FernUniversität in Hagen
Fakultät für Psychologie

bezirkskliniken mittelfranken

Modifikation und Validierung eines Persönlichkeitsinventars für Patienten des Maßregelvollzugs gemäß § 64 StGB (PI-MRV-64)

Projektbeschreibung

Verfasser: Michael Schwarz

Psychologe (M.Sc.)

Klinik für Forensische Psychiatrie

Bezirksklinikum Ansbach

Kontakt: michael.schwarz@bezirkskliniken-mfr.de

Betreuer: Prof. Dr. Andreas Mokros

Lehrgebietsleitung Persönlichkeits-, Rechtspsychologie und Diagnostik

Fakultät für Psychologie

FernUniversität in Hagen

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1. Einleitung: initiale Konzeption des PI-MRV-64

Das deutsche Strafgesetzbuch sieht in § 64 StGB die Unterbringung von substanzabhängigen Straftätern in forensischen Entziehungsanstalten als Maßregel der Besserung und Sicherung vor. Patienten dieser Einrichtungen des Maßregelvollzugs (MRV) durchlaufen eine mehrjährige Sucht- und Kriminaltherapie mit dem Ziel der dauerhaften Abstinenz und Straffreiheit (Schaumburg, 2010, S. 10 – 13). Zur differenzierten Behandlungsplanung ist eine gute Kenntnis der Persönlichkeit der Patienten bedeutsam, da diese den Aufbau, die Stabilität und den Ertrag der therapeutischen Arbeitsbeziehung beeinflusst (Schmidt-Quernheim, 2008, S. 97 – 99; Leygraf, 2006, S. 203 – 206; Schalast, 2014, S. 500 – 502). Im Rahmen einer umfassenden medizinisch-psychologischen Aufnahmediagnostik werden zur standardisierten Erfassung individueller Patientenmerkmale Persönlichkeitsfragebögen (Selbstbeurteilungsinventare) eingesetzt. Diese sind nach wissenschaftlich fundierten, testtheoretischen Kriterien konstruiert und ermöglichen eine ökonomische Messung von Eigenschaften und Einstellungen einer Person. Während im deutschen Sprachraum eine große Anzahl von Selbstbeurteilungsinventaren für die Normalpopulation existiert, wurden im forensischen Kontext vorrangig Instrumente für Inhaftierte in Justizvollzugsanstalten (Klemm, 2002; Seitz & Rautenberg, 2010; Niemeyer & Back, 2017) und zur Erfassung spezifischer Konstrukte (z. B. Aggressivität: Hampel & Selg, 1975; Psychopathie: Alpers & Eisenbarth, 2008) entwickelt. Ein Persönlichkeitsfragebogen, der an und für Patienten des MRV gem. § 64 StGB konstruiert wurde, liegt bislang nicht vor.

Gegenstand der Masterarbeit des Verfassers war die Konzeption und erste Evaluation eines derartigen Fragebogens (Titel: Konstruktion eines Persönlichkeitsinventars für Patienten des Maßregelvollzugs gem. § 64 StGB; Akronym: PI-MRV-64[1]). Das Instrument wurde mit dem Ziel entworfen, latente Merkmale und Einstellungen aus drei Bereichen abzubilden, die bei der Therapieplanung und -durchführung von Bedeutung sind: (a) personeninterne Grundlagen zur Therapiedurchführung, (b) Stellung von Delinquenz und Sucht im Leben eines Patienten und (c) forensisch relevante Persönlichkeitsaspekte einschl. soziale Erwünschtheitstendenz.

Der erstgenannte Bereich wurde als *Therapiegrundlagenbereich* (TGB), der zweitgenannte als *Delinquenz- und Suchtbereich* (DSB), der drittgenannte als *Persönlichkeits- und*

[1] PI-MRV-64: Persönlichkeitsinventar für Patienten des Maßregelvollzugs gem. § 64 StGB.

Kontrollbereich (PKB) bezeichnet. Jedem Bereich wurden drei Skalen zur Messung psychologischer Konstrukte zugeordnet, die den jeweiligen Bereichsschwerpunkt repräsentieren. Die Auswahl der Konstrukte erfolgte auf Basis einer Befragung von $N = 16$ Medizinern und Psychologen der Klinik für Forensische Psychiatrie am Städtischen Klinikum „St. Georg“ Leipzig. Die Fachkräfte wurden gebeten, Persönlichkeitsmerkmale und Einstellungen ihrer Patienten zu nennen, denen sie im therapeutischen Prozess besondere Beachtung schenken. Die erhobenen Konstrukte wurden, ergänzt um weitere Konstrukte aus der Literatur, in einen Ratingfragebogen aufgenommen. Der Ratingfragebogen wurde derselben Expertenstichprobe zur quantitativen Einschätzung der Relevanz der Konstrukte für die Therapieplanung vorgelegt. Bei einer Beurteilungsübereinstimmung von $ICC_{just,MW} = .82$, $F(86, 1290) = 5.68$, $p < .001$, 95 % KI für $ICC_{just,MW}$ [0.77; 0.87] wurden die Konstrukte des Ratingfragebogens gemäß den gemittelten Expertenurteilen pro Konstrukt in eine Rangreihe gebracht. Einem Vorschlag von Caspar und Wirtz (2002, S. 228) folgend, bildeten nicht die Rohwerte, sondern die am Mittelwert der einzelnen Experten *z*-standardisierten Werte die Grundlage der Mittelwertbildung. Jedem der drei Bereiche des zu konstruierenden Inventars wurden die drei am höchsten gerateten Konstrukte aus der Konstruktrangreihe zugeordnet (Tabelle 1).

Tabelle 1

Konzeptionell getrennte Bereiche und Konstrukte des PI-MRV-64 aus der Masterarbeit

Bereich	Konstrukte
Therapiegrundlagenbereich (TGB)	Vertrauen in die Mitarbeiter der MRV-Einrichtung (VM)
	Mitarbeits- und Kooperationsbereitschaft (MK)
	Leidensdruck (LD)
Delinquenz- und Suchtbereich (DSB)	Delinquenzhabitualisierung (DH)
	Substanzaffirmation (SA)
	Verantwortungs- und Schuldabwehr (VS)
Persönlichkeits- und Kontrollbereich (PKB)	Impulsive Destruktivität (ID)
	Interpersoneller Machiavellismus (IM)
	Ehrliche Beantwortung (EB)

Die neun Konstrukte wurden unter Einbezug forensisch-psychologischer Fachliteratur definiert (Konstruktexplikation). Zu ihrer Messung wurden zehn bis 14 Items abgeleitet (Skalenentwürfe). Die Erprobung der Items erfolgte an einer Stichprobe von $N = 57$ Patienten des MRV gem. § 64 StGB der Klinik für Forensische Psychiatrie am Städtischen Klinikum „St. Georg“ Leipzig. Das Antwortverhalten der Probanden bildete die Grundlage einer statistischen Skalenanalyse mittels konfirmatorischer Faktorenanalysen. Zur Konstruktion eindimensionaler (homogener) und messgenauer Skalen wurden in einem schrittweisen Prozess Items mit mangelhaften Kennwerten aus den Skalenentwürfen entfernt. Die finalen Skalen umfassten zwischen fünf und neun Items mit internen Konsistenzen von $.72 \leq \alpha \leq .86$ bzw. $.72 \leq \omega \leq .88$. Für die Skala *Delinquenzhabitualisierung* (DH) wurde ein erster Validitätsnachweis ermittelt: Zwischen den Testwerten auf der Skala und den Einträgen der $N = 57$ Probanden im Bundeszentralregister ergab sich ein hochsignifikanter positiver Zusammenhang von $r(54) = .55$, $p < .001$. Sämtliche Ergebnisse können bei Schwarz (2018) nachgelesen werden. Die Masterarbeit wurde anteilig vom Lehrstuhl für Klinische Psychologie und Psychotherapie sowie dem Lehrstuhl für Persönlichkeitspsychologie und psychologische Diagnostik der Fakultät für Psychologie der Universität Leipzig betreut. Die Veröffentlichung erfolgte über das Publikationsstipendienprogramm der Gesellschaft für Kriminologie, Polizei & Recht e. V.

2. Gegenstand des geplanten Projektes

Mit dem geplanten Projekt wird das Ziel verfolgt, das Persönlichkeitsinventar für Patienten des MRV gem. § 64 StGB zu modifizieren und zu validieren, um es für den praktischen Einsatz im Maßregelvollzug nutzbar zu machen. In Anlehnung an Kriterien, die Kunst (2004, S. 1) für Selbstbeurteilungsverfahren der Forensischen Psychologie formuliert hat, soll der zu konstruierende Fragebogen

- individualdiagnostische Einschätzungen von Persönlichkeitsmerkmalen und Einstellungen erlauben, die mit der Sucht- und Delinquenzproblematik von Patienten des MRV gem. § 64 StGB assoziiert und für die Behandlungsplanung bedeutsam sind,

- zur Zustands- und Verlaufsdiagnostik genutzt werden können,

- durch die Entwicklung an der Zielklientel nach testtheoretischen Standards eine objektive und reliable Differenzierung innerhalb dieser Population ermöglichen,

- konvergent, diskriminant und prognostisch validiert sein,

- zur Lückenschließung zwischen Persönlichkeitsfragebogen für die Normalpopulation und dem JVA-spezifischen Inventar PFI+ (Seitz & Rautenberg, 2010) beitragen,

- bei den Patienten auf Akzeptanz stoßen.

Zusätzlich ist das Nebenkriterium der Einfachheit der Itemformulierungen zu berücksichtigen (Schalast, 2000, S. 58). Haupt- und Nebensatzkonstruktionen, Fachwörter und alltagsferne Ausdrücke sind zu vermeiden. Auch mit der Einschätzung hypothetischer Sachverhalte ist sparsam umzugehen. Die Items sind nach Möglichkeit in der 1. Person Singular abzufassen und inhaltlich direkt auf die lesende Person und ihre Situation zu beziehen. Auf diese Weise soll sichergestellt werden, dass persönliche, selbstbezogene Beurteilungen – im Gegensatz zu unspezifischen, allgemeinen Sichtweisen – erfasst werden. Zur Differenzierung der Patientenantworten wird ein sechsstufiges Antwortformat im Sinne einer endpunktbenannten Likert-Skala (Likert, 1932) mit den Außenkategorien 0 = *trifft gar nicht zu* bis 5 = *trifft vollkommen zu* verwendet. Die Anzahl der Antwortstufen folgt einer Empfehlung von Krosnick und Presser (2010), nach der mit einer fünf- bis siebenstufigen Antwortskala die besten Messresultate erzielt werden. Auf eine Mittelkategorie oder „Weiß nicht"-Kategorie wird verzichtet, einerseits aus Gründen der uneindeutigen Interpretierbarkeit, andererseits, um die Probanden zur Selbstreflexion anzuregen.

2.1 Theoretischer Hintergrund: Risk-Need-Responsivity-Modell

Den theoretischen Rahmen der Modifikation des PI-MRV-64 bildet das *Risk-Need-Responsivity (RNR) Model of Offender Assessment and Treatment* von Bonta und Andrews (2017, S. 175 – 184). Ausgehend von der Allgemeinen Persönlichkeits- und kognitiv-sozialen

Lerntheorie delinquenten Verhaltens (General Personality and Cognitive Social Learning Theory of Criminal Conduct [GPCSL]; ebd., S. 43 – 52) formulieren die Autoren 15 empirisch fundierte Prinzipien, welche sowohl bei der Begutachtung und Therapie von Straftätern als auch der Prävention devianten Verhaltens anzuwenden sind. Im Zentrum stehen hierbei das Bedürfnis-, Risiko- und Ansprechbarkeitsprinzip (Risk-Need-Responsivity Principle; RNR):

- Das **Risikoprinzip** besagt, dass die Behandlungsintensität dem Gefährdungspotenzial und Rückfallrisiko eines Straftäters entsprechen soll. Je höher das Risikopotenzial und je schwerwiegender die zu erwartenden Straftaten, desto intensivere therapeutische Maßnahmen sind angezeigt.

- Dem **Bedürfnisprinzip** zufolge können Rückfälligkeit und Gefährdungspotenzial über die Veränderung kriminogener Risikofaktoren (criminogenic risk/need factors) reduziert werden. Über zahlreiche Einzelstudien und Metaanalysen hinweg erwiesen sich acht Merkmalskategorien Straffälliger als besonders hoch mit dem Aufbau, der Aufrechterhaltung und dem Rückfall in delinquentes Verhalten korreliert: (a) eine Vorgeschichte dissozialen/kriminellen Verhaltens, (b) prodelinquente Einstellungen (Techniken der Neutralisierung, Identifikation mit delinquenten Anderen, Ablehnung von Konventionen bzw. Institutionen), (c) ein prodelinquentes Umfeld, (d) eine dissoziale Persönlichkeitsstruktur, (e) Probleme in der Herkunftsfamilie und/oder Partnerbeziehungen im Erwachsenenalter, (f) Probleme in Schule und Beruf, (g) Substanzmissbrauch/-abhängigkeit sowie (h) unstrukturiertes Freizeitverhalten (ebd., 2017, S. 45 – 46). Während sich manche dieser *Central Eight* auf Ereignisse der Vergangenheit beziehen (z. B. Vorgeschichte dissozialen/kriminellen Verhaltens), welche nicht mehr veränderbar sind, beschreiben andere Merkmale dynamische Risikofaktoren, die gemäß dem Bedürfnisprinzip im Zentrum einer Kriminaltherapie stehen sollen.

- Das **Ansprechbarkeitsprinzip** umschreibt, dass Behandlungsmaßnahmen den Eigenheiten eines Patienten anzupassen sind. Dazu gehört einerseits, die Therapie auf allgemein wirksamen Lernmethoden aufzubauen (behaviorale, kognitive und sozi-

ale Lernstrategien mit dem Ziel des systematischen Kompetenzaufbaus) und andererseits, individuelle Faktoren und Präferenzen zu berücksichtigen (Alter, kultureller Hintergrund, Intelligenz, Stärken, Therapiemotivation und -erfahrung, Beziehungsfähigkeit, Reifegrad, Psychopathy etc.).

Im Gesamten setzt sich das RNR-Modell aus drei übergeordneten Prinzipien (1. Respect for the Person and the Normative Context; 2. Basis on psychological Theory; 3. General Enhancement of Crime Prevention Services), den Kernprinzipien und zentralen klinischen Aspekten (4. Principle of Human Service; 5. Risk; 6. Need; 7. General Responsivity; 8. Specific Responsivity; 9. Breadth/Multimodality; 10. Assessment of Strengths; 11. Structured Assessment; 12. Professional Discretion) sowie drei organisatorischen/betrieblichen Prinzipien (13. Community-based Services; 14. GPCSL-based Staff Practices; 15. Management) zusammen.

2.2 Modifikation der internen Struktur des PI-MRV-64

2.2.1 Semantische und strukturelle Modifikation

Die Bündelung der Skalen des Fragebogens in drei distinkte Bereiche wird beibehalten. Folgende semantische Modifikationen werden vor dem Hintergrund des RNR-Modells vorgenommen:

- Die Skala *Vertrauen in die Mitarbeiter der Maßregelvollzugseinrichtung* wird in *Vertrauen in das Behandlungsteam* (VB) umbenannt und ihre Items überarbeitet. Sie korrespondiert mit den RNR-Prinzipien 1 (Respect for the Person and the Normative Context), 4 (Human Service), 8 (Specific Responsivity) und 14 (GPCSL-based Staff Practices), nach welchen der Qualität der therapeutischen Beziehung insbesondere bei gering therapiemotivierten Straffälligen besondere Bedeutung zukommt.

- Die Skala *Mitarbeits- und Kooperationsbereitschaft* wird in *Kooperation im Therapieprozess* (KT) umbenannt und ihre Items überarbeitet. Sie gibt Aufschluss über

die spezifische Ansprechbarkeit eines Patienten (RNR-Prinzip 8): bei geringer Kooperation sollten Therapiemaßnahmen zur Anwendung kommen, die geeignet sind, persönliche Vorbehalte gegenüber der Behandlung ab- und Mitarbeitsbereitschaft aufzubauen (z. B. motivierende Gesprächsführung; Miller & Rollnick, 2015).

- Die Items der Skala *Leidensdruck* (LD) werden überarbeitet. Die Skala erfasst ebenfalls einen Aspekt der spezifischen Ansprechbarkeit: ein Patient mit hohem Leidensdruck zeigt eine andere Empfänglichkeit gegenüber therapeutischen Maßnahmen als ein Patient, dessen Problemverhalten wenig negativen Einfluss auf sein psychisches Befinden auszuüben scheint. Des Weiteren ist anzunehmen, dass hoher Leidensdruck mit Offenheit für eine größere Spannweite von Behandlungsmaßnahmen einhergeht (RNR-Prinzip 9: Breadth/Multimodality).

- Die Skala *Delinquenzhabitualisierung* wird getilgt. Sie erfasst vergangenes Erleben und Verhalten im Sinne der Central Eight-Kategorie „Vorgeschichte dissozialen/kriminellen Verhaltens“ und ist demnach ungeeignet zur Verlaufsmessung. Sie wird durch die neu zu konstruierende Skala *Identifikation mit delinquentem Lebensstil* (ID) ersetzt. Dieses Maß repräsentiert einen zentralen Risk/Need-Faktor. Es korrespondiert insbesondere mit den Central Eight-Aspekten „Identifikation mit delinquenten Anderen“ und „Ablehnung von Konventionen/Institutionen“ als Bestandteile der Merkmalskategorie „prodelinquente Einstellungen“.

- Über eine Verlängerung der Skala *Substanzaffirmation* (SA) wird die Erhöhung ihrer internen Konsistenz angestrebt. Das Maß bezieht sich auf den Central Eight-Faktor „Substanzmissbrauch“. Allerdings wird nicht das Vorliegen eines schädlichen Gebrauchs oder einer Abhängigkeit von psychotropen Substanzen einschließlich Alkohol eruiert – diese Fragestellung wurde im Eingangsgutachten behandelt und ist Grundlage der Anordnung des § 64 StGB –, sondern die gegenwärtige *Einstellung* eines Patienten zu seinem Hang, berauschende Mittel im Übermaß zu sich zu nehmen. Sie adressiert das RNR-Prinzip 5 „Risk“ (affirmative Haltung zum eigenen Konsum: hohe Rückfallgefahr) und die spezifische Ansprechbarkeit.

- Die Skala *Verantwortungs- und Schuldabwehr* wird in *Bagatellisierungstendenz* (BT) umbenannt. Ihre theoretische Grundlage bilden die fünf von Sykes und Matza (1957, 1968) formulierten Techniken der Neutralisierung (a) Ablehnung der Verantwortung (Denial of Responsibility), (b) Verneinung des Unrechts (Denial of Injury), (c) Ablehnung/Abwertung des Opfers (Denial of the Victim), (d) Verdammung der Verdammenden (Condemnation of the Condemners) und (e) Berufung auf höhere Instanzen (Appeal to higher Loyalties) welche von Bonta und Andrews (ebd., S. 126 – 127) in die Central Eight-Kategorie „prodelinquente Einstellungen" integriert wurden. Über eine Verlängerung der Skala wird die Erhöhung ihrer internen Konsistenz angestrebt.

- Die Skala *Impulsive Destruktivität* wird in *Mangelnde Impulskontrolle* (MI) umbenannt und neu konzipiert. In ihrer ersten Fassung diente sie zur Messung spontaner Aggressivität. Die Neigung eines Patienten, auf Frustrationen und vermeintliche Provokationen mit fremdschädigendem Verhalten zu reagieren, bildet sich jedoch häufig bereits in dessen forensischer Vorgeschichte ab. Wie die klinische Erfahrung zeigt, ist eine weniger offensichtliche Art von Impulsivität im therapeutischen Alltag von größerer Bedeutung: die Eigenschaft, sprunghaft, unreflektiert und ohne Bedacht auf langfristige Konsequenzen zu handeln. Diese Neigung spiegelt sich u. a. in vorzeitigen Therapieabbrüchen, Entweichungen oder Konsumrückfällen wider und entspricht dem Verständnis von Impulsivität im Psychopathy-Konzept nach Hare (2003; Mokros, Hollerbach, Nitschke & Habermeyer, 2017).[2] Die Items der Skala werden der neuen Konzeption angepasst. Eingeordnet in das RNR-Modell geben sie Aufschluss über einen Aspekt der Central Eight-Kategorie „dissoziale Persönlichkeitsstruktur" als zentralen Risk/Need-Faktor sowie die spezifische Ansprechbarkeit und das Prinzip „Strength" (angemessene Impulskontrolle als persönliche Stärke, die im Therapieprozess genutzt werden kann).

- Die Skala *Interpersoneller Machiavellismus* wird allgemeinverständlicher in *Manipulative Beziehungsgestaltung* (MB) umbenannt und ihre Items überarbeitet. Sie

[2] PCL-R-Item 14: Sprunghaftigkeit.

adressiert einen Aspekt der Central Eight-Kategorie „dissoziale Persönlichkeitsstruktur“, der spezifischen Ansprechbarkeit und ist in ähnlicher Weise im Psychopathy-Konzept nach Hare (ebd.) als „Betrügerisch/manipulatives Verhalten“ verankert (PCL-R-Item 5).

- Die neu zu konstruierende Skala *Selbstwerterleben* (SW) wird dem Fragebogen hinzugefügt. Bei der Persönlichkeitsvariable „self-esteem“ handelt es sich nach Bonta und Andrews (2017, S. 181) um ein „noncriminogenic minor need“, d. h. ein Bedürfnis, das in einer Kriminaltherapie allenfalls ergänzend zu den dynamischen Central Eight-Kategorien berücksichtigt werden sollte. Das Selbstwerterleben eines Patienten ist jedoch sowohl bei der Behandlungsplanung als auch im therapeutischen Umgang von Bedeutung – z. B. bei der Frage, ob ein stabilisierender oder ein konfrontativer Interaktionsstil den gegenwärtig größeren Nutzen zur Erreichung der Therapieziele mit sich bringt (RNR-Prinzip 8: spezifische Ansprechbarkeit). Mit der Skala wird ferner dem Aspekt der Ressourcenorientierung Rechnung getragen (RNR-Prinzip 10: Strength). Bei der Formulierung der Items ist darauf zu achten, dass normalpsychologisches, resozialisierungsförderliches Selbstwerterleben erfasst wird, keine Form pathologischer Übersteigerung (z. B. Narzissmus).

- Die Items der Kontrollskala *Ehrliche Beantwortung* (EB) werden überarbeitet. Die Skala wird nicht länger einem der drei Bereiche des Inventars zugeordnet, sondern als Zusatzskala geführt. Ihre Ergebnisse machen Aussagen über die Interpretierbarkeit der Selbstbeurteilung: ein Patient mit niedrigen Werten neigt zu bewusster und/oder unbewusster Selbst- und Fremdtäuschung. Diese Erkenntnis ist wiederum wichtig für den Behandlungsprozess und die Beziehungsgestaltung (RNR-Prinzip 14: GPCSL-based Staff Practices).

- Der *Delinquenz- und Suchtbereich* wird in *Devianzbereich* (DV) umbenannt.

- Der *Persönlichkeits- und Kontrollbereich* wird nach Ausgliederung der Kontrollskala EB in *Persönlichkeitsstilbereich* (PS) umbenannt. Nach Neukonzeption

der Skala *Mangelnde Impulskontrolle* (MI) und Aufnahme der Skala *Selbstwerterleben* (SW) sind in diesem Bereich ausschließlich allgemeinspsychologische Konstrukte (gegenüber den klinisch-forensischen Konstrukten in den Bereichen TG und DV) gebündelt. Nach diesen Modifikationen sind die drei Bereiche konzeptionell klarer voneinander abgegrenzt.

Die modifizierte Struktur ist in Tabelle 2 dargestellt.

Tabelle 2

Modifizierte Struktur des PI-MRV-64 für das geplante Projekt

Bereiche	Skalen
Therapiegrundlagen (TG)	Vertrauen in das Behandlungsteam (VB)
	Kooperation im Therapieprozess (KT)
	Leidensdruck (LD)
Devianz (DV)	Identifikation mit delinquentem Lebensstil (ID)
	Substanzaffirmation (SA)
	Bagatellisierungstendenz (BT)
Persönlichkeitsstile (PS)	Mangelnde Impulskontrolle (MI)
	Manipulative Beziehungsgestaltung (MB)
	Selbstwerterleben (SW)
Kontrollskala	Ehrliche Beantwortung (EB)

2.2.2 Indexwerte

Die Bündelung der Skalen zu drei distinkten Bereichen erfolgte in der Konzeptionsphase der ersten Version des Fragebogens auf Basis theoretischer Überlegungen (s. Abschn. 1). Explorativ soll im Rahmen der geplanten Studie geprüft werden, ob die Skalen entsprechend ihren Bereichen zur Berechnung folgender Indexwerte herangezogen werden können:

- Index Therapiegrundlagen (Index TG) = $\frac{(VB + KT + LD)}{3}$

- Index Devianz (Index DV) = $\frac{(ID + SA + BT)}{3}$

- Index Sozial-funktionaler Persönlichkeitsstil (Index SP) = $\frac{(MI[-] + MB[-] + SW)}{3}$

- Index Intrinsischer Leidensdruck (Index ILD) = $LD - \text{Index DV}$

Der **Index Therapiegrundlagen (Index TG)** zeigt das Vorliegen anerkannter Grundvoraussetzungen zur Durchführung einer Therapie gem. § 64 StGB an. Je höher der TG-Indexwert, desto günstiger werden die patientenspezifischen Voraussetzungen gesehen. Ein hoher Indexwert sollte mit gutem Therapieverlauf einhergehen.

Der **Index Devianz (Index DV)** ist ein Kennwert der Ausprägung kriminogener Haltungen und Einstellungen. Patienten mit hohen DV-Indexwerten zeigen sich subkulturell geprägt (ggf. auch prisonisiert), stehen dem Gebrauch von Rauschmitteln unkritisch gegenüber und neigen zur Rechtfertigung bzw. Verharmlosung ihrer Straftaten. Hohe Indexwerte sollten mit dissozialem Verhalten im Stationsalltag, ablehnender Haltung gegenüber dem Behandlungsprogramm und höheren Raten von Therapieabbrüchen einhergehen. Nach dem RNR-Modell sollte bei Patienten mit hohem Index DV eine entsprechend intensive psychotherapeutische Intervention angesetzt werden.

Der **Index Sozial-funktionaler Persönlichkeitsstil (Index SP)** gibt Aufschluss über die zu erwartende Verträglichkeit eines Patienten. Damit ist sowohl die soziale, nach außen gerichtete Verträglichkeit, als auch die Selbstverträglichkeit (einschl. Selbstkontrolle) gemeint. Ein hoher Indexwert sollte auf eine stabile, verlässliche Patientenpersönlichkeit hindeuten, der Vertrauen entgegengebracht und Konfrontation zugemutet werden kann.

Der **Index Intrinsischer Leidensdruck (Index ILD)** basiert auf dem Modell der Therapiemotivation von inhaftierten Straftätern nach Dahle (1994, 1995). In diesem Modell wird Leidensdruck in die Facetten *Problembelastung*, *Internale Problemverarbeitung* und *Belastung durch den Strafvollzug* unterteilt. Die ersten beiden Facetten beschreibt Dahle (ebd.) als interne Anreize zur Teilnahme an einer Kriminaltherapie (Therapieerfolgsmotiv), während die letztgenannte Facette einen externen Anreiz zur Therapiebeteiligung darstellt (Hafterleichterungsmotiv). Leidensdruck ist demnach intrinsisch, wenn ein Inhaftierter einerseits eine globale Problembelastung (z. B. soziale Konflikte, Schuldgefühle) wahrnimmt und andererseits die Ursache dieser Belastung auf sich selbst und sein Handeln zurückführt („als in ihm liegend erkennt“, Dahle, 1994, S. 228). Zur standardisierten Erfassung von intrinsischem Leidensdruck bei forensischen Populationen führen Carl, Breuer und Endres (2016, S. 12) aus, dass es „[a]lles in allem an zielgruppenspezifischen Verfahren [mangelt], die beide Komponenten des Leidensdrucks berücksichtigen und sowohl die Problembelastung als auch die Form der Problemverarbeitung erfassen.“

Mit dem Index ILD wird der Anteil der intrinsischen Problemverarbeitung eines Patienten des MRV gem. § 64 StGB an seinem globalen Leidensdruck (erfasst mit der Skala LD) spezifiziert. Der Grundgedanke ist, dass sich die Art der Problemverarbeitung in Testwerten auf den Skalen *Identifikation mit delinquentem Lebensstil* (ID), *Substanzaffirmation* (SA) und *Bagatellisierungstendenz* (BT) widerspiegelt. Es wird postuliert, dass ein Patient mit hohen Werten auf der Skala *Leidensdruck* (LD) bei gleichzeitig niedrigen Werten auf den Skalen ID, SA und BT seine Problembelastung auf die eigene Person und das eigenen Handeln bezieht. Im Umkehrschluss würde hoher Leidensdruck in Verbindung mit hohen Werten auf den genannten Skalen dafürsprechen, dass das Ausmaß der Problembelastung durch externe Faktoren (z. B. Belastung durch die Unterbringungssituation) bedingt ist.

2.2.3 Reflektive und formative Indikatoren

Als Grundlagen der Indexwerte werden die Skalen des zu konstruierenden Inventars als *Grundskalen* bezeichnet. Während sich die Grundskalen aus reflektiven Indikatoren (d. h. korrelierenden Items, die exakt ein latentes Konstrukt messen) zusammensetzen, werden die Indexwerte als das Ergebnis von formativen Indikatoren aufgefasst: die Ausprägungen auf den jeweiligen drei Grundskalen *formen* das übergeordnete Indexkonstrukt (Edwards & Bagozzi, 2000, S. 162; Bühner, 2011, S. 34 – 37). Die Bündelung von formativen Indikatoren zur Anzeige eines Konstrukts höherer Ordnung hängt von der Definition und den Facetten dieses Konstrukts ab, nicht – im Gegensatz zu Indikatoren reflektiver Konstrukte – von deren Interkorrelation oder Homogenität.[3] Die jeweiligen drei Grundskalen, aus denen sich die Indexwerte zusammensetzen, müssen nicht dieselbe Eigenschaft messen (nicht homogen/eindimensional sein) und nicht zwingend miteinander korrelieren, da sie auf keiner gemeinsamen Varianzquelle basieren (Welpe, 2014, S. 1021). Die Güte ihrer Eignung zur Erklärung des jeweiligen übergeordneten Indexwertes ist unabhängig davon, welchen Wert ihre Korrelationskoeffizienten im Intervall [-1; +1] annehmen. So wird bei den Grundskalen ID, SA und BT aus theoretischen Überlegungen heraus angenommen, dass sie alle eine Form von gesellschaftlicher Abweichung (Devianz) messen. Das Konstrukt Devianz ist ebenfalls ein latenter Faktor, doch gehen Änderungen in seinen Ausprägungen nicht zwangsläufig mit Änderungen in den Ausprägungen aller drei formativen Indikatoren einher (wie dies bei einer Gruppe reflektiver Indikatoren der Fall wäre, wenn sich Änderungen in den Ausprägungen des zugrundeliegenden latenten Faktors ergeben). Erhöht sich z. B. die Identifikation eines Patienten mit delinquentem Lebensstil von einer Messung 1 zu einer Messung 2, fällt der Index DV bei Messung 2 entsprechend höher aus. Aus einer Erhöhung des Index DV ist jedoch nicht ablesbar, auf welcher der drei zugeordneten Grundskalen sich Erhöhungen ergaben. Es ist sowohl denkbar, dass sich die Ausprägung auf allen drei Skalen moderat erhöht hat, als auch, dass sie auf einer einzigen Skala stark gestiegen, auf den anderen beiden Skalen hingegen leicht gesunken ist.

[3] Beispiel: die formativen Indikatoren Bildungsgrad, Höhe des Einkommens und Reputation des Berufs sind nicht homogen und korrelieren nicht zwangsläufig miteinander, doch sind sie definitorische Facetten des Indexwertes *Sozioökonomischer Status* (Hauser, 1973, S. 268).

2.2.4 Statistisch fundierte Skalenmodifikation

Zur Erstellung eines Testentwurfs werden pro Grundskala ca. 15 Items formuliert. Items aus den finalen Skalenformen der Masterarbeit, welche gute bis sehr gute statistische Kennwerte erreicht haben, werden als Referenz herangezogen, jedoch unter Einbezug klinischer Erfahrung überarbeitet. Aus validierten Skalen etablierter Testverfahren mit ähnlicher Messintention werden weitere Items abgeleitet und in Inhalt und Form an die Zielklientel angepasst. In jedem Fall handelt es sich bei den Skalen um Neukonstruktionen, keine Übernahmen aus der Masterarbeit.

Bezüglich des Umfangs des Testentwurfs ist auf die kognitive und psychische Belastbarkeit der Probanden Rücksicht zu nehmen. Ein Fragebogen mit mehr als 150 Items kann von einem Teil derselben womöglich nicht mit gleichbleibender Aufmerksamkeit, Motivation und Introspektion bearbeitet werden. Die finale Form des Fragebogens ist – unter Berücksichtigung testtheoretischer Kriterien – in ihrer Länge so zu gestalten, dass der Großteil aller deutschsprachigen Patienten des MRV gem. § 64 StGB die Items ohne kognitive Überlastung und bedeutsamem Motivationsverlust in einer einzigen Sitzung bearbeiten kann (ca. 90 – 110 Items; Nebengütekriterium der Zumutbarkeit).

Die theoretisch entworfenen zehn Grundskalen sind an einer Patientenstichprobe empirisch zu erproben. Es werden hierbei allen Probanden alle Items der Grundskalen in randomisierter Form zur Beantwortung vorgelegt. Eine Randomisierung der Items zwischen Probanden(gruppen) ist nicht vorgesehen. Über eine statistische Prüfung mittels konfirmatorischer Faktorenanalysen und anschließendem schrittweisen Itemausschluss sind die Grundskalen dahingehend zu verändern, dass sie homogene, die empirischen Daten repräsentierende und testtheoretischen Kriterien genügende Messmodelle bilden (iterativ-probatorisches Vorgehen). Zur Modellbewertung werden die klassischen Maße der internen Konsistenz Cronbachs Alpha (α; Cronbach, 1951) und McDonalds Omega (ω; McDonald, 1970; 1999, S. 90), korrigierte Trennschärfen r_{it} und Faktorladungen λ, der χ^2-Test sowie die deskriptiven Passungsmaße (Fit-Indizes) RMSEA, SRMR und CFI herangezogen.

Insgesamt soll jede der zehn Grundskalen in ihrer finalen Form folgenden Kriterien genügen (vgl. Hu & Bentler, 1999):

- Reliabilität: interne Konsistenz von $\alpha \geq .70$ bzw. $\omega \geq .70$

- Korrelationen der Items mit dem Testwert der jeweiligen Skala: korrigierte Item-Trennschärfen von $.40 \leq r_{it} \leq .80$

- Differenzierungsfähigkeit zwischen Personen: Faktorladungen von $.40 \leq \lambda \leq .80$[4]

- Kongruenz des finalen Messmodells (Skala) mit dem Populationsmodell (Daten der Probanden): Nichtsignifikanz des χ^2-Tests (Beibehaltung der H_0: Messmodell und Populationsmodell sind identisch)

- Kongruenz der implizierten Kovarianzmatrix Σ des finalen Messmodells mit der empirischen Kovarianzmatrix S: $.05 \leq \text{RMSEA} \leq .08$ (bei $N \leq 250$)

- Niedrige Werte in der Residualmatrix nach Subtraktion $S - \Sigma$: SRMR < .11 (in Kombination mit $\text{RMSEA} \leq .08$)

- Maximale Unterschiedlichkeit des finalen Messmodells von einem Unabhängigkeitsmodell/Nullmodell: $\text{CFI} \geq .95$

Aus der Klassischen Testtheorie abgeleitete Prüfkriterien eigenen sich nicht zur Anwendung bei formativen Indikatoren, da diesen keine gemeinsame Varianzquelle und homogene Struktur zugrunde liegen muss (s. Abschn. 2.2.3). Die Aussagekraft der Indexwerte wird über

[4] Im Einzelfall werden Items mit geringeren Faktorladungen beibehalten, wenn sie gemäß ihrem semantischen Gehalt für die Messung des intendierten Konstrukts wichtig sind. Bei schwerpunktmäßiger Berücksichtigung objektiv-numerischer Kennwerte ist darauf zu achten, dass die Repräsentativität der Items für das jeweilige Zielkonstrukt nicht verloren geht (Sicherstellung über Inhaltsvalidierung; s. Abschn. 2.3.1).

Validitätsnachweise bestimmt (s. Abschn. 2.3.2.1 und 2.3.4); ihr Hauptgütekriterium stellt jedoch die theoretische Plausibilität dar (Welpe, 2019, S. 1021). Grundsätzlich sind die den Indexwerten zugeordneten Grundskalen – gemäß der Natur formativer Indikatoren – je nach Definition der Indexkonstrukte TG, DV, SP und ILD austauschbar (Diamantopoulos & Winklhofer, 2001).

2.2.5 Test-Retest-Reliabilität

Neben Cronbachs Alpha und McDonalds Omega soll als weiteres Gütekriterium der Messgenauigkeit der finalen Skalenformen die Test-Retest-Reliabilität r_{tt} ermittelt werden. Hierzu wird der Testentwurf der gleichen Patientenstichprobe nach einem Zeitintervall ein zweites Mal zur Bearbeitung vorgelegt und der Pearson-Korrelationskoeffizient[5] der Testwerte aus beiden Messzeitpunkten berechnet (Bühner, 2011, S. 61). Die Test-Retest-Reliabilität einer Grundskala ist hoch, wenn die beiden Messungen hoch miteinander korrelieren.

Die Höhe der Korrelation kann durch unsystematische Veränderungen der wahren Werte beeinflusst werden (Schermelleh-Engel & Werner, 2012, S. 123). Neben Lern-, Übungs- und Erinnerungseffekten spielt im vorliegenden Fall vor allem die zeitliche Stabilität der zu messenden Eigenschaften eine Rolle. Es ist zu erwarten, dass nicht alle Zielkonstrukte die Stabilität von Persönlichkeitseigenschaften (*traits*) aufweisen, sondern als aktuelle, potenziell variable Zustände im Therapieprozess (*states*) zu verstehen sind. Patienteneigenschaften, die tief verwurzelt erscheinen (z. B. Identifikation mit delinquentem Lebensstil) können kurzfristigen Änderungen unterliegen, da sie als kriminogene Faktoren im Zentrum einzel- und gruppentherapeutischer Bearbeitung stehen. Das Vertrauen in das Behandlungsteam (VB) kann durch Nichtgewährung von Lockerungsstufen oder kritischen Äußerungen zum Behandlungsverlauf in einer Stellungnahme an die Staatsanwaltschaft gem. § 67e StGB erschüttert werden, was wiederum mit Veränderungen in der Kooperation im Therapieprozess (KT) einhergehen kann. Bei *states* sind aufgrund von deren zeitlicher Variabilität niedrigere Test-Retest-Reliabilitäten

[5] Voraussetzungen: Intervallskalierung der Variablen, Normalverteilung der Residuen, linearer Zusammenhang zwischen den Variablen. Alternatives Verfahren: Rangkorrelation nach Spearman (verteilungsfrei).

erwart- und vertretbar, ohne dass diese ein Hinweis auf mangelnde Skalengüte sind (vgl. Test-Retest-Koeffizienten der State-Ärger-Skalen im *State-Trait-Ärgerausdrucks-Inventar-2*: $.14 \leq r_{tt} \leq .29$; Rohrmann, Hodapp, Schnell, Tibubos, Schwenkmezger & Spielberger, 2013). Grundsätzlich sind niedrige Test-Retest-Korrelationskoeffizienten – bei gleichzeitig hohem Cronbachs α – auch als Hinweis auf die Änderungssensitivität einer Skala interpretierbar.

In Tabelle 3 (s. nächste Seite) wird ein Überblick über hypothetische Stabilitätskategorien (*state* vs. *trait*) der intendierten Zielkonstrukte des PI-MRV-64 gegeben. Ausgehend von der jeweiligen Kategorie wird die Höhe der zu erwartenden Test-Retest-Reliabilität basierend auf einer Literaturreferenz gelistet.

Tabelle 3

Stabilitätskategorie (state vs. trait) der Zielkonstrukte mit zu erwartenden Test-Retest-Reliabilitäten (r_{tt}) der Grundskalen basierend auf Korrelationskoeffizienten vergleichbarer Skalen (Referenz)

Ziel-konstukt	Stabilitäts-kategorie	zu erwar-tende r_{tt}	Referenz
VB	state	≥ .50	r_{tt} = .53 (VTT-TAB-Subskala *Soziales Vertrauen in der therapeutischen Arbeitsbeziehung*; Hewig, 2008)
KT	state	≥ .60	-
LD	trait	≥ .80	r_{tt} = .92 (PAREMO-Subskala *Seelischer Leidens-druck*; Nübling, Kriz, Herwig, Wirtz, Fuchs, Hafen, Töns & Bengel, 2005, S. 94)
ID	trait	≥ .80	r_{tt} = .79 (PSSI-Subskala *selbstbehauptend-antiso-zial*; Kuhl & Kazén, 2009, S. 111)
SA	state	≥ .50	r_{tt} = .48 (TCU-MS-d-Subskala *Problemerkennung*; Buchholz, Glöckner-Rist, Scherbaum & Rist, 2014)
BT	trait	≥ .80	r_{tt} = .91 (HIT; Barriga & Gibbs, 1996)
MI	trait	≥ .80	.81 ≤ r_{tt} ≤ .91 (UPPS-P-Subskalen; Weafer, Baggot & de Wit, 2013)
MB	trait	≥ .80	r_{tt} = .83 (Machiavellismus; Henning & Six, 2014)
SW	trait	≥ .80	r_{tt} = .88 (RSE; Robins, Hendin & Trzesniewski, 2001)
EB	trait	≥ .80	.81 ≤ r_{tt} ≤ .87 (Lügen und Leugnen; Ling, 2014)

Anmerkungen. Die Referenzangaben sind in den meisten Fällen als grobe Richtwerte zu sehen. Es handelt sich mitunter um Skalen, die den zu konstruierenden Grundskalen konzeptuell nur ansatzweise verwandt sind und deren Test-Retest-Reliabilitäten an nicht-forensischen Stichproben mit teils kurzen Retest-Intervallen ermittelt wurden. Ist keine Referenz gelistet, konnte bis dato keine Stabilitätsbestimmung einer vergleichbaren Skala in der Literatur gefunden werden. VTT-TAB: Vertrauens-Trias in der therapeutischen Arbeitsbeziehung, PAREMO: Patientenfragebogen zur Erfassung der Reha-Motivation, PSSI: Persönlichkeits-Stil- und Störungs-Inventar, TCU-MS-d: Deutsche TCU (Texas Christian University)-Behandlungsmotivationsskalen, HIT: „How I Think"-Questionnaire, UPPS-P: Urgency, Premeditation (lack of), Perseverance (lack of), Sensation Seeking, Positive Urgency, Impulsive Behavior Scale, RSE: Rosenberg Self-Esteem Scale.

Allgemein gültige Empfehlungen zur optimalen Länge eines Test-Retest-Intervalls bei Persönlichkeits- und Einstellungstests liegen in der Literatur nicht vor. In der geplanten Untersuchung soll die zweite Bearbeitung des Testentwurfs durch dieselbe Patientenstichprobe nach vier Wochen erfolgen.

2.3 Validierung

Für den Einsatz des modifizierten Fragebogens in der klinisch-forensischen Einzelfalldiagnostik ist sicherzustellen, dass das Instrument diejenigen Konstrukte misst, deren Messung intendiert ist (Validität). In der Testtheorie wird zwischen Inhalts-, Konstrukt- und Kriteriumsvalidität unterschieden. Im Rahmen des geplanten Projektes sollen Validitätsnachweise aus allen drei Bereichen erbracht werden.

2.3.1 Inhaltsvalidität

Der **Inhaltsvalidität** wird, da es sich um ein neuartiges Instrument für eine spezielle, klar definierte Zielpopulation handelt, besondere Beachtung geschenkt. Sie gilt gemeinhin als empirisch nicht prüfbar (Bühner, 2011, S. 61). Streng genommen entspricht jedoch nur Inhaltsvalidität der Grunddefinition von Validität (Murphy & Davidshofer, 2001) – ein Test misst, was er seinem Inhalt gemäß erfasst. Lawshe (1975) schlägt als Versuch zur quantitativen Evaluation der Inhaltsvalidität eines Persönlichkeitsinventars die Bewertung der Testitems durch eine Gruppe von Experten vor. Dieser Ansatz soll in der geplanten Studie zur Anwendung kommen. Als Experten werden Psychologinnen und Psychologen akquiriert, die im Maßregelvollzug gem. § 64 StGB tätig sind.[6] Die Akquise soll schriftlich an mindestens drei bayerischen Maßregelvollzugseinrichtungen erfolgen. Die Experten erhalten im Vorfeld der Erprobung der Skalenentwürfe einen Ratingfragebogen mit den Konstruktexplikationen (Definitionen) der zehn Dimensionen des PI-MRV-64 und den jeweiligen Items. Sie werden gebeten,

[6] Von dieser Berufsgruppe ist zu erwarten, dass sie aufgrund ihrer Ausbildung und ihrer praktischen Tätigkeit mit der Bewertung psychometrischer Erhebungsinstrumente vertraut ist und die Tauglichkeit der Items fundiert einschätzen kann.

die Relevanz der Items zur Messung des jeweiligen Zielkonstrukts auf einer sechsstufigen Likert-Skala mit den Außenkategorien 0 = *gar nicht relevant* und 5 = *sehr relevant* einzuschätzen (MZP IV 1)[7]. Mittels der Intraklassenkorrelationsmethode $ICC_{just,MW}$ (Wirtz & Caspar, 2002, S. 190) bzw. ICC3,*k* (Shrout & Fleiss, 1979) wird die Ähnlichkeit der Ratingurteile der Experten geprüft (Beurteilungsübereinstimmung). Eine hohe $ICC_{just,MW}$ (z. B. $ICC_{just,MW} \geq .80$) zeugt von hoher Beurteilungsübereinstimmung, d. h. die Experten sind sich einig darin, welche Items zur Erfassung des jeweiligen Zielkonstrukts relevant sind. Die Items werden gemäß den am Mittelwert der Ratings der einzelnen Experten *z*-standardisierten Werten pro Skala in eine Rangreihe gebracht und ein Gesamtmittelwert errechnet. Aus den resultierenden zehn Rangreihen ist ablesbar, welche Items nach Expertenmeinung die jeweiligen Konstrukte gut, mittelmäßig oder ungenügend repräsentieren.

Die Rangreihen werden, sofern ihnen hinreichend hohe Beurteilungsübereinstimmungen zugrunde liegen, mit den Ergebnissen der Konfirmatorischen Faktorenanalysen auf Basis der Patientendaten abgeglichen. Dadurch soll sichergestellt werden, dass sich die Grundskalen im Verlauf des statistischen Homogenisierungsprozesses (Itemausschluss) in ihrem semantischen Gehalt nicht von den Zielkonstrukten entfernen (Rossiter, 2008). Bei schwerpunktmäßiger Berücksichtigung objektiv-mathematischer Kennwerte darf die Repräsentativität der Items für ein Konstrukt nicht verloren gehen. Im Einzelfall kann ein Item mit wenig befriedigenden statistischen Kennwerten beibehalten werden, wenn aus den Expertenurteilen sowie „logischen und fachlichen Überlegungen“ (Michel & Conrad, 1982, S. 57) hervorgeht, dass es für die Messung des intendierten Konstrukts hoch relevant ist.

Nach der Skalenhomogenisierung ist erneut eine Prüfung der Inhaltsvalidität durchzuführen. Das Vorgehen bei der Datenerhebung (Akquise etc.) soll analog zum oben genannten sein. Zu diesem zweiten Erhebungszeitpunkt (MZP IV 2) werden den Experten jedoch nur die selektierten Items, d. h. die Items der finalen Skalenformen, vorgelegt. Erneut ist die Repräsentativität der Items für die jeweiligen Zielkonstrukte einzuschätzen. Zusätzlich zu Beurteilungsübereinstimmung und Itemrangreihe wird wiederum ein Gesamtmittelwert der Expertenratings pro Skala gebildet.

[7] MZP IV 1 = Messzeitpunkt Inhaltsvalidität 1.

Von Inhaltsvalidität der finalen Skalenformen ist zu sprechen, wenn zu MZP IV 2

- ein deskriptiver Anstieg in den Beurteilungsübereinstimmungswerten gegenüber MZP IV 1 erreicht wird ($ICC_{MZP\ IV\ 2} > ICC_{MZP\ IV\ 1}$; Signifikanzprüfung nicht erforderlich),

- die Expertenurteile mittlere bis hohe Relevanz der einzelnen Items anzeigen ($\mu \geq 2.50$ bei sechsstufiger Likert-Skala),

- der Gesamtmittelwert der Ratings pro Skala gegenüber MZP IV 1 gestiegen ist (H_0: $\mu_2 = \mu_1$; H_1: $\mu_2 > \mu_1$).[8]

Der Verfasser ist sich darüber im Klaren, dass es sich bei den erhobenen Daten um Mehrebenen- bzw. hierarchische Daten handelt. Die Experten (Level-1-Daten) sind in Kliniken (Level-2-Daten) geschachtelt. Bei der Frage, ob bei der Auswertung statt klassischer statistischer Tests eine Mehrebenenanalyse (Multilevel-Analyse) zur Anwendung kommen sollte, ist zu berücksichtigen, dass diese sehr elaborierte Form statistischen Testens hohe Anforderungen an die Daten stellt. Mehrebenenanalysen erfordern zur korrekten Schätzung der Modellparameter eine angemessene Anzahl an Level-2-Einheiten. In der Literatur sind Empfehlungen von mindestens $N \geq 20$ (Snijders & Bosker, 2011, S. 48) bzw. $N \geq 30$ (Kreft & De Leeuw, 1998) zu finden, Maas und Hox (2005) bezeichnen Stichprobengrößen von $N \geq 50$ als „small sample size“ (ebd., S. 86).

[8] Prüfung mittels Zweistichproben-*t*-Test für unabhängige Stichproben, Welch-Test (bei Heteroskedastizität der Stichproben) oder Mann-Whitney-*U*-Test (bei fehlender Normalverteilung oder kleinen Fallzahlen). Eine Signifikanz bzw. eine hohe Effektstärke bei Signifikanz des Anstiegs des Gesamtmittelwerts ist bei Skalen, die bereits zu MZP IV 1 einen hohen Gesamtmittelwert erreicht haben, nicht erforderlich (Deckeneffekt). Es wird von unabhängigen Stichproben ausgegangen, da nicht anzunehmen ist, dass zu MZP IV 2 die gleichen Experten und die gleiche Anzahl an Experten wie zu MZP IV 1 erhoben werden (anonymisierte Befragung).

Übertragen auf die geplante Studie müssten zur Schätzung eines Mehrebenenmodells Experten von mindestens $N = 20$, besser $N \geq 50$ Maßregelvollzugseinrichtungen akquiriert werden – eine unrealistisch hohe Zahl.[9] Eine Mehrebenenanalyse ist darüber hinaus aus inhaltlicher Sicht nicht zwingend erforderlich: die Bewertung der Items ist ausschließlich im Hinblick auf die Passung mit den explizit vorgegebenen Konstruktdefinitionen vorzunehmen. Ein Einfluss der Klinikzugehörigkeit auf die Bewertung ist nicht a priori zu erwarten. Da dieser Punkt dennoch von Interesse ist, soll mit einer univariaten einfaktoriellen Varianzanalyse mit Zufallseffekten geprüft werden, ob signifikante Unterschiede in den Mittelwerten der Expertenratings pro Skala in Abhängigkeit von der Klinikzugehörigkeit vorliegen.[10] Die H_0 lautet: Die Mittelwerte der Itembewertungen der Expertengruppen unterscheiden sich nur zufällig (H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$). Die H_1 lautet: Zwischen mindestens zwei der Expertengruppen besteht hinsichtlich der Itembewertungen ein Unterschied. Eine Zurückweisung der H_0 und Annahme der H_1 spricht für einen Einfluss der Klinikzugehörigkeit. In diesem Fall wird mit multiplen Post-Hoc-Mittelwertvergleichen (z. B. Tukey oder Scheffé) explorativ bestimmt, zwischen welchen Kliniken Unterschiede vorliegen. Ebenso sind Varianzaufklärung (korrigiertes R^2) und Effektstärke (partielles η^2) zu betrachten.

[9] Bayern: 14 MRV-Einrichtungen insgesamt, davon eine Forensik für Jugendliche und Heranwachsende, eine Forensik für Patientinnen, eine Forensik ohne § 64-Patienten = 11 infrage kommende Kliniken.

[10] Faktor: Klinikzugehörigkeit. Kategoriale unabhängige Faktorstufen: Kliniken. Intervallskalierte abhängige Variable: Mittelwerte der Expertenratings pro Skala. Zufallseffekte: die Kliniken sind eine Zufallsstichprobe aus der Gesamtheit aller deutschen MRV-Kliniken. Voraussetzungen: Normalverteilung der abhängigen Variablen in allen Gruppen (Normalverteilung der Fehler), Homoskedastizität (gleiche Varianz der abhängigen Variablen in allen Gruppen). Die einfaktorielle ANOVA entspricht mathematisch der einfachsten Mehrebenenanalyse, dem *Random-Intercept-Only-Modell*.

2.3.2 Kriteriumsvalidität

Als Kriteriumsvalidität wird der Zusammenhang der Ergebnisse eines psychometrischen Tests mit Außenkriterien bezeichnet, mit denen der Test aufgrund seines Messanspruchs korrelieren sollte (Bühner, 2011, S. 63). Es wird zwischen prognostischer, retrospekiver und Übereinstimmungsvalidität unterschieden.

2.3.2.1 Prognostische Validität

Zur Ermittlung der **prognostischen Validität** (auch: Vorhersagevalidität) werden Zusammenhänge der Testergebnisse mit zeitlich später erhobenen Kriterien ermittelt. Im vorliegenden Fall wird der Zusammenhang zwischen Testwerten auf den zehn Grundskalen bzw. den vier Indexwerten mit späterer Beantragung der Erledigung der Maßregel wegen Aussichtslosigkeit gem. § 67d (5) StGB[11] betrachtet. Diese Form des Therapieendes wird vom Behandlungsteam eines Patienten bei der zuständigen Staatsanwaltschaft angeregt, wenn aus Gründen, die in der Person des Untergebrachten liegen, innerhalb der gesetzlichen Unterbringungsfrist keine hinreichende Aussicht auf regelhafte Beendigung der Unterbringung mit bedingter Entlassung (= Therapieerfolg) besteht.

Es wird angenommen, dass

- hohe Werte auf folgenden Grundskalen und Indexwerten mit geringerer Wahrscheinlichkeit einer späteren BEA einhergehen (BEA_0-Skalen):
    - Vertrauen in das Behandlungsteam (VB)
    - Kooperation im Therapieprozess (KT)
    - Leidensdruck (LD)
    - Index Therapiegrundlagen (Index TG)
    - Index Sozial-funktionaler Persönlichkeitsstil (Index SP)
    - Index intrinsischer Leidensdruck (Index ILD)

[11] Im Folgenden abgekürzt BEA (= Beantragung der Erledigung der Maßregel wegen Aussichtslosigkeit) genannt.

- hohe Werte auf folgenden Grundskalen mit erhöhter Wahrscheinlichkeit einer späteren BEA einhergehen (BEA_1-Skalen):
    - Identifikation mit delinquentem Lebensstil (ID)
    - Substanzaffirmation (SA)
    - Bagatellisierungstendenz (BT)
    - Mangelnde Impulskontrolle (MI)
    - Manipulative Beziehungsgestaltung (MB)
    - Index Devianz (Index DV)

Zwischen den Werten auf der Grundskala *Selbstwerterleben* (SW) und der Kontrollskala *Ehrliche Beantwortung* (EB) werden keine spezifischen Vorhersagen in Bezug auf spätere BEA formuliert. Es sind Effekte in beide Richtungen denkbar, die empirischen Gegebenheiten sind explorativ zu ermitteln.

Zu prüfen sind die formulierten Zusammenhänge über eine binär-logistische Regression[12] (Methode: Einschluss) mit Kreuzvalidierung und Betrachtung von Sensitivität und Spezifität. Mit der logistischen Regression wird die Wahrscheinlichkeit des Auftretens der binären Variable BEA (P($Y = 0$): BEA nein; P($Y = 1$): BEA ja) durch die Testwerte auf den Grundskalen sowie der Indexwerte (Prädiktoren) vorhergesagt (Field, 2018, S. 880). Die Modellgüte wird mit den globalen Fit-Maßen χ^2 und Nagelkerkes R^2_N bewertet. Über eine Klassifizierungstabelle werden die Trefferraten und der Anteil fehlklassifizierter Fälle bestimmt. Die Vorhersagegüte der einzelnen Prädiktoren wird mit den Regressionskoeffizienten b_1 und den Odds Ratios (*Exp*(B)) evaluiert. Odds Ratios sind Maße der Stärke des Effekts eines Prädiktors, die bei einem Wert von *Exp(B)* = 1 minimal ausfällt (Diehl & Staufenbiel, 2007, S. 487).

[12] Voraussetzungen: keine Ausreißerwerte, keine Multikollinearität der Prädiktorvariablen (Prüfung mittels Toleranzwert oder Varianzinflationsfaktor), Linearität des Logits.

Folgende Kenngrößen werden betrachtet:

- Globale Gütemaße:

    - Likelihood Ratio-Test. H_0: der χ^2-Test ist nicht signifikant: Das spezifizierte Modell mit Prädiktoren ist gleich einem Modell, das nur die Regressionskonstante enthält (Nullmodell). H_1: Die Hinzunahme der Prädiktoren trägt signifikant zur Verbesserung der Modellanpassung bei. Inhaltlich: Die dichotome Kriteriumsvariable BEA wird von den Grundskalen und Indexwerten besser vorhergesagt als durch ein Modell ohne Prädiktoren.

    - Pseudo-R^2-Statistik. H_0: Nagelkerkes $R^2_N = .00$: Das spezifizierte Modell erklärt keinerlei Varianz der Kriteriumsvariable. H_1: Nagelkerkes $R^2_N > .00$: Durch die Hinzunahme der Prädiktoren steigt die Erklärung von Varianz der Kriteriumsvariable ($R^2_N \geq .20$: akzeptabel, $R^2_N \geq .40$: gut, $R^2_N \geq .50$: sehr gut).

    - Klassifizierung. Sensitivität: auf Basis des spezifizierten Modells wird ein angemessen hoher Anteil von Patienten mit späterer BEA richtig erkannt (Richtig-positiv-Rate; Trefferquote). Spezifität: auf Basis des spezifizierten Modells wird ein angemessen hoher Anteil von Patienten ohne spätere BEA richtig erkannt (Richtig-negativ-Rate). Der Prozentsatz der richtig klassifizierten Fälle am Gesamtprozentsatz soll hoch, der Prozentsatz an Fehlklassifizierungen niedrig sein.

    - Kreuzvalidierung. Per Zufallsauswahl werden 50 % der Probanden einer Kreuzvalidierungsstichprobe zugewiesen. Anhand der anderen 50 % der Fälle wird die logistische Regression berechnet. Die Regressionsgleichung wird an der Kreuzvalidierungsstichprobe überprüft. Für die Stabilität der Gleichung (und damit der Vorhersagegüte des Modells) spricht, wenn sich in der

Kreuzvalidierungsstichprobe ein ähnliches Klassifizierungsergebnis (Sensitivität und Spezifität) ergibt wie in der Originalstichprobe (Diehl & Staufenbiel, 2007, S. 502).[13]

- Lokale Gütemaße:

  - Regressionsgewichte b_1 bei BEA_0-Grundskalen und -Indexwerten: H_0: $b_1 = .00$: Zwischen der Ausprägung eines BEA_0-Prädiktors und der Wahrscheinlichkeit $P(Y = 1)$ besteht kein Zusammenhang. H_1: $b_1 < .00$: Der Anstieg des Wertes eines BEA_0-Prädiktors führt zur Verringerung der Wahrscheinlichkeit von $P(Y = 1)$. Inhaltlich: Höhere Testwerte auf BEA_0-Grundskalen und -Indexwerten bedeuten eine Abnahme der Wahrscheinlichkeit einer späteren BEA.

  - Regressionsgewichte b_1 bei BEA_1-Grundskalen und -Indexwerten: H_0: $b_1 = .00$: Zwischen der Ausprägung eines BEA_1-Prädiktors und der Wahrscheinlichkeit von $P(Y = 1)$ besteht kein Zusammenhang. H_1: $b_1 > .00$: Der Anstieg des Wertes eines BEA_1-Prädiktors führt zur Erhöhung der Wahrscheinlichkeit von $P(Y = 1)$. Inhaltlich: Höhere Testwerte auf BEA_1-Grundskalen und -Indexwerten bedeuten eine Zunahme der Wahrscheinlichkeit einer späteren BEA.

  - Odds Ratios *Exp*(B) bei BEA_0-Grundskalen und -Indexwerten: H_0: *Exp*(B) = 1.00: Steigt ein BEA_0-Prädiktor um eine Einheit und werden alle anderen Prädiktoren konstant gehalten, ergibt sich keine Veränderung der relativen Wahrscheinlichkeit von $P(Y = 1)$. H_1: *Exp*(B) < 1.00: Steigt ein BEA_0-Prädiktor um eine Einheit und werden alle anderen Prädiktoren konstant gehalten, ergibt sich eine Verringerung der relativen Wahrscheinlichkeit von $P(Y = 1)$; sie sinkt um den Faktor *Exp*(B). Inhaltlich: Bei einer Erhöhung des Testwertes eines BEA_0-Prädiktors um eine Einheit verringert sich die relative Wahrscheinlichkeit einer späteren BEA um den Faktor *Exp*(B).

[13] Die Kreuzvalidierung wird nur durchgeführt, wenn sich in allen Teilstichproben angemessen viele Fälle befinden. Als Faustregel gilt für jede Ausprägung des dichotomen Kriteriums $N \geq 25$ in jeder Teilstichprobe.

- Odds Ratios *Exp*(B) bei BEA_1-Grundskalen und -Indexwerten: H_0: *Exp*(B) = 1.00: Steigt ein BEA_1-Prädiktor um eine Einheit und werden alle anderen Prädiktoren konstant gehalten, ergibt sich keine Veränderung der relativen Wahrscheinlichkeit von P($Y = 1$). H_1: *Exp*(B) > 1.00: Steigt ein BEA_1-Prädiktor um eine Einheit und werden alle anderen Prädiktoren konstant gehalten, ergibt sich eine Erhöhung der relativen Wahrscheinlichkeit von P($Y = 1$); sie steigt um den Faktor *Exp*(B). Inhaltlich: Bei einer Erhöhung des Testwertes eines BEA_1-Prädiktors um eine Einheit erhöht sich die relative Wahrscheinlichkeit einer späteren BEA um den Faktor *Exp*(B).

Für die zehn Grundskalen und die drei Indexwerte TG, DV und SP werden getrennte Regressionsanalysen durchgeführt (= zwei Analysen: 1. Grundskalen; 2. Indexwerte). Andernfalls gehen Testwerte doppelt in die Berechnung ein, da sich die Indexwerte aus den Grundskalen formieren (Folge: hohe Multikollinearität der Prädiktoren). Der Einfluss des Index ILD ist demnach gänzlich separat zu ermitteln.

Es ist an dieser Stelle zu betonen, dass mit dem zu konstruierenden Fragebogen kein Prognoseinstrument geschaffen werden soll. Zeigen die Analyseergebnisse, dass mit (einzelnen) Grundskalen und Indexwerten zuverlässige prognostische Einschätzungen möglich sind, so erhöht dies erfreulicherweise die Anwendungsbreite des Verfahrens. Grundsätzlich ist jedoch *Persönlichkeitsmessung zur Therapieplanung* intendiert; fehlende Zusammenhänge zwischen Testergebnissen und späterer Beantragung der Erledigung der Maßregel wegen Aussichtslosigkeit sprechen nicht zwangsläufig gegen die Validität der Skalen, sondern sind unter dem Zeitbezug der Messungen zu interpretieren. Erreicht ein Patient zu Beginn der Therapie hohe Werte auf der Grundskala *Identifikation mit delinquentem Lebensstil* (ID) und wird drei Jahre später nach regelhaftem Lockerungsprocedere aus der Maßregelvollzugseinrichtung auf Bewährung entlassen, so deutet dies weniger auf Invalidität der Grundskala ID, als vielmehr auf erfolgreiches therapeutisches Bearbeiten der anfangs hohen delinquenten Identifikation hin.

2.3.2.2 Übereinstimmungsvalidität

Bei der **Übereinstimmungsvalidität** (auch: konkurrente/diagnostische Validität) werden Zusammenhänge der Testwerte mit zeitgleich erfassten Kriterien berechnet.

Im vorliegenden Fall sollen Fremdeinschätzungen der mit den Grundskalen erfassten Merkmale durch die Bezugstherapeuten[14] der Probanden erhoben werden. Den Bezugstherapeuten werden die Konstruktexplikationen (Definitionen) der Zielkonstrukte der Skalen vorgelegt. Sie werden gebeten, auf einer sechsstufigen endpunktbenannten Likert-Skala (Außenkategorien 0 = *trifft gar nicht zu* und 5 = *trifft vollkommen zu*) einzuschätzen, inwiefern die jeweils beschriebene Eigenschaft auf ihre an der Untersuchung teilnehmenden Patienten zutrifft. Stärke und Richtung des Zusammenhangs zwischen der Fremdeinschätzung und der Selbsteinschätzung (Testwerte der Probanden auf den Grundskalen) werden über partielle Pearson-Korrelationen mit der bisherigen Dauer der therapeutischen Arbeitsbeziehung (in Wochen) als Kontrollvariable[15] ermittelt. Die H_0 lautet: Zwischen der Selbst- und Fremdeinschätzung besteht kein Zusammenhang, $r = .00$. Die gerichtete H_1 lautet: Zwischen Selbst- und Fremdeinschätzung besteht ein positiver linearer Zusammenhang, $r > .00$. Inhaltlich: Selbst- und Fremdeinschätzung weisen Übereinstimmung auf, weil sie sich auf dasselbe latente Konstrukt beziehen. Die Skalen gelten als umso valider, je höher die jeweiligen Korrelationen ausfallen (Effektstärken). Orientiert an den klassischen Konventionen nach Cohen (1988) ist mindestens ein Effekt von $r \geq .30$ (mittlere Effektstärke) wünschenswert, ab $r \geq .50$ wäre von

[14] Der Begriff *Bezugstherapeut* wird, wie im Maßregelvollzug üblich, für alle Mitarbeiterinnen und Mitarbeiter verwendet, die einen Untergebrachten im therapeutischen Einzelsetting betreuen. Hierbei ist es unerheblich, ob es sich um Mitarbeitende des Ärztlichen, Psychologischen oder Sozialpädagogischen Dienstes handelt.

[15] Die Qualität einer Personenbeurteilung ist abhängig von der Dauer der Bekanntschaft des Beurteilenden mit der zu beurteilenden Person (Connelly & Ones, 2010; Asendorpf & Neyer, 2012, S. 97). Durch Herausrechnen der Dauer der bisherigen therapeutischen Zusammenarbeit aus dem Zusammenhang von Selbst- und Fremdeinschätzung wird für diese Einflussgröße kontrolliert. Voraussetzungen der partiellen Korrelation: Intervallskalierung der Variablen, Normalverteilung der Residuen, linearer Zusammenhang zwischen den Variablen. Alternatives Verfahren: Rangkorrelation nach Spearman (verteilungsfrei).

einem starken Effekt und somit hoher Übereinstimmungsvalidität zu sprechen. In einer Metaanalyse konnten Gignac und Szorodai (2016) jedoch zeigen, dass Effektstärken von $r \geq .50$ meist unrealistisch hoch sind, weshalb die empirisch ermittelten Werte $r \geq .10$, $r \geq .20$ und $r \geq .30$ als Richtlinien für niedrige, mittelgradige und hohe Effektstärken herangezogen werden sollten. In der Domäne von Selbst- und Fremdbeurteilungen werden bei der Messung von stabilen Persönlichkeitseigenschaften mit validen Instrumenten typischerweise Korrelationen von $r = .36$ erreicht (Connolly, Kavanagh & Viswesvaran, 2007).

Als Mischform von Übereinstimmungsvalidität und konvergenter Konstruktvalidität (s. Abschn. 2.3.3.1) werden die Testwerte der Grundskalen mit Fremdeinschätzungen vergleichbarer Konstrukte aus etablierten forensischen Verfahren korreliert. Herangezogen werden die Instrumente PCL-R (Psychopathy Checklist–Revised; Mokros, Hollerbach, Nitschke & Habermeyer, 2017) und HCR-20 (Historical-Clinical-Risk Management-20 Violence Risk Assessment Scheme; Webster, Douglas, Eaves & Hart, 1997; dt. Bearb.: Müller-Isberner, Jöckel & Cabeza, 1998). Die Antwortkodierung der Items erfolgt bei beiden Inventaren auf einer dreifach gestuften ordinalen Beurteilungsskala (0 = *trifft nicht zu*, 1 = *trifft in gewissem Ausmaß/teilweise zu*, 2 = *trifft völlig zu*). Als Zusammenhangsmaß wird die Rangkorrelation nach Spearman berechnet.[16] Die Hypothesen sind analog zu den im vorigen Absatz genannten. In Tabelle 4 (s. nächste Seite) sind den Grundskalen geeignete Validierungsitems aus PCL-R und HCR-20 mit erwarteten Zusammenhängen gegenübergestellt.

[16] Lineare, verteilungsfreie Korrelation einer metrischen und einer ordinal skalierten Variable.

Tabelle 4

Den Grundskalen zugeordnete Validierungsitems aus PCL-R und HCR-20 mit erwarteten Korrelationen (r_{xy})

Grundskalen	Validierungsitems	Quelle	erwartete r_{xy}
Vertrauen in das Behandlungsteam (VB)[a]	-	-	-
Kooperation im Therapieprozess (KT)	Fehlende Compliance (Item R4)	HCR-20	$\rho \leq -.20$
Leidensdruck (LD)	Mangel an Einsicht (Item C1)	HCR-20	$\rho \leq -.10$
Identifikation mit delinquentem Lebensstil (ID)	Negative Einstellungen (Item C2)	HCR-20	$\rho \geq .20$
Substanzaffirmation (SA)[a]	-	-	-
Bagatellisierungstendenz (BT)	Fehlende Verantwortungsübernahme für eigenes Handeln (Item 16)	PCL-R	$\rho \geq .30$
Mangelnde Impulskontrolle (MI)	Sprunghaftigkeit (Item 14)	PCL-R	$\rho \geq .30$
Manipulative Beziehungsgestaltung (MB)	Betrügerisch/Manipulativ (Item 5)	PCL-R	$\rho \geq .20$
Selbstwerterleben (SW)	Grandioses Selbstwertgefühl (Item 2)	PCL-R	$\rho \geq .30$
Ehrliche Beantwortung (EB)	Pathologisches Lügen (Item 4)	PCL-R	$\rho \geq -.30$

[a] Kein inhaltlich vergleichbares Item in den Validierungsverfahren vorhanden. Mit dem HCR-20-Item *Substanzmissbrauch* (H5) wird eine deskriptive Aussage über das Vorliegen eines missbräuchlichen Konsums legaler und illegaler Rauschmittel gemacht, keine Einschätzung der emotionalen/kognitiven Bewertung des Konsums.

Die zum Teil niedrig anmutenden Korrelationen werden nicht aufgrund ungenügender Reliabilität oder Validität der Items und Skalen erwartet, sondern müssen infolge von Methodenfaktoren (Selbst- vs. Fremdbeurteilung), Kriteriumskontamination und -defizienz (Grundskalen und Validierungsitems überlappen sich inhaltlich nur teilweise) sowie mangelnder Symmetrie (Grundskalen und Validierungsitems erfassen Merkmale unterschiedlich breit) angenommen werden (Bühner, 2011, S. 68 – 69). Während *Mangelnde Impulskontrolle* (MI) und *Sprunghaftigkeit* (PCL-R-Item 14) einander inhaltlich entsprechen, weisen *Leidensdruck* (LD) und *Mangel an Einsicht* (HCR-20-Item C1) nur eine grobe konzeptionelle Schnittmenge auf. Im Fall von *Selbstwerterleben* (SW) und *Grandioses Selbstwertgefühl* (PCL-R-Item 2) erfasst das Validierungsitem eine pathologische Überzeichnung des Zielkonstrukts der Grundskala. Ferner beziehen sich die Validierungsitems – gemäß den Beschreibungen in den Testmanualen – mitunter auf vergangenes Verhalten, während die Items der Grundskalen auf Merkmale und Einstellungen der Gegenwart ausgerichtet sind. Kriteriumsvalidität liegt nach einer Daumenregel von Bühner (2011, S. 80) vor, wenn Korrelationsnachweise mit relevanten Kriterien in Höhe von $r > .20$ erbracht werden können.

2.3.2.3 Retrospektive Validität

Retrospektive Validität (Zusammenhänge von Testwerten mit objektiven, zeitlich früher vorliegenden Kriterien) wird für die Grundskalen ID, SA, BT und den Index DV über Korrelationen mit den Eintragungen der Probanden im Bundeszentralregister (BZR) ermittelt. Die H_0 lautet: Zwischen den Eintragungen im BZR und den Testwerten auf den Grundskalen ID, SA, BT und dem Index DV gibt es keinen Zusammenhang, $r = .00$. Die gerichtete H_1 lautet: Zwischen den Eintragungen im BZR und den Testwerten auf den Grundskalen ID, SA, BT und dem Index DV besteht ein positiver linearer Zusammenhang, $r > .00$. Inhaltlich: Die delinquente Sozialisation eines Probanden spiegelt sich in den Ausprägungen der Devianz-relatierten Skalen wider.

Hinsichtlich der Effektgrößen sind mittlere bis hohe Werte zu erwarten. Bei der vormaligen Konzeptualisierung der Skala *Identifikation mit delinquentem Lebensstil* (ID) als *Delinquenzhabitualisierung* ergab sich bei Korrelation der Testwerte mit den BZR-Einträgen von $N = 57$ Probanden ein hochsignifikanter starker Effekt von $r(54) = .55$, $p < .001$ (Schwarz,

2018, Kap. 4.4.6). In der geplanten Studie ist bei der Grundskala ID aufgrund ihres stärkeren kriminobiographischen Bezugs eine höhere Effektstärke ($.30 \leq r \leq .50$) zu erwarten als bei den Grundskalen SA und BT ($.20 \leq r \leq .40$).

Zur Ermittlung der retrospektiven Validität werden Probanden in einer frühen Phase der Therapie einbezogen (z. B. in der Aufnahmephase oder am Beginn der intensivtherapeutischen Behandlung). Mit zunehmender Anzahl an Interventionseinheiten ist ein Abbau devianter Denk- und Handlungsstile zu erwarten. Der Zusammenhang zwischen entsprechenden Selbstaspekten und aktenkundigen Gesetzesverstößen dürfte bei längerer Unterbringungdauer zusehends verblassen, was sich mindernd auf die retrospektiven Validitätskoeffizienten auswirken würde.

2.3.3 Konstruktvalidität

Ein psychometrischer Test weist Konstruktvalidität auf, wenn er dasjenige Merkmal misst, dessen Messung intendiert ist (Bühner, 2011, S. 63). Es wird zwischen konvergenter, diskriminanter und faktorieller Konstruktvalidität unterschieden.[17]

2.3.3.1 Konvergente Konstruktvalidität

Zur Bestimmung von **konvergenter Konstruktvalidität** werden Korrelationen mit Tests ähnlicher Messintention bzw. eines ähnlichen Gültigkeitsbereichs bestimmt, für die bereits Validitätsnachweise vorliegen. In Tabelle 5 (s. nächste Seite) wird wiedergegeben, mit welchen publizierten Selbstbeurteilungsverfahren die Grundskalen konstruktvalidiert werden sollen.

[17] Teilweise werden in der Literatur Inhalts- und Konstruktvalidität unter Konstruktvalidität, konvergente und diskriminante Validität unter Kriteriumsvalidität gefasst (z. B. Asendorpf & Neyer, 2012, S. 98).

Tabelle 5

Den Grundskalen zugeordnete Validierungsskalen mit Angabe von Itemanzahl, Cronbachs Alpha (α) bzw. McDonalds Omega (ω) und Testautoren

Grundskala	Validierungsskala	Item-anzahl	α/ω	Testautoren
Vertrauen in das Behandlungsteam (VB)	Soziales Vertrauen in der Therapie	14	α = .91	Hewig, Hank, & Krampen (2009)
Kooperation im Therapieprozess (KT)	Vorsatz zu kooperieren	7	α = .68	Schalast (2000)
Leidensdruck (LD)	Leidensdruck	9	α = .75	Carl, Breuer & Endres (2014)
Identifikation mit delinquentem Lebens-stil (ID)	Selbstbehauptend-antisozialer Stil	10	α = .85	Kuhl & Kazén (2009)
Substanzaffirmation (SA)	Problemerkennung	9	α = .85	Buchholz, Glöck-ner-Rist, Scher-baum & Rist (2014)
Bagatellisierungsten-denz (BT)	Schuldexternalisierung	15	α = .88	Alpers & Eisen-barth, 2008
Mangelnde Impulskontrolle (MI)	Impulsives-Verhalten-8	8	.70 ≤ ω ≤ .87	Kovaleva, Beier-lein, Kemper, & Ramstedt (2014)
Manipulative Bezie-hungsgestaltung (MB)	Machiavellistischer Egoismus	15	α = .68	Alpers & Eisen-barth (2008)
Selbstwerterleben (SW)	Rosenberg Self-Esteem Scale	10	.84 ≤ α ≤ 85.	Von Collani & Herzberg (2003)
Ehrliche Beantwortung (EB)	Soziale Erwünschtheit-Gamma	6	.71 ≤ ω ≤ .78	Kemper, Beierlein, Bensch, Kovaleva, & Rammstedt (2014)

Für die zu erwartenden Höhen der Pearson-Korrelationen gelten ähnliche Einschränkungen wie die in Abschn. 2.3.2.2 genannten. Während Methodeneffekte hier nicht zu erwarten sind, mindern Messfehler (1 – α) die Korrelationsstärken. Durch Anwendung der doppelten Minderungskorrektur werden die Größen der Korrelationen der gemessenen Eigenschaften auf Konstruktebene ermittelt (Asendorpf & Neyer, 2012, S. 93). Bühner (2011, S. 80) zufolge ist konvergente Konstruktvalidität gegeben, wenn Korrelationen mit verwandten Konstrukten in Höhe von $r > .50$ vorliegen. Während im Fall einiger Grundskalen bereits ähnliche Skalen publiziert wurden, die eine zuverlässige Validierung erlauben, konnten für Grundskalen wie *Identifikation mit delinquentem Lebensstil* (ID) keine Äquivalente im deutschen Sprachraum gefunden werden. In diesen Fällen würden Korrelationen von $r < .50$ nicht zwangsläufig gegen die Validität der neu konstruierten Skalen sprechen, sondern für Kriteriumskontamination bzw. –defizienz und mangelnde Skalensymmetrie. Die Skala *Vorsatz zu kooperieren* von Schalast (2000) wird trotz geringer Reliabilität zur Validierung der Grundskala *Kooperation im Therapieprozess* (KT) herangezogen, da sie aufgrund ihrer Konstruktion an und für Patienten des MRV gem. § 64 StGB hohe Inhaltsvalidität aufweist.

Des Weiteren soll der Bezug der Grundskalen zu Dimensionen eines anerkannten, empirisch fundierten Persönlichkeitsstrukturschemas exploriert werden. Hierzu werden Korrelationen mit den sechs Skalen des HEXACO-60 (Ashton & Lee, 2009) berechnet. Dieses Persönlichkeitsinventar mit 60 Items wurde zur ökonomischen Erfassung der sechs grundlegenden Persönlichkeitsfaktoren *Ehrlichkeit-Bescheidenheit* (engl. Honesty-Humility = H), *Emotionalität* (engl. Emotionality = E), *Extraversion* (engl. Extraversion = X), *Verträglichkeit vs. Ärger* (engl. Agreeableness vs. Anger = A), *Gewissenhaftigkeit* (engl. Conscientiousness = C) und *Offenheit für Erfahrungen* (Openness to Experience = O) entwickelt. In Tabelle 6 (s. nächste Seite) werden Annahmen über die Richtung des Zusammenhangs zwischen den Dimensionen des HEXACO-60 und den Dimensionen des PI-MRV-64 aufgeführt.

Tabelle 6

Annahmen über Korrelationsrichtungen (positiv vs. negativ) von Testwerten auf den Skalen des HEXACO-60 mit Testwerten auf den Grundskalen des PI-MRV-64

	VB	KT	LD	ID	SA	BT	MI	MB	SW	EB
H	+	+	+	-	-	-	-	-	+	+
E	-	+	+	-	-	-	-	-	-	+
X	+	+	-	+	-	+	+	+	+	-
A	+	+	+	-	-	-	-	-	+	+
C	+	+	+	-	-	+	-	+	+	-
O	+	+	-	-	+	-	-	-	+	+

Anmerkung. H: Ehrlichkeit-Bescheidenheit, E: Emotionalität, X: Extraversion, A: Verträglichkeit vs. Ärger, C: Gewissenhaftigkeit, O: Offenheit für Erfahrungen, VB: Vertrauen in das Behandlungsteam, KT: Kooperation im Therapieprozess, LD: Leidensdruck, ID: Identifikation mit delinquentem Lebensstil, SA: Substanzaffirmation, BT: Bagatellisierungstendenz, MI: Mangelnde Impulskontrolle, MB: Manipulative Beziehungsgestaltung, SW: Selbstwerterleben, EB: Ehrliche Beantwortung.

Die Ermittlung von Stärke und Richtung der Zusammenhänge erlaubt die Einordnung der zehn Dimensionen des zu konstruierenden Inventars in ein allgemeines Strukturmodell von Normalvarianten der menschlichen Persönlichkeit. Hierdurch wird das Verständnis der persönlichkeitspsychologischen Hintergründe der patientenspezifischen Ausprägungen auf den Grundskalen und Indexwerten des PI-MRV-64 gefördert. Dies ist neben theoretischer Relevanz auch von praktischer Bedeutung. So bietet das Wissen um jene Hintergründe Alternativerklärungen für bestimmte Ausprägungen auf den Grundskalen an, die über die unmittelbare Situation der Unterbringung im Maßregelvollzug hinausgehen. Beispielsweise würden sich niedrige Werte eines Patienten in *Kooperation im Therapieprozess* (KT) bei positiver Korrelation mit *Offenheit für Erfahrungen* (O) nicht nur über die spezifische Ablehnung der forensischen Unterbringung erklären lassen, sondern auch über einen generellen Mangel dieses Patienten, sich Neuem zu öffnen und ausgetretene Pfade zu verlassen. Über eine therapeutische Bearbeitung der mangelnden Offenheit kann wiederum die Kooperation im Therapieprozess positiv beeinflusst werden.

2.3.3.2 Diskriminante Konstruktvalidität

Zur Bestimmung der **diskriminanten Konstruktvalidität** (niedrige Korrelationen von Skalen mit Skalen anderer Gültigkeitsbereiche) wird jede Grundskala mit allen in Tabelle 5 gelisteten Validierungsskalen korreliert. Das Kriterium der diskriminanten Konstruktvalidität gilt als erfüllt, wenn die Korrelationen der Testwerte auf einer Grundskala mit Testwerten auf den Validierungsskalen der anderen Grundskalen niedriger ausfallen als die Korrelation der Testwerte jener Grundskala mit Testwerten auf der ihr zugeordneten Validierungsskala. Nach Bühner (2011, S. 80) sollten Koeffizienten von $r < .40$ nachgewiesen werden.

2.3.3.3 Faktorielle Validität

Faktorielle Validität liegt vor, wenn eine Skala aus homogenen, konstruktnahen Indikatoren (Items) besteht, deren Varianz und Interkorrelation durch einen gemeinsamen Faktor erklärt werden. Durch Skalenmodifikation mittels konfirmatorischer Faktorenanalysen (s. Abschn. 2.2.4) ist faktorielle Validität der finalen Skalenformen a priori gegeben.

2.3.4 Weitere Analysen

Sind validitätsbezogene Nachweise für das zu konstruierende Inventar erbracht, werden zwei weitere explorative Analysen durchgeführt.

Über drei Ein-Item-Maße beurteilen die Bezugstherapeuten der Probanden auf sechsstufigen endpunktbenannten Likert-Skalen (Außenkategorien 0 und 5) die gegenwärtige *Behandlungsbilanz.* Als Indikatoren der Behandlungsbilanz werden (a) das Erreichen bisheriger Teilziele der Therapie gemäß Behandlungsplan, (b) eine Prognose zur dauerhaften Abstinenz und (c) eine Prognose zur zukünftigen Straffreiheit erfragt (angelehnt an Schalast, 2000, S. 210).[18]

[18] Ausführliche validierte Fragebogenverfahren zur Evaluation des Therapieprozesses und -erfolgs (z. B. BFTB: Bonner Fragebogen für Therapie und Beratung; Fuchs, Sidiropoulou, Vennen & Fisseni, 2003; STEP: Stundenbogen für die Allgemeine und Differenzielle Einzelpsychotherapie; Krampen, 2002) eignen sich im vorliegenden Fall nicht zur Messung der

Über multiple lineare Regressionen[19] wird mit der Methode Einschluss die Vorhersagegüte der Grundskalen und Indexwerte für jedes der drei Ein-Item-Maße separat ermittelt. Die H_0 lautet jeweils: $b_{1;\ ...;\ k} = .00$: Wenn der Wert einer Grundskala oder eines Indexwertes (Prädiktor) um eine Einheit steigt, so ändert sich der Wert des betrachteten Indikators der Behandlungsbilanz (Kriterium) um 0 Einheiten, gegeben alle anderen Werte der Grundskalen/Indexwerte werden konstant gehalten. Die H_1 lautet: $b_{1;\ ...;\ k} \neq .00$: Wenn der Wert einer Grundskala oder eines Indexwertes (Prädiktor) um eine Einheit steigt, so verändert sich der Wert des betrachteten Indikators der Behandlungsbilanz um $b_{1;\ ...;\ k}$ Einheiten, gegeben alle anderen Werte der Grundskalen/Indexwerte werden konstant gehalten. Inhaltlich: Die Messwerte der Grundskalen/Indexwerte sagen die Messwerte der drei Ein-Item-Maße der Behandlungsbilanz vorher. Positive Vorzeichen der Beta-Gewichte zeigen eine Zunahme, negative Vorzeichen eine Abnahme der Werte der Indikatoren an. Für folgende Grundskalen und Indexwerte werden bei Zurückweisung der H_0 positive Vorzeichen der Beta-Gewichte erwartet:

- Vertrauen in das Behandlungsteam (VB)
- Kooperation im Therapieprozess (KT)
- Leidensdruck (LD)
- Index Therapiegrundlagen (Index TG)
- Index Sozial-funktionaler Persönlichkeitsstil (Index SP)
- Index Intrinsischer Leidensdruck (Index ILD)

zu erfassenden Indikatoren der Behandlungsbilanz, da sie einerseits meist die Patientenwahrnehmung adressieren und andererseits auf Quantifizierung von Symptomreduktion (z. B. Depressivität, interpersonelle Probleme) oder allgemeinen Wirkfaktoren (z. B. motivationale Klärung, therapeutische Beziehung) ausgerichtet sind.

[19] Voraussetzungen: Intervallskalierung der Prädiktoren und des Kriteriums (Indikatoren der Behandlungsbilanz als stetige Outcome-Variablen), Linearität des Zusammenhangs, Homoskedastizität und Normalverteilung der Residuen, keine Multikollinearität der Prädiktorvariablen (Prüfung mittels Toleranzwert oder Varianzinflationsfaktor).

Für folgende Grundskalen und Indexwerte werden bei Zurückweisung der H_0 negative Vorzeichen der Beta-Gewichte erwartet:

- Identifikation mit delinquentem Lebensstil (ID)
- Substanzaffirmation (SA)
- Bagatellisierungstendenz (BT)
- Mangelnde Impulskontrolle (MI)
- Manipulative Beziehungsgestaltung (MB)
- Index Devianz (Index DV)

Keine spezifische Vorhersage wird in Bezug auf *Selbstwerterleben* (SW) und *Ehrliche Beantwortung* (EB) getroffen. Es sind Effekte in beide Richtungen denkbar.

Stärke und Richtung des Zusammenhangs zwischen Grundskalen- bzw. Indexwerten und der *Unterbringungsdauer* der Probanden zum Zeitpunkt der Erhebung werden über separate Pearson-Korrelationen berechnet (bivariater Zusammenhang). Die H_0 lautet: Zwischen der Unterbringungsdauer und den Testwerten besteht kein Zusammenhang, $r = .00$. Die H_1 lautet: Änderungen in der Unterbringungsdauer gehen mit Änderungen in den Testwerten einher, $r \neq .00$. Generell wird angenommen, dass sich mit zunehmender Therapiedauer die Ausprägungen der Testwerte auf Grundskalen des Index Devianz vermindern (negative Korrelationskoeffizienten), ebenso Testwerte in *Leidensdruck* (LD) und *Mangelnde Impulskontrolle* (MI). Keine spezifische Vorhersage wird in Bezug auf die übrigen Skalen getroffen. Es ist sowohl denkbar, dass sich mit zunehmender Therapiedauer Werte in *Vertrauen in das Behandlungsteam* (VB), *Kooperativität im Therapieprozess* (KT) und *Selbstwerterleben* (SW) erhöhen als auch abschwächen. Ebenso ist möglich, dass längere therapeutische Intervention einen Rückgang manipulativer Tendenzen in der Beziehungsgestaltung und eine Zunahme von Ehrlichkeit zu bewirken vermag, als auch, dass sich Werte in *Manipulativer Beziehungsgestaltung* (MB) erhöhen und Werte in *Ehrliche Beantwortung* (EB) vermindern (z. B. weil sich Patienten durch manipulatives, unehrliches Agieren Vorteile im Lockerungsprozess erhoffen). Über Streudiagramme wird visuell exploriert, ob nichtlineare Zusammenhänge zwischen den Variablen vorliegen, die mit linearen Verfahren unentdeckt bleiben.

3. Erforderliche Stichprobenumfänge

Bei **Konfirmatorischen Faktorenanalysen** werden, als Spezialfälle von Strukturgleichungsmodellen, Stichprobenumfänge von mind. $200 \leq N \leq 250$ Probanden empfohlen (Bühner, 2011, S. 432). Eine Heuristik nennt fünf bis zehn Probanden pro Item einer Skala (15 Items: 75 bis 150 Probanden). Eine vom Verfasser durchgeführte Power-Analyse für den Fit-Index RMSEA erbrachte unter Verwendung der Parameter $H_{0:}$ RMSEA $\leq$ .05, $H_{1:}$ RMSEA $\leq$ .08, $df = 90$ (Messmodell mit 15 Items), $\alpha = .05$ und einer Power von $(1 - \beta) = .80$ einen erforderlichen Stichprobenumfang von $N = 141$ Probanden (Kalkulator von Preacher & Coffman, 2006). Dieser Wert wird als Minimum der Probandenanzahl zur Skalenkonstruktion angesetzt.

Zur Ermittlung der **prognostischen Validität** der Skalen sind pro Ausprägung des dichotomen Kriteriums BEA[20] mindestens $N = 25$ Fälle einzubeziehen (Backhaus, Erichson, Plinke, & Weiber, 2008, S. 288). Andere Autoren empfehlen deutlich höhere Fallzahlen zur präzisen Schätzung eines logistischen Regressionsmodells (Hosmer & Lemeshow, 2000, S. 346). In Anbetracht dessen, dass die Beantragung der Erledigung der Maßregel wegen Aussichtslosigkeit (§ 67d (5) StGB) keinen Normal- sondern einen Sonderfall darstellt, der sich im Zeitraum des geplanten Projektes bei der Konstruktionsstichprobe kaum in großer Anzahl ereignen wird, sei der Wert $N \geq 25$ als Zielwert der Probandenanzahl mit späterer BEA angesetzt.

Zur Prüfung von **Übereinstimmungs-, retrospektiver, konvergenter und diskriminanter Validität sowie Test-Retest-Reliabilität** der Skalen mittels Pearson-Korrelationen ist unter Verwendung der Parameter $H_{0:} r = .00$, $\alpha = .05$ und $(1 - \beta) = .80$ bei Erwartung eines hohen Effekts H_1: $r = .50$ die Erhebung von $N \geq 23$ Fällen pro Validierungskonstrukt (Fremdeinschätzungen, BZR-Eintragungen, PCL-R/HCR-20-Items, Konstrukte aus Validierungsskalen) erforderlich, bei Erwartung eines mittleren Effekts (bzw. hohen Effekts nach Connolly et al., 2016) H_1: $r = .30$ die Erhebung von $N \geq 67$ Fällen (Kalkulation mit G*Power; Faul, Erdfelder, Lang & Buchner, 2007). Der Wert $N = 23$ sei als Minimum angesetzt, der Wert $N = 67$ als Maximum. Zur Ermittlung der Test-Retest-Reliabilität wird eine Probandenanzahl von $N \geq 67$

[20] Beantragung der Erledigung der Maßregel wegen Aussichtslosigkeit.

angestrebt, da je nach Konstruktstabilität (*state* vs. *trait*) unterschiedlich hohe Effekte bei den Einzelkorrelationen zu erwarten sind.

Die multiplen Regressionsanalysen zur Bestimmung der Vorhersagegüte der Grundskalen und Indexwerte bezüglich der **Indikatoren der Behandlungsbilanz** (Therapiefortschritt, Prognose Abstinenz, Prognose Straffreiheit) erfordern im Fall der Grundskalen (zehn Prädiktoren) unter Verwendung der Parameter H_0: $f^2 = .00$, $\alpha = .05$ und $(1 - \beta) = .80$ bei Erwartung eines hohen Effekts H_1: $f^2 = .35$ (Cohen, 1988) die Erhebung von $N \geq 57$ Einschätzungen pro Indikator. Im Fall der Indexwerte (vier Prädiktoren) sind unter Verwendung analoger Parameter $N \geq 40$ Einschätzungen erforderlich, um einen derartigen Effekt zu finden (Kalkulation mit G*Power; ebd., 2007).

4. Probandenakquise

Erhoben werden Daten von männlichen Patienten des MRV gem. § 64 StGB in den Lockerungsstufen 0 bis C3 in Maßregelvollzugseinrichtungen des Freistaates Bayern. Patienten in höheren Lockerungsstufen (C4 bis D3) sind für die Studienteilnahme faktisch nicht mehr erreichbar, da sie aufgrund einer unbegleiteten Beschäftigung außerhalb der Maßregelvollzugseinrichtung nur in den Abend- und Nachtstunden auf ihren jeweiligen Stationen anzutreffen sind (Lockerungsstufe C4) oder sich in Serienbeurlaubungen zum Zwecke des Probewohnens befinden (Lockerungsstufen D1 – D3).

Zur Erprobung des Testentwurfs werden an mindestens zwei bayerischen Maßregelvollzugseinrichtungen Probanden akquiriert. Während an der Klinik für Forensische Psychiatrie am Bezirksklinikum Ansbach, an welcher der Verfasser selbst tätig ist, eine fortlaufende Erhebung umgesetzt werden kann, ist an anderen Standorten aufgrund räumlicher Distanz und zeitlich begrenzter Ressourcen nur eine block-/tageweise Datensammlung möglich. Aus diesem Grund soll die Validierung des Inventars vorrangig an der Klinik für Forensische Psychiatrie am Bezirksklinikum Ansbach erfolgen. Zur Erbringung validitätsbezogener Nachweise sind nicht derart umfangreiche Stichprobentestungen erforderlich wie für konfirmatorische

Faktorenanalysen (s. Abschn. 3). Der Einbezug von Validierungsdaten teilnehmender Patienten aus anderen Kliniken ist dennoch in hohem Maße wünschenswert und soll, wann immer logistisch möglich, umgesetzt werden.

Zu beachten ist, dass es sich bei der Zielklientel um eine Population von klinisch-forensischen Fällen handelt und demnach, bezogen auf die Grundgesamtheit aller in Deutschland lebenden Menschen, um eine hoch spezielle Randpopulation. Bei der Datenerhebung soll deshalb der Grundsatz Qualität vor Quantität gelten. Es ist zielführender, Daten unter persönlicher Anleitung und sorgfältiger Beachtung von Probandenmerkmalen (ausreichende Sprachkenntnisse, kognitive Fähigkeiten, Motivation und Ernsthaftigkeit) an einer Mindestfallzahl zu erheben, als den Testentwurf wahllos an Kliniken zu versenden, um den Stichprobenumfang zu maximieren. Durch eine geringere, aber valide Datenmenge sind stabilere Faktorstrukturen zu erwarten als durch zahlreiche qualitativ fragwürdige Testbearbeitungen.

Patienten teilnehmender Maßregelvollzugseinrichtungen werden in regelmäßig stattfindenden Stationsversammlungen über Art und Ablauf der modular angelegten Untersuchung informiert. Sie erhalten die Möglichkeit, an insgesamt drei Messzeitpunkten (Modulen) teilzunehmen. Modul 1 umfasst die Bearbeitung des Testentwurfs (ca. 30 – 45 min). Hierzu wird die größte Probandenanzahl benötigt (s. Abschn. 3); es ist das Mindestmaß der Studienteilnahme. Zusätzlich kann an Modul 2, der Bearbeitung des Validierungstests, mitgewirkt werden (ca. 20 – 30 min.). Modul 1 und 2 sollten in geringem zeitlichen Abstand, idealerweise am gleichen Tag, angeboten werden. Modul 3 umfasst abschließend die erneute Bearbeitung des Testentwurfs zur Bestimmung der Test-Retest-Reliabilität (ca. 30 – 45 min.; s. Abschn. 2.2.5). Es soll ca. vier Wochen nach Modul 1 angeboten werden. Modul 2 und 3 kann durchlaufen, wer an Modul 1 teilgenommen hat. Die Teilnahme an Modul 3 setzt jedoch nicht die Beteiligung an Modul 2 voraus. Abbildung 1 (s. nächste Seite) veranschaulicht den Studienablauf als Flussdiagramm.

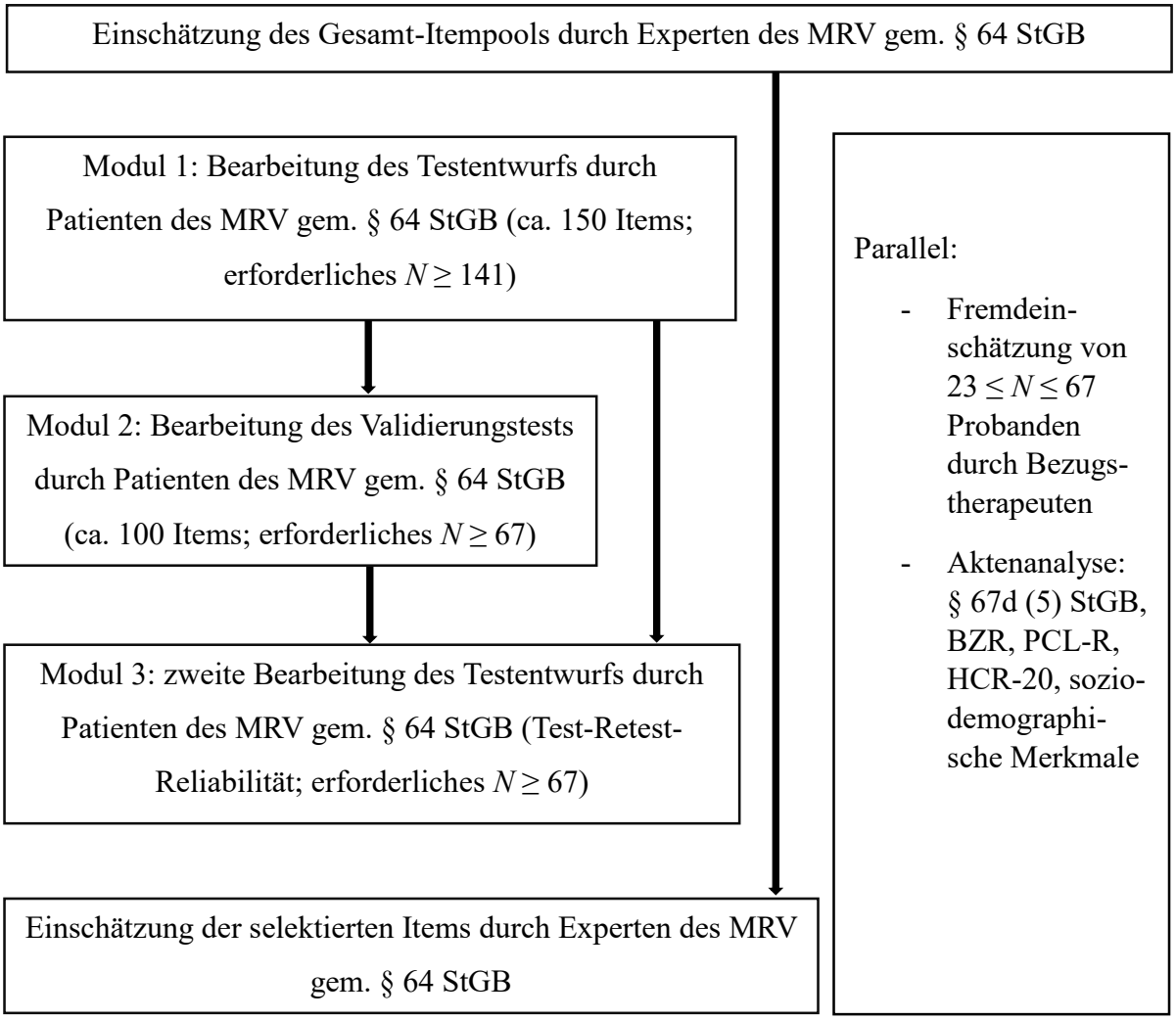

Abbildung 1. Flussdiagramm zum Studienablauf

Die teilnehmenden Patienten sollen für jedes Modul eine Vergütung erhalten. Die Vergütung sollte für Modul 1 am höchsten sein, für Modul 2 und 3 geringer. Für die Experten des MRV gem. § 64 StGB und die Bezugstherapeuten der Probanden ist keine Vergütung geplant.

5. Datenschutz und ethische Aspekte

Alle Probanden erklären sich schriftlich mit der Gewährung von Akteneinsicht bereit. Es werden das Alter in Jahren (nicht: Geburtsdatum), der Bildungsstand (höchster erreichter Schulabschluss), Vorstrafen (Eintragungen im Bundeszentralregister), Unterbringungsdauer bis zum Erhebungszeitpunkt in Monaten, Anlassdelikt, Substanzkonsummuster (Präferenzdrogen, Alter bei Erst- und Regelkonsum), Testwerte aus PCL-R und HCR-20[21] sowie psychiatrische Diagnosen (F10 – F19 u. a.) erhoben. Die Studie beschränkt sich auf männliche Probanden. Zur Rekrutierung stehen mehrere bayerische Maßregelvollzugseinrichtungen potenziell zur Verfügung (u. a. Klinik für Forensische Psychiatrie am Bezirksklinikum Ansbach). Da weder die Geburtsdaten, noch die jeweilige Klinik und Station erhoben bzw. numerisch codiert werden, ist es für Dritte nicht möglich, die Identität eines Studienteilnehmers zu bestimmen. Ferner erklären sich die Probanden mit der Fremdeinschätzung durch ihre Bezugstherapeuten bereit. Die Bezugstherapeuten erhalten keine Informationen über die Itembeantwortung der Probanden. Die Probanden erhalten keine Informationen über das Ergebnis der Fremdeinschätzung durch ihre Bezugstherapeuten.

Ein Pseudonymisierungsverfahren gewährleistet, dass die Weitergabe, Speicherung und Auswertung der probandenbezogenen Daten nach gesetzlichen Bestimmungen ohne Namensnennung erfolgt. Dabei wird jedem Probanden durch fortlaufende Nummerierung eine persönliche Codenummer auf Papier zugeordnet. Aus den Codenummern sind weder Rückschlusse auf Patientennamen noch auf Stationen und Kliniken möglich. Damit wird sichergestellt, dass die im Rahmen der Studie gemachten Angaben keinen Einfluss auf die weitere Behandlung haben können. Die Liste, auf der jedem Probandennamen eine Codenummer zugordnet ist, wird in einem abschließbaren Schrank im Sekretariat des Bezirksklinikum Ansbach, Klinik für Forensische Psychiatrie, aufbewahrt. Sie ist nur dem Verfasser und seinem Betreuer zugänglich und wird im Anschluss an die Erhebung vernichtet.

[21] Sofern im Erhebungszeitraum in den Patientenakten vorliegend.

Die soziodemographischen, kriminalbiographischen und testpsychologischen Probandendaten aus der Patientenakte sowie die Ergebnisse der Fremdeinschätzugnen gehen ausschließlich in aggregierter Form (Gruppenebene) in den schriftlichen Projektbericht ein. Die Angaben werden zu keinem Zeitpunkt innerhalb oder außerhalb der Klinik über den Klarnamen der Probanden mit den Itemantworten verbunden. Im gesamten Prozess der Untersuchung wird ausnahmslos mit den Patientencodes gearbeitet.

Die Studienteilnahme erfolgt vollkommen freiwillig. Die Probanden werden darüber aufgeklärt, dass sie ihre Einwilligung zur Teilnahme jederzeit ohne Angabe von Gründen und ohne negative Folgen für den Therapieverlauf wiederrufen können. Bereits erhobene Daten werden in diesem Fall gelöscht. Jeder Teilnehmer wird vor der Testung vollständig über das Ziel und den Ablauf des Projektes aufgeklärt. Es wird betont, dass die Studie keine individuelle Persönlichkeitsdiagnostik der Teilnehmer zum Ziel hat, sondern das Antwortverhalten der Gesamtstichprobe auf die Items im Fokus steht. Untersuchungsgegenstand sind die zu konstruierenden Skalen, nicht die Probanden. Nach der Testbearbeitung werden die Eigenschaften, die mit den Items gemessen werden sollen, gegenüber den Teilnehmern transparent gemacht (Debriefing).

Es ist davon auszugehen, dass bei den Probanden eine moderate psychische Belastung durch Items, die sich auf ihre Sucht- und Kriminalbiographie beziehen, auftritt. Des Weiteren ist eine ebenfalls moderate kognitive Belastung durch Verständnis- und Urteilsbildung beim Beantworten der Items wahrscheinlich. Die Probanden werden darauf hingewiesen, dass sie sich an ihre Bezugsmitarbeiter auf Station wenden können, wenn während oder nach der Testung Gesprächsbedarf auftritt. Die Stationsteams der Kliniken werden im Vorfeld vollständig über die Inhalte des Projektes aufgeklärt.

6. Open Science

Zur Gewährleistung von Transparenz und Nachvollziehbarkeit des Forschungsprozesses verpflichtet sich der Verfasser zur Online-Veröffentlichung der vorliegenden Projektbeschreibung vor Beginn der Datensammlung und -analyse auf einer Open Science-Plattform (Präregistrierung; z. B. www.psycharchives.org).

7. Literaturverzeichnis

Alpers, G. W. & Eisenbarth, H. (2008). *Psychopathic Personality Inventory-Revised (PPI-R). Deutsche Version*. Göttingen: Hogrefe.

Asendorpf, J. B. & Neyer, F. J. (2012). *Psychologie der Persönlichkeit* (5., vollst. überarb. Aufl.). Berlin u. a.: Springer.

Ashton, M. C. & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*(4), 340–345.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2008). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung* (12. Aufl.). Berlin u. a.: Springer.

Barriga, A. Q. & Gibbs, J. C. (1996). Measuring cognitive distortion in antisocial youth: Development and preliminary validation of the "How I Think" questionnaire. *Aggressive Behavior, 22,* 333–343.

Bonta, J. & Andrews, D. A. (2017). *The psychology of criminal conduct* (6th ed.). New York: Routledge.

Buchholz, A., Glöckner-Rist, A., Scherbaum, N., & Rist, F. (2014). Deutsche TCU Behandlungsmotivationsskalen (TCU-MS-d). *Zusammenstellung sozialwissenschaftlicher Items und Skalen.* doi:10.6102/zis214

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte u. erweiterte Aufl.). München: Pearson.

Carl, L. C., Breuer, M. M. & Endres, J. (2016). Leidensdruck und Behandlungsmotivation bei Gewaltstraftätern. *Forensische Psychiatrie und Psychotherapie, 23*(1), 8–36.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale: Erlbaum.

Connelly, B. S. & Ones, D. S. (2010). An other perspective on personality: meta-analytic integration of observers‘ accuracy and predictive validity. *Psychological Bulletin, 136*(6), 1092-1122.

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The Convergent Validity between Self and Observer Ratings of Personality: A meta-analytic review. *International Journal of Selection and Assessment, 15*(1), 110–117. doi:10.1111/j.1468-2389.2007.00371.x

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Dahle, K.-P. (1994). Therapiemotivation inhaftierter Straftater. In M. Steller, K.-P. Dahle & M. Basqué (Hrsg.), *Straftäterbehandlung. Argumente für eine Revitalisierung in Forschung und Praxis.* (Bd. 2, S. 227–246). Pfaffenweiler: Centaurus.

Dahle, K.-P. (1995). *Therapiemotivation hinter Gittern. Zielgruppenorientierte Entwicklung und Erprobung eines Motivationskonstrukts für die therapeutische Arbeit im Strafvollzug.* Regensburg: Roderer.

Diamantopoulos, A. & Winklhofer, H. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38,* 269–277.

Diehl, J. M. & Staufenbiel, T. (2007). *Statistik mit SPSS für Windows Version 15.* Eschborn bei Frankfurt a. M.: Dietmar Klotz.

Edwards, J. R. & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155–174.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. G*POWER 3.1.9.2 (Shareware). Verfügbar unter: http://gpower.hhu.de/

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). Los Angeles u. a.: Sage.

Fuchs, T., Sidiropoulou E., Vennen, D. & Fisseni, H. J. (2003). *Bonner Fragebogen für Therapie und Beratung (BFTB).* Göttingen: Hogrefe.

Gignac, G. E. & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102,* 74–78. doi:10.1016/j.paid.2016.06.069

Hampel, R. & Selg, H. (1975). *FAF – Fragebogen zur Erfassung von Aggressivitätsfaktoren. Handanweisung.* Göttingen: Hogrefe.

Hare, R. D. (2003). *Hare Psychopathy Checklist-Revised (PCL-R): 2nd Edition*. Toronto: Multi-Health Systems.

Hauser, R. M. (1973). Disaggregating a social-psychological model of educational attainment. In A. S. Goldberger & O. D. Duncan (Hrsg.), *Structural Equation Models in the Social Sciences* (S. 255 – 284). New York: Seminar Press.

Henning, H. & Six, B. (2014). Machiavellismus. *Zusammenstellung sozialwissenschaftlicher Items und Skalen.* doi:10.6102/zis126

Hewig, M. (2008). *Generalisierte und spezielle Vertrauensaspekte in der Psychotherapie. Eine empirische Studie zur prognostischen Bedeutung der Vertrauens-Trias für das Ergebnis stationärer Psychotherapie.* Unveröffentlichte Dissertation, Universität Trier.

Hewig, M., Hank, P. & Krampen, G. (2009). VTT-TAB – Kurzskalen zur Erfassung von Vertrauen in der Psychotherapie (Skalen zur Vertrauens-Trias in der therapeutischen Beziehung). *Klinische Diagnostik und Evaluation, 2*(3), 175–193.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley & Sons.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. doi:10.1080/10705519909540118

Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A. & Rammstedt, B. (2014). Soziale Erwünschtheit-Gamma (KSE-G). *Zusammenstellung sozialwissenschaftlicher Items und Skalen.* doi:10.6102/zis214

Klemm, T. (2002). *KV-S – Konfliktverhalten situativ. Verfahren zur Erfassung von Persönlich-keitsauffälligkeiten in Konfliktsituationen* (1. Aufl.). Göttingen: Hogrefe.

Kovaleva, A., Beierlein, C., Kemper, C. J. & Rammstedt, B. (2014). Die Skala Impulsives-Verhalten-8 (I-8). *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. doi:10.6102/zis183

Krampen, G. (2002). *Stundenbogen für die Allgemeine und Differenzielle Einzelpsychotherapie (STEP).* Göttingen: Hogrefe.

Kreft, I. G. G. & De Leeuw, J. (1998). *Introducing multilevel modeling.* Newbury Park: Sage.

Krosnick, J. A., & Presser, S. (2010). Questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of Survey Research* (2nd ed., pp. 263–313). West Yorkshire: Emerald Group.

Kuhl, J. & Kazén, M. (2009). *Persönlichkeits-Stil- und Störungs-Inventar (PSSI). Manual* (2., überarb. u. neu normierte Aufl.). Göttingen: Hogrefe.

Kunst, H. (2004). *Psychometrische Analysen zur Erfassung von Persönlichkeitsmerkmalen bei Straftätern – Übersetzung und Überprüfung des Antisocial Personality Questionaire*. Unveröffentlichte Dissertation, Technische Universität Dresden.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563–575. doi:10.1111/j.1744-6570.1975.tb01393.x.

Leygraf, N. (2006). Psychiatrischer Maßregelvollzug (§ 63 StGB). In H.-L. Kröber, D. Dölling, N. Leygraf & H. Sass (Hrsg.), *Handbuch der Forensischen Psychiatrie. Band 3: Psychiatrische Kriminalprognose und Kriminaltherapie* (S. 193–221). Darmstadt: Steinkopff.

Likert, R. (1932). A technique for measurement of attitudes. *Archives of Psychology, 140,* 1–55.

Ling, M. (2014). Lügen und Leugnen. *Zusammenstellung sozialwissenschaftlicher Items und Skalen.* doi:10.6102/zis74

Maas, C. J. M. & Hox J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology 1*(3), 86–92. doi: 10.1027/1614-1881.1.3.86

McDonald, R. P. (1970). The theoretical foundations of common factor analysis, principal factor analysis and alpha analysis. *British Journal of Mathematical and Statistical Psychology, 22,* 165–175.

McDonald, R. P. (1999). *Testtheory: A unified treatment.* Mahwah: Erlbaum.

Michel, L. & Conrad, W. (1982). Testtheoretische Grundlagen psychometrischer Tests. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B: Methodologie und Methoden, Serie II: Psychologische Diagnostik, Band 1: Grundlagen Psychologischer Diagnostik* (S. 1–129). Göttingen: Hogrefe.

Miller, W. R. & Rollnick, S. (2015). *Motivierende Gesprächsführung. Motivational Interviewing: 3. Auflage des Standardwerks in Deutsch.* Freiburg: Lambertus.

Mokros, A., Hollerbach, P., Nitschke, J. & Habermeyer, E. (2017). *Deutsche Version der Hare Psychopathy Checklist - Revised (PCL-R) von R. D. Hare.* Göttingen: Hogrefe.

Murphy, K. R. & Davidshofer, C. O. (2001). *Psychological testing: principles and applications* (5th ed.). Upper Saddle River: Prentice Hall.

Niemeyer, L. M. & Back, M. (2017, January 20). *Development and validation of an imprisonment-appropriate version of the Personality Inventory for DSM-5 (PID-5). Project description.* Abgerufen am 29. September 2019 von https://osf.io/eefvq/

Nübling, R., Kriz, D., Herwig, J., Wirtz, M., Fuchs, S., Hafen, K., Töns, N. & Bengel, J. (2005). *Normierung des Patientenfragebogens zur Erfassung der Reha-Motivation PAREMO. Abschlussbericht.* Unveröffentlichtes Manuskript, Albert-Ludwigs-Universität Freiburg. Abgerufen am 27. Oktober 2019 von https://www.gfqg.de/assessment/paremo.html?file=files/content/downloads/Assessment/PAREMO%20Projektabschlussbericht.pdf

Preacher, K. J., & Coffman, D. L. (2006). Computing power and minimum sample size for RMSEA [Computer software]. Verfügbar unter http://quantpsy.org/

Robins, R. W., Hendin, H. M. & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 27*(2), 151–161. doi:10.1177/0146167201272002

Rohrmann, S., Hodapp, V., Schnell, K., Tibubos, A., Schwenkmezger, P. & Spielberger, C. D. (2013). *STAXI-2. Das State-Trait-Ärgerausdrucks-Inventar – 2. Deutschsprachige Adaptation des State-Trait Anger Expression Inventory – 2 (STAXI-2) von Charles D. Spielberger.* Götttingen: Hogrefe.

Rossiter, J. R. (2008). Content validity of measures of abstract constructs in management and organizational research. *British Journal of Management, 19,* 380–388. doi:10.1287/isre.2.3.192

Schalast, N. (2000). Therapiemotivation im Maßregelvollzug gem. § 64 StGB: Patientenmerkmale, Rahmenbedingungen, Behandlungsverläufe. In F. Schaffstein, H. Schöch & H. Schüler-Springorum (Hrsg.), *Neue Kriminologische Studien* (Bd. 21). München: Wilhelm Fink.

Schalast, N. (2014). Behandlung substanzabhängiger Straftäter. In T. Bliesener, F. Lösel & G. Köhnken (Hrsg.), *Lehrbuch Rechtspsychologie* (S. 489–511). Bern: Hans Huber.

Schaumburg, C. (2010). *Basiswissen: Maßregelvollzug* (1. Aufl. der Neuausgabe). Bonn: Psychiatrie-Verlag.

Schermelleh-Engel, K. & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktualisierte u. überarbeitete Aufl., S. 119–141). Berlin u. a.: Springer.

Schmidt-Quernheim, F. (2008). Behandlung im Maßregelvollzug. In F. Schmidt-Quernheim & T. Hax-Schoppenhorst (Hrsg.), *Professionelle forensische Psychiatrie. Behandlung und Rehabilitation im Maßregelvollzug* (2., vollständig überarbeitete und erweiterte Aufl., S. 91–198). Bern: Hans Huber.

Schwarz, M. (2018). Konstruktion eines Persönlichkeitsinventars für Patienten des Maßregelvollzugs gem. § 64 StGB (Masterarbeit). In Gesellschaft für Kriminologie, Polizei und Recht e.V. (Hrsg.), *Schriftenreihe der GKPR e.V.* (Bd. 9). Frankfurt a. M.: Verlag für Polizeiwissenschaften.

Seitz, W. & Rautenberg, M. (2010). *Persönlichkeitsfragebogen für Inhaftierte (PFI+). Manual* (1. Aufl.). Göttingen: Hogrefe.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Snijders, T. A. B. & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London u. a.: Sage.

Sykes, G. M. & Matza, D. (1957). Techniques of neutralization: A theory of delinquency. American *Sociological Review, 22*(6), 664–670.

Sykes, G. M. & Matza, D. (1968). Techniken der Neutralisierung. Eine Theorie der Delinquenz. In F. Sack & R. König (Hrsg.), *Kriminalsoziologie* (S. 360–371). Frankfurt am Main: Akademische Verlagsgesellschaft.

Von Collani, G. & Herzberg, P. Y. (2003). Zur internen Struktur des globalen Selbstwertgefühls nach Rosenberg. *Zeitschrift für Differentielle und Diagnostische Psycholo*gie, *24*(1), 9–22. doi:10.1024//0170-1789.24.1.3

Weafer, J., Baggott, M. J. & de Wit, Harriet. (2013). Test-retest reliability of behavioral measures of impulsive choice, impulsive action, and inattention. *Exp Clin Psychopharmacol., 21*(6), 475–481. doi:10.1037/a0033659

Webster, C. D., Douglas, K. S., Eaves, D. & Hart, S. D. (1998). *Die Vorhersage von Gewalttaten mit dem HCR-20* (R. Müller-Isberner, D. Jöckel & S. G. Cabeza, Hrsg. und Übers.). Haina: Institut für Forensische Psychiatrie (Originalarbeit erschienen 1997).

Welpe, I. (2014). Messung, formative vs. reflektive. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie* (18. Aufl., S. 1021). Bern: Hogrefe.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen: Hogrefe.