Tripletts können aufgrund ihrer hohen Kovarianz die Ergebnisse, nicht nur die innere Konsistenz, sondern auch die Faktorenanalyse hinsichtlich Eigenwerten, relativer Größenordnung und Abfolge massiv beeinflussen. Die Reihe der erhaltenen Eigenwerte gibt keinen generellen Maßstab hinsichtlich der *Anzahl* der *psychologisch interessanten Faktoren,* und die relativ größten Eigenwerte müssen auch nicht – unabhängig von der Fragestellung – die psychologisch wichtigsten sein – nach welchen Maßstäben? Diese grundsätzlichen Einwände betreffen zwar alle faktorenanalytischen Testkonstruktionen, doch hinsichtlich der lexikalisch-induktiven Strategie ist es ein *zentraler* Einwand, falls hier eine kleine und von ihrer Datengrundlage her – genau genommen – psychologisch weitgehend undefinierte Auswahl den Rang der *größten und wichtigsten* Persönlichkeitsfaktoren erhält. 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?

Im Manual zum NEO-PI-R (Ostendorf & Angleitner, 2004) wird nur auf die intensive Übersetzungsarbeit hingewiesen, aber nicht detailliert auf den Inhalt der besonders bedenklichen bzw. „idiomatischen" Items, die ja nur im Kontext der vorherrschenden amerikanischen Lebensbedingungen und Einstellungen zu verstehen wären. Die genaue Protokollierung dieser transkulturellen Übertragung psychologischer Item-Inhalte sollte in solchen Fällen zur wissenschaftlichen Dokumentation im Manual gehören, und diese Erfahrungen wäre in kulturpsychologischer Hinsicht interessant (siehe Andresen & Beauducel, 2008; Becker, 1996, 2000; Paunonen & Ashton, 2001). – Andresen (2015; Andresen & Stemmler, 1982) hat in seiner sehr umfangreichen Übersicht, Methodendiskussion und Reanalyse eine fundierte Auseinandersetzung mit dem „Mythos Big Five" gegeben. Seine eigenen Reanalysen aufgrund einer Sammlung solcher Inventare bzw. Primärfaktoren, allerdings anhand eines nicht repräsentativen Datensatz von nur 340 Personen, führten zur Beschreibung und vergleichenden Diskussion von bis zu 22 Faktoren und modellartigen Konzepten, auch zu eigenen Strukturmodellen, doch ohne Bezug auf externe Kriterien, Validitätshinweise oder Anwendungsstrategien.

Kulturpsychologische Kritik an amerikanischen Persönlichkeitsmodellen FFM

Wenn von Fragebogenautoren die *universelle* Geltung und Gültigkeit der von ihnen so hervorgehobenen Persönlichkeitsfaktoren behauptet wird, verlangt dieser Anspruch, sich mit den kontroversen Voraussetzungen und problematischen Konstruktionsverfahren auseinanderzusetzen. Ein noch zu wenig behandeltes Teilproblem sind auch die kulturpsychologischen Tendenzen bzw. Vorurteile solcher Fragebogen. Der *universalistische* Anspruch (McCrae & Costa, 1997) erinnert an den früheren, gelegentlich fast „imperial" anmutenden Unterton einiger 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 bzw. kritisiert worden (vgl. Hofstede, 2006; Marsella, Dubanoski, Hamada & Morse, 2000; sowie zu den kulturellen Unterschieden der *Social Axioms*, Leung et al., 2002). Da die statistischen Vergleiche nicht selten anhand von Gelegenheitsstichproben, beispielsweise „College Students" (Englisch als Zweitsprache) mit einem starken Selektionsbias, oder anhand problematischer, nur verbaler, aber nicht psychologisch äquivalenter Übersetzungen, unternommen wurden (McCrae, Terracciano & 79 Members of the Personality Profiles of Cultures Project, 2005), ergeben sich fundamentale Zweifel an solchen oberflächlichen inter-kulturellen Projekten. Diese hatten aufgrund von Daten aus 51 Kulturen die universelle Gültigkeit des 5-Faktoren-Konzepts behauptet. Die betreffende Publikation gibt kaum Hinweise auf die Übersetzung und wenig zur Datenerhebung, die durch freiwillig und anonym am Projekt beteiligte Collegestudenten erfolgte. Es sollten möglichst 50 (tatsächlich zwischen 22 und 305) Bekannte erfasst werden, wobei jedem Mitarbeiter eine Quotierung vorgeschrieben wurde nach vier Kategorien College-Studenten (Männer bzw. Frauen) oder Personen über

40 Jahre alt (Männer und Frauen). Die Effekte aufgrund fehlender Repräsentativität werden nicht diskutiert, aber die generelle Gültigkeit der Konzeption behauptet.

Die „großen Fünf" wurden durch van der Linden, Nijenhuis und Bakker (2010) zunächst auf zwei Meta-Faktoren *Stability* und *Plasticity* zurückgeführt und schließlich auf einen einzigen Super-Faktor der „general social effectiveness" reduziert, der als universeller second order *General Factor of Personality* postuliert wird. Um zu belegen, dass die psychologischen Items nicht allein für die „Western educated, industrialized, and rich democracies" gelten, wurden Mitglieder des indigenen Stammes der Tsimane im Amazonasgebiet Boliviens befragt, anhand einer jeweils mündlich erfolgenden Übersetzung aufgrund einer spanischen Übersetzung aus dem Englischen. Wie jene Befragten auf dieses wahrscheinlich nicht allein aus ihrer Lebenssicht seltsame amerikanische Menschenbild reagierten, wird nicht mitgeteilt, jedoch werden Übersetzungsprobleme berichtet, sogar die Eliminierung eines völlig unpassenden Items. Die mageren Befunde der statistischen Analysen werden trotzdem als Beleg für die Universalität dieses Generalfaktors gewertet.

Bereits zwischen den europäischen Ländern bzw. Kulturen scheinen über die sprachlichen Übersetzungsprobleme hinaus einige beträchtliche Unterschiede in der *psychologischen Bedeutung* von Fragebogenitems zu bestehen. Aus diesen, auch den eigenen Erfahrungen ist die Kritik an den simplifizierenden inter-kulturellen Übersetzungen von Persönlichkeitsfragebogen 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 Beispiele psychologisch zu erläutern, empirisch zu begründen und im Detail zu dokumentieren. In der dritten Phase wäre heute nicht bloß nach einer *sprachlich äquivalenten* Übersetzung zu fragen, sondern die *psychologische Äquivalenz* der Items zu untersuchen. Diese Schwierigkeiten sind allgemein unterschätzt worden. Hier sollen Richtlinien vorbeugen, die eventuell auch in Reaktionen auf solche expansiven „weltweiten" Studien zum Nachweis der universellen Gültigkeit des FFM entstanden sind.

Gegen diese Art von Persönlichkeitsforschung lautet die naheliegende Erwiderung aus chinesischer Sicht (Cheung, Fan und Cheung, 2017, p. 105): „Indigenous psychology arose in non-Western countries as a reaction to the dominance of Western psychology models, which are based on values of individualism, rationality, and objectivity. Given their emphasis on universal truths, these Western constructs are implicitly assumed by some to be universally applicable without recognizing possible cultural differences. According to the experience of indigenous psychologists, these presumed universal models may not provide adequate, relevant, or meaningful understanding of human behavior that is contextualized in local cultural contexts. The indigenous psychology movement emphasizes studies of human behavior from the natives' perspective using local cultural concepts and culturally relevant methodologies." – Chinesische Vergleichsstudien konnten beispielsweise zeigen, dass der Faktor „Interpersonelle Bezogenheit (Interpersonal Relatedness)" des chinesischen Fragebogens von keiner Big-Five-Facette erfasst wurde (Cheung, Fan & Cheung, 2017; Cheung, Fan & To, 2009; Cheung, van de Vijver & Leong, 2011). – Konnte denn erwartet werden, dass auch die Aspekte des traditionellen konfuzianischen Menschenbildes mit der Verpflichtung gegenüber Familie und Gesellschaft oder das „bescheidene Verhalten" (Chen et al., 2017) in dem vorherrschenden amerikanischen FFM-Menschenbild repräsentiert sein könnten? – Die chinesische Psychologin Fanny M. Cheung (Cheung et al., 2017; Cheng, Cheung et al., 2014), die durch mehrere Untersuchungen zu dieser kulturkritischen Beurteilung beitrug und inzwischen einen chinesischen Persönlichkeitsfragebogen, den *Chinese Personality Assessment Inventory* (CPAI, Cheung, Fan & To, 2009), entwickelte, erhielt für diese Forschung 2012 den Award for Distinguished Contributions to the International Advancement of Psychology" der *American Psychological Association* (Cheung, 2012): „By adopting a combined emic-etic approach to developing the *Chinese Personality Assessment Inventory*, the first Asian personality measure translated into six other languages, she overcame the ethnocentricism found in both the etic and emic approaches." – Neyer und Asendorpf (2018, S. 405–408) referieren einige der neueren Arbeiten aus der kulturvergleichenden Persönlichkeitsforschung, die in aktuellen Arbeiten anstrebt, den „emischen Ansatz zu verfolgen, der nach kulturspezifischen Persönlichkeitskonstrukten sucht" und „westliche Persönlichkeitskonstrukte nur teilweise bestätigen" konnte. – Seit 2005 wurde auch an der Entwicklung des *South African Personality Inventory* (SAPI) gearbeitet (Meiring, van de Vijver, Rothmann, de Bruin mit einem Team von Studierenden), ebenfalls mit dem Ergebnis wesentlicher kulturpsychologischer Unterschiede (Verfügbar unter: https://sapiproject. co.za/home/index).

Ausblick

In seinem Lehrbuch Psychologische Diagnostik geht Kubinger (2009, S. 218–259) auf Persönlichkeitsfragebogen ein; sein Vorbild ist der Typ des „Big Five-Persönlichkeitsmodells". Er meint, dass auch angesichts einiger neuerer und weniger einheitlicher Ergebnisse gilt: „Nichtsdestotrotz bedeutet das Big Five-Persönlichkeitsmodell zum aktuellen Forschungstand

die Basis allen psychologischen Diagnostizierens im Persönlichkeitsbereich“ (S. 219).

Neyer und Asendorpf (2018, S. 112) sehen den Nutzen von lexikalisch begründeten Faktorenmodellen u. a. für schnelle oberflächliche Persönlichkeitsbeschreibungen, für den Vergleich von Selbst- und Bekanntenbeurteilungen oder für allgemeine Aussagen über die Stabilität von Eigenschaften. Zweitens „können solche Modelle dazu dienen, die inzwischen uferlose Zahl von Persönlichkeitsskalen übersichtlich zu klassifizieren. Die meisten dort erfragten Eigenschaften lassen sich als Unterfaktoren der Faktoren eines lexikalisch abgeleiteten Modells oder der Kombination dieser Faktoren auffassen. … Der Nutzen des lexikalischen Ansatzes für die Persönlichkeitspsychologie ist begrenzt. Die wichtigste und unüberwindliche Grenze liegt darin, dass es sich nur um eine Beschreibung der Ähnlichkeitsstruktur von Eigenschaften handelt, die alltagspsychologisch repräsentiert sind. Was sich nach alltagpsychologischer Wahrnehmung ähnlich sieht, muss sich aber nach wissenschaftlichen Kriterien lange noch nicht ähneln.“ Die Verfasser illustrieren diesen grundsätzlichen Einwand mit einer Parabel, die den Unterschied zwischen der Alchemie aufgrund äußerlicher sensorischer Merkmale von Substanzen und der wissenschaftlich erklärenden Chemie meint. – Der Nutzen von *lexikalisch-induktiv abgeleitete* Konzepte für die Persönlichkeitspsychologie ist begrenzt. Sie können als alltagspsychologische Konzepte eine *deskriptive* Funktion haben (S. 112). – Abgesehen von der Wünschbarkeit oberflächlicher Beschreibungen, ist die behauptete ordnungsstiftende Funktion empirisch unbegründet, da zumindest der NEO-Fragebogen nicht die lexikalische Vielfalt repräsentiert, sondern einen psychologisch undefinierten Ausschnitt und insgesamt nicht einmal näherungsweise die Domäne bzw. den Eigenschaftsraum der Persönlichkeitspsychologie, wie die anderen Kapitel in den Lehrbüchern der Persönlichkeitspsy-chologie demonstrieren. – Anders formuliert handelt es sich um deskriptive Begriffe (oder sprachliche Schemata) die sich auf einige relativ häufi vorkommende Merkmalsmuster in Selbstbeurteilungen beziehen, wobei viele Merkmalsbereiche bzw. Eigenschaften fehlen. Psychologisch sind Aussagen weder über den Funktionszusammenhang noch über die Dynamik und Entwicklung der Persönlichkeit möglich.

Die Inkonsequenz, dass Eysencks *kriterienorientiert* und nicht lexikalisch-induktiv entwickelten Dimensionen E und N übernommen wurden und sozusagen das Rückgrat der FFM bilden, wird jedoch nicht diskutiert. Die Testautoren der FFM scheinen weiterhin außerordentlich überzeugt zu sein, ungeachtet der *Selbst-Widerlegungen* innerhalb dieser Test-Familie: ohne Rücksicht auf die testmethodische Kritik an der psychologisch unzureichend begründeten Datenbasis oder auf die überzeugende kulturpsychologen Kritik an dem zugrundeliegenden „amerikanischen Menschenbild“.

Eysencks Persönlichkeitskonstrukte E und N bilden das herausragende Vorbild und Lehrbeispiel. In der Domäne der Persönlichkeitsfragebogen existieren kaum Skalen, die als robuster, zeitlich stabiler, formal besser reproduzierbar oder inhaltlich-persönlichkeitspsychologisch konsistenter gelten können als Eysencks E und N. Eysencks anfängliche Konzeption und die mehrstufige testmethodische Entwicklung der Fragebogen innerhalb seiner Persönlichkeitstheorie wurden ausführlich von Amelang und Bartussek (1997, S. 324–360) beschrieben. Wissenswert ist, dass nicht interne Analysen oder Faktorenanalyse eines Itempools am Anfang standen, sondern die sog. Kriterien-Rotation hinsichtlich der Diskrimination der zuvor im Maudsley Hospital gebildeten klinisch psychologischen Gruppen. Die Items der ersten Neurotizismus-Skala (Maudsley Medical Questionnaire MMQ, Eysenck, 1958b) wurden auch mit den von Guilford übernommenen Items für den Bereich Extraversion-Introversion zu den beiden Skalen E und N des Maudsley Personality Inventory MPI (Eysenck, 1959) zusammengefasst und mit kleineren Item-Revisionen weiter ausgebaut zum Eysenck Personality Inventory EPI (Eysenck, 1964) und Eysenck Personality Questionnaire EPQ (Eysenck & Eysenck, 1975) und Nachfolgern.

Auch die Analysen des FPI-Datensatzes aufgrund der zweiten und dritten Normierung des FPI-R haben die von Eysenck ausgearbeiteten Konfigurationen von Items bestätigt. Empirisch belegen auch die zahlreichen Befunde der Züricher Längsschnittstudie die herausragende kriterienbezogene Bedeutung. Die markante Abbildung dieser Datenstruktur könnte als der Vergleichsmaßstab für die Überzeugungskraft alternativer testmethodischer Strategien gelten. Dass Eysencks jahrzehntelanges Forschungsprogramm zur neuro- und psychophysiologischen Fundierung beider Dimensionen nicht erfolgreich war, ist ein anderes Thema.

Auffällig bleibt, dass Lehrbücher der Persönlichkeitspsychologie den Postulaten und Kontroversen über „die“ Persönlichkeitsfaktoren so viel Raum geben (z. B. Amelang & Bartussek, 1997; Hagemann et al., 2016; Neyer & Asendorpf, 2018). Andererseits werden sehr viel wichtigere Themen der Konstruktion und Evaluation von Fragebogen sowie der psychologischen Diagnostik oft nur kurz oder nur abstrakt, aber kaum im Hinblick auf die Konsequenzen für neue Assessmentstrategien

behandelt werden: u. a. das erwähnte Reliabilitäts-Validitäts-Dilemma, die Enttäuschungen durch kritische Multitrait-Multimethod-Analysen, die Kriterienvalidierung und deren Symmetrieprinzipien (Brunswik-Linse), der Entscheidungsnutzen, die Generalisierbarkeitstheorie, die multivariate Reliabilitätstheorie oder die multimodale Diagnostik und das innovative Ambulante Assessment.

Funder (2006) sah 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 Funder in den zugrundliegenden und undiskutierten philosophischen Überzeugungen hinsichtlich Individualität versus Gleichheit, Willensfreiheit versus Determinismus, Konsistenz versus Flexibilität. – Im Hinblick auf Funder ist jedoch anzumerken, dass die fast ausschließlich auf Fragebodendaten beruhende Person-Situation-Debatte durch die innovative Methodik des ambulanten Assessment überholt ist. – Wird es vielleicht einmal einen ähnlichen wissenschafts*psychologischen* Rückblick geben hinsichtlich der essentialistisch anmutenden Kontroverse über die *richtige Anzahl* der *wichtigsten* Persönlichkeitsfaktoren in Fragebogen? Verbunden mit der Einsicht, dass durch Faktorenanalysen oder Messmodelle weder über die *psychologisch adäquate* Auswahl von Persönlichkeitskonstrukten entschieden noch die strukturelle Subjektivität der Selbstbeurteilungen aufgehoben werden kann?

Die zunehmend kritische Strömung hinsichtlich der Konstruktion und des Geltungsanspruchs von FFM-Persönlichkeitsfragebogen ist auch außerhalb der Psychologie bemerkt worden wie etwa in dem wissenschaftsjournalistischen Beitrag von Eva Tenzer (2013) zu lesen ist: *Big Five unter Beschuss.*
(Verfügbar unter*:* https://www.wissenschaft.de/geschichte-archaeologie/big-five-unter-beschuss/).
„Seit Jahrzehnten sind Psychologen davon überzeugt, dass alle Menschen auf der ganzen Welt sich durch fünf Persönlichkeitsmerkmale charakterisieren lassen. Neue Studien stellen das in Zweifel. Von Tokio bis Timbuktu, von Frankfurt bis zu den Fidschi-Inseln – überall auf der Welt bestimmen Psychologen die Persönlichkeit eines Menschen anhand von fünf Merkmalen. Diese sogenannten Big Five gelten als universell für Frau und Mann, für Jung und Alt, für Stadt- und Landbewohner. Demnach zeigt sich die Persönlichkeit darin, wie verträglich, gewissenhaft, extravertiert, emotional stabil und offen für Neues jemand ist. Erst 2005 hat eine große Studie in 50 Ländern auf allen Kontinenten mit rund 12 000 Befragten die kulturübergreifende Gültigkeit des Modells belegt. – Doch trifft es wirklich für alle Menschen auf der ganzen Welt zu? Nein, sind amerikanische Forscher inzwischen überzeugt.“ Tenzer geht auf die Kritik aus chinesischer Sicht (CPAI von Cheung et al., 2009) ein und auf die Entwicklung des *South African Personality Inventory* (SAPI von Meiring, van de Vijver, Rothmann, de Bruin, https:// sapiproject.co.za/home/index): „Mitgefühl zählt“. In dieser afrikanischen Gesellschaft sind die sozialen Beziehungen der Menschen fundamental: „Das Ergebnis: Zu den Big Five, die auch in Südafrika relevant sind, kommen die vier landestypischen Merkmale Mitgefühl („soft-heartedness“), Vertrauenswürdigkeit („integrity“), Harmoniebedürfnis („relationship harmony“) und Hilfsbereitschaft („facilitating“). … Die Antworten zeigen, dass das Individuum stark als Teil der Gemeinschaft begriffen wird, also im Kontext seiner sozialen Beziehungen und Situationen.“ Auf der anderen Seite fand das Postulat der fünf Faktoren in den USA und auch in Deutschland eine große populäre Resonanz, und bei einem Blogger im Internet erreicht diese Überzeugung eine fast missionarische Intensität.

6. Bereitschaft zu offener Auskunft und Selbstschilderung

„Der Mensch, der es bemerkt, dass man ihn beobachtet und zu erforschen sucht, wird entweder verlegen (geniert) erscheinen, und da kann er sich nicht zeigen, wie er ist; oder er verstellt sich, und da will er nicht gekannt sein, wie er ist.“ Und es ist zweifelhaft, ob „ein anderes denkendes Subjekt sich unseren Versuchen der Absicht angemessen von uns unterwerfen lässt.“ (Kant 1798/1983, BA X– XII, S. 401; A X– XI, S. 15–16).

Antworttendenzen in Fragebogen: Selbstdarstellung in der Selbstbeurteilung

Das Problem der Antworttendenzen wird immer wieder angesprochen, insbesondere auf dem Gebiet der Personalpsychologie und der forensischen Psychologie, bildet jedoch ein durchgängiges Methodenproblem. Es stellt sich zweifellos auch bei Interviews und bei den meisten anderen psychologisch-diagnostischen Methoden, wird dort aber seltener vertieft. Die Gründe liegen wohl in der weiten Verbreitung der Fragebogen, in den scheinbar durchsichtigeren Verhältnissen und in der zeitweilig verbreiteten Auffassung, diese Tendenz wie einen Fehler kontrollieren zu können. Viele Darstellungen vermitteln

den Eindruck, es handle sich um ein spezifisches Problem der Persönlichkeitsfragebogen, und übersehen, dass methodische Reaktivität, Reaktanz und situationsabhängige Effekte ein fundamentales Problem nahezu jeder psychologischen Methodik (und sogar vieler medizinischer Untersuchungsmethoden) bilden.
Die FPI-Autoren hatten sich bereits zuvor skeptisch hinsichtlich der Anwendung von Fragebogen in der Personalpsychologie ausgesprochen, denn hier liegt es sehr nahe, einen möglichst positiven Eindruck zu machen. Wie dies zu erreichen sei, wird in populären psychologischen Schriften erläutert, z. B. in einem Ratgeber für ein erfolgreiches Auftreten bei Bewerbungen bzw. in einem Assessment-Center. Aus der Sicht der Personalpsychologie geben Hossiep, Paschen und Mühlhaus (2000) – vor dem Hintergrund ihrer Entwicklung des Bochumer Inventars zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) – eine ausführliche Übersicht zum Thema. Für die methodenbewusste Anwendung von Fragebogen sind die Gründe deutlich zu machen, weshalb es keine methodisch befriedigende Lösung gibt, sondern nur pragmatische Regeln für die Anwendung.

Universalität der Antworttendenzen, Definitionsversuche und unzureichende Untersuchungen

Je nach Autor werden unterschiedliche Aufzählungen von verschiedenen formalen sowie inhaltlichen Antworttendenzen gegeben. Die meisten sind schon in den 1950er Jahren bei Cronbach und Guilford zu finden, . die Ja-sage-Tendenz (Akquieszenz), die Tendenz zur unentschiedenen Mitte oder zu den Extrema, die Tendenz einen guten Eindruck zu machen (siehe Amelang & Schmidt-Atzert, 2006; Borkenau & Amelang, 1986; Amelang & Borkenau, 1981, 1982; Buse, 1976, 1980; Dilchert, Ones, Viswesvaran & Deller, 2006; Herzberg, 2011; Krauth, 1995; Lösel, 1995; Meier, 1985; Moosbrugger & Kelava, 2007; Mummendey, 1995, 1999; Pohl, 2004; Richman, Kiesler, Weisband & Drasgow, 1999; Rosch et al., 1984; Schwarz & Sudman, 1992; Seiwald, 2002, 2003; Semin et al., 1981; Vagt & Wendt, 1978; Westmeyer, 1995; Wiggins, 1973). Solche Antworttendenzen wurden aus verschiedenen fachlichen Perspektiven untersucht, u. a. testmethodisch, differenziell-psychologisch, sozial-psychologisch, kognitions-psychologisch. Zu diesem Thema gehören auch schematische Urteilstendenzen, charakteristische Urteilsfehler und systematische Erinnerungstäuschungen. Dagegen fanden die deutlichen, teils systematisch negativ gefärbten Effekte retrospektiver Selbstbeurteilungen von Befinden und Verhalten, trotz reichhaltiger empirischer Nachweise aus dem ambulanten Assessment, bisher nur relativ geringe Aufmerksamkeit. Wie eng solche Tendenzen miteinander konfundiert und mit Persönlichkeitseigenschaften verknüpft sind, zeigt sich etwa bei der Akquieszenz, der sozialen Erwünschtheit oder beim sog. Konsistenzeffekt, ähnlich klingende Aussagen stimmig zu beantworten. Sind nicht fast alle Antworttendenzen zugleich wichtige Charakteristika von Selbstbeschreibungen, also Facetten von Persönlichkeitseigenschaften? Auch deshalb ist es schwierig, die Antworttendenzen im engeren Sinn von anderen unerwünschten Varianzanteilen abzugrenzen: den möglichen Einflüssen kognitiver Stile, semantischer und sprachlicher Schwierigkeiten, den Reihenfolgeeffekten und typischen formalen Fehlerquellen bei Antwortauswahl und Antwortprotokollierung. Die begriffliche Unterscheidung darf nicht darüber hinwegtäuschen, dass eine operationale Definition von Antworttendenzen praktisch schwierig und eine statistische Separierung kaum möglich ist – noch weniger eine hinreichend zuverlässige Absicherung durch sog. Kontrollitems oder Kontrollskalen im Einzelfall. Kaum wird berücksichtigt, dass solche Antworttendenzen untereinander konfundiert sind, kontextspezifisch und personenspezifisch variieren, d. h. von bestimmten Iteminhalten provoziert werden. Folglich käme es darauf an, möglichst durch eine vorbeugende Auswahl die Iteminhalte, Itemund Antwortformate zu optimieren. Das Konzept der multiple-choice Items zur Balanzierung solcher Effekte hat allerdings für Persönlichkeitsfragebogen nicht überzeugen können.
Die Diskussion von Antworttendenzen konzentriert sich meistens auf drei Aspekte: die Ja-sage-Tendenz (Akquieszenz), die Präferenz für mittlere oder extreme Antwortkategorien und die Tendenz zur sozialen Erwünschtheit bzw. zur absichtlichen Verfälschung. Wahrscheinlich spielt hier die zeitweilig verbreitete Meinung mit, gerade diese Effekte könnten wie ein Messfehler kontrolliert oder auspartialisiert werden: durch Paare gegensätzlich gepolter Items (was sprachlich oft kompliziert ist) bzw. durch dichotome Antwortformate oder durch eine „Lügenskala“ (bzw. „Offenheit“). Das größte Interesse fanden die vermuteten Effekte der sozialen Erwünschtheit, doch wäre nur bei einem relativ kleinen Anteil der Items eindeutig zu sagen, welche der Antworten in jedem Fall wirklich „sozial erwünscht“ ist. Es mangelt an überzeugenden Versuchsplänen und klärenden Untersuchungen unter Alltagsbedingungen.
Zu diesem Thema gehören Begriffe wie Selbstkonzept und Selbstwahrnehmung, Selbstbild und Fremdbild, Effekte der sozialen Interaktion und des Feedbacks, Selbstdarstellung und Impression-Management, Eindrucksbildung, Beurteilungsfehler, kognitive und motivationale Verzerrungen, implizite Persönlichkeitstheorien, Beobachtbarkeit von Eigenschaften, subjektive Konzepte der Alltagspsychologie usw. (Forgas, 1999; Herzberg, 2011; Mummendey, 1995; Neyer & Asendorpf, 2018; Paulhus, 1989, 2002). Es handelt sich um eine sehr umfangreiche, z. T. widersprüchliche Literatur. Die meisten

Untersuchungen beruhen auf den Daten von Studierenden oder von anderen Personen, die Fragebogen unter verschiedenen, meist sehr durchsichtigen Instruktionen verschiedener Art ausfüllen sollen, d. h. insgesamt fragwürdigen Projekten mit geringer ökologischer Validität. Die tatsächliche Wirkung solcher Instruktionen und die begrifflichen Unterscheidungen sind gewöhnlich nicht durch unabhängige operationale Definitionen, sondern in zirkulärer Weise wieder durch andere Fragebogendaten belegt.

Innerhalb des sozial erwünschten Verhaltens möchte Paulhus (2002) vier Aspekte unterscheiden: egoistischer Bias, d. h. positive soziale und intellektuelle Qualitäten zu behaupten, und moralistischer Bias, d. h. positive moralische Qualitäten zu behaupten und negative, sozial abweichende Eigenschaften zurückzuweisen. Der egoistische Bias soll mit Extraversion, der moralistische mit Emotionalität (Neurotizismus) assoziiert sein. Auf einer zweiten Ebene werden beide Einstellungen in eine bewusste und eine unbewusste Form aufgeteilt. Dieses begrifflich, nicht empirisch, konstruierte Schema soll mit den vier Skalen des Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1994; 1998) repräsentiert werden. Wie dies ohne den sozialen und inhaltlichen Kontext der realen Testsituation und ohne Bezug auf die Vielfalt individueller und gruppen- und aufgabenspezifischer Wertdisposition möglich ist – und ohne externe Kriterien gültig sein soll – bleibt völlig offen. Ein Anwendungsversuch des BIDR von Allbutt, Ling, Heffernan und Shafiullah (2008) führte nicht über die internen Korrelationen von Fragebogendaten mit anderen Selbstbeurteilungen hinaus. Für das NEO-PI-R Inventar wurde nachträglich eine weitere Skala Positive Presentation Management (PPM) durch Vergleichsuntersuchungen in Analogsituationen mit unterschiedlicher Instruktion, vor allem mit Studierenden in simulierter Bewerbungssituation, entwickelt, um Antwortverzerrungen im beruflichen Bereich bzw. in der Personalpsychologie zu erkunden. Korrelationen zwischen PPM-Werten und den Skalen des BIDR wurden dahingehend interpretiert, dass PPM eher die Tendenz, sich selbst positiv zu sehen, anzeigt und kaum eine explizite Strategie abbildet, einen guten Eindruck machen zu wollen. Die PPM-Skala sei nicht für eine Adjustierung des „faking“ geeignet. Bemerkenswert sind die Hinweise, dass auch beim Beantworten von Listen mit nicht verbalen Items (Figuren, Symbolen usw.), die zur Persönlichkeitsbeschreibung dienen sollen, unter verschiedenen Instruktionen interessante differenzielle Effekte auftreten (Amelang, Schäfer & Yousfi, 2002). Um die Diskussion weiterzuführen, werden hier einige Untersuchungen referiert und die testkonstruktiven Aspekte erörtert. Weiterhin existieren keine Konventionen für den Umgang mit diesem Problem in der psychologischen Praxis.

Empirische Studien

Im Hinblick auf das Freiburger Persönlichkeitsinventar hatten die Untersuchungen von Hampel und Klinkhammer (1978) und Kury (1983a,b) über Verfälschungstendenzen dazu geführt, dass die Testautoren dieses Problem ausführlich diskutierten, vor unkritischer Interpretation warnten und einen expliziten Verfahrensvorschlag machten. Der Einfluss sozialer Erwünschtheit auf Testwerte des FPI wurde von Kury (1983a) bei jugendlichen Straftätern in Analogsituationen untersucht. Die vier Instruktionsbedingungen bezogen sich auf den Zweck der Fragebogenerhebung: Forschungssituation anonym, Forschungssituation nicht-anonym, Fragebogen zu den Vollzugsakten, Fragebogen zu den Vollzugsakten verbunden mit einer Warnung vor Verfälschung der Antworten. In dieser Abstufung fielen die Testwerte der Skala Offenheit sehr signifikant geringer aus. Diese Tendenz war begleitet von einer Zunahme der Varianz. Folglich reagierten die Untersuchungsteilnehmer unterschiedlich auf die Instruktionsbedingungen. Die Möglichkeit zu positiver Selbstdarstellung bedeutet nicht, dass sich tatsächlich jeder so verhält. Ein weiteres instruktives Ergebnis stammt ebenfalls von Kury (2002), der den Einfluss der Datenerhebungsmethode auf kriminologisch-viktimologische Untersuchungen analysierte. Außer dem FPI wurde ein umfangreiches selbstentwickeltes Inventar mit Fragen zur Opferwerdung und zu den Einstellungen hinsichtlich kriminalpolitisch relevanter Themen verwendet. Erwartet wurde, dass in einem Interview eher die Verfälschungstendenz im Sinne einer allgemein möglichst positiven Beschreibung der eigenen Person auftreten wird als bei anonymer schriftlicher Befragung. Zwei Zufallsstichproben wurden verglichen: 1420 Probanden im Alter ab 14 Jahren mit postalisch-schriftlicher Befragung (Rücklauf gleich 49 %) und 542 Probanden mit vollstandardisiertem Interview und FPI-R (Teilnahmebereitschaft gleich 58 %). Die direkt befragte Gruppe äußerte sehr signifikant höhere Lebenszufriedenheit, Soziale Orientierung und Extraversion. In den Skalen Erregbarkeit, Aggressivität, Emotionalität bestand kein Unterschied. In der Skala Offenheit zeigten sich nur eine Tendenz zu geringerer Offenheit sowie signifikante Effekte hinsichtlich Geschlecht und Altersgruppe, wobei die ältere Frauengruppe die niedrigsten Testwerte hatte. Demnach führt das persönliche Interview zu mehr Antworten, die sozial erwünscht sein könnten, als die postalische Erhebung. Der Autor plädierte erneut für eine bessere methodische Absicherung der Ergebnisse der empirischen Kriminologie.

Aus völlig anderer Perspektive kann in der Medizin und der Rehabilitationspsychologie nach Tendenzen und subjektiven

Theorien der Symptomwahrnehmung, nach der Diskrepanz von Beschwerden und objektiven Befunden bei chronisch Kranken oder nach möglichen Konsequenzen einer tendenziell negativen Selbstdarstellung (siehe Franke, 2002) gefragt werden. In diesem Zusammenhang sind u. a. die Begriffe Krankheitsverhalten und Rentenneigung wichtig. Das Thema Krankheitsverhalten hat Myrtek (1998a; Myrtek & Fahrenberg, 1998) aufgrund der Literatur und sehr umfangreicher eigener Untersuchungen beschrieben. Follow-up Ergebnisse in der Rehabilitationspsychologie zeigten substanzielle Zusammenhänge mit Persönlichkeitsfragebogen auf. Franke (2002) befasste sich nur mit dem sog. „faking bad", obwohl auch im klinischen Bereich mit dem Gegenteil umzugehen sein wird, der Dissimulation von Beschwerden, u. a. um eine Hospitalisierung zu vermeiden oder andere unerwünschte Konsequenzen einer psychiatrischen oder somatischen Diagnose. Statt auf die ökologisch valideren deutschen Untersuchungen einzugehen, werden nur amerikanische Arbeiten, nur bei Studierenden mit Instruktionsvariation durchgeführt, zitiert. Das faking-bad Verhalten sei ein komplexer Prozess, dessen Beurteilung umfangreiches Fachwissen erfordere. Als Standard empfiehlt Franke die flexible Nutzung von speziellen Assessmentverfahren und Einzelfall-Beurteilungen im Hinblick auf Bezugsgruppen von Patienten und Gesunden sowie allgemein ein multimodales Vorgehen.

Merten, Friedel, Mehren und Stevens (2007) befassten sich mit den möglichen Antworttendenzen in Persönlichkeitstests innerhalb der nervenärztlichen Begutachtung und stellten fest, dass es im deutschen Sprachraum kaum praktikable Ansätze gebe, während für die neuropsychologische Diagnostik ihres Erachtens durchaus Verfahren zur Überprüfung der Beschwerdenvalidität vorlägen. Die Autoren wollten prüfen, in wie weit sich „negativ verzerrtes Antwortverhalten auf die Validität von Daten der standardisierten Persönlichkeitsdiagnostik auswirkt". Als Maßstab diente ihnen der *Strukturierte Fragebogen Simulierter Symptome* (SFSS). Der SFSS ist die deutsche Version des Fragebogens Structured Inventory of Malingered Symptomatology (SIMS) von Smith und Burger (1997) und wurde von Cima et al. (2003) publiziert. Sie enthält 75 Items, die – abgesehen von einigen Items zur verbalen Intelligenz – vor allem psychopathologische Symptome enthalten, darunter mehrere aus dem Konstruktbereich „Neurotizismus". Für den SFSS wird eine empirische Validität behauptet, weil in Gruppen von Studierenden verschiedene Instruktionsbedingungen, d. h. „ehrlich" zu antworten bzw. zu simulieren, unterschiedliche Testwerte lieferten. Merten et al. (2007) analysierten die Daten von 93 Probanden, die neurologisch-psychiatrisch begutachtet waren, und von denen Testwerte des SFSS, des FPI-R und Ergebnisse des *Word Memory Test* (WMT) vorlagen. Das erhaltene FPI-R-Persönlichkeitsprofil zeigte eine deutliche Abhängigkeit vom Ergebnis der Klassifikation des WMT bzw. SFSS (negativ verzerrtes vs. unauffälliges Antwortverhalten). Die FPI-R-Offenheitsskala korrelierte nicht mit den Variablen SFSS und WMT, die ihrerseits mäßig zusammenhingen. Die Autoren schlossen, dass aufgrund der Offenheitsskala des FPI-R keine Aussage über negative Antwortverzerrungen getroffen werden könne. Sie fordern größere Anstrengungen zur Entwicklung von Fragebogenskalen zur Erkennung solcher Verzerrungen – ohne jedoch gründlicher auf die immanenten Schwierigkeiten und die ökologische Validität einzugehen. Auch andere Beiträge (Kubinger, 2002; Seiwald, 2002, 2003) folgen immer noch dem einfachsten Verfahren, Studierenden (oft aus dem Fach Psychologie) einen Fragebogen mit durchsichtigen Instruktionsvarianten vorzulegen. Wenn Aussagen über die externe Validität fehlen, muss nach dem möglichen Erkenntnisgewinn, der sich auch durch die Wiederholung solcher Studien nicht steigern lässt, gefragt werden. Wenn Kubinger dann schreibt: "It is concluded that psycho-diagnosis based on personality inventory might not be risked" (S. 10), aber nicht auf Effektstärken oder alternative Assessmentstrategien eingeht, bleibt diese Beurteilung sehr fragwürdig.

Antworttendenzen und Antwortformat

Erwähnenswert ist der Versuch, die Antworttendenzen durch neue Antwortformate zu verringern. Für die Vorgabe am Bildschirm wurden anstelle der kategorialen Antwortformate verschiedene Varianten von visuellen Analogskalen verwendet, z. B. eine keilförmige Skala oder der Pfeil eines Schiebereglers. Dadurch sollen eine differenziertere Skalierung erreicht und zugleich Antworttendenzen verringert werden. Für die zweite Behauptung gibt es bisher keine hinreichenden Belege, was angesichts der definitorischen und versuchsplanerischen Schwierigkeiten einsichtig ist. Das Design von Fragebogenitems und Antwortformaten ist ein aktuelles Thema, denn die computer-unterstützte Darbietung erlaubt viele Varianten, auch auf einem hand-held PC bzw. heute Smartphone (siehe u. a. Palmblad & Tiplady, 2004; Kuhlmann, Dantlgraber & Reips, 2017; Schwenkmezger & Hank, 1993). Bereits Richman, Kiesler, Weisband & Drasgow (1999) hatten eine Metaanalyse zahlreicher Untersuchungen durchgeführt, um zu klären, ob die Testdarbietung am Bildschirm oder als Fragebogen (anonym oder im Gegenüber) Unterschiede liefert. Es wurde kein Effekt „Computer versus Fragebogen" gefunden.

Mit dem konventionellen Format des FPI befasste sich Hambros (2002). Sie hatte die Erwartung, dass das dichotome Format (stimmt – stimmt nicht) für die Befragten beanspruchender und unangenehmer sei als eine in vier Kategorien abgestufte Form. Die Untersuchung geht auf Kubingers Behauptung zurück, dass die dichotomen Antworten in

Persönlichkeitsfragebogen angeblich ungeeigneter sind als mehrstufige Formate. Statt einen Fragebogen in zwei Varianten zu geben, versuchte Hambros jedoch die Hypothese durch den Vergleich zwischen zwei verschiedenen Inventaren FPI (zweistufig) und *Trierer Persönlichkeitsfragebogen* (TPF; vierstufig) zu prüfen. In einem 2 × 2 Plan mit 265 Teilnehmern erhielten vier Gruppen nach Zufallsprinzip entweder das FPI-A1 oder den TPF, entweder vor oder nach zwei anstrengenden Leistungstests. Unmittelbar nach den Fragebogen sollten die Teilnehmer drei Eigenschaften des Tests nennen und als positiv oder negativ bewerten sowie in einem Satzergänzungstest weitere Kommentare liefern. Außerdem wurden die im Fragebogen vorgenommenen Korrekturen und die Anzahl fehlender Antworten ausgewertet. Auf die erhaltenen Kommentare hatte die Variation der Reihenfolge keinen Einfluss, signifikant war nur die geringere Anzahl fehlender Antworten im FPI verglichen mit dem TPF. In den Testwerten zeigten sich nur beim FPI zwei Reihenfolgeeffekte: die Testwerte in der Skala zur Aggressivität und Emotionalität waren am Ende der langen Testserie – wahrscheinlich belastungsbedingt – höher. Die erhaltenen freien Beschreibungen unterschieden sich nur in einer von fünf Kategorien: das FPI wurde als *weniger* frustrierend eingestuft. Die Auswertung ergab zahlreiche sprachliche und inhaltliche Kritikpunkte, und es wurde pauschal die Revision einer größeren Anzahl von Items empfohlen. – Bemerkenswert ist jedoch, dass hier die veraltete Version des FPIA1 ausgewählt wurde statt der revidierten Version FPI-R. Weder die vorausgegangenen Studien zu diesem Thema noch die anlässlich der Normierungen gewonnenen, bevölkerungsrepräsentativen Ergebnisse über Verständlichkeit, Zumutbarkeit und allgemeine Akzeptanz des FPI wurden einbezogen. Die Untersuchung zeigte erwartungswidrig eine höhere Akzeptanz des FPI. Das Ergebnis bleibt jedoch unklar, weil in dem Versuchsplan unterschiedliche Fragebogeninhalte und Antwortformate konfundiert sind. Überzeugender könnte ein innovatives cross-over Design mit geeignet ausgewählten Sets von Items und Antwortformaten sein und – wie bereits durch andere Untersucher – eine gründlichere Analyse der semantischen Probleme im individuellen Urteilsprozess.

Die Antworttendenzen beim Ärgerausdruck im deutschen State Trait Anger Expression Inventory (STAXI) (Schwenkmezger et al., 2000) wurden von Gollwitzer, Eid und Juergensen (2005) untersucht. Ein großer Datensatz (*N*=4497 Patienten) ermöglichte Analysen nach einem „mixture distribution item response model". Sie führten zu zwei bzw. drei Klassen von Antwortstilen; diese wurden anhand der verfügbaren Skalen des FPI-R psychologisch als soziale Erwünschtheit und als allgemeine Kategorienpräferenz interpretiert. Eine Summation der Items führe zu falschen Interpretationen, falls die unterschiedlichen Antworttendenzen nicht berücksichtigt werden. – Solche Modellierungen könnten an sich bekannte Antworttendenzen auf neue Weise beschreiben. Unbeantwortet bleiben jedoch die Frage der praktischen Effektstärke und die entscheidende Frage nach der Separierung der Antworttendenzen im Einzelfall. Welche Alternativen werden, abgesehen vom Verzicht auf Persönlichkeitsfragebogen, vorgeschlagen? So bleibt erst noch überzeugender zu belegen, welchen Nutzen solche Modellierungen für die innovative Konstruktion *neuer* Persönlichkeitsfragebogen haben könnten.

Objektive Persönlichkeitstests statt Fragebogen?

Einen Ausweg aus dem Methodenproblem der Antworttendenzen sieht Kubinger (2003c) in den *Objektiven Persönlichkeitstests* im Sinne von R. B. Cattell (1957; Cattell & Warburton, 1967). Wenn hier eine neue experimentalpsychologische Verhaltensdiagnostik gesehen wird, weil damit seines Erachtens nicht „Persönlichkeit", sondern Verhalten gemessen würde, entspricht diese optimistische Einschätzung (Ortner, Proyer & Kubinger, 2006) nicht den Schwierigkeiten und den früheren Erfahrungen und kritischen Evaluation in diesem Bereich (vgl. Amelang & Schmidt-Atzert, 2006; Fahrenberg, 1964; Häcker, 1982; Häcker, Schwenkmezger & Utz, 1979). Cattells Ideen zur Entwicklung objektiver Persönlichkeitstests waren anregend, doch die behauptete *Undurchschaubarkeit* muss differenziert und relativiert werden, und die *Auswertungsobjektivität* darf nicht mit der *Durchführungsobjektivität* verwechselt werden. Außerdem waren für Cattell, und noch mehr für Eysenck (1967), die *apparativen* Messmethoden und die physiologischen Parameter wesentlich. Cattell wollte, wie er sagte, diesen Bereich weiter ausbauen, fand jedoch damals keinen Psychophysiologen für sein Labor. Typische objektive Tests waren für Eysenck und Cattell: negatives Nachbild der Exner-Spirale, Reminiszenzphänomene, d. h. die individuelle Leistungssteigerung am Pursuit Rotor nach einer Pause, Fingergeschicklichkeit, Tapping im persönlichen Tempo u. a. Sättigungsphänomene, visuelle, akustische, kinästhetische und andere Schwellen, Flimmerverschmelzung, Wahrnehmungstäuschungen, Lidschlag-Konditionierung, physiologische Aktivierungsparameter bei belastenden Aufgaben.

Bei genauer Überprüfung im Labor zeigten sich bei jedem dieser apparativen Tests wichtige Einflussgrößen: der erhebliche Einfluss der technischen Gerätespezifikationen, die Auswahl und exakte Messung von Stimulusparametern und Hintergrund, die durch Instruktionselemente und methodenbedingte Reaktivität bedingte Varianz, die gravierenden Effekte von – nachträglich explorierten – individuellen Strategiewechseln und impliziten Hypothesen der Versuchspersonen usw. Diese

labormethodisch entscheidenden Parameter waren in der Objektiv-Analytischen Testbatterie (O-A-Testbatterie) nicht beschrieben und nicht einmal für die wichtige Lidschlagkonditionierung im Londoner Labor so prägnant definiert, dass eine genaue Replikation möglich gewesen wäre. Ein besonderes Problem bilden die in Zeitreihenstudien gut zu erkennenden, massiven Eingewöhnungs- und Übungseffekte bei vielen dieser Tests. Die eigenen Ergebnisse waren so inkonsistent und die Bemühungen um Kontrolle dieser verschiedenen „Methodenfaktoren" zur testmethodischen Standardisierung so frustrierend, dass nur wenige Hinweise aus diesen umfangreichen Untersuchungen publiziert wurden (Fahrenberg, Kuhn, Kulick & Myrtek, 1977; Fahrenberg, Myrtek, Kulick & Frommelt, 1977; Fahrenberg, 1977; Fahrenberg, 1987a; Myrtek, 1984). Einige der Papier-und-Bleistift-Tests in Computerprogramme umzusetzen ist naheliegend, aber nicht neu. Die Vergleichbarkeit von Untersuchungsergebnissen könnte durch beliebige Software-Varianten noch stärker beeinträchtigt sein als früher; es fehlen in dieser Hinsicht internationale Standards wie sie auch für die medizinische Labordiagnostik erst zum Teil erreicht sind. Bei all diesen Tests, auch neuropsychologischen Untersuchungsmethoden, einschließlich von MRT-Untersuchungen, kann die „Sozialpsychologie" des Experimentierens nicht übersehen werden, wie im Vergleich zu einem physikalischen Festkörper-Experiment unmittelbar einsichtig ist. Psychologische Experimente verlangen in jedem Fall sprachliche Kommunikation, Instruktionsverständnis und Compliance, d. h. Bereitschaft, die Ruhebedingung und die gestellte Aufgabe bzw. den intendierten emotionalen Zustand, so zu realisieren, wie vom Untersucher (ohne wirkliche Kontrollchancen) gewünscht.
Das Buch von Ortner et al. (2006) geht jedoch auf die kritischen Themen wie Durchführungsobjektivität, Durchschaubarkeit, Ausklammerung der wichtigen apparativen Tests, viel zu wenig ein; die notwendigen Prüfungen der konvergenten und diskriminanten Validität sowie die persönlichkeitspsychologischen Kriterienkorrelationen dieser Tests kommen höchstens am Rande vor. Dass viele der unerwünschten Methodenfaktoren als *averbale* Antworttendenzen zu verstehen sind, wird nicht erörtert. Ein zu diskutierender Aspekt ist schließlich, dass die gewünschte Undurchschaubarkeit solcher Persönlichkeitstests implizit als Absicht zur Täuschung der Probanden verstanden werden könnte.

Schlussfolgerungen

Skeptiker könnten fragen, ob die Debatte über Antworttendenzen bzw. Verfälschungstendenzen seit dem erinnerungswürdigen Urteil von Fiske und Pearson (1970) über den *chaotischen Forschungszustand* dieses Bereichs wesentliche Fortschritte gemacht hat. Noch nicht einmal eine befriedigende Taxonomie der vielfältigen Antworttendenzen existiert, und dieser Mangel hängt mit den schwierigen operationalen Definitionen und natürlich mit der Eigenart von Selbstbeurteilungen und den grundsätzlich objektivierbaren Selbstauskünften zusammen. – An dieser Stelle kann nur versucht werden, die Forschungslage in einigen allgemeineren Thesen und kritischen Anmerkungen zusammenzufassen und einen pragmatischen Verfahrensvorschlag für das FPI zu geben.

Mangel an realitätsnahen Untersuchungen zur Sozialen Erwünschtheit

Der Vorrang des Themas Soziale Erwünschtheit mit der schematischen alltagspsychologischen Unterstellung verbreiterter Fälschungsabsichten scheint plausibel zu sein. Mögliche Vorteile einer positiven Selbstdarstellung sind offensichtlich und dieser generelle Einwand wird vor allem dort vorgebracht, wo aktuelle Absichten anzunehmen sind, sich in einem positiven Sinn darzustellen: Bei der Personalauswahl und bei der psychologischen Beurteilung von Angeklagten bzw. Inhaftierten. Im klinischen Bereich könnte zwar dieses Motiv fehlen, doch werden auch Patienten Anlässe und Absichten haben, sich in einer bestimmten Weise darzustellen. – Ein eindimensionales Verständnis dieser Selbstauskünfte mit der Absicht, eine Antworttendenz *Soziale Erwünschtheit* methodisch zu isolieren, unterschätzt das psychologische Bedingungsmuster bei weitem. Bereits unter den Studierenden der Psychologie, die an solchen fiktiven Erprobungen in Pseudo-Situationen teilnehmen, wird es ein breites Meinungsspektrum geben, welche Antworten einen guten Eindruck machen könnten. Andere Erwartungen werden sich bei einer realen Bewerbung (je nach Tätigkeit) oder im Strafvollzug oder im Verlauf einer medizinischen oder psychologischen Rehabilitation manifestieren können. Den Untersuchungen fehlt in der Regel eine situationsgerechte Definition dieses Konzepts von „erwünscht". *Soziale Erwünschtheit* ist ein *multi-referentielles Konstrukt*, das einen Zusammenhang herstellt: zwischen Testsituation, Testmotivation, Persönlichkeitseigenschaften, stilistischen Merkmalen, verbaler Intelligenz, individuellen Erwartungen und Motiven, Anpassungsbereitschaft, Nutzenabwägungen, Furcht vor Nachteilen oder Furcht vor Entdeckung sowie die individuellen und populationsbezogen-generellen Wertdispositionen wie Ehrlichkeit und Offenheit. Zweifellos können in Persönlichkeitsfragebogen und Interviews Antworttendenzen manifest werden. Hinweise für Einflüsse der Testsituation und der Erwartungen auf die Antworten liegen aus vielen empirischen Untersuchungen vor. Wichtig sind Untersuchungspläne, die eine näherungsweise realistische Variation der Testbedingungen in Analog-Situationen herstellen – was heute aus

forschungsethischen Gründen noch schwieriger ist als vor 30 Jahren. Deswegen sind die älteren quasi-naturalistischen Ergebnisse überzeugender als die simplen und redundanten Befragungen von Studierenden in Lehrveranstaltungen. Ohne den realen Kontext und ohne die erlebte Relevanz der Testsituation bleiben die meisten neueren Ergebnisse höchst artifiziell, d. h. ohne ökologische Validität.

Inhaltliche Prägnanz und Effektstärke der Sozialen Erwünschtheit?

Untersuchungsergebnisse aufgrund variierter Instruktion stimmen darin überein, dass die gemeinten Effekte *Sozialer Erwünschtheit* zwar häufig auftreten, jedoch z. B. in einem Persönlichkeitsinventar wie dem FPI, kein prägnantes Muster generieren oder zuverlässig reproduzieren. Die meisten Untersucher berichten jedoch nur die Signifikanz der Ergebnisse und keine Effektstärken. Die Tabellen von Kury (1983a,b) sprechen für eine mittlere bis große Effektstärke. Außerdem war die Varianz unter der Ernstbedingung tendenziell höher, d. h. die unterschiedliche Bereitschaft, solchen Instruktionen bzw. einem „Situationsdruck" zu folgen, könnte hier eine größere Rolle spielen. Auch weil nur wenige – näherungsweise – naturalistische Analogstudien vorliegen, ist die statistische *Effektstärke* in pragmatischer Hinsicht kaum zu beurteilen. In den typischen Untersuchungen erreichen sie eventuell nur eine mittlere Größenordnung. Das schließt nicht aus, dass im Einzelnen beträchtliche Einflüsse der aktuellen Selbstdarstellung existieren und entsprechende Risiken der diagnostischen Urteilsbildung.

Defizite der Operationalisierung

Während der vergangenen Jahrzehnte scheint konzeptionell und empirisch nur wenig Neues hinzugekommen zu sein. Die in der Fachliteratur der psychologischen Methodenlehre deutliche Einengung auf das Thema *Soziale Erwünschtheit* und der Mangel an einem breiteren methodologischen Horizont waren hinderlich. Die externe bzw. ökologische Validität der konventionellen Analog-Situationen und Instruktionsvarianten wurde weitgehend vernachlässigt. Auf der anderen Seite sind Simulation, Aggravation, Dissimulation alte Stichworte der psychiatrischen Begutachtungspraxis, und die häufigen Diskrepanzen zwischen körperlichen Beschwerden und objektiven Befunden ein wichtiges Thema der Psychologie des Krankheitsverhaltens. Insbesondere in kognitionspsychologischen Beiträgen zum Thema Urteilstendenzen fehlt häufig ein breiterer Horizont im Hinblick auf die differenziell-psychologische und testmethodische Forschung, d. h. die multi-referenziellen Konstruktionen, die interaktionistische Sichtweise (Person-Situation-Debatte), die Forschung zur methodenbedingten Reaktivität und Reaktanz sowie empirische Untersuchungen zur Compliance.

Aussichtslosigkeit einer Fragebogen-immanenten Kontrolle der Selbstbeurteilung

Wenn Aussagen in einem Fragebogen durch andere Selbstbeurteilungen in *demselben* oder in einem anderen Fragebogen erfasst und korrigiert werden sollen, läuft dieses Verfahren auf ein zirkuläres Unternehmen hinaus. Außerdem ist psychologisch kaum abzugrenzen, welche Antworttendenzen *nicht* zu einem bestimmten Eigenschaftskonstrukt gehören, denn solche Tendenzen, u. a. die Ja-Sage-Tendenz, die Neigung zu extremen oder zu mittleren Antwortkategorien, der Wunsch nach sozialer Geltung können auch Merkmale einzelner Persönlichkeitseigenschaften sein. Das MMPI enthielt drei zusätzliche formale Skalen: die Anzahl unentschiedener Antworten, die Anzahl von nicht eingeräumten, aber eigentlich sehr verbreiteten allgemeinen Schwächen (Lügenskala) und die Anzahl sehr seltener (psychopathologischer) Antworten. Demgegenüber sind in einigen der neueren Offenheitsskalen Items zusammengestellt, die beispielsweise beim FPI-R nahezu von 50 Prozent der Bevölkerung bejaht werden. Die Kombination solcher Items soll einen Hinweis auf die individuelle Ausprägung eines sehr häufigen oder eines sehr seltenen, atypischen Antwortverhaltens geben. Einige Autoren scheinen noch an der früher zeitweilig vertretenen Auffassung festzuhalten, die Antworttendenz des „faking good" sei eine Art Fehlerkomponente, die das Ergebnis kontaminiert, und die durch eine geeignete statistische Operation, aufgrund eines „Lügenwertes", entfernt (auspartialisiert) werden könne. Allerdings wäre für solche „Korrekturen" ein Testwert der „Offenheit" ungeeignet, weil solche Skalen, um eine individuelle Auspartialisierung zu ermöglichen, wesentlich länger sein müssten. Demgegenüber legen die Forschungserfahrungen eine psychologisch differenziertere Konzeption nahe. Der Versuch, bewusste und unbewusste Anteile (vgl. Paulhus, 2002) zu trennen, oder der Wunsch nach einer objektiven, quasi-experimentalpsychologischen Isolierung und statistischen Korrektur überfordern die Fragebogenmethodik grundsätzlich. Zu fragen, wie eine Person „eigentlich" sei, erinnert an eine frühere, essentialistische Auffassung von einer unveränderlichen Persönlichkeit und entspricht nicht dem Verständnis von Persönlichkeitseigenschaften als Dispositionsprädikaten mit zeit- und situationsabhängiger Variabilität. In diesen allgemeinen methodologischen Bezugsrahmen sind die speziellen Annahmen über Erwünschtheit, Täuschung bzw. vorsätzliche Verzerrung, mangelnde Compliance, Offenheit und Privatheit einzuordnen und zu relativieren.

Die Selbstdarstellung ist ein integraler Bestandteil der geäußerten Selbstbeurteilungen. Wer Selbstbeurteilungen in der

psychologischen Diagnostik und Prognostik verwendet, muss auch die strukturelle Subjektivität dieser Berichte hinnehmen. Einzelne Facetten der *Selbstdarstellung innerhalb der Selbstbeurteilung* eines Menschen abgrenzen zu wollen, scheint nur heuristisch und interpretativ möglich zu sein, sofern nicht objektivierbare Variablen wie Lebenslaufdaten, Verhaltensbeobachtungen oder – zum Vergleich mit der Tendenz zu körperlichen Beschwerden – medizinische Befunde einbezogen werden können. In der sozialpsychologischen Einstellungsforschung z. T. vorkommende Objektivierungsversuche wie psychophysiologische Messungen oder das Bogus-Pipeline-Verfahren mit der Täuschung über einen angeblichen Lügendetektor sind aus unterschiedlichen berufsethischen oder methodischen Gründen völlig ungeeignet (siehe Mummendey & Bolten, 1981).

Insgesamt fällt der hohe Spezialisierungsgrad dieser Debatte auf. Sie bleibt oft auf die soziale Erwünschtheit der Antworten beschränkt und geht kaum auf die schwierige Definition dieses Konstrukts oder auf die verwandten Methodenprobleme ein. Als Frage nach der Verzerrung bzw. Verfälschung erscheint das Thema hauptsächlich in den Bereichen der Personalpsychologie, d. h. bei Eignungsuntersuchungen, und im Zusammenhang von forensischen Begutachtungen und Kriminologie. Zweifellos können sich auch in allen anderen Bereichen psychologischer Diagnostik spezielle Antworttendenzen, methodenbedingte Reaktivität und Reaktanz auswirken, auch im klinisch-psychologischen oder pädagogischen Bereich – und dies natürlich nicht nur in Persönlichkeitsfragebogen, sondern in den meisten Datenquellen. Realitätsferne Untersuchungen im üblichen Format werden jedoch kaum noch etwas beitragen können.

Mögliche Strategien und das Fehlen von Konventionen

Weder die Autoren testkritischer Artikel noch die Lehrbücher vermitteln im Hinblick auf die sehr verbreitete Fragebogenmethodik ein deutliches Interesse an geeigneten Verfahrensvorschlägen oder Konventionen für die Praxis. Wie können aus dem Wissen über Erwartungshaltungen und Antworttendenzen und über die immanente Selbstdarstellung in Selbstbeurteilungen praktische Regeln für konvergente Strategien der Interpretation diagnostischer Befunde abgeleitet werden? Die Richtlinien zur Qualitätssicherung (DIN 33420: *Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*; vgl. Diagnostik- und Testkuratorium 2018) verlangen ja, dass die Ergebnisse der Untersuchung so wenig wie möglich durch den Kandidaten selbst verfälscht werden können. Wie dies verhindert oder erkannt werden soll, wird jedoch nicht weiter diskutiert oder zu Vorschlägen für Konventionen verdichtet.

Die neueren Beiträge, so lässt sich zusammenfassend sagen, bekräftigen die Auffassung der FPI-Autoren und die wiederholt formulierten *Empfehlungen*, die als Annäherung an dieses Methodenproblem gemeint sind: Erstens wurde das nicht unproblematische erste FPI-R-Item „Ich habe die Anleitung gelesen und bin bereit, jeden Satz offen zu beantworten" beibehalten und zweitens die FPI-R-*Skala Offenheit*. Das Item 1 ist auch deswegen problematisch, weil offene Antworten verlangt werden, ohne dass die Testteilnehmer bereits im Einzelnen wissen können, was gefragt werden wird. Wenn dann einzelne Antworten eher defensiv ausfallen, ist das eine plausible Strategie, die zumindest in bestimmten Testsituationen durchaus der Lebenserfahrung entsprechen kann.

Die *Skala Offenheit* repräsentiert ein komplexes Persönlichkeitsmerkmal und ist keine „Korrekturskala" und kein direktes Maß der Antworttendenz im Sinne der sozialen Erwünschtheit. Extrem niedrige oder hohe Werte sprechen für ein – auch insgesamt – auffälliges Antwortverhalten und sind ein warnender Hinweis, wenn möglich, zusätzliche Informationen einzuholen oder vorsichtshalber auf die diagnostische Auswertung zu verzichten. Für die deutsche *Personality Research Form* (PRF) gaben Stumpf et al. (1985) eine entsprechende Empfehlung. Grundsätzlich sind Testwerte aus Persönlichkeitsfragebogen – wie in den allermeisten anderen psychologischen Verfahren auch – von der Testsituation, der Testmotivation, der Wiederholung und anderen Bedingungen beeinflusst und legen deshalb eine möglichst gründliche Gewichtung und Interpretation im Einzelfall nahe – es sei denn, dass nur ein *erstes* Screening einer größeren Personenzahl beabsichtigt ist. – Alle diese Beschränkungen bewusst zu halten, bleibt wichtig, um zur vorsichtigen Interpretation zu motivieren. Weder Verharmlosung noch formale Kontrolle können aus diesem Dilemma herausführen. 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? In Kapitel 4.1 des FPI-R-Manuals (2020) wurde eine praktische Regel formuliert und auf die berufsethischen Aspekte hingewiesen. „Die Testautoren empfehlen, Testergebnisse bei geringer Ausprägung der Skala FPI-R 10 Offenheit (Stanine 1, 2, oder 3) vorsichtig und möglichst erst nach Einholen zusätzlicher Informationen über die Testmotivation des Probanden zu interpretieren. Auch die Verneinung oder Nicht-Beantwortung des Items Nr. 1 müssen solche Bedenken auslösen."

7. Prinzipien der Assessmenttheorie

Assessment in der Differentiellen Psychologie entspricht dem Begriff von Diagnostik, meint jedoch allgemeiner die Erfassung von psychologischen Merkmalen nach bestimmten methodischen Prinzipien zu einem praktischen Zweck, welcher eine rationale Entscheidung verlangt. *Assessmentstrategien* legen in einem Datenerhebungsplan fest, welches Konstrukt mit welchem Untersuchungs- und Auswertungs-Konzept erfasst werden soll – so wurde in der Einleitung dieses Kapitels definiert. Grundlagen und wichtige Prinzipien der Assessmenttheorie haben u. a. Cattell, Cronbach, Campbell und Fiske, Brunswik und Wittmann entwickelt. Vor allem Cattell hat für die differenzielle Psychologie und Persönlichkeitsforschung vorbildlich das Denken in multivariaten Zusammenhängen propagiert. Für die kritische Anwendung und die Weiterentwicklung von Persönlichkeitsfragebogen sind die folgenden Themen bzw. Prinzipien besonders hervorzuheben, sodass hier einige Erläuterungen angebracht sind zu: Multitrait-Multimethod, Multimodale Diagnostik, Generalisierbarkeitstheorie, Brunswiks Linsenmodell und die Symmetrie von Prädiktoren und Kriterien, multivariate Reliabilitätstheorie (und entsprechende Aggregationstechniken), Entscheidungsnutzen des Assessments, alltagsnahe Evaluation und ökologische Validität.

Allgemeine Prinzipien der Operationalisierung

Grundsätzliche Fragen zur operationalen Definition von psychologischen Begriffen betreffen: allgemeine Kategorien, Adäquatheit von Definitionen, Konstruktdefinitionen und Operationalisierungsfehler, multiple Beschreibungen, Datentheorie, Allgemeine Prinzipien der Operationalisierungen, Mehrdeutigkeit, und generell die Gesichtspunkte zur Adäquatheit von Beschreibungen (siehe die Übersicht, Fahrenberg, 2013, S. 487–504). Das hohe Anspruchsniveau und der Schwierigkeitsgrad psychologischer Methodik folgen aus der Einsicht, welche Perspektiven sich hier verbinden müssen:

- Überlegungen zur Adäquatheit der expliziten Definition des theoretischen Konstrukts,
- Fachwissen über mögliche Operationen und typische Operationalisierungsprobleme,
- Bereitschaft zur Aufgabe sehr facettenreicher (polythetischer) zugunsten einfacherer Konstrukte,
- konsequente Methodenentwicklung nach Prinzipien des Assessmentheorie.

Die Prinzipien der Assessmentheorie sind in die deutsche Fachliteratur erst relativ selten *systematisch* aufgenommen und entwickelt worden; einige neure Lehrbücher enthalten Kapitel oder Hinweise, jedoch kaum realistische bzw. relevante Beispiele empirischer Forschung. – Im Buchtitel kommt, trotz des wegweisenden Werks *Behavioral Assessment* (Haynes & Wilson, 1979), „Psychologisches Assessment“ nicht vor, höchstens Assessment in der Personalpsychologie, d. h. für das Assessment von Führungskräften (Assessmentcenter), oder für Assessment in der Rehabilitation. Generell werden weiterhin die herkömmlichen Begriffe *Psychologische Diagnostik* und *Testkonstruktion* bevorzugt. – Außerdem gibt es im Buchhandel von deutschen Autoren mehr als ein Dutzend Lehrbücher über *Statistik für Psychologen* bzw. *Quantitative Methoden*, aber heute kein einziges Lehrbuch *Psychologische Interpretation.* So stellt sich die Frage: müssen *statistische* Ergebnisse *psychologisch* nicht weiter interpretiert werden oder geht das in der Forschung und in der Berufspraxis so einfach, dass Regeln, Prinzipien und methodenkritische Kontrollen im Studium nicht vermittelt und gelernt zu werden brauchen? Einige Publikationen zur Gutachtenmethodik weisen auf die notwendige Integration von Befunden hin, ohne jedoch die prinzipiellen Grundlagen psychologischer Interpretation zu geben.

Multitrait-Multimethod

Angesichts der von Selbstbeurteilungen und Selbstauskünften ausgehenden und deshalb unsicheren Persönlichkeitsfragebogen muss gerade für solche Testergebnisse möglichst nach Bestätigungen durch andere Informationsquellen gesucht werden. – Die grundsätzliche Forderung ist in dem Prinzip der *multiplen Operationalisierung* enthalten. Campbell und Fiske (1959) entwickelten die Idee der Multitrait-Multimethod Matrix (MTMM-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,* also nicht allein mit Persönlichkeitsfragebogen, wurden nur relativ 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 scheint eher eine Ausnahme zu sein, Divergenzen bzw. unerwartet niedrige Korrelationen sind häufig. Deshalb hat insbesondere Fiske (1971, 1978, 1981, 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, 1985, 1987; vgl. Eid & Diener, 2005).

Definitionen von MTMM-Plänen und deren Auswertung mit Korrelationskoeffizienten oder konfirmatorischen Faktorenanalysen schildern u. a. Schermelleh-Engel 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. Solche Koeffizienten müssten eventuell noch abgewertet werden, denn die von Eltern, Freunden und Bekannten gegebenen Einstufungen stützen sich wahrscheinlich teilweise auch auf die Äußerungen und Selbstbeurteilungen der betreffenden Person in alltäglichen Mitteilungen. – Was bedeutet die 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, 2009; Eid, Nussbeck & Lischetzke, 2006; Ostendorf, Angleitner & Ruch, 1986), aber es mangelt in der Persönlichkeitsdiagnostik immer noch an überzeugenden empirischen Studien (Hamilton, 1971; Krüger, Rowold, Borgmann, Staufenbiel & Heinitz, 2011).

Wie würden die MTMM-Befunde erst aussehen bei einem eigentlich multimodalen Ansatz, der auch *unabhängige* Datenquellen, Verhaltensmessungen, nicht-reaktive Maße u. a. einschlösse? Schon Cattell (1950, 1957) hatte darauf gedrängt, strikter zwischen den Datenquellen Behavior *Rating* und Behavior *Measurement* zu unterscheiden. Er engagierte sich in seinem anspruchsvollen Vorhaben, die Entsprechungen von Persönlichkeitsfaktoren in den Bereichen der Verhaltensbeurteilungen, der Fragebogen und der objektiven Tests zu untersuchen, ohne jedoch überzeugende Konvergenzen erhaltenen Faktoren beschreiben zu können. – Der Begriff „Verhalten“ ist irreführend, wenn nicht kategorial zwischen dem intersubjektiv beobachtbaren, manifesten, aktuellen, auch symbolischen, aber averbalen Verhalten (wie in der Human-Ethologie oder Primatenforschung) und der verbalen Kommunikation über das individuelle Verhalten unterschieden wird. Auch die verbreitete Unterscheidung von Selbst-Einstufungen und Fremd-Einstufungen kann irreführend sein, weil zumindest die von Freunden und Bekannten gegebenen Einstufungen von der verbalen Kommunikation über Selbstbeurteilungen beeinflusst sein werden wie bei den wechselseitigen Einstufungen von Partnern (siehe Kapitel 5.3). 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; Spinath, 1999). Diese Abgrenzung ist wesentlich, um in Validitätsstudien anhand von Fremdeinstufungen das Risiko zirkulärer Schlüsse zu erkennen und vielleicht zu kontrollieren. Inwieweit sind die gelegentlich als befriedigend angesehenen Korrelationen zwischen Fragebogenergebnissen und Verhaltenseinstufungen durch konventionelle und stereotype Anteile der impliziten Persönlichkeitsauffassungen, durch populäre Alltagstheorien, Attributionen usw. bedingt? Wie ist eine Konfundierung der verbalen Aussagen, die einerseits als standardisierte Selbstbeschreibung im Fragebogen, andererseits als informelle Selbstbeschreibung in der sozialen Interaktion mit dem Einstufer bzw. Bekannten gegeben werden, zu vermeiden? Die Effekte solcher Methodenvarianz sind noch zu wenig untersucht, obwohl die meist unbefriedigenden Ergebnisse der früheren Multitrait-Multimethod-Analysen in der Persönlichkeitsforschung solche Differenzierungen nahelegen (siehe Amelang & Zielinski, 1994; Endler & Hunt, 1969; Fahrenberg, 1987b; Kenrick & Funder, 1988; McCrae, 1982; Ostendorf, Angleitner & Ruch 1986).

Viele der persönlichkeitstheoretischen Konstrukte sind komplex und facettenreich, polythetisch, sodass gravierende Operationalisierungsfehler vorkommen könnten. Höhere begriffliche Prägnanz ist nur durch Aufgliederung in Subkonstrukte zu erreichen und durch entsprechende explizite Konventionen, welche der Subkonstrukte vorrangig verwendet werden. Die Idee der *multiplen Operationalisierung* ist einleuchtend. Sie wurde sogar in den Bereich der qualitativen Methodik, d. h. in die Methodenlehre des interpretativen Paradigmas, 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 die Prädiktoren und das Kriterium unsymmetrisch geplant sind (siehe Brunswiks „Linsenmodell“).

Multimodale Diagnostik

Eine 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 Cattell (1957, 1973) hat ein sehr umfangreiches Forschungsprogramm zur Inventarisierung von Faktoren unternommen, und ein wichtiges Prinzip war dabei der *multimodale* Ansatz: Lebenslaufdaten, Selbstbeurteilungen (Einstufungen, standardisierte Fragebogen), Verhaltensbeurteilungen, Verhaltensbeobachtungen, wirkliche 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 seitdem auch keine systematische Fortsetzung. Dagegen hat Eysenck keine systematischen multimodalen Analysen durchgeführt, trotz der für seine Theorie wesentlichen psychophysiologischen Korrelate der Persönlichkeitsfaktoren E und N.

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 (Themenheft der *Diagnostica*, Fahrenberg, 1987b). Baumann und Stieglitz (2008) veröffentlichten 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 (2008) 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 auch hier viele der sogenannten Fremdbeurteilungen in mehr oder minder hohem Ausmaß auf die Selbstbeurteilungen und Selbstauskünfte der Patienten stützen, also methodisch konfundiert sind. Auch das AMDP-System zur Dokumentation psychiatrischer Befunde (Arbeitsgemeinschaft für Methodik und Dokumentation, 2018) 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).

Gerade für die Klinische Psychologie sind diese Ergebnisse bitter. Zwar gibt es speziell in diesem Bereich relativ viele parallele Informationen und Aufgaben der differenziellen Diagnostik, Prognose, Katamnese, Verlaufskontrolle und retrospektiven Therapieevaluation, doch anscheinend zu wenig Kooperation und Standardisierung. 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.

Bezogen auf Angststörungen gibt es in der klinisch-psychologischen Forschung und Evaluation von Angsttherapien seit langem das anspruchsvolle *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 jedoch 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 Messung geeigneter physiologischer Parameter selbstverständlich sein, insbesondere von kontinuierlicher Messung kardiovaskulärer Indikatoren der Beanspruchung (Herzfrequenz, Blutdruck) und Atmung (Pneumogramm), außerdem die Messung der Bewegungsaktivität, eventuell ergänzt durch zusätzliche, jedoch nur diskontinuierlich in Speichelproben nachträglich zu bestimmende Laborwerte der endokrinen Aktivität (Hormonwerte). – 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 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 Selbstbeurteilungen 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 nicht oder viel zu kurz ein (Hoyer, Beauducel & Franke, 2002; Hoyer & Helbig, 2005; Kubinger, 2009). – 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 Untersuchung angebracht. Erst solche Prozessanalysen könnten die offenen Fragen der differenziellen Indikation und Therapieevaluation beantworten. Therapieverläufe mit zunehmender bzw. hoher Kopplung (Konkordanz) von Funktionssystemen könnten im Vergleich zu diskordanten Prozessen effektiver und nachhaltiger sein (Wilhelm & Fahrenberg, 2018).
Die genannten Lehrbücher der Testmethodik vermitteln den Eindruck, dass im Konzept der multimodalen Diagnostik, falls der Begriff überhaupt vorkommt, eher ein abstraktes Problem gesehen wird statt auch die praktisch-diagnostischen Konsequenzen aufgrund der Reviews von Baumann und Stieglitz (2008) zu erörtern – trotz der großen Tragweite dieser Ergebnisse und der anschließenden Kritik bzw. der Vorschläge, 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 (2008).
Amelang und Schmidt-Atzert (2006) haben den Eindruck, dass institutionalisierte Diagnostik meist uni-modal und individuelle Diagnostik meist multimodal ist. Bei mäßiger Konkordanz von Daten aus verschiedenen Quellen gebe es Möglichkeiten der Verbesserung: Aggregation über Messzeitpunkte, über Kriteriumsbereiche 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." 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.

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 Generalisierbarkeitstheorie 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. (1972) 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, die für die Medikation und die Lebenserwartung entscheidend ist? Demgegenüber mangelt es in der Testpsychologie an solchen Generalisierbarkeitsstudien.

Symmetrieforderung hinsichtlich Indikatoren und Kriterien

In seinem „Linsen-Modell“ veranschaulichte Brunswik (1956) wie ein repräsentativer Versuchsplan aussehen sollte, und diese Prinzipien sind auch auf die Kriterienvalidierung (externe Validierung) von psychologischen Tests zu übertragen. Brunswik fordert 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) hat das Konzept von vier Datenboxen (Prädiktoren, experimentelles Treatment, nicht-experimentelles Treatment und Kriterien) in Anlehnung an Brunswik (1956) und Cattell (1957, 1966) entwickelt, um die notwendigen Präzisierungen von Assessmentstrategien und Validitäts- und Reliabilitätsaspekten zu erreichen (siehe auch Wittmann, 1988, 2002, 2009, 2012; Beauducel et al., 2005; Wittmann & Klumb, 2006; Wittmann, Nübling & Schmidt, 2002; Wittmann & Schmidt, 1983). Aggregiert werden kann über Zeitpunkte (Messwiederholungen), über Situationen (Settings, Untersuchungsbedingungen), über Items (Konstrukt-Facetten, Verhaltensweisen) und andere Dimensionen der Datenbox, sodass 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 Extraversion 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, sodass hier erweiterte, symmetrische Aggregationen notwendig sind. Das Verfahren kann pragmatisch kriterienorientiert (Indexmessung) oder theoretisch konstruktorientiert sein. Wittmann fordert, der Planung adäquater Validierungsuntersuchungen im Vergleich zu den oft überwertig diskutierten Reliabilitätsberechnungen mehr Gewicht zu geben. – Brunswiks Prinzipien von Repräsentativität und Symmetrie und Cronbachs Reliabilitätstheorie führten Wittmann zu den Prinzipien seiner „multivariaten Reliabilitätstheorie“. Diese innovative Konzeption ist für die Validierung psychologischer Tests (siehe oben zur Aggregation) und generell in der Evaluationsforschung wichtig. Während die traditionelle Reliabilitätstheorie in den heutigen Lehrbüchern oft zu ausführlich behandelt wird, fehlt meist eine Diskussion der *Generalisierbarkeitstheorie*, und von der multivariaten Reliabilitätstheorie Wittmanns wird höchstens eine der ersten Arbeiten, aber kaum die neuere Entwicklung zitiert (Wittmann, 2009, 2012).

Aggregation

Fragebogenitems haben typische Inhalte: Beschreibungen eigener Verhaltensweisen und Reaktionen anderer, eigene Zuschreibungen von Eigenschaften, Wünschen und Interessen, Einstellungen und Überzeugungen, außerdem biografische Fakten u. a. (Angleitner, 1976; Angleitner & Wiggins, 1986; Lösel, 1995). Die Itemformulierung und unscharfe Quantoren verlangen bei den Befragten unterschiedlich komplexe kognitive Prozesse. 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; Paunonen, 1984; ausführlicher Schweizer, 1986, 1990). Formal kann zwischen der Aggregation im Raum von Personen, Situationen, Variablen, Wiederholungen, Variabilitäten und Konsistenzen differenziert und dementsprechend zwischen speziellen persönlichkeitstheoretischen Konstrukten bzw. Komponenten der Kovarianz unterschieden werden (siehe Stemmler, 1996, 2001; siehe auch Wittmanns generalisierte Reliabilitätstheorie). Auf Aggregation basiert auch das Spearman-Brown-Prinzip der Reliabilitäts-Steigerung durch Verlängerung des Tests, d. h. Hinzufügen relativ homogener Items. Die Diskussion über Aggregation wurde u. a. durch „multiple act criteria“ im Sinne von Fishbein und Ajzen (1974) und durch die Mischel-Epstein-Kontroverse (Epstein, 1977, 1980; Mischel, 1968) über die angebliche Validitätsgrenze bei r=.30 angeregt, denn diese Kontroverse muss unter dem Gesichtspunkt der speziellen Datenaggregation präzisiert werden. Das Linsenmodell bzw. die multivariate Reliabilitätstheorie enthalten die Forderung, auf der Seite der Prädiktoren und der Kriterien ein vergleichbares Abstraktionsniveau anzustreben, d. h. die inhaltliche Breite (Facetten) und andere Spezifikationen zeitlicher, situativer und anderer Art anzupassen. Die Symmetrie dieser Auswahl bzw. die erreichte Aggregation verbessert die Chancen der Kriterienvalidierung von Testergebnissen. Diese Konzeption ist beispielsweise auch auf das Assessment in der Neuropsychologie übertragbar (Peper, 2018).

Diagnostische Urteilsbildung

Das diagnostische Gutachten (u. a. Amelang & Schmidt-Atzert, 2006; Fisseni, 2004, Westhoff, 2004, Westhoff & Kluck, 2003) 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 (Handlungen und Motive) und der Beobachtung des manifesten Verhaltens 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. *mit einem aktuellen Lebensbezug*, verwendet. Als Anwendungsmöglichkeiten beschrieb er hauptsächlich 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 differenziellen 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 vermutlich als veraltet geltenden „diagnostischen Psychologie“ (Heiß, 1982) 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. (Fahrenberg, 2002). 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 oder didaktische Beispiele. 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.

Entscheidungsnutzen und mögliche Schadensfunktion

Das bekannte Schema der richtigen und falschen Diagnosen (richtig bzw. falsch Positive und richtig bzw. falsch Negative) ist oft dargestellt, jedoch wird selten erläutert, weshalb neben der Nutzenfunktion auch die *Schadensfunktion* solcher

professionellen Entscheidungen wichtig ist. 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).

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 herausragenden und breiten Forschungserfahrung in Medizin und Psychologie einige Beispiele, wo ein sehr viel höheres Risiko akzeptiert wird, weil die Alternativen sehr unerwünscht sind. Auch ein Verfahren mit niedriger Reliabilität könnte 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).

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 im Vergleich zur klinisch-diagnostischen Interpretation von MMPI-Profilen im Vergleich zur „richtigen" psychiatrischen Diagnose – ein Evaluationsverfahren, das heute auf *beiden* Seiten zu revidieren ist (Engel, 2019). 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 idiografischer Auffassungen (Fisseni, 2004), 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.

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 und in einem Anwendungskontext (und dessen berufsethischen Regeln) 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 Persönlichkeitseigenschaften, Motiven und

Einstellungen getroffen werden? Sind nicht auf ihrer Ebene die Prinzipien der externen und „ökologisch“ überzeugenden Validierung sowie die Interpretation und die wissenschaftlichen Kontrollen ebenso ernsthaft zu diskutieren wie die Eigenschaftskonstrukte, die Persönlichkeitstheorien oder die „Messmodelle“ auf ihrer Ebene?

Ambulantes Assessment

Ambulantes Assessment bedeutet Datenerhebung im Alltag der Untersuchten mit mobilen digitalen Systemen. Diese Protokolle enthalten *aktuelle und kontextbezogene* Daten, die in dem angegebenen Setting mit einer genauen Zeitangabe verankert sind (electronic momentary assessment, real time data capture). In diesem Sinn hatte die 2008 gegründete Gesellschaft ihren Namen erhalten: *Society for Ambulatory Assessment – explaining behavior in context.* (Verfügbar unter*:* http://ambulatory-assessment.org/).Während das *Ambulante Monitoring* in der Medizin vorwiegend der Diagnostik und der Überwachung von Risikopatienten, u. a. bei Herz-Kreislauf-Erkrankungen, dient, sind die Aufgaben des *Ambulanten Assessment* in der Psychologie vielseitiger. Auch in der Arbeitspsychologie und Klinischen Psychologie gibt es Überwachungsaufgaben, z. B. an riskanten Arbeitsplätzen oder als Selbst-Monitoring bei bestimmten chronischen Gesundheitsstörungen oder Verhaltensproblemen. Für viele andere Fragestellungen sind Verhaltens- und Erlebnisdaten aus dem Alltag wesentlich. Ambulantes Assessment ist der umfassende Begriff für die Datenerfassung im Alltag. Dagegen beziehen sich andere Begriffe wie *Ecological Momentary Assessment (EMA), Experience Sampling Method (ESM), Time Sampling Diary (TSD)* oder *electronic diary* in der Regel nur auf die Methoden, aktuelle Selbstbeurteilungen sowie Auskünfte über Tätigkeiten und soziale Situationen zu erheben, ohne behaviorale, physiologische oder ambiente Messungen durchzuführen. Während der vergangenen Jahre hat international die Zahl von Publikationen in diesem Bereich der Psychologie und der Medizin stark zugenommen. Allein in PsycInfo hat während der letzten Jahre die Anzahl von Publikationen mit dem Begriffsfeld „ambulatory assessment“, „ecological momentary assessment“, „experience sampling method“ (im Abstract) auf ca. 500 im Jahr zugenommen. Deutet der offensichtliche Mangel an methodisch fortgeschrittenen und kriterienbezogenen Validierungsstudien für Persönlichkeitsfragebogen vielleicht an, dass diese Papier-und-Bleistift-Methoden im Vergleich zur alltagsnahmen Forschung stark an wissenschaftlichem Interesse verloren haben? Oder wird diese Herausforderung noch nicht gesehen?

Die Datenerhebung erfolgt heute meist mit Smartphones bzw. ähnlichen Systemen oder Rekordern mit spezieller Software, um Daten über Ereignisse und Verhaltensweisen, Befinden und andere Selbstbeurteilungen, Angaben über erlebte Situationen und objektive Settingmerkmale zu gewinnen. Darüber hinaus sind mit miniaturisierten Rekordern Bewegungsaktivität und Bewegungsmuster (Sitzen, Stehen, Treppensteigen usw.), physiologische Funktionen (kardiovaskuläre, respiratorische u. a. Funktionen), akustische Signale (Sprache, Spracheingaben, Geräusche) sowie Umgebungsbedingungen (Helligkeit, Geräusche, Temperatur, relative Feuchte) zu registrieren. Falls diese Daten on-line mit Recorder-Analyzer-Systemen ausgewertet werden, sind Rückmeldungen und andere interaktive Methoden möglich (Myrtek, 2004; Myrtek, Foerster & Brügner, 2001). Der Vorzug der computer-gestützten Protokollierung im ambulanten Assessment liegt gerade in dem Zugang zu dem momentanen Befinden, Verhalten, Situationen usw. Aktuelle Aussagen werden weitaus eher die individuelle Befindlichkeit und die Gewohnheiten repräsentieren als die retrospektiven, subjektiv auf unbekannte Weise aggregierten, summarischen Antworten im Fragebogen bzw. Notizen im Tagebuch. Im Allgemeinen werden sie durch ihre größere Verhaltensnähe gültiger und überzeugender sein als retrospektive und vom Kontext abgelöste Fragebogen-Antworten (siehe auch Baumann, Thiele & Laireiter, 2003; Ebner-Priemer, 2006; Fahrenberg, 1994b, 2010; Fahrenberg, Myrtek, Pawlik & Perrez, 2007; Fahrenberg, Leonhart & Foerster, 2002; Fahrenberg & Myrtek, 1996, 2001, 2005; Hektner & Csikszentmihalyi, 2002; Mehl & Conner, 2012; Pawlik und Buse, 1982, 1996, 2008; Wilhelm & Perrez, 2008).

Das ambulante Assessment kehrt auf eine innovative Weise zur Tradition des in den 1970er Jahren propagierten *Behavioral Assessment* zurück und ermöglicht eine psychologische bzw. psychophysiologisch orientierte Verhaltensanalyse im alltäglichen Leben. Die konsequente multimodale Diagnostik und Prozessforschung können ein differenziertes Bild vom aktuellen Befinden, von Verhaltensunterschieden und eventuellen Verhaltensstörungen oder registrierbaren Symptomen geben und Fehlschlüsse vermeiden helfen. Neben den Grundlagenstudien in verschiedenen Bereichen sind z. B. die realistische multimodale Erfassung von Arbeitsbelastungen (Myrtek et al., 2001) und die Diagnostik und Therapiekontrolle bei Patienten mit einer Panikstörung (Wilhelm & Fahrenberg, 2018) zu nennen. Wesentlich sind die einander ergänzenden Datenebenen: subjektives Befinden und Beschwerden, aktuelle Situationen und Tätigkeiten, berichtetes Verhalten und partiell auch Verhaltensmessungen (Aktivität und Bewegungsmuster) sowie gegebenenfalls physiologische Parameter und mit geeigneten Sensoren „ambiente“ Parameter der Umwelt, d. h. Helligkeit, Temperatur, Feuchte Sprechaktivität, Sprache und Geräuschpegel.

Die oft festgestellten Widersprüche dieser Datenebenen müssen von kritischen Untersuchern „ertragen“ und interpretativ bzw. durch weitere Untersuchungen und theoretische Differenzierungen bewältigt werden. Die konsequente multimodale Diagnostik und Prozessforschung ermöglicht ein viel differenziertes Bild von Verhaltensunterschieden und Verhaltensstörungen sowie deren subjektiver Repräsentation und kann Fehlschlüsse verhindern helfen. Unterstützung für die Methodik des ambulanten Assessment kommt auch aus einem anderen wichtigen Anwendungsbereich. Die amerikanische Food and Drug Administration FDA hat in einer Richtlinie empfohlen, dass die Befunde zum *patient reported outcome* in der Erprobungsphase von Pharmaka sich künftig nicht mehr auf konventionelle, retrospektiv ausgefüllte Fragebogen, sondern auf *aktuell* erhobene Daten stützen sollten (U. S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, 2006). In Deutschland werden künftig auch für die Evaluation von Psychotherapie Nachweise für die breite Wirksamkeit unter Alltagsbedingungen nahegelegt bzw. gefordert (Wissenschaftlicher Beirat Psychotherapie WBP, 2007; Nübling, 2008).

Externe Validierung durch Datenerfassung im Alltag (ambulantes Assessment)

Zwei Untersuchungsansätze sind erwähnenswert, weil das ambulante Assessment hier die gewöhnlich verborgenen Diskrepanzen zwischen geäußerten Selbstbeurteilungen und objektiven Messungen demonstriert hat. Das Ausmaß der körperlichen Bewegungsaktivität ist ein wichtiger Aspekt, u. a. in der Evaluation von Fitness-Programmen, in der Erziehung übergewichtiger Kinder und in der Mobilisierung von Senioren. Bei der Validierung von Fragebogen bzw. Skalen, welche die körperliche Aktivität erfassen sollen, ergab sich in der Regel eine beträchtliche subjektive Überschätzung der tatsächlichen Aktivität im Vergleich zur objektiven Aktivitäts- und Bewegungs- Messung. Es besteht nur eine geringe bis höchstens mittlere gemeinsame Varianz der Methoden (siehe das Review von Bussmann, Ebner-Priemer & Fahrenberg, 2009). Der Vergleich zwischen Episoden subjektiv erlebter Emotionen und den entsprechenden Segmenten physiologischer Registrierungen (mit Kontrollfragen in randomisiertem Versuchsplan) zeigte, dass beide Komponenten, der *erlebte* Aspekt und der *physiologische* Aspekt einer alltäglichen „Emotion“, keinen bzw. keinen substantiellen Zusammenhang haben (Myrtek, 2004). Eine vielversprechende Methodik ist die stichprobenweise oder kontinuierliche Aufzeichnung von Sprechaktivität und Umweltgeräuschen, falls die Teilnehmer dazu bereit sind. Auf diese Weise können nach einer Adaptationsphase alltagsnahe Daten, u. a. über die soziale Umgebung, soziale Interaktionen, Gewohnheiten, Hinweise z. B. auf depressive Stimmungsänderungen, und deren individuelle alltagspsychologische Interpretation gewonnen werden (Mehl & Holleran, 2007; Vazire & Mehl, 2008). Insbesondere wenn ein Fragebogen nach der Strategie des act-frequency Ansatzes konzipiert ist (siehe Hodapp et al., 2004), würde die Methodik des ambulanten Assessment einen wichtigen Zugang zur ökologischen Validität ermöglichen.

Der Retrospektionseffekt

Der Retrospektionseffekt wird durch den Vergleich der rückblickenden Einstufungen mit den momentanen bzw. den über einen Zeitabschnitt gemittelten Selbstbeurteilungen festgestellt. Retrospektive Einstufungen von Befinden, Beanspruchung (Stress), Beschwerden, Schmerzen usw. stimmen mit den aktuellen Einstufungen nicht gut überein. Methodenstudien haben gezeigt, dass bei Fragebogen, insbesondere bei tagebuchähnlich wiederholter Anwendung, substantielle Verzerrungen auftreten (vgl. Baumann, Thiele & Laireiter, 2003; Lucas & Baird, 2005). Erinnerungstäuschungen können eine systematische Verzerrung mit problematischen Konsequenzen verursachen. Die Compliance ist sehr eingeschränkt, d. h. ein hoher Prozentsatz der Fragebogen bzw. Skalen in Tagebuchform wird erst nachträglich ausgefüllt. Inzwischen existieren zahlreiche Untersuchungen und Reviews über solche Retrospektionseffekte (u. a. Gorin & Stone, 2001; Pohl, 2004; Schwarz, 2007) und zum *negativen* Retrospektionseffekt (Käppler, Becker & Fahrenberg, 1993; Fahrenberg, Bolkenius et al., 2002, Fahrenberg, Leonhart et al., 2002). Da weder der Anteil verspäteter Einträge noch das Ausmaß der retrospektiven Verzerrungen kontrolliert werden können, bestehen grundsätzliche Zweifel hinsichtlich aller Tageslaufund Längsschnitt-Untersuchungen mit konventioneller Papier- und Bleistift-Methode. Solche Diskrepanzen zwischen momentanen und rückblickenden Einstufungen liefern ein starkes Argument für die digitalen Systeme zur Datenerhebung.

Ambulantes Assessment und Validierung von Persönlichkeitsfragebogen und Stimmungsskalen

Die Nicht-Äquivalenz von Fragebogendaten und Felddaten wurde in den Pionierarbeiten von Pawlik und Buse (1992, 1996; Buse & Pawlik, 1984, 1994) gezeigt. In ihren „Felduntersuchungen zur transsituativen Konsistenz individueller Unterschiede“ hatten sie Persönlichkeits-Situations-Wechselwirkungen nicht an den üblichen retrooder prospektiven Fragebogendaten, sondern erstmals durch ambulantes Assessment geprüft. Als Ergebnis ist festzuhalten, dass die in Fragebogendaten

häufig festgestellte Situation-Person-Wechselwirkung für Persönlichkeit-Setting-Interaktionen in ambulanten Assessmentdaten nicht zu belegen war. Damit ist die gesamte, auf *Fragebogendaten* basierende Person-Situation-Interaktion- Kontroverse überholt. Entsprechende innovative Strategien wurden von Perrez, Schoebi und Wilhelm (2000) in der Stressforschung und von Baumann, Thiele und Laireiter (2003) in Untersuchungen zur sozialen Unterstützung entwickelt (zusammenfassend siehe Fahrenberg, Leonhart et al., 2002; Fahrenberg, Myrtek, Pawlik & Perrez; 2007; Pawlik & Buse, 1982, 1992; Wilhelm & Perrez, 2008).

Weiterhin existiert anscheinend keine Untersuchung, in der die Methodik des ambulanten Monitoring und Assessment *gezielt* zur *Validierung* von Persönlichkeitstests eingesetzt wurde. Bisher handelt es sich um Nebenergebnisse anderer Arbeiten. Wegen der innovativen Bedeutung dieser Methodik für die externe Validierung psychologischer Tests werden hier nach bereits referierten Freiburger Untersuchungsergebnissen nur wenige ausgewählte Hinweise auf die inzwischen sehr zahlreichen Untersuchungen gegeben.

Auch Schlafstörungen sind mit der Methodik des ambulanten Monitoring zu entdecken, falls die Bewegungsaktivität und das Elektrokardiogramm (EKG) fortlaufend registriert werden. Myrtek (2002) untersuchte 223 Herz-Kreislauf-Patienten und 89 Rheuma-Patienten während eines stationären Heilverfahrens sowie zwei Gruppen zu je 50 Studentinnen. Die objektive Schlafqualität wurde aufgrund der Registrierungen als gestört oder als normal eingestuft. Personen mit gestörtem Schlaf hatten im Vergleich zu solchen mit normalem Schlaf in allen vier untersuchten Gruppen sehr signifikant oder signifikant höhere Testwerte in der Skala Emotionalität des FPI-R, jedoch nicht in den Skalen Extraversion und Gesundheitssorgen. – Im Rahmen von zwei Forschungsprojekten der Arbeitsgruppe in Fribourg (Schweiz) berichteten insgesamt mehr als 550 Eltern und Kinder aus 173 Familien mithilfe des computer-unterstützten Familien-Selbst- Monitoring-Systems (FASEM-C) an sieben aufeinanderfolgenden Tagen sechsmal täglich ihr momentanes Befinden und gaben Auskunft über aktuelle körperliche Beschwerden (Perrez, Schoebi & Wilhelm, 2000; Perrez et al., 2005). Als Nebenergebnis wurden für 186 Eltern des ersten Projekts Korrelationen zwischen den intraindividuellen Mittelwerten und Standardabweichungen der verwendeten Items über alle Zeitpunkte (Wilhelm, 2004, S. 130, dort Tabelle 8.1) und den drei FPI-R-Skalen Lebenszufriedenheit, Emotionalität und Offenheit berechnet. Die Befunde stimmen weitgehend mit den zuvor referierten Ergebnissen überein: Der Testwert Emotionalität (und vergleichbar Lebenszufriedenheit) korreliert mit den entsprechenden Befindensitems in der Größenordnung zwischen r=.30 und .50, nicht jedoch mit den intraindividuellen Standardabweichungen der Items, die einen Index der täglichen Befindensvariabilität bilden. Welcher Zusammenhang zwischen individuellen Symptomberichten und der Persönlichkeitseigenschaft Emotionalität besteht, wurde von Michel (2006) auf der Basis der verfügbaren Daten aus beiden Projekten in Fribourg untersucht (335 Eltern und 213 Kinder [11–19 Jahre alt] aus 169 Familien). Die statistische multi-level-Analyse ergab keinen Haupteffekt für die Skala Emotionalität des FPI-R, aber eine Interaktion zwischen Emotionalität und Berichtszeitpunkt in kurvilinearer Form: bei durchschnittlicher Ausprägung der Eigenschaft eine verstärkte Symptomwahrnehmung am Morgen und am Abend relativ zur Tagesmitte. Personen mit relativ hoher Emotionalität zeichneten sich durch weitgehend konstant bleibende Symptomberichte aus. – Der Vorzug der computer-gestützten Protokollierung im ambulanten Assessment liegt gerade in dem Zugang zum momentanen Befinden und zu den Auskünften über Tätigkeiten, Situationen usw. Aktuelle Aussagen werden weitaus eher die individuelle Befindlichkeit und die Gewohnheiten repräsentieren als die retrospektiven, subjektiv auf unbekannte Weise aggregierten, summarischen Urteile im Fragebogen bzw. Tagebuch. Die computer-unterstützten Selbsteinstufungen sind in viel größerer Dichte der Informationen verfügbar, sie werden mit wesentlich höherer Compliance und technischer Zuverlässigkeit gewonnen als in einem Fragebogen, und diese Auskünfte sind kontextbezogene Daten. Gerade unter dem Aspekt der ökologischen Validität handelt es sich, dort wo es praktisch möglich ist, um die Methode der Wahl.

Inzwischen expandiert diese Forschungsrichtung in viele Richtungen und Fragestellungen (siehe die Übersichten, Ebner-Priemer, 2006; Fahrenberg & Myrtek, 1996, 2001; Mehl & Conner, 2012). Hier werden – außer Psychophysiologie und Bewegungsaktivität, nur einige Studien zu einem Spektrum von Themen der Persönlichkeitsforschung und Klinischen Psychologie genannt (Bussmann, Ebner-Priemer & Fahrenberg, 2009; Ebner-Priemer, Koudela, Mutz & Kanning, 2013; Ebner-Priemer & Trull, 2009; Fleeson & Noftle, 2012; Raselli & Broderick, 2007; Himmelstein, Woods & Wright, 2019; Hofmans, De Clercq, Kuppens, Verbeke & Widiger, 2019; Raugh, Chapman, Bartolomeo, Gonzalez & Strauss, 2019; Volkers et al., 2002; Wright et al., 2019). Die Persönlichkeitsfragebogen werden weiterhin nützlich sein, doch ist abzusehen, dass sich einige Gebiete der Grundlagenforschung neu orientieren werden: „Lab and/or Field? Measuring personality processes and their social consequences" (Wrzus & Mehl, 2015).

8. Kritik an der Dominanz der Fragebogenmethodik

In der psychologischen Methodik dominieren heute – abgesehen vom Bereich der Intelligenz- und Leistungstests einschließlich der Schultests – ganz allgemein Fragebogen in Form standardisierter *Skalen* oder mehrere Skalen umfassende *Inventare*. Dies gilt für die Klinische und die Gesundheitspsychologie, die Personalpsychologie und Arbeits-/ Organisationspsychologie ebenso wie für die Differentielle Psychologie oder die Sozialpsychologie (vgl. die Publikationen in den Fachzeitschriften und die Berichte über neue Verfahren).

Dies verwundert umso mehr, als es seit Jahrzehnten bekannt ist, dass solche Fragebogendaten nur sehr eingeschränkt verhaltensgültig sind, denn sie geben primär die subjektiven (mentalen) Repräsentation des individuellen Erlebens und Verhaltens wieder: Was Personen meinen, wie sie sich in der Vergangenheit gefühlt und verhalten haben oder wie sie sich in der Zukunft fühlen und verhalten werden. Deutlicher müsste von introspektiven Selbstbeurteilungen des individuellen Befindens und Verhaltens gesprochen werden. Diese strukturelle Subjektivität der Daten und die testkonstruktiven Konsequenzen zu betonen, ist unüblich – vielleicht wegen der zu deutlichen Abweichung von der (natur-)wissenschaftlichen Konzeption vieler Psychologen, zumindest im Hauptstrom der Allgemeinen Psychologie an den Universitäten (vgl. Rammsayer & Troche, 2005; Sänger & Schäfer, 2017).

In diesem Dilemma befinden sich die Anwender von Persönlichkeitsfragebogen: Die Fragebogen sind unentbehrlich, denn sie vermitteln wesentliche Informationen über die Selbstbeurteilungen einer Person, außerdem Selbstbeurteilungen und auch Selbstauskünfte, die grundsätzlich objektivierbar wären. Auf der anderen Seite sind die begrenzten und z. T. widersprüchlichen Resultate im Hinblick auf objektivierbare Kriterien seit langem bekannt; sie scheinen allerdings in manchen Anwendungssituationen unterschätzt oder ausgeklammert zu werden. Selbstbeurteilungen und verbale Auskünfte über eigenes Verhalten (berichtetes Verhalten) sind eben keine *Verhaltensdaten*. Durch Fragebögen und Interviews wird dennoch versucht, auch Informationen zu erhalten, die weit über die subjektive Repräsentation von Erleben, Verhalten und Einstellungen hinausgehen. Die Dominanz der Fragebogen als *Ersatzmethode*, wenn es in der praktischen Diagnostik nicht allein auf die durchaus wichtige Selbstbeurteilung, sondern wesentlich auch auf das manifeste Verhalten ankommt, ist umso merkwürdiger, als seit langem überwältigend zahlreiche und konvergente Befunde existieren, dass retrospektive Auskünfte unzuverlässig sind. In dieser Hinsicht stimmen experimentelle Arbeiten zu recall- und hindsight-Effekten, autobiografische Studien sowie alltagsnahe Untersuchungen mit computer-unterstützten Tagebüchern überein (Literatur siehe u. a. Gorin & Stone, 2001; Fahrenberg, Myrtek, Pawlik & Perrez, 2007; Käppler, Brügner & Fahrenberg, 2001; Pohl, 2004). Die Pionierarbeiten von Pawlik und Buse (1982, 1992, 1996, 2008) wurden zuvor erwähnt.

Wenn heute auf vielen Gebieten nahezu ausschließlich Fragebogen eingesetzt werden, reduziert sich die Psychologie nach Baumeister, Vohs und Funder (2007, S. 396) auf „Psychology as the science of self-report and finger movements“ (d. h. beim Ankreuzen von Items). Ohne Bezug auf die seit den 1980er Jahren verfügbare und auch von Psychologen genutzte, inzwischen weit entwickelte Methodik des Ambulanten Assessment üben Baumeister et al. Kritik an der der Dominanz der Fragebogenmethodik in der gegenwärtigen Psychologie: „Psychology calls itself the science of behavior, and the American Psychological Association's current 'Decade of Behavior' was intended to increase awareness and appreciation of this aspect of the science. Yet some psychological subdisciplines have never directly studied behavior, and studies on behavior are dwindling rapidly in other subdisciplines. We discuss the eclipse of behavior in personality and social psychology, in which direct observation of behavior has been increasingly supplanted by introspective self-reports, hypothetical scenarios, and questionnaire ratings. We advocate a renewed commitment to including direct observation of behavior whenever possible and in at least a healthy minority of research projects.” – Dies Autoren stellen eine anhaltende Fehlentwicklung fest: Seit der *kognitiven Wende* in den 1980er Jahren *ersetzen* die Selbstbeurteilungen innerer Zustände zunehmend die Verhaltensanalyse, statt sie adäquat zu *ergänzen*. Fragebogen sind zwar bequemer anzuwenden als die aufwändigeren Verhaltensbeobachtungen, doch sind die typischen Mängel unübersehbar: Erinnerungstäuschungen, Verzerrung durch Antworttendenzen u. a. motivationale oder kognitive Sets, systematisch verfälschende Retrospektionseffekte, d. h. wesentliche Diskrepanzen zwischen Selbsteinschätzungen und tatsächlichem Verhalten. – Im Rückblick könnte es tatsächlich als ein fast dramatisch zu nennender Einschnitt gelten, dass die psychologische Diagnostik trotz des Aufbruchs zum multi-methodischen *Behavioral Assessment* (Haynes & Wilson, 1979) in den 1970er Jahren in weiten Bereichen der Psychologie auf einen früheren Stand, d. h. auf Selbstbeurteilungen, zurückfiel.

Vielleicht darf diese Entwicklung nicht zu grundsätzlich als eine wissenschaftstheoretische Wende, als Präferenz der „kognitiven“ (mentalen) Repräsentationen auf Kosten der intersubjektiven Realitätsprüfung angesehen werden, sondern hat auch

triviale Gründe. In den Umfragen über die Verbreitung psychologischer Tests, wurde oft gesagt, dass andere Tests erstens zu aufwändig und zweitens für die Zielgruppe ungeeignet wären (Schorr, 1995, Steck, 1997). Die mono-methodisch angewendeten Fragebogen sind für einige psychologische Fragestellungen testökonomisch und unersetzlich, sollten aber in der Regel nur ein Bestandteil einer multimodalen Diagnostik sein, um krasse Fehlbeurteilungen zu vermeiden. 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. Wer auf die psychologischen Informationen aus Persönlichkeitsfragebogen nicht verzichten will, sollte grundsätzlich multimodal vorgehen und dabei die strukturelle Subjektivität von Selbstbeurteilungen akzeptieren. – 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.
Auffällig bleibt der Mangel an kritischen MTMM-Untersuchungen, in denen bestimmte Skalen von Persönlichkeitsfragebogen mit entsprechenden Verfahren aus mindestens zwei anderen Methodentypen verglichen werden: Fremdbeurteilungen (durch Bekannte und durch andere Personen), Verhaltensbeobachtungen, Daten aus breit angelegtem ambulantem Assessment nach verschieden Untersuchungsplänen. Das weitgehende Fehlen solcher multimethodischen Untersuchungen könnte vermuten lassen, dass nicht nur der größere Aufwand im Vergleich zu den typischen Publikationen aufgrund von Persönlichkeitsfragebogen, sondern auch die begründete kritische Erwartung hinsichtlich mangelnder Konvergenz hinderlich sind (siehe Campbell und Fiske, 1959). – Die Forschungsrichtung des *ambulanten Assessment* scheint sich in Teilgebieten bereits weitgehend von der noch relativ verbreiteten Fragebogenmethodik getrennt zu haben, zumindest was aktuelle Tätigkeiten, Befinden, Emotionen und Beanspruchung (Stress) betrifft. Umso mehr sind strategische Fragen und multimethodische Untersuchungen wichtig, um geeignete Methoden in ihren speziellen Vorteilen und Nachteilen zu beurteilen und, je nach Aufgabenstellung, strategisch kombinieren.

Strategische Konsequenzen – Ausblick von einer mittleren Position

Selbstbeurteilungen 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 testkökonomisch, sie haben eine unmittelbare oder auch scheinbare Validität. 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 nur die Verhaltensunterschiede zu erläutern helfen. Grundsätzlich kann zwischen verschiedenen Absichten und Strategien unterschieden werden.
Cronbach (1970, S. 555 f.) argumentierte, 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. Inwieweit 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 biografisch-diagnostischen Ansatz in der psychologischen Beratung und Fallarbeit. Auch viele der 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 soziodemografischer Merkmale, wie es mit der Methode statistischer Zwillinge, matched pairs, geschieht.) Die Wiederholung der repräsentativen Normierung des FPI-R hatte das überraschende Ergebnis, dass die Itemmuster und die Testwert-Verteilungen (Normen) nur wenig voneinander abwichen. Die FPI-R-Skalen repräsentieren also psychologische Konstrukte, die offensichtlich in den Selbstbeschreibungen der Durchschnittsbevölkerung einen herausragenden und überdauernden Platz haben. Es handelt sich um robuste Dimensionen eines differenziell-psychologischen Beschreibungssystems. Aus der Einsicht in die *strukturelle*

Subjektivität dieser vielschichtigen Selbstbeurteilungen 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 soziodemografischer, klinischer und beruflicher Art, zu anderen selbstberichteten Inhalten, zu Fremdeinstufungen und zu vielen empirisch prüfbaren Daten und Verhaltensweisen.

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 Unterscheidung 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. Über die im Inventar erfassten Persönlichkeitseigenschaften hinaus können jedoch störungsspezifische diagnostische Aussagen kaum erwartet werden. Mehrere Untersuchungen zeigten in Katamnesen, dass der subjektiv erfahrene – und teils auch der medizinisch beurteilte – Erfolg mit den Testwerten der Erstuntersuchung, u. a. der Emotionalität, Beanspruchung, Neigung zu körperlichen Beschwerden und Lebenszufriedenheit deutlich zusammenhängt. Die weit hervorragende Zürich- Studie hat in einer großen Längsschnittuntersuchung beeindruckend gezeigt, dass die Skalenwerte der Emotionalität (Neurotizismus), aber auch hinsichtlich Aggressivität und Extraversion, Prädiktoren von Verhaltensauffälligkeiten, von psychopathologischen bzw. psychosomatischen Befunden und von psychiatrischen Diagnosen sind. – Die zahlreichen Befunde (und auch einschränkenden Feststellungen) in vielen Anwendungsfeldern lassen sich inzwischen nicht mehr in wenigen Aussagen bilanzieren.

Die Persönlichkeitsforschung steht weiterhin vor der von Fiske erkannten konzeptuellen und methodischen Herausforderung, anstelle der globalen Eigenschaftskonstrukte wesentlich kleinere Beschreibungseinheiten zu entwickeln. Das Freiburger Arbeitsprogramm zur psychophysiologischen Persönlichkeitsforschung, dessen Ergebnisse kurz zitiert sind, hat die Divergenz zwischen den Selbstbeurteilungen der emotionalen Labilität (bzw. der vegetativen Labilität bzw. allgemeinen körperlichen Beschwerden) und den Messungen der psychophysischen Reaktivität aufgezeigt statt der erwarteten Korrelate im Sinne Eysencks. 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. Wenn vorgeschlagen wird, in der psychologischen Diagnostik nicht uni-methodisch vorzugehen, sondern multimethodisch, um verschiedene Verfahren zur Absicherung zu verwenden, dann müsste diese Empfehlung noch entschiedener lauten: nicht *multi-methodisch,* z. B. mit einer Kombination verschiedener Fragebogen, sondern *multi-modal* auf kategorial verschiedenen Datenebenen. Die *multimodale Strategie* ist jedoch gegenwärtig weder testtheoretisch und testmethodisch noch hinsichtlich praktisch-diagnostischer Standards hinreichend ausgearbeitet.

Die Fragebogenantworten konsequent als *Selbstbeurteilungen* anzusehen, kann die Absicht fördern, solche Befunde sinnvoll mit den entsprechenden Daten von Bezugsgruppen sowie mit *Verhaltensinformationen* zu vergleichen. Auch die zunächst abstrakten Hinweise auf die Nutzenfunktion und die potenzielle Schadensfunktion jeder diagnostischen Feststellung, verlangen eine vorsichtige Interpretationsweise und – wo immer möglich – eine *multimodale Strategie*, die möglichst viele Kontextinformationen und Absicherungen einzubeziehen versucht. Eine mittlere Position einzunehmen, bedeutet auch, bei pauschaler Kritik an dieser Methodik nach der wissenschaftlichen Überzeugungskraft der Argumente zu fragen, und noch einmal: Welche Alternative für die Persönlichkeitsforschung und Persönlichkeitsdiagnostik schlagen die Kritiker beim Verzicht auf die Fragebogenmethodik vor?

In der allgemeinen Kritik an Fragebogen 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.
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 oder Inter-Test-Korrelationen mit anderen Persönlichkeitsfragebogen 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 methodisch einfacher sein?
Überzeugende Validierungsstudien haben folglich 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 Evaluierungen 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?

9. Einstellungen zu Persönlichkeitsfragebogen, populäre Medienbeiträge und problematische Publikationen

Parallel zu den wissenschaftlichen Diskussionen und Kontroversen über Persönlichkeitsfragebogen gibt es auch auf der Seite der Befragten sowie von Journalisten zustimmende und mehr oder minder kritische Kommentare. Auch die breite Anwendung dieser Tests regte zu Leitfäden an, wie ein Persönlichkeitsfragebogen auszufüllen ist, um in einer Bewerbungssituation einen guten Eindruck zu machen. Zunehmend gibt es Pressemeldungen über die Ergebnisse von Umfragen, ohne die notwendigen Vorbehalte zu erläutern, falls die Autoren ihrerseits überhaupt die Methodik und die Generalisierbarkeit der Ergebnisse hinreichend erläutert hatten. Andererseits äußern sich Journalisten und Blogger im Internet kritisch zu Fragebogen, offensichtlich ohne sich auf eine fachliche Ausbildung stützen zu können.
Die Kritik der Persönlichkeitsfragebogen aus der Sicht der *Befragten* kann am besten erfasst werden, indem die Datenerhebung für Testkonstruktion und Normierung mit einigen Meta-Fragen, also Fragen über den Fragebogen, verbunden wird. Bei der Repräsentativerhebung 1982 wurden im Anschluss an die 240 Items des FPI-G sechs Meta-Fragen gestellt. Die Befragten wurden gebeten, die benötigte Zeit in Minuten zu schätzen. Es folgten Fragen nach der Sicherheit der Beantwortung und nach der Verständlichkeit der Fragen, nach dem Zutreffen der Fragen auf wesentliche Charaktereigenschaften und nach der möglichen „Zudringlichkeit" der Fragen (jeweils ein Balken von 0, 10, 20 … 100 %) und eine Einstufung, inwieweit solche Fragebogen für geeignet gehalten werden, Menschen besser zu verstehen (1 = sehr gut geeignet bis 5 = nicht geeignet). Die Schätzungen der benötigten Zeit für 240 Items lagen zwischen 10 und 150 Minuten, im Mittel bei 32 Minuten

(Standardabweichung 16, 1. Quartil 20, 3. Quartil 40 Minuten), sodass der mittlere Zeitbedarf für 138 Items des FPI-R in der Größenordnung von 18 Minuten liegt. Die Befragten halten im Mittel 76 % der Items für sicher beantwortbar, 86 % für verständlich, 57 % für persönlich zutreffend, 22 % für zu persönlich und zudringlich. Von den Befragten stuften 49 % den Fragebogen als gut oder sehr gut geeignet ein, 27 % als befriedigend, 10 % als ausreichend und 15 % als nicht geeignet zum besseren Verständnis von Menschen. – Die Beziehungen zwischen Meta-Fragen und FPI-R-Skalen (1982) sind geringfügig und machen höchstens ein Prozent gemeinsame Varianz aus. Die relativ höchste Korrelation besteht zwischen FPI-R 8 Körperliche Beschwerden und relativ hohem Zeitbedarf für das Ausfüllen des Fragebogens (r=.13).

Bei Verknüpfung der Meta-Fragen mit den IfD-Statusmerkmalen ergaben sich einige hochsignifikante Befunde (1982). Wer bei relativ vielen Items unsicher in der Beantwortung war, wurde auch vom Interviewer als „ziemlich unsicher" eingestuft. Wer relativ viele Items unverständlich fand, gehörte eher zur höchsten oder zur niedrigsten Bildungs- und Einkommens-Schicht und wurde vom Interviewer eher als unsicher eingestuft. Wer relativ viele Items als persönlich unzutreffend ansah, gehörte eher zu jenen Personen mit Abitur und mit Studium. Insgesamt ist die gemeinsame Varianz von Skalen und Meta-Fragen zu gering, um von den Antworten zu den Meta-Fragen eine praktisch verwertbare Moderatorfunktion erwarten zu können.

In den *Medien* wird zunehmend über die Ergebnisse von psychologischen Umfragen berichtet, die sich auf Fragebogen oder Internet-Umfragen stützen, wobei kaum zwischen fachlich qualifizierten und unseriösen Berichten unterschieden werden kann. Oft ist nicht zu erkennen, ob erst die journalistische Kurzdarstellung oder bereits der Bericht (bzw. die Pressemitteilung) der Autoren diese grundsätzlichen Mängel enthalten. Insofern ist auch die populäre Rezeption eines psychologischen Fragebogens ein Thema der psychologischen Methodik. Es sind Missverständnisse und Stereotypisierungen mit entsprechenden Rückwirkungen möglich, d. h. eine potenzielle „Schadensfunktion". Typische Kardinalfehler sind die *bedenkenlosen Verallgemeinerungen* der erhaltenen Ergebnisse und die *naiven Kausaldeutungen* von Korrelationen der Testwerte mit anderen Daten. Dass die Daten aus einem irgendwie gesammelten Fragebogenmaterial oder aus einer irgendwie angelegten Umfrageaktion im Internet bzw. in den Social Media einen fundamentalen Selektionsfehler haben müssen, wird vielen Journalisten, Redakteuren und Bloggern nicht ohne weiteres deutlich sein, sodass auch keine fachliche Beratung benötigt und gesucht wird.

Das *Internet* wird heute auch benutzt, um mit Persönlichkeitsfragebogen bzw. ihren Kurzformen Daten von interessierten Personen kostenlos zu gewinnen, ggf. wird eine Rückmeldung der Ergebnisse in Aussicht gestellt. Außerdem stehen in den Medien gelegentlich Berichte über Forschungsergebnisse, die mittels Fragebogen gewonnen wurden, in selteneren Fällen auch über die Konstruktion und Anwendung solcher Fragebogen. Allerdings sollte bei den zunehmend auch von Psychologen, teils mit Adressen einer Universität oder eines Instituts durchgeführten Internet-Umfragen der fundamentale Hinweis unerlässlich sein, dass solche Daten keine allgemeine Gültigkeit im Vergleich zu den wesentlich aufwändigeren bevölkerungsrepräsentativen Umfragen haben, zumal anscheinend oft auf Erhebung bzw. Mitteilung der Quotierungen und weitere Kontrollen verzichtet wurde. Folglich sind die Mitteilungen aufgrund solcher Internet-Umfragen nicht als grundsätzlich replizierbare Ergebnisse anzusehen, sondern als spekulative Mitteilungen mit einem groben systematischen Bias: wen erreichen solche Aktionen und wer antwortet darauf? (Selbst, wenn ein Honorar in Aussicht gestellt wird?)

Aus fachlicher Sicht sehr problematisch sind viele *populäre Artikel* in den Medien. Ein Blogger, Lars Lorber, der sich für Persönlichkeitspsychologie interessiert und ein Buch über „Der große Typentest" schrieb, veröffentlichte eine Kurzbeschreibung der 12 FPI-Skalen mit dem Titel „12 ungewöhnliche Eigenschaften im Test" und urteilte dann: „Heutige Relevanz. Das Freiburger Persönlichkeitsinventar wurde – im Gegensatz zu den meisten anderen Persönlichkeitstests – weder auf der Basis einer bestimmten Theorie zur menschlichen Persönlichkeit noch auf der Basis wissenschaftlicher Studien erstellt. … Es erfüllt wissenschaftliche Standards, kann jedoch auch kritisiert werden, dass seine 12 Eigenschaften etwas beliebig zusammengestellt wurden. Aufgrund seines Alters und der Dominanz der aussagekräftigeren Big Five ist es in der Persönlichkeitsforschung heute kaum noch präsent." Als Quelle dieser Einsichten gibt Lorber die Publikation von Borkenau und Ostendorf (1989) an. Verfügbar unter: http://www.typentest.de/blog/2015/11/das-freiburgerpersonlichkeitsinventar/

Lorber publizierte im Internet außerdem eine Liste „nicht empfehlenswerter Persönlichkeitstests" und im Jahr 2016 „Die große Liste der Persönlichkeitstests aktualisiert", in denen er Persönlichkeitstests bewertet und empfiehlt, wobei der Aspekt „kostenlos" wesentlich ist. Der naive Hinweis zu „BIG Five und die Varianten davon" ist bemerkenswert: „Ja. DER wissenschaftliche Gold-Standard seit Jahrzehnten. Kein anderer Test hat auch nur ansatzweise eine so hohe wissenschaftliche Bedeutung. Kostenlos, sowie in vielen Büchern, und von Anbietern." Lorber sieht im Modell Big Five den „eindeutigen Klassenbesten" und hat auf dieser Basis seinen eigenen „Typentest" entwickelt. (Gegen das FPI hat der Blogger auch den

Einwand, dass es nicht kostenlos verfügbar ist.) In den Worten des Bloggers äußert sich eine fast missionarisch zu nennende Überzeugung, sodass nach problematischen Konsequenzen gefragt werden kann. Diese und ähnliche Meinungen in den Medien, auch in der Wikipedia, werden in der Medien-Öffentlichkeit für Informationssuchende gewiss sehr viel wichtiger sein als die professionellen Test-Rezensionen. Ob es schon vorkam, dass eine Person, die fachpsychologischen Rat suchte, unbedingt mit dem (amerikanischen) BIG FIVE untersucht werden wollte – ähnlich jenen Patienten, die von ihrem Hausarzt unbedingt das Medikament verschrieben haben möchten, von dem im Internet so überzeugend die Rede war?

Problematische „Medien-Auftritte" gibt es auch von Universitätsangehörigen, wie ein aktuelles Beispiel zeigt. Ein Interview in der *Badischen Zeitung* am 2. 10. 2018 mit dem Titel „Freiburger sind neurotischer" wurde zum Anlass der eigenen Recherche bereits, bevor der sehr ausführliche Artikel in der *Psychologischen Rundschau* publiziert wurde (Obschonka et al., 2019). Der interviewte Wirtschaftswissenschaftler Wyrwich berichtet über eine großangelegte Studie, die zu einer „Deutschlandkarte" geführt habe: „Regionale Unterschiede der Verteilung von Personen mit unternehmerischem Persönlichkeitsprofil in Deutschland – ein Überblick". Die Autoren sind an der regionalen Verteilung des „Big Five Eigenschaftsprofils" dieser Personengruppe interessiert und organisierten im Zeitraum von 2003 bis 2015 als Teil eines internationalen Projekts eine Internet-basierte Umfrage, deren Repräsentativität angeblich durch statistische Gewichtung aufgrund der Quoten des Statistischen Bundesamts hergestellt wurde. Erfasst waren Geschlecht, Altersgruppe und Postleitzahlen des Geburtsorts und des gegenwärtigen Wohnortes. Die statistischen Unterschiede verschiedener Regionen hinsichtlich „unternehmerischer Persönlichkeitsmerkmale" wurden auf „selektive Migration" (und vermutete Erblichkeit) und auf „Sozialisationseffekte" zurückgeführt. Eine einfachere Interpretation wäre: die Autoren haben den kritischen Unterschied von p- und d-Statistiken unzureichend bedacht, sodass minimale Unterschiede überinterpretiert und kausal gedeutet werden, und sie haben vor allem den fundamentalen Selektionsbias nicht bemerkt: Wer beantwortet im Internet freiwillig solche Umfragen? So werden hier aus Mängeln der Datenerhebung Migrationsmuster der unternehmerischen Persönlichkeit, und „die Freiburger" (Studierende, die sich im Internet äußern?) sind wohl durch Sozialisationseffekte neurotischer oder doch durch Migrationseffekte?

Für problematisch halten die FPI-Autoren auch die zunehmende Verbreitung von sogenannten Testknackern. Bereits 1974 publizierte Susanne von Paczensky „Der Testknacker. Wie man Karriere-Tests erfolgreich besteht (Erstauflage bei Bertelsmann [Gütersloh], Neuauflage 2018 bei Rowohlt Repertoire). „Wer heute eine Stellung sucht oder aufsteigen will, muss sich immer häufiger sogenannten Persönlichkeitstests unterziehen. Anders als Eignungstests prüfen sie keine berufsbezogenen Fähigkeiten, sondern zielen mit fragwürdigen Mitteln auf den Intimbereich ab. Sie sollen Merkmale aufdecken wie Labilität, sexuelle Neigungen, Familienkonflikte, Trinkgewohnheiten, Ängste, Aggressionen oder Abhängigkeiten. Diese Testverfahren werden geheim gehalten, ihre Ergebnisse dem Getesteten nicht mitgeteilt, wohl aber im Bedarfsfall gegen ihn verwendet. Persönlichkeitstests mögen für den Therapeuten trotz ihrer Schwächen ein brauchbares psychodiagnostisches Hilfsmittel sein. Bei der Personalauslese in Industrie und Verwaltung haben sie, wie Susanne v. Paczensky schlüssig belegt, aus moralischen, arbeitsrechtlichen und wissenschaftlichen Gründen nichts zu suchen. Ihr kritischer ‚Wegweiser durch das Testdickicht' deckt deshalb die Konstruktion der einzelnen Tests auf und lehrt, wie man sie trickreich und elegant unterläuft. (Verlagsinformation von Rowohlt Repertoire, siehe https://www.rowohlt.de. Abruf am 30. 8. 2019). Seitdem existieren in weiter Verbreitung zahlreiche Testknacker und Testtrainer, d. h. Anleitungen für Einstellungs- und Berufseignungstest: für die Einstellung im öffentlichen Dienst, zur erfolgreichen Darstellung im Assessment-Center, bei Führerscheinverlust usw. Als Beispiel dient das Portal „Career-Test.de". Angesichts der oft laienhaften und teils irreführenden Urteile ist der Hinweis bemerkenswert: „Career-Test wird von Diplom-Psychologen mit langjähriger Erfahrung in der Testerstellung und -durchführung entwickelt und gehört seit 2008 zu den beliebtesten Karriereportalen im deutschsprachigen Raum." (Abruf am 20. 8. 2018). Ein „Test-Training mit Lösungen" verlangt als „Schutzgebühr" 5.50 Euro. Als „Die wichtigsten Einstellungstests" werden u. a. der Big- Five-Persönlichkeitstest (B5T) von Satow (eine Variante der 5-Faktor-Modelle), der 16 PF, das FPI und weitere Verfahren genannt. Aus dem FPI werden auf dem Portal ohne genaue Quellenangabe ein Item jeder FPI-Skala wörtlich zitiert, und das naive Urteil lautet insgesamt: „Willkürliche Auswahl der Persönlichkeitsmerkmale durch die Autoren, Kein direkter Bezug zur Arbeitswelt, Kein Intervallskalenniveau, Veraltete Normen."

Angesichts der Bedeutung der Berufswelt (und auch der finanziellen Interessen vieler Testtrainer) ist es verständlich, dass die Stiftung Warentest dieses Angebot zu vergleichen versuchte. Der anonym verfasste Artikel *Persönlichkeitstests im Internet. Was bin ich?* (verfügbar unter test.de, Abruf am 9. 7. 2014) bewertet 10 „Persönlichkeitstests" und erwähnt hier einige sehr allgemeine Anmerkungen der ebenfalls anonymen Experten. Zwei Varianten des 5-Faktoren-Konzepts wurden bewertet: der Fragebogen zur Persönlichkeit, Psychologisches Institut der Universität Münster (www.uni-muenster.de/ psyweb: „Sie erhalten Studienergebnisse, Individuelles Feedback, ggf. Sachprämien") als befriedigend (Note 3.7) und der Fragebogen von

Psychomedia (verfügbar unter http://www.psychomedia.de) (Note 3.9).
Interessant sind die auf folgender Seite publizierten Rückmeldungen: https://www.test.de/Persoenlichkeitstests-im-Internet-Was-bin-ich- 4727586 -0/(Stand 9. 7. 2014; 4 negativ und 45 positiv), denn die Redaktion beantwortete einen der kritischen Kommentare: „Die Anlage unserer Prüfung können Sie dem Unterartikel „So haben wir getestet“ entnehmen. Daraus geht hervor, dass die Untersuchung und Bewertung der Testverfahren natürlich nicht auf einer subjektiven Sicht beruhen, sondern dass die Prüfung durch verschiedene Fachexperten, kombiniert mit Testdurchläufen mehrerer Nutzer erfolgte. Bei den Experten, die das Testverfahren überprüften, handelte es sich um ausgewählte und neutrale Fachleute auf dem Gebiet der psychologischen Diagnostik. Die Prüfkriterien orientierten sich an der DIN 33430 ‚Anforderung an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen‘ … (aci).“ – Hinweise auf berufsethische Überlegungen waren nicht einmal andeutungsweise enthalten. – Kann sich hier aus fachlicher Sicht die ironische Frage anschließen, ob nicht auch andere psychologische Tests einfacher durch die *Stiftung Warentest* statt mühselig nach *TBS-TK Standards* zu evaluieren sind?

10. Qualitätssicherung

Deutsche Richtlinien zur Testbeurteilung

„Das Diagnostik- und Testkuratorium (DTK) der Föderation Deutscher Psychologenvereinigungen (BDP, DGPs) setzt mit dem Testbeurteilungssystem TBS-TK Standards für die Rezension von psychologisch-diagnostischen Verfahren und trägt somit zur Qualitätsverbesserung diagnostischer Verfahren und Entscheidungen bei.“ (vgl. Erläuterungen zum TBS-TK auf folgender Internetseite: https://www. psyndex.de/index.php?wahl=Testkuratorium). Durch die Gründung des „Zentrum für wissenschaftlich-psychologische Dienstleistungen (DGPs)“ ergeben sich viele Chancen fachlicher Unterstützung und Kooperation, nachdem mit dem *Leibniz-Zentrum für Psychologische Information und Dokumentation* (ZPID) schon seit Jahren vorzügliche Möglichkeiten der Dokumentation, Literatursuche und Datenarchivierung bestehen.

Das *Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen* (TBS-TK) liegt inzwischen in der 3. Fassung vor (Diagnostik- und Testkuratorium, 2018b). Eine beträchtliche Anzahl von Testrezensionen, nach den früheren Richtlinien (Testkuratorium, 2007, 2010) oder nach dem neuen TBS-TKSystem verfasst, ist zu finden über den Link: https://www. bdp-verband.de/publikationen/testrezensionen. Diese Testrezensionen werden auch in der *Psychologischen Rundschau* publiziert. – In dem *Verzeichnis der Testrezensionen* (25., aktualisierte Auflage, Stand: Juni 2019) sind 15 Rezensionen des FPI aufgeführt. Verfügbar unter: https://www.psyndex.de/pub/tests/verz_teil5.pdf.
Qualitätssicherung durch unabhängige Test-Beurteilungen ist notwendig, weil psychologische Tests auch für Entscheidungen von hoher Tragweite verwendet werden. Eine *gesetzliche Grundlage* haben die Medizinisch-psychologischen Untersuchungen (MPU) bei der Feststellung der Fahreignung (nach § 71a FeV). Es gibt eine „Richtlinie zur Bestätigung der Eignung der Testverfahren und -geräte und der Eignung der Kurse zur Wiederherstellung der Kraftfahreignung“ vom 31. März 2017 (Bundesministeriums für Verkehr und digitale Infrastruktur der Bundesrepublik Deutschland, 2017; . S. 227 ff.). Dazu sollen die „Eignung der eingesetzten psychologischen Testverfahren und -geräte […] durch eine unabhängige Stelle bestätigt werden.“ – Die TBS-TK-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 wären ein Codebook bzw. eine Sammlung geeigneter Beispiele notwendig, um durch Veröffentlichung und öffentliche Diskussion problematischer Geschehnisse zu lernen und Regeln zu entwickeln.
Eine weitere Richtlinie, die DIN 33430 (Diagnostik- und Testkuratorium, 2018a), gilt genau genommen für die berufliche Eignungsdiagnostik und somit nicht für das Freiburger Persönlichkeitsinventar. Die Testautoren hatten bereits in den früheren Auflagen darauf hingewiesen, dass beim FPI typischer Weise an Personen in Forschungsvorhaben, an Patienten unter Beratungs- oder Therapiebedingungen oder an anonym bleibende Personen gedacht sei und nicht an Personen unter Prüfungs-, Bewerbungs- und Personalauslese-Bedingungen. Dennoch enthält Kerstings (2018) DIN-SCREEN Checkliste 1, Version 3, die sich primär auf die Eignungsdiagnostik und Personalpsychologie bezieht, zahlreiche testmethodische

Merkmale, die generell wichtig sind.

In der Konsequenz solcher Schritte auf dem Gebiet der Testrezensionen läge es, wie in der medizinischen Diagnostik und Labormedizin, auch an vertiefte fachliche Kooperation, an verpflichtende Vereinbarungen, eventuell auch an (freiwillige) Kontrollverfahren zu denken. Es gibt methodisch unzureichend entwickelte Instrumente und ungenügend normierte Tests, und es gibt die Perspektive, dass zur Evaluation des erwarteten psychologischen Nutzens logisch zwingend auch die Reflektion einer möglichen Schadensfunktion gehört. Ist zu erwarten, 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 anonym. Gibt es eine Vision, dass die Fachverbände der Psychologie ein modernes Fehler-Reportingsystem einführen? In der psychologischen Diagnostik bilden die Testrezensionen (und die Lehrbücher) nur die eine Seite, lassen jedoch die praktische Seite, die tatsächliche Anwendung und die Compliance hinsichtlich der Maßstäbe der Qualitätssicherung unberührt. – Vielleicht erinnern sich auch andere Testautoren, von betroffenen Personen Hinweise auf sehr problematische Testanwendungen durch Fachpsychologen erhalten zu haben? An wen können eventuelle Beschwerden gerichtet werden?

Ein anderer und direkt zu verwirklichender Ansatz wäre die *Reanalyse* der zur Testkonstruktion verwendeten Datensätze, deren Archivierung *open access* selbstverständlich sein sollte. Tatsächlich sind in PsychData des ZPID erst wenige Datensätze dieser Art archiviert. Sie stehen nach schriftlicher Vereinbarung open access zur Verfügung. Die Besorgnis, diese Daten könnten leicht zu einer konkurrierenden Konstruktion abgezweigt werden, ist weitgehend unbegründet, denn diese Aktion würde, abgesehen davon, dass sie leicht zu erkennen wäre, gegen berufsethische Verpflichtungen und das Copyright des Verlags verstoßen. – Eine Umfrage unter den wenigen „Datennehmern" der FPI-Daten aus PsychData ergab allerdings im Jahr 2017, dass es zumeist um die Verwendung im akademischen Unterricht ging; der Versuch einer Reanalyse mit Rekonstruktion wurde bisher nicht unternommen bzw. publiziert. Tatsächliche Replikationsversuche, u. a. auf Datenbanken gestützt, sind auch auf diesem Gebiet der Psychologie noch extrem selten.

International Test Commission Guidelines

Den deutschen Richtlinien waren die von Kommissionen der *American Psychological Association* ausgearbeiteten Regeln *(Standards for educational and psychological testing)* vorausgegangen (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1985; vgl. Häcker, Leutner & Amelang, 1998).

Heute sind an erster Stelle die Originalfassungen der *ITC Guidelines on Test Use, Version 1.2* (International Test Commission, 2013) zu nennen. Sie liegen in verschiedenen Sprachen vor, die auf der Seite der ITC heruntergeladen können: *ITC Guidelines on Test Use (translations and description* (https://www.intestcom.org/ page/17). Eine deutsche Fassung der Guidelines (*Internationale Richtlinien für die Testanwendung, Version 2000)* wurde 2001 veröffentlicht (International Test Commission, 2001). Das Projekt der ITC (International Test Commission) hat das Ziel, Richtlinien für die fachgerechte Testanwendung zu erarbeiten und eine bestmögliche Durchführungspraxis in der Diagnostik anzuregen. Die bisherige Arbeit der ITC zur Förderung fachgerechter Testadaptionen (Hambleton, Merenda & Spielberger, 2005; Hambleton & Patsula, 1999) trägt wesentlich dazu bei, die Übereinstimmung in der Qualität von Tests zu gewährleisten, die zum Einsatz in unterschiedlichen Kultur- und Sprachbereichen adaptiert werden. Auf seinem Kongress in Athen 1995 griff der ITC-Beirat die Anregung auf, dieses Anliegen durch die Erarbeitung von Richtlinien für eine faire und ethisch korrekte Anwendung von Tests zu erweitern, aus denen sich Standards für die Ausbildung und spezifische Kompetenzanforderungen an die Testanwender ableiten lassen." (International Test Commission, 2001, S. 4). Hier wird ausdrücklich auf die erforderliche Qualitätskontrolle bei Adaptationen für unterschiedliche Kulturund Sprachbereiche hingewiesen. – Offensichtlich haben in den Kommissionen keine Psychologin und kein Psychologe aus Deutschland mitgearbeitet, auch die Literaturhinweise erwähnen keinen einzigen deutschen Beitrag. Diese Beobachtung ist auch deshalb erwähnenswert, weil es sich bei einem großen Anteil der in deutscher Sprache publizierten Tests um Adaptionen amerikanischer Vorbilder handelt. Die *International Test Commission* hat eine Anzahl von Arbeitsgruppen eingerichtet. Auf der Internetseite der ITC (www.intestcom.org/page/5) heißt es dazu: „six projects have produced guidelines that have gained wide international acceptance. These are:

1. The ITC Guidelines for Translating and Adapting Tests

2. The ITC Guidelines on Test Use
3. The ITC Guidelines on Computer-Based and Internet-delivered Testing
4. The ITC Guidelines on Quality Control in Scoring, Test Analysis and Reporting of Test Scores
5. The ITC Guidelines on the Security of Tests, Examinations, and Other Assessments
6. The ITC Guidelines on Practitioner Use of Test Revisions, Obsolete Tests, and Test Disposal
7. The ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations
   We are currently developing new guidelines on Technology-Based Assessments, in collaboration with the Association of Test Publishers (ATP).” Alle genannten Richtlinien können auf der Internetseite der ITC (www.intestcom.org) heruntergeladen werden.

Es gibt Übersetzungen der Guidelines in offiziellen Fassungen, welche ebenfalls auf der Seite der ITC (www. intestcom.org) heruntergeladen werden können (siehe auch Evers et al., 2017; Merenda, 2006). Wichtige Webadressen im Zusammenhang mit der Diskussion um Qualitätssicherung in der Psychodiagnostik sind weiterhin auf folgender Seite aufgelistet: „PsychLinker: Qualitätssicherung in der Psychodiagnostik“ (https:// www.psychlinker.de/category.php?cat=52)

– – – – – – – – – – – – – – – – – –

Testpflege: Aspekte der Qualitätssicherung

Weitere Aspekte zur Qualitätssicherung wurden im Artikel von Fahrenberg & Hampel (2020) diskutiert: „Was bedeutet „Testpflege“? – Zur Qualitätssicherung von Persönlichkeitsfragebogen“. – siehe (https://psydok.psycharchives.de/jspui/handle/20.500.11780/3785).

„Qualitätssicherung verfolgt das Ziel durch kontinuierliche Beschäftigung mit Aspekten der Struktur- und Prozessqualität die Erreichung von Qualitätsanforderungen (Qualität) sicherzustellen. … Charakteristisch ist eine Orientierung am PDCA-Zyklus [Planen-Ausführen-Prüfen-Handeln-Zyklus], der die systematische Erfassung von Qualitätsindikatoren verlangt, und nach Identifikation von Defiziten die Einleitung potenziell qualitätsverbessernder Maßnahmen einfordert, deren Effektivität möglichst unmittelbar kritisch im Hinblick auf die Zielsetzungen geprüft werden muss.“ (Wirtz, 2019, S. 1465). Die Aufgabe der Qualitätssicherung stellt sich in sehr vielen Bereichen, und ein Seitenblick beispielsweise auf die Entwicklung von Softwaresystemen zeigt, welcher Aufwand erforderlich ist. Die gesamten Kosten verteilen sich nach einer Schätzung zu 30 Prozent auf die Entstehungs- und Entwicklungsphase, zu 40 Prozent auf die Evolutionsphase (Korrekturen, Änderungen, Optimierungen, Erweiterungen), zu 25 Prozent auf die Erhaltungsphase und zu 5 Prozent auf die Ablösungsphase bzw. Ausmusterung; die Erhaltungsphase beansprucht etwa den doppelten Zeitaufwand der Evolutionsphase (Sneed, Baumgartner & Seidl, 2011, S. 199).

Im Hinblick auf einen psychologischen Test interessiert die systematische Verbesserung der Qualität aufgrund der gewonnenen Erfahrungen. Für dieses Qualitätsmanagement sind außer den Testautoren und dem Verlag auch die Anwender und die Test-Rezensenten wichtig, auch wenn sie keinen organisierten Qualitätszirkel bilden. – Gibt es von einem eingeführten Test überhaupt eine ergänzte, überarbeitete oder sogar revidierte Neuauflage und nicht bloß einen unveränderten Nachdruck?

Hauptsächliche Themen wären:

- Überprüfung der Normen;
- Replizierbarkeit der Skalen und andere testkonstruktive Aspekte;
- weiterführende Diskussion der zugrundeliegenden psychologischen Konstrukte;
- Integration von Validitätshinweisen aufgrund eigener Untersuchungen der Testautoren und aufgrund der Fachliteratur;
- Testrezensionen und Reviews ggf. mit einem kritischen Vergleich ähnlicher Tests;
- Stellungnahme der Testautoren zu Testrezensionen und zu ähnlichen Test-Publikationen;
- Engagement des Testverlags hinsichtlich Neuauflagen und Unterstützung bevölkerungsrepräsentativer Normierungen;
- Verfügbarkeit der Datensätze für unabhängige Reanalysen;
- Umfang und Zugänglichkeit der Dokumentation.

Zusammenfassung: Für psychologische Tests und deren Anwendung wurden Richtlinien der Qualitätssicherung entwickelt, insbesondere in dem *Testbeurteilungssystem TBS-DTK des Diagnostik- und Testkuratorium* (2018) und in den *Guidelines* der *International Test Commission* (2005, 2013, 2015, 2017). Im weiteren Sinn umfasst Qualitätssicherung den gesamten Prozess eines kontinuierlichen Qualitätsmanagements, an dem die Testautoren und der Verlag sowie die Anwender und die Rezensenten Anteil haben. Jede neue Auflage eines Tests bietet die Gelegenheit, qualitätsverbessernde Befunde und Überlegungen aufzunehmen. Einige dieser Aspekte werden am Beispiel der dritten Normierungsstudie zum *Freiburger Persönlichkeitsinventar FPI* diskutiert. Nach den bevölkerungsrepräsentativen Normierungen in den Jahren 1982 (nur Westdeutschland) und 1999, die keine erheblichen Abweichungen ergaben, war es zur aktuellen Qualitätskontrolle angebracht, die Normen und die Skalenkonstruktion erneut zu überprüfen. Tatsächlich sind die Normwerte der Altersgruppen 16-19 und 20-29 anzupassen. Die Unterschiede sind hauptsächlich in den Bereichen Leistungsorientierung, Aggressivität, Extraversion und Emotionalität zu erkennen. Weitere Aspekte der Qualitätssicherung werden nur kurz diskutiert. Über die testkonstruktiven Analysen und die Integration von neueren Validitätshinweisen wird im Testmanual der 9. Auflage (Fahrenberg, Hampel & Selg, 2020) ausführlich berichtet – auch in der Absicht, die geforderte kritisch-methodenbewusste Anwendung von Persönlichkeitsfragebogen zu unterstützen. Der zeitliche und finanzielle Aufwand für Skalenkonstruktion, Normierung und Validierung hinsichtlich externer Kriterien ist so hoch, dass kooperative Projekte zur Qualitätskontrolle – und künftig bereits zur Entwicklung – mehrdimensionaler Persönlichkeitsfragebogen zu erwägen sind.

– – – – – – – – – – – – – – – – –

Psychologiegeschichtliche Anmerkung

In seiner von Kraepelin in Leipzig angeregten Untersuchung über Träume und Schlaf verwendete Heerwagen (1889) 500 Exemplare eines Fragebogens. Der erste standardisierte Persönlichkeitsfragebogen ist wahrscheinlich der von Heymans und Wiersma (1906) entwickelte und nach prozentualen Antworthäufigkeiten ausgewertete Fragebogen zum Thema Vererbung psychischer Dispositionen. Binet und Simon publizierten ihren Intelligenztest in den Jahren 1905 bis 1909. In jenen Jahren gab es auch die ersten korrelationsstatistischen Untersuchungen von geistigen Leistungsmerkmalen im Hinblick auf einen Zentralfaktor der Aufmerksamkeit durch Krueger und Spearman (1907) in Wundts Leipziger Institut – der Vorläufer der späteren Korrelations- und Faktorenanalysen. Wird auch bei Fragebogen die korrelationsstatische Analyse des internen Zusammenhangs der Antworten als Kriterium gewählt, so gilt Lankes (1915) Publikation des *Interrogatory on Perseveration Tendency* als Pionierleistung – noch vor dem bekannteren *Personal Data Sheet* von Woodworth (1918), d. h. einem Fragebogen, der ein psychiatrisch orientiertes Interview von Rekruten ersetzen sollte.

Die grundsätzlichen Diskussionen über Psychometrie und über die Verständlichkeit von Fragen reichen bis zu Wilhelm Wundt und Immanuel Kant zurück. Kants (1798) kurzgefasste Methodenkritik an psychologischer Selbstbeurteilung übertrifft in ihrer Prägnanz noch heute viele einführende Lehrbücher der Psychologie. Er nennt (in heutiger Terminologie ausgedrückt) u. a. methodenbedingte Reaktivität, Konfundierung von Selbstbeobachtung und Selbstdarstellung, durch Einstellungen bedingte systematische Verzerrungseffekte der Selbstbeobachtung sowie der Selbst- und Fremdbeurteilung, Urteilsprozesse in der Wahrnehmung, selektive Beschränkung auf bewusste Vorgänge, fragliche Compliance und mögliche Reaktanz sowie die Unmöglichkeit, solche Aussagen und Zusammenhänge tatsächlich zu messen und mathematisch zu konstruieren, weshalb die Psychologie zwar eine empirische, aber keine exakte Wissenschaft sein könne (Quellenangaben siehe Fahrenberg, 2018, S. 124–125).

Zutiefst skeptisch war auch Wilhelm Wundt (1907, 1908) gegenüber Befragungen und „Ausfrageexperimenten“; demgegenüber formulierte er erstmals Prinzipien einer methodisch anspruchsvollen Interpretationslehre für Texte und psychologische Befunde (1921). 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 kontrollierten 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 äußeren Fragen beschränken, zu deren Beantwortung überhaupt keine psychologischen Beobachtungen erforderlich sind“ (Wundt, 1902, 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).

Literaturverzeichnis (Auswahl)

Das vollständige Verzeichnis steht im FPI-Manual (2020, S. 198-221) und auf der Homepage https://jochen-fahrenberg.de/

Amelang, M. & Bartussek, D. (1997). *Differentielle Psychologie und Persönlichkeitsforschung* (4. Aufl.). Stuttgart: Kohlhammer.

Andresen, B. (2015). *Mythos Big Five*. Norderstedt: Books on Demand (Herstellung und Verlag).

Baumeister, R. F., Vohs, K. D. & Funder, D. C. (2007). Psychology as the science of self-report and finger movements. Whatever happened to actual behavior? *Perspectives on Psychological Science*, 2, 396–403.

Borkenau, P. & Ostendorf, F. (2008). *NEO-Fünf-Faktoren Inventar nach Costa und McCrae. Handanweisung.* (2., neu normierte und vollständig überarbeitete Auflage). Göttingen: Hogrefe.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. 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. (1973). *Personality and mood by questionnaire*. San Francisco: Jossey-Bass.

Cheung, F. M., Fan, W. & Cheung, S. F. (2017). Indigenous measurement of personality in Asia. In: A. T. Church (Ed.) *The Praeger handbook of personality across cultures: Trait psychology across cultures* (Vol. 1, pp. 105–135). Santa Barbara, CA.: Praeger.

Cheung, F. M., van de Vijver, F. J. R. & Leong, F. T. L. (2011). Toward a new approach to the study of personality in culture. *American Psychologist*, 66 (7), 593–603.

Diagnostik- und Testkuratorium (2018b). TBS-DTK. Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 3. Januar 2018. *Psychologische Rundschau*, 69 (2), 109–116. https://doi.org/10.1026/0033-3042/a000401.

Engel, R. R. (2019). *Minnesota Multiphasic Personality Inventory – 2 – Restructured Form™ (MMPI-2-RF™). Deutschsprachige Adaptation des Minnesota Multiphasic Personality Inventory – 2 – Restructured Form™ von Yossef Ben-Porath und Auke Tellegen.* Bern: Hogrefe.

Eysenck, H. J. (1950). Criterion analysis: An Application of the hypothetico-deductive Method as Factor Analysis. *Psychological Review*, 57, 38–53.

Eysenck, H. J. (1964). *The Eysenck Personality Inventory (E-P-I).* London: University of London Press Ltd. (Eysenck-Persönlichkeits-Inventar (E-P-I), hrsg. von D. Eggert (1974, 2. Aufl., 1983). Göttingen: Hogrefe.)

Fahrenberg, J. (Hrsg.). (1987b). Multimodale Diagnostik [Themenheft]. *Diagnostica*, 33 (4).

Fahrenberg, J. (2018). Wilhelm Wundt (1832–1920). Gesamtwerk: Einführung, Zitate, Kommentare, Rezeption, Rekonstruktionsversuche. Lengerich: Pabst Science Publishers. ZPID PsyDok http://hdl.handle.net/20.500.11780/3782

Fahrenberg, J., Hampel, R. & Selg, H. (2020). *Freiburger Persönlichkeitsinventar*. 9., vollständig überarbeitete Auflage mit neuer Normierung und Validitätshinweisen, Prinzipien der Testkonstruktion und modernen Assessmenttheorie. Göttingen: Hogrefe.

Fahrenberg, J. & Hampel, R. (2020). *Was bedeutet „Testpflege"? – Zur Qualitätssicherung von Persönlichkeitsfragebogen*. Dokumentenserver für die Psychologie (PsyDok) des ZPID. Verfügbar unter: https://psydok.psycharchives.de. http://hdl.handle.net/20.500.11780/3785.

Fahrenberg, J., Leonhart, R. & Foerster, F. (2002). *Alltagsnahe Psychologie mit hand-held PC und physiologischem Mess- System.* Bern: Huber. ZPID PsyDok http://hdl.handle.net/20.500.11780/667

Fiske, D. W. (1978). *Strategies for personality research*. San Francisco: Jossey-Bass.

Goldberg, L. R. (1981). Language and individual differences: the search for universals in personality lexicon s. *Review of Personality and Social Psychology*, 2, 141–165.

Hampel, R. & Klinkhammer, F. (1978). Verfälschungstendenzen beim Freiburger Persönlichkeitsinventar in Bewerbungssituationen. *Psychologie und Praxis*, 22, 58–69.

Hampel, R. (2020). *Freiburger Persönlichkeitsinventar FPI-R. Dokumentation zur 9. Auflage.* Dokumentenserver für die Psychologie (PsyDok) des ZPID. Verfügbar unter: https://psydok.psycharchives.de. http://hdl.handle.net/20.500.11780/3786.

Hathaway, S. R. & McKinley, J. C. (1943). *Manual for the Minnesota Multiphasic Personality Inventory*. New York: Psychological Cooperation.

Haynes, S. N. & Wilson, C. L. (1979). *Behavioral assessment. Recent advances in methods*. San Francisco, CA.: Jossey-Bass.

Heerwagen, F. (1889). Statistische Untersuchungen über Träume und Schlaf. *Philosophische Studien (Leipzig)*, 5, 301–320.

Herzberg, P. Y. (2011). Selbstdarstellung in Persönlichkeitsfragebögen: Das Phänomen der sozialen Erwünschtheit. In L. F. Hornke, M. Amelang, M. Kersting (Hrsg.). *Persönlichkeitsdiagnostik* (Enzyklopädie der Psychologie. Methodologie und Methoden. Serie Psychologische Diagnostik. Bd. 4, S. 121–155). Göttingen: Hogrefe.

Heymans, G. & Wiersma, E. (1906). Beiträge zu einer speziellen Psychologie auf Grund einer Massenuntersuchung. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, 42, 81–127.

International Test Commission (2005). *ITC Guidelines on Computer-Based and Internet Delivered Testing*. Verfügbar unter: www.intestcom.org/files/guideline_computer_based_testing.pdf.

International Test Commission (2013). ITC Guidelines on Test Use (Version 1.2). Verfügbar unter: www.intestcom.org/ files/guideline_test_use.pdf

International Test Commission (2017). *The ITC Guidelines for Translating and Adapting Tests* (Sec. Edition, Version 2.4). Verfügbar unter: https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

Kant, I. (1798/1983). *Anthropologie in pragmatischer Hinsicht. In: Immanuel Kant Werkausgabe in 6 Bänden*. Band 6 (S. 395– 690). (Hrsg. Wilhelm Weischedel). Darmstadt: Wissenschaftliche Buchgesellschaft.

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

Krueger, F. & Spearman, Ch. (1907). Die Korrelation zwischen verschiedenen geistigen Leistungsfähigkeiten. Zeitschrift für Psychologie, 44, 50–114.

Kubinger, K. D. (2009). *Psychologische Diagnostik, Theorie und Praxis psychologischen Diagnostizierens* (2. Aufl.). Göttingen: Hogrefe.

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

Kury, H. (2002). Das Freiburger Persönlichkeitsinventar und sein Einsatz bei kriminologischen Fragestellungen. Das Problem der Verfälschungstendenzen (S. 249–270). In M. Myrtek (Hrsg.). *Person im biologischen und sozialen Kontext*. Göttingen: Hogrefe.

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

Lienert, G. A. & Raatz, U. (1994). *Testaufbau und Testanalyse* (5. Aufl.). 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.

Neyer, F. J. & Asendorpf, J. (2018). *Psychologie der Persönlichkeit* (6. Aufl.). Berlin: Springer.

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

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau*, 50 (3), 140–156.

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

Roth, M., Schmitt, V. & Herzberg, P. Y. (2010). Psychologische Diagnostik in der Praxis: Ergebnisse einer Befragung unter BDP-Mitgliedern. *reportpsychologie*, 35 (3), 118–128.

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.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik* (5. Aufl.). Berlin: Springer.

Schorr, A. (1995). Stand und Perspektiven diagnostischer Verfahren in der Praxis. Ergebnisse einer repräsentativen Befragung westdeutscher Psychologen. *Diagnostica*, 41, 3–20.

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.

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

Sneed, H.M. & Baumgartner, M. & Seidl, R. (2011). Testpflege und -fortschreibung. In: R. Richard (Hrsg.). Der Systemtest. Von den Anforderungen zum Qualitätsnachweis (S. 199-219). München: Carl Hanser.

Steck, P. (1997). Psychologische Testverfahren in der Praxis. Ergebnisse einer Umfrage unter Testanwendern. *Diagnostica*, 43, 267–284.

Stemmler, G., Hagemann, D. Amelang, M. & Spinath, F. (2016). *Differentielle Psychologie und Persönlichkeitsforschung* (8. Aufl. des Lehrbuchs von Amelang u. Bartussek). Stuttgart: Kohlhammer.

Testkuratorium (2010). TBS-TK-Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009. *Psychologische Rundschau*, 61 (1), 52–56.

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.

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.

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

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.

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.

Woodworth (1918), Woodworth, R. S. (1918). Personal Data Sheet. Chicago: Stoelting.

Wundt, W. (1902). Grundzüge der physiologischen Psychologie. Band 2 (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.