WOGE

Unvertretbar nach 40 Jahren Anwendung?

Meinungen über MMPI-2 gehen weit auseinander

Das Testkuratorium hat bei Dr. Petra Hank und Prof. Dr. Peter Schwenkmezger von der Uni Trier eine Testbesprechung des Minnesota Personality Inventory-2 in Auftrag gegeben. Die beiden Wissenschaftler kommen in ihrer kritischen Wertung zu dem Ergebnis, dass das Verfahren, das seit mehreren Jahrzehnten in Deutschland angewendet wird, unvertretbar sei. In einer Stellungnahme reagiert Prof. Dr. Rolf R. Engel, Klinikum der Uni München und Mitherausgeber einer revidierten, deutschsprachigen Fassung des MMPI-2, auf diese Einschätzung. Report Psychologie veröffentlicht beide Texte.

Das Minnesota Personality Inventory–2 (MMPI)* Testbesprechung im Auftrag des Testkuratoriums

Petra Hank & Peter Schwenkmezger

Die revidierte, deutschsprachige Fassung des Minnesota Multiphasic Personality Inventory (MMPI-2), herausgegeben von Hathaway, McKinley und Engel (2000), wird einer kritischen Rezension unterzogen. Ausgehend von den Zielsetzungen des MMPI-2 wird der Aufbau des Verfahrens, differenziert nach den Skalensets Validitäts-, Basis-, Inhalts- und Zusatzskalen beschrieben. Entstehungshintergrund und Nachvollziehbarkeit der Konstruktion des MMPI-2 werden vor dem Hintergrund der gegenwärtigen psychologisch-diagnostischen Praxis diskutiert. Unter den Aspekten Objektivität, Transparenz, Zumutbarkeit, Verfälschbarkeit und Störanfälligkeit wird die Durchführung des Inventars bewertet. Es folgen Überlegungen zu Verwertung und Evaluation des MMPI-2 unter Berücksichtigung der klassischen Haupt- und Nebengütekriterien. Der Beitrag schließt mit einer kritischen Würdigung des Verfahrens.

* In der deutschen Überarbeitung von Rolf R. Engel (2000)

1. Einleitung

Mit dem Ziel, psychische Störungen ähnlich wie in einem psychiatrischen Interview, aber ökonomischer zu erfassen, publizierten Hathaway und McKinley 1940 das Minnesota Multiphasic Personality Inventory als eines der ersten mehrdimensionalen Fragebogeninventare. Als MMPI-Saarbrücken wurde es von Spreen (1963) für deutschsprachige Verhältnisse adaptiert. Überarbeitet und neu normiert gaben Hathaway, McKinley und Engel 2000 die Nachfolgeversion, den MMPI-2, heraus. Mit dem apriori dimensionalisierten Fragebogen (Kubinger, 1995) sollen alle wesentlichen Persönlichkeitsbereiche psychometrisch erfasst werden können. Über die ursprüngliche Zielsetzung der psychiatrischen Kategorisierung hinaus, erhebt das Verfahren zwischenzeitlich den Anspruch, auch für eignungsdiagnostische Fragestellungen tauglich zu sein.

Nach Einschätzung von Amelang und Zielinski (1997) handelt es sich beim Minnesota Multiphasic Personality Inventory (MMPI) um das weltweit gebräuchlichste Persönlichkeitstestsystem. Jährlich erscheinen rund 1000 anwendungsbezogene Forschungsuntersuchungen zum MMPI, und zwar insbesondere an klinisch auffälligen Gruppen.

2. Testgrundlage

2.1 Diagnostische Zielsetzungen und Aufbau

Der MMPI-2, die Revision des MMPI-Saarbrücken, ist ein »Breitbandverfahren zur Beschreibung wichtiger Persönlichkeitseigenschaften und psychischer Störungen« (S. 1). Ziel ist es, zu einer kohärenten, psychometrischen Erfassung der Persönlichkeitsdynamik und ggf. zu einer abschließenden Stellungnahme zu gelangen. Anwendungskontexte sind gemäß dem Manual u.a. medizinische und psychiatrische Beurteilungen sowie die Personalauswahl. Der MMPI-2 kann als Einzel- oder Gruppentest vorgegeben werden.

Zum Aufbau: Der MMPI-2 umfasst insgesamt 567 dichotome Items, die im Hinblick auf das Zutreffen auf die eigene Person mit »richtig« oder »falsch« zu beantworten sind. Die Items verteilen sich auf Validitäts-, Basis-, Inhalts- und Zusatzskalen.

Die Validitätsskalen

Zu ihnen gehören die klassischen Skalen wie Lügen-Skala [L], Seltenheits-Skala [F] und Korrektur-Skala [K] mit 15, 60 und 30 Items. Die rational konstruierte für die Testperson (Tp) transparent gehaltene *L-Skala* misst die Neigung, sich in der Testsituation zu verstellen und marginale Charakterschwächen zu leugnen, um als ideale Persönlichkeit zu erscheinen. Die Testautoren heben hervor, dass die Skala keine allgemeine Tendenz zur Lüge misst, sondern einen von mehreren Indizes dafür liefert, ein Testprotokoll durch eine bestimmte Antworttendenz zu entstellen. Sie schließen: »Deutlich erhöhte Werte (T-Wert > 70) spiegeln mit hoher Wahrscheinlichkeit ein durchgängiges Testverhalten wider, das die normale Aussagekraft der klinischen Skalen negativ beeinflusst« (S. 26).

Die *F-Skala* enthält Items, die von knapp 10% der Tpn der Normierungsstichprobe in Schlüsselrichtung beantwortet wurden. Sie dient zur Identifikation von Tpn mit zufälligem Antwortverhalten, verursacht durch mangelnde Testmotivation, intellektuelle Einschränkung, gravierende Leseschwäche oder fehlenden Realitätskontakt. Bei gewissenhafter Bearbeitung und klinisch-unauffälligem Antwortprotokoll sollte der zugehörige T-Wert < 55 sein.

Die Items der *K-Skala* thematisieren Eigenschaften, die viele Personen von sich und ihrer Familie leugnen. Hohe Scores auf dieser Skala »können… daher die Tendenz wiedergeben, auf subtile Art und Weise Antworten so zu wählen, dass sie möglichst wenig auf psychische Probleme hinweisen« (S. 27).

Die Differenz zwischen dem Rohwert der F- und dem der K-Skala, der sog. *F-minus-K-Index*, offenbart Simulationstendenzen der Tpn. Übersteigt diese Differenz einen T-Wert > 40, rät der Autor von einer Interpretation des Testprofils ab.

Weiteren Aufschluss über die Testmotivation geben die neu eingeführten Validitätsindikatoren Back-F-Wert, (FB-Wert, 40 Items), Beantwortungsinkonsistenz (VRIN, 67 Item-Antwort-Paare) und Zustimmungstendenz (TRIN, 23 Item-Antwort-Paare). Analog zur F-Skala sagt der Wert der F_B*-Skala*, deren Items im letzten Drittel des Fragebogens platziert sind, etwas über die Verlässlichkeit des Antwortverhaltens im Hinblick auf die Zusatzskalen aus.

Inkonsistentes Antwortverhalten sowie die Tendenz zur Zustimmung ungeachtet des Iteminhalts decken die *VRIN*- resp. *TRIN*-Skala auf, die jeweils aus Itempaaren bestehen. Hohe VRIN-Werte (Rohwerte ≥ 17) sowie auffällig hohe (Rohwerte ≥ 14) bzw. niedrige (Rohwerte ≤ 4) TRIN-Werte lassen auf ein wahlloses Antwortverhalten und damit ein ungültiges Testprotokoll schließen. Diese Skalen sind als Ergänzung der klassischen Validitätsskalen gedacht und »sollten vorsichtig angewendet werden bis mehr empirische Evidenz vorliegt« (S. 29). Im Übrigen wird empfohlen, alle Validitätsindikatoren im Kontext der Biographie und gegenwärtigen Lebenssituation der Tpn zu bewerten.

Die klinischen Skalen

Zu ihnen zählen die zehn Skalen Hypochondrie (Hd), Depression (D), Hysterie/Konversionsstörung (Hy), Psychopathie, Soziopathie, antisoziale Persönlichkeitsstörung (Pp), männliche/weibliche Interessen (Mf), Paranoia (Pa), Psychasthenie (Pt), Schizophrenie (Sc), Hypomanie (Ma) und soziale Introversion (Si), für die sich mit Ausnahme der Skalen Hd, Mf und Si verschiedene Subskalen bilden lassen (s. Tab. 1): Nach Wiener und Harmon (1946) können die Items der einzelnen Skalen in nicht subtile vs. subtile Teilskalen zusammengefasst werden, je nachdem, ob sie für die Tpn offensichtlich bzw. durchschaubar sind oder nicht. Darüber hinaus liegen nach Harris und Lingoes (1955) für die einzelnen Skalen vier bis sechs inhaltlich homogene Teilskalen vor.

Die *Hd-Skala* misst mit 32 Items eine allgemeine Besorgtheit um die körperliche Gesundheit und die Beschäftigung mit der eigenen Person.

Tabelle 1: Übersicht zu den klinischen Skalen des MMPI-2

Skalenname	Itemzahl	Subskalen nach Wiener und Harmon (1946)	Subskalen nach Harris und Lingoes (1955)
Hypochondrie (Hd)	32	Nein	---
Depression (D)	57	Ja	1. Niedergeschlagenheit 2. Psychomotorische Verlangsamung 3. Körperbeschwerden 4. Geistige Leere 5. Grübelei
Hysterie/ Konversionsstörung (Hy)	60	Ja	1. Leugnung sozialer Ängste 2. Bedürfnis nach Zuneigung 3. Unpässlichkeit 4. Körperbeschwerden 5. Agressionshemmung
Psychopathie (Pp), Soziopathie, antisoziale Persönlichkeitsstörung	50	Ja	1. Familiäre Disharmonie 2. Autoritätsprobleme 3. Unbeirrbarkeit durch soziale Probleme 4. Soziale Entfremdung 5. Mangelndes Selbstvertrauen
männliche/weibliche (Mf-m) Interessen (Mf-f)	56	Nein	---
Paranoia (Pa)	40	Ja	1. Verfolgungsgedanken 2. Sensitivität 3. Naivität
Psychasthenie (Pt)	48	Ja	---
Schizophrenie (Sc)	78	Ja	1. Mangelndes Vertrauen in andere 2. Inadäquater Affekt 3. Ich-Mangel im Denken 4. Ich-Mangel im Wollen 5. Ich-Mangel im Sinne von Hemmungsverlust 6. Bizarre Sinneswahrnehmungen
Hypomanie Ma	46	Ja	1. Mangelnde Moral 2. Antriebssteigerung 3. Unerschütterlichkeit 4. Größenwahn
soziale Introversion Si	69	Nein	---

T-Werte $\geq$ 65 auf der D-Skala (57 Items) weisen laut Autoren auf eine depressive Störung hin. Über das klinische Bild der Depression – mit Gefühlen der Trauer und Niedergeschlagenheit, Hilflosigkeit, Unzulänglichkeit und Hoffnungslosigkeit – hinaus, erfasst diese Skala auch die für Depressive typischen Persönlichkeitszüge wie Überverantwortlichkeit, hohes Anspruchsniveau und Intrapunitivität. Nach Harris und Lingoes (1955) lassen sich die Items dieser Skala den fünf homogenen Inhaltsbereichen Niedergeschlagenheit, psychomotorische Verlangsamung, Körperbeschwerden, geistige Leere und Grübelei zuordnen.

Die *Hy-Skala* (60 Items) identifiziert laut Autoren Tpn mit sensorischen oder motorischen Störungen ohne organische Grundlage. Ihre Items lassen sich gemäß Harris und Lingoes (1955) in die Komponenten Leugnung sozialer Ängste, Bedürfnis nach Zuneigung, Unpässlichkeit, Körperbeschwerden und Aggressionshemmung gliedern.

Psychopathische, soziopathische und antisoziale Persönlichkeiten weist die *Pp-Skala* mit 50 Items aus. Kennzeichnend für Tpn mit hohen Scores auf dieser Skala ist die Missachtung sozialer und moralischer Verhaltensregeln. Die fünf Harris-Lingoeschen Inhaltskomponenten lauten: familiäre Disharmonie, Autoritätsprobleme, Unbeirrbarkeit durch soziale Probleme, soziale Entfremdung und mangelndes Selbstvertrauen.

Die *Mf-Skala* (Mf-m bzw. Mf-f mit je 56 Items), »nur noch aus traditionellen Gründen im MMPI-2 enthalten« (S. 33), erhebt geschlechtsspezifische Gefühle, Interessen und Einstellungen zum Arbeitsleben, sozialen Beziehungen und Hobbies. Ihre ursprüngliche Messintention, die Geschlechtsrollenorientierung zu diagnostizieren, wird durch die beiden neuen Skalen GM und GF realisiert (s.u.).

Mit 40 Items erhebt die *Pa-Skala* Misstrauen, Feindseligkeit im zwischenmenschlichen Umgang, Selbstbezogenheit und Unsicherheit. Für diese Skala können die Subskalen Verfolgungsgedanken, Sensitivität und Naivität nach Harris-Lingoes gebildet werden.

Die 48 Items der *Pt-Skala* beinhalten »eine allgemeine Angst und Verzweiflung, einen hohen moralischen Anspruch, Selbstbeschuldigungen für Misserfolge und harte Bemühungen um Impulskontrolle« (S. 34).
Die Iteminhalte der *Sc-Skala* (78 Items) spiegeln die Bandbreite schizophrener Symptome wider. Sie lassen sich den sechs Harris-Lingoes Subskalen mangelndes Vertrauen auf andere, inadäquater Affekt, Ich-Mangel im Denken, Ich-Mangel im Wollen, Ich-Mangel im Sinne von Hemmungsverlust und bizarre Sinneswahrnehmungen zuordnen.
Die *Ma-Skala* (46 Items) erfasst manische und hypomanische Charakteristika. Mangelnde Moral, Antriebssteigerung, Unerschütterlichkeit und Größenwahn ergeben sich als inhaltliche Teilskalen nach Harris und Lingoes.
Ein Hang zur Eigenbrötelei, soziale Schüchternheit und Mangel an Durchsetzungsfähigkeit sind kennzeichnend für Tpn mit unterdurchschnittlichen Werten auf der *Si-Skala* (69 Items). Umgekehrt nehmen Tpn mit Werten über dem Mittelwert aktiv am Gesellschaftsleben teil und verhalten sich sozial geschickt.
Die Items der Validitäts- und Basisskalen machen zusammen die ersten 370 des Testinventars aus.

Die Inhaltsskalen
Insgesamt 366 Items steuern die 15, im Vergleich zum MMPI-Saarbrücken zum Teil neuen, rational entwickelten Inhaltsskalen zum MMPI-2 bei, unter ihnen die *Angst-Skala* (ANX-Skala). Sie erhebt mit 23 Items generelle Angstsymptome, »also Gespanntheit, körperliche Beschwerden, wie Herzklopfen und Atemnot, Schlafstörungen, Sorgen und Konzentrationsschwäche« (S. 47). Spezifische angstauslösende Situationen, wie z.B. vor Tieren, Höhen oder auch dem Verlassen der Wohnung misst die Skala *Phobien* (FRS-Skala, 23 Items). Eine weitere inhaltliche Dimension, *Zwanghaftigkeit* (OBS-Skala), wird mit 16 Items abgebildet. Sie beinhaltet Entscheidungsschwierigkeiten und Grübeleien über alltägliche Dinge und Angelegenheiten. Die Skala *Depression* (DEP-Skala, 33 Items) stellt eine depressive Symptomatik fest. Somatische Beschwerden (u.a. gastrointestinale und kardiovaskuläre Symptome, Wahrnehmungsstörungen) sowie eine allgemeine Besorgnis um die körperliche Gesundheit erfasst die Skala *Körperbeschwerden* (HEA-Skala, 36 Items). Die Skala *Bizarre Angaben* (BIZ-Skala, 23 Items) erhebt gestörte Denkprozesse wie sie für psychotische Störungen charakteristisch sind. Mit 16 Items misst die Skala *Ärger* (ANG-Skala) die Fähigkeit zu einem kontrollierten Umgang mit Ärger. Die Skala *Zynismus* (Cyn-Skala, 23 Items) liefert Hinweise auf misstrauische und feindselige Personen. Die Skala Typ-A (Typ-A-Skala, 19 Items) identifiziert Personen vom Typ-A, die sich durch ein ausgeprägtes Konkurrenzverhalten, Ungeduld und Reizbarkeit auszeichnen. Personen mit geringem Selbstvertrauen und einer abwertenden Einstellung der eigenen Person gegenüber werden auf der Skala *Negatives Selbstwertgefühl* (Lse-Skala, 24 Items) auffällig. Zur Messung von *sozialem Unbehagen* (Sod-Skala) sind weitere 24 Items vorgesehen. *Familiäre* resp. *berufliche Schwierigkeiten* und Probleme erfassen die Fam-Skala bzw. Wrk-Skala mit 25 und 33 Items. Schließlich gibt die Skala *Negative Behandlungsindikatoren* (Trt-Skala, 26 Items) Hinweise auf negative Einstellungen gegenüber medizinischer oder psychotherapeutischer Behandlung.
Jede der genannten Skalen haben Ben-Porath und Sherwood (1993) in ihre Inhaltskomponenten differenziert. Im Ergebnis entstanden pro Skala bis zu vier der sog. Inhaltskomponentenskalen, die ebenfalls in der deutschen Form adaptiert wurden.

Die Zusatzskalen
Im Sinne einer Ergänzung und Erweiterung seines Anwendungsbereiches enthält der MMPI-2 darüber hinaus eine Reihe von zusätzlichen Skalen, unter ihnen die schon in der Vorläuferversion enthaltenen Skalen zur Messung von Angst, Verdrängung, Ich-Stärke und Allgemeine Suchtgefährdung. Zusätzlich wurden die Skalen *Überkontrollierte Feindseligkeit, Dominanz, Soziale Verantwortung, Psychische Probleme von Studierenden, Geschlechtsrollen-Skalen, Posttraumatische Belastungsstörungen, Eheproblem-Skala, Suchtpotential* und *Suchteingeständnis* neu gebildet. Da diese Skalen ausschließlich für Forschungszwecke gedacht sind und darüber hinaus (noch) keine empirischen Untersuchungen im deutschen Sprachraum vorliegen, wird auf ihre differenzierte Darstellung verzichtet.

2.2 Theoretische Grundlagen und Nachvollziehbarkeit der Testkonstruktion

Die theoretischen Überlegungen zum Entstehungshintergrund des MMPI-2 sind im Handbuch nicht enthalten. Demgegenüber betonen Hathaway, McKinley und Engel dessen Verankerung im amerikanischen Original und in seiner Vorläuferversion, dem MMPI-Saarbrücken. Überholte Normen und veraltete Items im Hinblick auf Inhalt und Formulierung benennen die Autoren als plausible Gründe für eine Neuauflage des Verfahrens. Dabei schien es ihnen »bei einem ‚Klassiker' wie dem MMPI nicht sinnvoll, eine nationale Version zu entwickeln, die zwar nach den Grundsätzen des Originals, ansonsten aber vollständig neu konstruiert ist« (S. 6). Im Vordergrund stand vielmehr die Absicht, die Kontinuität zum amerikanischen Interpretationsmanual zu wahren. Die Überarbeitung beschränkt sich daher – was die alten Items betrifft – auf Veränderungen im sprachlichen Ausdruck, der Grammatik sowie der Wahl einer geschlechtsneutralen Formulierung. Das Ergebnis: Ein update mit insgesamt 567 Items, »für die sich vom reinen Inhalt her auch elegantere Formulierungen angeboten (hätten), damit hätte man aber zugleich größere Abweichungen zum MMPI-Saarbrücken und zum US-MMPI riskiert« (S. 6).
310 Items wurden unverändert von der amerikanischen bzw. deutschen Originalversion übernommen, weitere 149 nach sprachlicher Überarbeitung. Die restlichen 108 Items wurden in einem aufwendigen, mehrstufigen Übersetzungsprozess neu kreiert: In einem ersten Schritt übersetzten zwei unabhängige

Übersetzer/innen die neuen Items aus dem Englischen ins Deutsche. Ausgehend von der Konsensfassung beider Übersetzungen wurden die Items anschließend ins Englische rückübersetzt, bevor sie nach erneuter Überarbeitung zusammen mit den übrigen Items vom Language Department der University of Minnesota ins Deutsche übersetzt wurden. Anzahl der revidierten Items pro Skala und Art ihrer Veränderung sind überblicksartig in tabellarischer Form im Manual dargestellt. Für die amerikanische Version konnten Ben-Porath und Butcher (1989) keine bedeutsamen psychometrischen Unterschiede zwischen »alten« und »neuen« Items feststellen. Eine entsprechende Äquivalenzüberprüfung steht für die deutschsprachige Fassung aus.
Wer mehr über die Itemgenerierung erfahren will und darüber, wie die Skalen entstanden sind, muss auf das Handbuch des MMPI-Saarbrücken zurückgreifen.

Konstruktion der Validitätsskalen
Hierzu finden sich im Manual des MMPI-2 nur spärliche Hinweise und dies obwohl die »sechs Validitätsskalen in ihrer Gesamtheit über die Interpretierbarkeit des Profils entscheiden« (S. 27). Die L-Skala geht auf Hathaway und McKinley (1940), basierend auf den Forschungsarbeiten von Hartshorne und May (1928) sowie Hartshorne, May und Shuttleworth (1930) zurück. Zur Entstehungsgeschichte der F-Skala erfährt man leider wenig. Gleiches gilt für die K-Skala. Die neuen Validitätsindikatoren F_B, VRIN und TRIN werden mit Arbeiten von Tellegen (1982, 1988) begründet.

Konstruktion der klinischen Skalen
Den klinischen Skalen ist eine empirische Konstruktion gemeinsam. Die Antworten psychiatrischer Diagnosegruppen wurden mit denen gesunder Kontrollpersonen kontrastiert und anschließend kreuzvalidiert. Die Items selbst stammen aus unterschiedlichsten Quellen, u.a. aus Arbeiten zu psychiatrischen, medizinischen und neurologischen Störungen, zur Differentialdiagnostik, zu sozialen und emotionalen Einstellungen und zur Persönlichkeitsentwicklung.
Hier stellt sich die Frage nach der Angemessenheit der klinischen Skalen. Ihre ursprüngliche Konstruktion liegt ungefähr 60 Jahre zurück und orientiert sich am Konzept der inhaltlichen Gültigkeit und zwar aus Kraepelinscher Sicht klinischer Syndrome (Kraepelin, 1909). Explizite und operational definierte Kriterien haben zwischenzeitlich jedoch die nosologischen Klassen des Kraepelinschen Klassifikationssystems abgelöst und multiaxiale Beschreibungssysteme wie das ICD-10 oder das DSM IV mit substanziellen Veränderungen bei den Kriterien zu einzelnen Störungen, einer differenzierteren Bezeichnung der Subtypen, psychopathologischen, taxonomischen und nosologischen Entscheidungsregeln sowie Veränderungen in der Nomenklatur hervorgebracht. So gesehen wird der MMPI-2 der modernen psychiatrischen Diagnostik nicht mehr gerecht.

Konstruktion der Inhaltsskalen
Dieser Skalensatz wurde faktorenanalytisch gewonnen, mit der Begründung, »den Inhalt von Fragebogenitems nicht nur bei der Erstellung eines Itempools zu beachten, sondern ihn auch zur Grundlage für die Skalenkonstruktion zu benutzen« (S. 45). Nach Meinung der Autoren sind sie völlig unabhängig von den Basisskalen. Dass den Autoren daran gelegen ist, einen so umfangreichen Itemsatz vollständig auszuschöpfen, ist verständlich. Warum dies nicht stärker theoriegeleitet und empirisch fundiert geschieht jedoch nicht: Von welchen theoretischen Modellvorstellungen zu den einzelnen Konstrukten ausgegangen wurde, bleibt ebenso offen wie das faktorenanalytische Vorgehen selbst. Datenmaterial, das die postulierte (statistische) Unabhängigkeit zwischen Basis- und Inhaltsskalen belegte, fehlt erst recht.

Konstruktion der Zusatzskalen
Die »alten« Zusatzskalen werden »als schon eingeführt und validiert« befunden und mit dem Hinweis auf das Handbuch des MMPI-Saarbrücken abgehandelt. Die neuen Zusatzskalen seien »auf der Grundlage von empirischen Konstruktionsmethoden« (S. 23) entstanden, um neue Zielbereiche für den MMPI zu erschließen. Auch hier wird nicht näher auf die Konstruktionsprinzipien eingegangen.

3. Testdurchführung

3.1 Durchführungsobjektivität

Der Einschätzung der Autoren, wonach der MMPI-2 »relativ leicht und einfach vorzugeben und auszuwerten ist« (S. 13), ist zuzustimmen. Dies gewährleisten differenzierte (Verhaltens-)Empfehlungen zur Testvorgabe sowie Hinweise auf mögliche Störquellen seitens der Testanwender, Tpn und dem Setting. Ausdrücklich weisen sie darauf hin, dass eine ruhige ablenkungsfreie Atmosphäre unverzichtbar für die Testbearbeitung ist. Positiv hervorzuheben ist die geforderte Supervision der Testanwender durch Fachkräfte sowie die Evaluation der Testpraxis zur Sicherung von Qualitätsstandards, ein Hinweis, der für die Durchführung aller Testverfahren eigentlich selbstverständlich sein sollte, in der Praxis aber häufig nicht eingehalten und auch in Manualen viel zu wenig betont wird.

3.2 Transparenz

Während die Validitätsskalen für die Testpersonen nicht zu durchschauen sind, sind die Messintentionen der Basisskalen durch ihre empirische Konstruktion offensichtlich. Und auch die Inhalts- und Zusatzskalen lassen eine direkte Beziehung zwischen Items, Antwortverhalten und Schlussfolgerungen für die Tpn zu.

3.3 Zumutbarkeit

Der »Wegfall einiger als unangebracht empfundener Items und die Hinzunahme von Items aus neuen Inhaltsbereichen« (S. 6) hat nicht zu einem Gewinn im Hinblick auf die Zumutbarkeit des Inventars geführt. Wie seine Vorläuferversion enthält der MMPI-2 eine Reihe von Items, die befremdend wirken, den Intimbereich betreffen und solche, die allen Prinzipien der

Anschrift
Dr. Petra Hank
und
Prof. Dr. Peter Schwenkmezger,
Universität Trier,
Fachbereich I
Psychologie,
54286 Trier

Itemkonstruktion zuwiderlaufen (s.a. Punter & Kubinger, 2002). Darüber hinaus sollte in Anbetracht der hohen Bearbeitungszeit, die der MMPI-2 mit seinen 567 Items erfordert, im Einzelfall genau überlegt werden, ob unter der jeweils definierten Zielsetzung und dem zu erwartenden Nutzen die Anwendung dieses umfangreichen Verfahrens tatsächlich gerechtfertigt ist.

3.4 Verfälschbarkeit

Besteht seitens der Testperson die Bereitschaft zur Mitarbeit, ist der MMPI-2 nicht mehr und nicht weniger gefährdet als andere Persönlichkeitsverfahren auch, verfälscht zu werden. Allerdings verfügt dieser Test mit den L-, F- und K-Skalen über interne Kontrollmechanismen, mit denen sich – nach Meinung der Autoren – verzerrende Antworttendenzen aufdecken lassen.

3.5 Störanfälligkeit

Eine fundierte und sorgfältige Handhabung des Untersuchers vorausgesetzt sowie Durchführungsbedingungen, welche Kooperation und Aufmerksamkeit der Testperson fördern, erscheint der MMPI-2 nicht besonders störanfällig. Nach Einschätzung seiner Konstrukteure verträgt er selbst eine Testunterbrechung. Allerdings fehlen (empirische) Hinweise, wie groß die Zeitintervalle maximal sein können, um eine Unterbrechung mit nachteiligem Einfluss auf die Gütekriterien ausschließen zu können. Ebenso wenig gibt es Kriterien, unter denen eine Unterbrechung angezeigt ist.

4. Testverwertung

4.1 Auswertungsobjektivität

Die manuelle Auswertung erleichtern Schablonen. Alle Markierungen, die in einer Schablone erscheinen, werden zum jeweiligen Skalenrohwert aufsummiert und in die dafür vorgesehene Spalte des Antwortbogens eingetragen, bevor sie im nächsten Schritt nach teilweiser K-Korrektur auf dem Profilblatt markiert und mit einer Linie zur sog. Profildarstellung verbunden werden. Die zugehörigen (uniformen) T-Werte sind an der linken und rechten Seite des Profilblatts abgetragen. Neben der manuellen ist auch eine computerunterstützte Auswertung und Interpretation möglich. Zusätzlich wird die Kodierung des Basisprofils nach Dahlstrom, Welsh und Dahlstrom (1972) beschrieben. Zusammenfassend ist festzustellen, dass der MMPI-2 auch im Hinblick auf seine Auswertung objektiv ist.

4.2 Zuverlässigkeit

Zur Überprüfung der Messgenauigkeit schätzen die Autoren die interne Konsistenz der Skalen, getrennt für die Männer und Frauen der deutschen Normierungsstichprobe (N = 958) und im Vergleich zur amerikanischen. Für die gesamte Stichprobe erreichen die Basisskalen im Mittel eine nur mäßige Konsistenz von α = .75, was jedoch angesichts der empirischen Skalenkonstruktion nicht überrascht. Insgesamt gesehen sind für die deutsche Stichprobe etwas höhere Konsistenzkoeffizienten im Vergleich zur amerikanischen festzustellen. Nur ein Drittel der Zusatzskalen hat einen befriedigenden Homogenitätsgrad. Die Werte streuen von α = .34 - .90 für die männliche und α = .37 - .90 für die weibliche Teilstichprobe. Selbst den faktorenanalytisch konstruierten Inhaltsskalen kann nur knapp zur Hälfte eine zufriedenstellende Homogenität ($\alpha \geq$ .80) attestiert werden – und dies trotz der großen Itemanzahl pro Skala. Entsprechend der im Vergleich zu den Inhaltsskalen reduzierten Itemanzahl der zugehörigen Komponentenskalen fallen die Konsistenzschätzungen für diese deutlich niedriger aus.

Als ein weiteres Maß der Messgenauigkeit werden die Ergebnisse einer Retestuntersuchung von 49 Männern und 56 Frauen mitgeteilt, denen das Inventar im Abstand von 10 Tagen zweimal zur Bearbeitung vorgegeben wurde. Wie dem Manual zu entnehmen ist, ergeben sich, über alle Skalen hinweg betrachtet, Niveauunterschiede in den Skalenmittelwerten von maximal 2.5 Punkten mit tendenziell höheren Werten zum ersten im Vergleich zum zweiten Messzeitpunkt. Für die Basisskalen betragen die Stabilitätskoeffizienten im Durchschnitt r_{tt} = .83, liegen damit in einem mittleren Bereich und entgegen der Einschätzung der Testautoren durchaus nicht »am oberen Rand dessen, was man bei Persönlichkeitsfragebögen erwarten kann« (S.35). Es resultieren Standardmessfehler zwischen drei und fünf T-Werten. Im Hinblick auf die Zusatzskalen ist bei der weiblichen Stichprobe eine geringe Merkmalsstabilität mit Retestkoeffizienten r_{tt} <. 80 zu verzeichnen, bei der männlichen fallen sie insgesamt etwas günstiger aus. Übereinstimmend instabil erweisen sich die Skalen VRIN und TRIN bei beiden Teilstichproben. Vergleichbare Ergebnisse ergeben sich für die Inhaltsskalen. Die Ursache hierfür sehen die Autoren in der »Varianzbeschränkung der Reteststichprobe« (S. 45). Hier wäre es sinnvoll, die Varianzen von Test- und Reteststichprobe zahlenmäßig mitzuteilen, damit der/die Testinterpret/in sich ein Bild von der Lage machen könnte, oder aber die Reliabilitätskoeffizienten entsprechend den vorliegenden Varianzverhältnissen zu korrigieren. Selbstkritisch relativieren die Autoren die Ergebnisse – gut so, allerdings trifft die angeführte Begründung nicht: Nicht die Retestreliabilität, sondern die Konsistenz steigt proportional mit der Itemanzahl einer Skala.

Insgesamt gesehen können die berichteten Reliabilitätsergebnisse nicht überzeugen. Die Werte der Konsistenzschätzungen vieler Skalen genügen den Empfehlungen nicht, sieht man einmal von den Basisskalen ab, bei denen ohnehin die Angemessenheit von Konsistenzschätzungen zu hinterfragen ist. Die Stabilitätskoeffizienten sind zwar, zumindest was die Basisskalen betrifft, in Ordnung, aufgrund der ihnen zugrunde liegenden kleinen Stichprobe, bei der außerdem nicht klar wird, ob es sich um klinisch auffällige oder um gesunde Personen handelt, sind sie jedoch nur von marginaler Aussagekraft.

4.3 Gültigkeit

Interkorrelationen zwischen den Basisskalen werden unter Rekurs auf Hobi und Klär (1973; Hobi, 1983) als Hinweis auf die Validität des MMPI-2 ins Feld geführt. Für ihre Interpretation wären Angaben zur Signifikanzgrenze hilfreich. Als ein weiterer Beleg für die Validität

des Verfahrens werden die Ergebnisse einer Hauptkomponentenanalyse mit anschließender Varimax-Rotation mitgeteilt. Danach ergeben sich die vier Faktoren psychotische Gedankeninhalte, neurotische Verhaltensweisen, Geschlechtsrollenidentifikation und Introversion, deren »Existenzberechtigung« die Autoren »wegen der Itemüberlappung« (S. 37) bezweifeln. Schwerer noch wiegen hier die Überlegungen von Moosbrugger und Hartig (2002), wonach auf der Grundlage dichotomer Items entsprungene Faktoren weniger über die psychologische Struktur der fraglichen Dimension Aufschluss geben als vielmehr Ausdruck eines methodischen Artefakts sind. Betrachtet man sich ungeachtet dieser Einwände dennoch die Faktorladungsmatrix und legt als Kriterium für die Ladungskoeffizienten ≥ .60 an, ist die inhaltliche Beschreibung der Autoren nur für den ersten Faktor, auf dem die Skalen F, Hd, Hy, Pp, Pa, Pt und Sc laden, nachvollziehbar. Faktor zwei vereint die Skalen L und K auf sich. In diesem Zusammenhang von neurotischen Verhaltensweisen zu sprechen ist zumindest diskussionswürdig, wenn nicht sogar fraglich. Die Skala D macht den dritten Faktor aus, die Skala Mf stellt den vierten dar und entspricht damit dem von den Testautoren benannten dritten Faktor Geschlechtsrollenidentifikation.

4.4 Normierung

In die Normierungsstichprobe gingen die Daten von 458 Männern und 500 Frauen ein, die in einem geschichteten, dreistufigen Zufallsauswahlverfahren erhoben wurden. Den demographischen Charakteristika zufolge ist diese Stichprobe als repräsentativ für die bundesdeutsche Bevölkerung hinsichtlich Alter, Geschlecht und geographischer Herkunft zu bewerten. Zusätzlich zum MMPI-2 beantworteten diese Personen, vermutlich als externes Validitätskriterium gedacht, den 16 PF-R (Schneewind & Graf, 1999) im Rahmen eines Interviews. Um einen Hinweis auf die Übereinstimmung zwischen schriftlicher resp. mündlicher (16 PF-R) Testvorgabe (MMPI) zu erhalten, waren im 16 PF-R nochmals 16 Items des MMPI-2 eingestreut. Abgesehen von der äußerst fragwürdigen Vorgehensweise, die 16 doppelt vorgegebenen Items – sie machen weniger als 3% der gesamten Itemzahl des MMPI-2 aus – als Indikator für die Äquivalenz zwischen mündlicher und schriftlicher Testvorgabe werten zu wollen, stellt sich die Frage, warum der 16-PF-R überhaupt erhoben wurde, da ohnehin keine Ergebnisse hierzu mitgeteilt werden.

Mit dem Hinweis auf die angestrebte Kontinuität zwischen dem MMPI-Saarbrücken bzw. der amerikanischen Originalversion und dem MMPI-2 stellen die Autoren einen Mittelwertsvergleich über insgesamt drei amerikanische und zwei deutsche Datensätze an. Mitgeteilt werden lediglich Mittelwerte und korrespondierende Standardabweichungen. Will man der Interpretation der Testautoren nicht nur Glauben schenken, sondern sich selbst ein Bild von der Äquivalenz machen, kommt man nicht umhin, eigene Berechnungen anzustellen, da die entsprechenden inferenzstatistischen Ergebnisse nicht mitgeteilt werden. Aber auch diese erleichtern einem die Autoren nicht unbedingt: Fehlende Angaben zur Stichprobengröße machen es notwendig, dass man sich die über die verschiedenen Publikationen hinweg verstreuten Angaben zur Stichprobengröße selbst zusammensucht.

Als positive Neuerung im Vergleich zum MMPI-Saarbrücken ist die Transformation der Skalenrohwerte in uniforme T-Werte nach Tellegen und Ben-Porath (1992) zu werten. Gegenüber den herkömmlichen linearen T-Werten haben sie den Vorteil, unabhängig vom Ausmaß der Schiefe der jeweiligen Rohwertverteilung dem gleichen Prozentrang zu entsprechen. Allerdings ist damit die von Angleitner (1997) festgestellte »psychometrische Anomalie« des MMPI-2 nicht behoben: Ungeachtet der externalen Konstruktion führen die Autoren auch mit dieser Normierung eine dimensionale Interpretation des Verfahrens durch und suggerieren so kontinuierliche Übergänge zwischen an sich distinkten Klassen.

4.5 Bandbreite

Die Verwendung des MMPI-2 bietet sich nach dem Manual für unterschiedliche Fragestellungen in der klinischen Anwendung an. Angesichts der Weiterentwicklung der modernen Klassifikationsforschung und der damit verbundenen operationalen psychiatrischen Diagnostik stellt sich jedoch die Frage, ob der MMPI-2 hier noch indiziert ist.

Der diagnostische Anspruch des MMPI-2 auf die Domäne der Personalselektion ist aufgrund der mangelnden Validitätsnachweise nur im Falle einer hohen Selektionsrate gepaart mit einer hohen Grundquote zu verantworten. Darüber hinaus wird ohnehin vor dem Einsatz von Persönlichkeitsinventaren im Rahmen von eignungsdiagnostischen Situationen mit selektivem Charakter gewarnt, da diese Inventare relativ anfällig für die Effekte spezieller Testmotivationen sind (Fahrenberg, Hampel & Selg, 1994; Janke, 1973). Auch die Zielsetzung, den MMPI-2 zur mehrdimensionalen Erfassung von habituellen Personmerkmalen einsetzen zu wollen, erscheint aufgrund der pathologisch orientierten Skalen nicht angemessen.

Über diese »alten«, bereits mit dem MMPI-Saarbrücken beanspruchten, Anwendungskontexte hinaus, wurde versucht, mit dem MMPI-2 den Geltungsbereich des Inventars durch Hinzufügen weiterer Zusatz- und Inhaltsskalen auszudehnen. Diese Erweiterung ist für die deutsche Fassung bisher nicht in Sicht, weil die Gültigkeit der entsprechenden Skalen noch nicht empirisch belegt ist.

4.6 Informationsausschöpfung

Bemisst man die Informationsausschöpfung an der Menge der begründet ableitbaren Indikatoren, ist sie für den MMPI-2 als eher gering zu bewerten, da umfassende Nachweise von Reliabilität und Validität der Skalen fehlen.

5. Testevaluation

5.1 Ökonomie

Der MMPI-2 ist für den Testanwender nicht zuletzt durch den computerunterstützten Auswertungsdienst einfach zu handhaben, stellt für die zu testende Person allerdings eine nicht unerhebliche zeitliche Belastung dar. Die Bearbeitungszeit wird für gesunde Personen mit etwas mehr als einer Stunde veranschlagt, bei klinisch auffälligen Personen liegt sie nach Angaben der Autoren bis zu 30% darüber. Seinen Nutzen hinsichtlich der inkrementellen Validität abwägen zu wollen, erscheint zum gegenwärtigen Zeitpunkt verfrüht.

5.2 Fairness

Der MMPI-2 wäre »unfair«, wenn durch seine Art der Testung bestimmte Personengruppen systematisch benachteiligt würden. Entsprechend »gefährdete« Personen werden von den Autoren benannt. Befolgt der Testanwender jedoch die sorgfältigen Empfehlungen der Autoren zur Testfähigkeit der Testperson, ist davon auszugehen, dass der MMPI-2 fair misst.

5.3 Akzeptanz

Die verschwindend geringe Zahl der Itemauslassungen in der Normierungsstichprobe spricht für die Akzeptanz des Verfahrens. Es gibt aber keine Informationen darüber, welche Anreize den Tpn der Normierungsstichprobe geboten wurden, die möglicherweise die Akzeptanz des Verfahrens begünstigt haben.

5.4 Vergleichbarkeit

Der MMPI-2 wurde nach der klassischen Testtheorie konstruiert, seine psychometrische Qualität anhand der entsprechenden Gütekriterien untersucht. (Uniforme) T-Werte erlauben die Vergleichbarkeit von Ergebnissen des MMPI-2 mit denen anderer Verfahren.

5.5 Bewährung

Ginge man von der Zahl der Veröffentlichungen zu seinem Vorläufer als Prädiktor für den Bewährtheitsgrad des MMPI-2 aus, bestünde kein Anlass zur Sorge um dessen Bewährung. Ohne den Wert dieser »allgemeinen Anerkennung (eignungs-)diagnostischer Arbeit« (Jäger, 1966) unterschätzen zu wollen, enthebt sie nicht von dem empirischen Gütenachweis entsprechender (Eignungs-)Prognosen. In Anbetracht der aufgeführten Kritikpunkte erscheint es mehr als fragwürdig, dass der MMPI-2 sich tatsächlich je bewähren wird.

6. Äußere Testgestaltung

Das Testmanual ist leicht verständlich geschrieben und übersichtlich gegliedert. Die Items werden in einem wiederverwendbaren Testheft präsentiert, ihre Antworten auf einem separaten Antwortbogen markiert, die Ergebnisse auf einem Profilblatt dargestellt.

7. Abschließende Bewertung

Zusammenfassend bleibt festzuhalten: Der MMPI-2 stellt den Versuch einer längst überfälligen Überarbeitung des MMPI-Saarbrücken dar. Oberste Zielsetzung war es, die Kontinuität zum MMPI-Saarbrücken und zum amerikanischen Original beizubehalten, eine Strategie, die durchaus in Frage gestellt werden kann. Über die geglückte inhaltliche Kontinuität hinaus findet sich in der Nachfolgeversion nämlich auch eine Reihe der vielfach kritisierten Schwächen des MMPI-Saarbrücken wieder.

Ein Beweggrund für die Überarbeitung des MMPI-Saarbrücken, seine Neustandardisierung an einer repräsentativen Eichstichprobe, ist nur teilweise erreicht worden. Zwar handelt es sich bei der jetzigen um eine repräsentative geschichtete Zufallsstichprobe, ihr Umfang lässt aber nach wie vor zu wünschen übrig. Zum zweiten nutzten die Autoren die Überarbeitung, um schwer verständliche und antiquierte Items auszusondern. Schade, dass dies nicht konsequenter erfolgte, denn das Inventar zählt nach wie vor 567 Items. Hier ist nicht nachzuvollziehen, warum die Autoren sich der Tradition verhaftet fühlen und die Mf-Skala mit immerhin 56 Items in den MMPI-2 hinüberretten. Anders ausgedrückt: Die Chance, das Verfahren mit der Revision zu entschlacken, ist vertan worden.

Eine immer wieder gegen den MMPI-Saarbrücken vorgebrachte Kritik der unzureichenden psychometrischen Gütenachweise (Kubinger, 1995; zsf. Angleitner, 1997) gilt nach wie vor: Bedeutsame empirische Nachweise zu Reliabilität und Validität des MMPI-2 fehlen. Im Hinblick auf seine Reliabilität präsentieren die Autoren in ihrer Aussagekraft unbefriedigende Konsistenzschätzungen und äußerst schwache Stabilitätskoeffizienten. Als Validitätsnachweis begnügen sie sich mit den Ergebnissen einer Faktorenanalyse. Hier wünschte man sich weitergehende Untersuchungen mit externen diagnostischen und prognostischen Gütekriterien.

Eine Folge dieser Nachlässigkeit berührt einen weiteren, alten Kritikpunkt, nämlich die Frage nach der Gültigkeit eines Testprotokolls (Angleitner, 1997). Mit dem Ziel, die Entscheidung hierüber auf sichereren Boden zu stellen, präsentieren die Autoren drei neue Validitätsindikatoren (Fb, VRIN und TRIN), um ihren Nutzen gleich wieder zu relativieren: »Die Skalen VRIN und TRIN sind derzeit noch in der Erprobung und sollten vorsichtig angewendet werden bis mehr empirische Evidenz vorliegt« (S. 29). Es wäre sinnvoller gewesen, die klassischen Validitätsindikatoren zu begründen.

Doch damit nicht genug. Ungeachtet der mangelnden Reliabilitäts- und Validitätsnachweise lassen sich die Autoren eine Profildarstellung nicht nehmen, die folglich weder messgenau noch gültig sein kann. Dies wiegt um so schwerer als u.a. Angleitner (1997) im Zusammenhang mit dem MMPI-Saarbrücken auf diese Problematik hingewiesen hat.

Fazit. Selbst wenn die Einschätzung der Testautoren zutrifft, dass sich »der MMPI in der ganzen Welt als der wohl wichtigste Fragebogen zur Selbsteinschätzung von Personen mit psychischen Problemen und Störungen bewährt hat,…« (S. Vii), enthebt sie dies nicht von einer empirischen Tauglichkeitsüberprüfung des MMPI-2. Ganz im Gegenteil, gerade weil der MMPI so häufig eingesetzt wird, sind Nachweise zu seiner psychometrischen Gültigkeit unerlässlich. Entsprechenden Forschungsarbeiten sollten allerdings kri-

tische Überlegungen zu Messintentionen und Anwendungsbereichen des MMPI-2 vorangestellt werden. Solange diese empirischen Gültigkeitsnachweise ausstehen, erscheint der Einsatz des Verfahrens nicht vertretbar.

LITERATUR

AMELANG, M. & ZIELINSKI, W. (1997). *Psychologische Diagnostik und Intervention* (2. Auflage). Berlin: Springer.
ANGLEITNER, A. (1997). Testrezension zum Minnesota Multiphasic Personality Inventory (MMPI). *Zeitschrift für Diffeentielle und Diagnostische Psychologie, 18*, 4-10.
BEN-PORATH, Y.S. & BUTCHER, J.N. (1989). Psychometric stability of rewritten MMPI items. *Journal of Personality Assessment, 53*, 645-653.
BEN-PORATH, Y.S. & SHERWOOD, N. (1993). *The MMPI-2 Content Component Scales*. Minneapolis: University of Minnesota Press.
DAHLSTROM, W.G., WELSH, G.S. & DAHLSTROM, L.E. (1972). *An MMPI Handbook. Volume I. Clinical Interpretation*. Minneapolis: University of Minnesota Press.
FAHRENBERG, J., HAMPEL, R. & SELG, H. (1994). *Das Freiburger Persönlichkeitsinventar (FPI) – Revidierte Fassung FPI-R und teilweise geänderte Fassung FPI-A1.* Göttingen: Hogrefe.
HARRIS, R.E. & LINGOES, J.C. (1955). *Subscales for the MMPI: An aid to profile interpretation*. Mimeographed Materials. Department of Psychiatry, University of California (Corrected version, 1968; documented in Dahlstrom, Welsh and Dahlstrom, 1972).
HARTSHORNE, H. & MAY, M.A. (1928). *Studies in the nature of character. I. Studies in deceit*. New York: Macmillan.
HARTSHORNE, H., MAY, M.A. & SHUTTLEWORTH, F.K. (1930). *Studies in the nature of character. III. Studies in the organization of character*. New York: Macmillan.
HATHAWAY, S.R. & MCKINLEY, J.C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology, 10*, 249-254.
HATHAWAY, S.R., MCKINLEY, J.C & ENGEL, R.R. (Hrsg.). (2000). *Manual zum Deutschen MMPI-2™*. Göttingen: Huber.
HOBI, V. (1983). Zur Faktorenstruktur der mehrdimensionalen Persönlichkeitsinventare MMPI, 16-PF, FPI und GT. *Psychiatrie, Neurologie und Medizinische Psychologie, 35*, 236-243.
HOBI, V. & KLÄR, A. (1973). Eine gemeinsame Faktorenanalyse von MMPI, FPI und 16-PF. *Zeitschrift für Klinische Psychologie, 2*, 27-48.
JÄGER, A.O. (1966). Prognose und Bewährung in der Eignungsdiagnostik. *Psychologische Rundschau, 17*, 185-208.
JANKE, W. (1973). Das Dilemma von Persönlichkeitsfragebogen. Einleitung des Symposiums über Konstruktion von Fragebogen. In G. Reinert (Hrsg.), *Bericht über den 27. Kongress der Deutschen Gesellschaft für Psychologie in Kiel 1970* (S. 44-48). Göttingen: Hogrefe.
KRAEPELIN, E. (1909). *Psychiatrie*. Leipzig: Barth.
KUBINGER, K.D. (1995). *Einführung in die Psychologische Diagnostik*. Weinheim: Psychologie Verlags Union.
MOOSBRUGGER, H. & HARTIG, J. (2002). Factor analysis in personality research: Some artifacts and their consequences for psychological assessment. *Psychologische Beiträge, 44*, 136-158.
PUNTER, J. F. & KUBINGER, K. (2002). Was ist aus der Kritik der »Testrezension: 25 einschlägige Verfahren« (Zeitschrift für Differentielle und Diagnostische Psychologie, 18, Heft 1-2) geworden? *Psychologie in Österreich, 22*, 24-32.
SCHNEEWIND, K.A. & GRAF, J. (1999). *Der 16-Persönlichkeits-Faktoren-Test, revidierte Fassung (16 PF-R)*. Bern: Huber.
SPREEN, O. (1963). MMPI. *Saarbrücken. Handbuch zur deutschen Ausgabe des MMPI von S.R. Hathaway and J.C. McKinley*. Bern: Huber.
TELLEGEN, A. (1982). *Brief manual for the Differential Personality Questionnaire*. Minneapolis: University of Minnesota Press.
TELLEGEN, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56*, 621-663.
TELLEGEN, A. & BEN-PORATH, Y.S. (1992). The new uniform T scores for the MMPI-2: rationale, derivation, and appraisal. *Psychological Assessment, 4*, 145-155.
WIENER, D.N. & HARMON, L.R. (1946). Subtitle and obvious keys for the MMPI item pool. *Psychological Monographs, 80* (2, Whole No. 630).

Stellungnahme zur Testrezension des MMPI-2

durch Hank und Schwenkmezger (2003)

Rolf R. Engel

Prof. Dr. Rolf R. Engel

Klnikum der Universität München, Psychiatrische Klinik, Abt. Klinische Psychologie und Psychophysiologie, Nußbaumstr. 7, 80336 München

re@psy.med.uni-muenchen.de

Im Auftrag des Testkuratoriums haben Hank & Schwenkmezger die vorstehende Testbesprechung des MMPI-2 verfasst, die mit dem Satz endet: »Solange diese empirischen Gültigkeitsnachweise ausstehen, erscheint der Einsatz des Verfahrens nicht vertretbar.« Man reibt sich die Augen und versteht es nicht: ein gut 50 Jahre altes Verfahren, das ohne jeden Zweifel das empirisch am besten belegte und recherchierte klinische Testinstrument ist, wird in einer Rezension, die sich vornehmlich an praktisch tätige Diplompsychologinnen und Diplompsychologen richtet, nicht etwa als »nicht empfehlenswert« oder, schön wär's, als »überholt«, sondern als »nicht vertretbar« beurteilt. Wäre der MMPI ein neuer Test, träfe eine solche Kritik nur den Testautor. Bei einem etablierten Verfahren ist das anders: Hier trifft sie alle praktisch tätigen klinischen Psychologen, die den Test (als MMPI-Saarbrücken oder als MMPI-2) in Deutschland seit über 40 Jahren einsetzen, Patienten damit beschreiben, Gutachten darauf begründen, Mitarbeiter und Kolleginnen darin unterrichten und mit all dem – nach dieser Rezension – offensichtlich professionell Unvertretbares praktizieren. Ist das nur ein Missverständnis? Sind akademisch ausgebildeten Praktikern die empirischen Gültigkeitsnachweise ihrer Verfahren egal? Oder haben die Regeln des Testkuratoriums Rezensentin und Rezensent in die Irre geführt?

Inhalt der Besprechung

Die sehr ausführliche Besprechung beschreibt zunächst auf gut fünf Manuskriptseiten das Testverfahren selbst (Abschnitt 1.1). Im Wesentlichen werden die einzelnen Skalen beschrieben. Dabei gibt es, mit einer Ausnahme, kaum inhaltlichen Dissens, den hier aufzugreifen sich lohnte. Die Ausnahme betrifft den Satz in der Einleitung: »Über die ursprüngliche Zielsetzung der psychiatrischen Kategorisierung hinaus erhebt das Verfahren zwischenzeitlich den Anspruch, auch für eignungsdiagnostische Fragestellungen tauglich zu sein.« Etwas später, unter 1.1, heißt es: »Anwendungskontexte sind gemäß dem Manual u.a. medizinische und psychiatrische Beurteilungen sowie die Personalauswahl.« Noch an zwei weiteren Stellen der Rezension wird Eignungsdiagnostik oder Personalauswahl als Anwendungsbereich genannt. Mal abgesehen davon, dass der MMPI-2 in erster Linie bei *klinisch-psychologischen* und nicht bei *medizinischen* Beurteilungen eingesetzt wird, kommt weder das Wort Personalauswahl noch Eignungsdiagnostik im deutschen Manual vor. Schon im dritten Satz des Vorworts des deutschen Handbuchs wird der MMPI als »Fragebogen zur Selbsteinschätzung von Personen mit psychischen Problemen und Störungen« eingeführt und auch im Rest des Textes geht es immer um klinische Probleme. Auch bei sorgfältiger Kontrolle meines Manualtextes konnte ich keine Stelle finden, an der der MMPI oder MMPI-2 zur Personalauswahl oder Eignungsdiagnostik empfohlen würde. Mir ist ein Rätsel, woher diese Vermutung stammt.

Die gewichtigsten Kritikpunkte der weiteren Rezension sind die folgenden: Im Abschnitt 1.2 wird gerügt, dass die theoretischen Überlegungen zum Entstehungshintergrund des MMPI-2 im Handbuch nicht enthalten sind. Die Konstruktion der Validitätsskalen sei kaum beschrieben, die Konstruktion der klinischen Skalen sei veraltet, über den Konstruktionshintergrund der Inhaltsskalen und der Zusatzskalen werde nichts berichtet.

Eine positive Beurteilung (eigentlich die einzige) erfährt der recht banale Abschnitt des Manuals, der Hinweise für die Testvorgabe enthält und zum Beispiel die Supervision der Testanwender durch Fachkräfte fordert. Zu den sonstigen Kriterien der Testdurchführung gibt es durchwachsene Beurteilungen.

Im Abschnitt 3 wird die Zuverlässigkeit als »nicht überzeugend« beurteilt. In der Sektion Gültigkeit wird nur die im Handbuch angegebene Faktorenanalyse angeführt, was als völlig ungenügend gewertet wird. (Diese Faktorenanalyse wurde übrigens auf Skalen-, nicht auf Itemebene durchgeführt, weshalb der an sich richtige Verweis auf die methodischen Probleme von Faktorenanalysen dichotomer Items ins Leere geht.) Bei der Normierung wird die Stichprobengröße bemängelt.

Bei der abschließenden Bewertung sticht vor allem die »Kritik der unzureichenden psychometrischen Gütenachweise« ins Auge, die auch mit Verweisen auf Kubingers (1995) Lehrbuch und Angleitners (1997) Rezension des MMPI in der Zeitschrift für Differentielle und Diagnostische Psychologie bekräftigt wird. Waren die Stabilitätskoeffizienten im Hauptteil der Rezension der Höhe nach noch »in Ordnung«, wenn auch wegen kleiner Stichprobe und fehlender Information über die Stichprobenzusammensetzung nur von »marginaler Aussagekraft«, so werden daraus in der Zusammenfassung schon »äußerst schwache Stabilitätskoeffizienten«. Nachweise zur psychometrischen Gültigkeit seien unerlässlich und würden fehlen. Kritisiert werden vor allem schwache Reliabilitäts- und fehlende Gültigkeitsdaten.

Stellungnahme

Nachvollziehbare Darlegungen zum theoretischen Hintergrund einer Skalenkonstruktion und empirische Daten über deren Gelingen sind die Basis einer wissenschaftlichen Testkonstruktion. Verfügte der MMPI

nicht über solche Daten, hätten sich Generationen von empirisch forschenden differentiellen und klinischen Psychologen nicht damit beschäftigt.

Die Konstruktion der Basisskalen des MMPI wurde beginnend im Jahr 1940 in einer Serie von Aufsätzen dargestellt, die jeweils ausführlich sowohl den theoretischen Hintergrund als auch das (empirische) Konstruktionsprinzip erläutern. Es gibt je einen eigenen Artikel für die Konstruktion der Skalen Hd, D, Pt, Si. Die Konstruktion der Skalen Hy, Pp und Ma beschreibt eine zusammenfassende Arbeit, die der Skalen Mf, Pa und Sc eine weitere, die Konstruktionsarbeit an den (klassischen) Validitätsskalen sind in zwei weiteren Arbeiten dargestellt. Das alles muss weder ein Anwender noch ein Rezensent in den Originalartikeln nachlesen: wie im deutschen Handbuch erwähnt, sind diese Arbeiten in einem Reader (Welsh & Dahlstrom, 1956) zusammengefasst, der in jeder deutschen UB zu finden sein dürfte, und sie werden auch ausführlich in diversen neuen, leicht von jeder Buchhandlung zu besorgenden Lehrbüchern zum MMPI-2 zusammenfassend dargestellt (am ausführlichsten in Greene, 2000, einem höchst empfehlenswerten Buch, das ebenfalls im deutschen Handbuch zitiert wird). Auch die Konstruktion der neuen Zusatzskalen ist selbstverständlich in Publikationen ausführlich beschrieben, mindestens eine (die jeweils wichtigste) ist im deutschen Manual genannt. Über die Konstruktion der Inhaltsskalen und der Inhaltskomponentenskalen für den MMPI-2 gibt es je eine eigene Monographie, auch die im Manual zitiert. Wenn sich also jemand über die Hintergründe und Konstruktionsprinzipien der MMPI-2-Skalen informieren will, hat er alle Möglichkeiten dazu, die einem akademisch gebildeten Benutzer offen stehen. Es kann nicht der Sinn eines deutschen Manuals der Neuausgabe sein, noch einmal abzuschreiben oder auch nur zusammenzufassen, was in derart vielen Publikationen enthalten ist. Die Menge der Literaturstellen, die man dabei mit einbeziehen müsste, wäre viel zu groß, Greenes Buch zum Beispiel hat knapp 700 Seiten. Das deutsche Handbuch zum MMPI-2 schildert die Neunormierung, beschreibt die Skalen und gibt dem Neuling Hinweise zu deren Interpretation (erheblich ausführlicher als im alten Handbuch des MMPI-Saarbrücken), mehr will und kann es bei einem im Prinzip 60 Jahre alten Verfahren nicht leisten.

Ähnlich verhält es sich mit dem zweiten Kritikpunkt, den angeblich unzureichenden psychometrischen Gütenachweisen. Nur weil im Handbuch keine Validierungsstudien berichtet sind, kann man nicht folgern, es gäbe sie nicht. Für die klassischen Skalen gibt es eine kaum noch zu erfassende Fülle von empirischem Datenmaterial, selbst für viele der neueren Skalen des MMPI-2 gibt es schon umfangreichere Validitätsbelege als für viele andere Instrumente. Eine schnelle Abfrage in der Datenbank Psycinfo ergab für den Zeitraum 1990 bis 2002 zum Stichwort MMPI 3175 Zitate, also rund ein paar Hundert pro Jahr. Schränkt man die Abfrage mit dt=book nur auf die publizierten Bücher ein, erhält man in diesem Zeitraum 70 Bücher, in denen als Schlagwort MMPI vorkommt. Darunter sind viele, die sich ausschließlich der Interpretation des MMPI oder MMPI-2 widmen. Die wichtigsten Lehrbücher, mit denen ein ernsthafter Benutzer anfangen sollte, sind im deutschen Manual erwähnt. Es sind auch nicht alle Publikationen auf englisch: im deutschen Psyndex wurden immerhin im gleichen Zeitraum und zum gleichen Stichwort ebenfalls 291 Zitate gefunden. Auch hier gilt also: ein deutsches Handbuch, das die Neunormierung des MMPI-2 beschreiben soll, kann das nicht alles wieder aufarbeiten und zitieren. Genauso wie das amerikanische MMPI-2-Manual verfolgt auch das deutsche den Zweck, dem Leser das Zahlenmaterial zur Neunormierung möglichst vollständig, ansonsten aber nur die wesentlichen Informationen über den praktischen Gebrauch des Verfahrens an die Hand zu geben, ergänzt um die Literaturstellen, in denen Weiteres zu finden ist.

Bei den wenigen Daten, die speziell für die deutsche Adaptation neu erhoben wurden, tut die Rezension so, als seien sie die einzige Grundlage zur Beurteilung des Verfahrens. Etwa hinsichtlich der Reliabilität: Empirische Daten zur Reliabilität der klassischen MMPI-Skalen werden seit Jahrzehnten gesammelt und jede neue Stichprobe ist bestenfalls als Ergänzung zu betrachten, nicht als Ersatz. Die im Handbuch vorgelegten Reliabilitätsdaten unterscheiden sich nicht nennenswert von dem, was man auf Grund der Literatur erwarten kann (auch wenn die Erwartungen der Rezensenten scheinbar andere sind). Und selbst wenn sie sich unterschieden, würde das eher eine Aussage über die Stichprobenzusammensetzung der neuen Untersuchung als über den Test machen.

Die deutsche Neunormierung wurde an einer repräsentativen Stichprobe in Deutschland wohnender deutscher Staatsbürger im Alter von 18 bis 70 Jahren durchgeführt. Von 1052 Probanden wurden Daten erhoben, die von 958 Personen gingen in die Normierung ein (die Details stehen im Handbuch). Derart aufwändige Normierungen wurden bisher nur für wenige Fragebögen und meines Wissens für keinen einzigen Leistungstest vorgelegt. Selbst Kubinger, der eigentlich nicht dafür bekannt ist, real existierende Verfahren mit Lob zu überhäufen, bezeichnet die Normierung des MMPI-2 in einer von den Rezensenten selbst zitierten Publikation (Punter & Kubinger, 2002) als beispielhaft. Was ist das Fazit der Rezension: Der Stichprobenumfang lässt zu wünschen übrig. Warum? Wofür? Um wie viel Euro hätte der Test denn teurer werden dürfen, wenn dafür eine größere (wie groß?) Stichprobe untersucht worden wäre? Hier ist Augenmaß gefragt.

Die Interpretationshinweise im deutschen (wie im amerikanischen) Handbuch sind nicht aus der Luft gegriffen, auch wenn nicht jeder Satz im Handbuch mit Daten unterlegt ist. Die verbale Interpretation von MMPI-Profilen hat eine lange Tradition. Die Daten, auf denen sie beruht, stammen in der Mehrzahl aus empirischen Arbeiten, in denen deskriptive Korrelate von erhöhten Skalenwerten oder bestimmten Profiltypen gefunden wurden. Auch hierzu kann man in dem Buch von Greene (2000) das notwendige Hintergrundwissen finden. Eine sehr vollständige Zusammenstellung der vorhandenen empirischen Validitätsinformationen für die klinischen Profiltypen der Basisskalen publizierte Lachar (1974), der in der

Literatur

ANGLEITNER, A. (1997). Testrezension zu Minnesota Multiphasic Personality Inventory (MMPI). *Zeitschrift für Differentielle und Diagnostische Psychologie, 18*, 4-10.

ENGEL, R.R. (1980). Validierung eines klinischen Routine-Systems zur computerisierten Erstellung von MMPI-Befunden bei psychiatrischen Patienten. *Archiv für Psychiatrie und Nervenkrankheiten, 229*, 165-177.

ENGEL, R.R. (1997). Replik zu A. Angleitners Testrezension des Minnesota Multiphasic Personality Inventory (MMPI). *Zeitschrift für Differentielle und Diagnostische Psychologie, 18*, 10-15.

GREENE, R.L. (2000). *The MMPI-2. An Interpretive Manual.* Second Edition. Boston: Allyn & Bacon.

HANK, P. & SCHWENK-MEZGER, P. (2003). Testbesprechung: Das Minnesota Multiphasic Personality Inventory – 2 (MMPI-2) in der deutschen Überarbeitung von Rolf R. Engel (2000). *Report Psychologie (in der vorliegenden Ausgabe)*.

KUBINGER, K.D. (1995). *Einführung in die Psychologische Diagnostik.* Weinheim: Psychologie Verlags Union.

LACHAR, D. (1974). *The MMPI: Clinical Assessment and Automated Interpretation.* Los Angeles, CA: Western Psychological Services.

MOREY, L.C. (1991). *Personality Assessment Inventory.* Odessa, FL: Psychological Assessment Resources.

PUNTER, J.F. & KUBINGER, K.D. (2002). Was ist aus der Kritik der »Testrezensionen: 25 einschlägige Verfahren« (Zeitschrift für Differentielle und Diagnostische Psychologie, 18, Heft 1-2) geworden? *Psychologie in Österreich, 2-3*, 24-33.

TESTKURATORIUM DER FÖDERATION DEUTSCHER PSYCHOLOGENVERBÄNDE (1986). Beschreibung der einzelnen Kriterien für die Testbeurteilung. *Diagnostica, 32*, 358-360.

WELSH, G.S. & DAHLSTROM, W.G. (Eds.) (1956). *Basic readings on the MMPI in psychology and medicine.* Minneapolis: University of Minnesota Press.

gleichen Monographie zusätzlich auf der Grundlage der berichteten Studienergebnisse ein formalisiertes Interpretationssystem entwickelte. Es ist in seiner Gesamtheit publiziert, damit sich jeder ein Bild von der empirischen Basis eines einzelnen interpretativen Statements machen kann. Zusammen mit dem MMPI-2 wird (bei der Faxauswertung) eine deutsche Adaptation dieses klinischen Interpretationssystems angeboten, das über viele Jahre an der Psychiatrischen Klinik der LMU München entwickelt und auf seine Gültigkeit überprüft wurde (Engel, 1980). Es beruht noch komplett auf den Skalen und Normen des alten MMPI, eben deshalb, weil es bisher nur dafür Gültigkeitsdaten gibt. Das innovative Angebot einer Auswertung und Interpretation des MMPI-2 per Fax (der Einsender faxt das Antwortblatt an einen Auswertungsservice und erhält in wenigen Minuten Auswertung und Interpretation per Fax zurück) wird in der Rezension überhaupt nicht erwähnt, obwohl es für die praktische Anwendung des Verfahrens von hohem Wert ist.

Man kann nun eigentlich nicht annehmen, dass all dies den Rezensenten unbekannt wäre. Dies würde ja heißen, dass die Rezensenten eines Testverfahrens weniger davon wüssten als der durchschnittliche Anwender. Nein, es scheint sich um den Fall einer beabsichtigten Ahnungslosigkeit zu handeln. Die Rezensenten gehen offensichtlich nur von den Angaben aus, die im deutschen Handbuch wörtlich zu finden sind und halten es nicht für ihre Aufgabe, den zahlreichen Verweisen auf weitere Publikationen nachzugehen. Ich halte das für ein absurdes Vorgehen. Ein Verfahren wie der MMPI, zu dem über 60 Jahre hinweg weltweit namhafte Vertreter der klinischen und differentiellen Psychologie beigetragen haben, wird in einer Rezension ausschließlich nach seinem deutschen Handbuch beurteilt! Ist das gewollt provinziell? Wenn ein Verfahren, zu dem in den letzten 12 Jahren weltweit über 3000 Publikationen erschienen sind (wohlgemerkt, nicht in grauer Literatur, sondern in wissenschaftlichen Fachzeitschriften, die meisten davon mit empirischer Ausrichtung), in dieser Rezension als »Anwendung nicht vertretbar« eingestuft wird, dann wird hier jedenfalls ein Sonderweg beschritten und der internationale Standard verlassen.

Es wäre allerdings nur die halbe Wahrheit, würde man das (nur) der Autorin und dem Autor der vorliegenden Rezension anlasten. Die Rezension wurde vom Testkuratorium in Auftrag gegeben, einem Gremium, das (irgendwie) von der Deutschen Gesellschaft für Psychologie und dem Berufsverband Deutscher Psychologinnen und Psychologen besetzt wird und keinen eigenen Rechtsstatus hat. Dieses Testkuratorium hat 1986 in einer dreiseitigen Publikation einen Kriterienkatalog für die Beurteilung von psychologischen Tests verfasst. Seither werden Verfasser von Rezensionen im allgemeinen gebeten, sich an diese Richtlinien zu halten. In der Präambel der Richtlinien (S. 358) steht: »Grundlage für die Testbewertung ist prinzipiell das Testmanual; dieses muss so beschaffen sein, dass die wichtigsten Aussagen zu den für die Beurteilung relevanten Punkten daraus erarbeitet werden können.« Nun gut, diese Empfehlung ist dann nicht falsch, wenn man das »erarbeiten« so interpretiert, dass die notwendigen Literaturverweise enthalten sein müssen, damit man sich sachkundig machen kann. Allerdings wird die Empfehlung in der vorliegenden Rezension (und in vielen anderen zuvor) so interpretiert, dass nur diejenigen Daten zur Grundlage der Rezension gemacht werden, die explizit im Handbuch abgedruckt sind. Ich habe diesen unseligen Satz wegen der Gefahr seiner Fehlinterpretation schon in meiner Replik auf Angleitners MMPI-Rezension von 1997 angeprangert (Engel, 1997). Ein 60 Jahre altes Verfahren mit Tausenden von Publikationen und ein 60 Wochen oder Monate altes Verfahren, zu dem es außer dem Handbuch nichts gibt, kann man nicht dadurch gleich behandeln, dass man beide nach ihrem Testhandbuch beurteilt. Das wäre sowohl wissenschaftlicher als auch professioneller und berufspolitischer Unsinn. Es wäre etwa so, als wollte man – in Zeiten der DIN-Normierung von Tests mögen solche Vergleiche erlaubt sein – die Zuverlässigkeit eines Volkswagens nach der Güte seiner portugiesischen Gebrauchsanweisung beurteilen. Es bereitet intellektuelle Qualen, wenn in der Rezension der Handbuchhinweis auf die notwendige Qualifizierung der Benutzer des MMPIs besonders lobend erwähnt wird, gleichzeitig aber diesem qualifizierten Benutzer offensichtlich nicht zugetraut wird, bei der Anwendung eines Tests etwas anderes als das Testhandbuch zu lesen geschweige denn schon von seiner akademischen Ausbildung her zu kennen! Hier ist dringend Änderung notwendig und die Vorstände von Berufsverband und Deutscher Gesellschaft wären gut beraten, den Wortlaut der damaligen Richtlinien vernünftig zu überarbeiten, damit zukünftige Rezensenten sich nicht etwa aufgefordert sehen, auf eine wissenschaftliche Arbeitsweise und den notwendigen Einbezug von Quellen zu verzichten. Wenn Tests so einfach wären, dass alles Notwendige in ein 50seitiges Handbuch passte, bräuchte man vermutlich auch keine Diplompsychologen, um sie anzuwenden.

Fazit

Auch jenseits der im internationalen Kontext absurden Diskussion über die »Vertretbarkeit« seiner Anwendung hat der MMPI als praktisches Routineverfahren viele Stärken und viele Schwächen. Beides erschließt sich dem Leser allerdings nicht nach Lektüre der vorstehenden Rezension – und wohl auch nicht nach dieser Replik. In Angleitners alter Rezension von 1997 und meiner Replik dazu wurden einige wichtige Punkte diskutiert (der MMPI-2 ist nicht so viel anders als der MMPI und vieles von dem dort Gesagten gilt nach wie vor), ansonsten empfehle ich englischsprachiges Material, am besten Greene's neues Buch (2000). Im nächsten Jahr wird mit dem »Verhaltens- und Erlebensinventar«, der deutschen Adaptation des Personality Assessment Inventory von Morey (1991), ein klinischer Fragebogen neu auf den deutschen Markt kommen, der in USA in direkter Konkurrenz zum MMPI-2 steht und bei gleichem Indikationsbereich item- und skalenmetrisch vieles besser macht. Ob er auch in der Praxis der bessere Test ist, kann sich dann zeigen.