Marcel Guéridon

Ist eine Evaluation der Wirksamkeit von Sozialtherapie überhaupt möglich?

Eine überwiegend frustrierende Diskussion

Zusammenfassung

Spätestens seit der Entscheidung des Bundesverfassungsgerichts 2006 ist die Evaluation von Behandlungsmaßnahmen in den Strafvollzugsgesetzen verankert. In der Praxis ist die angewandte Forschung aber mit einer Reihe methodischer Probleme konfrontiert, die die grundsätzliche Möglichkeit verlässlicher und nützlicher Ergebnisse in Frage stellen. Dieser Beitrag diskutiert am Beispiel der Evaluation der Sozialtherapie einige dieser Probleme und die Frage, ob eine sinnvolle Evaluation möglich ist und was dafür getan werden kann.
Schlüsselwörter: Evaluationsforschung, Sozialtherapie, Validität, Nützlichkeit, Wirksamkeit

Abstract:

At least since the decision of the German Federal Constitutional Court in 2006 evaluation research on offender treatment is mandatory for the penitentiary administration. In practice applied research is faced with numerous methodological problems. Taken together this raises the question if sound and useful results can be obtained at all. This article discusses some of these problems with reference to the example of the evaluation of German Social Therapy and tries to answer the question whether a meaningful evaluation is possible and what could be done in favor.
Key Words: Evaluation research, Social Therapy, Validity, Usefulness, Effectiveness

1. Einleitung

Die Forderung nach „evidenzbasierter" (Kriminal-)Politik (Cartwright & Hardie, 2012), die sich auf Fakten und Forschungsergebnisse stützen soll, hat spätestens seit dem Urteil des Bundesverfassungsgerichtes (BVerfG, Urteil des Zweiten Senats vom 31. Mai 2006 – 2 BvR 1673/04 – Rn. [1-77]) ihren festen Platz auch in den (Jugend-)Strafvollzugs- und Sicherungsverwahrungsgesetzen der Länder gefunden. Auf

DOI: 10.5771/2365-1083-2016-3-285

Verhaltensänderung abzielende Einzelmaßnahmen und Maßnahmenkomplexe wie die Sozialtherapie und auch der Strafvollzug im Ganzen müssen entsprechend der gesetzlichen Regelungen wissenschaftlich begleitet und überprüft werden. Die Evaluationsforschung hat eine entsprechend hohe Verantwortung zu tragen (Cronbach, 1982; Weisburd & Hinkle, 2012). Bei teuren Maßnahmen wie der sozialtherapeutischen Behandlung geht es einerseits um die Rechtfertigung der Verwendung öffentlicher Mittel, besonders ausschlaggebend scheint jedoch, dass das Festhalten an nachweislich unwirksamen Praktiken aus spezialpräventiver Sicht unsinnig ist. Auch gutgemeinte Maßnahmen können zudem Schaden anrichten, was sowohl in der Psychotherapie lange Zeit keine Aufmerksamkeit bekommen hat (vgl. Lilienfeld, 2007) als auch im Bereich der sozialen Interventionen auf einen „code of silence" stößt (McCord, 2003, S. 17). Immerhin finden solche Diskussionen auf Inhaftierungen insgesamt bezogen teilweise statt (vgl. Durlauf & Nagin, 2011; Häßler, 2012).

Die Diskussion, ob Behandlung von Straftätern und von Sexualstraftätern im Besonderen wirksam ist, wird spätestens seit Martinsons (1974) „Nothing-Works"-These geführt. Während generelle Wirksamkeit eher bejaht wird (Andrews & Bonta, 2010; Hanson, Bourgon, Helmus & Hodgson, 2009; Marshall & Marshall, 2010; Schmucker & Lösel, 2015), ist der Tenor der Skeptiker besonders in Bezug auf Sexualstraftaten, dass bisher überzeugende Argumente für die Wirksamkeit fehlen (Dennis et al., 2012; Farabee, 2005; Hoberman, 2015; Rice & Harris, 2013).

Die bisher trotz umfangreicher Forschung unklare Datenlage führt besonders aus der meta-analytischen Forschung zur Forderung, endlich randomisierte Experimente (randomised controlled trials, RCTs) durchzuführen und damit Forschung gemäß „Goldstandard" zu betreiben (Lösel, 2008; Rice & Harris, 2013; Weisburd, 2010; Weisburd & Hinkle, 2012). Im Bereich der deutschen Strafvollzugsforschung ist die angewandte Forschung dem bisher nur in Ausnahmen nachgekommen (z.B. Ansorge, 2011; Ortmann, 2002). Die unbedingte Überlegenheit und die Versprechungen von RCTs wurden dagegen wiederholt von theoretischer Seite in Frage gestellt (Berk, 2005; Cartwright, 2007; Hammersley, 2014; Marshall & Marshall, 2007; Sampson, 2010).

Die angewandte Forschung steht aber nicht nur bei der Wahl des Forschungsdesigns vor bedeutenden Problemen. Darüber hinaus handelt um ein politisch besonders heikles Feld, in dem viele gesetzliche Regelungen bestehen (die etwa Randomisierung vielfach ausschließen), auch ist Strafvollzug ein komplexes Gebilde aus verschiedenen Professionen, Traditionen und Interessenlagen. An verschiedenen Stellen wurde aus der angewandten Forschung bereits darauf hingewiesen, welche vielfältigen Probleme bei der Evaluation der Sozialtherapie (Suhling, 2012; Wößner, 2014), des Jugendstrafvollzugs (Obergfell-Fuchs & Wulf, 2011) sowie bei kriminologischen Evaluationen insgesamt auftreten können (Budde & Suhling, 2016¸ Lösel, 2008).

Der vorliegende Beitrag führt diese Frage exemplarisch dahin weiter, ob eine Evaluation der Wirksamkeit der Sozialtherapie überhaupt möglich ist. Die Möglichkeiten sind, so die vorläufige These, nicht nur von Fähigkeiten und Kreativität der Evaluatoren abhängig. Auch tatsächliche Gegebenheiten bestimmen maßgeblich mit, welche Fragen überhaupt verlässlich beantwortet werden können und welche nicht. Mit

Hammersley (2014, S. 9) könnte man auch bezüglich der Evaluation der Sozialtherapie behaupten: „Unfortunateley, we must face the fact that research cannot always provide answers to questions that are seen as pressing by policymakers and practitioners. Nor can we guarantee the validity of our findings." Diese These geht über die Feststellung hinaus, dass keine „perfekte" Evaluationsstudie möglich ist (Spöhr, 2009, S. 169). Möglicherweise ist die Sozialtherapie, bei der sich die Probleme der Evaluationsforschung im Strafvollzug in aller Deutlichkeit zeigen, ein Beispiel, für das die Evaluationsforschung schlicht keine sinnvollen Antworten geben kann.

Um sich der These zu nähern, werden im nächsten Abschnitt mögliche Kriterien diskutiert, bevor im Detail auf die Probleme eingegangen wird, die sich bei der Evaluation der Sozialtherapie zeigen. Dabei wird im Abschnitt 3 die Verlässlichkeit und in Abschnitt 4 die Nützlichkeit der möglichen Forschungsergebnisse besprochen. Im Abschnitt 5 werden einige Lösungsmöglichkeiten thematisiert, bevor ein Fazit gezogen wird.

2. *Welche Kriterien muss eine Evaluation der Sozialtherapie erfüllen?*

Bisherige Diskussionen um Schwierigkeiten der Evaluation im Strafvollzug sind oft eher unstrukturiert. Auf die Sozialtherapie bezogen diskutieren verschiedene Autoren (z.B. Wößner, 2014, Guéridon & Suhling 2015, Niemz, 2015) einzelne Probleme der Evaluation, ohne ein bestimmtes System von Kriterien zu verwenden. Am strukturiertesten diskutierten bisher Spöhr (2009) und Suhling (2012) die Herausforderungen bei einer Evaluation der Sozialtherapie.

Qualitätsstandards wie die Standards der Deutschen Gesellschaft für Evaluation e.V. (DeGEval, 2008) oder die Orientierung an einem idealtypischen Verlauf eines Evaluationsprojektes, wie etwa im normativen Modell von Balzer (2005) dargestellt, betreffen eher wünschenswertes und korrektes Vorgehen der Evaluatoren. Für eine Debatte über die Möglichkeiten einer Evaluation sind sie daher nur begrenzt hilfreich.

Lösel (2008) und Farrington (2003) verwenden als Ausgangspunkt die Gefährdungen der vier Arten der Validität nach Shadish, Cook und Campbell (2002). Statistische Validität beschreibt die Verlässlichkeit des Zusammenhangs zwischen Maßnahme (treatment) und Ergebnis (outcome) sowie die Eignung der verwendeten statistischen Verfahren, während interne Validität angibt, ob ein Zusammenhang kausal interpretiert werden kann. Konstruktvalidität betrifft die Möglichkeit theoretischer Verallgemeinerung und externale Validität die Möglichkeit einer Verallgemeinerung über Personen und Kontexte. Die Systematik kann um die deskriptive Validität erweitert werden (Farrington, 2003; Lösel, 2008), welche die Notwendigkeit der genauen Beschreibung des Programms und seiner Ziele umfasst. Auch wenn alle Arten der Validität für eine verlässliche Evaluation unerlässlich sind, ist doch besonders die interne Validität entscheidend. Alle Evaluationsbemühungen helfen nichts, wenn nicht ausreichend sicher gesagt werden kann, ob wirklich die Maßnahme den Effekt verursacht hat. Das ist und bleibt die zentrale Frage der Evaluation (Lösel, 2008, S. 143).

Minimal verkürzt kann man vielleicht fordern, dass eine Evaluation ausreichend verlässlich und nützlich sein muss. Unter Verlässlichkeit fallen Fragen der Implementation, der Validität und kausalen Inferenz sowie Fragen der Datenerhebung und -auswertung. Unter Verlässlichkeit wird in diesem Beitrag also nicht nur die Reliabilität der Bewertung verstanden, sondern ebenso ihre Objektivität und Gültigkeit (Validität). Nützlichkeit betrifft die Fragen, ob und inwieweit man von der Evaluation lernen kann. Beide Aspekte sind nicht unabhängig: Eine Evaluation, die nicht verlässlich ist, kann nicht nützlich sein, aber eine sehr verlässliche Evaluation ist unter Umständen nicht nützlich und damit letztlich auch ohne Wert, wenn weder die richtigen Fragen beantwortet werden noch Wirkung von der Evaluation ausgehen kann (vgl. Cartwright, 2007).

3. Ist eine verlässliche Evaluation der Sozialtherapie möglich?

3.1 Evaluationsgegenstand, Implementation und Integrität: Ist Sozialtherapie gleich „Sozialtherapie"?

Ob eine Maßnahme gemäß ihrer Konzeption umgesetzt wird (Implementation) und ob die Maßnahme in verschiedenen Einrichtungen über die Zeit gleich ist (Integrität) sind zwei Weichen dafür, welche Fragen eine Evaluation beantwortet kann. Ist eine Maßnahme nicht gemäß Konzept umgesetzt, kann die Wirksamkeit des konkreten Programms nicht evaluiert werden, sondern nur die tatsächliche Praxis. Wird die Maßnahme ständig angepasst und in den Einrichtungen unterschiedlich umgesetzt, können die Effekte nicht einfach verglichen oder aggregiert werden. Aus dieser doppelten Unterscheidung ergeben sich vier Fragen, die je nach konkreter Situation überhaupt nur beantwortet werden können (siehe Tabelle 1).

Tabelle 1: Fragen, die eine Evaluation sinnvoll beantworten kann in Abhängigkeit von Implementation und Integrität

Implementation gemäß Konzept?	**räumliche / zeitliche Integrität?**	
	Nein: einzelne Sozialtherapeutische Abteilung	**Ja: Sozialtherapie insgesamt**
Ja: Sozialtherapie als Programm	(1) Ist die Sozialtherapie in der Anstalt X im Zeitraum Y wirksam gewesen?	(2) Ist Sozialtherapie wirksam?
Nein: Behandlung in sozialtherapeutischen Abteilungen	(3) Ist die Behandlung in der sozialtherapeutischen Abteilung X im Zeitraum Y wirksam gewesen?	(4) Ist Behandlung in Sozialtherapeutischen Abteilungen wirksam?

3.1.1 Implementation: Was ist eigentlich Sozialtherapie?

Ob die Wirksamkeit eines spezifischen Programms evaluiert werden kann, hängt schlicht davon ab, ob es das Programm gibt. Alternativ kann nur die Wahrscheinlichkeit von Wirksamkeit auf Grundlage von Konzepten abgeschätzt werden. Eine solche „State of the Art"-Evaluation (Gollwitzer & Jäger, 2014, S. 22) ist vor der Implementation einer Maßnahme sinnvoll, kann aber auch zur Überprüfung alteingesessene Maßnahmen in Hinblick auf aktuelles Wissen nützlich sein. Tatsächliche Wirksamkeit kann daraus aber nicht abgeleitet werden. Eine Evaluation der Wirksamkeit von Sozialtherapie muss daher bei der Frage beginnen, was Sozialtherapie ist und ob es diese Sozialtherapie eigentlich gibt. Als Grundlage kann dabei das Konzept der integrative Sozialtherapie (vgl. Wischka und Specht, 2001; Wischka, 2014) dienen. Das Konzept benennt als gemeinsame Merkmale:

- Die Berücksichtigung und Einbeziehung des gesamten Lebensumfeldes in und außerhalb der Sozialtherapeutischen Einrichtung bis zur Entlassung
- Die Gestaltung der Handlungsmöglichkeiten und Beziehungsformen innerhalb der Sozialtherapeutischen Einrichtungen im Sinne einer therapeutischen Gemeinschaft
- Die Modifizierung und Verknüpfung psychotherapeutischer, pädagogischer und arbeitstherapeutischer Vorgehensweisen

Neben diesen Leitlinien beinhalten die vom Arbeitskreis Sozialtherapeutischer Anstalten formulierten Mindeststandards strukturelle Rahmenbedingungen, die als notwendige Voraussetzungen sozialtherapeutischer Arbeit eingeschätzt werden (Arbeitskreis Sozialtherapeutischer Anstalten im Justizvollzug e.V., 2012).

Zur Bewertung der Implementation kann auf die jährlichen Erhebungen der Kriminologischen Zentralstelle (zuletzt Elz, 2015) zurückgegriffen werden, die auch die Erfüllung der Mindeststandards untersucht. Das Ergebnis ist allerdings Gegenstand von Diskussionen: Während Egg und Niemz (2012) positive Entwicklungen sehen, betont Rehn (2012, 2014) klare Diskrepanzen zum Konzept der Sozialtherapie, besonders bei der Unabhängigkeit der Einrichtungen. Wenn es sich bei einer sozialtherapeutischen Einrichtung in der Tat um eine „behandlungsorientierte Abteilung des Regelvollzugs" (Rehn, 2014, S. 247) und nicht um die Sozialtherapie im eigentliche Sinne handeln sollte, dann ist die Beantwortung der Frage, ob „Sozialtherapie" wirkt, nicht zu beantworten. Beantwortet werden kann dann nur die Frage, ob die tatsächliche Praxis wirksam ist, wie man diese auch immer nennen möchte (vgl. Wößner, 2014).

Ob eine Evaluation möglich ist, hängt daher von den Interessen der Auftraggeber ab: Wenn die integrative Sozialtherapie bewertet werden soll, muss diese gegebenenfalls erst implementiert werden. Geht es darum, die lediglich als Sozialtherapie benannte „Behandlungsabteilung" zu evaluieren, ist die konkrete Ausgestaltung als notwendiger Aspekt der deskriptiven Validität Ausgangspunkt. Welche Fragen verlässlich beantwortet werden können, hängt dann auch davon ab, wie gut die tatsächliche Praxis erhoben und beschrieben werden kann.

3.1.2 Räumliche und zeitliche Integrität: Wie unterschiedlich ist Sozialtherapie?

Räumliche und zeitliche Integrität sind entscheidende Aspekte, wenn eine definierte Maßnahme insgesamt bewertet werden soll. Ist etwa eine Einschätzung der Sozialtherapie in Deutschland das Ziel, muss Sozialtherapie in Bayern mit Sozialtherapie in Brandenburg vergleichbar sein. Tatsächlich finden sich schon in den gesetzlichen Grundlagen der Bundesländer klare Unterschiede (vgl. Suhling, 2008). Darüber hinaus weisen die Einrichtungen eigene Konzepte auf. Die Analysen von Niemz (2015, S. 25) zeigen, dass sich bereits auf dieser konzeptuellen Ebene eine „ausgesprochene Heterogenität zwischen den Bundesländern und den einzelnen Einrichtungen" findet. Suhling und Keßler (2015) berichten ähnliches für die Sozialtherapie im Jugendstrafvollzug. Im Bundesvergleich zeigt sich daher deutlich, dass „Sozialtherapie nicht gleich Sozialtherapie" ist (vgl. Suhling, 2008, S. 330; Spöhr, 2009; Wößner, 2014).

Selbst wenn innerhalb eines Bundeslands ein gemeinsames Konzept besteht, bestehen Unterschiede zwischen den Einrichtungen. So berichten Suhling und Guéridon (2016) von Unterschieden zwischen Sozialtherapeutischen Einrichtungen in Niedersachsen, die unabhängig von Merkmalen der Klienten bestehen. Die Ergebnisse weisen auf relevante Einflüsse der Einrichtung und damit verbundener Merkmale (z.B. ländlicher vs. urbaner Raum) bei verschiedenen Prozess- und Strukturmerkmalen hin, etwa bei der Häufigkeit der Gewährung von Vollzugslockerungen und der durchschnittlichen Behandlungsdauer. Solche Befunde sind durchaus zu erwarten. Die Psychotherapieforschung weist explizit auf die Bedeutung von Therapeuteneffekten hin (Lambert, 2013 a, 2013 b) und in der Bildungsforschung wird Erfolg von Schülern auch mit Unterschieden zwischen Lehrern und Schulen erklärt (Hattie, 2003). Es wäre überraschend, wenn sich verschiedene Sozialtherapeutische Einrichtungen, Wohngruppen und selbst unterschiedliche Behandler durch identische Praxis auszeichnen würden. Unterschiedliche Sicherheitsstufen und Klima der Einrichtung sind nur zwei weitere Aspekte, die sich auf die Praxis auswirken können.

Für die Wirksamkeit der Behandlung müssen solche Unterschiede kein Problem darstellen, auch wenn sich Implementation und Integrität als relevante Faktoren für die Wirksamkeit erwiesen haben (vgl. Suhling, 2008). Es kann durchaus für Spezialisierung argumentiert werden, wenn möglichst individuell „passende" Behandlung das Ziel ist. In Bayern wird diese Idee etwa mit getrennten Einrichtungen für Sexual- und Gewaltstraftäter umgesetzt. Inhaltlich ist die Aggregation der einzelnen Einrichtungen aber nicht ohne weiteres sinnvoll. Wenn verschiedene „Versionen" einer Maßnahme zusammengefasst werden, müssen die Ergebnisse entsprechend interpretiert werden. Man kann vermuten, dass bei recht groß skalierten Maßnahmen wie der Sozialtherapie das „Iron Law of Evaluation" greift (Rossi, 1987, 2003), nach dem im Mittel ein Nulleffekt über alle Einrichtungen erwartet werden kann. Ludwig, Kling und Mullainathan (2011, S. 24) vermuten explizit inkonsistente Implementation sozialer Maßnahmen als eine Erklärung solcher Befunde.

Eine Auswertung auf Ebene einzelner Einrichtungen könnte eine Lösung sein, reduziert aber die verfügbare Fallzahl stark. Dieses Problem erschwert auch die Nutzung

spezieller statistischer Methoden. Die Verwendung von Mehrebenenmodellen (vgl. Suhling & Guéridon, 2016) zur gemeinsamen Modellierung von Personen- und Einrichtungsmerkmalen ist etwa davon abhängig, dass genügend Datenpunkte auf allen Ebenen vorliegen. Die Empfehlungen liegt bei mindestens zehn Einheiten auf Ebene 2 (Maas & Hox, 2005), was bei der durchschnittlichen Anzahl sozialtherapeutischer Einrichtungen innerhalb eines Bundeslandes ein wesentliches Hindernis darstellt. Auch bei diesen Verfahren muss zudem die Integrität der Maßnahme gewährleistet ist.

Auch zeitliche Integrität ist zu beachten. Wie Wößner (2014, S. 55) betont, ist Sozialtherapie von „ständiger Weiterentwicklung und Anpassung" betroffen. Ergebnisse aus dem Jahr 2005 sind nicht ohne weiteres auf die heutige Sozialtherapie übertragbar, wenn sich die Praxis verändert hat. Eine Einschätzung über die Jahre hinweg ohne Modellierung der Zeit kann so positive oder negative Entwicklungen verdecken. Besonders deutlich wird dieser Punkt angesichts gesetzlicher Veränderungen wie der Neuschaffung der Landesjustizvollzugsgesetze.

Zuletzt betrifft die Forderung nach Integrität auch die Kontrollbedingung. Wird Sozialtherapie mit dem Regelvollzug verglichen, müssen Entwicklungen beider Seiten ebenso berücksichtigt werden wie das Problem gegenseitiger Beeinflussung (treatment contamination). Ein Beispiel dafür wäre die Einführung von Maßnahmen oder Teilen von Maßnahme im Regelvollzug. Umgekehrt beeinflusst der Regelvollzug auch die Sozialtherapie, wenn anstaltsweite Regelungen geändert werden. Auch Kontakte zwischen Klienten der Sozialtherapie und dem Regelvollzug sind bei sozialtherapeutischen Abteilungen eher die Regel (Niemz, 2015). Diese Interferenz zwischen den Bedingungen ist ein wesentliches Problem für die Schätzung kausaler Effekte (Heckman & Smith, 1995; Weisburd & Hinkle, 2012; siehe Abschnitt 3.4).

Schließlich ist auch die korrekte Implementation und Integrität der Evaluation selbst kritisch: Ein Eingangsstatus in den ersten drei Monaten nach Verlegung in die Sozialtherapie erfasst werden soll, dann muss das auch so geschehen oder aber der Eingangsstatus hat eine andere Bedeutung als ihm zugeschrieben wird.

3.2 Ziele der Maßnahme und Kriterien der Zielerreichung

Ob eine Maßnahme durch sozialwissenschaftliche Forschung evaluiert werden kann, hängt auch davon ab, ob sie klare Ziele hat, die sich in verlässlich messbare Kriterien übersetzen lassen. Die gesetzlichen Formulierungen benennen explizit die Wirksamkeit als Kriterium, wobei eine Maßnahme als wirksam bezeichnet werden kann, wenn sie ihre beabsichtigte Wirkung erreicht (vgl. Suhling, 2012). Darüber hinaus sollte die Zielerreichung wirtschaftlich angemessen gelingen, keine unerwünschten Nebeneffekte erzeugen und ethisch gerechtfertigt sein. Es geht also sowohl um das „ob" als auch das „wie" der Zielerreichung.

Sozialtherapeutische Einrichtungen teilen die Ziele des Strafvollzugs (vgl. Guéridon & Suhling, 2015; Suhling, 2012; Spöhr, 2009). Während bei einigen Maßnahmen, etwa beruflichen Ausbildungen, in der Regel geeignete Ziele identifiziert und

Kriterien recht gut erfasst werden können, ist dies bei komplexen Maßnahmen wie Sozialtherapie schwieriger. Diese zeichnet sich gerade durch die Kombination verschiedener Bausteine aus, was Bussmann, Seifert und Richter (2007, S. 280) auch als „Therapiecocktail" bezeichnen. Entsprechend vielseitig sind die Ziele der Sozialtherapie formuliert.

Zentrales Kriterium der Wirksamkeit bleibt formal die Legalbewährung, auch wenn das Rückfallkriterium inhaltlich hochumstritten und methodisch problematisch ist, vor allem wegen der Dunkelfeldproblematik. Mit wenigen Ausnahmen (z.B. Lauterbach, 2009) beruhen Rückfallanalysen ausschließlich auf Daten aus dem Hellfeld, also nur den „Erwischten", zudem wird eine Fehlerquote von bis zu 10 % in Bundeszentralregisterauszügen vermutet (Obergfell-Fuchs und Wulf, 2008, 2011). Zur Absicherung der Ergebnisse ist daher eine Ausdifferenzierung des Rückfallkriteriums sinnvoll (Kerner, 2013), zudem wären Sensitivitätsanalysen (vgl. Rosenbaum, 2010) eine Möglichkeit, die Verlässlichkeit der Ergebnisse abzusichern.

Als Alternative könnte die Veränderung des Inhaftierten während der Haft und so Risiko einer erneuten Straffälligkeit als Kriterium festgelegt werden gemessen (vgl. Suhling, 2012). Eine dafür notwendige standardisierte Abschlussuntersuchung findet im Strafvollzug bisher nicht statt, auch wenn es erste Pilotprojekte gibt (z.B. MeWiS in Hessen und Niedersachsen, vgl. Budde & Suhling, 2016). In der Sozialtherapie ist die Situation etwas besser, da zum Teil Falldokumentationssysteme bestehen, etwa in Bayern und Niedersachsen. Solche Messungen am Ende der Inhaftierung bzw. am Ende der Maßnahme sind dennoch keine alleinige Lösung. Es bleibt fraglich, ob man eine Maßnahme wirksam nennen kann, wenn zwar die Impulsivität reduziert wird, dies aber keinen Effekt auf erneute Straftaten hat. Veränderungen auf allgemeinen Persönlichkeitsfragebögen haben sich beispielsweise bisher als schlechte Prädiktoren erwiesen (Schwedler & Schmucker, 2012). Auch die Nachhaltigkeit der Effekte bleibt fraglich, wenn nur am Ende der Maßnahme gemessen wird. Im Kontext Strafvollzug sind Katamnesen kaum möglich, da Klienten nach Entlassung in der Regel möglichst wenig Kontakt mit der Justiz haben möchten. Wößner, Wienhausen-Knezevic und Gauder (2014) berichten etwa eine Teilnahme von nur 145 von ursprünglich 399 in Haft Befragten (36 %) bei einer Befragung ein Jahr nach Entlassung, obwohl es sich um externe Forschung handelt. Es ist zudem mit positiver Selektion zu rechnen (vgl. Kaiser & Byrka, 2011), wenn nur die generell prosozialeren Entlassenen teilnehmen.

Die angesprochenen Probleme betreffen ebenso die Idee, soziale Integration nach Entlassung als Erfolgskriterium zu verwenden (vgl. Suhling, 2012). Eine weitere, wenn auch nur begrenzt aussagefähige Alternative sind Verhaltensindikatoren während der Haft und die Integration am Haftende (vgl. Wakeling & Travers, 2010). Über Nachhaltigkeit hinaus geht das Problem echter Langzeiteffekte. Sampson, Winship & Knight (2013) geben Beispiele für Effekte, die sich erst nach vielen Jahren zeigen. Solche Effekte können besonders bei jungen Inhaftierten vermutet werden (vgl. Obergfell-Fuchs & Wulf, 2008, S. 232).

3.3 Datenerhebung und Datenqualität: Lassen sich verlässliche Daten gewinnen?

Wenn geeignete Kriterien gefunden werden, bleibt die Frage, ob eine verlässliche Messung möglich ist. Betrachtet man exemplarisch die im niedersächsischen Rahmenkonzept Sozialtherapie (Niedersächsisches Justizministerium, 2011) oder das Behandlungskonzept der Sozialtherapeutischen Anstalt in Gelsenkirchen (2006) benannten Behandlungsziele, finden sich schwer messbarer psychologischer Konstrukte. Benannt werden unter anderem soziale und kognitive Kompetenzen (empathische Fähigkeiten, Problemlösefähigkeit, emotionale Kontrolle) sowie metakognitive Fähigkeiten (persönliche Risikofaktoren erkennen). Neben der Frage geeigneter Messinstrumente ist zumindest auch an die Eignung der Datenquellen sowie an die Datenqualität zu denken.

Die am häufigsten verwendeten Methoden zur Einschätzung der Persönlichkeit und Fähigkeiten sind sicherlich Selbst- und Fremdauskunft. Beide Methoden haben bekannte Probleme. Selbstauskunftsverfahren setzen ehrliche Antworten voraus, zudem muss der Klient überhaupt die Fähigkeit haben, sich selbst korrekt einschätzen zu können. Das ist nicht selbstverständlich: Einerseits werden den Klienten Schwierigkeiten bei der Metakognition zugeschrieben, dennoch wird davon ausgegangen, dass sie ihre Kompetenzen verlässlich einschätzen können. In Bereichen wie der Emotionserkennung ist die Übereinstimmung von Selbsteinschätzungen und tatsächlicher Leistung generell umstritten (Ames & Kammrath, 2004; Kelly & Metcalfe, 2011). Aussagen über eigene Fähigkeiten scheinen daher auch jenseits von sozialer Erwünschtheit nur begrenzt verlässlich. Allgemeine Persönlichkeitsfragebögen scheinen zudem nicht einfach auf den Strafvollzug übertragen werden zu können, wie Hosser, Lauterbach & Camehn (2008) am Beispiel des Freiburger-Persönlichkeits-Inventars zeigen.

Auch Fremdeinschätzungen sind nicht ohne Probleme. Die Psychotherapieforschung legt nahe, dass auch Therapeuten Veränderungen ihrer Klienten nicht immer korrekt einschätzen (Lambert, 2013 a), zudem besteht die latente Gefahr von Verfälschungen. Schon Campbell (1979) befürchtete, dass die Messung von Indikatoren umso eher korrumpiert wird, je mehr diese für Entscheidungs- und Bewertungsprozesse herangezogen werden. Verzerrungen sind auch möglich, wenn Spielräume in der Interpretation bestehen: Ist die Zusage auf eine Arbeit gleichwertig zu einem unterschriebenen Arbeitsvertrag? Ist Übergangsgeld ein Ausdruck davon, dass die finanzielle Situation geregelt ist?

Diese Prozesse betreffen auch Einschätzungen durch die Evaluatoren. McCord (2003, S. 17) führt als mögliche Erklärung für den „Code of Silence“ bei negativen Wirkungen von Maßnahmen die Sorge an, dass Behandlung insgesamt diskreditiert werden könnte. Evaluatoren sind nicht frei von Überzeugungen, die Entscheidungen im Forschungsprozess beeinflussen können. Auch solche „researcher degrees of freedom“ (Simmons, Nelson & Simonsohn, 2011, S. 1359) können ein Problem für die Validität der Analysen sein.

Alternative Datenquellen wie Gutachten und die Gefangenenpersonalakte sind zwar nützlich, aber nicht vor falschen, widersprüchlichen oder fehlenden Informationen si-

cher. Obwohl häufig genutzt, ist die Güte dieser Datenquellen wenig erforscht und vorhandene Ergebnisse sind eher kritisch. Wolter und Häufle (2014) berichten beispielsweise von erheblichen Diskrepanzen zwischen dem Ausmaß selbstberichteter Gewalthandlungen und den registrierten Vorkommnissen. Die Reliabilität der Aktenanalysen dürfte ein weiteres Problem sein.

Implizite Verfahren und Leistungstests sind zwar schwieriger zu erheben, können in gewissem Maß durch bewusst schlechte Leistungen verfälscht werden und sind möglicherweise mit größeren Hemmungen der Inhaftierten konfrontiert, stellen aber dennoch eine Alternative mit insgesamt höherer Validität dar, wenn ausreichende Konstruktvalidität gegeben ist (vgl. Babchishin, Nunes & Hermann, 2013). Gerade implizite Verfahren könnten nützlich sein, um schwer objektivierbare Konstrukte wie sexuelle Präferenzen abzubilden (vgl. Schmidt, Banse & Imhoff, 2015; Schmidt, Gykiere, Vanhoeck, Mann & Banse, 2014).

Relevant ist auch die Frage der Datenqualität, besonders der Aspekt fehlender Daten. Aus eigener Forschungspraxis sind Gefangenenpersonalakten gerade für die Entlassungssituation oft wenig ertragreich. Regelmäßig ist der letzte Vollzugsplan die einzige Informationsquelle, aber zwischen diesem und der Entlassung können Monate liegen. Dazu kommt, dass viele wichtige Informationen in der Behandlungsuntersuchung zwar enthalten sind, aber nicht im Verlauf berichtet werden. Eine mögliche, aber hochproblematische Lösung ist die Gleichsetzung dieser fehlender Werte mit dem letzten bekannten Wert (last obervation carried forward). Für die Behandlung fehlender Veränderungswerte ist diese Methode sehr kritisch zu sehen (Carpenter, Kenward, Evans, & White, 2004; Cook, Zeng & Yi, 2004; Molenberghs et al., 2004). Das Verfahren ist auch nicht per se konservativ, da nicht vermerkte negative Entwicklungen maskiert werden können. Fehlende Daten sind auch darüber hinaus ein Problem. In der Studie von Schwedler und Schmucker (2012) reduziert sich die Stichprobe durch fehlende Daten in Selbstauskunftsverfahren um 24 %. Bei der Evaluation der Sozialtherapie in Sachsen (Wößner et al., 2014) sind am Ende der Haft nur noch 303 von 399 Befragten (76 %) in der Stichprobe, bei der Befragung ein Jahr nach der Entlassung nur noch 36 %. Für die Problematik fehlender Werte gibt es einige gute statistische Lösungsmöglichkeiten (Multiple Imputation, Maximum Likelihood Schätzungen, vgl. Enders, 2010, Graham, 2009), diese setzen aber ausreichend große Stichproben voraus und machen zum Teil starke Annahmen.

Für verlässliche Schlüsse sind ausreichend große Datensätze notwendig, weshalb die die angesprochene Fallzahl ein generelles Problem ist. Auch wegen langer Behandlungsdauern sind größere Stichproben selten: Bei Hosser, Bosold und Lauterbach (2006) liegen lediglich 17 vollständige Datensätze vor. Schwedler und Schmucker (2012) geben für ihre Untersuchung in Erlangen im Zeitraum 1996-2005 eine Fallzahl von 92 an. Bieschke (2014) evaluiert die Sozialtherapie Neustrelitz anhand von rund 30 Fällen in der Behandlungsgruppe. Ortmann (2002) erreichte nur eine Fallzahl von 114 in der Experimentalgruppe in den Jahren 1982-1990.

Barbaree (1997) argumentiert, dass bei so geringen Fallzahlen und geringen Basisraten die Wahrscheinlichkeit eines Typ-II-Fehlers erhöht ist. Gerade bei Sexualstraftaten,

wo eine geringe (einschlägige) Rückfälligkeit im Hellfeld gefunden wurde (Jehle, Albrecht, Hohmann-Fricke & Tetal, 2013), ist das Risiko daher höher, einen tatsächlich bestehenden Effekt fälschlich zurückzuweisen. Lösungen wie ein längerer Betrachtungszeitraum zur Vergrößerung der Stichprobe (Barbaree, 1997) erhöhen umgekehrt die Schwierigkeit, die korrekte Implementation und Integrität der Maßnahme sicherzustellen.

3.4 Kausalität: Lassen sich Effekte verlässlich auf die Sozialtherapie zurückführen?

Dass Ziel einer Evaluation der Wirksamkeit ist letztlich eine Schätzung kausaler Zusammenhänge (Lösel, 2008, S. 143), die Veränderung soll durch die Maßnahme erzeugt werden. Eine solche Schätzung steht immer vor dem sogenannten fundamentalen Problem der Kausalität (Holland, 1986; Imbens & Rubin, 2015; Morgan & Winship, 2014), dass eine Person niemals gleichzeitig in zwei sich ausschließenden Zuständen beobachtet werden kann. Es ist daher nicht möglich, den Effekt einer Maßnahme auf Einzelpersonen als einfache Differenz aus zwei Beobachtungen zu schätzen. Stattdessen muss auf einen Vergleich mit einer (unbehandelten) Kontrollgruppe zurückgegriffen werden. Ein einfacher Vorher-Nachher-Vergleich ist angesichts der Komplexität der sozialtherapeutischen Behandlung schlicht kein geeignetes Verfahren, um verlässliche Aussagen zur Wirkung zu machen.

Das Fehlen einer geeigneten Kontrollgruppe ist daher vermutlich das größte Problem der Evaluation der Sozialtherapie, obwohl auf den ersten Blick verschiedene Kontroll- oder Kontrastgruppen zur Verfügung zu stehen scheinen. Historische Daten früherer Inhaftierter stammen aus anderen Kontexten und sind oft nicht ausreichend dokumentiert. Andere Vergleichsmöglichkeiten bieten Personen, deren Antrag auf eine Verlegung in die Sozialtherapie abgelehnt wurde oder die eine Verlegung verweigern. Beide Gruppen weisen inhaltlichen Besonderheiten auf und sind zahlenmäßig eher zu klein, um geeignete Kontrollen darzustellen. Auch eine Warteliste oder eine anfallende Kontrollgruppe wegen geringer Kapazitäten (beispielsweise Bieschke, 2014) ist nur in Ausnahmefällen realisierbar. Es bleibt die Möglichkeit, andere „Unbehandelte" zum Vergleich heranzuziehen, die aus verschiedenen Gründen nicht in die Sozialtherapie verlegt wurden. Dabei sollte die Ähnlichkeit zur Interventionsgruppe möglichst groß sein und zumindest eine ähnliche, idealerweise identische Wahrscheinlichkeit bestehen, die Intervention zu erhalten. Genau dies ist die große Stärke der Randomisierung, die für alle Personen die gleiche Zuweisungswahrscheinlichkeit erzeugt. Es verwundert nicht, dass die Forderung an die Forschung daher lautet, zur Überprüfung der Wirksamkeit randomisierte Experimente durchzuführen. Einige immer wieder gegen randomisierte Experimenten vorgebrachte Einwände wie ethische Bedenken sind bei genauer Betrachtung zudem nicht gültig (Weisburd, 2010).

Ein randomisiertes Experiment ist im Kontext der Sozialtherapie aber schon durch die gesetzlichen Grundlagen nicht möglich, da es sich in der Regel um „ist"-Reglungen handelt (Suhling, 2008). Selbst wenn Randomisierung möglich wäre, bestehen indes

Schwierigkeiten. Zum einen ist erst bei ausreichend großen Stichproben davon auszugehen, dass Randomisierung erfolgreich ist. Wie schon beschrieben (Abschnitt 3.2), wird es sich bei der Evaluation der Sozialtherapie in den meisten Fällen um kleinere Fallzahlen handeln. Dazu kann festgehalten werden, dass Therapieabbrüche selektiv sind (vgl. Olver, Stockdale & Wormith, 2011). Endres, Breuer & Schmucker (2016) fordern daher auch im Falle von randomisierten Experimenten sogenannte Intent-to-Treat-Analysen (ITT). Bei solchen Analysen werden auch Abbrecher in die Auswertung einbezogen, um tatsächliche Effekte einer Maßnahme zu schätzen. Der Nachteil an ITT-Analysen ist, dass nicht die Leistungsfähigkeit einer Maßnahme, sondern im wesentlichen Effekte der Zuweisungspraxis betrachtet werden (vgl. Imbens & Rubin, 2015, S. 514). Eine Kombination beider Ansätze scheint daher sinnvoll.

Verzerrungen durch die Randomisierung selbst sind ebenfalls möglich (Heckman & Smith, 1995), wenn durch die Randomisierung eine andere Klientel behandelt wird als im Realbetrieb (randomization bias) oder die Kontrollgruppe Therapiesubstitute erhält (substitution bias). Letztere sind im Falle der Sozialtherapie wahrscheinlich, da viele Bestandteile der Sozialtherapie zum Teil oder sogar identisch auch im Regelvollzug angeboten werden.

Eine weitere Quelle von Einschränkungen verlässlicher Schätzungen sind Verletzungen der sogenannten SUTVA (stable unit treatment value assumption). Das Konzept aus dem Kontext des kontrafaktischen Kausalmodells definiert zwei Grundannahmen, die für unverzerrte Schätzungen gegeben sein müssen (vgl. Guo & Fraser, 2014, Imbens & Rubin, 2015, Morgan & Winship, 2014). Zum einen darf die Zuweisung einer Person keinen Einfluss auf die potentiellen Ergebnisse einer anderen Person haben („No Interference“, Imbens & Rubin, 2015, S. 10). Die Zuweisung einer Person X in die Sozialtherapie darf also keine Wirkung auf die Ergebnisse der anderen Personen in Sozialtherapie- und Kontrollgruppe haben. Diese Annahme dürfte im Fall der Sozialtherapie kaum erfüllt sein. Personen sollen in Gruppenmaßnahmen ja gerade interagieren und die Wohngruppenzusammensetzung ist für das Ergebnis des Einzelnen natürlich relevant. Die Analysen von Niemz (2015) haben auch gezeigt, dass soziale Interaktion zwischen Sozialtherapie und Regelvollzug üblich ist. Solche Interferenzeffekte bedeuten auch, dass eine Änderung der Zuweisungspraxis, etwa die Aufnahme drogenabhängiger Inhaftierter, natürlich auch die Wirksamkeit der Sozialtherapie verändern kann. Die zweite Annahme fordert, dass keine versteckten Variationen der Maßnahme existieren („No Hidden Variations of Treatments“, Imbens & Rubin, 2015, S. 11). Der mittlere erwartete Effekt einer Person würde in einem solchen Fall nicht nur durch die Zuweisung zur Sozialtherapie bestimmt, sondern auch durch die konkrete Zuweisung zu einer bestimmten Abteilung, die sich durch bestimmte Merkmale von anderen Einrichtungen unterscheidet (z.B. die Länge der Einzeltherapiesitzungen). In diesem Fall würden zwei Versionen der Maßnahme existieren und die Schätzung der Effekte wäre nicht verlässlich, wenn ein Zusammenhang zwischen dem jeweiligen Merkmal und dem Kriterium besteht.

Die Annahmen der SUTVA wurden verschiedentlich als zu streng kritisiert, was zu alternativen Konzepten und Lösungsvorschlägen bei Verstößen gegen die SUTVA ge-

führt hat (vgl. Heckman & Vytlacil, 2005; Rosenbaum, 2007; Manski, 2013; VanderWeele & Hernan, 2013; VanderWeele, Tchetgen & Halloran, 2014). Letztlich ist selbst perfekte Randomisierung aber nichts wert, wenn Fragestellung, Ziele und Messung der relevanten Konstrukte unklar oder unverlässlich sind. Eine schlecht konzeptualisierte und gemessene Persönlichkeitseigenschaft bleibt unzuverlässig, unabhängig vom Forschungsdesign.

Alternativen zu randomisierten Experimenten bieten die zahlreichen Varianten quasiexperimenteller Designs (vgl. Legewie, 2012; Shadish & Cook, 2009; Shadish et al., 2002), wobei die Probleme randomisierter Experimente überwiegend in gleichem Maße zutreffen. Als wahrscheinlich beste Alternative zur Randomisierung werden sogenannte Regression Discontinuity Designs (RDD, Berk, 2010; Shadish et al., 2002) angeführt, bei denen die Zuweisung zu einer Maßnahme explizit über einen Cutoff-Wert festgelegt wird. Zur Schätzung von Effekten werden jene Personen aus Kontroll- und Interventionsgruppe vergleichen, deren Zuweisungswahrscheinlichkeit direkt neben dem Trennwert liegen. Ein solches Verfahren wäre möglich, wenn nur Personen ab einem bestimmten Risikoscore in die Sozialtherapie verlegt würden. Man würde dann jene Personen miteinander vergleichen, die um diesen Wert herum liegen. RDD können ebenso verlässlich sein wie randomisierte Experimente (Berk, Barnes, Ahlman & Kurtz, 2010; Shadish, Galindo, Wong, Steiner, & Cook, 2011). Die Schwäche der Designs ist die Bindung an den Cutoff-Wert: Ist die Wahrscheinlichkeit einer Zuordnung zu unscharf („fuzzy"), ist eine zuverlässige Schätzung der Effekte nicht mehr gesichert. Ob RDD im Falle der Sozialtherapie anwendbar sind, ist bisher offen, auch da Forschungsergebnisse zur Selektion in die Sozialtherapie kaum verfügbar sind. Seifert (2014) konnte für die Sozialtherapie Halle aber zeigen, dass das eingeschätzte Risiko einer erneuten Straftat nicht gut vorhergesagt hat, wer in die Sozialtherapie verlegt wurde.

Ist die Wahrscheinlichkeiten einer Verlegung in die Sozialtherapie in Behandlungs- und Kontrollgruppe ausreichend ähnlich, werden Stratifizierung und andere Matching-Verfahren zu attraktiven Alternativen. Gerade Propensity Score Matching wird zunehmend häufiger verwendet (vgl. Guo & Fraser, 2014). In der Forschung zur Sozialtherapie wurde Matching zum Teil verwendet, wobei meist paarweises Matching auf wenigen Merkmalen verwendet wurde. Die Evaluation der Sozialtherapie in Sachsen nutzt etwa nur das Alter und das Delikt (Wößner, 2013). Auch Matching-Verfahren sind nicht ohne Einschränkungen (King & Nielsen, 2016; Shadish, 2013; Steiner, Cook & Shadish, 2011), vor allem sind reliable und valide Messungen der verwendeten Merkmale nötig.

4. Ist eine nützliche Evaluation möglich?

Eine wissenschaftlich hervorragende, randomisiert experimentell durchgeführte Studie an einer große Zahl von Probanden mit ausgezeichnet gemessenen Daten ist letztlich nichts wert, wenn Praxis und Öffentlichkeit nicht von den Ergebnissen erfahren, die

eigentlichen Fragen nicht beantwortet werden, das Ergebnis keinen Einfluss auf die Praxis haben kann oder die Ergebnisse hoffnungslos veraltet sind.

Schon Cronbach (1982) betonte, dass das Ziel von Evaluation letztlich in der Förderung von Lernprozessen bestehe. Auch und gerade negative Ergebnisse sind für solche Lernprozesse notwendig. Es ist damit eine Frage der Verantwortung, auch unerfreuliche Ergebnisse breit zu kommunizieren Wer Evaluationen in Auftrag gibt, sollte zudem wissen, dass nicht jede Frage beantwortet werden kann, dass Ergebnisse oft uneindeutig und Empfehlungen vage sein werden. Rossi, Freeman und Lipsey (2003, S. 297) argumentieren entsprechend, dass Evaluatoren den Auftraggeber im Falle der Verwendung von Forschungsdesigns ohne Randomisierung darüber unterrichten sollten, dass die Ergebnisse nicht endgültig sicher sind. Angesichts der zahlreichen Probleme, die auch randomisierte Studien im Falle der Sozialtherapie mit sich bringen, kann diese Empfehlung wohl generell gelten.

Ergebnisse von Evaluationen sollten die Praxis auch dann unterstützen und verändern können, wenn es sich um summative Evaluationen der Wirksamkeit handelt (Cronbach, 1982). Wenn eine Studie nahelegt, das Sozialtherapeutische Behandlung nicht wirkt, dann ist diese Information für sich allein genommen zwar wichtig, beantwortet aber nicht die daran anschließenden Fragen: Warum funktioniert die Maßnahme nicht? Was muss anders gemacht werden, damit die Maßnahme funktioniert? Die entscheidende Frage ist oft eher das praktische „What will work?“ (Sampson, Winship & Knight, 2013; Cartwright & Hardin, 2012) als die einfache Frage, ob etwas wirkt. Auch Heckman und Smith (1995) kritisieren die Durchführung von „black-box“-Experimenten, bei denen zwar die Größe eines mittleren Effekts bestimmt wird, aber Wirkprozesse nicht untersucht werden und so keine Richtung für Verbesserungen mitgegeben werden kann.

Gerade wenn erneute Straffälligkeit als Kriterium für den Erfolg von Maßnahmen herangezogen wird, besteht zudem die Gefahr, über längst vergangene Zeiten zu sprechen. Obergfell-Fuchs und Wulf (2008) benennen die Studie von Ortmann (2002) als Beispiel für exakt diese Situation: Die methodisch hochwertige Studie konnte kaum praktische Relevanz entfalten, da bei Veröffentlichung eine der beiden untersuchten Einrichtungen geschlossen und die zweite Einrichtung bereits stark verändert worden war. Wenn der internationale Richtwert für den Beobachtungszeitraum nach Entlassung von fünf Jahren zugrunde gelegt wird, kann man einfach errechnen, dass jede Rückfalluntersuchung vor diesem Risiko steht. Inklusive der Bearbeitungszeit durch das Bundeszentralregister nach Beantragung und einer sehr schnellen Auswertung der BZR-Auszüge, die ausreichend Personalkapazität voraussetzt, könnte vermutlich frühestens Mitte 2022 mit Ergebnissen für 2016 Entlassene gerechnet werden. Es ist anzunehmen, dass sich die Sozialtherapie in diesem Zeitraum ebenfalls verändert und die Ergebnisse daher hinsichtlich ihrer Gültigkeit hinterfragt werden sollten.

5. Was kann getan werden?

Nimmt man alle bis hierhin angesprochenen Problem und Schwierigkeiten zusammen, scheint es vielleicht so, als ob eine sinnvolle Evaluation der Sozialtherapie kaum möglich ist. Dem Aufwerfen von Problemen sollten aber auch Vorschläge folgen, was zur Rettung getan werden könnte. Ohne Anspruch auf Vollständigkeit sind dafür verschiedene Ansätze denkbar:

a) In jedem Fall muss eine Kontrolle der Ergebnisse stattfinden. Da die potentiellen Kontrollgruppen allesamt Schwierigkeiten aufweisen und Randomisierung nicht möglich scheint, ist ein Design mit multiplen Kontrollgruppen (Shadish et al., 2002) eine Möglichkeit zur Erhöhung der Verlässlichkeit. Die Verwendung elaborierter Methoden wie Regressions-Diskontinuitäten-Designs oder Matching-Verfahren kann eine zusätzliche Absicherung darstellen, gerade wenn die Methoden um Sensitivitätsanalysen ergänzt werden.
b) Evaluation der Sozialtherapie kann aus den dargestellten Gründen nicht nur Wirksamkeitsevaluation sein. Ebenso wichtig sind die formative Evaluation der Implementation sowie die genaue Analyse von Strukturen und Prozessen. Dazu gehört primär der Zuweisungsmechanismus, aber ebenfalls die Analyse von Therapieabbrüchen, der Unterschiede zwischen den verschiedenen Einrichtungen, des Klimas in den Einrichtungen, aber auch von Merkmalen und Einstellungen des Personals.
c) Statt allgemein und global die Wirksamkeit zu untersuchen, könnte auf spezifischere Fragestellungen ausgewichen werden: Statt Sozialtherapie im Ganzen zu untersuchen, könnte der Fokus mehr auf die Bewertung einzelner Maßnahmen innerhalb der Sozialtherapie gelegt werden. Dabei muss klar sein, dass Rückfälligkeit nicht im Zentrum dieser Fragestellungen stehen kann.
d) Der vielleicht wichtigste Punkt ist die Diagnostik. Wenn es gelingt, verlässliche psychometrische Instrumente zu entwickeln und zumindest am Anfang und Ende der Maßnahme bzw. der Haft Erhebungen durchzuführen, wäre die Datenbasis sehr viel besser. Implizite Tests für Einstellungen und Präferenzen sowie Leistungstests für Fähigkeiten wie Selbstkontrolle und Empathie hätten dabei Vorteile gegenüber problematischen Messungen über Selbstauskünfte und Einschätzungen der Behandler. Zumindest sollten multiple Methoden verwendet werden. Es ist auch an andere Formen der standardisierten oder gar „blinden" Diagnostik zu denken, etwa in Form zentraler Diagnostikeinrichtungen, die keine Information darüber erhalten, ob die untersuchte Person am Anfang oder Ende der Haft steht.
e) Eine Analyse der Entwicklungsprozesse wird möglich, wenn während der Maßnahme weitere Messungen durchgeführt werden. Dazu muss ein Weg gefunden werden, der für die Praxis verträglichen Aufwand bedeutet und dennoch ausreichend umfangreiche Daten zur Verfügung stellt. Auch katamnestische Messungen zur Ermittlung der Nachhaltigkeit der Maßnahmen und der sozialen Integration können eine sinnvolle Ergänzung sein.

f) Wenn geeignet, sollten auch qualitative Methoden verwendet werden. Der größte Nutzen ist von einer pluralen Forschungslandschaft zu erwarten, die alle möglichen Erkenntniszugänge nutzt, solange sie verlässliche Einschätzungen zulassen (vgl. Dollinger, 2015; Maruna, 2015).

6. *Evaluation der Sozialtherapie – zum Scheitern verurteilt?*

Am Anfang dieses Beitrags wurde die Frage aufgeworfen, ob die Wirksamkeit sozialtherapeutischer Behandlung möglicherweise durch sozialwissenschaftliche Forschung nicht sinnvoll eingeschätzt werden kann. Obwohl für das Beispiel Sozialtherapie verschiedene Probleme immer wieder benannt werden (Lösel, 2008; Spöhr, 2009; Suhling, 2012; Wößner, 2014), wurde diese „letzte" Frage nach aktueller Kenntnis bisher nicht systematisch diskutiert.

Als Fazit könnte man festhalten: Es kommt darauf an. Diese Antwort ist vielleicht unbefriedigend, trifft aber genau die Problematik wissenschaftlichen Bewertung des Strafvollzugs und besonders der Sozialtherapie. Vermieden werden sollte eine einseitige Betonung eines Lösungsvorschlags. Natürlich können durch Randomisierung Einschränkungen der internen Validität am ehesten ausgeschlossen werden, aber ohne Prozess-, Struktur- und Verlaufsanalysen mit geeigneten psychometrischen Instrumenten bleiben Aussagekraft und Nützlichkeit beschränkt. Umgekehrt helfen die beste Datenerhebung und aktuellste statistische Methoden dem Justizvollzug nicht weiter, wenn der Erfolg ausschließlich am problematischen Rückfallkriterium festgemacht wird. Schließlich nützt Forschung nichts, wenn die Evaluation nicht ergebnisoffen ist und Lernprozesse nicht unterstützt werden können, weil Ergebnisse nicht veröffentlicht werden.

Eine Evaluation der Sozialtherapie wird immer ein etwas unsicheres Fundament haben, ob randomisierte Experimente verwendet werden oder nicht. Es ist dennoch deutlich erkennbar, dass die aktuelleren Studien wie die Evaluation in Sachsen, (Wößner et al., 2014), in Neuburg-Herrenwörth (Klein, Schmucker & Lösel, 2015), Bayern (Endres, 2014) oder Niedersachsen (Guéridon & Suhling, 2015; Suhling & Guéridon, 2016) nicht nur methodisch aktuelle Entwicklungen aufgreifen (Matching-Verfahren, Mehrebenenanalysen, multiple Kontrollgruppen), sondern auch Schwerpunkte in der Prozess- und Strukturevaluation setzen (Klima, Mitarbeiter, Unterschiede zwischen Einrichtungen, Therapieabbrüche) und verschiedene Kriterien verwenden (Rückfall, Veränderung in Therapie, soziale Integration).

Das Wichtigste für die Weiterentwicklung der Evaluationsforschung scheint zu sein, dass Evaluationsforschung und Selbstreflektion von Praxis und Justizverwaltung wirklich gewollt werden und nicht notwendiges Übel mit Alibifunktion sind. Von Inhaftierten wird verlangt, eigenes Handeln zu hinterfragen, Verhaltensänderung und Behandlung auch wirklich zu wollen – und nicht etwa nur teilzunehmen, um Vergünstigungen zu erhalten und früher entlassen zu werden. Das sollte dann aber auch für die Beziehung von Justizvollzug, Politik und Evaluationsforschung gelten. Dann ist eine

sinnvolle, verlässliche und nützliche Evaluation im Rahmen genereller wissenschaftlicher Grenzen auch möglich.

Literatur

Ames, D. R., & Kammrath, L. K. (2004). Mind-reading and metacognition: Narcissism, not actual competence, predicts self-estimated ability. *Journal of Nonverbal Behavior*, *28*(3), 187-209.

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct*. New Providence: LexisNexis.

Ansorge, N. (2011). Evaluation von Naikan im Justizvollzug – Erste Zwischenergebnisse. In N. Saimeh (Hrsg.*). Kulturelle und therapeutische Vielfalt im Maßregelvollzug*. Bonn: Psychiatrie Verlag.

Arbeitskreis Sozialtherapeutischer Anstalten im Justizvollzug e.V. (2012). Sozialtherapeutische Anstalten und Abteilungen im Justizvollzug: Mindestanforderungen an Organisation und Ausstattung sowie Indikation zur Verlegung. Revidierte Empfehlungen. In Wischka, B., Pecher, W. & van den Boogaart, H. (Hrsg.), *Behandlung von Straftätern. Sozialtherapie, Maßregelvollzug, Sicherungsverwahrung* (S. 20-26). Herbolzheim: Centaurus.

Babchishin, K. M., Nunes, K. L., & Hermann, C. A. (2013). The validity of Implicit Association Test (IAT) measures of sexual attraction to children: A meta-analysis. *Archives of sexual behavior*, *42*(3), 487-499.

Balzer, L. (2005). Wie werden Evaluationsprojekte erfolgreich. *Ein integrierender theoretischer Ansatz und eine empirische Studie zum Evaluationsprozess*. Landau: Verlag Empirische Pädagogik.

Barbaree, H. (1997). Evaluating treatment efficacy with sexual offenders: The insensitivity of recidivism studies to treatment effects. *Sexual Abuse: A Journal of Research and Treatment*, *9*(2), 111-128.

Berk, R. A. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology*, *1*(4), 417-433.

Berk, R. (2010). Recent Perspectives on the Regression Discontinuity Design. In: Piquero, A. R., & Weisburd, D. (Hrsg.). *Handbook of quantitative criminology* (563-579). New York: Springer.

Berk, R., Barnes, G., Ahlman, L., & Kurtz, E. (2010). When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology*, *6*(2), 191-208.

Bieschke, V. (2014). Evaluation der Sozialtherapeutischen Abteilung in der Jugendanstalt Neustrelitz. Forum Strafvollzug. *Forum Strafvollzug, 63* (4), 232-237.

Budde, S. & Suhling, S. (2016). MeWiS – Wirksamkeitsmessung im Vollzug – Wie wirksam ist der Justizvoll-ZUG? In: Koop, G. & Kappenberg, B. (Hrsg.). *Weichen gestellt für den Justizvollzug?* (201-214). Wiesbaden: Gesellschaft für Fortbildung der Strafvollzugsbediensteten e.V.

Bussmann, K.-D., Seifert, S. & Richter, K. (2007). Sozialtherapie im Strafvollzug: Die kriminologische Evaluation der Sozialtherapeutischen Anstalt Halle (Saale). In Bender, D.; Jehle, J.-M.; Lösel, F. (Hrsg.) Kriminologie und wissensbasierte Kriminalpolitik: Entwicklungs- und Evaluationsforschung (279-294). Mönchengladbach: Forum Verlag Godesberg.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and program planning*, *2*(1), 67-90.

Carpenter, J., Kenward, M., Evans, S., & White, I. (2004). Last observation carry-forward and last observation analysis. *Statistics in medicine*, *23*(20), 3241-3242.

Cartwright, N. (2007). Are RCTs the gold standard?. *BioSocieties*, *2* (1), 11-20.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Lodon: Oxford University Press.

Cook, R. J., Zeng, L., & Yi, G. Y. (2004). Marginal analysis of incomplete longitudinal binary data: a cautionary note on LOCF imputation. *Biometrics*, *60*(3), 820-828.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

DeGEval – Gesellschaft für Evaluation e.V. (2008) (Hrsg.): *Standards für Evaluation.* 4. unveränderte Auflage. Mainz.

Dennis JA, Khan O, Ferriter M, Huband N, Powney MJ, Duggan C. (2012). *Psychological interventions for adults who have sexually offended or are at risk of offending*. Cochrane Database of Systematic Reviews, Issue 12. Art. No.: CD007507.

Dollinger, B. (2015). Was wirkt aus wessen Perspektive? Aktuelle Tendenzen der »evidence-based criminology« und ihre Konsequenzen für Politik und professionelle Praxis. *Monatsschrift für Kriminologie und Strafrechtsreform, 98* (5), 428-443.

Durlauf, S.N. & Nagin, D.S. (2011). Imprisonment and crime. Can both be reduced? *Criminology & Public Policy, 10* (1), 13-54.

Egg, R., & Niemz, S. (2013). Die Entwicklung der Sozialtherapie im Justizvollzug im Spiegel empirischer Erhebungen. In Wischka, B., Pecher, W. & van den Boogaart, H. (Hrsg.), *Behandlung von Straftätern. Sozialtherapie, Maßregelvollzug, Sicherungsverwahrung* (S. 1-20). Herbolzheim: Centaurus.

Elz, J. (2015). *Sozialtherapie im Strafvollzug 2015: Ergebnisübersicht zur Stichtagserhebung zum 31.3.2015*. Wiesbaden: KrimZ.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Endres, J. (2014). Determinanten der Behandlungsteilnahme und des Behandlungsabbruchs bei inhaftierten Sexualstraftätern. *Forum Strafvollzug*, *40*(3), 237-243.

Endres, J., Breuer, M., & Stemmler, M. (2016). „Intention to treat" oder „treatment as received". *Forensische Psychiatrie, Psychologie, Kriminologie*, 10(1), 45-55.

Farabee, D. (2005). *Rethinking rehabilitation: Why can't we reform our criminals?* Washington, DC: AEI Press.

Farrington, D. P. (2003). Methodological quality standards for evaluation research. *The Annals of the American Academy of Political and Social Science*, *587*(1), 49-68.

Gollwitzer, M., & Jäger, R. S. (2014). *Evaluation kompakt*. Weinheim: Beltz.

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*, 549-576.

Guéridon, M. & Suhling, S. (2015). Sozialtherapie im Jugendstrafvollzug: Was ist das, was soll das und bringt das was? *Zeitschrift für Jugendkriminalrecht und Jugendhilfe, 26,* 130-139.

Guo, S., & Fraser, M. W. (2014). *Propensity Score Analysis: Statistical Methods and Applications*. Los Angeles: Sage Publications.

Hammersley, M. (2014). *Against 'Gold Standards' in Research: On the Problem of Assessment Criteria*. Verfügbar unter: http://www.degeval.de/fileadmin/users/Arbeitskreise/AK_Methoden/Hammersley_Saarbrucken.pdf (letzter Zugriff am: 7.3.2016):

Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). The principles of effective correctional treatment also apply to sexual offenders a meta-analysis. *Criminal justice and behavior*, *36*(9), 865-891.

Häßler, U. (2012). Gefängnisse produzieren Rückfall. *Forum Strafvollzug, 61* (6), 334-340.

Hattie, J. (2003). *Teachers make a difference: what is the research evidence?.* Melbourne: Australian Council for Educational Research.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives*, *9*(2), 85-110.

Heckman, J. J., & Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, *73*(3), 669-738.

Hoberman, H. M. (2016). Forensic Psychotherapy for Sexual Offenders: Has Its Effectiveness Yet Been Demonstrated?. In Phenix, A. & Hoberman, H.M. (Hrsg.) *Sexual Offending: Predisposing Antecedents, Assessments and Management* (605-666). New York: Springer.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945-960.

Hosser, D., Bosold, C. & Lauterbach, O. (2006). Sozialtherapeutische Behandlung von jungen Sexualstraftätern. Ergebnisse einer Evaluationsstudie. *Recht & Psychiatrie, 24*, 125-133.

Hosser, D., Lauterbach, O. & Camehn, K. (2008). Validität und Reliabilität des FPI-R beim Einsatz im Strafvollzug. *Diagnostica 54*, 129-137.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jehle, J. M., Albrecht, H. J., Hohmann-Fricke, S., & Tetal, C. (2013). *Legalbewährung nach strafrechtlichen Sanktionen: Eine bundesweite Rückfalluntersuchung 2007 bis 2010 und 2004 bis 2010*. Mönchengladbach: Forum Verlag Godesberg.

Kaiser, F. G., & Byrka, K. (2011). Environmentalism as a trait: Gauging people's prosocial personality in terms of environmental engagement. *International Journal of Psychology*, *46*(1), 71-79.

Kelly, K. J., & Metcalfe, J. (2011). Metacognition of emotional face recognition. *Emotion*, *11*(4), 896-906.

King, G. & Nielsen, R. (2016). *Why Propensity Scores Should Not Be Used for Matching*. Verfügbar unter: http://j.mp/1sexgVw (Letzter Zugriff am 7.3.2016)

Kerner, H.-J. (2013). Überlegungen zu einer differenzierten Rückfallforschung. Das Beispiel Jugendstrafvollzug. *Forum Strafvollzug 62* (6), S. 354-357.

Klein, R., Schmucker, M., & Lösel, F. (2015). Evaluation der sozialtherapeutischen Abteilungen der JVA Neuburg-Herrenwörth: Evaluationskonzept & erste Ergebnisse. In DVJJ e.V. (Hrsg.), *Jugend ohne Rettungsschirm. Dokumentation des 29. Jugendgerichtstages 14.-17.9.2013 in Nürnberg* (375-391). Mönchengladbach: Forum Verlag Godesberg GmbH.

Lambert, M.J. (2013 a). The Efficacy and Effectiveness of Psychotherapy. In Lambert, M. J. (Hrsg.). *Bergin and Garfield's handbook of psychotherapy and behavior change* (219-257). Hoboken: John Wiley & Sons.

Lambert, M. J. (2013 b). Outcome in psychotherapy: the past and important advances. *Psychotherapy, 50* (1), 42-51.

Lauterbach, O. (2009). Soziale Integration und Delinquenz nach Entlassung aus dem Jugendstrafvollzug. *Zeitschrift für Jugendkriminalrecht und Jugendhilfe, 20*, 44-50.

Legewie, J. (2012). Die Schätzung von kausalen Effekten: Überlegungen zu Methoden der Kausalanalyse anhand von Kontexteffekten in der Schule. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *64*(1), 123-153.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on psychological science*, *2*(1), 53-70.

Lösel, F. (2008). Doing evaluation research in criminology: Balancing scientific and practical demands. In King, R., & Wincup, E. (Hrsg.) *Doing research on crime and justice* (141-162). New York: Oxford University Press.

Ludwig, J., Kling, J. R., & Mullainathan, S. (2011). *Mechanism experiments and policy evaluations* (No. w17062). National Bureau of Economic Research. Verfügbar unter: http://www.nber.org/papers/w17062 (Letzter Zugriff am 7.3.2016):

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86-92.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, *16*(1), S1-S23.

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.

Marshall, W. L., & Marshall, L. E. (2010). Can treatment be effective with sexual offenders or does it do harm? A response to Hanson (2010) and Rice (2010). *Sexual Offender Treatment*, *5* (2).

Martinson, R. (1974). What works? – Questions and answers about prison reform. *Public Interest*, 35, 22–54.

Maruna, S. (2015). Qualitative Research, Theory Development, and Evidence-Based Corrections: Can Success Stories Be "Evidence"?. In Miller, J. & Palacios, W.R. (Hrsg.) *Qualitative Research in Criminology* (311-337). New Brunswick: Transaction Publishers.

McCord, J. (2003). Cures that harm: Unanticipated outcomes of crime prevention programs. *The Annals of the American Academy of Political and Social Science*, *587*(1), 16-30.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., & Carroll, R. J. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics*, *5*(3), 445-464.

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference*. Cambridge University Press.

Niedersächsisches Justizministerium (2011). Soziatherapie im niedersächsischen Justizvollzug. Rahmenkonzept. Verfügbar unter: www.justizportal.niedersachsen.de/download/56054/zum_Downloaden.pdf (Letzter Zugriff am 7.3.2016).

Niemz, Susanne (2013). *Sozialtherapie im Strafvollzug 2013: Ergebnisübersicht zur Stichtagserhebung zum 31.3.2013*. Wiesbaden: KrimZ

Niemz, S. (2015). *Evaluation sozialtherapeutischer Behandlung im Justizvollzug*. Wiesbaden: KrimZ (Kriminologie und Praxis; Bd. 68).

Obergfell-Fuchs, J. & Wulf, R. (2008). Evaluation des Strafvollzugs. *Forum Strafvollzug 57* (5), 231-236.

Obergfell-Fuchs, J., & Wulf, R. (2011). Methodische Folgerungen für die Evaluation des Jugendstrafvollzugs. Aus der Evaluation von Projekt Chance. In: *Gewaltdelinquenz – lange Freiheitsentziehung – Delinquenzverläufe (*273-287*). Mönchengladbach: Forum Verlag Godesberg.*

Olver, M. E., Stockdale, K. C., & Wormith, J. S. (2011). A meta-analysis of predictors of offender treatment attrition and its relationship to recidivism. *Journal of consulting and clinical psychology*, *79*(1), 6.

Ortmann, R. (2002). *Sozialtherapie im Strafvollzug. Eine experimentelle Längsschnittstudie zu den Wirkungen von Strafvollzugsmaßnahmen auf Legal- und Sozialbewährung.* Freiburg: edition iuscrim.

Rehn, G. (2012). Sozialtherapie im Justizvollzug – eine kritische Bilanz. In Wischka, B., Pecher, W. & van den Boogaart, H. (Hrsg.), *Behandlung von Straftätern. Sozialtherapie, Maßregelvollzug, Sicherungsverwahrung* (32-80). Herbolzheim: Centaurus.

Rehn, G. (2014). Was tun? Zur Gegenwart und Zukunft der Sozialtherapie. *Forum Strafvollzug, 63* (4), 244-248.

Rice, M. E. and Harris, G. T. (2013) Treatment for Adult Sex Offenders. May We Reject the Null Hypothesis? In: Harrison, K. & Rainey, B. (Hrsg.). *The Wiley-Blackwell Handbook of Legal and Ethical Aspects of Sex Offender Treatment and Management* (219-235). Chichester: John Wiley & Sons, Ltd.

Rosenbaum, P. R. (2007). Interference Between Units in Randomized Experiments. *Journal of the American Statistical Association*, *102* (477), 191-200.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York.

Rossi, P. H. (1987). The iron law of evaluation and other metallic rules. *Research in social problems and public policy*, *4*(1987), 3-20.

Rossi, P. H. (2003). The "Iron Law of Evaluation" Reconsidered. Verfügbar unter: http://www.welfareacademy.org/rossi/Rossi_Remarks_Iron_Law_Reconsidered.pdf (Letzter Zugriff am 7.3.2016).

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2003). *Evaluation: A systematic approach*. London: Sage.

Sampson, R. J. (2010). Gold standard myths: Observations on the experimental turn in quantitative criminology. *Journal of Quantitative Criminology*, *26*(4), 489-500.

Sampson, R. J., Winship, C., & Knight, C. (2013). Translating causal claims. *Criminology & Public Policy*, *12*(4), 587-616.

Schwedler, A. & Schmucker, M. (2012). Verlaufsmessung im sozialtherapeutischen Behandlungsvollzug – Wie sinnvoll sind allgemeine Persönlichkeitsmaße. *Monatsschrift für Kriminologie & Strafrechtsreform 95* (4), 269-280.

Seifert, S. (2014). *Der Umgang mit Sexualstraftätern: Bearbeitung eines sozialen Problems im Strafvollzug und Reflexion gesellschaftlicher Erwartungen*. Berlin: Springer.

Shadish, W.R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R. (2013). Propensity score analysis: promise, reality and irrational exuberance. *Journal of Experimental Criminology*, *9*(2), 129-144.

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual review of psychology*, *60*, 607-629.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological methods*, *16*(2), 179.

Sozialtherapeutische Anstalt Gelsenkirchen. (2006) *Behandlungskonzetion. Sozialtherapeutische Anstalt Gelsenkirchen.* Verfügbar unter: http://www.sotha-gelsenkirchen.nrw.de/Aufgaben/Behandlungskonzept/index.php (Letzter Zugriff am 7.2.2016).

Steiner, P. M., Cook, T. D., & Shadish, W. R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics*, *36*(2), 213-236.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22*(11), 1359 –1366.

Spöhr, M. (2009). *Sozialtherapie von Sexualstraftätern im Justizvollzug: Praxis und Evaluation*. Mönchengladbach: Forum Verlag Godesberg.

Suhling, S. (2008). Sozialtherapie im Jugendstrafvollzug: Prinzipien wirksamer Behandlung. *Zeitschrift für Jugendkriminalrecht und Jugendhilfe*, *19*, 330-335.

Suhling, S. (2012). Evaluation der Straftäterbehandlung und der Sozialtherapie im Strafvollzug. (162-232). In: Wischka, B., Pecher, W. & van den Boogaart, H. (Hrsg.). *Behandlung von Straftätern: Sozialtherapie, Maßregelvollzug, Sicherungsverwahrung* (162-232). Freiburg: Centaurus.

Suhling, S. & Keßler, S. (2015). Sozialtherapeutische Behandlung im Jugendstrafvollzug: ein Überblick. In DVJJ (Hrsg.), *Jugend ohne Rettungsschirm. Herausforderungen annehmen!* (S. 597-622). Mönchengladbach: Forum Verlag Godesberg.

Suhling, S., & Guéridon, M. (2016). Vergleiche zwischen sozialtherapeutischen Einrichtungen. *Forensische Psychiatrie, Psychologie, Kriminologie*, *10*(1), 32-44.

Schmidt, A. F., Banse, R., & Imhoff, R. (2015). Indirect Measures in Forensic Contexts. In T. M. Ortner & F. J. R. van de Vijver (Hrsg.). *Behavior-Based Assessment: Going Beyond Self-Report in the Personality, Affective, Motivation, and Social Domains* (173-194). Göttingen: Hogrefe.

Schmidt, A. F., Gykiere, K., Vanhoeck, K., Mann, R. E., & Banse, R. (2014). Direct and indirect measures of sexual maturity preferences differentiate subtypes of child sexual abusers. *Sexual abuse: a journal of research and treatment*, *26*(2), 107-128.

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.

VanderWeele, T. J., & Hernan, M. A. (2013). Causal inference under multiple versions of treatment. *Journal of causal inference*, *1*(1), 1-20.

VanderWeele, T. J., Tchetgen, E. J. T., & Halloran, M. E. (2014). Interference and sensitivity analysis. *Statistical science*, *29*(4), 687.

Wakeling, H. & Travers, R. (2010) Evaluating offending behaviour programmes in prison. In Brown, J. M., & Campbell, E. A. (Hrsg.) *The Cambridge handbook of forensic psychology* (820-829). Cambridge: Cambridge University Press.

Weisburd, D. (2010). Justifying the use of non-experimental methods and disqualifying the use of randomized controlled trials: challenging folklore in evaluation research in crime and justice. *Journal of Experimental Criminology*, *6*(2), 209-227.

Weisburd, D., & Hinkle, J. C. (2012). The importance of randomized experiments in evaluating crime prevention. In Welsh, B.C. & Farrington, D.P. (Hrsg.)*The Oxford handbook of crime prevention* (446-465). New York: Oxford University Press.

Wischka, B. (2014). *Entwicklung und Evaluation eines Behandlungsprogramms für Sexualstraftäter (BPS) im Kontext integrativer Sozialtherapie.* Hildesheim: Universität Hildesheim.

Wischka, B. & Specht, F. (2001). Integrative Sozialtherapie – Mindestanforderungen, Indikation und Wirkfaktoren. In G. Rehn, B. Wischka, F. Lösel & M. Walter (Hrsg.), *Behandlung „gefährlicher" Straftäter. Grundlagen, Konzepte, Ergebnisse* (249-263). Herbolzheim: Centaurus.

Wolter, D. & Häufle, J. (2014). Wie aussagekräftig sind Gefangenenpersonalakten als Entscheidungshilfe im Strafvollzug? *Monatsschrift für Kriminologie und Strafrechtsreform 97* (4), 280-293.

Wößner, G. (2013). Zielsetzung und Anlage der Untersuchung. In: Wößner, G., Hefendehl, R. & Albrecht, H.-J. (Hrsg.) *Sexuelle Gewalt und Sozialtherapie*. Berlin: Duncker & Humblot.

Wößner, G. (2014). Wie kann man in der Sozialtherapie Therapieerfolg feststellen oder messen? *Forensische Psychiatrie, Psychologie, Kriminologie*, *8*(1), 49-58.

Wößner, G., Wienhausen-Knezevic, E. & Gauder, K.-S. (2014). *„Ich wurde halt einfach ins kalte Wasser geworfen…" Der Übergang in Freiheit und seine Herausforderungen aus der Perspektive entlassener Strafgefangener.* Verfügbar unter: http://www.bvaj.de/docs/Perspektive_Entlassener.pdf (Letzter Zugriff am 7.3.2016).

Korrespondenzadresse:

Marcel Guéridon
Kriminologischer Dienst im Bildungsinstitut des nds. Justizvollzuges
Fuhsestraße 30
29221 Celle
marcel.gueridon@justiz.niedersachsen.de