Schluss- und Evaluationsbericht zum Modellversuch „Neue psychotherapeutische Interventionsprogramme und Evaluationskonzepte im Schweizer Strafvollzug“

Ergänzter Evaluationsbericht

Finale Fassung

Universität Konstanz, Arbeitsgruppe Forensische Psychologie
Universität Bern, Forschungsgruppe, Forensisch-Psychiatrischer Dienst

20. September 2018

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

ASAT®	Anti-Sexuelle-Aggressivität-Training®
ASAT®Suisse	Anti-Sexuelle-Aggressivität-Training®Suisse
aStGB	Schweizerisches Strafgesetzbuch in einer aktuell nicht mehr gültigen Fassung
AuG	Bundesgesetz über die Ausländerinnen und Ausländer (Ausländergesetz)
BetMG	Bundesgesetz über die Betäubungsmittel und die psychotropen Stoffe (Betäubungsmittelgesetz)
BJ	Bundesamt für Justiz
FPD	Forensisch-Psychiatrischer Dienst der Universität Bern
GST	Gruppe der Gewaltstraftäter
GST R&R	Experimentalgruppe innerhalb der Gruppe der Gewaltstraftäter, welche das Reasoning and Rehabilitation Programm absolviert
HAB	Hostile Attribution Bias
ICD-10	International Classification of Diseases, 10th revision
IIP-D	Inventar zur Erfassung interpersonaler Probleme – Deutsche Version
JVA	Justizvollzugsanstalt
K-FAF	Kurzfragebogen zu Aggressivitätsfaktoren
KG	Kontrollgruppe
M	Mittelwert
MV	Modellversuch
N	Anzahl
PCL-R	Psychopathy Checklist revised
R&R/ R&R2	Reasoning and Rehabilitation Programm/ revised
RNR	Risk-Need-Responsivity Prinzip
SA	Standardabweichung
SST	Gruppe der Sexualstraftäter
SST ASAT	Experimentalgruppe innerhalb der Gruppe der Sexualstraftäter, welche das Anti-Sexuelle-Aggressivität-Training®Suisse absolviert
StGB	Schweizerisches Strafgesetzbuch
TAR	Time at Risk
TAU	Treatment as Usual (Vergleichsgruppe, welche die jeweilige Standardbehandlung erhält)
T1	Erster Messzeitpunkt (Prä-Messung)
T2	Zweiter Messzeitpunkt (Post-Messung)
VRAG	Violence Risk Appraisal Guide
VÜ	Fragebogen zur Verantwortungsübernahme
WG	Bundesgesetz über Waffen, Waffenzubehör und Munition (Waffengesetz)

Der Modellversuch „Neue psychotherapeutische Interventionsprogramme und Evaluationskonzepte im Schweizer Strafvollzug“

Ziele

Die zentrale Zielsetzung des Modellversuchs (MV) bestand in der Implementierung und Wirksamkeitsprüfung als innovativ beurteilter Interventionsprogramme im Schweizer Strafvollzug, namentlich die Programme „Reasoning and Rehabilitation“ (R&R; Ross, Fabiano, & Ross, 1986; R&R2; Ross, Hilborn, & Liddle, 2007)[1] sowie das „Anti-Sexuelle-Aggressivität-Training®Suisse“ (ASAT®Suisse; Falk & Steffes-enn, 2014; siehe auch Steffes-enn, 2005; Steffes-enn, 2008).

Ein weiteres Ziel des MV bestand in der Implementierung geeigneter Strategien zur Evaluation von Behandlungsprogrammen für Straftäter. Zur wissenschaftlichen Begleitung, Evaluation und Berichterstattung wurde auf unabhängige externe Expertise zurückgegriffen.

Fragestellungen[2]

Die primäre Fragestellung des MV war, inwieweit eine störungs- und deliktorientierte Behandlung von Gewalt- und Sexualstraftätern zu messbaren Veränderungen von Einstellungen, Werthaltungen, Verhalten und persönlichkeitsbezogenen Variablen sowie des Rückfallrisikos führt. Im Einzelnen wurden folgende Fragestellungen formuliert:

(1) Sind die forensischen Therapien allgemein wirksam? Zeigen Straftäter, welche eine forensische Therapie (d.h. Einzel- und/oder Gruppentherapie) erhielten, im Vergleich zu Personen, die nicht behandelt wurden (Kontrollgruppe), Verbesserungen im Bereich von kriminogenen Einstellungen, Persönlichkeitseigenschaften und Verhaltensmassen?
(2) Sind die neuen Therapien wirksam? Lassen Straftäter, welche mittels R&R(2) oder ASAT®Suisse behandelt wurden, im Vergleich zur Kontrollgruppe positive Veränderungen in den genannten Merkmalen erkennen?

[1] Anmerkung der Evaluatoren: Zunächst wurde das klassische R&R angewandt (35 Sitzungen à 120 Minuten), im Verlauf des MV fand jedoch eine vollständige Umstellung auf das neuere R&R2 statt, das eine Kurzform des ursprünglichen Reasoning and Rehabilitation-Programms darstellt (14 Sitzungen à 90 Minuten). Auf Grundlage der zur Verfügung gestellten Daten ist eine Unterscheidung der Probanden nach Teilnehme an einer der beiden Varianten nicht möglich, sodass diese im vorliegenden Bericht zu einer einzigen Gruppe zusammengefasst werden.

[2] vgl. FPD Bern (2009)

(3) Sind die neuen Therapien den Einzeltherapien überlegen? Weisen Straftäter, die eine Gruppen- und Einzelbehandlung erhielten, aufgrund der Kombination mit den neuen Therapieprogrammen zum Zeitpunkt der Nachhermessung bessere Ergebnisse auf als Täter, welche ausschließlich Einzeltherapie erhielten?
(4) Welche differentiellen Effekte der forensischen Therapien auf Klienten-Gruppen sind zu beobachten? Zeigen die genannten Behandlungen unterschiedliche Effekte in Abhängigkeit von Merkmalen der Klientengruppen und dem Behandlungssetting (vgl. „What works for whom under which circumstances?“; Lösel & Schmucker, 2005)?

Ablauf des Modellversuchs

Abbildung 1 zeigt den Ablauf des Modellversuchs in der Übersicht (Nomenklatur und Fallzahlen gemäß vorliegendem Evaluationsbericht). Zur besseren Übersichtlichkeit finden darin ausschließlich diejenigen Aspekte im Untersuchungsablauf Erwähnung, die Eingang in den vorliegenden Evaluationsbericht fanden. Die im Rahmen des Modellversuchs durchgeführten Untersuchungen waren noch weitaus umfangreicher. Im folgenden Text wird der Ablauf ausführlich erläutert.

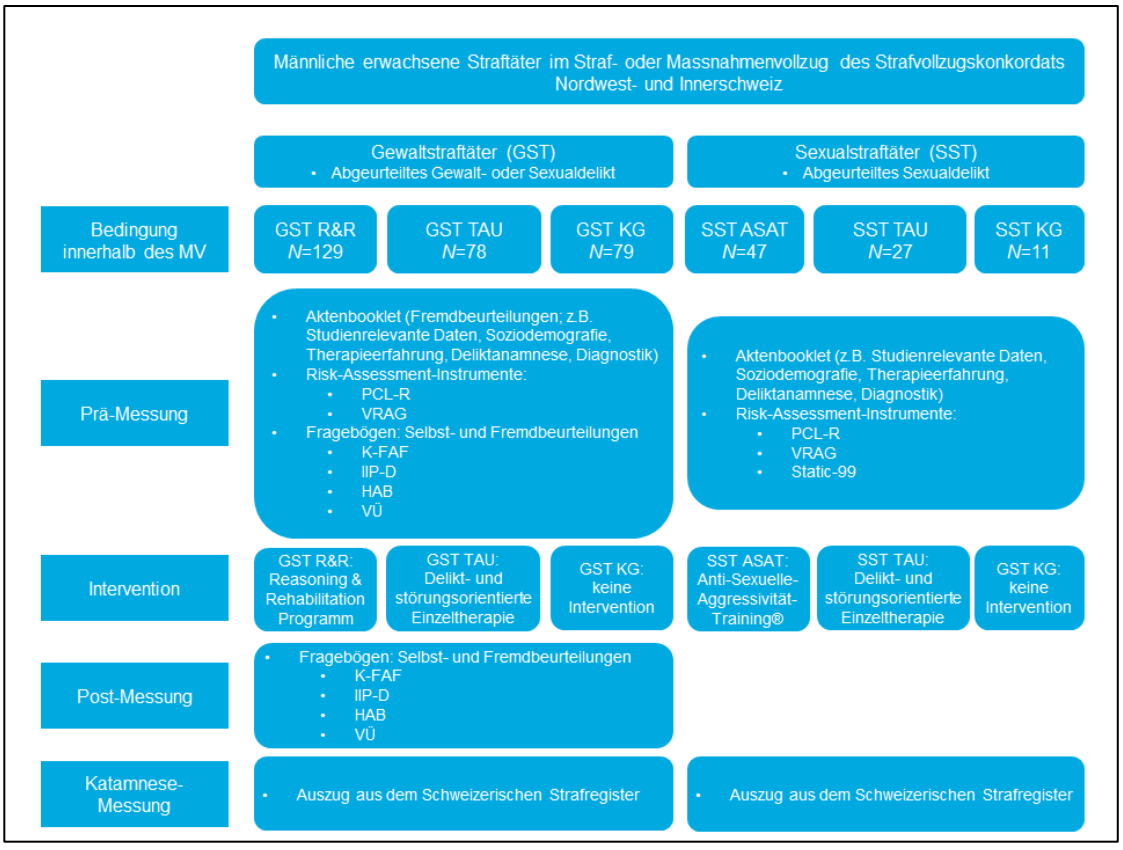

Abbildung 1. Schematischer Ablauf des Modellversuchs

Durchführungs- und Planungsphase des Modellversuchs

Die zentralen Schritte im Laufe des Planungs- sowie Durchführungsphase sind in Abbildung 2 aufgeführt.

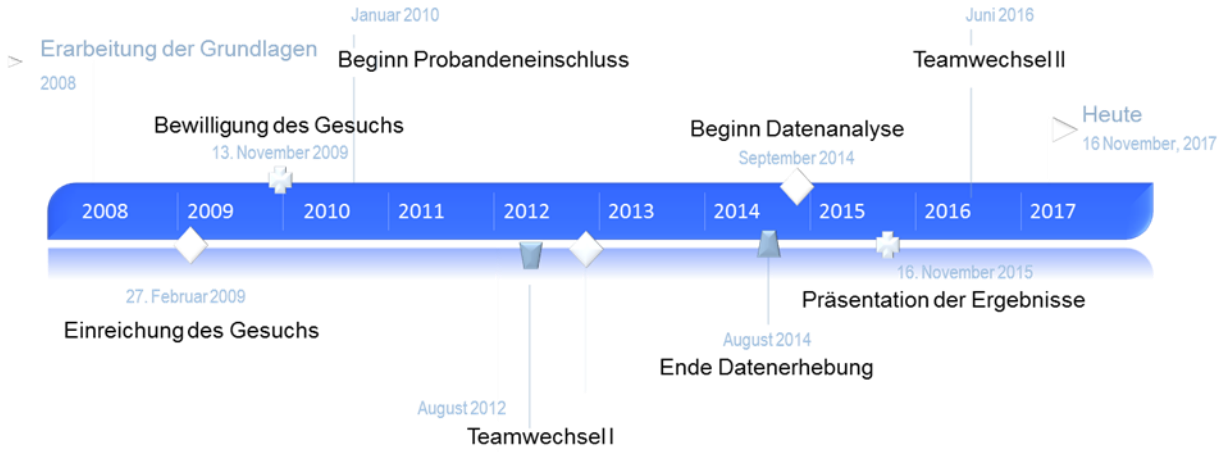

Abbildung 2. Projektschritte

Die Planung startete 2008, der Projektstart im Januar 2010 erfolgte durch Prof. Dr. Anneliese Ermer (Leitung), Prof. Dr. Dr. Martin Peper (Evaluationsverantwortlicher) und durch die wissenschaftlichen Mitarbeiterinnen lic. phil. Jenny Steinbach, Sandra Burkhard, MSc . und lic. phil. Nadia Loretan. Per August 2012 erfolgte eine Restrukturierung des Teams. Die Leitung oblag ab dann Dr. Dorothee Klecha, Prof. Dr. Dr. Martin Peper agierte weiterhin als Evaluationsverantwortlicher. Die neuen wissenschaftlichen Mitarbeitenden waren Dr. Sandy Krammer und Roger Gubelmann, MSc. Im Juni 2016 wurde die Projektleitung von Prof. Dr. Michael Liebrenz übernommen, die Evaluation oblag Prof. Dr. Jerôme Endrass von der Universität Konstanz. Abbildung 2 gibt einen Überblick über die Teammitglieder.

Die Dauer des Modellversuchs war durch das erste Projektteam auf 60 Monate angelegt mit Zwischenberichten in 2010, 2011 und 2012. Das Projektende war vorgesehen für 09/2014 mit Schlussbericht in 12/2014.

Abbildung 3. Teammitglieder

Ein- und Ausschlusskriterien[3]

Einschlusskriterien allgemein

Bei der untersuchten Klientel handelte es sich um männliche Straftäter, die sich zum Zeitpunkt der Studie entweder in einer Justizinstitution, in einem Maßnahmezentrum (geschlossen, halb-offen) oder in Freiheit (Ambulatorium oder Bewährungshilfe) befanden. Grundsätzlich wurden Personen in den MV aufgenommen, die als Indexdelikt gemäß Aktenlage eine Gewaltstraftat (im vorliegenden Bericht GST-Gruppe genannt) oder eine Sexualstraftat (im vorliegenden Bericht SST-Gruppe genannt) begangen hatten[4]. Es wurden nur solche Personen in die Studie einbezogen, welche hierzu ihre schriftliche Einwilligung gegeben hatten. Alle Probanden mussten zum Zeitpunkt ihres schriftlichen Einverständnisses zur Studienteilnahme 18 Jahre oder älter sein. Es wurde im Nachhinein kontrolliert, dass jeder Proband nur einfach in eine der Teilstichproben des MV einging.

Einschlusskriterien der Experimentalgruppen (GST R&R/ SST ASAT)
Bei den Probanden der Therapiegruppen R&R(2) und ASAT®Suisse handelte es sich um Gewalt- und/oder Sexualstraftäter mit gerichtlich oder vollzugsseitig angeordneten Therapien. Dabei musste es sich entweder um das R&R(2)- (GST-Gruppe) oder das ASAT®Suisse-Programm (SST-Gruppe) handeln. Therapieteilnehmer wurden durch die Therapieabteilungen gemeldet bzw. durch die Forschungsmitarbeitenden in Erfahrung gebracht und daraufhin den entsprechenden Untersuchungsgruppen zugewiesen.

Einschlusskriterien der Vergleichsgruppen (GST TAU/ SST TAU)
Bei den Probanden der Vergleichsgruppen handelte es sich um Gewalt- und/oder Sexualstraftäter, die ebenfalls gerichtlich oder vollzugsseitig eine Therapie angeordnet bekommen hatten. Sie mussten sich aktuell in Einzeltherapie befinden und weder am R&R(2)- noch am ASAT®Suisse -Programm teilnehmen. Analog zu den Experimentalgruppen wurden Probanden durch die Therapieabteilungen gemeldet bzw. die Forschungsmitarbeitenden erkundigten sich über Inhaftierte, die möglicherweise den Einschlusskriterien entsprachen, und den entsprechenden Untersuchungsgruppen zugewiesen.

Einschlusskriterien der Kontrollgruppen (GST KG/ SST KG)
Die Kontrollgruppe umfasste Gewalt- und/oder Sexualstraftäter, die sich während der Zeit des Vollzuges und zwischen den Untersuchungsterminen nicht in psychotherapeutischer Behandlung befanden, d.h. die zum Zeitpunkt der Studie keine Einzel- oder Gruppentherapie erhielten.

---

[3] Die folgenden Informationen zu den Ein- und Ausschlusskriterien des MV werden so genau dargestellt, wie dies auf Grundlage der zur Verfügung stehenden Informationen möglich ist.

[4] Anmerkung der Evaluatoren: Einige Personen waren dennoch in den MV aufgenommen worden, ohne dass sie ein Gewalt- oder Sexualdelikt begangen hatten. Dies war in allen Subgruppen der Fall. Diese Probanden wurden von der vorliegenden Analyse ausgeschlossen (siehe Abschnitt „Von der Auswertung ausgeschlossene Probanden“).

Ausschlusskriterien

Für sämtliche Gruppen und Bedingungen galten grundsätzlich folgende Ausschlusskriterien:
– Ungenügende Deutschkenntnisse
– Intelligenzminderung
– Akute Alkohol- oder Drogen-Intoxikation zum Zeitpunkt der Datenerhebung.

Überprüft wurden die genannten Kriterien nach Augenschein durch das Vollzugspersonal und/ oder durch die Forschungsmitarbeitenden des FPD Bern, mit Ausnahme der Intelligenzminderung, zu deren Überprüfung zudem die Akten des Straftäters hinzugezogen wurden bzw. auf das Betreuungs- oder Vollzugspersonal abgestellt wurde.

Methodisches Design[5]

Die Auswirkungen einer Teilnahme am R&R(2)- bzw. am ASAT®Suisse -Programm wurden in einem quasi-experimentellen Design durch den Vergleich von Probanden in jeweils drei Bedingungen erfasst:
Experimentalgruppen (Bezeichnung im vorliegenden Bericht: GST R&R und SST ASAT): Diese Probanden nahmen entweder am R&R(2)- (GST R&R) oder am ASAT®Suisse- (SST ASAT) Therapieprogramm teil. Neben den spezifischen Therapieprogrammen erhielten auch die Probanden der Experimentalgruppe eine in der jweiligen Institution übliche Einzeltherapie.
Vergleichsgruppen (Bezeichnung im vorliegenden Bericht: GST TAU und SST TAU): Diese Probanden erhielten das in der jeweiligen Anstalt übliche Behandlungsprogramm für Gewalt- und/oder Sexualstraftäter (GST TAU) bzw. Sexualstraftäter (SST TAU) in Form von Einzeltherapie.
Kontrollgruppen (Bezeichnung im vorliegenden Bericht: GST KG und SST KG): Diese Probanden erhielten über den Zeitraum der Untersuchung keine psychotherapeutische Behandlung.
Der Versuchsplan umfasste eine longitudinale Datenerhebung zu verschiedenen Messzeitpunkten: Eine Prä-Messung vor Beginn der Intervention, eine Post-Messung nach Beendigung der Intervention (jeweils Fragebogen) sowie eine Katamnese-Messung (Erhebung der Rückfälligkeit). Die Analyse wurde über Differenzwerte vollzogen.

[5] Anhand der zur Verfügung stehenden Daten konnten in der vorliegenden Evaluation nicht sämtliche vorgesehene Auswertungsaspekte beantwortet werden (z.B. Faktor „Behandlungssetting“), sodass im folgenden Abschnitt nur diejenigen Faktoren erläutert werden, welche im Rahmen der Evaluation ausgewertet werden konnten.

Instrumente

Im Folgenden werden die im Rahmen des MV eingesetzten Instrumente genannt, welche auch zum Zwecke der vorliegenden Evaluation Berücksichtigung fanden. Eine ausführlichere Beschreibung dieser Instrumente befindet sich in Anhang 3. Im MV wurden darüber hinaus noch zahlreiche weitere Instrumente eingesetzt (siehe FPD Bern, 2009). Diejenigen Instrumente, die nicht für die vorliegende Evaluation analysiert wurden, finden zum Zweck einer besseren Übersichtlichkeit keine Erwähnung im vorliegenden Bericht.

Aktenbooklet
Informationen, welche den zur Verfügung stehenden Akten über den Probanden entnommen wurden, wurden mithilfe eines Kodierbogens (Aktenbooklet) dokumentiert. Damit wurden allgemeine Daten zu Merkmalen, welche die Studie betreffen (z.B. Gruppenzugehörigkeit), persönliche sowie soziodemografische Angaben, Angaben zum familiären Hintergrund, zur Therapieerfahrung, zur Deliktanamnese, zur forensisch-psychiatrischen Diagnostik sowie Messwerte in Risk-Assessment-Instrumenten erhoben. In die vorliegende Evaluation finden nur ausgewählte Informationen aus dem Aktenbooklet Eingang.
Die Interrater-Reliabilität des selbst entwickelten Instruments wurde nicht empirisch untersucht. Anhand von regelmäßig zweifach ausgeführten Fall-Analysen wurde die Übereinstimmung der Rater anhand selektierter Fälle nach Augenschein überprüft.

Risk-Assessment-Instrumente
– Psychopathy Checklist revised (PCL-R; Hare, 2003)
– Violence Risk Appraisal Guide (VRAG; Quinsey, Harris, Rice, & Cormier, 2006)
– Static-99 (Hanson & Thornton, 1999)

Fragebögen
– Kurzfragebogen zur Erfassung von Aggressivitätsfaktoren (K-FAF; Heubrock & Petermann, 2008)
– Inventar zur Erfassung interpersonaler Probleme (IIP-D; Horowitz, Strauß, & Kordy, 2000)
– Hostile Attribution Bias (HAB; Tremblay & Belchevski, 2004)
– Fragebogen zur Verantwortungsübernahme (VÜ; Gabriel, Oswald, & Bütikofer, 2005; Oswald & Bütikofer, 2002)

Vorgehen bei der Datenerhebung

Das im Folgenden erläuterte allgemeine Vorgehen bei der Datenerhebung im Rahmen des MV ist weitestgehend analog für die Gruppen GST und SST, sodass es unabhängig von der Gruppenzugehörigkeit gilt, sofern nicht anders erwähnt.

Stichprobenselektion und Rekrutierung

Experimentalgruppen (GST R&R sowie SST ASAT). Die Selektion der Probanden der Gruppentherapien erfolgte durch die jeweiligen Therapeuten, die Studienleitung hatte hierauf keinen Einfluss. In der Regel lief in den Experimentalgruppen die Studienanfrage wie folgt ab: Eine Woche vor der ersten Sitzung der Gruppentherapie wurden die Teilnehmer zu einem ersten Termin eingeladen, meist zum Zeitpunkt der Gruppentherapie. In den meisten Fällen erschienen zu diesem Termin bereits alle späteren Gruppenteilnehmer. Aus Sicherheitsgründen waren stets zwei Versuchsleiter zugegen. In wenigen Fällen wurden zusätzliche Termine mit denjenigen Personen vereinbart, die den ersten Termin nicht wahrnehmen konnten. Die potenziellen Probanden wurden mündlich und schriftlich über die Studie informiert. Fragen wurden unmittelbar beantwortet. Insbesondere denjenigen Probanden, welche beim FPD in therapeutischer Behandlung waren, wurde versichert, dass die Forschungsbefragung unabhängig von der Therapie durchgeführt werde und dass die Therapeuten keinen Einblick in die erhobenen Daten erhalten. Sofern alle Fragen geklärt werden konnten, der potentielle Proband umfassend über die Studie aufgeklärt worden war und sich freiwillig für eine Teilnahme entschieden hatte, wurde ihm die schriftliche Einverständniserklärung zur Unterschrift vorgelegt, welche danach die Studienleitung gegenzeichnete. Anschließend erfolgte die Datenerhebung im Gruppen-Setting.

Vergleichs- und Kontrollgruppen (GST TAU/KG sowie SST TAU/KG). Der erste Schritt der Rekrutierung erfolgte je nach Justizinstitution leicht unterschiedlich: In einigen Institutionen erfolgte eine Vorselektion geeigneter Personen durch die dortigen Mitarbeiter, während in anderen Institutionen diese Selektion durch Mitarbeitende des FPD-Forschungsteams vorgenommen wurde. In letzterem Fall nahmen diese vor Ort Einsicht in die Stammblätter aller an diesem Stichtag in der Institution befindlichen Straftäter. Die Vorselektion basierte auf dem Indexdelikt, mutmaßlicher deutscher Sprachkenntnisse sowie ausreichender Aufenthaltsdauer in der Strafanstalt, sodass eine Teilnahme über die gesamte Dauer des MV gesichert erschien.
Auf diese Vorselektion folgte ein persönliches Gespräch, bei dem ein oder zwei Versuchsleiter anwesend waren. In wenigen Fällen fand dieses Informationsgespräch in einem Gruppen-Setting statt. In diesem Fall stellten zwei Versuchsleiter die Studie denjenigen Insassen vor, die zuvor von den Mitarbeitenden der Anstalt vorselektiert worden waren. In beiden Fällen gab es eine mündliche und schriftliche Information über die Studie. Bezüglich der speziellen Untersuchungsziele und -prozeduren wurden dabei grundlegende Informationen allgemeinverständlich vermittelt, um so die Probanden nicht in Bezug auf die Therapieevaluationsziele zu sensibilisieren oder zu beeinflussen. Es kam jedoch zu keiner Täuschung der Teilnehmer bezüglich der Studieninhalte. Diejenigen Personen, welche zur freiwilligen Teilnahme bereit waren, wurden informiert, dass sie zum nächstmöglichen Zeitpunkt für die Befragung aufgeboten würden. Während der Befragung wurde der potentielle Proband auf zwischenzeitlich aufgekommene Fragen angesprochen und ob er den Inhalt der Studie verstanden habe. Zudem wurde der Person eine schriftliche Studien-Information vorgelegt, welche sie durchlesen und mitnehmen konnte. In dieser befanden sich auch Kontaktangaben, für den Fall, dass zu einem späteren Zeitpunkt Fragen aufkämen.

Sofern alle Fragen geklärt werden konnten, der potentielle Proband umfassend über die Studie aufgeklärt war und sich freiwillig zur Teilnahme entschied, wurde ihm die schriftliche Einverständniserklärung zur Unterschrift vorgelegt, welche anschließend durch die Studienleitung gegengezeichnet wurde. Danach erfolgte die Datenerhebung.

Aufwandsentschädigung. Bei einem Großteil der Probanden der beiden Therapiegruppen fanden die Interventionen im Rahmen der verordneten Maßnahmenbehandlung statt, für welche sie keine Vergünstigungen erhielten. Aufgrund der Schwierigkeiten, Kontrollprobanden zu rekrutieren, wurde diesen in Absprache mit dem BJ und der Anstaltsdirektion im Laufe des MV eine Entschädigung in Form einer Telefonkarte in Aussicht gestellt. Dieses Vorgehen bezog sich ausschließlich auf Probanden aus der JVA Witzwil.

Den Probanden des Forensik-Ambulatoriums und der Bewährungshilfe, die zum Befragungsort (entweder Forensik-Ambulatorium oder Büros der Bewährungshilfe inkl. Regionalstellen) einen Weg zurücklegen mussten, der mit Fahrtkosten verbunden war, wurden die Fahrtspesen erstattet.

Messzeitpunkte und Dauer der Erhebung[6]

Die Datenerhebungsphase des MV dauerte für alle Datenquellen vom 04.02.2010 bis zum 05.05.2015. Die Prä-Messung T1 (1. Selbstbeurteilung) fand in der GST-Gruppe zwischen dem 04.02.2010 und dem 25.03.2014, in der SST-Gruppe zwischen dem 25.03.2010 und dem 18.06.2014 statt. Diese lag in jedem Fall vor Beginn der Intervention. Die Post-Messung T2 (2. Selbstbeurteilung) fand in der GST-Gruppe zwischen dem 31.05.2010 und dem 05.08.2014, in der SST-Gruppe zwischen dem 02.02.2011 und dem 23.07.2014 statt. Diese lag in jedem Fall nach der Beendigung der Intervention, wobei nicht ausgeschlossen werden kann, dass einzelne Probanden die jeweilige Intervention vorzeitig bzw. nicht erfolgreich beendet haben. Die Bearbeitung der Fragebögen fand in der Unterbringungs-Institution statt, je Institution entweder im Einzel- oder im Gruppensetting. Die Anleitung erfolgte durch Mitarbeitende des FPD-Forschungsteams, wobei im Einzelsetting stets ein Forschungsmitarbeitender anwesend war und im Gruppensetting stets zwei. Die Fremdbeurteilungen wurden bis zum 16.09.2014 (GST) bzw. 12.08.2014 (SST) eingeholt. Bearbeitet wurden die Fremdbeurteilungen durch diejenige Person, die dem jeweiligen Probanden am nächsten stand. Dabei konnte es sich um Psychotherapeut/-innen, aber auch um Vollzugsmitarbeitende handeln. Die Fragebögen wurden so entworfen, dass sie weitgehend selbsterklärend waren. Traten dennoch Rückfragen auf, konnte das FPD-Forschungsteam telefonisch kontaktiert werden. Im Mittel betrug die Differenz zwischen den beiden Messzeitpunkten in der GST-Gruppe 140 Tage (*SA*=60) und in der SST-Gruppe 393 Tage (*SA*=167).
Die offizielle Rückfälligkeit wurde anhand der Einträge in das Schweizerische Strafregister erfasst. Hierbei wurde die einschlägige Rückfälligkeit im Gewalt- oder Sexu-

[6] Die folgenden Angaben spiegeln die Informationen wider, wie sie dem finalen Datensatz zu entnehmen sind, der als Grundlage der vorliegenden Evaluation diente.

albereich von der allgemeinen Rückfälligkeit unterschieden. Es wurde kontrolliert, dass alle Probanden, für die ein Auszug aus dem Strafregister eingeholt wurde, zum Zeitpunkt der Ziehung aus dem Strafregister noch lebten und in der Schweiz wohnhaft waren. Eine Kontrolle für Rückversetzung in den Vollzug fand nicht statt. Die Strafregisterauszüge wurden unabhängig von der Gruppenzugehörigkeit in zwei Wellen (März 2014 und Mai 2015) erhoben, zwischen dem 11.03.2014 und dem 05.05.2015.
Folgende Einschränkungen im Wissen zur Datenerhebung sind hinsichtlich der Auswertungsmöglichkeiten zu nennen:

- Erstens kann der Zeitraum, in dem die Interventionen durchgeführt wurden, nur näherungsweise durch die beiden Messzeitpunkte bestimmt werden (die Prä-Messung lag in jedem Fall vor Beginn der Intervention, in der Regel eine Woche vor Beginn; die Post-Messung lag in jedem Fall nach Beendigung der Intervention, in der Regel eine bis zwei Wochen nach Beendigung).
- Zweitens wurde nicht erhoben, zu welchem Zeitpunkt der Haft bzw. der Maßnahme (erstes, zweites oder letztes Drittel) die Interventionen bzw. die Datenerhebung durchgeführt wurden.

Durchführung der Interventionen

Im Folgenden werden die Interventionen, die im Rahmen des MV zur Anwendung gekommen sind, kurz dargestellt. Alle Interventionen werden in Anhang 2 näher erläutert.

Reasoning and Rehabilitation Programm

Das R&R- und später das R&R2-Programm wurde in den jeweiligen Räumlichkeiten der am MV teilnehmenden Institutionen durchgeführt. Frequenz und Dauer der Intervention entsprachen den Vorgaben des Manuals.
Das R&R-Programm wurde einmal wöchentlich in 35 hochstrukturierten manualbasierten Sitzungen à 120 Minuten durchgeführt.
Das R&R2-Programm wurde in 14 Sitzungen à 90 Minuten durchgeführt.
Die Sitzungen fanden in der Regel einmal wöchentlich statt, wobei das Manual idealerweise eine Frequenz von zwei bis drei Sitzungen pro Woche empfiehlt, dabei jedoch explizit Anpassungen an Umstände vor Ort erlaubt, solange eine Regelmäßigkeit der Gruppensitzungen gewährleistet ist. Die ideale Gruppengröße besteht aus acht Teilnehmern; pro Sitzung dürfen nicht weniger als vier und nicht mehr als zehn Personen teilnehmen (Ross et al., 2007).
Die Durchführung erfolgte durch zwei zuvor offiziell geschulte und dadurch für die Anwendung des R&R bzw. des R&R2 zertifizierte Behandler. Eine Person war dabei stets Psychotherapeut/in, die zweite Person stammte aus dem Vollzugspersonal, in Ausnahmefällen handelte es sich um zwei Psychotherapeut/-innen.

Nach Auskunft von Prof. Robert Ross, dem Entwickler des Programms, wird das R&R bzw. das R&R2 derzeit in 22 Ländern angewandt. Er schätzt, dass etwa 90‘000 junge oder erwachsene Straftäter (inklusive solche mit psychischen Störungen) an diesen Programmen teilgenommen haben. Die Implementierung in einem weiteren

Land (Chile) ist für Mai 2019 angedacht. Weiterhin wird das R&R2 durch den Forensisch-Psychiatrischen Dienst der Universität Bern angeboten.

Anti-Sexuelle-Aggressivität-Training®Suisse
Das ASAT®Suisse-Programm wurde in den Räumlichkeiten der jeweiligen am MV teilnehmenden Institution durchgeführt. Es wurde im Gruppensetting mit sechs bis acht Teilnehmern durchgeführt. Entsprechend den Vorgaben des Manuals wird das ASAT®Suisse in insgesamt 40-45 Sitzungen à 150 Minuten mit einer wöchentlichen Frequenz durchgeführt. Dies erfolgte durch zwei offiziell geschulte und für die Durchführung des ASAT®Suisse zertifizierte Behandler, wovon stets eine Person Psychotherapeut/in war und die zweite Person Mitarbeiter/in im Vollzug. In Ausnahmefällen handelte es sich um zwei Psychotherapeut/-innen.

Nach Auskunft von Frau Rita Steffes-enn (Entwicklerin des ASAT) ist hervorzuheben, dass das ASAT Suisse vom ursprünglichen ASAT zu differenzieren ist. Das ASAT ist länger und deliktunspezifischer konzipiert als das ASAT Suisse. Im März 2018 waren für das ASAT sowie das ASAT Jugend 83 Fachkräfte zertifiziert. Weitere 12 befanden sich zu diesem Zeitpunkt in Ausbildung. Hinsichtlich ASAT Suisse waren 30 Fachkräfte zertifiziert. Inwiefern das ASAT Suisse schweizweit angewandt wird, kann dennoch nicht abgeschätzt werden. Als gesichert kann gelten, dass das Programm durch den Forensisch-Psychiatrischen Dienst der Universität Bern angeboten wird, wobei die letzte Durchführung 2015 erfolgte.

Einzeltherapie in den Vergleichsgruppen
Die psychotherapeutische Behandlung entsprach der Standardbehandlung der jeweiligen am MV teilnehmenden Institution. Sie wurde im Einzelsetting durchgeführt. Typischerweise fand sie einmal wöchentlich bei einer Psychotherapeutin/ einem Psychotherapeuten in den Räumlichkeiten der jeweiligen Institution statt.

Datenschutz und ethische Richtlinien

Die Fragebögen wurden ausschließlich mit einem verschlüsselten Code als Pseudonym gekennzeichnet. Alle Probanden wurden umfassend über Maßnahmen zum Datenschutz informiert. Die Freiwilligkeit der Teilnahme wurde schriftlich festgehalten. Die Datenauswertung erfolgte nur in aggregierter Form, sodass keine Rückschlüsse auf einzelne Patienten, Therapeuten oder Anstalten möglich sind. Versuchsethische Aspekte wurden durch die zuständige Kantonale Ethikkommission (KEK) Bern abgesichert (Schreiben vom 27.08.2012, Az: 150/12).

Methode der Evaluation

Fokus und Hypothesen als Grundlage des Evaluationsberichts

Fokus des vorliegenden Berichts

In Absprache mit der Leitung des FPD Bern sollte der Fokus des vorliegenden Berichts gegenüber den ursprünglich beabsichtigten Auswertungen deutlich eingegrenzt werden. Während der Durchführung des Modellversuchs wurden zahlreiche Änderungen nötig, die sich auch aus den schwierigen Bedingungen eines multizentrischen Ansatzes in einem forensischen Umfeld ergaben. So standen beispielsweise Daten nicht im erhofften Maß und in der erhofften Qualität zur Verfügung.

Das Behandlungsprogramm R&R(2) wurde im vorliegenden Bericht anhand der folgenden beiden Fragestellungen evaluiert:
1) Auswirkung einer Teilnahme am R&R(2)-Programm auf Messwerte in Fragebögen, die auf die Erfassung von Persönlichkeitseigenschaften, Einstellungen und Verhaltensweisen zielen und die mit dem Rückfallrisiko in Zusammenhang stehen;
2) Auswirkung einer Teilnahme am R&R(2)-Programm auf die Rückfälligkeit.

Das Behandlungsprogramm ASAT®Suisse wurde ausschließlich über die Auswirkung einer Programmteilnahme auf die Rückfälligkeit evaluiert.

Ferner sollten aus den Ergebnissen allgemeine Schlussfolgerungen zu Therapieevaluationsstudien und Implikationen für zukünftige Evaluationskonzepte abgeleitet werden.

Fragestellung 1: Veränderungsmessung über Fragebogen (Evaluation des R&R(2)-Programms)

Um die Fragestellung 1 zu untersuchen, wurde die Veränderung in den vom Auftraggeber zur Verfügung gestellten Messwerte der folgenden Fragebögen im zeitlichen Verlauf ausgewertet:

- Kurzfragebogen zur Erfassung von Aggressivitätsfaktoren (K-FAF; Heubrock & Petermann, 2008),
- Inventar zur Erfassung interpersonaler Probleme (IIP-D; Horowitz et al., 2000),
- Hostile Attribution Bias (HAB; Tremblay & Belchevski, 2004) sowie
- Fragebogen zur Verantwortungsübernahme (VÜ; Gabriel et al., 2005; Oswald & Bütikofer, 2002).

Alle genannten Fragebögen werden in Anhang 3 näher erläutert.

Fragestellung 2: Rückfälligkeit (Evaluation des R&R(2)- und des ASAT®Suisse-Programms)

Als objektives Außenkriterium für die Wirksamkeit einer Teilnahme am R&R(2)-Programm und am ASAT®Suisse-Programm wurde die Rückfälligkeit der Probanden in den verschiedenen Untersuchungsgruppen verglichen. Als Indikator für erneute Verurteilungen wurden die vom Auftraggeber zur Verfügung gestellten Daten aus den Auszügen aus dem Schweizer Strafregister verwendet.

Hypothesen R&R(2)

Folgende Hypothesen wurden in Bezug auf das R&R(2) im Indikationsbereich der Gewaltstraftäter (GST-Gruppe) geprüft:

(1) Allgemeine Wirksamkeit der forensischen Therapien:
(1.1) Bei psychotherapeutisch behandelten Gewaltstraftätern (Gruppentherapie (GST R&R) oder Einzeltherapie (GST TAU)) lassen sich im Vergleich zur Kontrollgruppe (GST KG) positive Veränderungen von Einstellungen, Werthaltungen, Persönlichkeitseigenschaften und Verhaltensmaßen aufzeigen.
(1.2) Psychotherapeutisch behandelte Gewaltstraftäter (Gruppentherapie (GST R&R) oder Einzeltherapie (GST TAU)) weisen geringere Rückfallraten auf als die Kontrollgruppe (GST KG).

(2) Wirksamkeit des R&R(2)-Programms:
(2.1) Bei Gewaltstraftätern, die mit dem Gruppentherapieprogramm R&R(2) behandelt wurden (GST R&R), lassen sich im Vergleich zur Kontrollgruppe (GST KG) positive Veränderungen von Einstellungen, Werthaltungen, Persönlichkeitseigenschaften und Verhaltensmaßen aufzeigen.
(2.2) Mit R&R(2) behandelte Gewaltstraftäter (GST R&R) weisen geringere Rückfallraten auf als die Kontrollgruppe (GST KG).

(3) Vergleichbarkeit des R&R2- Programms mit der Standardbehandlung
(3.1) Unterscheiden sich Gewaltstraftäter, die mit dem Gruppentherapieprogramm R&R(2) behandelt wurden (GST R&R) von Gewaltstraftätern, die die anstaltsübliche Standardbehandlung in Form von Einzeltherapien erhielten (GST TAU) im Hinblick auf Veränderungen von Einstellungen, Werthaltungen und Persönlichkeitseigenschaften und Verhaltensmaßen?
(3.2) Unterscheiden sich mit R&R(2) behandelte Gewaltstraftäter (GST R&R) im Hinblick auf die Rückfallrate von Gewaltstraftätern, die mit der anstaltsüblichen Einzeltherapie behandelt wurden (GST TAU)?

Hypothesen ASAT®Suisse

Analog zur GST-Gruppe wurden im vorliegenden Bericht folgende Hypothesen in Bezug auf die Evaluation des ASAT®Suisse-Behandlungsprogramms im Indikationsbereich der Sexualstraftäter (SST-Gruppe) geprüft:

(1) Allgemeine Wirksamkeit der forensischen Therapien:
Psychotherapeutisch behandelte Sexualstraftäter (Gruppentherapie (SST R&R) oder Einzeltherapie (SST TAU)) weisen geringere Rückfallraten auf als die Kontrollgruppe (SST KG).

(2) Wirksamkeit des ASAT®Suisse-Programms:
Sexualstraftäter, die mit dem Gruppentherapieprogramm ASAT®Suisse behandelt wurden (SST ASAT), weisen geringere Rückfallraten auf als die Kontrollgruppe (SST KG).

(3) Vergleichbarkeit des ASAT®Suisse-Programms mit der Standardbehandlung:
Unterscheiden sich Sexualstraftäter, die mit dem Gruppentherapieprogramm ASAT®Suisse behandelt wurden (SST ASAT), im Hinblick auf die Rückfallrate von Sexualstraftätern, die mit der anstaltsüblichen Einzeltherapie behandelt wurden (SST TAU)?

Ablauf der Evaluation

Arbeitsschritte

Die Evaluation beinhaltete folgende Arbeitsschritte:

1. Festlegen des Fokus der Evaluation (gemeinsam mit dem Auftraggeber)
2. Sichtung und Aufbereitung von Literatur zum Thema Standards in der Therapie-Evaluation
3. Erarbeiten eines methodischen Verständnisses vom Ablauf des Modellversuchs unter besonderer Berücksichtigung methodisch relevanter Merkmale (Einschlusskriterien, Sicherstellen von Treatment Integrity, inhaltliche Entscheide bei der Kodierung von Rückfälligkeit)
4. Aufbereitung des Datensatzes. Dies beinhaltete die folgenden Arbeitsschritte:
    a. Sichtung der Variablen
    b. Klärung der inhaltlichen Relevanz von Variablen mit dem Auftraggeber
    c. Überführen von Variablen in ein analysebereites Format
    d. Plausibilitäts-Überprüfungen
    e. Neubildung von Variablen, die zentral für die Auswertung sind.

    Die Aufbereitung der Daten erfolgte in einem iterativen Prozess, da sich z.B. in einigen Fällen erst während der Durchführung der Berechnungen zeigte, dass wichtige Variablen nochmals neu gebildet werden mussten. Gemeinsam mit dem Auftraggeber wurde über die Auswahl geeigneter Variablen zur Erfassung der definierten Effektmaße diskutiert. Innerhalb des Evaluationsteams wurden die Ein- und Ausschlusskriterien der für die vorliegende Evaluation verwendeten Fälle diskutiert, sowie die Strategie zur statistischen Auswertung.
5. Statistische Auswertung. Alle Analysen wurden mit Stata® SE Version 14 (StataCorp, 2015) berechnet.

6. Verfassen des Berichts unter Einschluss von Revisionsschleifen innerhalb des Evaluationsteams.

Treffen mit dem Auftraggeber

Im Verlauf der Evaluation fanden drei persönliche Treffen zwischen dem Auftraggeber und dem Evaluationsteam statt. Der Inhalt dieser Treffen wird im Folgenden kurz erläutert.

1. Kick-off-Meeting am 23.08.2016 in Bern
   - Festlegen des Fokus der Evaluation
   - Zeitliche Rahmenbedingungen
   - Festlegen konkreter Arbeitsschritte und Zuständigkeiten
2. Präsentation erster Evaluationsergebnisse am 08.03.2017 in Bern
   - Vorstellung der „Rohauswertung“ der Daten
   - Gemeinsame Diskussion aufgetretener Schwierigkeiten
   - Festlegen des weiteren Vorgehens uns des Zeithorizonts
3. Präsentation und Besprechung der finalen Evaluationsergebnisse am 25.08.2017 in Zürich
   - Kritische Diskussion der bisher gezogenen Schlussfolgerungen
   - Absprache des weiteren zeitlichen Vorgehens

Umgang mit fehlenden Werten

Im vorliegenden Datensatz fand sich zu einer Vielzahl von Variablen, die zum Teil wesentlich für die Evaluation der Behandlungsprogramme waren, ein substanzieller Anteil fehlender Werte (Missing Values). Variablen, bei denen der Anteil fehlender Werte über 20% lag, wurden aus allen inferenzstatistischen Analysen ausgeschlossen. Dieser Anteil kann als moderat interpretiert werden, spiegelt nach Ansicht des Evaluationsteams ein angemessenes Verhältnis zwischen Datenverlust und Genauigkeit der Aussagen wider und wird als gängige Praxis angewandt; mit steigendem Anteil fehlender Werte steigt auch das Risiko für Verzerrungen, insbesondere bei einem Ausfallmuster, das wie vorliegend nicht dem des MCAR (Missing Completely at Random) entspricht (vgl. Acuna & Rodriguez, 2004; Finch, 2016).

Ausgewählte Variablen werden trotz eines höheren Anteils fehlender Werte aufgeführt, um zumindest einen deskriptiven Überblick in zentralen Messwerten oder Eigenschaften der Probanden zu erhalten. Eine Interpretation dieser Werte sollte aus den genannten Gründen nur unter größtem Vorbehalt vorgenommen werden.

Kategorisierung des Delikts

Im folgenden Abschnitt wird zunächst definiert, welche Delikte als Gewalt- bzw. Sexualdelikt gelten. Anschließend folgt die Beschreibung der verwendeten Kriterien,

nach denen ein Delikt als Rückfall eingeordnet wurde[7]. Die im Folgenden dargelegten Definitionen wurden sowohl zur Überprüfung der Einschlusskriterien verwendet als auch zur Definition der Rückfälligkeit als Außenkriterium.

Kategorisierung als Gewalt- oder Sexualdelikt

Als Gewaltdelikt wurden Verstöße gegen folgende StGB-Artikel gewertet: 111 (Vorsätzliche Tötung), 112 (Mord), 113 (Totschlag), 117 (Fahrlässige Tötung), 122 (Schwere Körperverletzung), 123 (Einfache Körperverletzung), 126 (Tätlichkeiten), 129 (Gefährdung des Lebens), 132 (Aufreizung zum Zweikampf (aStGB)), 133 (Raufhandel), 134 (Angriff), 136 (Verabreichung gesundheitsgefährdender Stoffe an Kinder), 140 (Raub), 183 (Freiheitsberaubung und Entführung), 184 (Freiheitsberaubung und Entführung, erschwerende Umstände), 185 (Geiselnahme), und 285 (Gewalt und Drohung gegen Behörden und Beamte).
Als Sexualdelikt wurden Verstöße gegen die folgenden StGB-Artikel gewertet: 187 (Sexuelle Handlungen mit Kindern), 188 (Sexuelle Handlungen mit Abhängigen), 189 (Sexuelle Nötigung), 190 (Vergewaltigung), 191 (Schändung), 193 (Ausnützung der Notlage), 194 (Exhibitionismus), 195 (Förderung der Prostitution), 197 (Pornografie), 198 (Sexuelle Belästigungen) und 213 (Inzest).

Definition Rückfall

Rückfälligkeit wurde wie folgt definiert:

a) Für Probanden, die bis zum Ende am MV teilgenommen haben (treatment as delivered):
    a. Der Proband war in Freiheit entlassen worden und befand sich zum Zeitpunkt der erneuten Deliktbegehung in Freiheit (gemäß der hierfür relevanten Variable „austrittsdatum“[8]).
    b. Die erneute Deliktbegehung fand nach dem Messzeitpunkt 2 (Variable „t2_datum“) statt, zu welchem die Therapie sicher beendet war.
b) Für alle zu Beginn eingeschlossenen Probanden (Intent-to-treat):

[7] Die Kategorisierung des Delikts, die bereits im Datensatz enthalten war (Variable „delikt_kategorisch“), konnte nicht verwendet werden, da sie eine geringe Plausibilität aufwies: In der Mehrzahl der als "Rückfall" kodierten Delikte liegt das Deliktdatum des „Rückfalls“ vor dem T1-Datum. Insgesamt 448 Delikte, verteilt auf 80 Codes, wurden so fälschlicher Weise als Rückfall klassifiziert. Umgekehrt lagen einige Delikte, die als Vorstrafe oder Indexdelikt kodiert waren, zeitlich nach dem T1- bzw. dem T2-Datum.

[8] Diese Variable beinhaltete die durch das FPD-Forschungsteam als am reliabelsten eingeschätzte Information zur Frage, ob und wenn ja wann ein Austritt erfolgte und sollte demensprechend ausschließlich verwendet werden. Diese Information wurde retrospektiv auf Grundlage der jeweiligen Anstaltsakten erhoben. Dennoch kann nicht ausgeschlossen werden, dass bei einzelnen Probanden versäumt wurde oder es nicht möglich war, das Datum eines erfolgten Austritts in der Variable „austrittsdatum“ festzuhalten. Es ist daher möglich, dass zum Zeitpunkt der Strafregisterauszüge bereits mehr Probanden entlassen waren, als dies aus den vorliegenden Daten hervorgeht. Diese Probanden wären folglich bei der Analyse der Rückfälligkeit nicht berücksichtigt. Ob diese Einschränkung eher zu einer Über- oder einer Unterschätzung der Rückfallrate in einzelnen oder allen Untersuchungsbedingungen führt, oder ob sich dadurch entstandene Verzerrungseffekte letztlich ausmitteln, kann auf Grundlage der vorliegenden Daten nicht beantwortet werden.

a. Der Proband war in Freiheit entlassen worden und befand sich zum Zeitpunkt der erneuten Deliktbegehung in Freiheit (gemäß der hierfür relevanten Variable „austrittsdatum“).
b. Die erneute Deliktbegehung fand nach dem Messzeitpunkt 1 statt (in Ermangelung eines genaueren Schätzers für den Zeitraum der Behandlung; siehe Abschnitt „Messzeitpunkte und Dauer **der Erhebung**“).

Von der Analyse zur Rückfälligkeit ausgeschlossen wurden Fälle, für die kein Strafregisterauszug vorlag, sodass keine verlässlichen Angaben zu Art und Datum der begangenen Delikte vorlagen (siehe Abschnitt „Von der Auswertung ausgeschlossene Probanden“).

Fehlende Möglichkeit zur Unterscheidung zwischen Vorstrafe und Indexdelikt

Die Klassifikation als Indexdelikt oder Vorstrafe konnte durch das Evaluationsteam nicht reliabel vorgenommen werden. Zum einen kann diese Unterscheidung anhand der vorhandenen Daten im Gegensatz zur Klassifikation als Rückfall nicht eindeutig getroffen werden: Selbst bei bekanntem Delikt- und Urteilsdatum kann es z.B. durch ein weitergezogenes Urteil vor das Bundesgericht mit entsprechender Verfahrensdauer und zwischenzeitlicher erneuter Straffälligkeit zur Diffusion kommen. Zum anderen wird die Entscheidung der Zuteilung eines Probanden zu einer bestimmten Intervention auch nach klinischen Gesichtspunkten getroffen. Konkret besteht z.B. die Möglichkeit, dass Personen, die bereits in der länger zurückliegenden Vergangenheit wegen Gewaltdelikten auffällig waren, bei denen es sich eigentlich um Vorstrafen handelte, in die R&R2-Experimentalgruppe eingeteilt wurden, wenn dies aus therapeutischer Sicht sinnvoll erschien.

Von der Auswertung ausgeschlossene Probanden

Unklare Indikation für Teilnahme am Therapieprogramm

Der Fokus bei der Implementierung und Evaluation von Therapien sollte auf spezifischen Subgruppen von Straftätern liegen. Zwar wurde weder das R&R- noch das R&R2-Programm ausschließlich für Gewalt- oder Sexualstraftäter entwickelt (Ross et al., 1986; Ross et al., 2007), jedoch bestand das Zielklientel des MV für die Evaluation des R&R2 in Gewaltstraftätern (FPD Bern, 2009). Sexual- und Gewaltdelikte werden für die Indikation zur Teilnahme am R&R2-Programm zusammengefasst, da Sexualdelikte eine spezielle Variante von Gewalt darstellen.

Probanden, die dem R&R2-Programm zugewiesen worden waren, ohne dass eine Verurteilung für ein Gewalt- oder Sexualdelikt dokumentiert ist (*N*=31), wurden aus diesen Gründen von der Analyse ausgeschlossen. Ebenso wurden in der Folge Probanden ohne dokumentiertes Gewalt- oder Sexualdelikt aus der Vergleichs- (*N*=10)

und aus der Kontrollgruppe (*N*=15) ausgeschlossen, um ein vergleichbares Zielklientel in allen drei Untersuchungsbedingungen zu erhalten.

Tabelle 1 zeigt den Anteil der Probanden der GST-Gruppe, die gemäß Strafregister-Auszug vor dem ersten Messzeitpunkt weder für ein Gewalt- noch für ein Sexualdelikt verurteilt worden waren und aus diesem Grund von den Analysen ausgeschlossen wurden.
Bei den Fällen, zu denen kein Strafregisterauszug vorlag, wurde wie folgt vorgegangen: Anhand der relevanten Variablen im Aktenbooklet „gewalt“ (Anlassdelikt Gewaltdelikt), „sexual“ (Anlassdelikt Sexualdelikt), „gewaltkat“ (Anlassdelikt Gewaltdelikt kategorisiert) und „sexualkat“ (Anlassdelikt Sexualdelikt kategorisiert) wurde überprüft, ob hinreichend Informationen vorlagen, ob es sich beim Anlassdelikt um ein Gewalt- oder Sexualdelikt handelte. War dies der Fall, wurden diese Fälle in der Stichprobe belassen (sofern sie nicht aus einem der in den folgenden Abschnitten genannten Gründe ausgeschlossen wurden). Ergaben sich auch aus dem Aktenbooklet keine hinreichenden Informationen, dass es sich um einen Gewalt- oder Sexualstraftäter handelte, wurden diese Fälle von sämtlichen Analysen ausgeschlossen.
Ein Chi-Quadrat-Test ergab keine Hinweise auf einen unterschiedlichen Anteil an Probanden zwischen den drei Bedingungen der GST-Gruppe, die mindestens ein Gewalt- oder Sexualdelikt vor dem ersten Messzeitpunkt begangen hatten ($\chi^2(2)=2.73$; $p=.255$).

Tabelle 1. Anteil Probanden mit fehlender Indikation für das Therapieprogramm: Gewaltstraftäter

	GST Bedingung	Fehlende Werte [a)]	*N*	Anteil	*p*
Kein Gewalt- oder Sexualdelikt vor T1	GST R&R	3.7% (*N*=6)	157	19.6% (*N*=31)	.255
	GST TAU	2.3% (*N*=2)	87	11.5% (*N*=10)	
	GST KG	8.2% (*N*=8)	90	16.7% (*N*=15)	
	GST Gesamt	**4.6% (*N*=16)**	**335**	**16.7% (*N*=56)**	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [a)] Es handelt sich hierbei um Fälle, in denen kein Strafregisterauszug vorliegt, sodass keine zuverlässigen Angaben zu Art und Datum der begangenen Delikte vorliegen.

Probanden, die dem ASAT®Suisse-Programm zugewiesen worden waren, ohne dass eine Verurteilung für ein Sexualdelikt dokumentiert ist (*N*=1), wurden von der Analyse ausgeschlossen. Neben den bereits genannten Gründen wurden diese Probanden auch aufgrund der nicht gegebenen Indikation für die Teilnahme am ASAT®Suisse gegeben: Das ASAT®(Suisse) wurde ausschließlich für erwachsene männliche Sexualstraftäter entwickelt (Falk & Steffes-enn, 2014; Steffes-enn, 2005, 2008). Um dem Therapieprogramm methodisch die größtmögliche Chance für einen Wirksamkeitsnachweis zu geben, sollte das Programm auch an der Population unter-

sucht werden, für die es entwickelt worden ist (vgl. Lösel & Schmucker, 2005). Um eine Vergleichbarkeit in allen Untersuchungsgruppen zu gewährleisten, musste diese Indikation ebenso für die Vergleichs- (Ausschluss: *N*=1) und die Kontrollgruppe (*N*=1) angewandt werden. Tabelle 2 zeigt den Anteil an Probanden der SST-Gruppe, die gemäß Strafregisterauszug kein Sexualdelikt vor dem ersten Messzeitpunkt begangen hatten und die daher von sämtlichen Analysen ausgeschlossen wurden.
Ein Chi-Quadrat-Test ergab keine Hinweise auf Unterschiede zwischen den drei Bedingungen der SST-Gruppe im Anteil derjenigen Probanden, die mindestens ein Sexualdelikt vor dem ersten Messzeitpunkt begangen hatten ($\chi^2(2)=3.97$; $p=.137$).

Tabelle 2. Anteil Probanden mit fehlender Indikation für das Therapieprogramm: Sexualstraftäter

SST Bedingung	Fehlende Werte[1)]	*N* Gesamt	Kein Sexualdelikt vor T1	*p*
SST ASAT	7.7% (*N*=4)	48	2.1% (*N*=1)	
SST TAU	6.7% (*N*=2)	28	3.6% (*N*=1)	.137
SST KG	7.7% (*N*=1)	12	8.3% (*N*=1)	
SST Gesamt	**7.4% (*N*=7)**	**88**	**3.4% (*N*=3)**	

Anmerkungen. ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [1)] Es handelt sich hierbei um die oben genannten Fälle, in denen kein Strafregisterauszug vorliegt, sodass keine zuverlässigen Angaben zu Art und Datum der begangenen Delikte vorliegen. Diese Fälle wurden, wie im folgenden Abschnitt erläutert, von den Analysen ebenfalls ausgeschlossen.

Fehlende Möglichkeit zur Bestimmung der Effektkriterien

Für 16 Fälle der GST-Gruppe lag kein Strafregisterauszug vor, sodass diese Probanden nicht für die Analyse zur Rückfälligkeit berücksichtigt werden konnten (siehe Tabelle 1): Da in diesen Fällen im Datensatz auch kein Datum zum Strafregisterauszug vorlag, ist davon auszugehen, dass für diese Personen tatsächlich kein Strafregisterauszug vorlag und nicht, dass der betreffende Proband keine Einträge im VOSTRA hat.
Sechs dieser Fälle ohne Strafregisterauszug hatten zusätzlich fehlende Angaben zum zweiten Messzeitpunkt, sodass für diese Probanden keine der beiden Fragestellungen beantwortet werden konnte. Diese sechs Fälle wurden von sämtlichen Analysen ausgeschlossen, einschließlich der Stichprobenbeschreibungen. Dabei handelte es sich um drei Probanden aus der Experimental-, einen Probanden aus der Vergleichs- und zwei Probanden aus der Kontrollgruppe. Zehn der Fälle ohne vorliegenden Strafregisterauszug hatten jedoch sowohl an der Prä- als auch der Post-Messung teilgenommen, sodass sie für die Analyse der Fragestellung 1 (Veränderungen in den Messwerten der Fragebögen) berücksichtigt werden konnten. Aus diesem Grund wurden sie auch in die Stichprobenbeschreibungen eingeschlossen.

In der SST-Gruppe lagen in sieben Fällen keine Strafregisterauszüge vor (siehe Tabelle 2). Da die Analyse in dieser Gruppe ausschließlich die Rückfälligkeit umfasste, wurden diese sieben Probanden von allen Auswertungen, einschließlich der Stich-

probenbeschreibungen, ausgeschlossen. Die ausgeschlossenen Probanden stammten in vier Fällen aus der Experimental-, in zwei Fällen aus der Vergleichs- und in einem Fall aus der Kontrollgruppe.

Ausschlüsse aufgrund anderer Gründe

In Fall 1286 (SST R&R) ist das Datum des ersten und zweiten Messzeitpunktes identisch. Dieser Fall wurde von allen Analysen ausgeschlossen.

Zusätzlich zum vollständigen Ausschluss von Probanden wurden in einigen Fällen aus verschiedenen Gründen Rohdaten verändert. Diese Einzelfälle sind unter genauer Angabe des jeweiligen Grundes in Anhang 4 aufgeführt.

Ergebnisse

Deskriptive Beschreibung der Stichproben

Alle Probanden waren männlich, volljährig und wurden aus den vom FPD Bern betreuten Justizinstitutionen sowie weiteren Anstalten des Strafvollzugskonkordats Nordwest- und Innerschweiz rekrutiert. Im Einzelnen stammten die Probanden aus den folgenden Anstalten: Thorberg, Witzwil, St. Johannsen, Wauwilermoos, Im Schachen, Lenzburg, Schöngrün, Pöschwies, Grosshof Luzern, Bewährungsdienste Kanton Bern, Ambulatorium FPD Bern[9]. Tabelle 3 zeigt die Anzahl der für die vorliegende Auswertung grundsätzlich berücksichtigten Probanden, separat aufgeführt nach Gruppenzugehörigkeit (GST oder SST) und Bedingung (Experimental-, Vergleichs- oder Kontrollgruppe), d.h. die Stichprobenzusammensetzung nach Ausschluss der oben genannten Fälle.

Tabelle 3. Stichprobenzusammensetzung: Für die Evaluation berücksichtigte Fälle

GST Bedingung	***N***	**SST Bedingung**	***N***
GST R&R	129	SST ASAT	47
GST TAU	78	SST TAU	27
GST KG	79	SST KG	11
GST Gesamt	**286**	**SST Gesamt**	**85**

Anmerkungen. GST: Gruppe der Gewaltstraftäter; SST: Gruppe der Sexualstraftäter; R&R: Reasoning and Rehabilitation Programm; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe.

[9] Eine genauere Angabe ist auf Grundlage der vorliegenden Daten nicht möglich. Insbesondere erlauben die vorliegenden Daten keine Identifikation derjenigen Probanden, die sich in einer ambulanten Behandlung befanden. Diese Probanden konnten daher auch für die Analyse zur Rückfälligkeit nicht berücksichtigt werden: Sie befanden sich schon während der Interventionsphase in Freiheit, d.h. „at risk“. Als Startpunkt der „time at risk“ kann jedoch ausschließlich auf das Entlassdatum (gemäß Variable „austrittsdatum“) zurückgegriffen werden. Für die ambulanten Probanden liegt in dieser Variablen jedoch per definitionem keine Angabe vor. In welche Richtung die dadurch zustande gekommenen Verzerrungen weisen und ob sie zu differenziellen Effekten in den unterschiedlichen Bedingungen führen, kann auf Grundlage vorliegender Daten ebenfalls nicht beantwortet werden.

Ergebnisse R&R(2)

Im Folgenden werden die Ergebnisse der Evaluation der GST-Gruppe dargestellt.

In einem ersten Schritt werden die den Analysen zugrundeliegenden Stichproben vergleichend beschrieben. Es werden Merkmale untersucht, welche einen Einfluss auf die Messwerte der Fragebögen (Fragestellung 1) bzw. auf das Rückfallrisiko (Fragestellung 2) haben können. Dies beinhaltet demografische Merkmale, die Vorstrafenbelastung, die psychiatrische Belastung, Summenwerte in etablierten Risk-Assessment-Instrumenten sowie die Therapieerfahrung der Probanden.

Zunächst werden alle in die Studie eingeschlossenen Probanden der GST-Gruppe auf vorbestehende Unterschiede zwischen den drei Bedingungen in zentralen Merkmalen überprüft. Eine Vergleichbarkeit der Bedingungen ist ein wichtiges Kriterium für die Einschätzung der Qualität eines quasi-experimentellen Designs (siehe auch Abschnitt „Einschätzung der methodischen Qualität auf der Maryland Scientific Methods Scale“): In einem solchen Design besteht die Gefahr eines Selektions-Bias, d.h. durch die nicht randomisierte Gruppenzuweisung besteht eine erhöhte Wahrscheinlichkeit, dass sich die Probanden der drei Bedingungen bereits a priori in Merkmalen unterscheiden, welche einen Einfluss auf die festgelegten Effektkriterien haben können. Anschließend wird der Stichprobenschwund präsentiert, um das Verhältnis von Probanden, die nur am ersten Messzeitpunkt, nicht aber am zweiten Messzeitpunkt teilgenommen haben (im Folgenden „Abbrecher“ genannt) zu Probanden, die den MV bis zum Ende durchlaufen haben (im Folgenden „Vollender“ genannt), in den drei Bedingungen aufzuzeigen. Die Vollender werden danach mit den Abbrechern verglichen, um Rückschlüsse zu ziehen, ob das Ausscheiden aus dem MV mit bestimmten Merkmalen der Probanden zusammenhängt. Dieser Vergleich wird sowohl innerhalb der einzelnen Bedingungen durchgeführt als auch in Bezug auf die GST-Gesamtgruppe. Dieser Zusammenhang kann somit auf der einen Seite hinsichtlich des Ausscheidens aus dem R&R(2)-Behandlungsprogramm überprüft werden, auf der anderen Seite aber auch hinsichtlich des Ausscheidens aus dem MV als Ganzes. Darüber hinaus wurde ein Vergleich zwischen den Bedingungen durchgeführt, der ausschließlich die Vollender berücksichtigt. Damit soll überprüft werden, ob eine Veränderung in den Messwerten der Fragebögen (Fragestellung 2) mit Merkmalen der Probanden zusammenhängt. Aus Gründen der Übersichtlichkeit findet sich dieser Vergleich in Anhang 1.

Nach diesen Stichprobenvergleichen werden zunächst die Ergebnisse zur Fragestellung 1 präsentiert: Durch die Analyse der Differenzwerte in den ausgewählten Fragebögen soll untersucht werden, ob sich die Messwerte der Probanden in den drei Bedingungen zwischen den beiden Messzeitpunkten verändert haben, um so auf einen Effekt der (Nicht-) Behandlung rückschließen zu können. Schließlich werden die Ergebnisse der Analyse zur Rückfälligkeit (Fragestellung 2) dargestellt. Durch den Ver-

gleich der Rezidivraten zwischen den drei Bedingungen soll auf Effekte einer (Nicht-) Behandlung rückgeschlossen werden.

Vergleichende Stichprobenbeschreibungen GST

Vorbestehende Unterschiede zu T1

Im Folgenden werden vorbestehende Unterschiede zwischen den Probanden der drei Bedingungen der GST-Gruppe hinsichtlich demografischer Merkmale, Art des Indexdelikts und Merkmalen mit Einfluss auf das Rückfallrisiko dargestellt. Diese Auswertung soll die Vergleichbarkeit der Probanden in den drei Bedingungen der GST-Gruppe statistisch überprüfen.

Demografische Merkmale: Vorbestehende Unterschiede

Nationalität

Tabelle 4 zeigt den Anteil an Probanden mit Schweizer Staatsangehörigkeit in den drei Bedingungen der GST-Gruppe zu Beginn des MV. Die Probanden in den drei Bedingungen der GST-Gruppe unterschieden sich zu Beginn signifikant in ihrem Anteil an Schweizer Staatsbürgern ($\chi^2(2)=13.80$; $p=.001$): In der Kontrollgruppe befanden sich signifikant mehr Ausländer als in der Experimental- ($\chi^2(1)=12.18$; $p<.001$) und Vergleichsgruppe ($\chi^2(1)=8.08$; $p=.004$).

Tabelle 4. Vorbestehende Unterschiede in der Nationalität GST

GST Bedingung	Fehlende Werte [% (*N*)]	*N*	Schweizer Nationalität [% (*N*)]	*p*
GST R&R	0.8% (*N*=1)	128	71.1% (*N*=91)	
GST TAU	0.0% (*N*=0)	78	69.2% (*N*=54)	.001
GST KG	0.0% (*N*=0)	79	46.8% (*N*=37)	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe.

Alter

Tabelle 5 zeigt das Alter der in die Auswertung eingeschlossenen Probanden der GST-Gruppe zum Zeitpunkt des Indexurteils. Kruskal-Wallis-H-Tests ergaben keine Hinweise auf Altersunterschiede zwischen den Probanden in den drei Bedingungen der GST-Gruppe zu Beginn der Studie ($\chi^2(2)=0.04$; $p=.980$).

Tabelle 5. Vorbestehende Unterschiede im Alter GST

GST Bedingung	Fehlende Werte [% (*N*)]	*N*	Alter Indexurteil [*M* (*SA*)]	*p*

GST R&R	3.1% (*N*=4)	125	31.4 (8.9)	
GST TAU	3.9% (*N*=3)	75	32.7 (11.2)	.980
GST KG	5.1% (*N*=4)	75	31.9 (9.5)	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe.

Bildungsniveau

Tabelle 6 zeigt das höchste erreichte Bildungsniveau aller eingeschlossenen Probanden der GST-Gruppe[10].

Ein exakter Test nach Fisher ergab, dass sich das Bildungsniveau der Probanden in den drei Bedingungen signifikant unterscheidet (p=.004). Dieser lässt sich wahrscheinlich vor allem auf den erhöhten Anteil an ausländischen Probanden in der Kontrollgruppe im Vergleich zu den beiden anderen Bedingungen zurückführen: In der Experimentalgruppe ging der niedrigere Anteil ausländischer Probanden gegenüber der Kontrollgruppe mit einem höheren Anteil an Probanden mit abgeschlossener Lehre, Matura oder Hochschulstudium einher ($\chi^2(1)=10.26$; p=.001). Der höhere Anteil ausländischer Probanden in der Kontrollgruppe ging ferner mit einem niedrigeren Anteil an Probanden, die zumindest ihre Schulpflicht abgeschlossen harten, einher. Dies gilt sowohl im Vergleich zur Experimental- ($\chi^2(1)=8.09$; p=.004) als auch zur Vergleichsgruppe ($\chi^2(2)=8.46$; p=.004).

Werden ausschließlich die Probanden mit Schweizer Nationalität eingeschlossen, ist das Ergebnis eines exakten Tests nach Fisher zur Überprüfung von Unterschieden im Bildungsniveau zwischen den drei Bedingungen der GST-Gruppe nicht mehr signifikant (p=.359).

Tabelle 6. Vorbestehende Unterschiede im Bildungsniveau GST

GST Bedingung	Fehlende Werte [% (*N*)]	*N*	Höchster Abschluss	Anteil [% (*N*)]	*p*
GST R&R	4.7% (*N*=6)	123	< 7 Jahre Schulbildung*)	2.4% (*N*=3)	.004
			Schulpflicht abgeschl. ggf. zzgl. Anlehre	32.5% (*N*=40)	
			Mindestens abgeschl. Lehre	35.0% (*N*=43)	
			Nicht-Schweizer Probanden	30.1% (*N*=37)	
GST TAU	6.4% (*N*=5)	73	< 7 Jahre Schulbildung	0% (*N*=0)	
			Schulpflicht abgeschl. ggf. zzgl. Anlehre	39.7% (*N*=29)	
			Mindestens abgeschl. Lehre	27.4% (*N*=20)	
			Nicht-Schweizer Probanden	32.9% (*N*=24)	

[10] Ausländische Probanden werden separat ausgewiesen, da bei diesen eine valide Zuordnung des Bildungsniveaus anhand vorliegender Daten nicht möglich ist.

GST KG	5.1% (*N*=4)	75	< 7 Jahre Schulbildung	2.7% (*N*=2)
			Schulpflicht abgeschl. ggf. zzgl. Anlehre	21.3% (*N*=16)
			Mindestens abgeschl. Lehre	20.0% (*N*=15)
			Nicht-Schweizer Probanden	56.0% (*N*=42)

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; *) Die Antwortkategorien beruhen auf dem „Bildungsstatus nach Wilson“, der im Aktenbooklet des FPD zur Einteilung des Bildungsniveaus verwendet worden war. Weiterführende Informationen zum Bildungsstand der Probanden sind auf Grundlage der zur Verfügung stehenden Daten nicht darstellbar.

Psychiatrische Belastung: Vorbestehende Unterschiede

Tabelle 7 zeigt den Anteil derjenigen für die Auswertung berücksichtigten Probanden in den drei Bedingungen der GST-Gruppe, die eine Diagnose in verschiedenen Bereichen des ICD-10 aufweisen: Persönlichkeitsstörungen (Kategorie F6), Schizophrenie, schizotype und wahnhafte Störungen (Kategorie F2), affektive Störungen (Kategorie F3) sowie Störungen durch psychotrope Substanzen (Kategorie F1). Es wurden hiermit Diagnosebereiche gewählt, die in forensischen Kontexten typischerweise und im Vergleich zur Allgemeinbevölkerung gehäuft zu beobachten sind. Die Diagnosen wurden dem (aktuellsten) psychiatrischen Gutachten entnommen. Der Zeitpunkt der Diagnosestellung wurde nicht spezifisch erfasst.
Aufgrund des hohen Anteils an fehlenden Werten, vor allem in der Kontrollgruppe, lassen sich keine stichhaltigen Aussagen zur psychiatrischen Belastung der Probanden in den drei Bedingungen treffen. Der hohe Anteil fehlender Werte in der Kontrollgruppe ergibt sich aus der Tatsache, dass sich die Probanden in dieser Bedingung nicht in Therapie befanden und in vielen Fällen auch keine Akteninformationen zur psychiatrischen Belastung in Form von Gutachten vorlagen. Einzelvergleiche zwischen der Experimental- und der Vergleichsgruppe, sofern der Anteil fehlender Werte maximal 20% beträgt, ergaben keine Hinweise auf signifikante Unterschiede.

Tabelle 7. Vorbestehende Unterschiede in der psychiatrischen Vorbelastung GST

	GST Bedingung	Fehlende Werte [% (*N*)]	*N*	Anteil [% (*N*)]	*p*
Persönlichkeitsstörung (ICD-10 Kategorie F6)[1)]	GST R&R	11.6% (*N*=15)	114	57.0% (*N*=65)	N/A
	GST TAU	21.8% (*N*=17)	61	45.9% (*N*=28)	
	GST KG	46.8% (*N*=37)	42	40.5% (*N*=17)	
Schizophrenie, schizotype und wahnhafte Störungen (ICD-10 Kategorie F2)	GST R&R	12.4% (*N*=16)	113	6.2% (*N*=7)	N/A
	GST TAU	19.2% (*N*=15)	63	12.7% (*N*=8)	
	GST KG	48.1% (*N*=38)	41	4.9% (*N*=2)	
Störungen durch psychotrope Substanzen	GST R&R	10.1% (*N*=13)	116	56.9% (*N*=66)	N/A
	GST TAU	18.0% (*N*=14)	64	43.8% (*N*=28)	

(ICD-10 Kategorie F1)	GST KG	45.6% (*N*=36)	43	41.9% (*N*=18)	
Affektive Störungen (ICD-10 Kategorie F3)	GST R&R	13.2% (*N*=17)	112	8.9% (*N*=10)	N/A
	GST TAU	20.5% (*N*=16)	62	4.8% (*N*=3)	
	GST KG	48.1% (*N*=38)	41	7.3% (*N*=3)	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [1)] Bei der Mehrheit der diagnostizierten Persönlichkeitsstörungen handelte es sich um die dissoziale PS (ICD-10: F60.2): R&R: 35% (*n*=23); TAU: 25% (*n*=7); KG: 35% (*n*=6).

Summenwerte in Risk-Assessment-Instrumenten: Vorbestehende Unterschiede

Die angewandten Risk-Assessment-Instrumente VRAG und PCL-R zeichnen sich durch einen substanziellen Anteil an fehlenden Werten aus, sodass zu den vorbestehenden Unterschieden der Probanden hinsichtlich ihres (Baseline-) Risikos keine verlässliche Aussage getroffen werden kann. Die Mittelwerte der Probanden in den drei Bedingungen der GST-Gruppe im Summenwert von VRAG und PCL-R sind dennoch in Tabelle 8 deskriptiv aufgeführt.

Tabelle 8. Vorbestehende Unterschiede in Risk-Assessment-Instrumenten GST

	GST Bedingung	Fehlende Werte [% (*N*)]	*N*	*M*(*SA*)	*Median*
VRAG Summenwert[*)]	GST R&R	42.6% (*N*=55)	74	9.3 (12.5)	8.5
	GST TAU	52.6% (*N*=41)	37	7.9 (13.2)	10.0
	GST KG	76.0% (*N*=60)	19	3.3 (11.9)	1.0
PCL-R Summenwert	GST R&R	31.8% (*N*=41)	88	19.4 (8.5)	19.5
	GST TAU	33.3% (*N*=26)	52	17.1 (7.7)	15.4
	GST KG	74.7% (*N*=59)	20	18.5 (76.6)	18.5

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [*)] Hierbei wurde überprüft, ob der VRAG-Summenwert bereits korrigiert wurde (dies ist bis zu einer maximalen Anzahl von vier fehlenden Angaben je Bewertung möglich). Dies ist der Fall: Der Anteil an fehlenden Werten im Item „vragtot“ ist identisch mit dem Anteil derjenigen Fälle, die in mehr als vier Items des VRAG fehlende Angaben haben.

Therapieerfahrung: Vorbestehende Unterschiede

Tabelle 9 zeigt den Anteil der Probanden aus der GST-Gruppe, die bereits vor Beginn des MV mindestens einmal in psychotherapeutischer Behandlung waren.

Tabelle 9. Vorbestehende Unterschiede in der Therapieerfahrung GST

	GST	Fehlende Werte	*N*	Anteil [% (*N*)]	*p*

	Bedingung	[% (*N*)]			
Psychotherapie-Erfahrung	GST R&R	14.7% (*N*=19)	110	56.4% (*N*=62)	*N/A*
	GST TAU	16.7% (*N*=13)	65	38.5% (*N*=25)	
	GST KG	36.7% (*N*=29)	50	40.0% (*N*=20)	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; Rot = Mehr als 20% fehlende Werte.

Im Einzelvergleich zwischen der Experimental- und der Vergleichsgruppe, in denen der Anteil fehlender Werte bei maximal 20% liegt, ergab ein Chi-Quadrat-Test signifikante Unterschiede: Der Anteil derjenigen Probanden, die vor Beginn des MV mindestens einmal psychotherapeutisch behandelt worden waren, ist in der Experimentalgruppe größer als in der Vergleichsgruppe ($\chi^2(1) = 5.24$; p=.022).

Stichprobenschwund GST

Tabelle 10 zeigt die Anzahl der in die Evaluation eingeschlossenen Studienteilnehmer (ST) der GST-Gruppe sowie den Anteil derjenigen Probanden, welche an der Studie bis zum Ende (Vollender: Daten liegen sowohl für Prä- als auch Post-Messung vor) bzw. nicht bis zum Ende (Abbrecher: Nur Daten der Prä-Messung liegen vor) teilgenommen haben. Eine weiterführende Unterscheidung nach den Gründen für einen Abbruch, insbesondere, ob es sich bei den Abbrechern um Therapie- oder um Studienabbrecher handelte, lässt sich nicht vornehmen, da hierzu keine Daten vorliegen. Auch der Anteil derjenigen angefragten potenziellen Probanden, die ihre Teilnahme bereits vor Studienbeginn abgelehnt hatten, wurde nicht erhoben.
Ein Chi-Quadrat-Test ergab signifikante Unterschiede in den drei Bedingungen der GST-Gruppe hinsichtlich des Anteils an Studienteilnehmern, welche die Studie nicht bis zum Ende absolviert haben ($\chi^2(2)=12.19$, p=.002): Post-Hoc-Tests zeigten, dass in der Experimentalgruppe signifikant weniger Abbrecher zu verzeichnen sind als in der Kontrollgruppe ($\chi^2(1)=12.02$, p=.001). (Der Unterschied zwischen R&R und TAU ist nach Bonferroni-Korrektur des Alpha-Niveaus (α=.025) knapp nicht signifikant: $\chi^2(1)=4.64$, p=.031).

Tabelle 10. Anteil Vollender und Abbrecher GST

GST Bedingung	In die Auswertung aufgenommene Probanden (*N*)	Vollender [% (*N*)]	Abbrecher [% (*N*)]	*p*
GST R&R	129	85.3% (*N*=110)	14.7% (*N*=19)	.002
GST TAU	78	73.1% (*N*=57)	26.9% (*N*=21)	
GST KG	79	64.6% (*N*=51)	35.4% (*N*=28)	
GST Gesamt	**286**	**76.2% (*N*=218)**	**23.8% (*N*=68)**	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe.

Unterschiede zwischen Abbrechern und Vollendern

Demografische Merkmale: Unterschiede zwischen Abbrechern und Vollendern

Nationalität

Tabelle 11 zeigt den Anteil Schweizer Probanden in den drei Bedingungen der GST-Gruppe sowie für die GST-Gesamtstichprobe, separat für Abbrecher und Vollender. Ein Chi-Quadrat-Test ergab signifikante Unterschiede zwischen Abbrechern und Vollendern über alle Bedingungen der GST-Gruppe hinweg ($\chi^2(2)=3.89$, $p=.048$): Die Vollender des Modellversuchs besaßen zu einem größeren Anteil die Schweizer Staatsbürgerschaft. Innerhalb der einzelnen Bedingungen fanden sich hingegen keine Hinweise auf Unterschiede zwischen Abbrechern und Vollendern.

Tabelle 11. Nationalität GST Abbrecher versus Vollender

GST Bedingung	Fehlende Werte [% (*N*)]	*N*	Schweizer Nationalität [% (*N*)]	*p*
R&R alle zugeordneten Probanden	0.8% (*N*=1)	128	71.1% (*N*=91)	
R&R Abbrecher	5.3% (*N*=1)	18	61.1% (*N*=1)	.314 [a)]
R&R Vollender	0% (*N*=0)	110	72.7% (*N*=80)	
TAU alle zugeordneten Probanden	0.0% (*N*=0)	78	69.2% (*N*=54)	
TAU Abbrecher	0.0% (*N*=0)	21	71.4% (*N*=15)	.799 [a)]
TAU Vollender	0.0% (*N*=0)	57	68.4% (*N*=39)	
KG alle zugeordneten Probanden	0.0% (*N*=0)	79	46.8% (*N*=37)	
KG Abbrecher	0.0% (*N*=0)	28	35.7% (*N*=10)	.142 [a)]
KG Vollender	0.0% (*N*=0)	51	52.9% (*N*=27)	
GST alle Probanden	0.4% (*N*=1)	285	63.9% (*N*=182)	
GST Abbrecher	1.5% (*N*=1)	67	53.7% (*N*=36)	**.048** [a)]
GST Vollender	0.0% (*N*=0)	218	67.0% (*N*=146)	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [a)] Ergebnis des Chi-Quadrat-Tests. Dieser Test dient der Überprüfung des Zusammenhangs zwischen kategorialen abhängigen Variablen.

Alter

Tabelle 12 zeigt das Alter der Probanden zum Zeitpunkt des Indexurteils in den drei Bedingungen der GST-Gruppe sowie für die GST-Gesamtstichprobe, separat für Abbrecher und Vollender. Wilcoxon-Mann-Whitney-Tests ergaben in der Vergleichsgruppe signifikante Unterschiede im Alter zum ersten Messzeitpunkt zwischen den Probanden, welche die Studie bis zum Ende absolviert haben und denjenigen, die nach dem ersten Messzeitpunkt ausgeschieden sind (z=-2.17; p=.030): Letztere waren im Mittel älter. In den anderen beiden Bedingungen sowie in der Gesamt-Stichprobe GST ergaben sich keine Hinweise auf derartige Unterschiede.

Tabelle 12. Alter GST Abbrecher versus Vollender

GST Bedingung	Fehlende Werte	*N*	Alter Indexurteil	*p*
R&R alle zugeordneten Probanden	3.1% (*N*=4)	125	*M*=31.4 (*SA*=8.9)	
R&R Abbrecher	0.0% (*N*=0)	19	*M*=31.3 (*SA*=8.9)	.951 [a)]
R&R Vollender	3.6% (*N*=4)	106	*M*=31.5 (*SA*=8.9)	
TAU alle zugeordneten Probanden	3.9% (*N*=3)	75	*M*=32.7 (*SA*=11.2)	
TAU Abbrecher	0.0% (*N*=0)	21	*M*=37.5 (*SA*=11.6)	**.030** [a)]
TAU Vollender	5.3% (*N*=3)	54	*M*=30.8 (*SA*=10.5)	
KG alle zugeordneten Probanden	5.1% (*N*=4)	75	*M*=31.9 (*SA*=9.5)	
KG Abbrecher	0.0% (*N*=0)	28	*M*=30.6 (*SA*=9.5)	.381 [a)]
KG Vollender	7.8% (*N*=4)	47	*M*=32.6 (*SA*=9.6)	
GST alle Probanden	3.9% (*N*=11)	275	*M*=31.9 (*SA*=9.7)	
GST Abbrecher	0.0% (*N*=0)	68	*M*=32.9 (*SA*=10.3)	.434 [a)]
GST Vollender	5.1% (*N*=11)	207	*M*=31.6 (*SA*=9.5)	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; *M* = Mittelwert; *SA* = Standardabweichung. [a)] Ergebnis des Wilcoxon-Mann-Whitney-Tests. Dieser Test ist die non-parametrische Entsprechung zum T-Test für unabhängige Stichproben und wird angewandt, wenn wie im vorliegenden Fall die abhängige Variable nicht normalverteilt ist (starke Rechtsschiefe).

Bildungsniveau

Tabelle 13 zeigt das höchste erreichte Bildungsniveau der Probanden in den drei Bedingungen der GST-Gruppe sowie für die GST-Gesamtstichprobe, separat aufgeführt für Abbrecher und Vollender. Abbrecher und Vollender unterscheiden sich hinsichtlich ihres höchsten erreichten Bildungsniveaus in der Bedingung TAU (Exakter Test nach Fisher: p=.025) sowie in Bezug auf alle eingeschlossenen Probanden (Exakter Test nach Fisher: p=.035). Die Abbrecher hatten zu einem größeren Anteil mindestens eine Lehre erfolgreich abgeschlossen als Vollender; der Anteil an Probanden, die lediglich die Pflichtschule absolviert hatten, war im Gegenzug in der

Gruppe der Abbrecher niedriger (TAU: $\chi^2(1)=7.60$; $p=.006$; Gesamtgruppe: $\chi^2(1)=5.70$; $p=.017$).
Anmerkung: Die berichteten Unterschiede zwischen Abbrechern und Vollendern bleiben bestehen, wenn man die höchsten erreichten Bildungsabschlüsse nur bei Probanden mit Schweizer Staatsangehörigkeit vergleicht.

Tabelle 13. Bildungsniveau GST Abbrecher versus Vollender

			Höchster Bildungsabschluss				
GST Bedingung	**Fehlende Werte [% (*N*)]**	***N***	**< 7 Jahre Schule [% (*N*)]**	**Abgeschl. Schul-pflicht [% (*N*)]**	**Mind. Lehre [% (*N*)]**	**Nicht-Schweizer [% (*N*)]**	***p***
R&R alle zugeordneten Probanden	4.7% (*N*=6)	123	2.4% (*N*=3)	32.5% (*N*=40)	35.0% (*N*=43)	30.1% (*N*=37)	
R&R Abbrecher	0.0% (*N*=0)	19	0.0% (*N*=0)	15.8% (*N*=3)	47.4% (*N*=9)	36.8% (*N*=7)	.283 [a)]
R&R Vollender	5.5% (*N*=6)	104	2.9% (*N*=3)	35.6% (*N*=37)	32.7% (*N*=34)	28.9% (*N*=30)	
TAU alle zugeordneten Probanden	6.4% (*N*=5)	73	0.0% (*N*=0)	39.7% (*N*=29)	27.4% (*N*=20)	32.9% (*N*=24)	
TAU Abbrecher	4.8% (*N*=1)	20	0.0% (*N*=0)	20.0% (*N*=4)	50.0% (*N*=10)	30.0% (*N*=6)	**.025**[a)]
TAU Vollender	7.0% (*N*=4)	53	0.0% (*N*=0)	47.2% (*N*=25)	18.9% (*N*=10)	34.0% (*N*=18)	
KG alle zugeordneten Probanden	5.1% (*N*=4)	75	2.7% (*N*=2)	21.3% (*N*=16)	20.0% (*N*=15)	56.0% (*N*=42)	
KG Abbrecher	0.0% (*N*=0)	28	3.6% (*N*=1)	17.9% (*N*=5)	14.3% (*N*=4)	64.3% (*N*=18)	.640 [a)]
KG Vollender	7.8% (*N*=4)	47	2.1% (*N*=1)	23.4% (*N*=11)	23.4% (*N*=11)	51.1% (*N*=24)	
GST alle Probanden	5.2% (*N*=15)	271	1.9% (*N*=5)	31.4% (*N*=85)	28.8% (*N*=78)	38.0% (*N*=103)	
GST Abbrecher	1.5% (*N*=1)	67	1.5% (*N*=1)	17.9% (*N*=12)	34.3% (*N*=23)	46.3% (*N*=31)	**.035** [a)]
GST Vollender	6.4% (*N*=14)	204	2.0% (*N*=4)	35.8% (*N*=73)	27.0% (*N*=55)	35.3% (*N*=72)	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [a)] Ergebnis des exakten Tests nach Fisher. Dieser Test wird anstelle des Chi-Quadrat-Tests angewandt, wenn dessen Voraussetzung einer erwarteten Häufigkeit von ≥ 5 je Zelle verletzt ist, wie im vorliegenden Fall.

Psychiatrische Belastung: Unterschiede zwischen Abbrechern und Vollendern
Tabelle 14 zeigt den Anteil derjenigen Probanden, bei denen laut (aktuellstem) psychiatrischen Gutachten die diagnostischen Kriterien einer Persönlichkeitsstörung (Diagnose aus ICD-10-Kategorie F6), einer Störung der ICD-10-Kategorie F2 (Schizophrenie, schizotype und wahnhafte Störungen), einer affektiven Störung (ICD-10 Kategorie F3) oder einer Störung durch psychotrope Substanzen (ICD-10 Kategorie F1) erfüllt sind. Da zu einem substanziellen Anteil der Probanden keine Angaben zur psychiatrischen Belastung vorliegen, ist eine inferenzstatistische Auswertung lediglich teilweise möglich. Vergleiche zwischen Abbrechern und Vollendern in den drei Bedingungen der GST-Gruppe ergaben, sofern sie durchgeführt werden konnten, keine Hinweise auf signifikante Unterschiede (siehe Tabelle 14).

Tabelle 14. Psychiatrische Vorbelastung GST Abbrecher versus Vollender

	Persönlichkeitsstörung (ICD-10 F6)				Schizophrenie (ICD-10 F2)				Affektive Störungen (ICD-10 F3)				Substanzstörung (ICD-10 F1)			
GST Bedingung	**Fehlende Werte [% (*N*)]**	***N***	**Anteil [% (*N*)]**	***p***	**Fehlende Werte [% (*N*)]**	***N***	**Anteil [% (*N*)]**	***p***	**Fehlende Werte [% (*N*)]**	***N***	**Anteil [% (*N*)]**	***p***	**Fehlende Werte [% (*N*)]**	***N***	**Anteil [% (*N*)]**	***p***
R&R alle zugeordneten Probanden	11.6% (*N*=15)	114	57.0% (*N*=65)		12.4% (*N*=16)	113	6.2% (*N*=7)		13.2% (*N*=17)	112	8.9% (*N*=10)		10.1% (*N*=13)	116	56.9% (*N*=66)	
R&R Abbrecher	5.3% (*N*=1)	18	61.1% (*N*=11)	.702 a)	5.3% (*N*=1)	18	5.6% (*N*=1)	.691 b)	5.3% (*N*=1)	18	16.7% (*N*=3)	.201 b)	0.0% (*N*=0)	19	63.2% (*N*=12)	.547 a)
R&R Vollender	12.7% (*N*=14)	96	56.3% (*N*=54)		13.6% (*N*=15)	95	6.3% (*N*=6)		14.5% (*N*=16)	94	7.4% (*N*=7)		11.8% (*N*=13)	97	55.7% (*N*=54)	
TAU alle zugeordneten Probanden	21.8% (*N*=17)	61	45.9% (*N*=28)		19.2% (*N*=15)	63	12.7% (*N*=8)		20.5% (*N*=16)	62	4.8% (*N*=3)		18.0% (*N*=14)	64	43.8% (*N*=28)	
TAU Abbrecher	14.3% (*N*=3)	18	39.9% (*N*=7)	N/A	14.3% (*N*=3)	18	16.7% (*N*=3)	.412 b)	14.3% (*N*=3)	18	5.6% (*N*=1)	N/A	14.3% (*N*=3)	18	44.4% (*N*=8)	.944 a)
TAU Vollender	24.6% (*N*=14)	43	48.8% (*N*=21)		21.1% (*N*=12)	45	11.1% (*N*=5)		22.8% (*N*=13)	44	4.5% (*N*=2)		19.3% (*N*=11)	46	43.5% (*N*=20)	
KG alle zugeordneten Probanden	46.8% (*N*=37)	42	40.5% (*N*=17)		48.1% (*N*=38)	41	4.9% (*N*=2)		48.1% (*N*=38)	41	7.3% (*N*=3)		45.6% (*N*=36)	43	41.9% (*N*=18)	
KG Abbrecher	50.0% (*N*=14)	14	28.6% (*N*=4)	N/A	50.0% (*N*=14)	14	14.3% (*N*=2)	N/A	50.0% (*N*=14)	14	7.1% (*N*=1)	N/A	50.0% (*N*=14)	14	50.0% (*N*=7)	N/A
KG Vollender	45.1% (*N*=23)	28	46.4% (*N*=13)		47.1% (*N*=24)	27	0.0% (*N*=0)		47.1% (*N*=24)	27	7.4% (*N*=2)		43.1% (*N*=22)	29	37.9% (*N*=11)	
GST alle Probanden	24.1% (*N*=69)	217	50.7% (*N*=110)		24.1% (*N*=69)	217	7.8% (*N*=17)		24.3% (*N*=53)	215	7.4% (*N*=16)		22.0% (*N*=63)	223	50.2% (*N*=112)	
GST Abbrecher	26.5% (*N*=18)	50	44.0% (*N*=22)	N/A	26.5% (*N*=18)	50	12.0% (*N*=6)	N/A	26.5% (*N*=18)	50	10.0% (*N*=5)	N/A	25.0% (*N*=17)	51	52.9% (*N*=27)	N/A
GST Vollender	23.4% (*N*=51)	167	52.7% (*N*=88)		23.4% (*N*=51)	167	6.6% (*N*=11)		24.8% (*N*=71)	165	6.7% (*N*=11)		21.1% (*N*=46)	172	49.4% (*N*=85)	

Anmerkungen. [a)] Ergebnis des Chi-Quadrat-Tests; [b)] Ergebnis des exakten Tests nach Fisher; Rot = mehr als 20% fehlende Werte.

Therapieerfahrung: Unterschiede zwischen Abbrechern und Vollendern

Tabelle 15 zeigt den Anteil der Probanden aus der GST-Gruppe, die bereits vor ihrer Inhaftierung mindestens einmal in psychotherapeutischer Behandlung waren, separat dargestellt für Abbrecher und Vollender. Chi-Quadrat-Tests zeigten weder innerhalb der Experimental- noch der Vergleichsgruppe Unterschiede zwischen Probanden, die den MV vorzeitig abbrachen gegenüber Probanden, die bis zum Ende an der Studie teilnahmen. Für die Kontrollgruppe und die R&R-Gesamtgruppe kann diesbezüglich aufgrund des hohen Anteils fehlender Werte keine Aussage getroffen werden.

Tabelle 15. Therapieerfahrung GST Abbrecher versus Vollender

GST Bedingung	Fehlende Werte [% (*N*)]	*N*	Psychotherapie-Erfahrung	*p*
R&R alle zugeordneten Probanden	14.7% (*N*=19)	110	56.4% (*N*=62)	
R&R Abbrecher	5.3% (*N*=1)	18	38.9% (*N*=7)	.102 [a)]
R&R Vollender	16.4% (*N*=18)	92	59.8% (*N*=55)	
TAU alle zugeordneten Probanden	16.7% (*N*=13)	65	38.5% (*N*=25)	
TAU Abbrecher	19.0% (*N*=4)	17	52.9% (*N*=9)	.153 [a)]
TAU Vollender	15.8% (*N*=9)	48	33.3% (*N*=16)	
KG alle zugeordneten Probanden	36.7% (*N*=29)	50	40.0% (*N*=20)	
KG Abbrecher	35.7% (*N*=10)	18	33.3% (*N*=6)	N/A
KG Vollender	37.3% (*N*=19)	32	43.8% (*N*=14)	
GST alle Probanden	21.3% (*N*=61)	225	47.6% (*N*=107)	
GST Abbrecher	22.1% (*N*=15)	53	41.5% (*N*=22)	N/A
GST Vollender	21.1% (*N*=46)	172	49.4% (*N*=85)	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [a)] Ergebnis des Chi-Quadrat-Tests; Rot = mehr als 20% fehlende Werte.

Fragestellung 1: Veränderungen in den Messwerten der Fragebögen in Abhängigkeit der Therapie

Im Folgenden wird die Fragestellung 1 der Evaluation des R&R2-Behandlungsprogramms für Gewaltstraftäter beantwortet. Dazu werden die Differenzwerte zwischen der Prä- und der Postmessung in den Messwerten verschiedener Fragebögen inferenzstatistisch zwischen den drei Bedingungen der GST-Gruppe verglichen. Definitionsgemäß kann für die Beantwortung der Fragestellung 1 ausschließlich auf die Vollender des MV zurückgegriffen werden. Ein Vergleich zwischen den Vollendern und den Abbrechern in wichtigen Merkmalen befindet sich in Anhang 1.

Kurzfragebogen zur Erfassung von Aggressivitätsfaktoren (K-FAF)

K-FAF: Selbstbeurteilung

Abbildung 4 zeigt die Differenzwerte (T2 – T1) im Aggressivitäts-Summenwert des Kurzfragebogens zur Erfassung von Aggressivitätsfaktoren (K-FAF) in der GST-Gruppe gemäß Selbstauskunft. In Tabelle 16 werden zusätzlich die Differenzwerte der einzelnen Skalen des K-FAF unter Angabe von Stichprobengröße, Standardabweichung und statistischer Signifikanz aufgeführt. Fehlende Werte wurden durch den Median der jeweiligen Bedingung ersetzt, sofern der Anteil fehlender Angaben je Skala und Fall bei maximal 20% lag. Dieser Wert wurde festgelegt, da das Manual des K-FAF keine Empfehlung zum Umgang mit fehlenden Werten beinhaltet, der Grenzwert von 20% nach Ansicht des Evaluationsteams ein angemessenes Verhältnis zwischen Datenverlust und Genauigkeit der Aussagen widerspiegelt, er durchgehend für die vorliegende Evaluation angewandt wurde und gängige Praxis darstellt (z.B. Acuna & Rodriguez, 2004; Finch, 2016).
Als Methode der Berechnung wurde eine einfaktorielle Varianzanalyse (ANOVA) gewählt. Diese wurde über die Differenzwerte der jeweiligen Skalen des K-FAF sowie auch über den Summenwert gerechnet. Die Berechnungen wurden explizit unter Einschluss von Ausreißer-Werten durchgeführt, um keine wertvolle Information zu verlieren und um den Besonderheiten der forensischen Klientel gerecht zu werden. (Ein Grubbs-Test ergab das Vorliegen von Ausreißer-Werten: Bezogen auf den Summenwert handelte es sich bei fünf Werten um Ausreißer nach Grubbs. In den einzelnen Skalen wurden zwischen null (Skala „Aggressionshemmung“) und fünf Ausreißer (Skala „spontane Aggressivität“) identifiziert).
In den Differenzwerten des Aggressivitäts-Summenwertes, gemäß Manual des K-FAF die Summe der Skalen „spontane Aggressivität“, „reaktive Aggressivität“ und „Erregbarkeit“, zeigten sich signifikante Unterschiede zwischen den drei Bedingungen der GST-Gruppe: $F(2, 206) = 3.52$, $p=.031$; $\eta^2 = .033$. Ein post hoc durchgeführter Tukey-Kramer-Test wird auf einem Alpha-Niveau von .05 nicht signifikant; der Einzelvergleich, der dem Alpha-Niveau am nächsten liegt und als Trend zu interpretieren ist, ist derjenige zwischen Experimental- und Kontrollgruppe ($p=.064$).

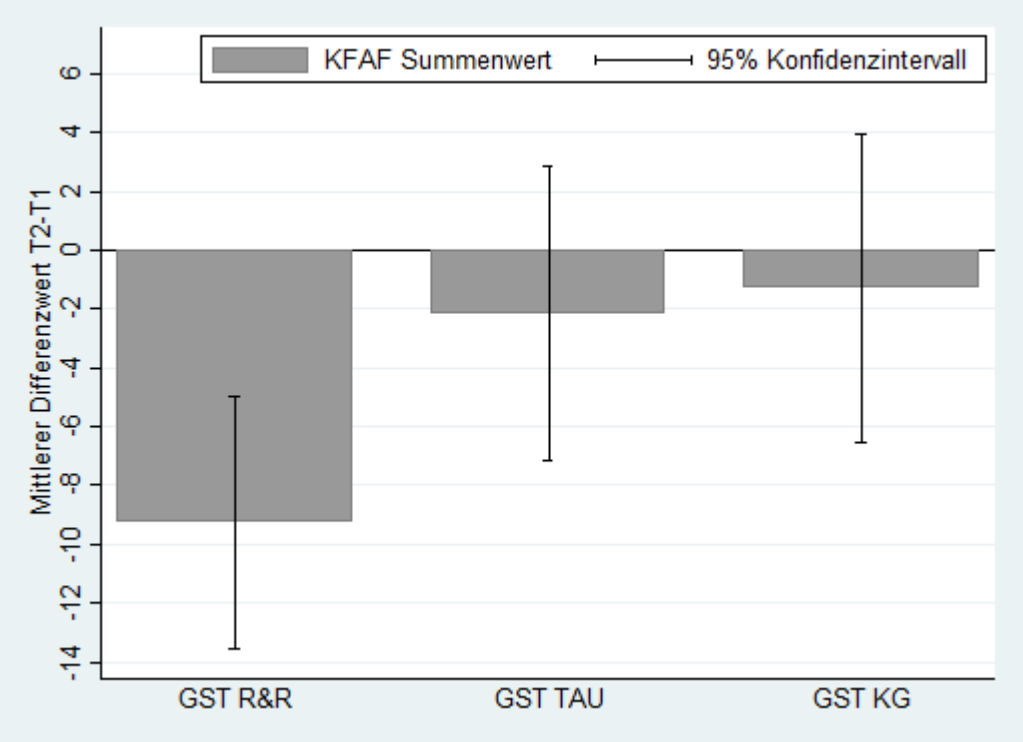

Abbildung 4. Mittlere Differenzwerte des K-FAF-Summenwerts

Tabelle 16. Differenzwerte im K-FAF (Selbstbeurteilung)

GST Bedingung	Fehlend[11] [% (*N*)]	*N*	*M*	*SA*	*p*
Skala „Spontane Aggressivität“					
GST R&R	1.8% (*N*=2)	108	-1.5	8.2	
GST TAU	5.3% (*N*=3)	54	1.1	6.8	.100 a)
GST KG	3.9% (*N*=2)	49	0.5	8.0	
Skala „Reaktive Aggressivität“					
GST R&R	2.7% (*N*=3)	107	-3.9	9.0	
GST TAU	5.3% (*N*=3)	54	-1.7	7.0	.075 a)
GST KG	2.0% (*N*=1)	50	-1.1	7.1	
Skala „Erregbarkeit“					
GST R&R	3.6% (*N*=4)	106	-3.7	8.4	
GST TAU	3.5% (*N*=2)	55	-1.5	8.6	.079 a)
GST KG	3.9% (*N*=2)	49	-0.7	7.9	

[11] „Fehlend“ beschreibt in den folgenden Tabellen denjenigen Anteil an Probanden, bei denen es sich nicht um Abbrecher handelt und für die dennoch keine Werte vorliegen. Sofern nicht anders berichtet, wurden fehlende Werte durch den Median ersetzt, wenn je Fall für einen substanziellen Anteil der Items eines Fragebogens Werte vorlagen. Das genaue Vorgehen ist separat für jeden Fragebogen dargelegt.

Skala „Selbstaggressivität“					
GST R&R	5.5% (*N*=6)	104	-2.3	6.8	
GST TAU	5.3% (*N*=3)	54	-1.8	4.5	.451 [b)]
GST KG	5.9% (*N*=3)	48	-1.0	5.6	
*Skala „Aggressionshemmung“**					
GST R&R	3.6% (*N*=4)	106	1.3	6.4	
GST TAU	5.3% (*N*=3)	54	-0.0	4.9	.206 [a)]
GST KG	3.9% (*N*=2)	49	-0.3	5.7	
Summenwert Aggressivität (= Summe der ersten drei Skalen des K-FAF)					
GST R&R	3.6% (*N*=4)	106	-9.2	22.4	
GST TAU	5.3% (*N*=3)	54	-2.2	18.3	**.031** [a)]
GST KG	3.9% (*N*=2)	49	-1.3	18.2	

Anmerkungen. Angegeben sind die Differenzwerte T2 – T1: Negative Werte bedeuten eine relative Abnahme in der Ausprägung der Items einer Skala; * Inhaltlich inverse Skala; a) Ergebnis einer einfaktoriellen Varianzanalyse; b) Ergebnis des Welch-Tests, der durchgeführt wird, wenn die Annahme der Homogenität der Varianzen der ANOVA wie im vorliegenden Fall verletzt ist.

K-FAF: Fremdbeurteilung (reduziert: 2 Items je Skala)

Für eine substanzielle Anzahl von Fällen liegt kein Fremdrating des K-FAF vor. Zum ersten Messzeitpunkt betrifft dies 27 Fälle der Experimentalgruppe (20.2%), 19 Fälle der Vergleichsgruppe (31.2%) und 30 Fälle der Kontrollgruppe (47.6%), zum zweiten Messzeitpunkt 21 Fälle der Experimentalgruppe (15.7%), 21 Fälle der Vergleichsgruppe (34.4%) und 31 Fälle der Kontrollgruppe (49.2%). Damit sind keine reliablen Aussagen zur Fremdbeurteilung des K-FAF möglich.

Auf Wunsch des Auftraggebers werden im Folgenden die Differenzwerte zwischen den beiden Messzeitpunkten separat nach den drei Bedingungen der GST-Gruppe aufgeführt, um zumindest ein deskriptives Ergebnis zu erhalten. Bei den Werten in Tabelle 17 handelt es sich um die Rohwerte; es wurde in diesem Falle aufgrund der substanziellen Anzahl an fehlenden Werten keine Imputation und keine statistische Analyse durchgeführt.

Tabelle 17. Differenzwerte im reduzierten KFAF (Fremdbeurteilung)

GST Bedingung	**Fehlend [% (*N*)]**	***N***	***M***	***SA***	***Median***
K-FAF Fremdbeurteilung „Spontane Aggressivität“ (reduzierte Itemanzahl)					
GST R&R	26.4% (*N*=29)	81	-0.2	3.7	0
GST TAU	38.6% (*N*=22)	35	0.2	1.4	0
GST KG	68.6% (*N*=35)	16	-1.5	2.4	-1
Skala „Reaktive Aggressivität“					

GST R&R	27.3% (*N*=30)	80	0.2	2.0	0
GST TAU	36.8% (*N*=21)	36	-0.1	1.6	0
GST KG	74.5% (*N*=38)	13	0.0	3.3	-1
K-FAF Fremdbeurteilung Skala „Erregbarkeit“ (reduzierte Itemanzahl)					
GST R&R	25.5% (*N*=28)	82	-0.2	2.3	0
GST TAU	36.8% (*N*=21)	36	-0.2	1.5	0
GST KG	66.7% (*N*=34)	17	-1.1	2.5	-1
K-FAF Fremdbeurteilung Skala „Selbstaggressivität“ (reduzierte Itemanzahl)					
GST R&R	25.5% (*N*=28)	82	0.1	2.2	0
GST TAU	35.1% (*N*=20)	37	-0.1	1.6	0
GST KG	68.6% (*N*=35)	16	1.8	6.0	0.5
*K-FAF Fremdbeurteilung Skala „Aggressionshemmung“ (reduzierte Itemanzahl)**					
GST R&R	28.2% (*N*=31)	79	0.1	1.9	0
GST TAU	35.1% (*N*=20)	37	-0.3	1.8	0
GST KG	72.6% (*N*=37)	14	0.1	2.5	0
K-FAF Fremdbeurteilung Summenwert Aggressivität = Summe der ersten drei Skalen (reduzierte Itemanzahl)					
GST R&R	27.3% (*N*=30)	80	-0.3	5.7	0
GST TAU	42.1% (*N*=24)	33	-0.2	3.4	0
GST KG	74.5% (*N*=38)	13	-2.6	5.7	-4

Anmerkungen. Angegeben sind die Differenzwerte T2 – T1: Negative Werte bedeuten eine relative Abnahme in der Ausprägung der Items einer Skala; * Inhaltlich inverse Skala.

Inventar zur Erfassung interpersoneller Probleme (IIP-D)

IIP-D: Selbstbeurteilung

Abbildung 5 zeigt die mittleren Differenzwerte im Gesamtwert des IIP-D in den drei Bedingungen der GST-Gruppe. In Tabelle 18 werden zusätzlich die Differenzwerte der einzelnen Skalen des IIP-D unter Angabe von Stichprobengröße, Standardabweichung und Signifikanzniveau aufgeführt. Fehlende Werte wurden durch den jeweiligen Median der Gruppe ersetzt, sofern je Fall maximal 5% der Werte nicht vorlagen. Ab diesem Grenzwert sollte laut Manual des IIP von einer Auswertung der Ergebnisse abgesehen werden.

Sowohl über die Differenzwerte der einzelnen Skalen-Mittelwerte als auch über den Differenzwert des IIP-Gesamtwertes wurde eine einfaktorielle Varianzanalyse gerechnet. Die Berechnungen wurden explizit unter Einschluss von Ausreißer-Werten durchgeführt, um keine wertvolle Information zu verlieren und um den Besonderheiten der forensischen Klientel gerecht zu werden. (Ein Grubbs-Test ergab das Vorliegen von Ausreißer-Werten: Bezogen auf die Differenzwerte des IIP-Gesamtwerts

handelte es sich bei fünf Werten um Ausreißer nach Grubbs. In den einzelnen Skalen wurden zwischen 0 (Skala JK) und 4 Ausreißer (Skala FG) identifiziert.)
Die Differenzwerte der einzelnen Skalen und des IIP-Gesamtwertes sind näherungsweise normalverteilt und wurden auf Homogenität der Varianzen überprüft (= Voraussetzungen einer ANOVA).
Varianzanalysen (ANOVAs) ergaben keine Hinweise auf Unterschiede zwischen den Probanden in den drei Bedingungen der GST-Gruppe, weder im Differenzwert der einzelnen Skalen des IIP-D noch im Gesamtwert.

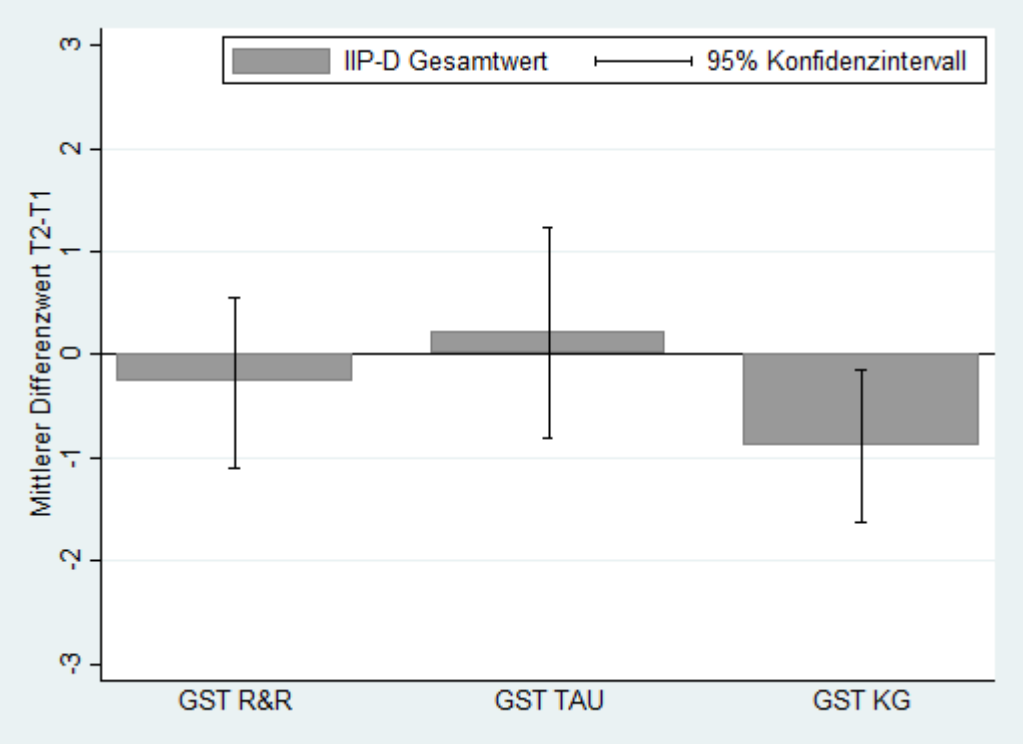

Abbildung 5. Mittlere Differenzwerte im Gesamtwert des IIP-D

Tabelle 18. Differenzwerte im IIP-D (Selbstbeurteilung)

GST Bedingung	Fehlend [% (*N*)]	*N*	*M*	*SA*	*p*
Differenzwerte Skala PA „Zu autokratisch/ dominant“					
GST R&R	2.7% (*N*=3)	107	-0.4	5.7	
GST TAU	3.5% (*N*=2)	55	0.4	4.4	.628 [a)]
GST KG	3.9% (*N*=2)	49	-0.6	4.1	
Differenzwerte Skala BC „Zu streitsüchtig/ konkurrierend“					
GST R&R	4.6% (*N*=5)	105	-0.9	5.0	
GST TAU	3.5% (*N*=2)	55	-0.6	5.1	.935 [a)]
GST KG	7.8% (*N*=4)	47	-1.0	4.2	

Differenzwerte Skala DE „Zu abweisend/ kalt“					
GST R&R	4.6% (*N*=5)	105	-0.1	6.2	
GST TAU	3.5% (*N*=2)	55	-0.1	6.4	.554 [a)]
GST KG	7.8% (*N*=4)	47	-1.2	5.0	
Differenzwerte Skala FG „Zu introvertiert/ sozial vermeidend“					
GST R&R	4.6% (*N*=5)	105	0.0	5.8	
GST TAU	3.5% (*N*=2)	55	1.1	6.2	.116 [a)]
GST KG	5.9% (*N*=3)	48	-1.1	3.8	
Differenzwerte Skala HI „Zu selbstunsicher/ unterwürfig“					
GST R&R	4.6% (*N*=5)	105	0.3	6.1	
GST TAU	5.3% (*N*=2)	54	0.3	6.4	.394 [a)]
GST KG	7.8% (*N*=4)	47	-1.1	5.1	
Differenzwerte Skala JK „Zu ausnutzbar/ nachgiebig“					
GST R&R	4.6% (*N*=5)	105	0.1	4.7	
GST TAU	5.3% (*N*=2)	54	0.4	5.1	.437 [a)]
GST KG	7.8% (*N*=4)	47	-0.7	2.9	
Differenzwerte Skala LM „Zu Fürsorglich/ freundlich“					
GST R&R	5.5% (*N*=6)	104	-0.8	5.4	
GST TAU	5.3% (*N*=3)	54	0.1	4.4	.368 [a)]
GST KG	5.9% (*N*=3)	48	-1.1	4.0	
Differenzwerte Skala NO „Zu expressiv/ aufdringlich“					
GST R&R	2.7% (*N*=3)	107	-0.4	5.3	
GST TAU	5.3% (*N*=3)	54	-0.2	4.2	.653 [a)]
GST KG	3.9% (*N*=2)	49	-1.0	3.7	
Differenzwerte IIP Gesamtwert (= Summe der acht Skalen geteilt durch acht)					
GST R&R	5.5% (*N*=6)	104	-0.3	4.2	
GST TAU	5.3% (*N*=3)	54	0.2	3.7	.356 [a)]
GST KG	9.8% (*N*=5)	46	-0.9	2.5	

Anmerkungen. Angegeben sind die Differenzwerte T2 – T1: Negative Werte bedeuten eine relative Abnahme in der Ausprägung der Items einer Skala; [a)] Ergebnis einer einfaktoriellen Varianzanalyse (ANOVA) über die Differenzwerte. Eine ANOVA dient der Überprüfung von Mittelwertsunterschieden einer intervallskalierten abhängigen Variablen in Abhängigkeit von einer kategorialen unabhängigen Variablen (mit zwei oder mehr Kategorien). Die einzelnen Differenzwerte der Skalen und des IIP-Gesamtwertes sind näherungsweise normalverteilt und die Varianzen sind hinreichend homogen (= Voraussetzungen einer ANOVA).

IIP-D: Fremdbeurteilung (reduziert: 1 Item je Skala)

Die Fremdbeurteilung des Inventars zur Erfassung interpersoneller Probleme besteht aus einem einzigen Item je Skala. Diese Items haben einen substanziellen Anteil an fehlenden Werten. Auf Wunsch des Auftraggebers werden in Tabelle 19 dennoch die mittleren Differenzwerte zwischen den beiden Messzeitpunkten in Abhängigkeit der drei Bedingungen der GST-Gruppe aufgeführt, um so einen deskriptiven Eindruck zu erlangen.

Tabelle 19. Differenzwerte im reduzierten IIP-D (Fremdbeurteilung)

GST Bedingung	Fehlend [% (*N*)]	*N*	*M*	*SA*	*Median*
Differenzwerte Skala PA (reduziert: 1 Item) „Zu autokratisch/ dominant“					
GST R&R	25.5% (*N*=28)	82	-0.0	0.8	0
GST TAU	35.1% (*N*=20)	37	-0.2	0.6	0
GST KG	64.7% (*N*=33)	18	-0.1	1.5	0
Differenzwerte Skala BC (reduziert: 1 Item) „Zu streitsüchtig/ konkurrierend“					
GST R&R	25.5% (*N*=28)	82	-0.1	0.9	0
GST TAU	35.1% (*N*=20)	37	0.0	0.7	0
GST KG	64.7% (*N*=33)	18	0.7	1.3	0
Differenzwerte Skala DE (reduziert: 1 Item) „Zu abweisend/ kalt“					
GST R&R	25.5% (*N*=28)	82	0.0	1.0	0
GST TAU	35.1% (*N*=20)	37	0.1	0.8	0
GST KG	62.8% (*N*=32)	19	-0.4	1.3	-1
Differenzwerte Skala FG (reduziert: 1 Item) „Zu introvertiert/ sozial vermeidend“					
GST R&R	25.5% (*N*=28)	82	0.2	1.0	0
GST TAU	35.1% (*N*=20)	37	0.1	0.7	0
GST KG	62.8% (*N*=32)	19	0.2	1.3	0
Differenzwerte Skala HI (reduziert: 1 Item) „Zu selbstunsicher/ unterwürfig“					
GST R&R	25.5% (*N*=28)	82	0.0	0.9	0
GST TAU	35.1% (*N*=20)	37	-0.2	0.8	0
GST KG	62.8% (*N*=32)	19	0.0	1.1	0
Differenzwerte Skala JK (reduziert: 1 Item) „Zu ausnutzbar/ nachgiebig“					
GST R&R	25.5% (*N*=28)	82	-0.0	0.8	0
GST TAU	35.1% (*N*=20)	37	0.1	0.8	0
GST KG	64.7% (*N*=33)	18	0.3	0.9	0
Differenzwerte Skala LM (reduziert: 1 Item) „Zu Fürsorglich/ freundlich“					
GST R&R	25.5% (*N*=28)	82	-0.1	0.8	0

GST TAU	35.1% (*N*=20)	37	-0.1	0.9	0
GST KG	64.7% (*N*=33)	18	0.2	1.2	0
Differenzwerte Skala NO (reduziert: 1 Item) „Zu expressiv/ aufdringlich“					
GST R&R	25.5% (*N*=28)	82	0.2	0.9	0
GST TAU	35.1% (*N*=20)	37	0.1	0.6	0
GST KG	62.8% (*N*=32)	19	0.1	1.1	0
Differenzwerte IIP Gesamtwert (= Summe der acht reduzierten Skalen geteilt durch acht)					
GST R&R	25.5% (*N*=28)	82	0.0	0.3	0
GST TAU	35.1% (*N*=20)	37	-0.0	0.3	0
GST KG	66.7% (*N*=34)	17	-0.1	0.4	0

Anmerkungen. Angegeben sind die Differenzwerte T2 – T1: Negative Werte bedeuten eine relative Abnahme in der Ausprägung der Items einer Skala.

Hostile Attribution Bias (HAB)

Die folgenden drei Tabellen zeigen die Messwerte im Fragebogen zum Hostile Attribution Bias (HAB) in der GST-Gruppe: Tabelle 20 beinhaltet die Items des HAB zu provozierenden Situationen, Tabelle 21 diejenigen zu Situationen mit unklarer Intention des Gegenübers und Tabelle 22 diejenigen zu eindeutig nicht provozierenden Situationen. Angegeben ist jeweils die Differenz des Summenwerts einer Skala bzw. des Gesamt-Summenwerts (T2 – T1). Fehlende Werte wurden durch den Median der jeweiligen Bedingung ersetzt, sofern der Anteil fehlender Angaben je Fall bei maximal 20% lag. Ausreißer wurden explizit mit in die Berechnung eingeschlossen. Für eindeutig provozierende Situationen ergab ein Grubbs-Test bezogen auf die einzelnen Skalen zwischen 1 und 2 Ausreißer-Werte, hinsichtlich des Summenwertes keinen Ausreißer. Für die Vignetten mit unklarer Absicht ergab ein Grubbs-Test zwischen 0 und 3 Ausreißerwerte bezogen auf die einzelnen Skalen und 1 Ausreißerwert bezüglich des Summenwertes. Für die eindeutig nicht provozierenden Vignetten handelt es sich gemäß Grubbs-Test bezogen auf die einzelnen Skalen bei 0 bis 5 Werten und bezogen auf den Summenwert bei 3 Werten um Ausreißer. Die Werte wurden auf annähernde Normalverteilung und Varianzhomogenität überprüft. War letztere nicht gegeben, wurde anstatt einer ANOVA ein Welch-Test als robustes Testverfahren zur Prüfung auf Gleichheit der Mittelwerte angewandt.

HAB: Selbstbeurteilung

Abbildung 6 zeigt die mittleren Differenzwerte im Summenwert des HAB für eindeutig provozierende Situationen in den drei Bedingungen der GST-Gruppe. In Tabelle 20 sind zusätzlich die Differenzwerte in den einzelnen Skalen des HAB unter Angabe von Stichprobengröße, Standardabweichung und Signifikanzniveau aufgeführt. Signifikante Gruppenunterschiede zeigten sich in der berichteten Wahrscheinlichkeit, die Situation als absichtliche Provokation wahrzunehmen: $F(2, 213) = 3.77$, $p=.025$; $\eta^2 = .034$. Ein post hoc durchgeführter Tukey-Kramer-Test zeigte, dass sich Experimental- und Kontrollgruppe auf einem Alpha-Niveau von .05 signifikant unterschei-

den: Während die Probanden der Experimentalgruppe zum zweiten Messzeitpunkt eine geringere Wahrscheinlichkeit berichten, eine provozierenden Situation als solche wahrzunehmen, geben die Probanden der Kontrollgruppe eine höhere Wahrscheinlichkeit an.
Ebenso zeigten sich Unterschiede in der Wahrscheinlichkeit, in einer provozierenden Situation verbal aggressiv zu reagieren: $F(2, 213) = 5.64$, $p=.004$; $\eta^2 = .050$. Ein post hoc durchgeführter Tukey-Kramer-Test zeigte, dass sich Experimental- und Kontrollgruppe auf einem Alpha-Niveau von .05 signifikant unterscheiden: Während die Probanden der Experimentalgruppe zum zweiten Messzeitpunkt eine geringere Wahrscheinlichkeit berichten, in einer provozierenden Situation mit Anschreien oder Beschimpfen zu reagieren, geben die Probanden der Kontrollgruppe eine höhere Wahrscheinlichkeit an.
Auch im Summenwert der Wahrscheinlichkeit für feindseliges Verhalten zeigten sich signifikante Unterschiede: $F(2, 213) = 4.03$, $p=.019$; $\eta^2 = .036$. Ein post hoc durchgeführter Tukey-Kramer-Test zeigte, dass sich Experimental- und Kontrollgruppe auf einem Alpha-Niveau von .05 signifikant unterscheiden: Während die Probanden der Experimentalgruppe zum zweiten Messzeitpunkt im Mittel eine Abnahme in der Ausprägung der HAB-Items für eindeutig provozierende Situationen zeigen, geben die Probanden der Kontrollgruppe eine geringfügig ausgeprägte Zunahme an.
Als Trend zeigten sich Gruppenunterschiede in der berichteten Wahrscheinlichkeit für unhöfliches Verhalten in einer provozierenden Situation: $F(2, 213) = 2.56$, $p=.080$; $\eta^2 = .023$. Ein post hoc durchgeführter Tukey-Kramer-Test zeigte auf einem Alpha-Niveau von .08 Unterschiede zwischen der Experimental- und der Vergleichsgruppe. Während die Wahrscheinlichkeit für unhöfliches Verhalten in der Experimentalgruppe zum zweiten Messzeitpunkt geringer ausfiel, blieb sie in der Vergleichsgruppe unverändert.

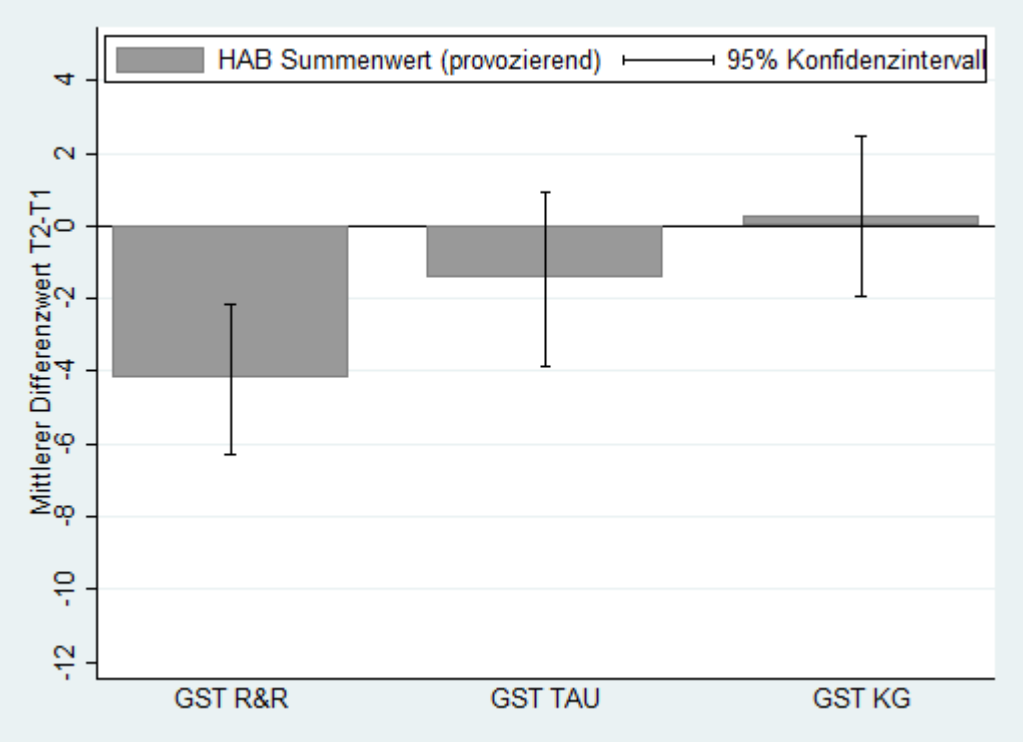

Abbildung 6. Mittlere Differenzwerte im Summenwert des HAB (provozierende Situationen)

Tabelle 20. Differenzwerte im HAB: Situationen mit klarer Provokation

GST Bedingung	Fehlend [% (*N*)]	*N*	*M*	*SA*	*p*
P (Wahrscheinlichkeit für Wahrnehmung als absichtliche Provokation), provozierende Absicht					
GST R&R	1.8% (*N*=2)	108	-0.5	2.0	
GST TAU	0% (*N*=0)	57	-0.3	1.6	**.025** [a)]
GST KG	0% (*N*=0)	51	0.4	2.1	
V (Wahrscheinlichkeit für Verärgerung), provozierende Absicht					
GST R&R	1.8% (*N*=2)	108	-0.5	2.1	
GST TAU	0% (*N*=0)	57	-0.2	1.5	.180 [a)]
GST KG	0% (*N*=0)	51	0.1	1.7	
M (Wahrscheinlichkeit für das Mitteilen der Verärgerung), provozierende Absicht					
GST R&R	1.8% (*N*=2)	108	-0.4	2.3	
GST TAU	0% (*N*=0)	57	0.1	1.6	.249 [b)]
GST KG	0% (*N*=0)	51	-0.0	1.7	
U (Wahrscheinlichkeit für unhöfliches Verhalten), provozierende Absicht					
GST R&R	1.8% (*N*=2)	108	-0.7	2.2	.080 [a)]

GST TAU	0% (*N*=0)	57	0.0	1.9	
GST KG	0% (*N*=0)	51	-0.2	1.9	
AB (Wahrscheinlichkeit für Anschreien/ Beschimpfen), provozierende Absicht					
GST R&R	1.8% (*N*=2)	108	-0.8	2.1	
GST TAU	0% (*N*=0)	57	-0.3	1.9	**.004** [a)]
GST KG	0% (*N*=0)	51	0.5	2.9	
D (Wahrscheinlichkeit für Drohen), provozierende Absicht					
GST R&R	1.8% (*N*=2)	108	-0.7	2.2	
GST TAU	0% (*N*=0)	57	-0.4	1.9	.294 [a)]
GST KG	0% (*N*=0)	51	-0.2	1.6	
G (Wahrscheinlichkeit für physische Gewalt), provozierende Absicht					
GST R&R	1.8% (*N*=2)	108	-0.5	2.2	
GST TAU	0% (*N*=0)	57	-0.5	2.1	.574 [a)]
GST KG	0% (*N*=0)	51	-0.2	1.6	
HAB Summenwert, provozierende Absicht					
GST R&R	1.8% (*N*=2)	108	-4.2	10.9	
GST TAU	0% (*N*=0)	57	-1.5	9.1	**.019** [a)]
GST KG	0% (*N*=0)	51	0.3	7.8	

Anmerkungen. Angegeben sind die Differenzwerte T2 – T1: Negative Werte bedeuten eine relative Abnahme in der Ausprägung der Items einer Skala; a) Ergebnis einer einfaktoriellen Varianzanalyse.

Abbildung 7 zeigt die mittleren Differenzwerte im Summenwert des HAB für Situationen mit unklarer Absicht des Gegenübers in den drei Bedingungen der GST-Gruppe. In Tabelle 21 sind zusätzlich die Differenzwerte in den einzelnen Skalen des HAB unter Angabe von Stichprobengröße, Standardabweichung und Signifikanzniveau aufgeführt.

Gruppenunterschiede zeigten sich in Form eines Trends in der berichteten Wahrscheinlichkeit für unhöfliches Verhalten in einer provozierenden Situation: *Welch's* $F(2, 124.25) = 2.59$, $p=.079$; $\eta^2 = .023$. Ein post hoc durchgeführter Tukey-Kramer-Test wurde auch auf einem Alpha-Niveau von .10 nicht signifikant.

Ebenfalls als Trend zeigten sich Unterschiede in der Wahrscheinlichkeit, in einer provozierenden Situation verbal aggressiv zu reagieren: $F(2, 212) = 2.75$, $p=.066$; $\eta^2 = .025$. Ein post hoc durchgeführter Tukey-Kramer-Test zeigte, dass sich Experimental- und Kontrollgruppe, ebenfalls als Trend, auf einem Alpha-Niveau von .06 unterscheiden: Während die Probanden der Experimentalgruppe zum zweiten Messzeitpunkt eine geringere Wahrscheinlichkeit berichten, in einer unklaren Situation mit Anschrei-

en oder Beschimpfen zu reagieren, geben die Probanden der Kontrollgruppe eine erhöhte Wahrscheinlichkeit an.

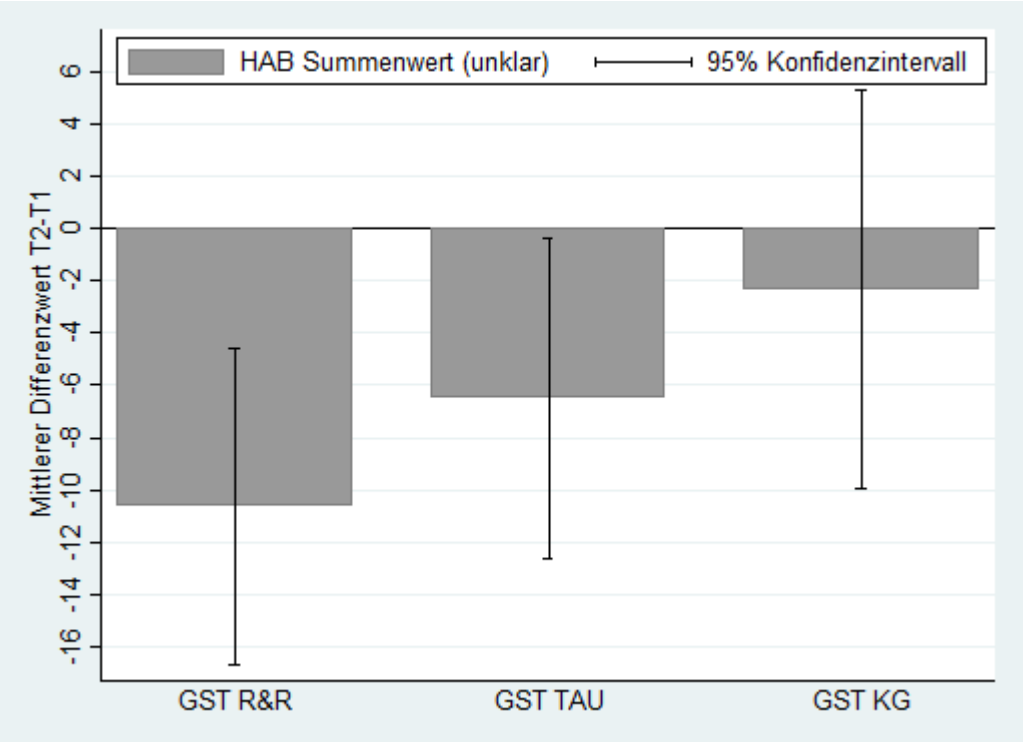

Abbildung 7. Mittlere Differenzwerte im HAB (Situationen mit unklarer Absicht)

Tabelle 21. Differenzwerte im HAB: Situationen mit unklarer Absicht

GST Bedingung	Fehlend [% (*N*)]	*N*	*M*	*SA*	*p*
P (Wahrscheinlichkeit für Wahrnehmung als absichtliche Provokation), unklare Absicht					
GST R&R	0.9% (*N*=1)	109	-1.0	5.2	
GST TAU	1.8% (*N*=1)	56	-1.4	4.8	.270 [a)]
GST KG	0% (*N*=0)	51	0.1	5.2	
V (Wahrscheinlichkeit für Verärgerung), unklare Absicht					
GST R&R	1.8% (*N*=2)	108	-1.5	5.4	
GST TAU	1.8% (*N*=1)	56	-1.2	4.7	.619 [a)]
GST KG	0% (*N*=0)	51	-0.6	4.0	
M (Wahrscheinlichkeit für das Mitteilen der Verärgerung), unklare Absicht					
GST R&R	1.8% (*N*=2)	108	-1.6	6.2	
GST TAU	1.8% (*N*=1)	56	-0.2	6.7	.305 [a)]
GST KG	0% (*N*=0)	51	-0.6	4.6	
U (Wahrscheinlichkeit für unhöfliches Verhalten), unklare Absicht					

GST R&R	1.8% (*N*=2)	108	-2.1	5.9	
GST TAU	1.8% (*N*=1)	56	-0.6	4.3	.079 [b)]
GST KG	0% (*N*=0)	51	-0.4	4.4	
AB (Wahrscheinlichkeit für Anschreien/ Beschimpfen), unklare Absicht					
GST R&R	1.8% (*N*=2)	108	-1.8	5.6	
GST TAU	1.8% (*N*=1)	56	-0.8	3.9	.066 [a)]
GST KG	0% (*N*=0)	51	0.2	4.9	
D (Wahrscheinlichkeit für Drohen), unklare Absicht					
GST R&R	1.8% (*N*=2)	108	-1.2	5.6	
GST TAU	1.8% (*N*=1)	56	-1.0	3.8	.798 [a)]
GST KG	0% (*N*=0)	51	-0.6	6.1	
G (Wahrscheinlichkeit für physische Gewalt), unklare Absicht					
GST R&R	1.8% (*N*=2)	108	-1.3	5.4	
GST TAU	1.8% (*N*=1)	56	-1.4	3.9	.549 [a)]
GST KG	0% (*N*=0)	51	-0.5	5.2	
HAB Summenwert, unklare Absicht					
GST R&R	1.8% (*N*=2)	108	-10.6	31.6	
GST TAU	1.8% (*N*=1)	56	-6.5	22.9	.227 [a)]
GST KG	0% (*N*=0)	51	-2.4	27.2	

Anmerkungen. Angegeben sind die Differenzwerte T2 – T1: Negative Werte bedeuten eine relative Abnahme in der Ausprägung der Items einer Skala; a) Ergebnis einer einfaktoriellen Varianzanalyse; b) Ergebnis des Welch-Tests.

Abbildung 8 zeigt die mittleren Differenzwerte im Summenwert des HAB in den drei Bedingungen der GST-Gruppe für Situationen, in denen das Verhalten des Gegenübers eindeutig als nicht provozierend einzuschätzen ist. In Tabelle 22 sind zusätzlich die Differenzwerte in den einzelnen Skalen des HAB unter Angabe von Stichprobengröße, Standardabweichung und Signifikanzniveau aufgeführt.
ANOVAs ergaben keine Hinweise auf Unterschiede in den Differenzwerten der Probanden in den drei Bedingungen der GST-Gruppe, weder in den einzelnen Aspekten aggressiver Wahrnehmung bzw. aggressiven Verhaltens noch im Summenwert.

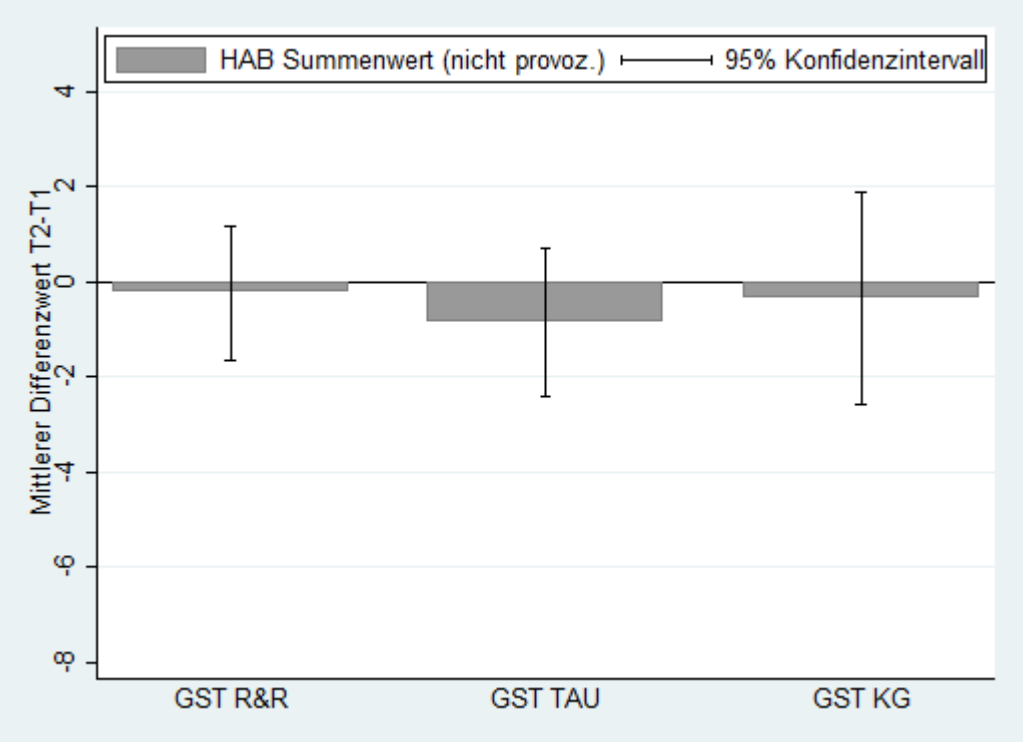

Abbildung 8. Mittlere Differenzwerte im Summenwert des HAB (nicht provozierende Situationen)

Tabelle 22. Differenzwerte im HAB: Situationen ohne Provokation

GST Bedingung	Fehlend [% (*N*)]	*N*	*M*	*SA*	*p*
P (Wahrscheinlichkeit für Wahrnehmung als absichtliche Provokation), nicht-provozierende Absicht					
GST R&R	1.8% (*N*=2)	108	-0.0	1.3	
GST TAU	0% (*N*=0)	57	-0.1	1.1	.718 [a)]
GST KG	0% (*N*=0)	51	0.1	1.4	
V (Wahrscheinlichkeit für Verärgerung), nicht-provozierende Absicht					
GST R&R	2.7% (*N*=3)	107	0.0	1.8	
GST TAU	0% (*N*=0)	57	-0.0	1.8	.902 [a)]
GST KG	0% (*N*=0)	51	-0.1	1.7	
M (Wahrscheinlichkeit für das Mitteilen der Verärgerung), nicht-provozierende Absicht					
GST R&R	3.6% (*N*=4)	106	0.0	1.8	
GST TAU	0% (*N*=0)	57	-0.2	2.2	.779 [a)]
GST KG	0% (*N*=0)	51	-0.1	1.5	
U (Wahrscheinlichkeit für unhöfliches Verhalten), nicht-provozierende Absicht					
GST R&R	3.6% (*N*=4)	106	-0.1	1.4	.862 [a)]
GST TAU	0% (*N*=0)	57	-0.0	1.2	

GST KG	0% (*N*=0)	51	-0.2	1.5	
AB (Wahrscheinlichkeit für Anschreien/ Beschimpfen), nicht-provozierende Absicht					
GST R&R	2.7% (*N*=3)	107	-0.1	1.3	
GST TAU	0% (*N*=0)	57	-0.2	0.9	.640 [a)]
GST KG	0% (*N*=0)	51	0.0	1.6	
D (Wahrscheinlichkeit für Drohen), nicht-provozierende Absicht					
GST R&R	2.7% (*N*=3)	107	0.0	1.0	
GST TAU	0% (*N*=0)	57	-0.2	0.7	.554 [a)]
GST KG	0% (*N*=0)	51	-0.1	1.4	
G (Wahrscheinlichkeit für physische Gewalt), nicht-provozierende Absicht					
GST R&R	2.7% (*N*=3)	107	0.0	1.1	
GST TAU	0% (*N*=0)	57	-0.1	0.8	.867 [a)]
GST KG	0% (*N*=0)	51	0.0	0.8	
HAB Summenwert, nicht-provozierende Absicht					
GST R&R	3.6% (*N*=4)	106	-0.2	7.3	
GST TAU	0% (*N*=0)	57	-0.8	5.9	.871 [a)]
GST KG	0% (*N*=0)	51	-0.4	8.0	

Anmerkungen. Angegeben sind die Differenzwerte T2 – T1: Negative Werte bedeuten eine relative Abnahme in der Ausprägung der Items einer Skala; a) Ergebnis einer einfaktoriellen Varianzanalyse.

HAB: Fremdbeurteilung (reduziert: 1 Item)

Die Fremdbeurteilung des HAB besteht aus einem Item. Dieses Item fragt nach der Einschätzung der Zuschreibung feindseliger Absichten des Patienten in uneindeutigen, möglichen Konfliktsituationen. Für dieses Item liegt ein substanzieller Anteil an fehlenden Werten vor. In Tabelle 23 werden dennoch zur deskriptiven Betrachtung die Differenzwerte zwischen beiden Messzeitpunkten in diesem Item nach Bedingung aufgeführt. Die Mittelwerte wurden einschließlich der Ausreißer-Werte berechnet.

Tabelle 23. Differenzwerte im Fremdbeurteilungs-Item des HAB

GST Bedingung	**Fehlend [% (*N*)]**	***N***	***M***	***SA***	***Median***
GST R&R	25.5% (*N*=28)	82	-0.1	0.9	0
GST TAU	35.1% (*N*=20)	37	0.6	5.2	0
GST KG	64.7% (*N*=33)	18	0.3	1.1	0

Anmerkungen. Angegeben sind die Differenzwerte T2 – T1: Negative Werte bedeuten eine geringer ausgeprägte Zuschreibung feindseliger Absichten zum zweiten Messzeitpunkt.

Fragebogen zur Verantwortungsübernahme (VÜ)

VÜ: Selbstbeurteilung

Abbildung 9 zeigt die mittleren Differenzwerte im Gesamtwert des Fragebogens zur Verantwortungsübernahme (VÜ) als Differenzwert der Summenwerte T2-T1. Tabelle 24 zeigt zusätzlich zum Gesamtwert die Veränderungen in den Subskalen „Rechtfertigung“ und „Entschuldigung“ unter Angabe von Stichprobengröße, Standardabweichung und Signifikanzniveau. Fehlende Werte wurden durch den Median der jeweiligen Bedingung ersetzt, sofern der Anteil fehlender Angaben je Fall bei maximal 20% lag. Die Berechnung wurde explizit unter Einschluss von Ausreißer-Werten durchgeführt (bezogen auf den VÜ-Gesamtwert ergab ein Grubbs-Test Hinweise auf zwei Ausreißer, bezogen auf die Subskala „Rechtfertigung“ keinen Ausreißer und bezogen auf die Subskala „Entschuldigung“ einen Ausreißer). Es wurde eine Überprüfung der Varianzhomogenität und der näherungsweisen Normalverteilung der Differenzwerte vorgenommen, um die Voraussetzungen zur Berechnung einer ANOVA zu überprüfen.

Der Differenzwert des Fragebogens zur Verantwortungsübernahme unterschied sich signifikant zwischen den Probanden der drei Bedingungen der GST-Gruppe: $F(2, 186) = 3.17$, $p=.044$; $\eta^2 = .033$. Ein post hoc durchgeführter Tukey-Kramer-Test zeigte auf einem Alpha-Niveau von .05 einen signifikanten Unterschied zwischen der Experimental- und der Kontrollgruppe: Während die Probanden der Experimentalgruppe eine erhöhte Verantwortungsübernahme zum zweiten Messzeitpunkt berichten, berichten diejenigen der Kontrollgruppe eine verminderte Verantwortungsübernahme. Bezogen auf die Differenzwerte der beiden Subskalen „Rechtfertigung“ und „Entschuldigung“ ergaben ANOVAs zur Überprüfung auf Unterschiede zwischen den drei Bedingungen keine signifikanten Resultate.

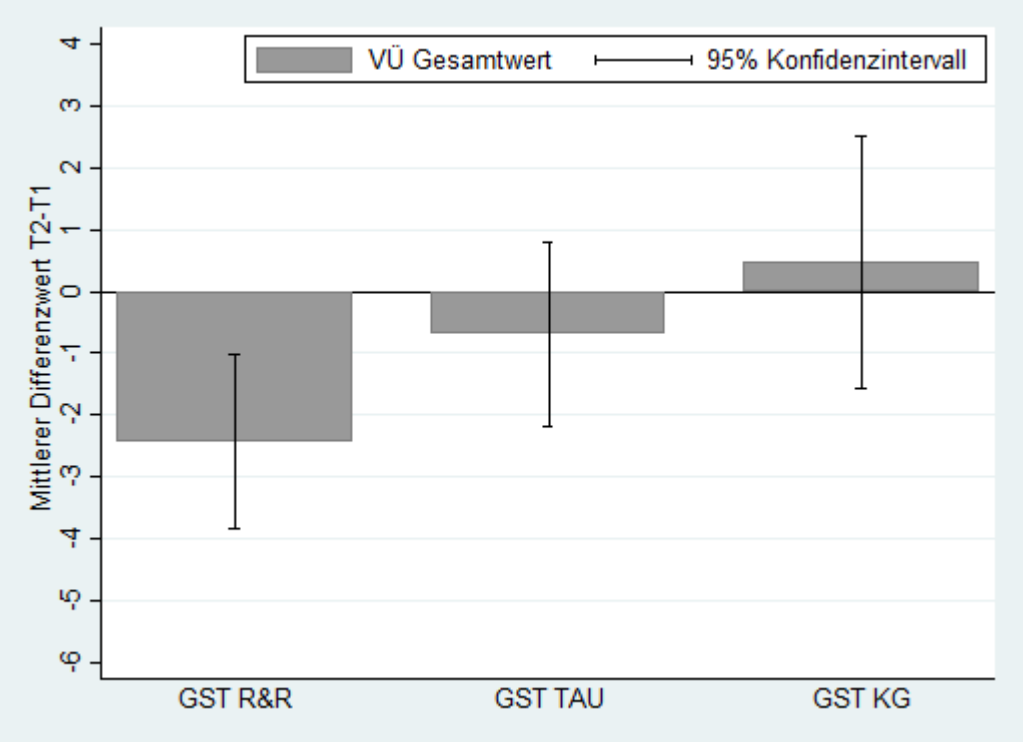

Abbildung 9. Mittlere Differenzwerte im Gesamtwert des VÜ

Tabelle 24. Differenzwerte im VÜ (Selbstbeurteilung)

GST Bedingung	**Fehlend [% (*N*)]**	***N***	***M***	***SA***	***p***
Subskala „Rechtfertigung“					
GST R&R	7.3% (*N*=8)	102	-0.5	2.9	
GST TAU	7.0% (*N*=4)	53	0.1	2.8	.456
GST KG	13.7% (*N*=7)	44	-0.6	3.7	
Subskala „Entschuldigung“					
GST R&R	6.4% (*N*=7)	103	-1.0	3.4	
GST TAU	7.0% (*N*=4)	53	-0.7	3.4	.332
GST KG	9.8% (*N*=5)	46	-0.1	3.1	
VÜ Gesamtwert					
GST R&R	10.9% (*N*=12)	98	-2.4	7.1	
GST TAU	10.5% (*N*=6)	51	-0.7	5.4	**.044**
GST KG	21.6% (*N*=11)	40	0.5	6.4	

Anmerkungen. Angegeben sind die Differenzwerte T2 – T1: Negative Werte bedeuten eine höhere Verantwortungsübernahme zum Zeitpunkt T2. **Rot** = Mehr als 20% fehlende Werte; da dies nur knapp über dem für diese Auswertung angewandten Grenzwert von 20% liegt, werden die Werte an dieser Stelle dennoch aufgeführt und die Berechnung durchgeführt.

VÜ: Fremdbeurteilung (reduziert: 1 Item)

Die Fremdbeurteilung des Fragebogens zur Verantwortungsübernahme besteht aus einem einzigen Item. Dieses fragt, ob der zu beurteilende Patient die Verantwortung für sein Hauptdelikt übernimmt. Dieses hat zu beiden Messzeitpunkten in einem substanziellen Teil der Fälle keine Angaben. In der Kontrollgruppe fehlen die Angaben jeweils in über der Hälfte der Fälle. In Tabelle 25 werden dennoch die Differenzwerte zwischen den beiden Messzeitpunkten zu einer deskriptiven Einschätzung aufgeführt, auch weil eine Fremdeinschätzung einen deutlichen Mehrwert zur reinen Selbsteinschätzung durch die Probanden selbst darstellt.

Tabelle 25. Differenzwerte des Fremdbeurteilungs-Items des VÜ

GST Bedingung	Fehlend [% (*N*)]	*N*	*M*	*SA*	*Median*
GST R&R	27.3% (*N*=30)	80	0.2	0.8	0
GST TAU	35.1% (*N*=20)	37	0.1	0.5	0
GST KG	66.7% (*N*=34)	17	0.4	1.0	0

Anmerkungen. Angegeben sind die Differenzwerte T2 – T1: Negative Werte bedeuten eine höhere Verantwortungsübernahme zum zweiten Messzeitpunkt.

Fragestellung 2: Rückfälligkeit GST

In der folgenden Auswertung werden sämtliche Probanden berücksichtigt, die in eine der drei Bedingungen der GST-Gruppe eingeschlossen worden waren und die zumindest an der Prä-Messung teilgenommen hatten. Dieser Ansatz entspricht einer Intent-to-treat-Analyse und ist die bevorzugte Art der Auswertung.
Anschließend werden die Ergebnisse einer zusätzlichen Auswertung der Rückfälligkeit präsentiert, in der ausschließlich diejenigen Probanden berücksichtigt wurden, die den MV komplett absolviert haben.
Für beide Arten der Rückfall-Bestimmung wurden ausschließlich diejenigen Probanden berücksichtigt, die zum Zeitpunkt der Ziehung aus dem Strafregister bereits in Freiheit entlassen worden waren (gemäß Variable „austrittsdatum“).

Rückfälligkeit GST: Intent-to-Treat-Analyse

In Tabelle 26 werden die Ergebnisse der Auswertung der Rückfälligkeit präsentiert, in der auch die Abbrecher des MV mit eingeschlossen wurden (Intent-to-treat-Analyse; siehe Abschnitt „Kategorisierung des Delikts“). Die „Time at Risk“ gibt die Differenz zwischen der Entlassung und dem Datum der Ziehung aus dem Strafregister an.

Von den 278 für die Evaluation der Rückfälligkeit berücksichtigten Probanden der GST-Gruppe, für die ein Auszug aus dem Strafregister vorliegt, erfüllten lediglich 63 die oben genannten Einschlusskriterien für eine Berechnung der Rückfälligkeit (22.7%). Von diesen 63 Probanden wurden insgesamt zwölf (19.1%) im zur Verfügung stehenden Beobachtungszeitraum rückfällig, davon fünf in der Experimental-, vier in der Vergleichs- und drei in der Kontrollgruppe. Exakte logistische Regressionen, trotz der geringen Fallzahl zur Überprüfung auf Unterschiede zwischen den drei Bedingungen berechnet, führten weder in Bezug auf die allgemeine Rückfälligkeit (p=.772) noch in Bezug auf die Rückfälligkeit wegen eines Gewalt- oder Sexualdelikts (p=.261) zu signifikanten Ergebnissen.

Die Time at Risk beträgt in dieser Auswertung im Mittel über alle Bedingungen hinweg 20 Monate, entsprechend 1.7 Jahren (SA=12.1 Monate; Range: 0.4 – 44.5 Monate). Ein Kruskal-Wallis-H-Test ergab keine Hinweise auf eine unterschiedlich lange Time at Risk in den drei Bedingungen ($\chi^2(2)=0.16$; p=.924).

Tabelle 26. Rückfälligkeit GST (Intent-to-Treat)

GST Bedingung	Fehlend [% (*N*)]	*N*	Time at risk [Monate] [*M* (*SA*)]	*p*	Rückfällig [% (*N*)]	*p*	Rückfällig (Gewalt- oder Sexualdelikt) [% (*N*)]	*p*
GST R&R	0% (*N*=0)	30	*M*=19.8 (*SA*=10.1)	.924[a)]	16.7% (*N*=5)	.772[b)]	3.3% (*N*=1)	.261[b)]
GST TAU	0% (*N*=0)	15	*M*=21.0 (*SA*=14.9)		26.7% (*N*=4)		13.1% (*N*=2)	
GST KG	0% (*N*=0)	18	*M*=19.6 (*SA*=13.4)		16.7% (*N*=3)		0% (*N*=0)	
GST Gesamt	**0% (*N*=0)**	**63**	***M*=20.0 (*SA*=12.1)**		**19.1% (*N*=12)**		**4.8% (*N*=3)**	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; TAR: Time at Risk; [a)] Ergebnis eines Kruskal-Wallis-H-Tests; [b)] Ergebnis einer exakten logistischen Regression.

Rückfälligkeit GST: Nur Vollender

In Tabelle 27 wird die Rückfälligkeit in Abhängigkeit der drei Bedingungen der GST-Gruppe präsentiert, wobei für diese Auswertung nur jene Probanden berücksichtigt wurden, die den MV bis zum Ende absolvierten und die vor der Ziehung aus dem Strafregister in Freiheit entlassen wurden (siehe Abschnitt „Kategorisierung des Delikts“).
Von den 211 Probanden der GST-Gruppe, die den MV vollendet haben und für die ein Auszug aus dem Strafregister vorliegt, erfüllten lediglich 42 die oben genannten Einschlusskriterien für eine Berechnung der Rückfälligkeit (19.9%). Von diesen 42 Probanden wurden insgesamt sieben (16.7%) im zur Verfügung stehenden Beobachtungszeitraum rückfällig, davon vier in der Experimental- und drei in der Vergleichs-

gruppe (15.4 bzw. 33.3% der für diese Berechnung eingeschlossenen ST). Exakte logistische Regressionen, trotz der geringen Fallzahl berechnet, ergaben, dass sich die Probanden in den drei Bedingungen der GST-Gruppe weder in der allgemeinen Rückfälligkeit (p=.197) noch in der Rückfälligkeit wegen eines Gewalt- oder Sexualdelikts (p=.178) voneinander unterscheiden.

Die Time at risk, definiert als die Dauer von der Entlassung in die Freiheit bis zum Datum der Ziehung aus dem Strafregister, beträgt über alle Bedingungen der GST-Gruppe hinweg im Mittel 19.3 Monate = etwa 1.6 Jahre (SA=11.5 Monate; Range: 1.8 – 44.5 Monate). Ein Kruskal-Wallis-H-Test ergab keine Hinweise auf eine unterschiedlich lange Time at Risk in den drei Bedingungen ($\chi^2(2)=1.62$; p=.446).

Tabelle 27. Rückfälligkeit GST (in Freiheit entlassene Vollender)

GST Bedingung	Fehlend [% (*N*)]	*N*	Time at risk [Monate] [*M* (*SA*)]	*p*	Rückfällig [% (*N*)]	*p*	Rückfällig (Gewalt- oder Sexualdelikt) [% (*N*)]	*p*
GST R&R	0% (*N*=0)	26	*M*=20.3 (*SA*=10.2)		15.4% (*N*=4)		3.9% (*N*=1)	
GST TAU	0% (*N*=0)	9	*M*=18.5 (*SA*=14.6)	.446[a)]	33.3% (*N*=3)	.197[b)]	22.2% (*N*=2)	.178[b)]
GST KG	0% (*N*=0)	7	*M*=17.0 (*SA*=13.5)		0% (*N*=0)		0% (*N*=0)	
GST Gesamt	**0% (*N*=0)**	**42**	***M*=19.3 (*SA*=11.5)**		**16.7% (*N*=7)**		**7.1% (*N*=3)**	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm (2); TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; TAR: Time at Risk; [a)] Ergebnis eines Kruskal-Wallis-H-Tests; [b)] Ergebnis einer exakten logistischen Regression.

Ergebnisse ASAT®Suisse

Im Folgenden werden die Ergebnisse der Gruppe SST präsentiert. Zunächst werden die den Analysen zugrundeliegenden Stichproben vergleichend beschrieben: Es werden Merkmale untersucht, welche einen Einfluss auf das Rückfallrisiko haben können. Dies beinhaltet demografische Merkmale, die Vorstrafenbelastung, die psychiatrische Belastung, Summenwerte in etablierten Risk-Assessment-Instrumenten sowie die Therapieerfahrung der Probanden. Den Vergleichsuntersuchungen liegen dieselben Ziele zugrunde wie für die GST-Gruppe erläutert: Die Untersuchung auf vorbestehende Unterschiede zwischen den drei Bedingungen der SST-Gruppe in Bezug auf alle in den MV eingeschlossenen Probanden soll eine Einschätzung der Vergleichbarkeit der Bedingungen ermöglichen. Die Darstellung des Stichprobenschwunds soll das Verhältnis zwischen Abbrechern und Vollendern der SST-Gruppe aufzeigen. Ein Vergleich zwischen Abbrechern und Vollendern anhand zentraler Merkmale der Probanden geht der Frage nach, ob das Ausscheiden aus dem MV mit den untersuchten Merkmalen korreliert, sowohl innerhalb der einzelnen Bedingungen als auch in Bezug auf die SST-Gesamtgruppe.

Vergleichende Stichprobenbeschreibung SST

Vorbestehende Unterschiede zu T1

Im Folgenden werden vorbestehende Unterschiede zwischen den Probanden der drei Bedingungen der SST-Gruppe hinsichtlich demografischer Merkmale, Art des Indexdelikts und Merkmalen mit Einfluss auf das Rückfallrisiko dargestellt. Diese Auswertung soll die Vergleichbarkeit der Probanden in den drei Bedingungen der SST-Gruppe statistisch überprüfen.

Demografische Merkmale: Vorbestehende Unterschiede

Nationalität

Tabelle 28 zeigt den Anteil an Schweizer Staatsbürgern in den drei Bedingungen der SST-Gruppe. Ein exakter Test nach Fisher ergab keine Hinweise auf Unterschiede zwischen den drei Bedingungen der Gruppe SST hinsichtlich des Anteils an Schweizer Probanden (p=.539).

Tabelle 28. Vorbestehende Unterschiede in der Nationalität SST

SST Bedingung	Fehlende Werte [% (*N*)]	*N*	Schweizer Nationalität [% (*N*)]	*p*
SST ASAT	2.1% (*N*=1)	46	95.7% (*N*=44)	.539
SST TAU	0% (*N*=0)	27	92.6% (*N*=25)	

SST KG	0% (*N*=0)	11	90.9% (*N*=10)

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe.

Alter

Tabelle 29 zeigt das Alter aller berücksichtigten Probanden in den drei Bedingungen der Gruppe SST zum Zeitpunkt des Indexurteils. Kruskal-Wallis-H-Tests zur Überprüfung auf Unterschiede im Alter der Probanden zum Zeitpunkt des Indexurteils zwischen den drei Bedingungen der SST-Gruppe ergaben kein signifikantes Ergebnis ($\chi^2(2)=1.91$; $p=.385$).

Tabelle 29. Vorbestehende Unterschiede im Alter SST

SST Bedingung	Fehlende Werte [% (*N*)]	Alter Indexurteil [*M* (*SA*)]	*p*
SST ASAT	2.1% (*N*=1)	40.4 (10.6)	
SST TAU	3.7% (*N*=1)	42.3 (11.2)	.385
SST KG	0% (*N*=0)	45.9 (12.3)	

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe.

Bildungsniveau

Tabelle 30 zeigt das höchste erreichte Bildungsniveau aller eingeschlossenen Probanden der SST-Gruppe[12]. Ein exakter Test nach Fisher ergab keine Hinweise auf Unterschiede im Bildungsniveau der Probanden in den drei Bedingungen der SST-Gruppe ($p=.952$).

Tabelle 30. Vorbestehende Unterschiede im Bildungsniveau SST

SST Bedingung	Fehlende Werte [% (*N*)]	Höchster Abschluss	Anteil [% (*N*)]	*p*
SST ASAT	4.3% (*N*=2)	< 7 Jahre Schulbildung	4.4% (*N*=2)	.952
		Schulpflicht abgeschl. ggf. zzgl. Anlehre	24.4% (*N*=11)	
		Mindestens abgeschl. Lehre	66.7% (*N*=30)	
		Nicht-Schweizer Probanden	4.4% (*N*=2)	
SST TAU	3.7% (*N*=1)	< 7 Jahre Schulbildung	3.9% (*N*=1)	

[12] Ausländische Probanden werden analog zum Vorgehen bezüglich der GST-Gruppe separat ausgewiesen, da bei diesen eine valide Zuordnung des Bildungsniveaus anhand vorliegender Daten nicht möglich ist.

		Schulpflicht abgeschl. ggf. zzgl. Anlehre	26.9% (*N*=7)
		Mindestens abgeschl. Lehre	61.5% (*N*=16)
		Nicht-Schweizer Probanden	7.7% (*N*=2)
SST KG	9.1% (*N*=1)	< 7 Jahre Schulbildung	0% (*N*=0)
		Schulpflicht abgeschl. ggf. zzgl. Anlehre	30.0% (*N*=3)
		Mindestens abgeschl. Lehre	60.0% (*N*=6)
		Nicht-Schweizer Probanden	10.0% (*N*=1)

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe.

Psychiatrische Belastung: Vorbestehende Unterschiede

Tabelle 31 zeigt den Anteil derjenigen Probanden in den drei Bedingungen der SST-Gruppe, die eine Diagnose gemäß (aktuellstem) psychiatrischem Gutachten in verschiedenen Bereichen des ICD-10 aufweisen: Persönlichkeitsstörungen (Kategorie F6), Schizophrenie, schizotype und wahnhafte Störungen (Kategorie F2), affektive Störungen (Kategorie F3) sowie Störungen durch psychotrope Substanzen (Kategorie F1). Der Zeitpunkt der Diagnosestellung wurde nicht spezifisch erfasst. Das Zustandekommen des hohen Anteils an fehlenden Werten in der Kontrollgruppe lässt sich analog zur GST-Gruppe durch fehlende Akteninformationen bzw. durch das gänzliche Fehlen eines psychiatrischen Gutachtens erklären.

Chi-Quadrat- bzw. exakte Tests nach Fisher zwischen der Experimental- und der Vergleichsgruppe, die aufgrund hinreichender Angaben durchgeführt werden konnten, ergaben keine Hinweise auf Unterschiede im Anteil an Probanden mit den aufgeführten Diagnosen.

Tabelle 31. Vorbestehende Unterschiede in der psychiatrischen Vorbelastung SST

	SST Bedingung	Fehlende Werte [% (*N*)]	*N*	Anteil [% (*N*)]	*p*
Persönlichkeitsstörung (ICD-10 Kategorie F6)	SST ASAT	10.6% (*N*=5)	42	54.8% (*N*=23)	N/A
	SST TAU	3.7% (*N*=1)	26	46.2% (*N*=12)	
	SST KG	36.4% (*N*=4)	7	57.2% (*N*=4)	
Schizophrenie, schizotype und wahnhafte Störungen (ICD-10 Kategorie F2)	SST ASAT	10.6% (*N*=5)	42	4.8% (*N*=2)	N/A
	SST TAU	3.7% (*N*=1)	26	7.7% (*N*=2)	
	SST KG	36.4% (*N*=4)	7	0% (*N*=0)	
Störungen durch psychotrope Substanzen (ICD-10 Kategorie F1)	SST ASAT	8.5% (*N*=4)	43	25.6% (*N*=11)	N/A
	SST TAU	3.7% (*N*=1)	26	15.4% (*N*=4)	
	SST KG	36.4% (*N*=4)	7	28.6% (*N*=2)	

Affektive Störungen (ICD-10 Kategorie F3	SST ASAT	10.6% (*N*=5)	42	7.1% (*N*=3)	
	SST TAU	3.7% (*N*=1)	26	3.8% (*N*=1)	N/A
	SST KG	36.4% (*N*=4)	7	14.3% (*N*=1)	

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; Rot = mehr als 20% fehlende Werte. Aufgrund der ohnehin kleinen Fallzahlen in der Kontrollgruppe und zur deskriptiven Übersicht über die psychiatrische Belastung der Probanden werden die Anteile dennoch aufgeführt.

Summenwerte in Risk-Assessment-Instrumenten: Vorbestehende Unterschiede

Die angewandten Risk-Assessment-Instrumente VRAG, PCL-R und Static-99 zeichnen sich durch einen substanziellen Anteil an fehlenden Werten aus, sodass zu den vorbestehenden Unterschieden der Probanden hinsichtlich ihres (Baseline-) Risikos keine verlässliche Aussage getroffen werden kann. Die Mittelwerte der Probanden in den drei Bedingungen der SST-Gruppe in den Summenwerten von VRAG, PCL-R und Static-99 sind dennoch in Tabelle 32 deskriptiv aufgeführt, da es sich bei den etablierten Instrumenten um wichtige Indikatoren für das vorbestehende Rückfallrisiko handelt.

Tabelle 32. Vorbestehende Unterschiede in Risk-Assessment-Instrumenten SST

	SST Bedingung	**Fehlende Werte [% (*N*)]**	***N***	***M*(*SA*)**	***Median***
VRAG Summenwert[*)]	SST ASAT	31.9% (*N*=15)	32	3.2 (11.7)	1.5
	SST TAU	37.0% (*N*=10)	17	1.4 (7.4)	-1
	SST KG	54.6% (*N*=6)	5	4.2 (18.9)	-2
PCL-R Summenwert	SST ASAT	25.5% (*N*=12)	35	17.0 (8.2)	17.0
	SST TAU	18.5% (*N*=5)	22	15.0 (7.5)	13.5
	SST KG	63.6% (*N*=7)	4	16.0 (11.3)	16.1
Static-99 Summenwert	SST ASAT	42.6% (*N*=20)	27	3.8 (2.1)	3
	SST TAU	40.7% (*N*=11)	16	3.8 (2.3)	4
	SST KG	54.6% (*N*=6)	5	3.2 (2.6)	4

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [*)] Hierbei wurde überprüft, ob der VRAG-Summenwert bereits korrigiert wurde (dies ist bis zu einer maximalen Anzahl von vier fehlenden Angaben je Bewertung möglich). Dies ist der Fall: Der Anteil an fehlenden Werten im Item „vragtot“ ist identisch mit dem Anteil derjenigen Fälle, die in mehr als vier Items des VRAG fehlende Angaben haben.

Therapieerfahrung: Vorbestehende Unterschiede

Tabelle 33 zeigt den Anteil der Probanden aus der SST-Gruppe, die bereits vor Beginn des MV mindestens einmal in psychotherapeutischer Behandlung waren. Ein

Chi-Quadrat-Test zwischen den Bedingungen ASAT und TAU, für die weniger als 20% der Werte fehlen, ergab einen signifikant höheren Anteil an Probanden in der Experimentalgruppe, die vor Beginn des MV bereits mindestens einmalig in psychotherapeutischer Behandlung waren ($\chi^2(1)=4.24$; $p=.039$).

Tabelle 33. Vorbestehende Unterschiede in der Therapieerfahrung SST

	SST Bedingung	Fehlende Werte [% (*N*)]	*N*	Anteil [% (*N*)]	*p*
Psychotherapie-Erfahrung	SST ASAT	17.0% (*N*=8)	39	82.1% (*N*=32)	N/A
	SST TAU	11.1% (*N*=3)	24	58.3% (*N*=14)	
	SST KG	45.5% (*N*=5)	6	66.7% (*N*=4)	

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; Rot = mehr als 20% fehlende Werte.

Stichprobenschwund SST

Tabelle 34 zeigt die Anzahl der für die vorliegende Evaluation berücksichtigten Probanden der SST-Gruppe zu Beginn des MV sowie den Anteil derjenigen Probanden, welche an der Studie bis zum Ende (Vollender: Daten liegen sowohl für Prä- als auch Post-Messung vor) bzw. nicht bis zum Ende (Abbrecher: Nur Daten der Prä-Messung liegen vor) teilgenommen haben. Eine weiterführende Unterscheidung nach den Gründen für einen Abbruch, insbesondere, ob es sich bei den Abbrechern um Therapie- oder um Studienabbrecher handelte, lässt sich nicht vornehmen, da hierzu keine Daten vorliegen. Auch der Anteil derjenigen angefragten potenziellen Probanden, die ihre Teilnahme bereits vor Studienbeginn abgelehnt hatten, wurde nicht erhoben. Ein Chi-Quadrat-Test ergab signifikante Unterschiede zwischen den Bedingungen im Anteil derjenigen Probanden, welche den MV vorzeitig abbrachen ($\chi^2(2)=8.75$; $p=.013$): Post-hoc-Tests zeigten, dass dieser Anteil in der Kontrollgruppe signifikant höher lag als in der Experimentalgruppe ($\chi^2(1)=8.79$; $p=.003$). Nach Bonferroni-Korrektur des Alpha-Niveaus ($\alpha=.025$) ergab der Einzelvergleich zwischen Vergleichs- und Kontrollgruppe knapp kein signifikantes Ergebnis ($\chi^2(1)=3.99$; $p=.046$).

Tabelle 34. Vollender und Abbrecher SST nach Bedingung

SST Bedingung	ST Gesamt (*N*)	Vollender [% (*N*)]	Abbrecher [% (*N*)]	*p*
SST ASAT	47	74.5% (*N*=35)	25.5% (*N*=12)	.013
SST TAU	27	63.0% (*N*=17)	37.0% (*N*=10)	
SST KG	11	27.3% (*N*=3)	72.7% (*N*=8)	
SST Gesamt	**85**	**64.7% (*N*=55)**	**35.4% (*N*=30)**	

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe.

Unterschiede zwischen Abbrechern und Vollendern

Demografische Merkmale: Unterschiede zwischen Abbrechern und Vollendern

Nationalität

Tabelle 35 zeigt den Anteil Schweizer Probanden in den drei Bedingungen der SST-Gruppe sowie für die SST-Gesamtstichprobe, separat für Abbrecher und Vollender. Exakte Tests nach Fisher ergaben keine Hinweise auf Unterschiede im Anteil an Probanden mit Schweizer Staatsbürgerschaft zwischen Abbrechern und Vollendern.

Tabelle 35. Nationalität SST Abbrecher versus Vollender

SST Bedingung	Fehlende Werte [% (*N*)]	*N*	Schweizer Nationalität [% (*N*)]	*p*
SST ASAT alle zugeordneten Probanden	2.1% (*N*=1)	46	95.7% (*N*=44)	
SST ASAT Abbrecher	8.3% (*N*=1)	11	100% (*N*=11)	.575 [a)]
SST ASAT Vollender	0% (*N*=0)	35	94.3% (*N*=33)	
SST TAU alle zugeordneten Probanden	0% (*N*=0)	27	92.3% (*N*=24)	
SST TAU Abbrecher	0% (*N*=0)	10	90.0% (*N*=9)	.613 [a)]
SST TAU Vollender	0% (*N*=0)	17	94.1% (*N*=16)	
SST KG alle zugeordneten Probanden	0% (*N*=0)	11	90.0% (*N*=9)	
SST KG Abbrecher	0% (*N*=0)	8	87.5% (*N*=7)	.727 [a)]
SST KG Vollender	0% (*N*=0)	3	100% (*N*=3)	
SST Gesamt alle zugeordneten Probanden	1.2% (*N*=1)	84	94.1% (*N*=79)	
SST Gesamt Abbrecher	3.3% (*N*=1)	29	93.1% (*N*=27)	.567 [a)]
SST Gesamt Vollender	0% (*N*=0)	55	94.6% (*N*=52)	

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [a)] Ergebnis des Exakten Tests nach Fisher. Dieser Test wird anstelle des Chi-Quadrat-Tests angewandt, wenn dessen Voraussetzung einer erwarteten Häufigkeit von ≥ 5 je Zelle verletzt ist, wie im vorliegenden Fall.

Alter

Tabelle 36 zeigt das Alter der Probanden zum Zeitpunkt des Indexurteils in den drei Bedingungen der SST-Gruppe sowie für die SST-Gesamtstichprobe, separat für Abbrecher und Vollender. Wilcoxon-Mann-Whitney-Tests ergaben keine Hinweise auf Unterschiede zwischen Abbrechern und Vollendern des MV hinsichtlich des Alters der Probanden zum Zeitpunkt des Indexurteils. Dies gilt sowohl für die Gesamtgruppe SST als auch für die drei Bedingungen.

Tabelle 36. Alter SST Abbrecher versus Vollender

SST Bedingung	Fehlende Werte [% (*N*)]	*N*	Alter Indexurteil [*M* (*SA*)]	*p*
SST ASAT alle zugeordneten Probanden	2.1% (*N*=1)	46	40.4 (10.6)	
SST ASAT Abbrecher	0% (*N*=0)	12	39.8 (9.2)	.901 [a)]
SST ASAT Vollender	2.9% (*N*=1)	34	40.6 (11.1)	
SST TAU alle zugeordneten Probanden	3.7% (*N*=1)	26	42.3 (11.2)	
SST TAU Abbrecher	0% (*N*=0)	10	42.3 (13.3)	.916 [a)]
SST TAU Vollender	5.6% (*N*=1)	16	42.3 (10.1)	
SST KG alle zugeordneten Probanden	0% (*N*=0)	11	45.9 (12.3)	
SST KG Abbrecher	0% (*N*=0)	8	44.6 (10.6)	.414 [a)]
SST KG Vollender	0% (*N*=0)	3	49.2 (18.7)	
SST Gesamt alle zugeordneten Probanden	2.4% (*N*=2)	83	41.7 (11.0)	
SST Gesamt Abbrecher	0% (*N*=0)	30	41.9 (10.9)	.909 [a)]
SST Gesamt Vollender	3.6% (*N*=2)	53	41.6 (11.2)	

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; *M* = Mittelwert; *SA* = Standardabweichung. [a)] Ergebnis des Wilcoxon-Mann-Whitney-Tests. Dieser Test ist die non-parametrische Entsprechung zum T-Test für unabhängige Stichproben und wird angewandt, wenn wie im vorliegenden Fall die abhängige Variable nicht normalverteilt ist (starke Rechtsschiefe).

Bildungsniveau

Tabelle 37 zeigt den höchsten erreichten Bildungsabschluss der Probanden in der Gruppe SST, aufgeführt nach Abbrechern und Vollendern des MV. Exakte Tests nach Fisher zur Überprüfung auf Unterschiede zwischen Abbrechern und Vollendern führten zu nicht signifikanten Ergebnissen. Dies gilt sowohl in Bezug auf die einzelnen Bedingungen als auch in Bezug auf die Gesamtgruppe.

Tabelle 37. Bildungsniveau SST Abbrecher versus Vollender

			Höchster Bildungsabschluss				
SST Bedingung	Fehlende Werte [% (*N*)]	*N*	< 7 Jahre Schule [% (*N*)]	Abgeschl. Schulpflicht [% (*N*)]	Mind. Lehre [% (*N*)]	Nicht-Schweizer [% (*N*)]	*p*
ASAT alle zugeordneten Probanden	4.3% (*N*=2)	45	4.4% (*N2*=)	24.4% (*N*=11)	66.7% (*N*=30)	4.4% (*N*=2)	

ASAT Abbrecher	8.3% (*N*=1)	11	9.1% (*N*=1)	27.3% (*N*=3)	63.6% (*N*=7)	0% (*N*=0)	.827 [a)]
ASAT Vollender	2.9% (*N*=1)	34	2.94% (*N*=1)	23.5% (*N*=8)	67.7% (*N*=23)	5.9% (*N*=2)	
TAU alle zugeordneten Probanden	3.7% (*N*=1)	26	3.9% (*N*=1)	26.9% (*N*=7)	61.5% (*N*=16)	7.7% (*N*=2)	
TAU Abbrecher	0% (*N*=0)	10	0% (*N*=0)	20.0% (*N*=2)	70.0% (*N*=7)	10.0% (*N*=1)	.894 [a)]
TAU Vollender	5.9% (*N*=1)	16	6.3% (*N*=1)	31.3% (*N*=5)	56.3% (*N*=9)	6.3% (*N*=1)	
KG alle zugeordneten Probanden	9.1% (*N*=1)	10	0% (*N*=0)	30.0% (*N*=3)	60.0% (*N*=6)	10.0% (*N*=1)	
KG Abbrecher	12.5% (*N*=1)	7	0% (*N*=0)	28.6% (*N*=2)	57.1% (*N*=4)	14.3% (*N*=1)	1.000 [a)]
KG Vollender	0% (*N*=0)	3	0% (*N*=0)	33.3% (*N*=1)	66.7% (*N*=2)	0% (*N*=0)	
SST Gesamt alle Probanden	4.7% (*N*=4)	81	3.7% (*N*=3)	25.9% (*N*=21)	64.2% (*N*=52)	6.2% (*N*=5)	
SST Abbrecher	6.7% (*N*=2)	28	3.6% (*N*=1)	25.0% (*N*=7)	64.3% (*N*=18)	7.1% (*N*=2)	1.000 [a)]
SST Vollender	3.6% (*N*=2)	53	3.8% (*N*=2)	26.4% (*N*=14)	64.2% (*N*=34)	5.7% (*N*=3)	

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [a)] Ergebnis des exakten Tests nach Fisher. Dieser Test wird anstelle des Chi-Quadrat-Tests angewandt, wenn dessen Voraussetzung einer erwarteten Häufigkeit von ≥ 5 je Zelle verletzt ist, wie im vorliegenden Fall.

Psychiatrische Belastung: Unterschiede zwischen Abbrechern und Vollendern

Tabelle 38 zeigt den Anteil derjenigen Probanden, bei denen laut Gutachten die diagnostischen Kriterien einer Persönlichkeitsstörung (Diagnose aus ICD-10-Kategorie F6), einer Störung der ICD-10-Kategorie F2 (Schizophrenie, schizotype und wahnhafte Störungen), einer affektiven Störung (ICD-10 Kategorie F3) oder einer Störung durch psychotrope Substanzen (ICD-10 Kategorie F1) erfüllt sind. Da zu einem substanziellen Anteil der Probanden in der Kontrollgruppe keine Angaben zur psychiatrischen Belastung vorliegen, ist lediglich eine teilweise Auswertung möglich.

Sämtliche Gruppenvergleiche zwischen Abbrechern und Vollendern hinsichtlich der psychiatrischen Belastung, die aufgrund ausreichend vorliegender Angaben berechnet werden konnten, führten zu nicht signifikanten Ergebnissen. Dies gilt sowohl für den Vergleich zwischen Abbrechern und Vollendern über alle Bedingungen hinweg ($.065 \leq p \leq .624$) als auch für den Vergleich zwischen Abbrechern und Vollendern innerhalb der Bedingungen ($.091 \leq p \leq .756$).

Tabelle 38. Psychiatrische Vorbelastung SST Abbrecher versus Vollender

	Persönlichkeitsstörung (ICD-10 F6)				Schizophrenie (ICD-10 F2)				Affektive Störungen (ICD-10 F3)				Substanzstörung (ICD-10 F1)			
SST Bedingung	**Fehlende Werte [% (*N*)]**	***N***	**Anteil [% (*N*)]**	***p***	**Fehlende Werte [% (*N*)]**	***N***	**Anteil [% (*N*)]**	***p***	**Fehlende Werte [% (*N*)]**	***N***	**Anteil [% (*N*)]**	***p***	**Fehlende Werte [% (*N*)]**	***N***	**Anteil [% (*N*)]**	***p***
SST ASAT alle zugeordneten Probanden	10.6% (*N*=5)	42	54.8% (*N*=23)		10.6% (*N*=5)	42	4.8% (*N*=2)		10.6% (*N*=5)	42	7.1% (*N*=3)		8.5% (*N*=4)	43	25.6% (*N*=11)	
SST ASAT Abbrecher	16.7% (*N*=2)	10	40.0% (*N*=4)	.283 a)	8.3% (*N*=1)	11	9.1% (*N*=1)	.460 b)	8.3% (*N*=1)	11	18.2% (*N*=2)	.163 b)	8.3% (*N*=1)	11	45.5% (*N*=5)	.091 b)
SST ASAT Vollender	8.6% (*N*=3)	32	59.4% (*N*=19)		11.4% (*N*=4)	31	3.2% (*N*=1)		11.4% (*N*=4)	31	3.2% (*N*=1)		8.6% (*N*=3)	32	18.8% (*N*=6)	
SST TAU alle zugeordneten Probanden	3.7% (*N*=1)	26	46.2% (*N*=12)		3.7% (*N*=1)	26	7.7% (*N*=2)		3.7% (*N*=1)	26	3.8% (*N*=1)		3.7% (*N*=1)	26	15.4% (*N*=4)	
SST TAU Abbrecher	0% (*N*=0)	10	50.0% (*N*=5)	.756 a)	5.9% (*N*=1)	16	10.0% (*N*=1)	.631 b)	0% (*N*=0)	10	0% (*N*=0)	.615 b)	0% (*N*=0)	10	30.0% (*N*=3)	.142 b)
SST TAU Vollender	5.9% (*N*=1)	16	43.8% (*N*=7)		0% (*N*=0)	10	6.3% (*N*=1)		5.9% (*N*=1)	16	6.3% (*N*=1)		5.9% (*N*=1)	16	6.3% (*N*=1)	
SST KG alle zugeordneten Probanden	36.4% (*N*=4)	7	57.1% (*N*=4)		36.4% (*N*=4)	7	0% (*N*=0)		36.4% (*N*=4)	7	14.3% (*N*=1)		36.6% (*N*=4)	7	28.6% (*N*=2)	
SST KG Abbrecher	37.5% (*N*=3)	5	60.0% (*N*=3)	N/A	37.5% (*N*=3)	5	0% (*N*=0)	N/A	37.5% (*N*=3)	5	20.0% (*N*=1)	N/A	37.5% (*N*=3)	5	20.0% (*N*=1)	N/A
SST KG Vollender	33.3% (*N*=1)	2	50.0% (*N*=1)		33.3% (*N*=1)	2	0% (*N*=0)		33.3% (*N*=1)	2	0% (*N*=0)		33.3% (*N*=1)	2	50.0% (*N*=1)	
SST Gesamt alle zugeordne-ten Probanden	11.8% (*N*=10)	75	52.0% (*N*=39)		11.8% (*N*=10)	75	5.3% (*N*=4)		11.8% (*N*=10)	75	6.7% (*N*=5)		10.6% (*N*=9)	76	22.4% (*N*=17)	
SST Gesamt Abbrecher	16.7% (*N*=5)	25	48.0% (*N*=12)	.624 a)	13.3% (*N*=4)	26	7.7% (*N*=2)	.432 b)	13.3% (*N*=4)	26	11.5% (*N*=3)	.223 b)	13.3% (*N*=4)	26	34.6% (*N*=9)	.065 a)
SST Gesamt Vollender	9.1% (*N*=5)	50	54.0% (*N*=27)		10.9% (*N*=6)	49	4.1% (*N*=2)		10.9% (*N*=6)	49	4.1% (*N*=2)		9.1% (*N*=5)	50	16.0% (*N*=8)	

Anmerkungen. [a)] Ergebnis des Chi-Quadrat-Tests; [b)] Ergebnis des exakten Tests nach Fisher.

Summenwerte in Risk-Assessment-Instrumenten: Unterschiede zwischen Abbrechern und Vollender

Die angewandten Risk-Assessment-Instrumente PCL-R, VRAG und Static-99 zeichnen sich durch einen substanziellen Anteil an fehlenden Werten aus, sodass zu den vorbestehenden Unterschieden der Probanden hinsichtlich ihres (Baseline-) Risikos keine verlässliche Aussage getroffen werden kann. Für eine deskriptive Übersicht sind die Mittelwerte der Probanden in den drei Bedingungen der SST-Gruppe im Summenwert von PCL-R, VRAG und Static-99 dennoch in Tabelle 39 aufgeführt.

Tabelle 39. Summenwerte Risk-Assessment-Instrumente SST Abbrecher versus Vollender

	PCL-R-Summenwert			VRAG-Summenwert [1)]			Static-99 Summenwert		
SST Bedingung	**Fehlende Werte [% (*N*)]**	***N***	***M* (*SA*)**	**Fehlende Werte [% (*N*)]**	***N***	***M* (*SA*)**	**Fehlende Werte [% (*N*)]**	***N***	***M* (*SA*)**
SST ASAT alle Probanden	25.5% (*N*=12)	35	17.0 (8.2)	31.9% (*N*=15)	32	3.2 (11.7)	42.6% (*N*=20)	27	3.8 (2.1)
SST ASAT Abbrecher	41.7% (*N*=5)	7	16.5 (7.6)	16.7% (*N*=2)	10	4.5 (12.4)	33.3% (*N*=4)	8	2.5 (1.6)
SST ASAT Vollender	20.0% (*N*=7)	28	17.1 (8.5)	37.1% (*N*=13)	22	0.4 (10.1)	45.7% (*N*=16)	19	4.4 (2.0)
SST TAU alle Probanden	18.5% (*N*=5)	22	15.0 (7.5)	37.0% (*N*=10)	17	1.4 (7.4)	38.5% (*N*=10)	16	3.8 (2.3)
SST TAU Abbrecher	10.0% (*N*=1)	9	15.8 (8.4)	20.0% (*N*=2)	8	1.8 (7.2)	40.0% (*N*=4)	6	4.7 (2.3)
SST TAU Vollender	23.5% (*N*=4)	13	14.5 (7.1)	47.1% (*N*=8)	9	1.0 (8.0)	37.5% (*N*=6)	10	3.3 (2.2)
SST KG alle Probanden	63.6% (*N*=7)	4	16.0 (11.3)	54.6% (*N*=6)	5	4.2 (18.9)	60.0% (*N*=6)	4	2.8 (2.8)
SST KG Abbrecher	62.5% (*N*=5)	3	14.3 (13.3)	50.0% (*N*=4)	4	5.8 (21.5)	57.1% (*N*=4)	3	2.3 (3.2)
SST KG Vollender	66.7% (*N*=2)	1	21.1 (-)	66.7% (*N*=2)	1	-2.0 (-)	66.7% (*N*=2)	1	4.0 (-)
SST Gesamt alle Probanden	28.2% (*N*=24)	61	16.2 (8.1)	36.5% (*N*=31)	54	2.7 (11.2)	43.5% (*N*=37)	48	3.8 (2.1)
SST Gesamt Abbrecher	36.7% (*N*=11)	19	15.8 (8.4)	26.7% (*N*=8)	22	1.9 (11.4)	40.0% (*N*=12)	18	3.3 (2.3)
SST Gesamt Vollender	23.6% (*N*=13)	42	16.4 (8.0)	41.8% (*N*=23)	32	3.3 (11.1)	45.5% (*N*=25)	30	4.0 (2.1)

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [1)] Hierbei wurde überprüft, ob der VRAG-Summenwert bereits korrigiert wurde (dies ist bis zu einer maximalen Anzahl von vier fehlenden Angaben je Bewertung möglich). Dies ist der Fall: Der Anteil an fehlenden Werten im Item „vragtot“ ist identisch mit dem Anteil derjenigen Fälle, die in mehr als vier Items des VRAG fehlende Angaben haben.

Therapieerfahrung: Unterschiede zwischen Abbrechern und Vollendern

Tabelle 40 zeigt den Anteil der Probanden aus der SST-Gruppe, die bereits vor Beginn des MV mindestens einmal in psychotherapeutischer Behandlung waren, separat dargestellt für Abbrecher und Vollender. Chi-Quadrat-Tests bzw. ein exakter Test nach Fisher ergaben keine Hinweise auf Unterschiede zwischen Probanden, die den MV vorzeitig abbrachen gegenüber Probanden, die bis zum Ende an der Studie teilnahmen. Dies gilt sowohl für die Gesamtgruppe als auch innerhalb der Bedingungen ASAT und TAU. Für die Kontrollgruppe kann diesbezüglich aufgrund des hohen Anteils fehlender Werte und der ohnehin kleinen Fallzahlen keine Aussage getroffen werden.

Tabelle 40. Therapieerfahrung SST Abbrecher versus Vollender

Bedingung	Fehlende Werte [% (*N*)]	*N*	Psychotherapie-Erfahrung [% (*N*)]	*p*
SST ASAT alle zugeordneten Probanden	17.0% (*N*=8)	39	82.1% (*N*=32)	
SST ASAT Abbrecher	8.3% (*N*=1)	11	90.9% (*N*=10)	.346 [a)]
SST ASAT Vollender	20.0% (*N*=7)	28	78.6% (*N*=22)	
SST TAU alle zugeordneten Probanden	11.1% (*N*=3)	24	58.3% (*N*=14)	
SST TAU Abbrecher	0% (*N*=0)	10	50.0% (*N*=5)	.484 [b)]
SST TAU Vollender	17.7% (*N*=3)	14	64.3% (*N*=9)	
SST KG alle zugeordneten Probanden	45.5% (*N*=5)	6	66.7% (*N*=4)	
SST KG Abbrecher	37.5% (*N*=3)	5	80.0% (*N*=4)	*N/A*
SST KG Vollender	66.7% (*N*=2)	1	0% (*N*=0)	
SST Gesamt alle Probanden	18.8% (*N*=16)	69	72.5% (*N*=50)	
SST Gesamt Abbrecher	13.3% (*N*=4)	26	73.1% (*N*=19)	.929 [b)]
SST Gesamt Vollender	21.8% (*N*=12)	43	72.1% (*N*=31)	

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; [a)] Ergebnis des exakten Tests nach Fisher. Dieser Test wird anstelle des Chi-Quadrat-Tests angewandt, wenn dessen Voraussetzung einer erwarteten Häufigkeit von ≥ 5 je Zelle verletzt ist; [b)] Ergebnis des Chi-Quadrat-Tests; Rot = mehr als 20% fehlende Werte. Im Falle der Gesamtgruppe wurde in diesem Fall eine Berechnung durchgeführt, da der Anteil fehlender Werte bei den Vollendern hier nur knapp über 20% liegt und bezogen auf alle Probanden der SST-Gruppe der festgelegte Grenzwert unterschritten bleibt.

Rückfälligkeit SST

Im Folgenden wird die Analyse der Rückfälligkeit in der SST-Gruppe analog zum Vorgehen bezüglich der GST-Gruppe auf zwei verschiedene Arten präsentiert: Zunächst werden dabei sämtliche Probanden einschließlich der Abbrecher berücksichtigt (Intent-to-Treat-Analyse). Anschließend wird die Rückfälligkeitsanalyse ausschließlich für diejenigen Probanden präsentiert, die den MV bis zum Ende durchliefen (Vollender).

Rückfälligkeit SST: Intent-to-Treat-Analyse

In der Literatur gibt es klare Hinweise auf ein höheres Rückfallrisiko für Studienteilnehmer, die eine Behandlung vorzeitig abbrechen, im Vergleich zu Teilnehmern, die das Behandlungsprogramm bis zu Ende durchlaufen (z.B. Hanson et al., 2002). Ein Ausschluss von Abbrechern würde somit das Risiko erhöhen, den Behandlungseffekt zu überschätzen. Zudem hat auch ein Teil der Abbrecher die Behandlung erhalten – wenn auch unterschiedlich lange, und mit der Einschränkung, dass der Zeitpunkt und genaue Grund des Ausscheidens in der vorliegenden Evaluation nicht ausgewertet werden konnten. Der Einschluss von ausgeschiedenen Studienteilnehmern in die Wirksamkeitsanalyse, also eine Intent-to-Treat-Analyse, ist daher erstrebenswert: Ein solches Vorgehen hat das geringste Risiko für Verzerrungen (z.B. Beech et al., 2007). In großen Meta-Analysen über Behandlungsprogramme für Sexualstraftäter werden auch die Abbrecher in die Auswertung mit eingeschlossen, wo immer das möglich ist (z.B. Hanson, Bourgon, Helmus, & Hodgson, 2009; Lösel & Schmucker, 2005; Schmucker & Lösel, 2015). Aus den genannten Gründen wurden auch die Abbrecher der SST-Gruppe mit in die Analyse zur Rückfälligkeit eingeschlossen. Die Ergebnisse sind in Tabelle 41 aufgeführt. Unter diesem Vorgehen befanden sich zum Zeitpunkt des Strafregisterauszugs neun der berücksichtigten 85 Probanden in Freiheit (10.6%), davon zwei Probanden aus der Experimental-, sechs aus der Vergleichs- und einer aus der Kontrollgruppe. Von diesen neun Probanden, die einer Time at Risk von durchschnittlich knapp zwei Jahren (*M* = 23.1 Monate; *SA* = 13.9; Range: 1.1 – 42.2 Monate) ausgesetzt waren, wurde ein Proband aus der Vergleichsgruppe rückfällig, was einer Rückfälligkeit von 16.7% innerhalb der Vergleichsgruppe entspricht (bzw. einer Rückfälligkeit von 11.1% bezogen auf die Gesamtgruppe der in Freiheit entlassenen Probanden der SST-Gruppe). Dieser Proband wurde wegen Verstoßes gegen Artikel 186 und 139 des StGB verurteilt (Hausfriedensbruch und Diebstahl). Ein Kruskal-Wallis-H-Test ergab keine Hinweise auf Unterschiede in der Time at risk zwischen den drei Bedingungen ($\chi^2(2) = 1.489$; p=.475).

Tabelle 41. Rückfälligkeit Gruppe SST (Intent-to-Treat)

SST Bedingung	Fehlend [% (*N*)]	*N*	TAR [Monate] [*M* (*SA*)]	*p*	Rückfällig [% (*N*)]	*p*	Rückfällig (Gewalt- oder Sexualdelikt) [% (*N*)]	*p*
SST ASAT	0% (*N*=0)	2	*M*=21.0 (*SA*=11.7)		0% (*N*=0)		0% (*N*=0)	
SST TAU	0% (*N*=0)	6	*M*=21.1 (*SA*=14.9)	.475 [a)]	16.7% (*N*=1)	N/A	0% (*N*=0)	N/A
SST KG	0% (*N*=0)	1	*M*=39.3 (*SA*=0)		0% (*N*=0)		0% (*N*=0)	
SST Gesamt	**0% (*N*=0)**	**9**	***M*=23.1 (*SA*=13.9)**		**11.1% (*N*=1)**		**0% (*N*=0)**	

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; TAR: Time at Risk; [a)] Ergebnis eines Kruskal-Wallis-H-Tests.

Rückfälligkeit SST: Nur Vollender

Tabelle 42 zeigt die Rückfälligkeit derjenigen Probanden der SST-Gruppe, die den MV bis zum Ende durchlaufen haben. Nur insgesamt sechs der 55 Probanden der SST-Gruppe, die den MV bis zum Ende absolviert haben (10.9%), befanden sich zum Zeitpunkt des Strafregisterauszugs in Freiheit (gemäß Variable „austrittsdatum“). Davon stammten zwei Probanden aus der Experimental- und vier aus der Vergleichsgruppe. Keiner dieser Probanden wurde innerhalb des zur Verfügung stehenden Follow-Up-Zeitraums von durchschnittlich etwa 1.5 Jahren (*M* = 18.4 Monate; *SA* = 12.2; Range: 1.1 – 35.0 Monate) rückfällig.

Tabelle 42. Rückfälligkeit Gruppe SST (in Freiheit entlassene Vollender)

SST Bedingung	Fehlend [% (*N*)]	*N*	TAR [Monate] [*M* (*SA*)]	*p*	Rückfällig [% (*N*)]	*p*	Rückfällig (Gewalt- oder Sexualdelikt) [% (*N*)]	*p*
SST ASAT	0% (*N*=0)	2	*M*=21.0 (*SA*=11.7)		0% (*N*=0)		0% (*N*=0)	
SST TAU	0% (*N*=0)	4	*M*=17.1 (*SA*=13.9)	N/A	0% (*N*=0)	N/A	0% (*N*=0)	N/A
SST KG	0% (*N*=0)	0	N/A		N/A		N/A	
SST Gesamt	**0% (*N*=0)**	**6**	***M*=18.4 (*SA*=12.2)**		**0% (*N*=0)**		**0% (*N*=0)**	

Anmerkungen. SST: Gruppe der Sexualstraftäter; ASAT: Anti-Sexuelle-Aggressivität-Training®Suisse; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; TAR: Time at Risk

Diskussion

Wirksamkeit forensischer Interventionen

Allgemeines Ziel forensischer Behandlungsprogramme ist eine Senkung des Rückfallrisikos. Eine hohe Wirksamkeit von Interventionen konnte vor allem für kognitiv-behaviorale Methoden gezeigt werden, die multimodal und hoch strukturiert angelegt sind und die auf einem wissenschaftlich fundierten Modell aufbauen (Lipsey & Cullen, 2007; Lösel & Schmucker, 2005). Im Bereich jugendlicher Sexualstraftäter konnte für die multisystemische Therapie (MST) eine besonders hohe Wirksamkeit nachgewiesen werden (Borduin, Henggeler, Blaske, & Stein, 1990; Borduin, Schaeffer, & Heiblum, 2009; vgl. Schmucker & Lösel, 2015).
Hinsichtlich der Wirksamkeit von forensischen Interventionen zeigt sich in der Literatur ein allgemeiner Trend: Lipsey und Cullen (2007) veröffentlichten eine systematische Review über alle Meta-Analysen, die den Effekt von Interventionen verschiedener Art auf die Rückfallraten von Straftätern untersuchten. Von den 18 Meta-Analysen, die sich eindeutig einem bestimmten Interventionstyp (strafend, unterstützend, therapeutisch) zuordnen lassen, untersuchten sieben den Effekt von strafenden und elf den Effekt von rehabilitativen bzw. therapeutischen Interventionen bei jugendlichen oder erwachsenen Straftätern.
Hierbei zeigten alle Meta-Analysen, die den Effekt von therapeutischen Interventionen untersuchten, eine Reduktion der Rückfallrate, im Mittel um 28%. Hingegen zeigte sich, dass durch sechs der sieben strafenden Interventionen das Rückfallrisiko nicht reduziert werden konnte (n=2 Meta-Analysen) bzw. sogar gesteigert wurde (n=4 Meta-Analysen; +4 bis +14%). Die einzige strafende Intervention, die mit einer Reduktion des Rückfallrisikos einherging (Mittlere Sanktionen; -2%), war der therapeutischen Intervention mit dem geringsten Effekt (Kognitive Verhaltenstherapie; -8%) deutlich unterlegen.
Die Meta-Analyse, die mit einer Senkung der Rückfallrate um 60% den deutlichsten Effekt von Therapien zeigte (Andrews et al., 1990) führte zur Entwicklung des Risk-Need-Responsivity-Modells (RNR). Dabei handelt es sich um ein heute allgemein anerkanntes Prinzip forensischer Behandlungsprogramme. Das Risk-Prinzip besagt, dass die Intensität der Intervention dem (valide erfassten) individuellen Rückfallrisiko entsprechen sollte. Nach dem Need-Prinzip hat eine Intervention einen größeren Effekt, wenn das individuelle kriminogene Bedürfnis hinsichtlich risikorelevanter Eigenschaften des Straftäters berücksichtigt wird. Das Responsivity- (Ansprechbarkeits-) Prinzip verlangt eine Passung der Intervention zu den Fähigkeiten und zum Lernstil des Straftäters. Insgesamt wird in der Literatur die Annahme gestützt, dass eine Berücksichtigung des RNR-Modells mit einer deutlichen Senkung des Rückfallrisikos einhergeht (Bonta & Andrews, 2016; Endrass, Rossegger, & Braunschweig, 2012).
Eine Meta-Analyse zur Wirksamkeit von Behandlungsprogrammen bei jungen Straftätern unter 25 Jahren (Koehler, Lösel, Akoensi, & Humphreys, 2013) führte zu vergleichbaren Ergebnissen: Im Mittel zeigten Therapien einen positiven Effekt in Bezug

auf die Rückfälligkeit, (kognitiv-) behaviorale Therapien lagen über dem Durchschnitt des Effekts aller Interventionen und rein strafende oder überwachende Interventionen zeigten eine schädliche Wirkung. Therapieprogramme, die in Übereinstimmung mit dem RNR-Modell durchgeführt wurden, zeigten im Mittel die größten Effekte (Koehler et al., 2013).

Wirksamkeit von Psychotherapien für Gewalt- und Sexualstraftäter

Schwere Rückfälle, die durch Gewalt- und Sexualstraftäter verübt werden, ziehen in besonderem Maße die öffentliche Aufmerksamkeit auf sich. Die Verhinderung von Rückfällen steht daher im Interesse von Entscheidungsträgern und Experten aus juristischen und medizinisch-psychologischen Fachgebieten. Gerade in den letzten Jahren ist das öffentliche Bewusstsein für Sexualstraftaten gewachsen. Dies geht damit einher, dass die Wirksamkeit von Therapien für Sexualstraftäter insgesamt deutlich umfangreicher überprüft ist als im Bereich von Gewaltstraftätern (z.B. Hanson et al., 2009; Lipsey & Cullen, 2007; Marques, Wiederanders, Day, Nelson, & Van Ommeren, 2005; Wößner & Schwedler, 2014).
So konnten Hanson et al. (2009) in ihre Meta-Analyse 129 Studien einschließen, die den Behandlungseffekt bei heranwachsenden oder erwachsenen Sexualstraftätern untersuchten. Hanson et al. (2009) nahmen dabei eine Bewertung der methodischen Qualität der Studien nach den Richtlinien des „Collaborative Outcome Data Committee“ (CODC; Beech et al., 2007) vor. Die Richtlinien wurden explizit für die Bewertung der Qualität von Evaluationsstudien zur Behandlung von Sexualstraftätern entwickelt. Die insgesamt 21 Items sind in sieben Bereiche geordnet: a) Kontrolle der unabhängigen Variablen (d.h. Inhalt des spezifischen Behandlungsprogramms und Programmintegrität sowohl für Interventions- als auch Vergleichsgruppe); b) Erwartungen des Versuchsleiters; c) Stichprobengröße; d) Stichprobenschwund; e) Vergleichbarkeit der Gruppen; f) Outcome-Variablen; g) durchgeführte Vergleichstests; sowie h) (nur bei Institutionen übergreifenden Versuchsplänen) Stichprobengröße in den Institutionen. Aus diesen Merkmalen werden die Kategorien „strong“, „good“, „weak“ und „rejected“ abgeleitet. Sie unterscheiden sich letztlich im Ausmaß an potenziellen Verzerrungen (Beech et al., 2007).
Von den 129 Studien wurden 105 (81%) aufgrund ungenügender methodischer Qualität („rejected“) aus der Meta-Analyse von Hanson et al. (2009) ausgeschlossen. Eingeschlossen wurden 18 als „weak“ und 5 als „good“ klassifizierte Studien. Die einzige Studie, welche die höchste Stufe erreichte, musste ausgeschlossen werden, weil das untersuchte Klientel aus Jungen mit sexuell aufdringlichem Verhalten bestand und somit qualitativ nicht vergleichbar war.
Die konkrete Wirksamkeit der Behandlung in den nach Hanson et al. (2009) „guten“ Studien konnte kaum nachgewiesen werden. Ausnahme davon stellte die Multisystemische Therapie (MST) bei jugendlichen Sexualstraftätern dar, die sich als deutlich überlegen gegenüber der Kontrollgruppe erwies, die ambulante kognitive Verhaltenstherapie erhalten hatte (Borduin et al., 1990; Borduin et al., 2009).
Die Prinzipien des RNR-Modells sind auch in der Behandlung von Gewalt- und Sexualstraftätern allgemein anerkannt (Bonta & Andrews, 2016).

Allgemein zum MV

Einschätzung der methodischen Qualität auf der Maryland Scientific Methods Scale

Zur Einschätzung der internen Validität und zur Klassifikation von Versuchsdesigns auf einer fünfstufigen Skala wurde die sogenannte *Maryland Scientific Methods Scale* (MSMS) entwickelt (Sherman et al., 1998).
In der konkreten Anwendung bei der Wirksamkeitsüberprüfung von forensischen Behandlungsprogrammen erreicht eine Studie die Stufe 1 der MSMS bei einem rein korrelativen Design, das die Wiederverurteilungsrate von Probanden einer Interventionsgruppe erfasst. Stufe 2 wird erreicht, wenn entweder die beobachtete mit der vorhergesagten Wiederverurteilungsrate einer Behandlungsgruppe verglichen wird oder die Rate mit der einer Kontrollgruppe verglichen wird, welche nicht äquivalent zur Behandlungsgruppe ist. Stufe 3 liegt bei einem Vergleich der Wiederverurteilungsrate vor, wenn eine äquivalente Ausprägung relevanter Faktoren bei Behandlungs- und Kontrollgruppe angenommen wird. Stufe 4 auf der MSMS wird erreicht, wenn die Wiederverurteilungsrate zwischen Behandlungs- und Kontrollgruppe verglichen wird und hierbei ein individuelles Matching der Probanden aufgrund theoretisch fundierter Faktoren oder aber eine statistische Kontrolle dieser Faktoren stattfindet. Stufe 5 wird nur bei einem vollständig randomisierten Kontrollgruppendesign (RCT) erreicht (Friendship, Street, Cann, & Harper, 2005; Hollin, 2008; Schmucker & Lösel, 2015).

Ein solches RCT-Design wird häufig als „Gold-Standard“ für die Evaluation medizinischer, psychiatrischer und auch psychotherapeutischer Interventionen betrachtet. Dies ist in der hohen internen Validität solcher Designs begründet. Allerdings wird deren Nutzen in forensischen Kontext kritisch diskutiert (Farrington & Welsh, 2005; Hollin, 2008; Marshall & Marshall, 2007): Auf der einen Seite garantiert eine randomisierte Gruppenzuweisung selbst bei Matching mehrerer relevanter Merkmale keine Vergleichbarkeit in weiteren zentralen Variablen, die stark mit dem Effektkriterium korrelieren (siehe Marques et al., 2005). Auf der anderen Seite kann ein RCT-Design kaum den tatsächlichen Bedingungen und häufigen Umständen im Strafvollzug gerecht werden (Hollin, 2008), womit keine hinreichende externe Validität, d.h. keine Übertragbarkeit auf andere Kontexte, erreicht wird. In der Forschungspraxis der forensischen Psychologie ist hier eine negative Korrelation von interner und externer Validität festzustellen, was ein gewichtiges Argument gegen RCT-Designs als Goldstandard darstellt (Endrass, Rossegger, & Braunschweig, 2012; Hollin, 2008). Hollin (2008) führt ein weiteres Argument auf, das zugleich die Wahl eines Quasi-experimentellen Designs aufwertet: Implizit wird häufig angenommen, dass aufgrund von Verzerrungs-Effekten in quasi-experimentellen Studien insbesondere die Gefahr besteht, das Behandlungseffekte überschätzt werden. Ergebnisse verschiedener Meta-Analysen, die den Zusammenhang zwischen der Qualität des Untersuchungsdesigns und der Kriteriumsvariablen untersucht haben, zeigten allerdings, dass in quasi-experimentellen Studien keine größeren Therapie-Effekte erreicht werden (Hollin, 2008). In einer weiteren Studie zum Vergleich von RCTs und quasiexperimentellen Designs konnte zwar gezeigt werden, dass RCTs im Mittel geringere Behandlungseffekte aufweisen; gleichzeitig korrelierten jedoch die Effekte beider Unter-

suchungsdesigns hoch positiv miteinander. Außerdem konnte in dieser Untersuchung gezeigt werden, dass der Unterschied zwischen RCTs und quasiexperimentellen Studien nur noch gering war, wenn ein prospektives Design angewendet wurden (Ioannidis et al., 2001). Weitere systematische Reviews und Meta-Analysen zur Wirksamkeit forensischer Interventionen, die sowohl RCTs als auch quasi-experimentelle Studien einschlossen, ergaben, dass das Studiendesign wenig bis gar keinen Einfluss auf die Höhe der berichteten Effektgrößen einer Behandlung hatte (Babcock, Green, & Robie, 2004; Lipsey & Cullen, 2007; Lösel & Schmucker, 2005; Schmucker & Lösel, 2015).

Trotz der diskutierten Schwächen erscheint die Anwendung eines quasi-experimentellen Designs zur Wirksamkeitsüberprüfung der Behandlung von Straftätern zweckmäßig, sofern einem möglichen Selektions-Bias durch geeignete statistische Methoden begegnet wird.

Ab Stufe 3 der MSMS kann vom Vorliegen einer hinreichend gewährleisteten Kontrolle der Vergleichbarkeit von Behandlungs- und Kontrollgruppe gesprochen werden (Lösel & Schmucker, 2005; Schmucker & Lösel, 2015). Diese Stufe erreichten jedoch nur 40% der Vergleichsstudien, die in die Metanalyse zur Wirksamkeit von Sexualstraftäter-Behandlungen eingeschlossen worden waren (Lösel & Schmucker, 2005). Für quasi-experimentelle Untersuchungen wie der vorliegenden wird empfohlen, Stufe 4 der MSMS anzustreben. In Bezug auf die vorliegende Studie ist zu diskutieren, ob diese eher Stufe 2 oder Stufe 3 der MSMS entspricht: Von einem individuellen Matching der Probanden in den drei Bedingungen musste aufgrund der schwierigen Rekrutierungsbedingungen zwar abgesehen werden, jedoch fand eine Kontrolle relevanter und theoretisch fundierter Variablen statt, die in Zusammenhang mit dem Rückfallrisiko stehen und somit einen Einfluss auf die Effektkriterien des MV haben können: Demografische Daten wie Alter, Nationalität und Bildungsniveau sowie, mit Einschränkung (aufgrund eines substanziellen Anteils fehlender Werte), die psychiatrische Vorbelastung, Werte in verschiedenen Risk-Assessment-Instrumenten sowie die Therapieerfahrung. Hierbei zeigten sich teilweise Unterschiede zwischen den Probanden der drei Bedingungen in den Vergleichsvariablen: So verfügten die Probanden in den beiden Experimentalbedingungen (GST R&R und SST ASAT) über mehr Therapieerfahrung als die Probanden in den Vergleichsgruppen, worauf im Folgenden noch näher eingegangen wird. In der GST-Gruppe war zudem ein höherer Ausländeranteil in der Kontrollgruppe im Vergleich zu den beiden Therapiebedingungen zu verzeichnen, was eine sehr relevante Variable darstellt. Dieser unterschiedlich hohe Ausländeranteil hatte wiederum Einfluss auf das Bildungsniveau, welches in der Experimentalgruppe höher war als in der Vergleichsgruppe. Keine Hinweise auf Unterschiede zwischen den Bedingungen der GST-Gruppe ergaben die durchgeführten Signifikanztests hinsichtlich folgender Faktoren: Alter, das um die Nationalität korrigierte Bildungsniveau und, soweit überprüfbar, die psychiatrische Belastung. Insgesamt entspräche der MV in Bezug auf die GST-Gruppe in Anbetracht der letztlich gefundenen Unterschiede zwischen den Bedingungen somit am ehesten Stufe 2 der MSMS.

Zwischen den drei Bedingungen der SST-Gruppe wurden keine Hinweise auf Unterschiede gefunden hinsichtlich der Faktoren Nationalität, Alter, Bildungsniveau, Vorstrafenbelastung sowie psychiatrische Belastung (letzteres konnte nur für die beiden Behandlungsgruppen SST ASAT und SST TAU überprüft werden). Da sich die Be-

dingungen der SST-Gruppe einzig in der Therapieerfahrung signifikant voneinander unterscheiden, kann hier von Stufe 3 der MSMS gesprochen werden.

Legt man die CODC-Richtlinien für die Qualitätsbeurteilung von Therapieprogrammen für Sexualstraftäter zugrunde (Beech et al., 2007; siehe oben), so können bezüglich der SST-Gruppe die Kriterien nicht beurteilt bzw. kaum als erfüllt betrachtet werden. Ein solches Nicht-Erreichen der Kriterien ist jedoch Standard in diesem Forschungsgebiet: So mussten auch vier von fünf Studien, die in der Metaanalyse von Hanson et al. (2009) gesichtet wurden, als „rejected" klassifiziert werden. Damit spiegelt der MV im Bereich der Evaluation von Therapien für Sexualstraftäter letztlich die Realität wider.

Sicherstellung der Programmintegrität (Treatment Integrity)

Die Programmintegrität ist ein zentrales Qualitätskriterium bei der Evaluation eines Behandlungsprogramms. Mangelhafte Programmintegrität wird insbesondere als einer der Hauptgründe für schwache Evaluationsergebnisse des R&R-Programms genannt (Antonowicz & Parker, 2012).
Andrews und Dowden (2005) veröffentlichten im Rahmen einer metaanalytischen Review zehn Kriterien, die für eine hohe Programmintegrität im forensischen Setting stehen. Diese Kriterien werden im Folgenden in Bezug auf die Durchführung von R&R bzw. R&R2 und ASAT®Suisse im vorliegenden MV besprochen.

1. *Spezifisches Modell.* Sowohl R&R(2) als auch ASAT®(Suisse) beruhen auf einem spezifischen Modell.
2. *Auswahl der Behandler.* Mindestens eine Person der Behandlungsteams hatte eine Ausbildung als Psychotherapeut/in absolviert. Zu Charakteristika der Behandler hinsichtlich Berufserfahrung und Erfahrung in der Anwendung der Behandlungsprogramme lässt sich keine Aussage treffen.
3. *Training der Behandler*. Sämtliche Behandler, die das R&R(2) oder das ASAT®Suisse durchführten, hatten die offizielle Schulung durchlaufen und waren dadurch für die Anwendung des jeweiligen Behandlungsprogramms zertifiziert.
4. *Klinische Supervision*. Es wurde im Rahmen des MV nicht erhoben, ob sich sämtliche Behandler in regelmäßiger klinischer Supervision befinden, auch wenn dies die Regel in den teilnehmenden Institutionen darstellt. Eine spezielle Supervision hinsichtlich der Anwendung von R&R(2) oder ASAT®Suisse gibt es nicht. Das Kriterium muss als nicht erfüllt betrachtet werden.
5. *Trainings-Manuale*. Sowohl R&R(2) als auch ASAT®(Suisse) sind strikt manualgebunden. Das Manual für das R&R-Programm ist ausschließlich im Rahmen von Zertifizierungskursen erhältlich (Eucker, 2013); ebenso ist das ASAT®-Manual nicht frei erhältlich.
6. *Monitoring des Behandlungs-Prozesses*. Zum Monitoring des Behandlungsprozesses sind keine Angaben möglich. Das Kriterium muss als nicht erfüllt betrachtet werden.
7. *Angemessene Therapiedosis*. Die Therapiedosis (Intensität) der beiden Behandlungsprogramme entspricht den Vorgaben der Manuale. Die Einhaltung der Manuale wurde abschließend nicht kontrolliert. Inwiefern eine Anpassung

der Therapiedosis an das Bedürfnis des jeweiligen Probanden stattfand, ist fraglich. Eine manualgetreue, für alle Probanden identische Therapiedosis steht einer solchen Forderung nach Berücksichtigung der individuellen kriminogenen Needs (vgl. RNR-Prinzip; Bonta & Andrews, 2016) jedoch eher entgegen. Insofern muss dieses Kriterium eher als nicht erfüllt betrachtet werden.

8. *Neuigkeit des Programms*. Für beide Programme kann von einem gewissen Neuigkeitswert ausgegangen werden, da zumindest noch keine publizierten Forschungsergebnisse betreffend ihres Einsatzes in der Schweiz existieren (FPD Bern, 2009). Nicht auszuschließen ist, dass die Programme ungeachtet ausstehender Evaluationsergebnisse an anderer Stelle in der Praxis des Schweizer Straf- und Maßnahmenvollzugs angewandt werden.
9. *Kleine Stichprobe*. Andrews und Dowden (2005) fordern für die Behandlungsgruppe eine Stichprobengröße von unter 100. (Dies liegt im Zusammenhang der Stichprobengröße mit der zu erwartenden Effektgröße begründet.) Diese Forderung ist für ASAT®Suisse erfüllt, die Stichprobengröße in der R&R(2)-Experimentalgruppe hingegen ist größer.
10. *Evaluation*. Ein Evaluator war ins Studiendesign und in die Begleitung des MV involviert.

Insgesamt wurden einige Anstrengungen ersichtlich, die im Rahmend des MV ergriffen wurden, um eine hohe Treatment Integrity sicherzustellen. Letztlich können die Kriterien nach Andrews und Dowden (2005) etwa zur Hälfte als erfüllt betrachtet werden: Vier der Kriterien können in vollem Umfang als erfüllt bezeichnet werden (1., 3., 5. und 10.), ein weiteres Kriterium mit einer gewissen Sicherheit (8.). Zu einem gewissen Anteil kann Kriterium 2 als erfüllt interpretiert werden (kontrollierte berufliche Ausbildung der ausgewählten Behandler). Ein sechstes Kriterium ist ausschließlich hinsichtlich des ASAT®Suisse erfüllt (9.).

Behandlungsstatus der Kontrollgruppen und Vorbehandlungen

– Es gibt einen großen Anteil an Probanden in allen Bedingungen, die bereits vor Beginn der Erhebung an mindestens einer Behandlungsmaßnahme teilgenommen hatten, d.h. auch die Kontrollgruppen können nicht als vollkommen unbehandelt gelten. Dies wurde versucht als Hilfsvariable zu erfassen, was jedoch nicht als zuverlässig zu betrachten ist, gerade falls diese auf Selbstauskünften der Probanden beruht und die Behandlung in der Vergangenheit stattfand. Die Hilfsvariable hat vor allem in der Kontrollgruppe einen hohen Anteil an fehlenden Werten, sodass eine Vergleichbarkeit schwierig ist.
– Da viele der forensischen Behandlungsformen (z.B. Suchttherapie, gezielte Behandlung spezifischer psychischer Störungen etc.) ähnliche Therapieziele anstreben, ist nicht auszuschließen, dass Probanden, die zuvor an diesen Interventionen teilgenommen hatten, bereits schon mehr Erfahrung mit den durch das Programm R&R(2) vermittelten Fertigkeitsbereichen gesammelt hatten und vergleichsweise weniger neue Fertigkeiten erwerben konnten. Dies erscheint insbesondere relevant, da die Probanden der Experimentalgruppen (GST R&R und SST ASAT) zu einem signifikant höheren Anteil über Psychotherapieerfahrung verfügen als die

Probanden der Vergleichsgruppe. Im Fall der GST-Gruppe gilt diese Einschränkung in Bezug auf sämtliche Effektkriterien (Fragestellung 1 und Fragestellung 2).
- Zudem kann nicht ausgeschlossen werden, dass während der Studiendauer in den Kontrollgruppen Einzelgespräche mit Vollzugsbediensteten in der KG als „inoffizielle“ Therapie stattfanden, die so nicht in der Akte vermerkt wurden.
- Für mögliche Behandlungen, die nach dem Ende des MV stattfanden und die somit Einfluss auf das Effektkriterium der Rückfälligkeit haben können, konnte nicht kontrolliert werden.
- Gering ausgeprägte Unterschiede zwischen den Untersuchungsgruppen können daher durchaus auch mit der vorausgegangenen Behandlung der Probanden im Sinne eines kumulierten Effekts der früheren Therapieerfahrung konfundiert sein. Andererseits könnten beobachtete Unterschiede zwischen der R&R(2)-Gruppe und der „unbehandelten“ Kontrollgruppe bei Kontrolle der bisherigen Therapieerfahrung für einen noch größeren zusätzlichen Nutzen des Programms in Bezug auf die Therapieziele sprechen.

Statistische Power (Teststärke)

Die Power ist ein Maß für die Aussagekraft eines statistischen Tests. Sie ist definiert als 1 – ß-Fehler. Das heißt, sie gibt an, mit welcher Wahrscheinlichkeit ein Fehler 2. Art vermieden wurde bzw. mit welcher Wahrscheinlichkeit ein Signifikanztest zugunsten der Alternativhypothese H1 entscheidet, falls diese Alternativhypothese korrekt ist.
Der vorliegende Modellversuch hat mehrere Limitationen hinsichtlich der erreichten Power. Zum einen gehen kleine Fallzahlen mit einer geringen Power einher. Weitere Einflussfaktoren auf die Power sind das gewählte Alpha-Niveau, die Effektstärke und die Art der verwendeten Testverfahren.
Zu diskutieren ist die erreichte Power insbesondere in Bezug auf die Analysen zur GST-Gruppe (die beabsichtigten inferenzstatistischen Analysen innerhalb der SST-Gruppe konnten aufgrund der geringen Fallzahlen nicht durchgeführt werden). Zu diesem Zweck wurde post hoc eine Power-Analyse mit dem Statistik-Programm G*Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) durchgeführt. Als fixe Größen angenommen wurden das Alpha-Niveau von .05 sowie eine Stichprobengröße von insgesamt N=286, was der maximalen Anzahl an zur Verfügung stehenden Probanden in der GST-Gruppe entspricht, sowie eine einfaktorielle ANOVA als statistisches Testverfahren. Variiert wurde die Effektgröße. Bei einem angenommenen kleinen Effekt von .10 beträgt die so berechnete Power .305. Wird eine Effektstärke von .05 angenommen, was dem größten Effekt im Rahmen der berichteten signifikanten Ergebnisse entspricht, reduziert sich die erreichte Power auf .107. Gerade unter der Einschränkung, dass diese Berechnungen sehr „wohlwollend“ durchgeführt wurden und die Power damit eher überschätzt wird, ist die im Rahmen des Modellversuchs erreichte Power als sehr gering einzuschätzen: Die Wahrscheinlichkeit, tatsächlich existierende Unterschiede *nicht* aufdecken zu können (Fehler 2. Art) beträgt in den beiden beispielhaften Berechnungen 69.5% bzw. 89.3%. Dies hat folgende Implikationen für die Interpretation der Ergebnisse: Unter den genannten Bedingungen des Modellversuchs ist von einer massiv eingeschränkten Fähigkeit auszugehen, real existierende Unterschiede statistisch überhaupt

nachweisen zu können. Dies ist insbesondere angesichts der zahlreichen berichteten Nullresultate zu bedenken. Darüber hinaus kann eine geringe Power bei einem ausschließlichen Fokus auf statistisch signifikante Ergebnisse, was dem wissenschaftlichen Standard entspricht und auch im vorliegenden Bericht der Fall ist, zu einer Überschätzung von Effektstärken führen, aufgrund der reduzierten Wahrscheinlichkeit bei Vorliegen kleiner Effektstärken in Verbindung mit geringen Fallzahlen ein statistisch signifikantes Ergebnis zu erhalten. Da die berichteten Effektstärken jedoch ohnehin allesamt als klein zu interpretieren sind, scheint letztere Limitation im vorliegenden Fall weniger relevant.

Abbrecher des Modellversuchs

Anteil an Abbrechern der GST-Gruppe

In der R&R-Experimentalgruppe wurde ein Abbrecher-Anteil in Höhe von 14.7% gefunden. Nach den von Thomas, Ciliska, Dobbins und Micucci (2004) formulierten Kriterien zur Einschätzung der methodischen Qualität von Literaturübersichten zur Wirksamkeit von Interventionen wird damit eine als „strong“ zu bezeichnende Güte erreicht (<20% Dropout-Rate). Auf den Kriterien von Thomas et al. (2004) beruhen ebenfalls die Empfehlungen des CODC-Komitees zur Evaluation von Interventionen bei Sexualstraftätern (Beech et al., 2007). Der hier berichtete Abbrecher-Anteil liegt im Bereich von Dropout-Raten in bisherigen internationalen R&R-Evaluationen, die Eingang in die Meta-Analysen von Tong und Farrington (2006, 2008) gefunden haben. Cullen, Soria, Clarke, Dean und Fahy (2011) berichten eine höhere Dropout-Rate von 50%. Die Autoren führten das R&R-Programm explizit bei Probanden mit psychischen Störungen durch, was relevant für die vorliegende Evaluation erscheint: Im vorliegenden Modellversuch scheint ein substanzieller Anteil der Probanden psychiatrisch belastet zu sein (Diagnosen aus Kapitel V des ICD-10), auch wenn hier aufgrund des hohen Anteils an fehlenden Werten keine zuverlässigen Aussagen getroffen werden können.
Ein unterschiedlicher Anteil an Abbrechern in den verschiedenen Untersuchungsgruppen kann auf eine mögliche Verzerrung hindeuten. Zum Beispiel kann es durch die hohen Anforderungen eines Behandlungsprogramms zu einem vermehrten Ausscheiden von Studienteilnehmern in der Experimentalgruppe mit hohem Rückfallrisiko kommen, während die Studienteilnehmer mit hohem Rückfallrisiko vermehrt in der Kontrollgruppe verbleiben, da hier keine spezifischen Anforderungen erfüllt werden müssen (Beech et al., 2007). Jedoch war in der vorliegenden Auswertung ein umgekehrtes Bild zu beobachten: Die Probanden des R&R(2)-Programms schieden zu einem signifikant geringeren (gegenüber der Kontrollgruppe) bzw. zu einem tendenziell geringeren Anteil (gegenüber der Vergleichsgruppe) vorzeitig aus dem MV aus. Letzteres weist auf eine erhöhte Compliance in der Experimentalgruppe hin und könnte ein Hinweis auf bessere motivierende Eigenschaften des R&R(2)-Programms im Vergleich zur Standardbehandlung sein. (In der Kontrollgruppe fehlen per definitionem Motivationsanreize, die sich aus den erwarteten positiven Konsequenzen einer Behandlung ableiten lassen).

Unterschiede zwischen Abbrechern und Vollendern der GST-Gruppe

In der Mehrheit der untersuchten Merkmale zeigten sich keine bedeutsamen Unterschiede zwischen Abbrechern und Vollendern. Zum vorbestehenden Rückfallrisiko kann aufgrund des hohen Anteils an fehlenden Werten keine Aussage getroffen werden. Unterschiede waren hinsichtlich folgender Vergleiche zu beobachten: Unabhängig von der jeweiligen Bedingung zeigte sich unter den Abbrechern des MV ein höherer Anteil an Probanden, die nicht die Schweizer Staatsbürgerschaft besitzen. Sowohl in der GST-Gesamtgruppe als auch innerhalb der Vergleichsgruppe verfügten die Abbrecher über ein höheres Bildungsniveau. Innerhalb der Vergleichsgruppe waren die Abbrecher außerdem älter als die Vollender, während sich in der Experimental-, der Kontroll- sowie der Gesamtgruppe keine Hinweise auf Unterschiede im Alter zeigten. Weder innerhalb der Bedingungen noch in Bezug auf die GST-Gesamtgruppe zeigten sich Hinweise auf Unterschiede in der psychiatrischen Belastung zwischen Abbrechern und Vollendern, soweit dies überprüft werden konnte, in ihrer psychiatrischen Belastung. In den beiden Behandlungsgruppen (R&R(2) sowie Standardbehandlung) zeigten sich keine Hinweise auf Unterschiede im Anteil an bereits psychotherapeutisch vorbehandelten Probanden. Das Ausscheiden stand somit nicht mit möglichen Vorbehandlungen in Verbindung.
Keine Hinweise auf Unterschiede zwischen den Abbrechern und den Vollendern des Modellversuchs zeigten sich innerhalb der Experimentalgruppe. Dies ist ein Indiz dafür, dass das vorzeitige Ausscheiden aus dem R&R(2)-Behandlungsprogramm nicht durch vorbestehende Unterschiede in den untersuchten Merkmalen (Alter, Nationalität, Bildungsniveau, Therapieerfahrung, Vorstrafenbelastung und psychiatrische Belastung) erklärt werden kann.

Abbrecher der SST-Gruppe

Etwas mehr als ein Viertel der Probanden des ASAT®Suisse-Programms brach den MV vorzeitig ab. Dieser Anteil lag signifikant niedriger als in der unbehandelten Kontrollgruppe und tendenziell niedriger als in der Vergleichsgruppe. Analog zur GST-Gruppe kann dies in der Tendenz ein Hinweis auf bessere motivierende Eigenschaften des ASAT® Suisse -Programms gegenüber der Standardbehandlung sein.
Zur Abbrecher-Quote in bisherigen Studien zum ASAT®(Suisse) finden sich keine Daten, auch zu Abbrecher-Quoten in Behandlungsprogrammen für Sexualstraftäter ist die Datenlage schwach. Hall (1995) fand in einer Meta-Analyse zu Behandlungsprogrammen für Sexualstraftäter im Mittel eine Abbrecher-Rate von 33%. Marques et al. (2005), die bereits beim Design der Studie die Wahrscheinlichkeit für ein vorzeitiges Ausscheiden zu minimieren versuchten, berichten eine Dropout-Rate in Höhe von 18%. Hingegen berichteten McGrath, Cumming, Livingston und Hoke (2003) für erwachsene Sexualstraftäter eine weitaus höhere Abbrecher-Rate von 47%, wobei in dieser Studie explizit hohe Anforderungen für einen Verbleib im Behandlungsprogramm stellte. Die in der vorliegenden Evaluation berichtete Quote von 25.5% liegt damit im mittleren bis unteren Bereich der genannten Arbeiten; aufgrund der geringen Fallzahlen besteht jedoch auch in diesem Punkt weiterer Forschungsbedarf. Der nicht zu erbringende Nachweis von Unterschieden hinsichtlich Alter, Nationalität, Bildungsniveau, Therapieerfahrung, Vorstrafenbelastung und psychiatrischer Belastung

zwischen Abbrechern und Vollendern ist ein Hinweis darauf, dass sich das vorzeitige Ausscheiden weder in den beiden Behandlungsbedingungen (SST ASAT und SST TAU), noch in der Kontrollbedingung (SST KG), noch in Bezug auf alle Probanden der SST-Gruppe mit vorbestehenden Unterschieden in den genannten Merkmalen erklären lässt.

Fragestellung 2: Effektkriterium Rückfälligkeit (R&R(2) und ASAT®Suisse)

Follow-Up-Dauer und Time at Risk

In Bezug auf die GST-Gruppe ist die im vorliegenden MV erreichte Follow-Up-Dauer positiv einzuschätzen: In bisherigen internationalen R&R-Evaluationen mit der Rückfallrate als Effektkriterium war diese im Durchschnitt eher kürzer als in der vorliegenden Evaluation. Die in die Meta-Analysen von Tong und Farrington (2006, 2008) eingeschlossenen Studien erreichten im Mittel eine Follow-Up-Dauer von etwa 14 Monaten, bei einem Range zwischen acht und 36 Monaten. Eine kurze Zeitspanne zur Einschätzung der Rückfälligkeit ist durchaus üblich, z.B. auch beim Bundesamt für Statistik: „Als rückfällig nach Entlassung werden alle Schweizer Erwachsenen bezeichnet, die innerhalb von drei Jahren nach einer Entlassung aus dem Strafvollzug ein Vergehen oder ein Verbrechen begehen, das eine erneute Verurteilung zur Folge hat“ (Bundesamt für Statistik, 2016). Der hier verwendete Zeitraum von drei Jahren wird im vorliegenden MV in beiden Gruppen jedoch nicht erreicht.
Als Vergleichswert: Die Häufigkeit eines so definierten Rückfalls bei wegen eines Gewaltdelikts verurteilten Straftätern, die im Jahr 2010 in Freiheit entlassen worden waren (aktuellste verfügbare Zahlen), betrug 41.6 % (Bundesamt für Statistik, 2016).

Als äußerst problematisch für die Analyse der Rückfälligkeit im vorliegenden MV hat sich jedoch der geringe Anteil an Probanden erwiesen, die innerhalb des Follow-Up-Zeitraums in Freiheit entlassen worden waren, und die somit überhaupt einer Time at Risk ausgesetzt waren. So standen insgesamt 63 der 278 Probanden (22.7%) aus den drei Bedingungen der GST-Gruppe für die Rückfälligkeitsanalyse zur Verfügung, womit die relativ umfangreiche Stichprobengröße nur teilweise für die Rückfälligkeitsanalyse ausgeschöpft werden konnte.

In der SST-Gruppe, die ohnehin deutlich weniger Probanden umfasst, ist der Anteil an Probanden, die einer Time at Risk ausgesetzt waren mit lediglich neun der in die Evaluation eingeschlossenen 85 Probanden (10.6%) noch geringer. Reliable Aussagen zur Rückfälligkeit sind in dieser Gruppe anhand dieser geringen Fallzahl nicht möglich. Auch wird die minimale Follow-Up-Dauer von 36 Monaten, die zur Erfassung sexueller Rückfälle gefordert wird (Beech et al., 2007), nicht erreicht.

Fragestellung 1: Veränderungen in den Messwerten der Fragebögen in Abhängigkeit der Therapie (R&R(2))

K-FAF: Aspekte der Aggressivität

Gefundene Unterschiede

Zwischen den drei Bedingungen der GST-Gruppe zeigten sich signifikante Unterschiede in den Differenzwerten des Summenwertes der selbst berichteten Aggressivität. Dieser Unterschied entspricht einem kleinen Effekt. Im Summenwert werden die nach außen gerichteten, eher impulsiven Aspekte von Aggressivität zusammengefasst: Spontane und reaktive Aggressivität sowie Erregbarkeit. Da die post hoc durchgeführten Einzelvergleiche kein signifikantes Ergebnis auf dem definierten Alpha-Niveau von 5% aufwiesen, lassen sich höchstens leichte Trends ausmachen, die zu einem späteren Zeitpunkt Ausgangspunkt für weitere Analysen darstellen: Die Probanden in allen drei Bedingungen berichteten zum zweiten Messzeitpunkt eine geringer ausgeprägte Aggressivität. Die höchste Reduktion zeigten dabei die Probanden der Experimentalgruppe (-9.2), gefolgt von der Vergleichs- (-2.2) und der Kontrollgruppe (-1.3). Der Einzelvergleich mit dem höchsten Effekt ist derjenige zwischen Experimental- und Kontrollgruppe: In der Größe eines Trends scheint nach der spezifischen Intervention in Form des Reasoning & Rehabilitation (2)-Programms die Aggressivität somit geringer ausgeprägt zu sein als bei Probanden, die keine therapeutische Intervention erhalten, gemäß Selbstauskunft.

Dieses Muster ist ebenso in den einzelnen Skalen des K-FAF ersichtlich (ohne das statistische 5% Signifikanzniveau zu erreichen): In allen Skalen, die Aspekte von Aggressivität messen, zeigen die Probanden der Experimentalgruppe zum zweiten Messzeitpunkt geringere Ausprägungen. In der Vergleichsgruppe zeigt sich durchweg eine geringer ausgeprägte Reduktion bzw. im Falle der spontanen Aggressivität ein höherer Wert zum zweiten Messzeitpunkt. Die Probanden der Kontrollgruppe berichten deskriptiv die am geringsten ausgeprägten Veränderungen in Form von reduzierten (reaktive Aggressivität, Erregbarkeit und Selbstaggressivität) bzw. erhöhten Werten (spontane Aggressivität). Im Fall der inhaltlich inversen Skala „Aggressionshemmung“ wird dieses Muster im Vergleich zwischen den drei Bedingungen des MV umgekehrt (ebenfalls ohne dabei das Signifikanzniveau zu erreichen): Während die Probanden der Experimentalgruppe eine erhöhte Aggressionshemmung zum zweiten Messzeitpunkt berichten, bewegt sich die Veränderung in den beiden anderen Gruppen um den Nullpunkt bzw. es wird eine reduzierte Aggressionshemmung berichtet.
Die Selbstauskünfte der Probanden zu Aspekten der Aggressivität konnten nicht durch eine Fremdbeurteilung verifiziert werden, da hier in zu vielen Fällen keine Angaben vorhanden sind (mehr als ¼ in der Experimentalgruppe, mehr als ⅓ in der Vergleichsgruppe und mehr als ⅔ in der Kontrollgruppe).

Beantwortung der Hypothesen

Die Hypothesen (1.1) und (2.1) des MV in Bezug auf das Effektkriterium der Aggressivität, wie sie mit dem K-FAF erhoben wird, müssen verworfen werden, da die gefundenen Unterschiede statistisch zu wenig robust sind.

In der Größenordnung eines Trends hängt die neue therapeutische Intervention des R&R(2) im Vergleich zu einer Nicht-Behandlung mit einem deutlicher ausgeprägten positiven Effekt auf Aspekte der nach außen gerichteten impulsiven Aggressivität zusammen.
Die Hypothese (3.1) ist dahingehend zu beantworten, dass sich keine Hinweise auf Unterschiede zwischen den mit dem R&R(2) und den mit der Standardtherapie behandelten Gewalt- und Sexualstraftätern zeigten.

IIP-D: Interpersonelle Schwierigkeiten

Die Hypothesen (1.1) und (2.1) des MV in Bezug auf das Effektkriterium interpersoneller Schwierigkeiten als Vermittler für aggressives Verhalten mussten verworfen werden: Eine positive Veränderung interpersoneller Probleme, wie sie mit dem IIP-D erfasst werden, durch die Interventionen im Rahmen des MV konnte nicht aufgezeigt werden. Dies kann mit der relativ kurzen Dauer zwischen den beiden Messzeitpunkten (140 Tage für GST R&R) zusammenhängen: Persönlichkeitstheorien wie sie auch dem IIP-D zugrunde liegen, konzeptualisieren die Persönlichkeit eines Menschen als eher zeitstabil, auch wenn das IIP-D mehr auf die verhaltensnahen Aspekte von Persönlichkeitsmerkmalen abzielt als andere klassische Persönlichkeits-Fragebögen. Somit erscheint es schwer, auf dieser Ebene in nur kurzer Zeit Veränderungen zu induzieren. Ein mehr verhaltensorientiertes Instrument wie z.B. der K-FAF könnte für eine kurzzeitige Veränderungsmessung geeigneter erscheinen, auch weil ein kognitiv-behaviorales Behandlungsprogramm wie das R&R(2) primär auf dieser Modalität ansetzt.
Die Hypothese (3.1) in Bezug auf das Effektkriterium interpersoneller Schwierigkeiten ist dahingehend zu beantworten, dass sich keine Hinweise auf Unterschiede zwischen den mit dem R&R(2) und den mit der Standardtherapie behandelten Gewalt- und Sexualstraftätern zeigten. Einschränkend ist dabei die oben genannte Verwerfung der Hypothese eines allgemeinen Behandlungseffektes auf die Differenzwerte des IIP-D zu bedenken.

HAB: Feindselige Attribution von Verhalten

Erwartete Effekte gemäß Hypothesen

Aus der Literatur zur menschlichen Aggressivität ist abzuleiten, dass eine hohe Aggressionsneigung mit der Tendenz einhergeht, einem Interaktionspartner feindselige Absichten zu unterstellen (Hostile Attribution). Diese subjektiv wahrgenommene feindselige Absicht ist ein wichtiges Element in der Genese aggressiven Verhaltens (Tremblay & Belchevski, 2004). Der Zusammenhang zwischen Aggressionsneigung und feindseliger Attribution konnte insbesondere in Situationen gezeigt werden, in denen die Absicht des Gegenübers nicht eindeutig ist (Crick & Dodge, 1994; Dill, Anderson, Anderson, & Deuser, 1997; Dodge, 1980; Matthews & Norris, 2002; vgl. Abschnitt "Beschreibung der verwendeten Instrumente" in Anhang 3). Daraus ergibt sich, dass eine Abnahme der Wahrscheinlichkeit für aggressives Verhalten sich insbesondere in den Fallvignetten des HAB niederschlagen sollte, die eine uneindeutige

Situation beschreiben; aus diesem Grund handelte es sich bei der Mehrzahl der im Rahmen des MV eingesetzten Fallvignetten um solche unklaren Situationen.

Situationen mit uneindeutiger Absicht des Interaktionspartners
Der erwartete Effekt der Behandlung konnte hier nicht gezeigt werden: Weder im Summenwert des HAB noch in den einzelnen Aspekten feindseliger Wahrnehmung oder aggressiven Verhaltens zeigten sich signifikante bedingungsabhängige Unterschiede in den Differenzwerten zu den beiden Messzeitpunkten. Lediglich in den Items, welche die Wahrscheinlichkeit für verbal aggressives Verhalten erfassen (unhöfliches Verhalten sowie Anschreien/ Beschimpfen), zeigten sich Unterschiede zwischen den Bedingungen in Form eines Trends (p<.10). In der selbst berichteten Wahrscheinlichkeit, auf eine unklare Situation mit Anschreien oder Beschimpfen zu reagieren, ergab sich in der Experimentalgruppe eine Verminderung, während in der Kontrollgruppe eine gering erhöhte Wahrscheinlichkeit berichtet wurde. Dieser Unterschied liegt ebenfalls nur in Form eines Trends vor.

Situationen mit eindeutig provozierender Absicht des Interaktionspartners
In diesen Fallvignetten zeigten sich signifikante Unterschiede zwischen den drei Bedingungen des MV: Die Probanden der Experimentalgruppe berichteten zum zweiten Messzeitpunkt über eine verringerte Wahrscheinlichkeit für aggressive Reaktionen in den Aspekten „Wahrnehmung der Situation als provozierend“, „Anschreien/ Beschimpfen“ sowie im Summenwert), während die Probanden der Kontrollgruppe jeweils eine erhöhte Wahrscheinlichkeit berichteten. Ein Unterschied zwischen den Bedingungen zeigte sich ferner in Form eines Trends in der selbst berichteten Wahrscheinlichkeit für unhöfliches Verhalten, hier berichtete die Experimentalgruppe eine reduzierte Wahrscheinlichkeit, während die Wahrscheinlichkeit in der Vergleichsgruppe unverändert blieb. Alle berichteten Unterschiede entsprechen einem kleinen Effekt.
Hervorzuheben ist hier die Veränderung in der Experimentalgruppe hinsichtlich der Wahrnehmung einer klar provozierenden Situation als solche: So besteht das Ziel einer kognitiv-behavioralen Therapie nicht darin, eine der Realität entsprechende Einschätzung einer Situation zu verändern; vielmehr sollte die Wahrscheinlichkeit reduziert werden, auf eine wie auch immer wahrgenommene Situation aggressiv zu reagieren.

Situationen mit eindeutig nicht provozierender Absicht des Interaktionspartners
In diesen Situationen konnten keine Hinweise auf Unterschiede zwischen den drei Bedingungen in den Differenzwerten der HAB-Items festgestellt werden, weder in Bezug auf die einzelnen Aspekte feindseliger Wahrnehmung oder aggressiven Verhaltens noch in Bezug auf den Summenwert.

Beantwortung der Hypothesen
Unterschiede zwischen den drei Bedingungen der GST-Gruppe konnten nicht im erwarteten Umfang gefunden werden. Die Hypothesen (1.1) und (2.1) in Bezug auf das Effektkriterium der feindseligen Attribution müssen daher verworfen werden, da die gefundenen Unterschiede statistisch nicht robust genug sind.

Die Veränderungen in einigen Items des HAB können jedoch auf einen (kleinen) Effekt in der R&R-Experimentalgruppe hinweisen, der in Zusammenhang mit der neuen Intervention steht: Gewaltstraftäter, die das R&R(2)-Programm durchlaufen haben, äußern zum Zeitpunkt der Post-Messung im Vergleich zu Gewaltstraftätern, die keine Therapie erhalten, in einigen Aspekten in der Tendenz eine reduzierte Wahrscheinlichkeit, in Situationen mit Konfliktpotenzial aggressiv zu reagieren. Dies betrifft zumindest die verbale Aggressivität (die Wahrscheinlichkeit hierzu wird in provozierenden Situationen als signifikant reduziert und in unklaren Situationen als tendenziell reduziert eingeschätzt). Dies erscheint insofern wichtig, da mit dem Äußern von Beschimpfungen zwar noch nicht die Handlungsschwelle zu physischer Gewalt überschritten wird, jedoch eine deutliche Eskalation des Konflikts erreicht wird. Zudem kann ab dieser Konfliktstufe strafrechtlich relevantes Verhalten in Form von Beleidigungen vorliegen. Somit erscheint im Kontext der Attribution von Absichten eine Verminderung der Bereitschaft zu verbaler Aggression durchaus zielführend, um eine Eskalation zu vermeiden, die zu körperlicher Gewalt und somit zu einem einschlägigen Rückfalldelikt führen kann.

Die Hypothese (3.1) ist dahingehend zu beantworten, dass sich keine Hinweise auf Unterschiede zwischen den mit dem R&R(2) und den mit der Standardtherapie behandelten Gewalt- und Sexualstraftäter zeigten – mit der Einschränkung, dass bereits der allgemeine Wirksamkeitsnachweis der forensischen Therapien in Bezug auf den HAB nicht im erwarteten Ausmaß erbracht werden konnte.

VÜ: Verantwortungsübernahme für das Delikt

Interpretation selbst berichteter Verantwortungsübernahme

Zunächst sind folgende theoretische Schwierigkeiten zu beachten, wenn das selbst berichtete Maß an Verantwortungsübernahme als Kriterium für eine erfolgreiche Behandlung herangezogen wird. Ein Grund für eine Verantwortungsübernahme kann ein hohes Ausmaß psychopathischer Eigenschaften sein, wie sie überdurchschnittlich häufig bei Gewalt- und Sexualstraftätern vorliegen – und entsprechend darf Verantwortungsübernahme nicht ohne weiteres als ein Anzeichen von Reue interpretiert werden (Henning & Holdford, 2006; Maruna & Mann, 2006). Im vorliegenden MV scheint diese Einschränkung jedoch hinreichend irrelevant: Zwar können reliable Aussagen zu psychopathischen Merkmalen aufgrund des hohen Anteils an fehlenden Werten nicht getroffen werden; bei denjenigen Probanden mit Angaben in der PCL-R ergaben sich jedoch augenscheinlich keine Hinweise auf Unterschiede zwischen den GST-Bedingungen hinsichtlich des PCL-R-Summenwertes und auch generell keine besonders ausgeprägten psychopathischen Eigenschaften der Probanden.

Gefundene Unterschiede

Im Gesamtwert des Fragebogens zur Verantwortungsübernahme berichten die Gewalt- und Sexualstraftäter, die das R&R(2)-Programm durchlaufen haben, ein höheres Ausmaß an Verantwortungsübernahme zum zweiten Messzeitpunkt. Diese Veränderung stellt einen signifikanten Unterschied zu denjenigen Gewalt- und Sexualstraftätern dar, die keine Therapie erhalten haben (bei eingeschränkter Aus-

sagekraft aufgrund des knapp über einem Fünftel liegenden Anteils an fehlenden Werten in der Kontrollgruppe): Diese Probanden berichteten im Gegenteil von einer geringer ausgeprägten Verantwortungsübernahme zum Zeitpunkt der Post-Messung. Der berichtete Effekt ist als klein einzustufen.

Keine Hinweise auf Unterschiede zwischen den drei Bedingungen der GST-Gruppe zeigten sich in den Differenzwerten der Subskalen „Rechtfertigung“ und „Entschuldigung“, sodass über die spezifische Art der Verantwortungsübernahme für das Delikt keine Aussage getroffen werden kann.

Die durch Selbstauskunft gewonnenen Angaben können nicht durch die Fremdbeurteilung überprüft werden, da hier in einem substanziellen Anteil der Fälle keine Angaben vorliegen. Im Mittel bewegt sich die Differenz zwischen den beiden Messzeitpunkten bei denjenigen Probanden, zu denen Fremdbeurteilungen vorliegen, um den Nullpunkt (rein deskriptiv, bei sehr eingeschränkter Aussagekraft aufgrund des hohen Anteils an fehlenden Werten).

Beantwortung der Hypothesen
Somit kann die Hypothese (2.1) teilweise gestützt werden, dass das in der Schweiz neuartige Interventionsprogramm R&R(2) wirksam ist in Bezug auf das Effektkriterium der Verantwortungsübernahme (als Teil der Fragestellung 1): Gewalt- und Sexualstraftäter, die das R&R(2) durchlaufen haben, zeigten die erwarteten positiven Veränderungen im Vergleich zu unbehandelten Gewalt- und Sexualstraftätern, und zwar bezogen auf eine allgemeine Verantwortungsübernahme, wie sie durch die Probanden selbst berichtet wird.
Die Hypothese (3.1) ist dahingehend zu beantworten, dass sich keine Hinweise auf Unterschiede in den Differenzwerten der VÜ zwischen den mit dem R&R(2) und den mit der Standardtherapie behandelten Gewalt- und Sexualstraftätern zeigten.
Zu verwerfen hingegen ist die Hypothese (1.1).

Zusammenfassung der Evaluationsergebnisse

Evaluation des R&R- bzw. R&R2-Behandlungsprogramms für Gewaltstraftäter

Es zeigten sich Hinweise auf kleine, positive Behandlungseffekte für Gewaltstraftäter durch die Teilnahme am R&R(2)-Programm im Vergleich zu unbehandelten Gewaltstraftätern. Dies betrifft vor allem die selbst berichtete Aggressivität, wie sie mit dem K-FAF erhoben wurde. Teilweise berichteten die Teilnehmer des R&R(2)-Programms außerdem eine geringer ausgeprägte feindselige Attribution in einer Alltagssituation mit Konfliktpotenzial zum Zeitpunkt der Post-Messung im Vergleich zur unbehandelten Kontrollgruppe. Dies konnte in Form einer reduzierten Wahrscheinlichkeit für verbale Aggression in konflikthaften Situationen (und in der Tendenz in mehrdeutigen Situationen) gezeigt werden. Zudem berichteten die Teilnehmer des R&R(2)-Programms von einer erhöhten Verantwortungsübernahme für ihr Delikt im Vergleich zu den unbehandelten Kontrollprobanden. Inwieweit die gefundenen positiven Veränderungen auf ein sozial erwünschtes Antwortverhalten der R&R(2)-Teilnehmer

zurückzuführen sind, muss jedoch unbeantwortet bleiben, da keine ausreichenden Daten für eine Fremdbeurteilung vorliegen. Keine Veränderungen, die auf einen Effekt der Teilnahme am R&R(2)-Programm zurückzuführen sein könnten, zeigten sich im Bereich interpersoneller Probleme: Wie bereits erwähnt, könnte es schwer sein, mit einer therapeutischen Intervention kurzfristig positive Effekte im Bereich der Persönlichkeit zu induzieren, auch wenn das IIP-D auf interpersonales Verhalten zielt und damit auf weniger statische Elemente als andere klassische Persönlichkeits-Fragebögen.

Anhand des „harten“ Effektkriteriums der Rückfälligkeit konnte hingegen keine Überlegenheit des R&R(2)-Programms gegenüber einer Nicht-Behandlung aufgezeigt werden. Auch im Vergleich zur Standardbehandlung in Form von delikt- und störungsorientierter Einzeltherapie zeigten sich keine Hinweise auf Unterschiede. Solch ein Nachweis ist im Rahmen der kurzen Time at Risk und angesichts des relativ geringen Anteils an Probanden, die überhaupt einer Time at Risk ausgesetzt waren, nur schwer zu erbringen. Anzumerken ist, dass keiner der zwölf rückfällig gewordenen Probanden in der Gesamtgruppe der Gewaltstraftäter ein schweres Rückfalldelikt beging: Neben einer sexuellen Belästigung (ein Proband der Experimentalgruppe), einer Körperverletzung und Tätlichkeiten (zwei Probanden der Vergleichsgruppe) handelte es sich mehrheitlich um Eigentumsdelikte oder um Fehlverhalten außerhalb des StGB. Gerade die sehr niedrige einschlägige Rückfälligkeit erschwert es, einen Überlegenheitseffekt einer bestimmten Intervention belegen zu können. Darüber hinaus muss bedacht werden, dass auch die formal unbehandelten Teilnehmer der Kontrollgruppe, die keine spezifische therapeutische Intervention erhalten, indirekt von einem Klima profitieren, das förderlich für die Rehabilitation ist und in dem bereits eine gute (psychiatrisch-psychologische) Versorgungsstruktur für sämtliche Klienten vorhanden ist, wie dies bereits in vielen Institutionen des Schweizer Strafvollzugssystems der Fall ist.

Über diese unmittelbaren Ergebnisse hinaus kann der geringe Stichprobenschwund als vorteilhaft für das R&R(2)-Therapieprogramm beurteilt werden. Der Anteil an Studienteilnehmern, die den Modellversuch vorzeitig abbrachen, lag niedriger als im Falle der Standardbehandlung. Zwar können anhand der vorliegenden Daten keine Aussagen über den Grund des Abbruchs getroffen werden (insbesondere bleibt die Frage ungeklärt, ob es sich um Therapie- oder lediglich um Studienabbrecher handelt), dennoch ist dies ein Indiz für eine erhöhte Compliance in der R&R(2)-Behandlungsgruppe.

Sämtliche hier getroffenen Aussagen sind mit der Einschränkung zu lesen, dass das R&R- und das R&R2-Behandlungsprogramm in einer gemeinsamen Gruppe zusammengefasst werden. Es ist letztlich anhand der vorliegenden Daten keine Aussage möglich, inwieweit Unterschiede zwischen den beiden Varianten des R&R zu den berichteten Effekten geführt haben könnten.

Evaluation des ASAT®Suisse-Behandlungsprogramms für Sexualstraftäter

Die Hypothesen des Modellversuchs in Bezug auf die Wirksamkeitsüberprüfung des ASAT®Suisse-Behandlungsprogramms für Sexualstraftäter konnten nicht empirisch überprüft werden. Dies lag in erster Linie an den Einschränkungen, die bereits bezüglich des R&R(2)-Programms genannt wurden: Die ohnehin geringe Fallzahl wurde zusätzlich durch die geringe Zahl von lediglich neun Studienteilnehmern reduziert, die überhaupt einer Time at Risk ausgesetzt waren. Positiv zu erwähnen ist dennoch, dass unter diesen neun Teilnehmern im zur Verfügung stehenden Beobachtungszeitraum von durchschnittlich knapp zwei Jahren lediglich ein geringfügiger Rückfall zu verzeichnen war; dies in der Gruppe der mit Standardtherapie behandelten Sexualstraftäter.
Analog zum R&R(2) zeichnete sich auch für das ASAT®Suisse eine tendenziell höhere Compliance im Vergleich zur TAU-Gruppe ab: Der Anteil an Abbrechern lag in der ASAT®Suisse-Gruppe niedriger, wobei auch hier keine Aussage über den genauen Grund des Ausscheidens getroffen werden kann, sodass der gezogene Schluss unter Vorbehalt steht.

Evaluation weiterer ausgewählter Aspekte des Modellversuchs

I Die Adäquanz des Versuchs- und Evaluationsdesigns:

Im Modellversuch wurden die Auswirkungen einer Teilnahme am R&R(2) bzw. am ASAT Programm anhand eines quasi-experimentellen Studiendesigns erfasst. Es wurden je Programm 3 Gruppen gebildet, namentlich die Experimentalgruppe (EG), die Vergleichsgruppe (VG) und die Kontrollgruppe (KG). Die Forschenden nahmen keinen Einfluss auf die Gruppenzuordnung, diese basierte auf vorhandenen Eigenschaften der Studienteilnehmenden (ST). Therapien wurden entweder durch das Gericht oder vollzugsseitig angeordnet, die Zuweisung zu R&R(2) bzw. ASAT erfolgte i.R. durch die Psychotherapeuten.

Quasi-experimentelle Studiendesigns sind auf der Scientific Methods Scale (SMS; Sherman et al., 1997), welche die Stärke der internen Validität widergibt, auf Stufe 3 (low-quality) bzw. 4 (high-quality) von insgesamt 5 lokalisiert. In Studiendesigns der Stufe 4 erfolgt die statistische oder methodische Kontrolle relevanter Faktoren, die Gruppendifferenzen bedingen können, während dies in Studiendesigns der Stufe 3 nicht erfolgt. Wilson et al. (2005) schlagen vor, dass Studien ab Stufe 3 von akzeptabler wissenschaftlicher Güte sind.

Die wesentliche Herausforderung dieses Studiendesigns ist die Sicherstellung vergleichbarer Gruppen bzw. die Vermeidung eines Selektions-Bias. In einem quasi-experimentellen Studiendesign ist es möglich, dass sich die ST a priori hinsichtlich einer Reihe von Merkmalen unterscheiden, welche die Evaluationsergebnisse beeinflussen können. Tatsächlich zeigten vorliegende Gruppenvergleiche (EG-VG-KG) eine Reihe von Gruppenunterschieden auf, welche die Ergebnisse der Evaluation potenziell konfundieren. Entsprechend ist es nur bedingt möglich, auf der Basis dieser Gruppenvergleiche für die Wirksamkeit sprechende Ergebnisse tatsächlich den

beiden Programmen zuzuschreiben, da sie ebenso gut auf Gruppenunterschiede rückführbar sein könnten. Diese potenzielle Konfundierung verringert die interne Qualität dieser Studie. Nichtsdestotrotz geben gruppenspezifische (intraindividuelle) Prä-Post-Vergleiche Auskunft über Veränderungen während der Gruppenteilnahme, die zumindest als Hinweise für oder gegen die Wirksamkeit gewertet werden können.

Zur Sicherstellung der internen Validität und Lösung der oben beschriebenen Herausforderung bietet sich ein randomisiert-kontrolliertes Studiendesign an (englisch: randomized controlled trial; RCT). Hier werden die ST den Versuchsbedingungen randomisiert zugewiesen. So hat jeder ST dieselbe Wahrscheinlichkeit in eine der Bedingungen zugeordnet zu werden, wodurch, bei genügend grosser Fallzahl, eine Gleichverteilung der Charakteristika der ST auf die Bedingungen erreicht wird.

RCT bergen nicht vernachlässigbare Risiken. Aus ethischer Sicht bedeutet die Zuteilung eines ST zu einer der Bedingungen auf Basis des Zufalls, dass einige der ST keine Behandlung oder erst zu einem späteren Zeitpunkt eine Behandlung erhalten, obwohl eine solche notwendig wäre. Besonders im forensischen Kontext, wo eine Psychotherapie mitunter gerichtlich angeordnet wurde und zu vollziehen ist, oder auch der Haftkontext eine sofortige psychotherapeutische Intervention notwendig machen kann (zu nennen wäre z.B. akute Suizidalität), ist dies eine grosse ethische Herausforderung und kaum zu vertreten. Denn was, wenn ein ST keine Therapie erhält, obwohl angeordnet, weil er am Versuch teilnimmt und der KG zugeordnet wurde, und in dieser Zeit erneut straffällig wird oder sich selbst oder jemand anderen verletzt oder sich suizidiert. Dies wirft die Frage auf, ob das Verbrechen, die Selbst- oder Fremdverletzung oder der Suizid durch die Teilnahme an der Therapie hätte verhindert werden können. Auch kann die Nicht-Teilnahme an der Therapie, wegen der Zuordnung zur Kontrollgruppe, dazu führen, dass die Vollzugsprogression behindert wird. In anderen Worten könnte es jemandem durchaus zum Nachteil ausgelegt werden, dass er aktuell, auf Basis einer randomisierten Zuordnung zur KG, an keiner Psychotherapie teilnimmt und so könnten ihm Vollzugslockerungen wie bspw. die Unterbringung bzw. Platzierung in einem Wohn- und Arbeitsexternat (WAEX) nicht zugestanden werden. Das würde die gesetzlich verankerte Rehabilitation behindern.

Eine weitere Schwierigkeit von RCTs besteht darin, dass ein Bias dadurch entstehen kann, dass mitunter Personen, die sich unter natürlichen Bedingungen nicht einer Therapiegruppe angeschlossen hätten, in eine eben solche zugeordnet werden. Dies setzt die externe Validität herab, die bei einem quasi-experimentellen Studiendesign als höher zu bewerten ist. Tatsächlich verhalten sich interne und externe Validität bei RCTs und bei quasi-experimentellen Designs in der Regel invers. Das birgt das Risiko, dass diese Personen im Verlauf entweder die Teilnahme verweigern, oder dass diese Personen, wenn sie das Programm durchlaufen, wegen ihrer geringen Therapiemotivation und damit einhergehend geringe Veränderungsmotivation die Ergebnisse verzerren. Als Beispiel dazu der von Van Voorhis et al. (2004) durchgeführte RCT über die Wirksamkeit vom R&R in den USA. Sie stellten eine Dropout-Rate von 40% fest. Mittels Intention-to-treat-Analyse stellten sie keine Wirksamkeit des Programms fest. Durch die Anwendung eines quasi-experimentellen Designs, dabei auf

diejenigen ST fokussierend, die das Programm bis zum Ende durchlaufen haben (completers) fanden sie Hinweise für die Wirksamkeit.

Grundsätzlich gilt, dass genuiner Zweifel über die Wirksamkeit eines Programms bestehen muss, wenn es mittels RCT untersucht werden soll. Liegen bereits Hinweise über die Wirksamkeit eines Programms vor, besteht kein genuiner Zweifel und es ist von der Implementierung des RCT eher abzusehen, insbesondere wegen der zuvor genannten ethischen Herausforderungen. Vorliegend lagen erste Hinweise über die Wirksamkeit von R&R und R&R2 vor. Entsprechend war ein aufwändiger und ethisch herausfordernder RCT zu Gunsten des Studiendesigns mit der nächstbesten Güte, nämlich ein quasi-experimentelles Studiendesign aus Sicht der Durchführenden zu verwerfen. Generell sind RCTs im kriminologischen Bereich unüblich. Farrington et al. (2002, S. 17) konstatierten in diesem Kontext: „While randomized experiments in principle have the highest internal validity, in practice they are uncommon in criminology and also often have implementation problems“.

Die im Rahmen dieses Modellversuchs getroffene Entscheidung für ein quasi-experimentelles Design und gegen ein RCT wird weiter dadurch unterstützt, dass beide Designs ähnlich grosse Effekte finden (Babcock et al., 2004; Lipsey et al., 2001; Lösel & Schmucker, 2005; Wilson et al., 2005).
Die Entscheidung für das vorliegende quasi-experimentelle Studiendesign erscheint folglich gerechtfertigt. Jedoch ist diese Studie auf der SMS auf Stufe 3 und nicht auf Stufe 4 anzusiedeln, da keine statistische oder methodische Kontrolle von Merkmalen erfolgte, die theoretisch Gruppendifferenzen a priori verursachten und somit einen Selektions-Bias evozierten. Die Anhebung der Studie auf Stufe 4 mittels eines Matchings/einer Parallelisierung wäre wünschenswert und geplant gewesen, doch wegen einer zu gering ausgefallenen Stichprobengrösse nicht umsetzbar. Gleichzeitig soll angemerkt sein, dass die SMS ausschliesslich eine Aussage über die interne Validität macht und weitere Arten an Validitäten, bspw. die externe Validität, unberücksichtigt lässt. Gerade die externe Validität, also die Entsprechung der experimentellen Situation mit der realen Situation, ist im therapeutischen Kontext besonders wichtig. Einzig auf die SMS abzustellen, greift aus Sicht der Durchführenden zu kurz, weil diese nicht alle relevanten Gütekriterien bewertet, sondern ausschliesslich eines.

II. Die Bedeutung der Ergebnisse des Modellversuchs für die Treatment as Usual (TAU-) Gruppe

Hinsichtlich der Gewaltstraftäter zeigten sich statistisch signifikante Unterschiede hinsichtlich einiger Skalen des HAB (Selbstbericht) sowie des Fragebogens zur Verantwortungsübernahme, wobei sich jeweils ausschliesslich die EG von der KG unterschied. Die EG zeigte gegenüber der KG in wenigen Skalen Verbesserungen in der Post- gegenüber der Prämessung. Hinsichtlich aller anderen Fragebögen sowie hinsichtlich der Rückfälligkeit zeigten sich keine statistisch signifikanten Gruppeunterschiede. Die VG (alias TAU) unterschied sich niemals statistisch signifikant von der EG und der KG.

In Anbetracht dessen, dass sich die EG (nicht die VG) tendenziell von der KG unterschied, und sich die EG von der VG theoretisch nur durch die Gruppenteilnahme unterscheidet, denn beide Gruppen haben begleitende (Einzel-)Psychotherapie, kann dies als vorsichtiger Hinweis darauf gewertet werden, dass die EG-Gruppe etwas besser als die VG-Gruppe abschneidet und eine Programmteilnahme vorteilhaft sein könnte. Entsprechend könnte der Schluss gezogen werden, dass sich in Psychotherapie befindliche Gewalt-Straftäter möglichst zusätzlich dieses Programm absolvieren. Gleichzeitig sind solche Schlüsse vor dem Hintergrund dessen, dass sich die EG von der KG hinsichtlich einer Reihe von a priori Merkmalen unterscheidet, schwierig zu ziehen. Zu den Unterschieden gehörten bspw. eine höhere Rate an Ausländern in der KG, ein Unterschied hinsichtlich des Bildungsniveaus und häufigere therapeutische Vorerfahrungen in der EG verglichen mit der VG. Es bleibt zukünftigen Studien vorbehalten diesen Schluss zu veri- oder zu falsifizieren.

Hinsichtlich Rückfälligkeit zeigten sich keine Gruppenunterschiede, weder in der intent-to-treat-Analyse noch in der Analyse mit den Vollendern. Jedoch ist anzumerken, dass die Stichprobengrösse gering ausfiel und idealerweise zu einem späteren Zeitpunkt weitere Strafregisterauszüge derer analysiert werden, dabei die Time-at-Risk erhöhend, die bis dann aus der Haft ausgetreten sind. Auf der Basis dieser Datenlage sind keine Schlüsse über die VG (alias TAU) möglich.

Hinsichtlich der Sexualstraftäter fand keine Auswertung der Fragebögen statt, es wurde ausschliesslich auf die Rückfälligkeit abgestellt. Doch auch diese Analyse war wegen der geringen zum Zeitpunkt der Einforderung der Strafregisterauszüge aus der Haft entlassenen ST nicht möglich (in der intent to treat Analyse waren 9 von 85 ST in Freiheit und in der Vollender-Analyse waren es 6 von 55). Auf der Basis dieser Datenlage ist es unmöglich Schlüsse über mögliche Konsequenzen über die TAU-Gruppe zu ziehen, das bleibt zukünftigen Studien vorbehalten.

Zusammenfassend ergaben sich erste vorsichtige Hinweise für ein besseres Outcome derjeniger, welche zusätzlich zur Einzel-Psychotherapie an der R&R(2)-Gruppe teilnahmen. Entsprechend sind die Konsequenzen für die TAU, dass die Teilnahme an dieser Gruppe indiziert ist.

III. Die Bedeutung der Katamnesedauer für den Modellversuch

Die durchschnittliche Time-at-Risk (TAR alias Katamnesedauer) betrug vorliegend über alle Bedingungen hinweg etwa 20 Monate, also 1 Jahr und 8 Monate. Das Bundesamt für Statistik (2016) verwendet für Rückfälligkeit eine Zeitdauer von drei Jahren. Daher ist der Zeitraum der TAR vorliegend als kurz zu bewerten. Gleichzeitig ist anzumerken, dass auch andere Therapie-Evaluations-Studien ihre Interpretation auf eher kurze TAR abstellten, bspw. Tong und Farrington (2006, 2008).
In Anbetracht der Notwendigkeit der Generierung von Rückfälligkeitsdaten, die zum Verständnis der Sinnhaftigkeit von Therapieprogrammen beitragen, muss es ein zentrales Anliegen sein, diesen Modellversuch insofern weiterzuführen, als dass zu einem späteren Zeitpunkt weitere Strafregisterauszüge als Indikatoren für Rückfälligkeit eingefordert werden. Einerseits steigert dies die Stichprobengrösse, da zu einem

späteren Zeitpunkt weitere Personen aus der Haft entlassen sind und somit Gelegenheit hatten sich zu bewähren. Anderseits verlängert sich dadurch die TAR. Beides erhöht die Aussagekraft der Daten und somit die Studienqualität, sodass weiterreichende Schlüsse hinsichtlich einer möglichen Beeinflussung der Rückfälligkeitsrate durch die Programmteilnahmen möglich sein könnten.

Stärken des Modellversuchs

Eine große Stärke des Modellversuchs stellt der multizentrische Ansatz dar, im Rahmen dessen Studienteilnehmer aus verschiedenen Institutionen des Strafvollzugskonkordats Nordwest- und Innerschweiz beteiligt waren. Dies zu ermöglichen, bedeutete einen entsprechenden personellen, zeitlichen und organisationalen Aufwand. Nur durch einen solchen Ansatz konnte die substanzielle Anzahl an Gewaltstraftätern gewonnen werden, aus denen sich die GST-Gruppe zusammensetzte. Letztlich bildet der Modellversuch damit ein großes Stück der Realität im (Nordwest- und Inner-) Schweizer Strafvollzug ab. Gerade im Hinblick auf die externe Validität der Ergebnisse ist dies eine große Stärke. Damit verbunden stellt das verwendete quasi-experimentelle prospektive Design eine adäquate Wahl dar (vgl. Hollin, 2008; Ioannidis et al., 2001). Positiv hervorzuheben ist außerdem, dass der Modellversuch zwar eine longitudinale Erhebung darstellt, der Zeitraum, in dem die Interventionen durchgeführt wurden, mit knapp viereinhalb Jahren jedoch eng umrissen bleibt. Dadurch bleibt der Einfluss von Änderungen in den formalen Rahmenbedingungen, in theoretischen Erwägungen und im Verständnis von Risikofaktoren und Veränderungsprozessen begrenzt, welcher in Studien, die eine mächtigere Stichprobengröße durch eine zeitliche Ausdehnung zu erreichen versuchen, nicht zu vernachlässigen ist. Ausnahme hiervon ist die Umstellung von R&R auf R&R2 im Laufe der Studie (siehe folgender Abschnitt).
So gut dies im Rahmen des Multicenter-Ansatzes möglich war, wurde versucht, die Programmintegrität einzuhalten; z.B. durch den Einbezug externer Evaluatoren bei der Planung, Begleitung und Auswertung des Modellversuchs. Sehr zielführend erwies sich die Wahl verhaltensnaher risikoassoziierter Effektkriterien (Fragestellung 1) als Ergänzung zum Außenkriterium der Rückfälligkeit (Fragestellung 2) in der GST-Gruppe. Gerade unter den Schwierigkeiten der wenigen Studienteilnehmer „at risk“ und der begrenzten Follow-Up-Dauer erscheinen risikorelevante Verhaltens-, Einstellungs- und Persönlichkeitsmaße als unmittelbare Effektkriterien eine sinnvolle Ergänzung zum „harten“ Effektkriterium Rückfälligkeit, solange diese psychologischen Maße theoretisch gut fundiert sind (Tong & Farrington, 2006).

Limitationen: „Lessons learned“

Auf die grundsätzlichen Schwierigkeiten von Therapieevaluationen im forensischen Setting wurde bereits an anderer Stelle eingegangen. Diese gelten ausnahmslos auch für den hier ausgewerteten Modellversuch. Dennoch soll nochmals darauf hingewiesen werden, dass auch die Experimentalgruppen, neben der manualisierten Therapie gemäss R&R bzw. ASAT®Suisse begleitende, institutionsspezifische Ein-

zeltherapien erhielten, sich die Experimentalgruppen von den TAU Gruppen also durch ein zusätzliches Therapieangebot unterschieden.
Eine allgemeine Limitation des Modellversuchs ist die begrenzte Follow-Up-Dauer: Während sie sich in der GST-Gruppe mit rund 20 Monaten im Bereich vergleichbarer internationaler R&R-Evaluationen befindet (siehe Antonowicz & Parker, 2012; Tong & Farrington, 2006; Tong & Farrington, 2008), wird in der SST-Gruppe das Kriterium einer Follow-Up-Dauer für Sexualstraftäter von mindestens drei Jahren nicht erreicht. Ein großer Anteil der Studienteilnehmer befand sich zudem zum Katamnese-Zeitpunkt noch nicht in Freiheit, sodass diese nicht für die Analyse der Rückfälligkeit berücksichtigt werden konnten und die ohnehin begrenzte Stichprobengröße noch weiter schrumpfte.
Eine wichtige Limitation stellt die geringe erreichte Power dar: Trotz der beachtenswerten Stichprobengröße in der GST-Gruppe ist diese vor dem Hintergrund des Anteils an Probanden, die für die Analysen berücksichtigt werden konnten, und in Verbindung mit den geringen Effektstärken letztlich nicht hinreichend groß, um tatsächlich bestehende Unterschiede anhand Unterschieden zwischen den Bedingungen im Rahmen des Modellversuchs statistisch aufdecken zu können.
Einerseits aus den genannten Gründen wünschenswert, führt eine lange Studiendauer auf der anderen Seite zu neuen Herausforderungen: So gab es im Laufe und unmittelbar nach Beendigung des Modellversuchs innerhalb des Forschungsteams mehrere Personalwechsel, die dazu führten, dass mit den Mitarbeitern auch gewisse Informationen verloren gingen, die für die Evaluation nicht mehr zur Verfügung standen. Zudem fand während der Studienlaufzeit innerhalb der Experimentalgruppe der Gewaltstraftäter eine vollständige Umstellung vom klassischen R&R (Ross et al., 1986) zur Kurzform des Programms (R&R2; Ross et al., 2007) statt. Anhand der vorliegenden Daten war eine Unterscheidung zwischen Teilnehmern an den beiden Programmen nicht möglich, sodass sie zu einer Gruppe (GST R&R) zusammengefasst werden mussten. Auch wenn diese Umstellung aus klinisch-praktischen Gesichtspunkten nachvollziehbar und sinnvoll erscheint, hat dies Konsequenzen für die Aussagekraft einer Wirksamkeitsevaluation. Zwar beruhen R&R und R&R2 auf ähnlichen Prinzipien und wurden vom gleichen Erstautor Robert D. Ross entwickelt, dennoch muss von Unterschieden ausgegangen werden. Dies betrifft z.B. die nicht zu vernachlässigende Zeitdauer von über 20 Jahren, die zwischen der Publikation der beiden Varianten liegt, wobei davon auszugehen ist, dass zwischenzeitlich aufgetretene Erkenntnisse zu einem gewissen Grad Eingang in die Weiterentwicklung des Programms gefunden haben. Auch bestehen deutliche Unterschiede in der Dauer der jeweiligen Intervention (gemäß R&R-Manual 35 Sitzungen, gemäß R&R2-Manual 14 Sitzungen). Letztlich ist im vorliegenden Bericht keine Aussage möglich, inwieweit das R&R- oder das R&R2-Programm einen differentiellen Einfluss auf die beobachteten Effekte ausüben.
Neben den genannten Vorteilen eines multizentrischen Ansatzes, liefert ein solcher auch viele Herausforderungen, die im Rahmen des Modellversuchs nicht sämtlich gelöst werden konnten. Diese beginnen mit der Schwierigkeit, zentrale Informationen über alle Zentren und alle Teilnehmer hinweg zu erheben. Dadurch konnten letztlich wichtige Daten wie Beginn und Ende der Intervention oder intramurale Vorkommnisse nicht ausgewertet werden. Als wesentliche Limitation ist zu nennen, dass keine Aussagen über Zeitpunkt und Grund des Ausscheidens aus den Therapieprogram-

men bzw. aus der Studie möglich sind. Das gleiche gilt für den Zeitpunkt innerhalb der Haft, zu welchem die Interventionen durchgeführt wurden.
Ein hoher Anteil an fehlenden Werten in vielen Variablen führte dazu, dass einige der beabsichtigten Auswertungen nicht oder nicht auf die gewünschte Weise vorgenommen werden können. Dies ist einerseits ein bekanntes Phänomen im Straf- und Maßnahmenvollzug. So wurden viele Informationen, wie z.B. Diagnosen, den psychiatrischen Gutachten entnommen. Gutachten lagen jedoch nicht zu jedem Probanden vor bzw. war nicht jeder einzelne Proband überhaupt psychiatrisch begutachtet worden. Dies gilt insbesondere für die Teilnehmer in den Kontrollgruppen. Andererseits scheint es aber auch bei verfügbaren Informationen in den verschiedenen Anstalten an personellen Ressourcen gemangelt zu haben – so waren beispielsweise in vielen Fällen die Fremdbeurteilungs-Versionen der Fragebogen gar nicht bearbeitet worden, insbesondere in den Kontrollgruppen. Hier wird die besondere Herausforderung von Forschung im Strafvollzug deutlich, wo sich die Mitarbeitenden neben den regulären Belastungen ihrer anspruchsvollen Tätigkeit noch zusätzlich mit Anforderungen konfrontiert sehen, die sich durch die Implementierung und Evaluation neuer Behandlungsprogramme ergeben.

Ausblick

Grundsätzlich passen die Anlage und die Ergebnisse des Modellversuches in eine gegenwärtige wissenschaftliche Diskussion, die sich mit der „replication crisis“ in der Klinischen Psychologie und dabei insbesondere in der Psychotherapieforschung auseinandersetzt (Hengartner, 2018). Dabei geht es um die zentrale Frage, warum Effekte von psychologischen Interventionen unter naturalistischen Bedingungen nur selten repliziert werden können. Ein Aspekt davon ist, dass die Wirksamkeit von psychotherapeutischen Interventionen schon aufgrund des sogenannten „publication bias“ systematisch zu hoch eingeschätzt wird (Cristea et al., 2017). Zudem hielten z.B. Fournier et al. (2015) fest, dass der grösste Vorteil der Psychotherapie, gegenüber psychopharmakologischen Interventionen möglicherweise in einer Verbesserung des sozialen Funktionsniveaus liege, dass die (replizierbare) Messung der Verbesserung in diesem Bereich aber noch schwerer sei, als eine Verbesserung der Symptombelastung abzubilden – eine Problematik, die auch im Modellversuch wieder zum tragen kam. Es werden verschiedene Lösungsansätze diskutiert, einer davon ist eine Verbesserung der Zugänglichkeit erhobener Daten, auch für andere Forschungsgruppen („data sharing“) und die Replikation von bereits geprüften Hypothesen/Ansätzen in nachfolgenden Studien. Insbesondere der erste Gedanke ist dabei ein Ansatz, den sich auch das Bundesamt für Justiz zu Nutzen machen könnte. So scheint es sinnvoll bei zukünftig geförderten Projekten die zugrunde liegenden Daten auch anderen Forschungsgruppen zugänglich zu machen, also ein „data availability statement“ zu verankern. Weiterhin könnte der Absicht des Gesetzgebers im Rahmen von Modellversuchen ebenso entsprochen werden, wenn zukünftig auch bereits erhobene Daten evaluiert werden, also retrospektive Studienanlagen akzeptiert werden würden.

Bezüglich des MV ist dennoch zu konstatieren, dass auch wenn die vorliegende Evaluation die erwarteten signifikanten Ergebnisse nicht im erhofften Maße aufzeigen konnte, daraus nicht geschlossen werden kann, dass strukturierte, kognitiv-behaviorale Behandlungsprogramme für Gewalt- und Sexualstraftäter keinen Nutzen haben. Wie bereits erwähnt, führten verschiedene Schwierigkeiten bei der Planung, Durchführung und Auswertung des MV zu einer eingeschränkten Überprüfbarkeit und letztlich zu einer begrenzten Möglichkeit, reliable Aussagen zur Wirksamkeit zu treffen. Gerade die genannten Schwierigkeiten können jedoch für die Planung zukünftiger Evaluationsstudien genutzt werden. Zunächst ist die Wichtigkeit zu betonen, möglichst alle relevanten Informationen detailliert und gründlich erheben zu können – dies ist bereits in der Planungsphase konkret zu bedenken. Für die Umsetzung müssen schließlich die nötigen personellen und finanziellen Ressourcen hierfür zur Verfügung stehen. Dazu sind umfangreiche Kooperationen nötig: Zum Generieren ausreichend großer Stichproben, anhand derer überhaupt erwartete Effekte nachgewiesen werden können, wurde bereits im vorliegenden MV ein multizentrischer Ansatz mit entsprechend hohem personellen, zeitlichen und organisationalen Aufwand gewählt, an dem zahlreiche Institutionen des Strafvollzugskonkordats Nordwest- und Innerschweiz beteiligt waren. Dennoch zeigen die noch immer geringen Fallzahlen, vor allem im Bereich der Sexualstraftäter, die weitere Notwendigkeit von kantons- und auch länderübergreifender Zusammenarbeit auf. Hier zeigte sich auch im Rahmen des MV, dass in der wissenschaftlichen Community kaum Standards für die Durchführung von Therapieevaluationen in forensischen Settings vorhanden sind: So muss bislang jede Forschergruppe ihre eigenen (negativen) Erfahrungen mit den Herausforderungen und Fallstricken in diesem ebenso komplexen wie gesellschaftlich hoch relevanten Bereich sammeln. Es erscheint im Hinblick auf zukünftige Studien zentral, das bestehende „Systemwissen“ der Community zugänglich zu machen. Dies kann zum Beispiel in Form internationaler Forschungsnetzwerke geschehen, aber auch durch das Publizieren „negativer“ Ergebnisse. Neben der Dissemination von Evaluationsergebnissen innerhalb der wissenschaftlichen Community sollte auch der Transfer in die Praxis ein zentrales Ziel zukünftiger Forschung sein (Cullen & Gendreau, 2000; Lipsey & Cullen, 2007). Weiterhin ist auch eine Fortführung des vorliegend evaluierten Modellversuchs denkbar: Z.B. könnte ein erneuter Auszug aus dem Schweizer Strafregister (bei Kontrolle von Haftentlassung, Tod und Wegzug aus der Schweiz) durchgeführt werden und dadurch zum einen die Follow-Up-Dauer verlängert und zum anderen die Anzahl an Studienteilnehmern vergrößert werden, die in Freiheit entlassen sind und damit einer Time at Risk ausgesetzt sind. Zentral ist dabei die Möglichkeit zur Finanzierung solcher Nachverfolgungen, die mehrere Jahre über das Ende der eigentlichen Studie hinausreichen können. Bestehende Studienergebnisse drohen ansonsten auch aus dem einfachen Grund fehlender personeller Ressourcen in der Schublade zu verschwinden. Dies kann wie bereits angedeutet zu einem regelrechten Circulus Vitiosus für die Therapieevaluation in forensischen Settings führen.

Darüber hinaus erscheint die sorgfältige Erfassung dynamischer Faktoren wie dem Institutionsklima oder dem therapeutischen Milieu, aber auch der Motivation der Teilnehmer oder dem Monitoring des Behandlungsprozesses als wichtige Voraussetzung für das Verständnis über die Wirkweise einer erfolgreichen Intervention und der ihr zugrundeliegenden Prozesse (z.B. Wößner & Schwedler, 2014).

Als ergänzendes Effektkriterium zukünftiger Therapieevaluationsstudien erscheinen ökonomische Abwägungen im Sinne von Kosten-Nutzen-Analysen interessant (Tong & Farrington, 2006). Zum einen konnte gezeigt werden, dass die Wirksamkeitsüberprüfung einer Intervention vom gewählten Effektkriterium abhängt (Lösel, 1995). Zum anderen zeigt sich gerade innerhalb staatlicher bzw. kantonaler Institutionen die gesellschaftliche Relevanz von Kostenfragen. Gerade für die Legitimation von aufwändigen Therapieprogrammen für Gewalt- und Sexualstraftäter innerhalb einer Gesellschaft könnten diese Überlegungen zuträglich sein. Bisherige Ansätze aus dieser Perspektive lieferten vielversprechende Ergebnisse, wonach bereits geringfügige Effekte – unabhängig von statistischer Signifikanz – kosteneffizient sein können (Aos, Miller, & Drake, 2006; Aos, Phipps, Barnoski, & Lieb, 2001; Endrass, Rossegger, & Kuhn, 2012).

Obwohl aufgrund der methodischen Bedingungen des Modellversuchs von einer massiv eingeschränkten Fähigkeit auszugehen ist, real existierende Unterschiede (- und damit z.B. Wirksamkeitsnachweise/Therapieerfolge) statistisch überhaupt nachweisen zu können, sind doch einige statistisch signifikante Resultate (aber auch Trends) zu verzeichnen, die dafür sprechen Gewalt- und/oder Sexualstraftätern, die in geschlossen oder offenen Schweizer Vollzugsinstitutionen untergebracht sind, auch zukünftig die untersuchten Therapieprogramme R&R und ASAT Suisse zugänglich zu machen. Bezogen auf den R&R (2) kann dabei auf die verringerte selbst berichtete Aggressivität, die geringer ausgeprägte feindselige Attribution in Alltagssituationen mit Konfliktpotenzial und auf die erhöhte Verantwortungsübernahme für das Delikt im Vergleich zu den unbehandelten Kontrollprobanden verwiesen werden. Sowohl für den R&R(2) als auch für das ASAT®Suisse sprechen zudem eine tendenziell höhere Compliance im Vergleich zur Treatment as Usual Vergleichsgruppe.

Anhang

Anhang 1. Stichprobenzusammensetzung GST (Nur Probanden mit erfolgreicher Teilnahme an der Studie bzw. Treatment as Delivered)

In der nachfolgenden Stichprobenbeschreibung des Teils der GST-Gruppe, welche den MV bis zum Ende durchlaufen hat (Vollender), werden die Probanden in den drei Bedingung im Hinblick auf demografische Merkmale sowie weitere Merkmale mit Einfluss auf das Rückfallrisiko miteinander verglichen. Damit soll auf vorbestehende Unterschiede hinsichtlich derjenigen Probanden überprüft werden, die für eine Beantwortung der Fragestellung 1 (Veränderungen in den Messwerten der Fragebögen) berücksichtigt werden konnten.

Demografische Merkmale: Unterschiede nach Bedingung (Vollender)

In den folgenden Tabellen sind demografische Merkmale der Studienteilnehmer (ST) der GST-Gruppe in den drei unterschiedlichen Bedingungen aufgeführt. Es wurden darin ausschließlich Probanden berücksichtigt, für die Daten zu T1 und T2 vorlagen, also ausschließlich diejenigen Probanden, die den MV bis zum Ende absolvierten.

Nationalität

Tabelle 43 zeigt den Anteil der Schweizer Probanden in den drei Bedingungen der GST-Gruppe, die die Studie bis zum Ende absolviert haben. Ein Chi-Quadrat-Test zeigte signifikante Unterschiede zwischen den drei Bedingungen der GST-Gruppe im Anteil der Vollender mit Schweizer Nationalität ($\chi^2(2)=6.24$; $p=.044$): In der Kontrollgruppe waren signifikant mehr Ausländer vertreten als in der Experimentalgruppe ($\chi^2(1)=6.12$; $p=.013$).

Tabelle 43. Nationalität GST (ohne Abbrecher)

GST Bedingung	Fehlende Werte [% (*N*)]	*N*	Schweizer Nationalität [% (*N*)]	*p*
R&R (Vollender)	0% (*N*=0)	110	72.7% (*N*=80)	
TAU (Vollender)	0% (*N*=0)	57	68.4% (*N*=39)	.044
KG (Vollender)	0% (*N*=0)	51	52.9% (*N*=27)	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe.

Alter

Tabelle 44 zeigt das Alter der bis zum Ende teilnehmenden Probanden der GST-Gruppe zum Zeitpunkt des Indexurteils. Ein Kruskal-Wallis-H-Test lieferte keine Hin-

weise auf Altersunterschiede der Probanden zum Zeitpunkt des Indexurteils zwischen den drei Bedingungen der GST-Gruppe ($\chi^2(2)=2.08$; $p=.354$).

Tabelle 44. Alter GST (ohne Abbrecher)

GST Bedingung	Fehlende Werte [% (*N*)]	*N*	Alter Indexurteil [*M* (*SA*)]	*p*
R&R (Vollender)	3.6% (*N*=4)	106	31.5 (8.9)	
TAU (Vollender)	5.3% (*N*=3)	54	30.8 (10.5)	.354
KG (Vollender)	7.8% (*N*=4)	47	32.6 (9.6)	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe.

Bildungsniveau

Tabelle 45 zeigt das Bildungsniveau der Probanden in den drei Bedingungen der GST-Gruppe, die die Studie bis zum Ende absolviert haben.

Ein exakter Test nach Fisher zeigte signifikante Unterschiede im höchsten erreichten Bildungsniveau zwischen den Probanden der GST-Gruppe, die den MV bis zum Ende absolviert haben ($p=.042$). Die durchgeführten Post-Hoc-Tests waren nach Bonferroni-Korrektur des Alpha-Niveaus allerdings nicht signifikant. Es gibt Hinweise, dass der erhöhte Anteil an ausländischen Probanden in der Kontrollgruppe tendenziell mit einem niedrigeren Anteil an Schweizer Probanden einhergeht, die die Pflichtschule abgeschlossen haben, sowohl in Bezug zur Experimental- ($\chi^2(1)=5.23$; $p=.022$) als auch zur Vergleichsgruppe ($\chi^2(1)=5.54$; $p=.019$).

Vergleicht man die höchsten erreichten Bildungsabschlüsse ausschließlich unter den Schweizer Probanden, gibt ein exakter Test nach Fisher keine Hinweise mehr auf signifikante Unterschiede zwischen den Bedingungen ($p=.186$).

Tabelle 45. Bildungsniveau GST (ohne Abbrecher)

GST Bedingung	Fehlende Werte [% (*N*)]	*N*	Höchster Abschluss	Anteil [% (*N*)]	*p*
R&R (Vollender)	5.5% (*N*=6)	104	< 7 Jahre Schulbildung	2.9% (*N*=3)	.042
			Schulpflicht abgeschl. ggf. zzgl. Anlehre	35.6% (*N*=37)	
			Mindestens abgeschl. Lehre	32.7% (*N*=34)	
			Nicht-Schweizer Probanden	28.9% (*N*=30)	
TAU (Vollender)	7.0% (*N*=4)	53	< 7 Jahre Schulbildung	0% (*N*=0)	
			Schulpflicht abgeschl. ggf. zzgl. Anlehre	47.2% (*N*=25)	
			Mindestens abgeschl. Lehre	18.9% (*N*=10)	
			Nicht-Schweizer Probanden	34.0% (*N*=18)	
KG (Vollender)	7.8% (*N*=4)	47	< 7 Jahre Schulbildung	2.1% (*N*=1)	
			Schulpflicht abgeschl. ggf. zzgl. Anlehre	23.4% (*N*=11)	

	Mindestens abgeschl. Lehre	23.4% (*N*=11)
	Nicht-Schweizer Probanden	51.1% (*N*=24)

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe.

Psychiatrische Belastung: Unterschiede zwischen den Bedingungen (Vollender)

Tabelle 46 zeigt deskriptiv die Diagnosen gemäß (aktuellstem) psychiatrischen Gutachten derjenigen Probanden der GST-Gruppe, die den MV bis zum Ende durchliefen. Die Diagnosebereiche umfassen Persönlichkeitsstörungen (Diagnose aus ICD-10-Kategorie F6), Störungen aus ICD-10-Kategorie F2 (Schizophrenie, schizotype und wahnhafte Störungen), Störungen durch psychotrope Substanzen (ICD-10 Kategorie F1) sowie affektive Störungen (ICD-10 Kategorie F3). Der Zeitpunkt der Diagnosestellung wurde nicht spezifisch erfasst. In der Kontroll- und teilweise auch der Vergleichsgruppe ist ein substanzieller Anteil an fehlenden Werten zu verzeichnen. Ein Chi-Quadrat Test zwischen den Bedingungen R&R und TAU, die berechnet wurden, sofern maximal 20% fehlende Werte vorlagen (im Fall von Substanzstörungen), ergab keine signifikanten Ergebnisse.

Tabelle 46. Psychiatrische Belastung GST (ohne Abbrecher)

	GST Bedingung	Fehlende Werte [% (*N*)]	*N*	Anteil [% (*N*)]	*p*
Persönlichkeitsstörung (ICD-10 Kategorie F6)	R&R Vollender	12.7% (*N*=14)	96	**56.3% (*N*=54)**	N/A
	TAU (Vollender)	24.6% (*N*=14)	43	**48.8% (*N*=21)**	
	KG (Vollender)	45.1% (*N*=23)	28	**46.4% (*N*=13)**	
Schizophrenie, schizotype und wahnhafte Störungen (ICD-10 Kategorie F2)	R&R Vollender	13.6% (*N*=15)	95	**6.3% (*N*=6)**	N/A
	TAU (Vollender)	21.1% (*N*=12)	45	**11.1% (*N*=5)**	
	KG (Vollender)	47.1% (*N*=24)	27	**0% (*N*=0)**	
Störungen durch psychotrope Substanzen (ICD-10 Kategorie F1)	R&R Vollender	11.8% (*N*=13)	97	**55.7% (*N*=54)**	N/A
	TAU (Vollender)	19.3% (*N*=11)	46	**43.5% (*N*=20)**	
	KG (Vollender)	43.1% (*N*=22)	29	**37.9% (*N*=11)**	
Affektive Störungen (ICD-10 Kategorie F3)	R&R Vollender	14.6% (*N*=16)	94	**7.4% (*N*=7)**	N/A
	TAU (Vollender)	22.8% (*N*=13)	44	**4.5% (*N*=2)**	
	KG (Vollender)	47.1% (*N*=24)	27	**7.4% (*N*=2)**	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; **Rot** = mehr als 20% fehlende Werte.

Therapieerfahrung: Unterschiede zwischen den Bedingungen (Vollender)

Tabelle 47 zeigt den Anteil derjenigen Probanden der GST-Gruppe, die bereits vor Beginn des MV eine psychotherapeutische Intervention erhalten hatten (nur Vollender). Ein Chi-Quadrat-Test zwischen den Bedingungen R&R und TAU ergab einen signifikant höheren Anteil an Probanden in der Experimentalgruppe, die bereits vor Beginn des MV mindestens einmalig eine psychotherapeutische Behandlung erhalten hatten ($\chi^2(1)=8.83$; $p=.003$).

Tabelle 47. Therapieerfahrung GST (ohne Abbrecher)

GST Bedingung	Fehlende Werte [% (*N*)]	*N*	Psychotherapie-Erfahrung [% (*N*)]	*p*
R&R Vollender	16.4% (*N*=18)	92	59.8% (*N*=55)	
TAU Vollender	15.8% (*N*=9)	48	33.3% (*N*=16)	*N/A*
KG Vollender	37.3% (*N*=19)	32	43.8% (*N*=14)	

Anmerkungen. GST: Gruppe der Gewaltstraftäter; R&R: Reasoning and Rehabilitation Programm; TAU: Treatment as Usual (Vergleichsgruppe); KG: Kontrollgruppe; **Rot** = Mehr als 20% fehlende Werte.

Anhang 2. Beschreibung der Interventionen

Reasoning and Rehabilitation Programm (R&R bzw. R&R2; Ross et al., 1986; Ross et al., 2007)[13]

Das „Reasoning and Rehabilitation“-Programm (R&R) ist ein multimodales, kognitiv-behaviorales Behandlungsprogramm für Gewaltstraftäter, das im Gruppensetting durchgeführt wird. Es beruht auf dem Modell von Ross und Fabiano (1985), wonach Straftäter Defizite in ihren kognitiven Fähigkeiten aufweisen und es ihnen dadurch an sozialen Kompetenzen mangelt.

Durch die Vermittlung kognitiver Fertigkeiten wird versucht, Denkprozesse dahingehend zu verändern, dass Straftäter angemessener auf Situationen reagieren können, die als Auslöser kriminellen Verhaltens auftreten (Tong & Farrington, 2006). Dabei handelt es sich um folgende Fertigkeiten Ebene von Denken und Handeln: Soziale Kompetenz, kreatives Denken, Selbstbehauptungstraining, Gesprächsführung (Verhandlungsfertigkeiten), zwischenmenschliches Training und soziale Perspektivübernahme (Ross, Fabiano, & Ewles, 1988). Diese Fertigkeiten werden in modularer Form trainiert, setzen jedoch Bezüge zueinander, um so eine Anwendung in verschiedenen Kontexten zu ermöglichen (Ross & Ross, 1995). Als Techniken kommen Gruppendiskussionen, Sokratische Dialoge, Modelling, Rollenspiele, Denkspiele, Problemlösen, Hausaufgaben, Dilemma-Diskussionen und Entspannungstechniken zum Einsatz. Den Gewaltstraftätern soll damit eine Perspektive zu kriminellem Verhalten in Form von prosozialem Verhalten aufgezeigt werden.
Entwickelt wurde das Programm in Kanada, mittlerweile ist es jedoch eines der weltweit am verbreitetsten Straftäterprogramme: Nach aktuellen Kalkulationen haben bisher über 70‘000 Probanden am R&R-Programm aus zahlreichen Ländern teilgenommen, darunter Australien, Dänemark, Deutschland, Estland, Großbritannien, Niederlande, Hongkong, Kanada, Japan, Jersey, Lettland, Libanon, Neuseeland, Norwegen, Schottland, Schweden, Schweiz, Spanien, USA und die Vereinigten Arabische Emirate (Gretenkord, 2015).
Die Wirksamkeit von R&R konnte über verschiedene Länder (Kanada, USA, Großbritannien und Schweden) und verschiedene Settings (ambulant und stationär) hinweg gezeigt werden: In einer Metaanalyse von Tong und Farrington (2008) ergab sich über 19 Studien, die insgesamt 32 Vergleiche vornahmen, eine gemittelte Reduktion der Rückfallrate um 14% für R&R-Teilnehmer im Vergleich zu Kontrollgruppen. Letztere waren etwa je zur Hälfte randomisiert ausgewählt bzw. hinsichtlich relevanter Variablen parallelisiert worden. Obwohl das Programm für Täter mit hohem Rückfallrisiko konzipiert worden war, zeigte sich der Effekt sowohl für diese als auch für Täter mit geringem Rückfallrisiko (Tong & Farrington, 2006, 2008).
Das Manual liegt auch in einer deutschsprachigen Version vor, die vom Institut für forensische Psychiatrie Haina e.V. (Deutschland) im Rahmen von Workshops zur Erlangung eines Zertifikats verwendet wird, das die Teilnehmer zur Anwendung des

[13] Die folgenden Informationen stellen eine Zusammenfassung der angeführten Primär- und Sekundärliteratur zum „Reasoning and Rehabilitation“-Programm dar, sofern nicht anders angegeben.

R&R-Programms befähigt und ermächtigt. Außerhalb der Zertifizierungskurse ist das Manual nicht erhältlich, um so eine hohe Programmintegrität zu gewährleisten (Eucker, 2013; Institut für forensische Psychiatrie Haina e.V., 2017).

Aufgrund des Nachweises der besseren Wirksamkeit kürzerer Interventionen fand im Verlauf des MV eine vollständige Umstellung auf die Kurzfassung des R&R-Programms, das R&R2 für Erwachsene (Ross et al., 2007) statt. Es besteht aus insgesamt 14 Sitzungen mit einer Dauer von je 90 Minuten. Das Manual empfiehlt idealerweise eine Frequenz von zwei bis drei Sitzungen pro Woche, erlaubt dabei jedoch explizit Anpassungen an die jeweiligen Bedingungen am Ort der Durchführung. Im Rahmen des MV fanden die Sitzungen in der Regel einmal wöchentlich statt.
Mit der Entwicklung des R&R2-Programms sollte insbesondere dem Responsivity- (Ansprechbarkeits-) Prinzip Rechnung getragen werden: Das Programm liegt in Versionen für verschiedene Klientengruppen vor, darunter diejenige für Erwachsene. Alle Techniken des R&R finden sich auch im R&R2 wieder, einige wurden jedoch nach neuen wissenschaftlichen Erkenntnissen aktualisiert und die Module „Emotionale Kompetenz“ sowie „Ausgleich zwischen Gedanken, Gefühlen und Verhalten“ hinzugefügt (Ross et al., 2007).

Anti-Sexuelle-Aggressivität-Training (ASAT®; Steffes-enn, 2005; 2008)[14]

Das ASAT® ist ein multimodales, strukturiertes, deliktorientiertes Trainingsprogramm zur Behandlung von Sexualstraftätern mit stark ausgeprägter Aggressionsproblematik. Es beinhaltet sowohl kognitiv-behaviorale als auch systemische Elemente und wird im Gruppensetting durchgeführt. Entwickelt wurde das ASAT® 2001 in der Sozialtherapeutischen Abteilung der Justizvollzugsanstalt Amberg (Steffes-enn, 2008). Ziel ist die Rückfallvermeidung durch Verbesserung der dynamischen kriminogenen Risikomerkmale und durch Steigerung der Empathie und Verantwortungsübernahme für die Tat. Das ASAT® ist eine Weiterentwicklung des Anti-Aggressivitäts-Trainings (AAT) von Weidner (1995), welches nach Farrelly und Matthews (1994) auf der Grundlage von Psychodrama nach Moreno, der Gestalttherapie nach Perls und der Provokativen Therapie entwickelt wurde. Das ASAT® orientiert sich konzeptuell sowohl am verhaltenstherapeutischen als auch am systemischen Ansatz. Der Bearbeitung des deliktspezifischen Verhaltens werden Täterklassifizierungen (Kraus & Berner, 2000; Rehder, 2001) zugrunde gelegt. Außerdem werden Erkenntnisse aus der Bindungstheorie berücksichtigt (Marshall, Anderson, & Fernandez, 1999; Rutrecht, Jagsch, & Kryspin-Exner, 2002). Das ASAT® soll die Teilnehmer befähigen, sich individueller Bewertungs- und Bindungsmuster bewusst zu werden, ihre Empathiefähigkeit zu steigern, und zu erkennen, dass die Schadenszufügung mit einer Steigerung des Selbstwertgefühls ein-hergeht (Steffes-enn, 2008). So geht auch (Sachse, 2006) bei persönlichkeitsgestörten Menschen von einem negativen Selbstkonzept aus, welches zu destruktivem Verhalten im Rahmen von Interaktionen führt. Im Sinne des Ansprechbarkeitsprinzips wird auch im Rahmen des ASAT® mit verschiedenen Methoden (Rollenspiele, audio-visuelle Präsentationen, Interaktions-

[14] Die folgenden Informationen stellen eine Zusammenfassung der angeführten Primärliteratur zum ASAT® dar, sofern nicht anders angegeben.

spiele, Meditationen) gearbeitet. Das ASAT® besteht aus 17 Modulen, welche in drei Phasen bearbeitet werden: der Integrationsphase (8 Module), der Konfrontationsphase (3 Module) und der Gewaltverringerungsphase (6 Module). Die Teilnehmer erhalten zudem pro Einheit/Block eine Hausaufgabe und müssen während des gesamten Kurses ein Tagebuch über Stärken und Emotionen führen.
Die Wirksamkeit des ASAT® wurde bisher im Rahmen zweier Diplomarbeiten überprüft, welche bei Absolventen des ASAT® eine höhere Empathiebildung im Vergleich zur Kontrollgruppe (Teilnehmer der anstaltsüblichen Sozialtherapie) nachweisen konnten (Teichmann & Müller, 2005, zit. nach Steffes-enn, 2008). Das ASAT® wurde 2008 erstmals durch den FPD in der Schweiz eingesetzt und zu einer für die Schweiz adaptierten Version weiterentwickelt, dem ASAT®Suisse (Falk & Steffes-enn, 2014), wofür eine Evaluation bislang noch aussteht (FPD Bern, 2009).

Einzeltherapien in den Vergleichsgruppen

Die Probanden der beiden Vergleichsgruppen durchliefen die in der jeweiligen Anstalt übliche Standard-Behandlung für Gewalt- bzw. Sexualstraftäter (Treatment as usual) in Form von Einzeltherapie. Dabei handelt es sich um eine Störungs- und deliktorientierte psychotherapeutische Einzelbehandlung. Diese Intervention ist als Alternativ-Behandlung zu den beiden zu evaluierenden Behandlungsprogrammen R&R2 und ASAT®Suisse konzipiert, da sie weitgehend dieselben übergeordneten Ziele verfolgt: Durch die psychotherapeutische Einzelbehandlung sollen die Symptome der psychischen Störung gebessert und die kriminogenen Faktoren verändert werden, welche mit der Straftat in Zusammenhang stehen. Zunächst erfolgt hierbei eine sorgfältige diagnostische Abklärung. Daraufhin wird ein Behandlungsplan erstellt, welcher die kurz-, mittel- und langfristigen Therapieziele benennt, regelmäßig evaluiert und, wenn nötig, modifiziert wird. Im Rahmen der deliktorientierten Behandlung werden beim FPD Bern Konfrontations- und Bewältigungsverfahren, sozial-lerntheoretisch und kognitiv geprägte Verfahren sowie Selbstkontroll- und Selbstmanagement-Verfahren eingesetzt (Ermer, 2008).

Anhang 3. Beschreibung der verwendeten Instrumente[15]

Risk-Assessment-Instrumente

Psychopathy Checklist revised (PCL-R; Hare, 2003)
Das Konstrukt der Psychopathie im Sinne von Hare („Psychopathy“) wurde mittels der PCL-R (Hare, 2003) erfasst. Die 20 Items messen Persönlichkeitseigenschaften und Verhaltensweisen, die mit traditionellen Konzeptionen von Psychopathie (Berrios, 1996; Cleckley, 1976; Hare & Cox, 1978) einhergehen. Die Informationen werden durch ein halbstrukturiertes Interview, Akteninformationen sowie itemspezifische Bewertungskriterien gewonnen. Jedes der Items wird auf einer dreistufigen Skala beantwortet (0, 1, 2). Der Summenwert (zwischen 0 und 40) gibt die Übereinstimmung des beurteilten Individuums mit einem prototypischen Psychopathen an (Hare, Black, & Walsh, 2013). Klassischerweise wird ab einem Summenwert von 30 ein Proband als psychopathisch klassifiziert (Hare, 2003). Verschiedene Cut-Off-Werte, unter anderem in Abhängigkeit vom jeweiligen Land, und auch dimensionale Interpretationen des PCL-Summenwertes werden jedoch diskutiert (z.B. Mokros et al., 2011). Mittels faktorenanalytischer Auswertungen konnte gezeigt werden, dass sich die anhand von 20 Items operationalisierte psychopathische Persönlichkeit auf zwei Faktoren abbilden lässt (Hare, 2003): Faktor 1 spiegelt interpersonale und affektive Auffälligkeiten wider (z.B. Empathiemangel), Faktor 2 einen sozial devianten Lebensstil (z.B. Verantwortungslosigkeit).
Sämtliche Projektmitarbeiter waren durch den Besuch eines Workshops in dem Verfahren geschult worden und sind dadurch zur Beurteilung mit der PCL-R zertifiziert. Zusätzlich wurden die Beurteiler supervidiert.

Violence Risk Appraisal Guide (VRAG; Quinsey et al., 2006)
Der VRAG ist ein aktuarisches Risk-Assessment-Instrument für Gewaltstraftäter, der statische Risikomerkmale erfasst. Straftäter können anhand des VRAG-Summenwertes einer Risikogruppe mit einer vergleichbaren Merkmalskombination zugeordnet werden, deren Rückfallrisiko für erneute Anklagen und Verurteilungen aufgrund eines Gewalt- oder Sexualdelikts für einen Sieben- und für einen Zehn-Jahres-Zeitraum bekannt ist. Trotz der retrospektiven Erfassung des VRAG-Summenwertes konnte die prädiktive Validität gezeigt werden. Der VRAG liegt in einer autorisierten deutschsprachigen Übersetzung vor (Rossegger, Urbaniok, Danielsson, & Endrass, 2009).

Static-99 (Hanson & Thornton, 1999)
Die Wahrscheinlichkeit sexueller und sexuell-gewalttätiger Rückfälle von Straftätern, die bereits mindestens einmal wegen einer Sexualstraftat angeklagt worden waren, kann durch das Static-99 (Hanson & Thornton, 1999, 2000) eingeschätzt werden. Das Instrument besteht aus 10 statischen Risikomerkmalen aus den Bereichen Demografie, Kriminelle Vorgeschichte und Opfermerkmale. Je nach Gesamtwert auf der

[15] Die Folgenden Beschreibung stellen eine Zusammenfassung der Primär- bzw. Sekundärliteratur zum jeweiligen Instrument dar, sofern nicht anders angegeben.

Skala wird der Straftäter einer von vier Risikokategorien (niedriges, niedriges bis moderates, moderates bis hohes, hohes Rückfallrisiko) zugeteilt. Die Grenzwerte für die Risikogruppen wurden empirisch ermittelt (Hanson & Thornton, 2000) und die prädiktive Validität des Static-99 konnte mehrfach bestätigt werden (Hanson, 2005; Hanson & Morton-Bourgon, 2005; Hanson & Thornton, 2000; Helmus & Hanson, 2007).

Fragebögen

Kurzfragebogen zur Erfassung von Aggressivitätsfaktoren (K-FAF; Heubrock & Petermann, 2008)

Der K-FAF ist die überarbeitete Kurzform des Fragebogens zur Erfassung von Aggressivitätsfaktoren (FAF; Hampel & Selg, 1998). Es handelt sich um ein Selbstbeurteilungsinstrument, das über 49 Aussagen verschiedene Aspekte aggressiven Verhaltens erfasst: Spontane Aggressivität (12 Items; Bsp.: „Manchmal gefällt es mir, andere zu quälen“), Reaktive Aggressivität (11 Items; Bsp.: „Ich schlage lieber zu, als feige zu sein“), Erregbarkeit (10 Items; Bsp.: „Leider werde ich schnell wütend“), Selbstaggressivität (9 Items; Bsp.: „Ich tue vieles, was ich hinterher bereue“) und Aggressions-Hemmung (7 Items; Bsp.: „Bevor es zum Streit kommt, gebe ich lieber nach“).

Die Probanden geben ihre Zustimmung bzw. Ablehnung zu den Aussagen über eine sechsstufige Likert Skala an, bei der lediglich die Pole eine inhaltliche Aussage besitzen (0 = „trifft überhaupt nicht zu“ und 5 = „trifft voll und ganz zu“).

Die Skalenwerte werden über den Summenwert der zur jeweiligen Skala gehörigen Items gebildet.

Als Maß für die nach außen gerichtete Aggressivität werden die Skalenwerte der ersten drei Skalen zu einem Summenwert zusammengefasst (Summe der Aggressivität).

Für die Zwecke des Modellversuchs kam zudem eine Fremdbeurteilungs-Version des K-FAF zur Anwendung (Quelle: Aktenbooklet): Die fünf Skalen sind analog zu den Skalen der Selbstbeurteilung konzipiert. Sie werden in der Fremdbeurteilung durch jeweils zwei Items gebildet (Bsp. für ein Item der Skala „Spontane Aggressivität“: „In der Gruppe hat der Patient oft Lust Schaden anzurichten“).

Inventar zur Erfassung interpersonaler Probleme (IIP-D; Horowitz et al., 2000)

Das IIP-D ist ein Fragebogen zur Selbsteinschätzung interpersonaler Probleme, d.h. zu Schwierigkeiten im Umgang mit anderen Menschen. Für die vorliegende Studie wurde die aus 64 Items bestehende Version verwendet. Für das Instrument liegen repräsentative Normstichproben vor. Das Inventar erfragt interpersonale Verhaltensweisen, die den Probanden entweder schwer fallen oder die die Probanden im Übermaß zeigen. Es beruht auf dem Circumplex-Modell interpersonalen Verhaltens nach Leary (1957; zit. nach Horowitz et al., 2000). Danach lassen sich interpersonale Verhaltensweisen in einem zweidimensionalen semantischen Raum abbilden, mit den Dimensionen Zuneigung (Extreme: feindseliges versus freundliches/ liebevolles Verhalten) und Kontrolle oder Dominanz (Extreme: dominierendes versus unterwürfiges Verhalten). Die acht Skalen des IIP-D sind innerhalb dieses zweidimensionalen

Raums angeordnet, wobei jede der Skalen eine unterschiedliche Kombination von Ausprägungen der zwei Dimensionen repräsentiert.
Bei den Skalen handelt es sich um die mittels Faktorenanalyse gebildeten Skalen „PA“ (Zu autokratisch/dominant; Bsp.: „Ich bin zu unabhängig“), „BC“ (Zu streitsüchtig/konkurrierend; Bsp.: „Ich streite mich zu viel mit anderen“), „DE“ (Zu abweisend/kalt, Bsp.: „Ich halte mir andere zu sehr auf Distanz“), „FG“ (Zu introvertiert/sozial vermeidend; Bsp.: „Es fällt mir schwer, mich Gruppen anzuschließen“), „HI“ (Zu selbstunsicher/unterwürfig; Bsp.: „Es fällt mir schwer, andere mit anstehenden Problemen zu konfrontieren“), „JK“ (Zu ausnutzbar/nachgiebig; Bsp.: „Es fällt mir schwer, anderen gegenüber ‚Nein‘ zu sagen“), „LM“ (Zu fürsorglich/freundlich; Bsp.: „Ich bin anderen gegenüber zu großzügig“) und „NO“ (Zu expressiv/aufdringlich; Bsp.: „Es fällt mir schwer, bestimmte Dinge für mich zu behalten“). Die Items werden auf einer fünfstufigen Likert-Skala beantwortet, die von 0 („nicht“) bis 4 („sehr“) reicht. Jeweils acht Items bilden eine Skala. Die Skalenwerte werden als Summenwerte der zugehörigen Items gebildet.
Die interne Konsistenz der Skalen konnte mittlerweile aufgezeigt werden: Nachdem Cronbachs Alpha der einzelnen Skalen zunächst zwischen α=.36 und α=.64 angegeben wurde (Horowitz et al., 2000), liegen die Konsistenzkoeffizienten der einzelnen Skalen inzwischen im Bereich von α=.71 und α=.82 (Horowitz, Strauß, Thomas, & Kordy, 2016). Neben den Skalenwerten wird auch ein Gesamtwert (Quotient der acht Skalensummen geteilt durch die Anzahl der Skalen) gebildet, der das Ausmaß an interpersonaler Problematik charakterisiert.

Indem die interpersonalen Probleme systematisch auf mehreren Dimensionen beschrieben werden, sollen interindividuelle Vergleiche von Personen in Bezug zu wichtigen Persönlichkeitsdimensionen ermöglicht werden. Es wird dabei angenommen, dass sich die geschilderten interpersonalen Probleme einer Person auf ein relativ invariantes, konsistentes und zeitlich stabiles Muster zurückführen lassen. Diesem Muster liegen bestimmte Eigenschaftsdimensionen zugrunde, denen die zugehörigen Skalenwerte einer Person zugeordnet werden können (Horowitz et al., 2000).

Für die Zwecke der vorliegenden Studie kam zudem eine Fremdbeurteilungs-Version des IIP-D zur Anwendung (Quelle: Aktenbooklet): Dazu wurde zu jeder der acht Skalen des IIP-D ein Item formuliert, welches eine Einschätzung der interpersonalen Probleme des Probanden durch einen Behandler entsprechend dem Aufbau des IIP-D ermöglicht. Die Skalen sind in diesem Item direkt benannt, z.B. lautet das Item für die erste Skala „Bitte geben Sie untenstehend an, inwieweit Sie Ihren Patienten als autokratisch/dominant erleben“. Die Antwort wurde analog zur Selbsteinschätzung auf einer fünfstufigen Likert-Skala mit Werten zwischen 0 („nicht“) und 4 („sehr“) gegeben.

Hostile Attribution Bias (HAB; Tremblay & Belchevski, 2004)
Mit dem HAB wird die Attribution feindseliger Absichten als Reaktion auf eine potenziell provozierende Situation erfasst. Die Aggressionsforschung konnte die Beziehung zwischen aggressivem Verhalten und der Aggressionsneigung als Persönlichkeitseigenschaft aufzeigen (z.B. Bushman, 1995; Giancola, 2002; Holtzworth-Munroe, Bates, Smutzler, & Sandin, 1997). Außerdem konnte gezeigt werden, dass

eine hohe Aggressionsneigung mit der Tendenz einhergeht, dem Interaktionspartner eine feindselige Absicht zu unterstellen, vor allem in uneindeutigen Situationen (Crick & Dodge, 1994; Dill et al., 1997; Dodge, 1980; Matthews & Norris, 2002). Eine wahrgenommene feindselige Absicht des Gegenübers ist zugleich ein wichtiger Prädiktor für zukünftiges aggressives Verhalten gegenüber dieser Person (Geen, 2001). Der Einfluss der Aggressionsneigung auf aggressives Verhalten wird hierbei durch die wahrgenommene feindselige Absicht des Gegenübers moderiert (Tremblay & Belchevski, 2004).

Die englischsprachige Original-Version des HAB von Tremblay und Belchevski (2004) besteht aus 24 Situationsvignetten, in die sich der Proband hineinversetzen soll. Die Vignetten beschreiben ein Verhalten einer Person oder einer Gruppe, das zur Provokation geeignet ist, in folgenden drei Varianten: 1) Situationen, in denen die Person/Gruppe die Zielperson klar provoziert, 2) Situationen, in denen die Absicht der Person/Gruppe unklar ist, und 3) Situationen, in denen die Person/Gruppe die Zielperson eindeutig nicht provoziert.

Der im Rahmen des Modellversuchs verwendete Fragebogen zum Hostile Attribution Bias besteht aus 12 der ursprünglich 24 Situationsvignetten, die vom FPD-Studienteam ins Deutsche übersetzt wurden. Zur Gewährleistung der Korrektheit erfolgte eine Rückübersetzung der 12 Vignetten ins Englische. Uneindeutige Formulierungen wurden in diesem Schritt entfernt.

Für jede der 12 Vignetten werden die Probanden gefragt, wie wahrscheinlich Sie in dieser Situation verärgert bzw. aggressiv reagieren würden. Die Wahrscheinlichkeit möglicher Reaktionen des Probanden wird in sieben Items erfragt (Quelle: Aktenbooklet):

a) Wie wahrscheinlich ist es, dass es sich bei dem Verhalten um eine absichtliche Provokation handelt?
b) Wie wahrscheinlich wären Sie über die Situation verärgert?
c) Für wie wahrscheinlich halten Sie es, dass Sie ihm sagen würden, dass Sie verärgert sind?
d) Für wie wahrscheinlich halten Sie es, dass Sie sich unhöflich ihm gegenüber verhalten würden?
e) Für wie wahrscheinlich halten Sie es, dass Sie ihn anschreien oder beschimpfen würden?
f) Für wie wahrscheinlich halten Sie es, dass Sie ihm drohen würden, wenn sich die Situation nicht klären lässt?
g) Für wie wahrscheinlich halten Sie es, dass Sie physische Gewalt gegen ihn einsetzen würden (ihn z.B. stossen oder packen), wenn sich die Situation nicht klären lässt?

Diese sieben Items werden auf einer fünfstufigen Likert-Skala beantwortet (0 = „Keinesfalls“; 1 = „Wahrscheinlich nicht“; 2 = „Vielleicht“; 3 = „Ziemlich wahrscheinlich“; 4 = „Sicher“).

Bei acht der zwölf ins Deutsche übersetzten Vignetten handelt es sich um Situationen mit unklarer Provokations-Absicht (z.B. „Sie gehen zur Arbeit und haben schlechte

Laune. Als Sie das Büro betreten, macht einer Ihrer Kollegen eine belustigende Bemerkung über Ihre Kleidung“). Jeweils zwei Vignetten beinhalten eine klar provozierende Absicht (z.B. „Sie überqueren zu Fuss eine belebte Kreuzung, wobei klar ist, dass Sie den Vortritt haben. Ein Mann, der mit seinem Auto nach rechts abbiegen möchte, fährt Sie dabei fast an. Er bremst mitten auf der Strasse und schreit Sie an: ‚Du blöder Idiot‘. Dann fährt er in eine Parklücke einige Meter entfernt“) bzw. eine eindeutig nicht provozierende Absicht (z.B. „Eine Gruppe Jugendlicher spielt im Park Frisbee. Der Frisbee trifft Sie ziemlich stark an der Nase. Einer der Jugendlichen läuft zu Ihnen hinüber, um zu schauen, ob es Ihnen gut geht“).

Für die Zwecke des Modellversuchs kam zusätzlich eine von den Behandlern vorzunehmende Fremdbeurteilung des Ausmaßes an Attribution feindseliger Absichten zur Anwendung (Quelle: Aktenbooklet): Dieses bezieht sich ausschließlich auf Situationen mit uneindeutigen Absichten. Es besteht aus einem einzigen Item („Zuschreibung feindseliger Absichten: Zeigt der Patient Tendenz, in uneindeutigen, möglichen Konfliktkonstellationen die Absicht des Gegenübers als feindselig zu interpretieren?“). Die Beantwortung erfolgt analog zur Selbstbeurteilung auf einer vierstufigen Likert-Skala mit den identischen Abstufungen.

Fragebogen zur Verantwortungsübernahme (VÜ; Gabriel et al., 2005; Oswald & Bütikofer, 2002)

Der Fragebogen zur Verantwortungsübernahme ist ein Selbstbeurteilungsinstrument. Er besteht aus 20 Aussagen, welche Rechtfertigungsgründe und Entschuldigungen in Bezug auf die Tat erfassen und somit das Ausmaß der Verantwortungsübernahme der Probanden hinsichtlich ihrer Delinquenz erfassen. Darin enthalten sind Items der Skala „Einstellung zum eigenen Delikt“ von Ortmann (1987) sowie weitere, in Anlehnung an bestehende Fragebögen (z.B. Schahn, Dinger, & Bohner, 1995) modifizierte und neu konstruierte Aussagen (Gabriel et al., 2005).

Die Probanden geben den Grad ihrer Zustimmung zu jeder der 20 Aussagen auf einer vierstufigen Likert-Skala an (1 = „Stimmt nicht“; 2 = „Stimmt eher nicht“; 3 = „Stimmt eher“; 4 = „Stimmt“). Je höher die Zustimmung ist, desto geringer ist die Verantwortungsübernahme ausgeprägt.

Faktorenanalytisch konnten die beiden Faktoren „Rechtfertigung“ und „Entschuldigung“ extrahiert werden, welche durch 13 der ursprünglich 20 Items des VÜ bestehen. Entschuldigung spiegelt dabei eine Strategie wider, mit der eine Person eine Norm anerkennt, sich aber von der Verantwortlichkeit für die Tat zu entbinden versucht. Rechtfertigung stellt eine Strategie dar, mit der eine Person die Verantwortlichkeit für die Tat übernimmt, diese jedoch nicht als Normverletzung wahrnimmt (Oswald & Bütikofer, 2002). Die zwei Faktoren erklären gemeinsam 42% der Varianz des Fragebogens (Gabriel et al., 2005). Sechs Items bilden demnach die Subskala „Rechtfertigung“ Bsp.: „Ich habe es eigentlich gut gemeint, aber dann ist alles schiefgegangen“), sieben Items die Subskala Entschuldigung (Bsp.: „Nach meiner Tat konnte ich mir selbst nicht mehr erklären, wie ich das habe machen können“). Die interne Konsistenz der beiden Subskalen liegt gemäß Gabriel et al. (2005) bei α=.72 (Rechtfertigung) bzw. α=.75 (Entschuldigung).

Für die Zwecke des Modellversuchs kam zudem eine Fremdbeurteilung des durch den Behandler wahrgenommenen Grades an Verantwortungsübernahme zur Anwendung (Quelle: Aktenbooklet): Diese besteht aus einem einzigen Item („Mein Patient übernimmt die Verantwortung für sein Hauptdelikt“). Die Antwort erfolgt auf einer gleich gearteten vierstufigen Likert-Skala wie bei der Selbstbeurteilung, mit dem Unterschied, dass in der Fremdbeurteilung ein höherer Grad an Zustimmung ein höheres Ausmaß an wahrgenommener Verantwortungsübernahme bedeutet.

Anhang 4. Ersetzen von Daten: Einzelfallspezifische Entscheidungen

Ersetzen von Daten aufgrund Problemen beim Reshapen ins Wide-Format

Beim Umformen des Datensatzes „Datensatz Modellversuch 161219.dta“ vom Long-Format ins Wide-Format kam es in einigen Fällen zu Fehlermeldungen bei Variablen, die nicht reshaped werden sollten (d.h. Variablen, die in allen zum jeweiligen Code gehörenden Zeilen die identische Information haben müssen). In diesen Fällen wurden die Rohdaten wie in Tabelle 48 angegeben verändert.

Tabelle 48. Umgang mit Problemen beim Reshaping

Variable	Code	Problem	Lösung
inschweiz, amleben	1048 1094	Einträge sind nur beim Indexdelikt vorgenommen worden, nicht bei der Vorstrafe.	Alle Zeilen auffüllen mit der vorhandenen Information, da diese auf den aktuellen Zeitpunkt (Datum SRA) bezogen ist.
datum_codebook	1034	In einer der 7 Zeilen steht ein abweichendes Datum.	Angleichen mit der Information aus den anderen 6 Zeilen.
quelle_urteil	1034	Angaben stehen nur in einer der zum Code gehörenden Zeilen.	Alle Zeilen auffüllen mit der vorhandenen Information.
el_tod	1034	Abweichende Angaben in einer Zeile („kein Elternteil gestorben“)	Angleichen mit der Information aus den anderen 6 Zeilen („Ein oder beide Elternteile gestorben“).
fremdunter	1013	Angaben nur beim als Rückfall codierten Delikt	Angleichen auf alle Zeilen, da es sich um eine Mehrinformation handelt.
gewalt	1036	In einer der 19 zum Code gehörenden Zeile „Ja“, in den übrigen Zeilen „Nein“.	Laut SRA handelt es sich um vollendeten Raub und Erpressung, d.h.um Gewaltdelikt → Alle Zeilen anpassen auf „Ja“.
gewaltvv	1036	In einer der 19 zum Code gehörenden Zeile „Ja“, in den übrigen Zeilen „Nein“.	Kein Versuch laut SRA → Alle Zeilen anpassen auf „Nein“.
gewaltkat	1036	Angaben stehen nur in einer der zum Code gehörenden Zeilen.	Anpassen in den anderen Zeilen, da Mehrinformation.
kons	1013	Angaben nur in einer der zum Code gehörenden Zeilen.	Anpassen in den anderen Zeilen, da Mehrinformation.

Ersetzen von Daten aus Gründen der Plausibilität

Für die Information für das Datum des Strafregisterauszugs wurde in Absprache mit dem FPD Bern primär die Variable „datum_sra“ („Strafregisterauszug Datum) verwendet. In folgenden Fällen, die keine Angabe in dieser Variable, jedoch in der Vari-

able „stichtag“ („Stichtag (Datum des Strafregisterauszugs)“) hatten, wurde die Information in „datum_sra“ durch die in „stichtag“ enthaltene Information ergänzt:
– 1030 (GST R&R)
– 1346 (SST ASAT)

In einigen Fällen war ein offensichtlich fehlerhaftes Jahrhundert bei den Variablen gespeichert, die gemäß Strafregisterauszug Informationen über das Deliktdatum enthalten („deliktdatum_von“; „deliktdatum_bis“). Dies kann zu bedeutsamen Verzerrungen bei der Klassifikation eines Delikts als Vorstrafe bzw. Indexdelikt oder Rückfall führen. Die Daten in diesen Variablen, welche zur neu kreierten Variable „datum_del_n“ („Deliktdatum kombiniert“) zusammengefasst wurden, wurden in den nachfolgend aufgeführten Fällen verändert. Zuvor wurde die Plausibilität der vermuteten korrekten Daten durch einen Abgleich mit dem entsprechenden Urteilsdatum sowie zeitlich angrenzenden Deliktdaten überprüft.
– „deliktdatum_von“:
  – Code 1093 (GST KG): 06.01.1900 wurde ersetzt durch 06.01.2000 (Falsches Jahrhundert)
  – Code 1087 (GST TAU): 11.01.1900 wurde ersetzt durch 11.01.2000 (Falsches Jahrhundert)
  – Code 1419 (GST R&R): 04.07.1900 wurde ersetzt durch 04.07.2000 (Falsches Jahrhundert)
  – Code 1331 (GST R&R): 21.02.1907 wurde ersetzt durch 21.02.2007 (Falsches Jahrhundert)
  – Code 1210 (GST R&R): 11.09.2998 wurde ersetzt durch 11.09.2008 (Falsches Jahrhundert bzw. Zahlenvertauschung)
  – Code 1419 (GST R&R): 30.10.2998 wurde ersetzt durch 30.10.2008 (Falsches Jahrhundert bzw. Zahlenvertauschung)
– „deliktdatum_bis“:
  – Code 1208 (GST KG): 16.01.1900 wurde ersetzt durch 16.07.2008 (Offensichtlicher Tippfehler, da in mehreren Delikten mit demselben Datum, die im selben Urteil abgeurteilt wurden, abweichende Informationen enthalten sind und da in dem fehlerhaften Eintrag „deliktdatum_von“ nach „deliktdatum_bis“ liegt, was sich logisch ausschließt.)

In einigen Fällen war ein offensichtlich fehlerhaftes Jahrhundert bei der Angabe zum Urteilsdatum (Variable „urteilsdatum“) gespeichert. In den folgenden Fällen wurde nach Überprüfung der Plausibilität (Abgleich vom jeweiligen Delikt- und Urteilsdatum und Vergleich mit zeitlich angrenzenden Verurteilungs-Daten) die Daten wie folgt verändert:
– Code 1076 (GST R&R): 26.09.2205 wurde ersetzt durch 26.09.2005
– Code 1440 (GST R&R): 21.01.1900 wurde ersetzt durch 21.01.2000
– Codes 1207 (GST KG) und 1222 (SST ASAT): 24.01.1900 wurde ersetzt durch 24.01.2000

In Fall 1425 (GST TAU) hat die Variable „gesetz“, die für die Bildung einer Variablen zur Art des Gesetzes benötigt wird, gegen das der Proband verstoßen hat, in einem Fall die Angabe „in Bearbeitung, Abklärung nötig“. Diese Angabe wird umcodiert zu

„Ausländisches Gesetz“, da dies eindeutig aus anderen Variablen (z.B. „deliktart“) hervorgeht.

Referenzen

Acuna, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In: D. Banks, L. House, F. R. McMorris, P. Arabie, & W. Gaul (Hrsg.), *Classification, clustering, and data mining applications* (S. 639-648). Heidelberg: Springer.

Andrews, D. A., & Dowden, C. (2005). Managing correctional treatment for reduced recidivism: A meta-analytic review of programme integrity. *Legal and Criminological Psychology, 10*(2), 173-187.

Andrews, D. A., Zinger, I., Hoge, R. D., Bonta, J., Gendreau, P., & Cullen, F. T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology, 28*(3), 369-404.

Antonowicz, D., & Parker, J. (2012). Reducing Recidivism Evidence from 26 Years of International Evaluations of Reasoning & Rehabilitation Programs. Abgerufen von https://www.researchgate.net/publication/283345614_26_years_of_international_evaluations_of_Reasoning_Rehabilitation_programs

Aos, S., Miller, M., & Drake, E. (2006). Evidence-based public policy options to reduce future prison construction, criminal justice costs, and crime rates. *Federal Sentencing Reporter, 19*(4), 275-290. Abgerufen von http://www.wsipp.wa.gov/ReportFile/952/Wsipp_Evidence-Based-Public-Policy-Options-to-Reduce-Future-Prison-Construction-Criminal-Justice-Costs-and-Crime-Rates_Full-Report.pdf

Aos, S., Phipps, P., Barnoski, R., & Lieb, R. (2001). *The Comparative Costs and Benefits of Programs To Reduce Crime. Version 4.0*. Olympia, WA: Washington State Institute for Public Policy. Abgerufen von https://eric.ed.gov/?id=ED453340

Babcock, J. C., Green, C. E., & Robie, C. (2004). Does batterers' treatment work? A meta-analytic review of domestic violence treatment. *Clinical psychology review, 23*(8), 1023-1053.

Beech, A., Bourgon, G., Hanson, R. K., Harris, A. J., Langton, C., Marques, J., . . . Seto, M. (2007). *The Collaborative Outcome Data Committee's Guidelines for the Evaluation of Sexual Offender Treatment Outcome Research. Part 2: CODC guidelines*der Reihe),Aufl. Abgerufen von https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/cllbrtv-tcmdt-gdlns/cllbrtv-tcmdt-gdlns-eng.pdf

Berrios, G. E. (1996). *The history of mental symptoms: descriptive psychopathology since the nineteenth century*: Cambridge University Press.

Bonta, J., & Andrews, D. A. (2016). *The psychology of criminal conduct* (6. Aufl.). London/ New York: Taylor & Francis.

Borduin, C. M., Henggeler, S., Blaske, D., & Stein, R. (1990). Multisystematic treatment of adolescent sexual offenders. *International Journal of Offender Therapy and Comparative Criminology, 34*(2), 105-113.

Borduin, C. M., Schaeffer, C. M., & Heiblum, N. (2009). A randomized clinical trial of multisystemic therapy with juvenile sexual offenders: effects on youth social ecology and criminal activity. *Journal of consulting and clinical psychology, 77*(1), 26-37.

Bundesamt für Statistik. (2016). Rückfallrate nach Entlassung aus dem Strafvollzug nach Geschlecht, Alter, Vorverurteilungen. https://www.bfs.admin.ch/bfs/de/home/statistiken/kriminalitaet-strafrecht/rueckfall.html, zuletzt abgerufen am 14.06.2017.

Bushman, B. J. (1995). Moderating role of trait aggressiveness in the effects of violent media on aggression. *Journal of personality and social psychology, 69*(5), 950-960.

Cleckley, H. (1976). *The mask of sanity* (5. Aufl.). St. Louis, MO: Mosby.

Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological bulletin, 115*(1), 74-101.

Cristea, I. A., Gentili, C., Cotet, C. D., Palomba, D., Barbui, C., and Cuijpers, P. (2017). Efficacy of psychotherapies for borderline personality disorder: a systematic review and meta-analysis. JAMA Psychiatry 74, 319–328. doi: 10.1001/jamapsychiatry.2016.4287.

Cullen, A. E., Soria, C., Clarke, A. Y., Dean, K., & Fahy, T. (2011). Factors predicting dropout from the reasoning and rehabilitation program with mentally disordered offenders. *Criminal Justice and Behavior, 38*(3), 217-230.

Cullen, F. T., & Gendreau, P. (2000). Assessing correctional rehabilitation: Policy, practice, and prospects. *Criminal justice, 3*, 109-175.

Dill, K. E., Anderson, C. A., Anderson, K. B., & Deuser, W. E. (1997). Effects of aggressive personality on social expectations and social perceptions. *Journal of Research in Personality, 31*(2), 272-292.

Dodge, K. A. (1980). Social cognition and children's aggressive behavior. *Child development, 51*(1), 162-170.

Endrass, J., Rossegger, A., & Braunschweig, M. (2012). Wirksamkeit von Behandlungsprogrammen. In: J. Endrass, A. Rossegger, F. Urbaniok, & B. Borchard (Hrsg.), *Interventionen bei Gewalt- und Sexualstraftätern: Risk-Management, Methoden und Konzepte der forensischen Therapie* (S. 43-69). Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft.

Endrass, J., Rossegger, A., & Kuhn, B. (2012). Kosten-Nutzen-Effizienz von Therapien. In: J. Endrass, A. Rossegger, F. Urbaniok, & B. Borchard (Hrsg.), *Interventionen bei Gewalt- und Sexualstraftätern: Risk-Management, Methoden und Konzepte der forensischen Therapie* (S. 77-88). Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft.

Ermer, A. (2008). Forensisch-psychiatrische Therapie: Störungs-und deliktorientierte Behandlung. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie, 56*(2), 79-87.

Eucker, S. (2013). Vorstellung des Behandlungskonzeptes Reasoning and Rehabilitation (R&R). *Vortrag für die 39. Arbeits- und Fortbildungstagung der Bundesvereinigung der Anstaltsleiter und Anstaltsleiterinnen im Justizvollzug e.V. vom 06. bis 10.Mai 2013 im Bildungszentrum Kirkel.* http://www.bvaj.de/ReasoningandRehabilitation.pdf, zuletzt abgerufen am 14.06.2017.

Falk, O., & Steffes-enn, R. (2014). *Das Anti-Sexuelle-Aggressivität-Training®Suisse. ASAT®Suisse-Arbeitshandbuch* (2. Aufl.): Universität Bern.

Farrelly, F., & Matthews, S. (1994). Provokative Therapie. In: R. J. Corsini (Hrsg.), *Handbuch der Psychotherapie* (S. 956-977). Weinheim: Beltz.

Farrington, D. P., & Welsh, B. C. (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology, 1*(1), 9-38.

Farrington, D. P. & Jolliffe, D. (2002). A feasibility study into using a randomised controlled trial to evaluate treatment pilots at HMP Whitemoor. Home Office Online Report 14/02. London: Home Office.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods, 41*(4), 1149-1160.

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*(2), 175-191.

Finch, W. H. (2016). Missing data and multiple imputation in the context of multivariate analysis of variance. *Journal of Experimental Education, 84*(2), 356-372.

Fournier, J. C., DeRubeis, R. J., Amsterdam, J., Shelton, R. C., and Hollon, S. D. (2015). Gains in employment status following antidepressant medication or cognitive therapy for depression. Br. J. Psychiatry 206, 332–338. doi: 10.1192/bjp.bp.113.133694

FPD Bern (2009). *Gesuch FPD Bern: Neue psychotherapeutische Interventionsprogramme und Evaluationskonzepte im Schweizer Strafvollzug (Beilagen Nr. 2 und 3)*. Bern: Forensisch-Psychiatrischer Dienst der Universität Bern

Friendship, C., Street, R., Cann, J., & Harper, G. (2005). Introduction: The Policy Context and Assessing the Evidence. In: G. Harper & C. Chitty (Hrsg.), *The Impact of Corrections on Re-offending: A Review of 'What Works'* (2. Aufl., S. 1-16). Los Angeles: Home Office Research, Development and Statistics Directorate.

Gabriel, U., Oswald, M. E., & Bütikofer, A. (2005). Freiwillige Teilnahme an Wiedergutmachungsprogrammen und die Bereitschaft zur Verantwortungsübernahme. *Zeitschrift für Sozialpsychologie, 36*(4), 239-249.

Geen, R. G. (2001). *Human aggression* (2. Aufl.). Philadelphia, PA: Open University Press.

Giancola, P. R. (2002). Alcohol-related aggression in men and women: the influence of dispositional aggressivity. *Journal of studies on alcohol, 63*(6), 696-708.

Gretenkord, L. (2015). R&R – Das „Reasoning and Rehabilitation Programm“. In: S. Eucker & R. Müller-Isberner (Hrsg.), *Praxishandbuch Maßregelvollzug: Grundlagen, Konzepte und Praxis der Kriminaltherapie* (2. Aufl., S. 197-203). Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft.

Hall, G. C. N. (1995). Sexual offender recidivism revisited: a meta-analysis of recent treatment studies. *Journal of consulting and clinical psychology, 63*(5), 802-809.

Hampel, R., & Selg, H. (1998). *Fragebogen zur Erfassung von Aggressivitätsfaktoren: FAF*. Göttingen: Hogrefe.

Hanson, R. K. (2005). *The validity of Static-99 with older sexual offenders*der Reihe),Aufl. Abgerufen von https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/vldty-sttc-99/index-en.aspx

Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). *A meta-analysis of the effectiveness of treatment for sexual offenders: Risk, need, and responsivity*der Reihe),Aufl. *User Report* Abgerufen von http://publications.gc.ca/collections/collection_2010/sp-ps/PS3-1-2009-2-eng.pdf

Hanson, R. K., Gordon, A., Harris, A. J., Marques, J. K., Murphy, W., Quinsey, V. L., & Seto, M. C. (2002). First report of the collaborative outcome data project on

the effectiveness of psychological treatment for sex offenders. *Sexual Abuse: A Journal of Research and Treatment, 14*(2), 169-194.
Hanson, R. K., & Morton-Bourgon, K. E. (2005). The characteristics of persistent sexual offenders: a meta-analysis of recidivism studies. *Journal of consulting and clinical psychology, 73*(6), 1154-1163.
Hanson, R. K., & Thornton, D. (1999). *Static 99: Improving actuarial risk assessments for sex offenders*. Ottawa, Ontario, Canada: Department of Justice Canada.
Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: a comparison of three actuarial scales. *Law and human behavior, 24*(1), 119.
Hare, R. D. (2003). *The Hare psychopathy checklist-revised* (2. Aufl.). Toronto, ON, Canada: Multi-Health Systems.
Hare, R. D., Black, P., & Walsh, Z. (2013). The Hare Psychopathy Checklist–Revised: Forensic Applications and Limitations. In: R. Archer & E. Wheeler (Hrsg.), *Forensic uses of clinical assessment instruments* (2. Aufl., S. 266-290). New York: Routledge.
Hare, R. D., & Cox, D. N. (1978). Clinical and empirical conceptions of psychopathy, and the selection of subjects for research. In: R. D. Hare & D. Schalling (Hrsg.), *Psychopathic behavior: Approaches to research* (S. 1-21). Chichester, UK: Wiley.
Helmus, L., & Hanson, R. K. (2007). Predictive validity of the Static-99 and Static-2002 for sex offenders on community supervision. *Sexual Offender Treatment, 2*(2), 1-14.
Hengartner MP (2018) Raising Awareness for the Replication Crisis in Clinical Psychology by Focusing on Inconsistencies in Psychotherapy Research: How Much Can We Rely on Published Findings from Efficacy Trials? Front. Psychol. 9:256. doi: 10.3389/fpsyg.2018.00256.
Henning, K., & Holdford, R. (2006). Minimization, Denial, and Victim Blaming by Batterers How Much Does the Truth Matter? *Criminal Justice and Behavior, 33*(1), 110-130.
Heubrock, D., & Petermann, F. (2008). *Kurzfragebogen zur Erfassung von Aggressivitätsfaktoren: K-FAF*. Göttingen: Hogrefe.
Hollin, C. R. (2008). Evaluating offending behaviour programmes: Does only randomization glister? *Criminology and Criminal Justice, 8*(1), 89-106.
Holtzworth-Munroe, A., Bates, L., Smutzler, N., & Sandin, E. (1997). A brief review of the research on husband violence part I: Maritally violent versus nonviolent men. *Aggression and violent behavior, 2*(1), 65-99.
Horowitz, L. M., Strauß, B., & Kordy, H. (2000). *Inventar zur Erfassung interpersonaler Probleme: IIP-D; deutsche Version*. Göttingen: Beltz.
Horowitz, L. M., Strauß, B., Thomas, A., & Kordy, H. (2016). *Inventar zur Erfassung interpersonaler Probleme: IIP-D; deutsche Version* (3. Aufl.). Göttingen: Beltz.
Institut für forensische Psychiatrie Haina e.V. (2017). Reasoning & Rehabilitation Programm. http://www.forensic-haina.de/r--r/index.html, zuletzt abgerufen am 16.06.2017.
Ioannidis, J. P., Haidich, A.-B., Pappa, M., Pantazis, N., Kokori, S. I., Tektonidou, M. G., . . . Lau, J. (2001). Comparison of evidence of treatment effects in randomized and nonrandomized studies. *Jama, 286*(7), 821-830.
Koehler, J. A., Lösel, F., Akoensi, T. D., & Humphreys, D. K. (2013). A systematic review and meta-analysis on the effects of young offender treatment programs in Europe. *Journal of Experimental Criminology, 9*(1), 19-43.
Kraus, C., & Berner, W. (2000). Die Klassifikation von Sexualstraftätern nach Knight und Prentky. *Monatsschrift für Kriminologie und Strafrechtsreform, 83*(6), 395-406.

Lipsey, M. W., & Cullen, F. T. (2007). The effectiveness of correctional rehabilitation: A review of systematic reviews. *Annual Review of Law and Social Sciences, 3*(1), 297-320.

Lipsey, M. W., Chapman, G. L. & Landenberger, N. A. (2001). Cognitive-behavioral programs for offenders. Annals of the American Academy of Political and Social Science, 578, 144-157.

Lösel, F. (1995). The efficacy of correctional treatment: A review and synthesis of meta-evaluations. In: J. McGuire (Hrsg.), *What works: Reducing reoffending: Guidelines from research and practice* (S. 79-114). Chichester, UK: Wiley.

Lösel, F., & Schmucker, M. (2005). The effectiveness of treatment for sexual offenders: A comprehensive meta-analysis. *Journal of Experimental Criminology, 1*(1), 117-146.

Marques, J. K., Wiederanders, M., Day, D. M., Nelson, C., & Van Ommeren, A. (2005). Effects of a relapse prevention program on sexual recidivism: Final results from California's Sex Offender Treatment and Evaluation Project (SOTEP). *Sexual Abuse, 17*(1), 79-107.

Marshall, W. L., Anderson, D., & Fernandez, Y. (Hrsg.). (1999). *Cognitive behavioural treatment of sexual offenders.* (Aufl.) Chichester, UK: Wiley.

Marshall, W. L., & Marshall, L. E. (2007). The utility of the random controlled trial for evaluating sexual offender treatment: The gold standard or an inappropriate strategy? *Sexual Abuse: A Journal of Research and Treatment, 19*(2), 175-191.

Maruna, S., & Mann, R. E. (2006). A fundamental attribution error? Rethinking cognitive distortions. *Legal and Criminological Psychology, 11*(2), 155-177.

Matthews, B. A., & Norris, F. H. (2002). When Is Believing “Seeing”? Hostile Attribution Bias as a Function of Self-Reported Aggression. *Journal of Applied Social Psychology, 32*(1), 1-31.

McGrath, R. J., Cumming, G., Livingston, J. A., & Hoke, S. E. (2003). Outcome of a treatment program for adult sex offenders from prison to community. *Journal of Interpersonal Violence, 18*(1), 3-17.

Mokros, A., Neumann, C. S., Stadtland, C., Osterheider, M., Nedopil, N., & Hare, R. D. (2011). Assessing measurement invariance of PCL-R assessments from file reviews of North American and German offenders. *International journal of law and psychiatry, 34*(1), 56-63.

Ortmann, R. (1987). *Resozialisierung im Strafprozess*. Freiburg i. Br.: Max-Planck-Institut für ausländisches und internationales Strafrecht.

Oswald, M. E., & Bütikofer, A. (2002). *Wissenschaftliche Evaluation des Modellversuches "Tataufarbeitung und Wiedergutmachung (TaWi) – Berner Modell"*. Psychologisches Institut der Universität Bern

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2. Aufl.). Washington DC: American Psychological Association.

Rehder, U. (2001). Sexualstraftäter: Klassifizierung und Prognose. In: G. Rehn, B. Wischka, F. Lösel, & M. Walter (Hrsg.), *Behandlung „gefährlicher Straftäter“. Grundlagen, Konzepte, Ergebnisse* (2. Aufl., S. 81-103). Herbolzheim: Centaurus.

Ross, R. R., & Fabiano, E. A. (1985). *Time to think: A cognitive model of delinquency prevention and offender rehabilitation*. Tennessee: Institute of Social Sciences and Arts.

Ross, R. R., Fabiano, E. A., & Ewles, C. (1988). Reasoning and Rehabilitation. *International Journal of Offender Therapy and Comparative Criminology, 32*, 29-35.

Ross, R. R., Fabiano, E. A., & Ross, R. D. (1986). *Reasoning and rehabilitation: A handbook for teaching cognitive skills*. Ottawa: Cognitive Center of Canada, University of Ottawa.

Ross, R. R., Hilborn, J., & Liddle, P. (2007). *Reasoning & Rehabilitation 2: short version for adults*. Ottawa: Cognitive Centre of Canada.

Ross, R. R., & Ross, R. (1995). The R&R programme. In: R. R. Ross & R. Ross (Hrsg.), *Thinking straight: The reasoning and rehabilitation program for delinquency prevention and offender rehabilitation* (S. 83-120). Ottawa, ON: Air Training & Publications.

Rossegger, A., Urbaniok, F., Danielsson, C., & Endrass, J. (2009). Der Violence Risk Appraisal Guide (VRAG)–Ein Instrument zur Kriminalprognose bei Gewaltstraftätern. *Fortschritte der Neurologie- Psychiatrie, 77*(10), 577-584.

Rutrecht, M., Jagsch, R., & Kryspin-Exner, I. (2002). *Bindungsstile bei Sexualstraftätern. Zusammenhang mit Aggression und Ängstlichkeit*. Frankfurt: Verlag für Polizeiwissenschaft.

Sachse, R. (2006). *Persönlichkeitsstörungen verstehen: Zum Umgang mit schwierigen Klienten*. Bonn: Psychiatrie-Verlag.

Schahn, J., Dinger, J., & Bohner, G. (1995). Rationalisierungen und Neutralisationen als Rechtfertigungsstrategien: Ein Vergleich zwischen Umwelt-und Delinquenzbereich. *Zeitschrift für Differentielle und Diagnostische Psychologie, 16*(3), 177-194.

Schmucker, M., & Lösel, F. (2015). The effects of sexual offender treatment on recidivism: An international meta-analysis of sound quality evaluations. *Journal of Experimental Criminology, 11*(4), 597-630.

Sherman, L. W., Gottfredson, D. C., MacKenzie, D. L., Eck, J., Reuter, P., & Bushway, S. D. (1998). *Preventing Crime: What Works, What Doesn't, What's Promising*. Washinton DC: Department of Justice, National Institute of Justice.

StataCorp. (2015). Stata Statistical Software: Release 14. College Station, TX: StataCorp.

Steffes-enn, R. (2005). Das „Anti-Sexuelle-Aggressivitäts-Training“ (ASAT): Stark genug, um schwach zu sein. In: B. Wischka, U. Rehder, F. Specht, E. Foppe, & R. Willems (Hrsg.), *Sozialtherapie im Justizvollzug - Aktuelle Konzepte, Erfahrungen und Kooperationsmodelle* (S. 245-259). Lingen: Kriminalpädagogischer Verlag.

Steffes-enn, R. (2008). Das Anti-Sexuelle-Aggressivität-Training (ASAT®). *Klinische Sozialarbeit, 4*(2), 9-12.

Thomas, B., Ciliska, D., Dobbins, M., & Micucci, S. (2004). A process for systematically reviewing the literature: providing the research evidence for public health nursing interventions. *Worldviews on Evidence-Based Nursing, 1*(3), 176-184.

Tong, L. J., & Farrington, D. P. (2006). How effective is the “Reasoning and Rehabilitation” programme in reducing reoffending? A meta-analysis of evaluations in four countries. *Psychology, Crime & Law, 12*(1), 3-24.

Tong, L. J., & Farrington, D. P. (2008). Effectiveness of "Reasoning and rehabilitation" in reducing reoffending. *Psicothema, 20*(1), 20-28.

Tremblay, P. F., & Belchevski, M. (2004). Did the instigator intend to provoke? A key moderator in the relation between trait aggression and aggressive behavior. *Aggressive Behavior, 30*(5), 409-424.

Weidner, J. (1995). *Anti-Aggressivitäts-Training für Gewalttäter (3. erweit. Aufl.)*. Bonn: Forum Verlag.

Wilson, D. B., Bouffard, L. A. & Mackenzie, D. L. (2005). A quantitative review of structured, group-orientated, cognitive-behavioral programs for offenders.

Criminal Justice and Behavior 32, 172-204.Wößner, G., & Schwedler, A. (2014). Correctional Treatment of Sexual and Violent Offenders Therapeutic Change, Prison Climate, and Recidivism. *Criminal Justice and Behavior, 41*(7), 862-879.