Was bedeutet „Testpflege“? – Zur Qualitätssicherung von Persönlichkeitsfragebogen [*]

Jochen Fahrenberg, Institut für Psychologie, und

Rainer Hampel, Arnold-Bergstraesser-Institut für kulturwissenschaftliche Forschung [**]
Universität Freiburg

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.

Schlüsselwörter: Qualitätskontrolle, Persönlichkeitsfragebogen, Normierung, Validierung, Freiburger Persönlichkeitsinventar

Quality Control of Personality Questionnaires

Abstract: Quality control broadly covers the entire process of continuous quality management, in which the test authors and the publisher, as well as the users and the reviewers participate. Some of these aspects are discussed using the example of the third standardization study of the Freiburg Personality Inventory FPI. The previous standardization studies (1982, West Germany only, and 1999), based on representative samples of the German adult population, did not show substantial deviations. For the current quality control, it appeared appropriate to re-examine the FPI-norms and the scale construction. In fact, the Stanine norms for the age groups 16-19 and 20-29 should be adjusted. Differences can be seen mainly with respect to ‘achievement orientation’, ‘aggressiveness’, ‘extraversion’ and ‘neuroticism’. Other aspects of quality control are only briefly discussed. The scale construction and new evidence of validity are reported in detail in the new test manual (Fahrenberg, Hampel & Selg, 2020, 9th edition). Time and financial efforts for scale construction, standardization, and validation with regard to external criteria is so complex that cooperative projects for quality control of multi-dimensional personality questionnaires have to be considered, even during the test development phase.

Keywords: quality control, personality questionnaire, test standardization, validation, Freiburg Personality Inventory

---

[*] Die Testautoren danken dem Verlag Hogrefe für die wichtige Unterstützung des Projekts, Herrn Jörg Hampe und Frau Dr. Anke Hensel für die fachliche Betreuung, und Herrn Michael Sommer und den Mitarbeitern im Institut für Demoskopie IfD Allensbach für die bewährte Zusammenarbeit auch bei dieser bevölkerungsrepräsentativen Erhebung.

[**] Dieser Text wurde nach Abschluss der 9. Auflage des Freiburger Persönlichkeitsinventars (2020) bei einer deutschen Fachzeitschrift für Psychologie zur Veröffentlichung eingereicht, aber vom Herausgeber „wegen unklarer Zielsetzung“ und ohne die Möglichkeit zur Neueinreichung abgelehnt. – Deshalb wird dieser Weg einer open access Veröffentlichung gewählt.
(1) Zur wichtigen Frage der Qualitätskontrolle von Tests gibt es eine pauschale Forderung im Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen, die ursprünglich aus einer Daumenregel stammt, aber, zumindest für deutschsprachige Persönlichkeitsfragebogen, keine *empirische* Basis hatte: Neue Normierung nach 8 bis 10 Jahren! (2) Nur für das Freiburger Persönlichkeitsinventar FPI existieren zwei bzw. jetzt drei Normierungsstudien auf bevölkerungsrepräsentativer Basis, jeweils verbunden mit replizierter Skalenkonstruktion und Aktualisierung der Validitätsbelege. (3) Das FPI ist auch nach der dritten Umfrage bei Mitgliedern des Berufsverbandes Deutscher Psychologinnen und Psychologen (Roth, Schmitt & Herzberg, 2010) der in der Praxis am weitesten verbreitete Persönlichkeitsfragebogen.

Korrespondenzadressen: jochen.fahrenberg@psychologie.uni-freiburg.de, rainer.hampel@abi.uni-freiburg.de

1. Aspekte der Qualitätssicherung

„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.

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.

Das Freiburger Persönlichkeitsinventar FPI wurde im Jahr 1970 publiziert, in der 4. Auflage (Fahrenberg, Hampel & Selg, 1984) durch weitere Skalen ergänzt und wiederholt hinsichtlich der Normierung und Replizierbarkeit der Skalenkonstruktion überprüft (siehe auch Bengel & Wittmann, 1983; Fahrenberg, Selg & Hampel, 1983). Nach 50 Jahren, anlässlich der 9. Auflage, wird hier aufgrund der aktuellen dritten Repräsentativerhebung über die Gesichtspunkte und Erfahrungen dieser „Testpflege" berichtet.

Richtlinien

Vorauszuschicken ist die Frage, inwieweit das *Testbeurteilungssystem TBS-DTK des Diagnostik- und Testkuratorium* (2018) oder die *Guidelines* der *International Test Commission* (ITC) bereits Grundsätze oder Beurteilungshinweise für diesen *Prozess der Qualitätsverbesserung* enthalten. Das TBS-DTK Testbeurteilungssystem fasst zahlreiche Bewertungskriterien zum Zweck von Testrezensionen zusammen. Das zugehörigen *DIN Screen* ist ein Bewertungsbogen, in dem diese Bewertungskriterien formalisiert sind. Sie wurden ursprünglich für berufsbezogene Eignungsbeurteilungen nach DIN 33430 entwickelt (Kersting, 2018a,b) und werden in Deutschland oft für andere Gebiete der psychologischen Diagnostik und Beurteilung übernommen. – Für Persönlichkeitsfragebogen und Klinische Skalen, so ist hier einzuschränken, können diese primär an Berufseignung und Berufsanforderungsanalysen orientierten Beurteilungen kaum ausreichen. So werden das für Persönlichkeitsfragebogen wesentliche Qualitätsmerkmal einer *bevölkerungsrepräsentativen* Normierung und die typischen Probleme der Skalenkonstruktion nur allgemein erwähnt.

Im TBS-TK (2018, S, 18-19) werden zahlreiche „Qualitätssichernde und qualitätsoptimierende Maßnahmen“ und Aspekte der „Qualitätsforderungen“ aufgezählt. Die differenzierte Bewertung hinsichtlich dieser qualitätsverbessernden Maßnahmen fehlt jedoch im TBS-TK weitgehend. Beispielsweise gilt die generelle Forderung, dass für die praktische Anwendung eines psychologischen Tests dessen Normierung eine zentrale Voraussetzung ist. Im *DIN Screen* (Kersting, 2018b) ist zu beurteilen „Die Angemessenheit der Normwerte wurde in den letzten acht Jahren überprüft“ (S. 239). Es handelt sich um eine Soll- und nicht um eine Muss-Forderung, und es folgt der Hinweis: „Es geht nur um eine Überprüfung der Angemessenheit der Normwerte. Ob eine Neunormierung durchgeführt werden muss, ergibt sich in Abhängigkeit von den Ergebnissen der Überprüfung. In der DIN 33430 wird nicht gefordert, dass spätestens alle acht Jahre neu normiert werden muss.“ Das folgende Beurteilungskriterium im *DIN Screen* lautet jedoch als Muss-Bestimmung: „In den Verfahrenshinweisen wird begründet, warum und unter welchen Umständen das Verfahren für einen Anwendungsfall ausgewählt werden kann, obwohl die Angemessenheit der Normwerte nicht in den letzten acht Jahren überprüft wurde.“

Demgegenüber hat die *International Test Commission* ausführliche *Guidelines for Practitioner Use of Test Revision, Obsolete Test, and Test Disposal* (2015) entwickelt, um die Beurteilung von revidierten Tests, aber auch die Gründe für den Verzicht auf überholte Tests zu vereinheitlichen: “Test Developers and Users Have a Reciprocal Relationship.” (p. 8). “Test revisions may be driven by knowledge that the assessed behaviours are subject to substantial change over time, by significant demographic changes, from research that leads to improvements in theories and concepts that should impact test use, from changes in diagnostic criteria, or in response to test consumer’s demands for improved versions. The presence of these qualities warrants consideration of whether a revised test may be more suitable than an older version.” (p. 9). “Revisions of tests that provide norm-referenced interpretations typically focus on updating norms. Thus, decisions regarding the purchase and use of a revised test may be more favourable when the norms for the revised test were obtained recently, are sufficient, and better represent the population.” (p. 11). Diese Guidelines enthalten keine Daumenregel für den zweckmäßigen Abstand von Testrevisionen. – „Practioners should have a broad knowledge of quality control practices, as these are critical to the accurate use of tests.” (ITC Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores, 2013, p. 6).

Die mehr oder minder differenzierten Richtlinien dieser Kommissionen werden die fachliche Diskussion wesentlich anregen können. Jedoch fehlen, beispielsweise zu den Anforderungen an bevölkerungsrepräsentative Normierungen, Hinweise auf empirische Untersuchungsergebnisse oder Vorbilder, so dass keine Maßstäbe oder Bedingungen solcher Maßstäbe vermittelt werden. An der Grundforderung nach regelmäßiger Überprüfung der Normen lässt sich zeigen, wie schwierig es bleibt, empirisch fundierte Anforderungen zu formulieren. Aus einer verbreiteten Daumenregel „8-10 Jahre“ wird bei Kersting (2018b) die Forderung nach einer Überprüfung im Intervall von 8 Jahren. Geeignete empirische Untersuchungen, um diese Forderung zu begründen, werden nicht zitiert. Regeln in dieser Allgemeinheit, ohne Rücksicht auf die untersuchten psychologischen Merkmale und ohne Konvention zur Evaluation von Effektstärken, kann es auch nicht geben.

Bemerkenswert sind die *ITC Guidelines for Translating and Adapting Tests* (2017), in denen die Methodenprobleme, u.a. der wesentliche Unterschied zwischen der sprachlichen und der psychologischen Äquivalenz der Übersetzung von Items, gründlich dargelegt sind. Das TBS-TK geht auf solche Fragen nicht näher ein, obwohl es sich bei vielen der in Deutschland publizierten Tests um einfache Übersetzungen angloamerikanischer Vorbilder handelt. (In den meisten der ITC-Kommissionen scheint kein Experte aus den deutschsprachigen Ländern mitgearbeitet zu haben.)

Zur Normierung von Persönlichkeitsfragebogen

Der vielfach beschriebene gesellschaftliche Wandel in Deutschland während der letzten Dekaden, eventuell zurückzuführen auf demografische Veränderungen, technologischen Fortschritt, Massenmedien und Globalisierung, könnte sich auch in veränderten Selbstbildern, Verhaltensweisen und Lebensstilen zeigen. In der Sozialforschung wurde u.a. auf eine Auflösung geschlechtsspezifischer Rollenbilder und Karrierepläne hingewiesen und erläutert, dass ältere Menschen heute aktiver seien als vor 20 Jahren, häufiger erwerbstätig, häufiger ehrenamtlich engagiert, mehr Sport treiben und mehrheitlich sozial gut eingebunden sind. Im Unterschied zu Intelligenz- und Leistungstests sind Persönlichkeitsfragebogen, Klinische Skalen und Einstellungsskalen stärker von Sprachgewohnheiten abhängig, so dass einzelne Wörter und ganze Items, auch die Rechtschreibung, bereits nach zwei oder drei Jahrzehnten „altmodisch“ wirken können. Auch hinsichtlich der Offenheit der Antworten oder der sozialen Erwünschtheit könnte es solche Trends geben.

Bei einem typischen Persönlichkeitsfragebogen ist eine bevölkerungsrepräsentative Normierung und bei einer Klinischen Skala die Normierung für eine bestimmte Patientenpopulation zu erwarten. Ein Vergleich der häufiger verwendeten (deutschen) Persönlichkeitsinventare zeigt, dass „repräsentativ“ unterschiedlich verstanden wird. Die bevölkerungsrepräsentative Normierung im engeren Sinn sollte aufgrund einer einheitlichen Erhebung durch ein auf Umfragen spezialisiertes Institut vorgenommen werden. Dieses Verfahren ist an dem Quotenplan, bezogen auf die offizielle Bevölkerungsstatistik, und an anderen Informationen zu erkennen. Je nachdem wie detailliert bestimmte Quotierungsmerkmale berücksichtigt sind und welche Zellenbesetzungen bzw. Schätzfehler vorgesehen sind, steigen der erforderliche Stichprobenumfang und die Kosten an. Für die Normierungen des FPI wurde die Größenordnung von 3000 Personen als hinreichend angesehen, um die Normen nach Geschlecht und sieben Altersgruppen differenzieren zu können. Der Datensatz wurde in bewährter Zusammenarbeit mit dem Institut für Demoskopie IfD, Allensbach, in drei Wellen erhoben. Jeweils waren ca. 300 routinierte freie Mitarbeiter des *Instituts für Demoskopie*

Allensbach beteiligt, um den Fragebogen direkt an die nach Quotenplan zu suchenden Teilnehmer auszugeben und das ergänzende mündliche Interview zu führen.

Daneben gibt es weit verbreitete Persönlichkeitsfragebogen, deren Normierungsbasis auf andere Weise gewonnen wurde. Es werden diverse Datensätze aus verschiedenen Erhebungen und laufenden Projekten als sogenannte „Eichstichprobe" zusammengesetzt, um eine möglichst breite und eventuell „repräsentative" Grundlage anzunähern. Es sind „zuhandene" Daten bzw. „Gelegenheitsstichproben", eventuell sogar durch telefonische Umfragen oder im Internet gewonnen. Über den methodischen Bias und dessen Effektstärken hinsichtlich der wichtigsten Personenmerkmale kann bei diesem Verfahren ohne eine tatsächliche Repräsentativerhebung nichts ausgesagt werden.

Auch die direkten Erhebungen durch routinierte Interviewer haben noch einige typische Methodenprobleme, doch sind sie zur Qualitätssicherung zweifellos angemessener als die Kombination diverser zuhandener Datensätze. Dass die Konstruktion und die Normierung von Persönlichkeitsfragebogen bevölkerungsrepräsentativ anzulegen sind, muss gerade angesichts der zunehmenden Publikation von Internet-basierten und deshalb hochselektiven Erhebungen betont werden (vgl. ITC-Guidelines 2005). Auch für die Publikation einzelner Forschungsvorhaben aufgrund solcher Daten muss auf den völlig unbekannten und unkontrollierbaren groben Selektions-Bias hinsichtlich der Kontaktaufnahme, der Bereitschaft und der technischen Voraussetzungen zur Teilnahme hingewiesen werden. Wenn außerdem in solchen Untersuchungen nur die Signifikanzen statt der Effektstärken von Gruppenunterschieden interpretiert werden (siehe u.a. Obschonka, Wyrwich, Fritsch, Gosling, Rentfrow & Potter, 2019) entstehen höchst fragwürdige Publikationen.

2. Zur Qualitätssicherung des Freiburger Persönlichkeitsinventars

2.1 Normierung

Trotz der nur geringen Differenzen der mittleren Testwerte zwischen den Erhebungen von 1982 und 1999, damals nur im Vergleich der westdeutschen Bevölkerung möglich, war für die 9. Auflage nach Auffassung der Autoren wie auch des Verlags eine Überprüfung erforderlich, um eventuell die Normen anzupassen. Für die aktuelle Normierungsstudie wurde außerdem, auch auf Anregung des IfD und des Verlags, die Formulierung von neun der 138 Items sprachlich leicht verändert bzw. modernisiert. Beispiele sind: Nr. 15: Ich kann mich erinnern, mal so wütend (statt: so zornig) gewesen zu sein, dass ich das nächstbeste Ding nahm und es zerriss oder zerschlug; Nr. 97: Ich werde leicht rot (statt: Ich erröte leicht). Ein möglicher differenzieller Effekt dieser Anpassung lässt sich ohne Kontrollgruppe nicht angeben. Die Antworthäufigkeiten bei diesen modifizierten Items differieren zwischen den Erhebungen 1999 und 2018 zwischen 3 und maximal 8 Prozent.

Die Daten der neuen bevölkerungsrepräsentativen Stichprobe würden die Aktualisierung der Normen-Tabellen des FPI-R für die heutige Bevölkerung ermöglichen. Diesem für die praktische Anwendung wichtigen Schritt gehen statistische Analysen voraus, ob die Veränderungen so deutlich sind, dass überhaupt eine Anpassung notwendig ist. Kriterien sind die Effektstärke ES und in praktischer Hinsicht ein anderer Normwert (Skalenstufe) für einen bestimmten Test-Rohwert. Es genügt jedoch nicht, die erhobenen Daten direkt zu vergleichen, denn der Bevölkerungsaufbau hat sich hinsichtlich Lebensalter und Bildungsabschluss kontinuierlich verändert. Im Prinzip könnten nur *repräsentative* Kohorten- bzw. Langzeitstudien eine zuverlässige Grundlage liefern.

Die Kohortenstudie des *Sozio-oekonomische Panel* (SOEP) konnte in den Jahren 2003, 2009 und 2013 mit der Kurzform eines Big Five Inventory verknüpft werden (Wagner, Lüdtke & Robitzsch, 2019). Die spezielle Auswahl und der spezielle Antwortmodus der Items lassen jedoch keine Verallgemeinerung auf andere Versionen dieser Fünf-Faktoren-Inventare zu. Der Effekt sozioökonomischer Merkmale wurde nicht untersucht, außerdem sind die Intervalle der Erhebung gering, so dass Schlussfolgerungen für die in der Praxis verwendeten Persönlichkeitsfragebogen kaum angebracht sind.

Zur *Schätzung* der Effekte mittelfristiger Veränderungen werden hier zwei statistische Verfahren verwendet (siehe Tabelle 1 und 2):

(1) die *Bildung statistischer Zwillinge* (matched pairs). Dabei standen für das Jahr 1999 N = 3 660 und für 2018 N = 3 450 Personen zur Verfügung. Die Suche nach statistischen Zwillingen ergab N = 2 304 Paare, die in Hinsicht auf Geschlecht, Alter in Jahren (7 Kategorien), Schulabschluss (3 Kategorien) und Bundesland (Ost/West) identisch sind. Dieser soziographisch kontrollierte Mittelwertvergleich ist zur Schätzung mittelfristiger Trends zweckmäßig, wenngleich der matched pairs-Datensatz nicht mehr bevölkerungsrepräsentativ ist.

(2) die *Konstruktion eines Längsschnitts (Pseudo*-Kohorte). Da auch die Veränderungen im FPI-Profil über den längeren Zeitraum von 1982, 1999 und 2018 interessierten, aber dieselben Personen nicht erneut befragt werden konnten, wurden ersatzweise Analysen mit konstruierten Alterskohorten durchgeführt. Beispielsweise wurde für den 20-jährigen männlichen Jugendlichen mit Hauptschulabschluss aus der Erhebung 1982 ein 37-jähriger Mann mit Hauptschulabschluss aus der Erhebung 1999 als statistischer Zwilling gesucht. Es verblieb schließlich eine Analyse-Stichprobe von N = 966 Paaren. Diese Stichprobe wurde dann in vier Altersgruppen unterteilt. Dieselbe Auswahlprozedur von Personen erfolgte für die Datensätze 1999 und 2018, wobei aber Ost- und Westdeutsche gemäß ihrem Anteil an der Gesamtbevölkerung berücksichtigt werden konnten. Diese Analysestichprobe besteht aus N = 1 113 Paaren.

Tabelle 1: Effektstärken in den FPI-Skalenwerten 1999 vs. 2018 nach Alter, Schulabschluss und West-/Ost-Deutschland

	1999 vs. 2018			**1999 vs. 2018**			**1999 vs. 2018**	
Einflussfaktoren	**Alter 16-29**	**Alter 30-64**	**Alter 65+**	**HS**	**RS**	**ABI+**	**West**	**Ost**
Effektstärke	**ES**	**ES**	**ES**	**ES**	**ES**	**ES**	**ES**	**ES**
Lebenszufriedenheit	-.01	.04	.03	.12	-.01	.00	.06	-.04
Soziale Orientierung	-.03	.12	.20	.10	.09	.11	.08	.15
Leistungsorientierung	.15	.11	-.03	-.03	.12	.14	.07	.13
Gehemmtheit	-.08	-.02	.20	.16	.04	-.13	.00	.05
Erregbarkeit	.08	.02	-.01	-.12	.12	.03	.00	.09
Aggressivität	.34	.05	.10	.00	.18	.15	.10	.18
Beanspruchung	.03	-.10	-.02	-.10	-.01	-.10	-.02	-.15
Körperliche Beschwerden	-.04	.02	.25	.12	.09	-.04	.01	.16
Gesundheitssorgen	-.07	.02	.29	.19	.06	-.05	.04	.11
Offenheit	.10	-.06	-.06	-.19	.00	.06	.00	-.10
Extraversion	.21	.08	-.18	-.12	.11	.11	.08	-.01
Emotionalität	.14	.01	.06	-.03	.13	.01	.05	.04
Stichprobengröße (N)	471	1.366	467	601	945	758	1.606	698

Anmerkungen: N = 2.304 statistische Zwillinge (matched pairs). HS = Hauptschule, RS = Realschule, ABI+ =, Abitur/Studium. Polung der ES positiv = Mittelwert im Jahr 1999 höher. ES ist die aus der Mittelwertdifferenz und s_M berechnete Effektstärke. Das Vorzeichen bei ES gibt die Richtung des Effekts an. Ein Minus-Zeichen bedeutet: höherer Mittelwert 2018 im Vergleich zu 1999.

Tabelle 2: Mittelwertvergleiche und Effektstärken für Alterskohorte 50-59 aus 1982 (1999) und 1999 (2018)

Bundesländer	**nur Westdeutschland**					**Deutschland gesamt**				
Jahr	**1982**		**1999**			**1999**		**2018**		
Alterskohorte	**50-59 Jahre**		**67-76 Jahre**			**50-59 Jahre**		**69-78 Jahre**		
Altersdurchschnitt	55		72			55		74		
Skala	M	s	M	s	ES	M	s	M	s	ES
Lebenszufriedenheit	7.1	3.2	7.7	2.9	-.19	7.8	3.1	7.7	2.8	.06
Soziale Orientierung	6.8	3.0	6.6	3.0	.07	6.6	3.1	6.6	2.7	-.01
Leistungsorientierung	6.8	3.2	5.3	3.3	**.49**	7.2	3.2	6.0	3.2	**.40**
Gehemmtheit	5.6	3.0	5.7	3.2	-.03	4.9	3.1	4.8	2.8	.02
Erregbarkeit	5.9	2.9	4.8	2.9	.38	5.4	3.0	4.6	3.0	.25
Aggressivität	3.9	2.6	3.4	2.9	.18	4.3	2.8	3.3	2.5	.35
Beanspruchung	6.4	3.6	3.8	3.1	**.73**	5.9	3.5	4.0	3.0	**.54**
Körperliche Beschwerden	4.8	3.2	4.8	3.1	.02	3.8	3.0	4.1	3.2	-.11
Gesundheitssorgen	6.9	3.0	8.2	2.8	**-.41**	6.7	2.8	7.6	2.7	**-.32**
Offenheit	5.4	2.8	4.8	2.9	.23	5.9	2.8	5.2	2.8	.26
Extraversion	6.1	3.6	4.9	3.4	.34	6.6	3.5	6.0	3.2	.20
Emotionalität	6.5	3.7	5.5	3.5	.28	6.0	3.7	5.5	3.6	.13
Stichprobengröße (N)	192		192			258		258		

Anmerkung: Vergleich konstruierter Kohorten nur Westdeutschland 1982 zu 1999 (N = 192) und Deutschland insgesamt 1999 zu 2018 (N = 258), siehe auch Tabelle 1.

Ergebnisse

Die Tabelle 1 zeigt, dass die *Unterschiede zwischen den FPI-Profilen 1999 und 2018* insgesamt eher gering sind, wenn Effektstärken statt t-Test-Signifikanzen zugrunde gelegt werden. Die höchste Effektstärke ergibt sich für die Skala *Aggressivität*: der Mittelwert hat bei 16-29-Jährigen abgenommen (d = .34), etwa in der Größenordnung einer Skalenstufe. In diesem Kontext ist daran zu erinnern, dass es sich nur um mittelfristige Veränderungen der Selbstbeurteilungen handelt und nicht um manifestes Verhalten. Deshalb ist kein direkter Bezug zur Gewaltkriminalität herzustellen. Die Studien von Pfeifer, Baier und Kliem (2018, S. 5) haben ergeben, dass – entgegen landläufiger Meinung und medialer Berichterstattung über schwerste Gewalttaten – die Gewaltkriminalität nicht zugenommen, sondern

abgenommen hat. – Auch die Testwerte für *Körperliche Beschwerden* und *Gesundheitssorgen* bei Älteren sind niedriger als im Jahr 1999 (d = .25 und d = .29). Dieser Trend ist in Anbetracht einer verbesserten Gesundheitsvorsorge und veränderter Lebensweise nachvollziehbar (Gesundheitsbericht des Robert Koch Instituts, 2015; S. 31). Die große Mehrzahl der berechneten Effektstärken ist unbedeutend.
Die Tabelle 2 enthält Ergebnisse des konstruierten Längsschnitts, der wegen dieser Konstruktion und mit nur zwei Zeitpunkten eine begrenzte Aussagekraft hat. – Die Mittelwertvergleiche der FPI-Skalen (1982/1999 und 1999/2018) ergaben zwar zahlreiche signifikante Differenzen, doch nur die Alterskohorte der 50-59-Jährigen wies im Vergleich mit den um 17 und 19 Jahre älteren Personen nennenswerte Veränderungen in drei FPI-Skalen auf: eine geringer ausgeprägte Leistungsorientierung, Beanspruchung und Gesundheitssorgen (Effektstärken mit Beträgen von |.32| bis |.73|). Personen in der zweiten Lebenshälfte beschreiben sich mit wachsendem Alter - nicht überraschend - als weniger beansprucht und nicht mehr „im Stress"; sie sind weniger leistungsorientiert und ihre Gesundheitssorgen nehmen zu.

Zusammenfassend ergibt sich aus der neuen Repräsentativerhebung: Die beobachteten Veränderungen sind hinsichtlich ihrer Effektstärke gering und würden sich nicht oder nur geringfügig auf die Normen auswirken. Für die Testpraxis hatten die Autoren die Stanine-Skala ausgewählt, d.h. relativ grobe Standardwerte, welche der Spannweite der Testwerte angemessen sind. – Ausnahmen bestehen jedoch hinsichtlich der Skalen *Leistungsorientierung, Aggressivität, Extraversion* und *Emotionalität.* Hier bestehen in den Mittelwerten der Jahre 1999 und 2018 Trends, die teils etwas mehr als einen Rohwertpunkt betragen, am deutlichsten in den zwei Altersklassen der 16-19 und der 20-29-Jährigen. Folglich sind die Normen für diese Altersgruppen zu ändern. Zugleich werden die Normen-Tabellen insgesamt aktualisiert, d.h. auch die geringen Effekte berücksichtigt. – Darüber hinaus wird eine zusätzliche Tabelle mit Normen aufgenommen, die nach vier Bildungskategorien, aber nur hinsichtlich der drei auffälligen Skalen *Lebenszufriedenheit*, *Gehemmtheit* und *Emotionalität* differenziert. Generell bleibt die Gliederung der Normen nach Männern und Frauen sowie nach sieben Altersklassen.

Da sich die Mittelwerte der FPI-Skalen seit der ersten Normierung 1982, d.h. in einem Zeitraum von 36 Jahren, nur geringfügig änderten, ist jene, in den zitierten TBS-TK Richtlinien pauschal erwartete Neu-Normierung im Abstand von 8 bis 10 Jahren, zu relativieren. Erst aufgrund weiterer Kontrollen bei ähnlichen Persönlichkeitsfragebogen könnten empirisch fundierte Empfehlungen gegeben werden.

2.2 Skalenkonstruktion

Die Konstruktion der 12 FPI-Skalen wurde durch Itemanalysen und Reliabilitätsanalysen, Faktorenanalysen (orthogonal, Varimax-Rotation) und durch Clusteranalysen (Ward-Algorithmus) überprüft. Explorative Versuche mit IR-Modellierungen wurden kritisch kommentiert. Hier wird nur eine zusammenfassende Tabelle wiedergegeben. Aufgrund der Faktorenanalysen sind 92 Prozent der FPI-Items nach ihren höchsten Ladungen den betreffenden Skalen zuzuordnen. In der Clusteranalyse sind es 85 Prozent der Items, wobei insbesondere die Skala Gesundheitssorgen durch Tendenz zu zwei Subclustern auffällt (Tabelle 3).

Die multistrategische Kontrolle der Skalenkonstruktion ergab keine perfekte, doch eine weitgehende Übereinstimmung der Faktoren- und Clusteranalysen untereinander und eine Kontinuität hinsichtlich der früheren Analysen. Deshalb sehen die Testautoren keinen hinreichenden Anlass, einzelne FPI-Skalen zu revidieren. Die Skalen entsprechen weiterhin den erklärten Absichten der Testentwicklung. Die Forschungsvorhaben der Testautoren gaben den theoretischen Rahmen: psychophysiologische Persönlichkeitsforschung im Sinne Eysencks, d.h. hinsichtlich Emotionalität und Extraversion, außerdem Aggressivitätsforschung, Soziale Orientierung (prosoziale Einstellung), Leistungsorientierung, Beanspruchung und Körperliche Beschwerden. In diesem theoretischen Rahmen ist die Konstruktionsstrategie hypothetisch-deduktiv und empirisch-induktiv. 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 – ohne Anspruch auf eine bestimmte, umfassende Theorie der Persönlichkeit und ohne Zuversicht in eine angeblich die gesamte Persönlichkeitsdomäne repräsentierende, lexikalisch-psychologische Reduktionsmethode (aufgrund englischer Wörterbücher?) mit wahlweise 15, 6, 5, 4 oder noch weniger Generalfaktoren – selbst wenn die Konstruktion und die Normierung bevölkerungsrepräsentativ durchgeführt wären (zur Diskussion, siehe Testmanual).

Einzelne FPI-Skalen weisen untereinander Korrelationen bis zu einer Größenordnung von etwa r = .45 (gemeinsame Varianz ca. 20%) auf, wie es für einige der Eigenschaften persönlichkeitspsychologisch erwartet werden kann. Es verbleibt jedoch ein großer Anteil eigenständiger und psychologisch-deskriptiv nützlicher, wahrer Varianz, wie aus den Reliabilitätskoeffzienten ersichtlich ist. Die Thematik des FPI-R stimmt mit einigen anderen Fragebogen mehr oder minder überein. Ein gründlicher Vergleich, der über bloß statistische Aussagen hinsichtlich der Kovarianzen hinausginge, erfordert jedoch ein, offensichtlich noch nie organisiertes, übergeordnetes Bezugssystem, das wesentliche Kriterieninformation (multimodal, aggregierend und als MTMM-Analyse) aus den Anwendungsbereichen einschließen müsste, um pragmatisch zu informieren.

Tabelle 3: Reliabilitätskoeffizienten, Faktorenanalysen der Daten aus drei Befragungswellen hinsichtlich der 10 Standardskalen und separat für E und N sowie eine Clusteranalyse (FPI-R Daten 2018, N = 3 450).

		Bevölkerungsrepräsentativer Datensatz 2018					
		# Items mit höchster Ladung auf betreffendem Faktor			Reliabilität	Clusterlösung (Anzahl Items)	
FPI-R-Skala	Anzahl Items	FA Welle 1	FA Welle 2	FA Welle 3	Cronbach α	korrekt zugeordnet	anderes Cluster
1 Lebenszufriedenheit	12	11	9	10	.77	11	1
2 Soziale Orientierung	12	12	12	12	.72	9	3
3 Leistungsorientierung	12	8	10	12	.77	11	1
4 Gehemmtheit	12	11	11	9	.77	12	0
5 Erregbarkeit	12	12	10	11	.78	12	0
6 Aggressivität	12	9	12	11	.77	10	2
7 Beanspruchung	12	11	11	11	.82	12	0
8 Körperliche Beschwerden	12	12	12	12	.77	12	0
9 Gesundheitssorgen	12	12	12	11	.75	5	7
10 Offenheit	12	11	11	12	.75	8	4
Itemsumme Skalen 1–10	120	109	110	111		102	18
E Extraversion	14	14	14	14	.80		
N Emotionalität	14	14	14	14	.82		

Anmerkungen: (1) Der Datensatz besteht aus drei repräsentativen Stichproben der deutschen Bevölkerung. Die Daten wurden in drei Erhebungswellen gesammelt (Juni und September-November 2018; N Welle 1 = 1093, N Welle 2 = 1137, N Welle 3 = 1163. (2) 12 zufällig ausgewählte Items aus dem gesamten Itempool der Standardskalen (n = 120) ergaben eine untere Grenze der Reliabilität von α = .27.

Zum Dilemma von Kontinuität und Vergleichbarkeit versus Verbesserung durch Revision sind als die bedeutendsten Vorbilder zu nennen: das *Minnesotas Multiphasic Personality Inventory* MMPI und die Entwicklung des *Maudsley Medical Questionnaire* MMQ zum *Maudsley Personality Inventory* MPI, *Eysenck Personality Inventory* EPI, *Eysenck Personality Questionnaire* EPQ (vgl. die Übersicht in Schmidt-Atzert & Amelang, 2012). Aus der ersten Etappe der Entwicklung des FPI bleibt die Erinnerung, dass bei Revisionen auch ein möglicher Schaden zu bedenken ist. Für die herausragende *Zürich Studie*, eine von der Psychiatrischen Universitätsklinik organisierte Kohortenstudie, wählten Jules Angst und Mitarbeiter das FPI in der ersten Version aus. Die Zürich-Studie, ein vielseitiges Projekt mit verschiedenen Patientengruppen sowie repräsentativ erhobenen Vergleichsgruppen, führte zu einer Serie wichtiger Publikationen (u.a. Angst, Dobler-Mikola & Binder, 1984; Hengartner, Tyrer, Ajdacic-Gross, Angst & Rössler, 2018) – hinsichtlich des FPI jedoch nur teilweise mit dessen heutigen Skalen vergleichbar. Insofern könnte es zur Qualitätssicherung gehören, dass die Testautoren bei einer eingreifenden Revision wenigstens eine Kombination von Marker-Items entwickeln, um den Anschluss an solche seltenen Forschungsvorhaben zu erleichtern.

2.3 Integration von Validitätshinweisen

Von den Autoren eines eingeführten Tests ist zu erwarten, dass bei jeder Neuauflage mindestens eine Fortschreibung der Validitätshinweise erfolgt und in adäquaten Abständen eine neue Normierung, die mit der Kontrolle der Testkonstruktion verbunden wird. Diese „Pflege“ ist umso mehr erforderlich, wenn ein Test relativ weit verbreitet ist.

Die Chancen der dritten bevölkerungsrepräsentativen Normierung des FPI wurden erneut genutzt, um durch schriftliche Zusatzfragen und durch Mitwirkung der Interviewer weitere Daten und Validierungshinweise zu gewinnen:

- 24 soziodemografische Merkmale, einschließlich Bildungsabschluss, Schichtzugehörigkeit, Kirchenbindung, aufgrund des standardmäßigen mündlichen Interviews (IfD);
- Angabe der Parteipräferenz (Selbstbericht);
- Berufliche Belastung, u.a. Anforderungen und Dauer, und Gesundheitszustand, u.a. Arztbesuche, chronische Krankheit, Psychotherapie, Medikamente (Selbstberichte);

- Politisch-soziale Einstellungen und religiös-weltanschauliche Einstellungen mit abgeleiteten Miniskalen bzw. Aggregaten, u.a. autoritär-konservative Einstellung, ideelle Einstellung hinsichtlich Gottesglaubens, Religiosität und Sinnfragen (Selbstbeurteilungen);
- Gegenwärtige Sorgen und Ängste (Selbstbeurteilungen);
- verhaltensnahe Frage hinsichtlich Aggressivität, Extraversion, Leistungsorientierung, Beanspruchung und Soziale Orientierung, z.B. erfahrene körperliche Angriffe seit dem Alter von 16 Jahren (Anzahl), Kontakt mit Freunden in der Freizeit (Häufigkeit), leistungsorientierter Sport, tatsächlich gegebene Spenden (Selbstberichte).
- Eindruck vom Verhalten der Befragten auf acht 4-stufigen Skalen, z. B. wie selbstsicher, freundlich, lebhaft jemand wirkte (Verhaltensbeurteilungen durch die Interviewer);
- Einstufungen der Wohnung und Wohnumgebung sowie Beobachtungen hinsichtlich des Körperschmucks, d.h. der Anzahl sichtbarer Tätowierungen, Piercings und Ringe (Beobachtungen der Interviewer).

Es sind Daten unterschiedlicher Quellen bzw. Datenebenen, wobei zu unterscheiden ist zwischen Selbstbeurteilungen der Befragten und Selbstberichten, die im Prinzip objektiviert werden könnten, beispielsweise die Antworten hinsichtlich beruflicher Belastung oder Psychotherapie. Die Beurteilungen und Beobachtungen durch die Interviewer bilden eine besondere Datenklasse. Den Interviewern wurden keine Hinweise auf typische Methodenprobleme solcher Beobachtungen gegeben. Zweifellos werden diese Einstufungen durch sprachlich-semantische Schwierigkeiten und durch Beurteiler-Tendenzen beeinflusst sein, doch ist angesichts der sehr großen Anzahl von ca. 300 erfahrenen Interviewern anzunehmen, dass sich solche Tendenzen teilweise „ausmitteln". Jedenfalls können auf diese Weise alltagsnahe psychologische Informationen zusätzlich gewonnen werden, wie sie sonst bei der Konstruktion und Normierung psychologischer Tests nicht möglich sind.

Die hypothesengeleiteten und die explorativen Analysen lieferten psychologisch wichtige Beiträge zur Validierung, die auch in dieser Kombination von Datenquellen neu sind. Die Ergebnisse sind im Testmanual tabellarisch und graphisch dargestellt, teils auch durch die *Aggregation* ausgewählter Kriterieninformation weitergeführt. Beispiele sind:

- Gesundheitszustand, d.h. ein Aggregat aufgrund der angegebenen Häufigkeit von Arztbesuchen, Psychotherapie, Medikamenten, mit den FPI-Testwerten *Körperliche Beschwerden* Effektstärke ES = +0.82; Emotionalität +0.65; Lebenszufriedenheit –.52.
- Verhaltensaussage „bereits Opfer körperlicher Angriffe gewesen" mit dem FPI-Testwert *Aggressivität* ES = +1.82.
- Autoritarismus und AfD-Parteipräferenz mit dem FPI-Testwert *Soziale Orientierung* ES = –0.80.
- Ideelle Einstellung, d.h. ein Aggregat aufgrund der Aussagen über Gottesglauben, Religiosität, Sinnfragen, mit dem FPI-Testwert *Soziale Orientierung* ES = +0.58.
- Verhaltensbeurteilung „gehemmt" durch den Interviewer mit den FPI-Testwerten *Gehemmtheit* ES = +0.87; Leistungsorientierung = +0.85; Extraversion –0.80.

Die Sekundärliteratur mit empirischen und methodischen Beiträgen zum FPI ist heute nicht mehr überschaubar. Die FPI-Bibliografien wurden für insgesamt 655 Arbeiten bis zum Jahr 1994 geführt (Potreck, 1983; Jehle & Fahrenberg, 1994). Diese Dokumentation wird heute durch die Literaturbanken PubPsych, PsycINFO, MEDLINE u.a. geleistet, jedoch nur unvollständig, denn die unveröffentlichten Diplom-, Master- und Doktorarbeiten fehlen. Im Manual der 9. Auflage werden neben früheren, aber herausragenden Beiträgen insbesondere jene referiert, die zur Persönlichkeitsforschung und zu zentralen Anwendungsgebieten, insbesondere Klinische Psychologie, Psychotherapieforschung einschließlich Sozialtherapie, wichtig erscheinen. Hervorgehoben wurden auch Studien, die eine Verbindung zwischen Persönlichkeitsskalen und Ergebnissen des Ambulanten Assessment herstellen oder ungewöhnliche Befunde enthalten, wie die FPI-Daten von auffälligen Personen mit extremistischer Orientierung.

In zahlreichen Untersuchungen wurden Unterschiede in den FPI-Testwerten bestimmter Personengruppen beschrieben. Diese Gruppierungen erfolgten je nach Fragestellung der Untersucher aufgrund von bestimmten Statusmerkmalen, psychosozialen oder klinisch-psychologischen Merkmalen. Demgegenüber sind Publikationen über die Vorhersage praktisch relevanter Kriterien, über die Evaluation individueller Begutachtungen oder Evaluation von Klassifikations- und Selektionsentscheidungen in bestimmten Anwendungsfeldern selten, und anspruchsvolle persönlichkeitspsychologische Grundlagenforschung zu den theoretischen Konstrukten (Persönlichkeitsdispositionen), ihrer funktionellen Dynamik und zur biographischen Entwicklung fehlen weitgehend. Oft handelt es sich um kleine Personengruppen, oft um eine Gelegenheitsauswahl, d.h. mit unklarer Präselektion, oder es mangelt an einer äquivalenten Kontrollgruppe. Falls wichtige Befunde mitgeteilt werden, fehlen in der Regel Replikationen. Die

erwähnte Züricher Kohortenstudie ist auch durch ihren Bezug auf ein sehr breites Spektrum klinisch-psychologisch relevanter Befunde eine weit herausragende Forschung.

Insgesamt mangelt es jedoch an methodisch anspruchsvolleren und zugleich praxisnahen Untersuchungen der externen Validität im Hinblick auf wichtige, diagnostisch und prognostisch wichtige Kriterien bzw. den tatsächlichen „Entscheidungsnutzen". Eine *vertiefte Qualitätskontrolle* und Evaluation dieser Art übersteigt die Möglichkeiten der meisten Testautoren, so dass Kooperationen notwendig sind.

2.4 Künftige Evaluationen

Die Frage nach geeigneten Forschungsprojekten zur Kriterienvalidierung von FPI-Skalen, insbesondere im klinisch-psychologischen, rehabilitationspsychologischen und sozialtherapeutischen Bereich liegt nahe. Die Testautoren waren durchaus motiviert, sich weiterhin dieser Aufgabe zu stellen. Zwei relativ große Projekte scheiterten, trotz der Kooperationsbereitschaft der Ärzte und Psychologen in den betreffenden Kliniken und der bereits gesicherten Finanzierung, an der Universitätsbürokratie und an einem Datenschutzbeauftragten und dessen zusätzlicher Forderung, die Projektleiter dürften nicht mit den eingestellten Projektmitarbeitern sprechen, da andernfalls die Anonymität der von diesen Mitarbeitern untersuchten Patienten nicht sicher gewährleistet sei. Projekte dieser Art können heute wahrscheinlich nur noch nur aus den großen Institutionen der Praxisfelder (Kliniken, soziale Einrichtungen, 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 zu gewinnen und die Prinzipen der modernen Assessmenttheorie adäquat anzuwenden.

Das Drängen auf gründlichere externe Kriterienvalidierung kann im Zusammenhang mit der erwähnten Umfrage unter Mitgliedern des Berufsverbandes BDP (Roth u.a., 2010) gesehen werden und der vorrangigen Forderung, dass in der universitären Diagnostikausbildung eine kritische Diagnostik gelehrt wird. Gerade die so häufig verwendeten Persönlichkeitsfragebogen verlangen 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 Testkonstruktion übersehen werden.

Die kritische Diskussion der Vorzüge und der Probleme von Persönlichkeitsfragebogen gehören nach der Auffassung der Autoren auch in das 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. Auch eine Einführung in wichtige Prinzipien der modernen Assessmenttheorie ist notwendig – in Erinnerung an Cattells (1973) *Personality and Mood by Questionnaires*. Wichtige Leitkonzepte der modernen Assessmenttheorie werden an dieser Stelle nur genannt: Multitrait-Multimethod, Multimodale Strategie, Generalisierbarkeitstheorie, Aggregation und Brunswiks Linsenmodell, Wittmanns Multivariate Reliabilitätstheorie. Hinzu kommt die Frage nach der ökologischen (externen) Validität von Testergebnissen hinsichtlich der tatsächlich im Alltag registrierten Indikatoren des individuellen Verhaltens und Befindens mit der Methodik des *Ambulanten Assessment* (Ecological Momentary Assessment, Experience Sampling Method), technisch unterstützt durch miniaturisierte, ggf. auch interaktive Systeme. Diese innovative Methodik, die in wichtigen Ansätzen in den deutschsprachigen Ländern entwickelt wurde, ist auf einigen Gebieten bereits die Methode der Wahl gegenüber den natürlich viel einfacheren Persönlichkeitsfragebogen. Die Methodik und Anwendung des *ambulanten Assessment* haben sich international mit jährlich mehr als 500 Publikationen zu einer breiten und höchst aktiven Forschungsrichtung entwickelt, die jedoch in Lehrbüchern der Differentiellen Psychologe und der psychologischen Diagnostik auch nach 40 Jahren noch kaum rezipiert ist.

Diese Informationen als Grundlage kritisch-methodenbewusster Diagnostik sind heute noch wichtiger, da im Vergleich zum früheren Diplom-Studiengang wahrscheinlich nur ein Teil der Absolventen eines Psychologie-Studiums (mit dem speziellem Vertiefungsgebiet) über eine hinreichende diagnostische Ausbildung und Gutachtenpraxis unter Supervision, einschließlich Testkonstruktion und Einführung in die wichtigsten Prinzipien der modernen Assessmenttheorie verfügen wird.

2.5 Rezensionen und Kommentare, Datensätze open access

Zum FPI wurden im Laufe der Jahre 15 Testrezensionen, außerdem eine Anzahl von Reviews in Lehrbüchern und in Handbuchbeiträgen verfasst. Rezensionen tragen zur Qualitätssicherung bei, insbesondere wenn sie explizit methodenkritisch und nicht pauschal formuliert sind. Möglichst prägnante Testrezensionen, die bestehende Defizite hervorheben, könnten tatsächlich eine motivierende Wirkung auf Testautoren und Testverlag haben. – Andererseits gab es in einigen der FPI-Rezensionen und Reviews sachliche Fehler und Missverständnisse. So wurde die

Skalenkonstruktion unter anderem als „faktorenanalytisch“, „blind faktorenanalytisch“ oder „willkürlich“ beurteilt statt das hypothetisch-deduktive und empirisch-induktive Verfahren nachzuvollziehen, dem bereits Eysenck in der Entwicklung der Dimensionen Neuroticism und Extraversion-Introversion hinsichtlich klinischer Kriterien folgte (Eysenck, 1950). Die beiden Dimensionen bilden noch immer das Rückgrat der meisten heutigen Persönlichkeitsinventare. Die erwähnten Rezensionen motivierten dazu, in der Neuauflage des Testmanuals vier verschiedene Konstruktionsstrategien zu beschreiben: die hauptsächlich kriterienbezogene Strategie, die auf eine spezielle Persönlichkeitstheorie bezogene deduktive Strategie, die lexikalisch-induktive Strategie und die kombinierte, hypothetisch-deduktive und empirisch-induktive Strategie.

Eine bessere fachliche Kooperation zwischen den Testrezensenten und den Testautoren ist wünschenswert. Selbst wenn den Testautoren von der Redaktion intern die Möglichkeit einer Stellungnahme gegeben wird, könnte eventuell ein gedruckter Kommentar der wechselseitigen Qualitätskontrolle dienlich sein.

Zum FPI wurden seit der 1. Auflage Datensätze zu Vergleichszwecken zur Verfügung gestellt. Sie stammten aus eigenen Untersuchungen oder konnten von anderen Autoren anonymisiert zur Verfügung gestellt werden. Für solche Vergleichszwecke und für kritische Reanalysen oder Unterrichtszwecke sind die Datensätze der Repräsentativerhebungen zur 4. und 7. Auflage des FPI-R in ZPID PsychData archiviert und können dort aufgrund eines Datennehmer-Antrags mit dem ZPID angefordert werden. Diese Datensätze enthalten eine beträchtliche Anzahl von zusätzlichen psychologischen und sozioökonomischen Daten. Auch für den aktuellen Datensatz der 9. Auflage ist diese open access Regelung vorgesehen. Rückblickend ist jedoch deutlich, dass die einzelnen Datensätze nur sehr selten angefordert wurden; eine gezielte Umfrage bei den Datennehmern hinsichtlich der Normierungs-Datensätze ergab eine nur minimale Nutzung, am ehesten noch für Unterrichtszwecke.

3. Ausblick

Ist „Testpflege“ ein zusätzliches Gütekriterium? Die Normierung von Persönlichkeitsfragebogen sollte regelmäßig überprüft werden wie es in der produktiven Kooperation mit dem Testverlag und einem auf Umfragen spezialisierten Institut zu leisten ist. Eine Daumenregel hinsichtlich des adäquaten Zeitintervalls kann nicht vorgeschlagen werden, denn es mangelt noch an weiteren Erfahrungsberichten über die Effektstärke solcher Trends. In der „Testpflege“ erfordern bevölkerungsrepräsentative Normierungen den größten finanziellen und zeitlichen Aufwand. Deshalb ist es zweckmäßig, in einer Repräsentativerhebung mehrere Fragebogen, je nach Länge und Inhalt, zu kombinieren. An eine fachliche Kooperation ist bereits beim kompetenten Entwurf eines mehrdimensionalem Persönlichkeitsfragebogen zu denken, um die speziellen Erfahrungen aus den verschiedenen Bereichen der Persönlichkeitsforschung zu verbinden. Methodisch überzeugende Validitätsstudien und Evaluationen des Entscheidungsnutzens erfordern gemeinsame Projekte von Testautoren und von Psychologen, die in größeren Organisationen tätig sind und die im Interesse dieser Organisationen anonymisierte Kriterieninformationen verwenden dürfen. Zunehmend wichtig sind auch die mit der Methodik des ambulanten Assessment unter Alltagsbedingungen durchgeführten Studien. Auch solche Projekte sind ohne Kooperation kaum möglich.

Künftige Neukonstruktionen von mehrdimensionalen *Persönlichkeitsfragebogen* sollten sich grundsätzlich nicht mit einer Datenreduktion auf wenige Sekundärfaktoren oder mit einem angeblich universellen System mit minimaler Anzahl von Skalen begnügen. Die eigenschaftstheoretisch, im deduktiv-induktiven Verfahren oder primär kriterienorientiert konstruierten Skalen haben den Vorzug, dass sie eventuell sogar in Modulen – näherungsweise wie bei einem Intelligenztest – zusammengesetzt werden könnten, um dem Anwender eine aufgabenrelevante Auswahl zu ermöglichen. Wegen der erforderlichen Forschungskompetenz in vielen Bereichen und wegen des hohen Arbeits- und Kostenaufwands wären solche Vorhaben nur auf eine neue Weise kooperativ möglich – wie es sich in der Labormethodik der Medizin und Naturwissenschaften bewährt hat.

Literatur

Angst, J., Dobler-Mikola, A. & Binder, J. (1984). The Zurich Study – A prospective epidemiological study of depressive, neurotic and psychosomatic syndromes. I. Problem, Methodology. *European Archives of Psychiatry and Neurological Sciences*, 234, 13-20.

Bengel, J. & Wittmann, W. W. (1983). Urteilsfehler aufgrund zeitlicher Veränderungen von Testnormen. *Diagnostica*, 29, 101–117.

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

Cattell, R. B. (1973). *Personality and mood by questionnaire*. San Francisco: Jossey-Bass.

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

Fahrenberg, J., Selg, H. & Hampel, R. (1983). Die bevölkerungsrepräsentative Normierung des Freiburger Persönlichkeitsinventars FPI-A*. Diagnostica*, 29, 336–343.

Fahrenberg, J., Hampel, R. & Selg, H. (1984–2001). *Das Freiburger Persönlichkeitsinventar (revidierte Fassung FPI-R und teilweise geänderte Fassung FPI-A1,* 4. Aufl. 1984; ergänzte 5. Aufl. 1989, ergänzte 6. Aufl. 1994, revidierte 7. Auflage 2001; 1. bis 3. Aufl. siehe Fahrenberg & Selg). Göttingen: Hogrefe.

Fahrenberg, J., Hampel, R. & Selg, H. (2020, im Druck). *Freiburger Persönlichkeitsinventar FPI-R. Neue Normierung und Validitätshinweise, Prinzipien der Testkonstruktion und modernen Assessmenttheorie.* (9. erweiterte Aufl.). Göttingen: Hogrefe.

Fahrenberg, J. & Selg, H. (1970). *Das Freiburger Persönlichkeitsinventar FPI* (1. Auflage 1970, 2. erweiterte Aufl. 1973, ergänzte 3. Aufl. 1978). Göttingen: Hogrefe.

Hengartner, M. P., Tyrer, P., Ajdacic-Gross, V., Angst, J. & Rössler, W. (2018). Articulation and testing of a personality-centered model of psychopathology: evidence from a longitudinal community study over 30 years. *European Archives Psychiatry & Clinical Neuroscience*, 268 (5), 443–454.

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

International Test Commission (2013). ITC Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores. https://www.intestcom.org/files/guideline_quality_control.pdf

International Test Commission (2015). The ITC Guidelines on Practitioner Use of Test Revisions, Obsolete Tests, and Test Disposal. https://www.intestcom.org/files/guideline_test_disposal.pdf

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

Jehle, M. & Fahrenberg, J. (1994). *Literaturverzeichnis zum Freiburger Persönlichkeitsinventar FPI 1983–1993*. (Forschungsberichte des Psychologischen Instituts der Albert-Ludwigs-Universität Nr. 105). Freiburg i. Br.: Psychol. Institut. http://www.jochen-fahrenberg.de/uploads/media/FPI_Bibliographie_1983-1993_01.pdf

Kersting, M. (2018a). Qualitätssicherung und -optimierung in der Eignungsdiagnostik. In: Diagnostik- und Testkuratorium (Hrsg.) *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430* (S. 2-20). Berlin: Springer. (DOI 10.1007/978-3-662-53772-5)

Kersting, M. (2018b). Zur Information über und Dokumentation von Instrumenten zur Erfassung menschlichen Erlebens und Verhaltens – Die DIN SCREEN Checkliste 1, Version 3. In: Diagnostik- und Testkuratorium (Hrsg.) *Personalauswahl kompetent gestalten: Grundlagen und Praxis der Eignungsdiagnostik nach DIN 33430 (S. 223-244)*. Berlin: Springer. http://kersting-internet.de/pdf/Kersting_2017_in_DTK_DIN-SCREEN_Checkliste_1.pdf

Obschonka, M., Wyrwich, M., Fritsch, M., Gosling, S. D., Rentfrow, P. J. & Potter, J. (2019). Von unterkühlten Norddeutschen, gemütlichen Süddeutschen und aufgeschlossenen Großstädtern: Regionale Persönlichkeitsunterschiede in Deutschland. *Psychologische Rundschau*, 70(3), 173-194.

Pfeifer, C., Baier, D. & Kliem, S. (2018). *Zur Entwicklung der Gewalt in Deutschland. Schwerpunkte: Jugendliche und Flüchtlinge. Zentrale Befunde eines Gutachtens im Auftrag des Bundesministeriums für Familie, Senioren, Frauen und Jugend*. https://www.zhaw.ch/storage/shared/sozialearbeit/News/gutachten-entwicklung-gewalt-deutschland.pdf

Potreck, F. (1983). *Kommentiertes Literaturverzeichnis zum Freiburger Persönlichkeitsinventar FPI. Veröffentlichungen seit 1978*. Forschungsberichte des Psychologischen Instituts der Universität Freiburg, Nr. 11. http://www.jochen-fahrenberg.de/uploads/media/FPI_Bibliographie_seit_1978_01.pdf

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

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

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, S. 199

Wagner, J., Lüdtke, O. & Robitzsch, A. (2019). Does personality become more stable with Age? Disentangling state and trait effects for the Big Five across the life span using local structural equation modelling. *Journal of Personality and Social Psychology*, 116(4), 666-680.

Wirtz, M.A. (2019). Qualitätssicherung. In: M.A. Wirtz (Hrsg.). *Dorsch. Lexikon der Psychologie* (S. 1465). (19. überarbeitete Aufl.). Göttingen: Hogrefe.