Article

Sexual Abuse
2020, Vol. 32(4) 452–475


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1079063219871576
journals.sagepub.com/home/sax


On the Effectiveness of Sexual Offender Treatment in Prisons: A Comparison of Two Different Evaluation Designs in Routine Practice

Friedrich Lösel[1,2], Eva Link[1], Martin Schmucker[1], Doris Bender[1], Maike Breuer[3], Lena Carl[1], Johann Endres[3], and Lora Lauchs[1]

Abstract

Although there is less continuity of sexual offending in the life course than stereotypes suggest, treatment should lead to a further reduction of reoffending. Contrary to this aim, a recent large British study using propensity score matching (PSM) showed some negative effects of the core sex offender treatment program (SOTP) in prisons. International meta-analyses on the effects of sex offender treatment revealed that there is considerable variety in the results, and methodological aspects and the context play a significant role. Therefore, this study compared different designs in the evaluation of sex offender treatment in German prisons. PSM was compared with an exact matching (EM) by the Static-99 in a sample of 693 sex offenders from Bavarian prisons. Most results were similar for both methods and not significant due to low base rates. There was a treatment effect at $p < .05$ on general recidivism in the EM and at $p = .06$ on serious reoffending in the PSM. For sexual recidivism, EM showed a negative trend, whereas PSM suggested the opposite. Overall, the study underlines the need for more replications of evaluations of routine practice, methodological comparisons, sensitive outcome criteria, and differentiated policy information.

Keywords

sex offender treatment, prison, evaluation design, propensity score matching

[1]Institute of Psychology, Friedrich-Alexander University Erlangen-Nürnberg, Germany
[2]Institute of Criminology, University of Cambridge, UK
[3]Criminological Research Unit of the Bavarian Ministry of Justice, Erlangen, Germany

Corresponding Author:
Eva Link, Institute of Psychology, Friedrich-Alexander University Erlangen-Nürnberg, Nägelsbachstr, 49c, D-91052 Erlangen, Germany.
Email: eva.link@fau.de

Introduction

This article is part of a special issue on developmental and life course research on sexual offenders. Various international studies clearly showed that the typical sexual reoffending rates are rather low at about 10% or less (e.g., Hanson & Morton-Bourgon, 2009; Jennings, 2015; McCann & Lussier, 2008; Nisbet, Wilson, & Smallbone, 2004; see also the articles in this issue). This is much less than in other fields of crime and violence, but nonsexual reoffending of sexual offenders is also more frequent. Of course, the official data on sexual reoffending underestimate actual rates due to undetected offenses, but this is a general problem of the dark figure in criminological research. In spite of relatively low recidivism figures, sexual offending remains a highly relevant topic for crime policy, the mass media, and general public. As a consequence, the empirical message of not much continuity of sexual offending in the life course is an important scientific finding. However, the general population, victim organizations, and justice ministers will not fully be satisfied by this message. Although science is committed to a rational and realistic approach, it cannot ignore the understandable view that every single case is one too many. Accordingly, crime policy in Western countries aims for a further reduction of the risk of reoffending, although low base rates may lead to a statistical floor effect that limits the potential impact of any kind of treatment, particularly when small sample sizes do not yield enough statistical power for detecting a significant treatment effect.

Against this background, the organizers of the 2018 conference at Brussels have asked the first author of this article to give an overview of the current state on the effects of treatment. His presentation summarized that there has been international progress in research and practice of sex offender treatment, but the question of "what works" for sex offenders is still discussed controversially (e.g., Ho, 2015; Ho & Ross, 2012; Koehler & Lösel, 2015; Mann, Carter, & Wakeling, 2012; Marshall & Marshall, 2010; Rice & Harris, 2003; Seto et al., 2008). The discussion is complicated by many factors that may have an impact on empirical findings of evaluations, for example, different types of offenses, offender groups, comorbidities, treatment contents, implementation quality, evaluation designs, outcome criteria, legal regulations, and institutional contexts (e.g., Lösel & Schmucker, 2017; Seto, 2018). Although meta-analyses showed overall positive treatment effects, there is no clear evidence for successful treatment in *prisons* (Hanson, Bourgon, Helmus, & Hodgson, 2009; Lösel & Schmucker, 2005; Schmucker & Lösel, 2015). There is a large variation between different primary studies, and due to the abovementioned reasons, many did not reveal a significant effect. In other fields of violent offender treatment, the evidence seems to be more robust (e.g., Lösel, 2012; Wilson, 2017).

As Lösel and Schmucker (2017) and Schmucker and Lösel (2015, 2017) already presented overviews of the international findings on sex offender treatment, we decided to address a more specific issue in this article. In particular, we deal with the political context of sex offender treatment evaluation and the potential impact of methodology on the results. This approach is triggered by recent experiences in England and Wales, where practice of sexual offender treatment has been at the international forefront.

The Recent British Experience

The political relevance of sex offender treatment became recently evident when a research team of the British Ministry of Justice (MoJ) reported negative results on the core sex offender treatment program (SOTP) in England and Wales (Mews, Di Bella, & Purver, 2017). As this finding was the trigger for the empirical part of the present article, we briefly describe the British study and its political consequences.

The group format of the core SOTP in England and Wales has been developed by international leaders in the field (Mann & Thornton, 1998). The program based on sound analyses of the literature and was accredited by the then Correctional Services Accreditation Panel (CSAP; now Correctional Services Accreditation and Advice Panel [CSAAP]). The widespread implementation of the program was supported by international knowledge on the appropriateness of structured cognitive–behavioral treatment for sexual offenders. There was no solid evidence base in Britain, but a study of Friendship, Mann, and Beech (2003) suggested some positive effects. This study contained a matched comparison between treated and untreated sexual offenders that used the risk bands of the Static-99 actuarial risk assessment for stratification. Perhaps due to the relatively low prevalence of sexual recidivism and not very large sample sizes, the evaluation revealed no significant effect on sexual reoffending but a tendency in this direction. Only for the combined outcome of sexual and/or other violent reoffending, there were significant treatment effects. Although this study had weaknesses, such as controlling for only a few static risk factors and a rather short follow-up-period of 2 years, it contained at least some findings that seemed to support the use of the SOTP.

The cold shower came in 2017. Because no randomized controlled trial (RCT) has been implemented in Britain, researchers from the Analytical Services Department of the MoJ carried out a large quasi-experimental study on the SOTP using methods of propensity score matching (PSM). The study contained 2,562 convicted sex offenders who started SOTP in prison between 2000 and 2012 in England and Wales. This group was compared with 13,219 untreated sex offenders using 87 matching factors for PSM. The main outcomes were sobering: The rate of overall sexual reoffending was larger in treated sex offenders (10.0%) than for the control group (8.0%). Child image reoffending was also not in favor of treatment (treatment group [TG] = 4.4%, comparison group [CG] = 2.9%). Due to the large sample sizes, these differences were highly significant, whereas in other sexual offense outcomes, there were no significant effects. Because of the alerting message of these findings, the Justice Ministry appointed an expert panel that reviewed several drafts of the study before publication. This group consisted of experts in statistics and evaluation from Britain and Karl Hanson and Friedrich Lösel from the field of sex offender treatment research. The panel emphasized that an RCT would have been preferable, but an RCT on sex offender treatment had not been carried out for legal, practical, financial, ethical, and other reasons. After intensive discussions of draft reports of the quasi-experimental PSM study, the panel agreed upon methodological limitations, for example, sexual deviance could not be included as a matching factor due to lacking information, and some matching criteria were not theoretically relevant.

PSM is a widely used matching strategy, but embraces different methodical approaches and is discussed controversially (Jann, 2017; King & Nielsen, 2016). For example, PSM can be biased when theoretically and content-relevant covariates are neglected (Luellen, Shadish, & Clark, 2005). Although the panel considered these and other problems, it felt that the Mews et al. (2017) study had merits in comparison with many weaker evaluations and should be published. It took a while until the release of the study, but the delivery of the core SOTP was terminated when the first findings were known in the MoJ. British mass media got information about the undesirable findings and blamed the Ministry of wasting about £100 million for sex offender treatment that did not work, but even increased the risk of reoffending. The study was quickly published after this media campaign. Since then, the government produced a report on offending behavior programs and established a working group to promote empirical program evaluation (what was always emphasized by CSAP/CSAAP).

This case report shows that the issue of the effectiveness of sexual offender treatment goes far beyond academic discussions. It also raises the question of the relevance and validity of one single study. The "replication crisis" is currently a hot topic in psychology, medicine, criminology, and other sciences (Baker, 2016; Lösel, 2018; Open Science Collaboration, 2015). Although one should be cautious with using the term "crisis", it is rightly emphasized that singular studies may not tell "the truth" and replications are extremely important. With regard to sexual offender treatment, the finding of Mews et al. (2017) was not a totally unexpected surprise. Meta-analyses had shown that sex offender treatment in prisons did not reveal a mean significant effect (Lösel & Schmucker, 2005; Schmucker & Lösel, 2015). Some single studies on sex offender treatment pointed in the same direction, but found positive effects on delayed or less harmful sexual recidivism (e.g., Olver, Nicholaichuk, Gu, & Wong, 2012). Meta-analyses did also not support a pure group treatment format. Although Ware, Mann, and Wakeling (2009) provided sound reasons for a group format, they stated a lack of clinical research on this issue. Meta-analytic results on sex offender treatment showed no significant results of group-only programs and the inclusion of individual sessions was more effective (Schmucker & Lösel, 2015). Issues of privacy and other clinical considerations also support at least some one-to-one sessions or more individualized treatment modules.

Although meta-analyses were partially in line with the zero-effect findings of Mews et al. (2017), they deviated from the British *negative* results. Some sound international studies had shown a failure of sex offender treatment (Marques, Wiederanders, Day, Nelson, & van Ommeren, 2005), but these were older studies and they had not shown significantly negative effects. Although negative treatment effects are not frequent in criminology, the Mews et al. (2017) finding should remind us to the fact that psychosocial interventions can be harmful in spite of best intensions (McCord, 2003). However, we must also ask whether one single study with a controversially discussed methodological design should lead to far-reaching conclusions on the practical and political levels.

Evaluation Methods and Effect Size Variability

Some of our meta-analytic findings (Lösel & Schmucker, 2005) suggested that not only treatment participants' and setting characteristics but also methodological features had an impact on the variation in outcomes of sex offender treatment. Methodological characteristics such as the sample size, the source of recidivism data and the recidivism base rate, or the definition of the treatment and control groups (e.g., handling of refusers and dropouts) explained part of the effect size variance. In a broad systematic review of criminological interventions, Weisburd, Lum, and Petrosino (2001) showed that more internally valid evaluation designs (i.e., RCTs) revealed smaller effects than weaker designs. However, with regard to sexual and other offender treatment, the relation between design quality and effect size is less clear and not linear (Lipsey & Cullen, 2007; Lösel, 2012; Schmucker & Lösel, 2015).

Although RCTs are seen as the gold standard in program evaluation, in sexual offender treatment, randomized designs often cannot be realized due to legal, practical, ethical, and other reasons (Marshall & Marshall, 2007). RCTs are particularly difficult to implement in evaluations of routine practice, although these are more relevant for policy than often better controlled demonstration or model projects (Lipsey, 2018). Quasi-experimental research is rather common in evaluations of sex offender treatment despite its limitations with regard to causal inference (Schmucker & Lösel, 2015). To alleviate these limitations, matching procedures aiming for equivalent treatment and control groups even after incidental assignment have been advocated (Stuart, 2010). Originally, the most common matching method was an exact pairwise matching on selected variables that were relevant with regard to recidivism. It is a major problem of the variable-oriented exact matching (EM) procedures that increasing the number of matching variables inevitably results in losing cases because there is no match for many cases (or choosing the "nearest neighbor" may also reduce the equivalence between TG and CG). To elude this problem, only a single or a few particularly important variables may be used. In evaluations of sexual offender treatment, the Static-99 is often used as a single matching variable because it can be easily coded from file information and, despite its simplicity, is a relatively valid measure for future recidivism. With PSM, Rosenbaum and Rubin (1983) proposed a method that allowed matching treatment and control group members on a large number of variables via a propensity score, thereby avoiding the major problems of single variable matching. PSM has become the method of choice in various recent evaluations of sexual offender treatment (e.g., Duwe & Goldman, 2009; Grady, Edwards, & Pettus-Davis, 2017; Mews et al., 2017), whereas in earlier studies, matching typically referred to a single or a few selected variables (e.g., Friendship et al., 2003; Procter, 1996; Rice, Harris, & Quinsey, 1991). This shift from single variable to PSM triggered our study.

Our Study

We will analyze data on sex offender treatment in social–therapeutic prisons in Bavaria (Germany). A previous evaluation used the same Static-99 stratification design as

Friendship et al. (2003) in Britain. In the German study, there were by trend desirable results among low-risk offenders, but negative effects for high-risk cases and on overall sexual recidivism (Breuer & Endres, 2016). This analysis differed not only from Friendship et al. (2003) but also from international research on an inverse relation between risk level and effect size (Schmucker & Lösel, 2015). The latter is plausible when one takes the higher base rates in more risky offenders into account.

Our main research question in the present article is whether two widely used evaluation designs lead to similar or different results. In addition, we aim to provide data on sexual offender treatment in a country that is underrepresented in international research. On one hand, we used an EM procedure based on a single control variable, the Static-99 risk score (i.e., de facto a sum of a few objective items). On the other hand, we carried out an analysis by using PSM, which allows to simultaneously balance a large number of confounding variables. These two approaches reflect the historical development of matching methods in sexual offender treatment as shown above. Even today, controlling for only a few relevant confounders is still common in criminological evaluation studies, because researchers often have to rely on limited retrospective data on the offenders.

To our knowledge, this is the first study that explicitly compares different matching procedures on the same data set of sexual offender treatment evaluation.

Method

Sample

Our data stem from adult male sex offenders with a prison sentence of at least 2 years who had been released from Bavarian prisons between 2004 and 2008. For this particular group, biographical, criminological, and other data had been gathered by the prison staff and forwarded to the Criminological Research Unit (CRU) of the Bavarian MoJ. Data on recidivism were available for $n = 829$ offenders of whom 136 had to be excluded from further analyses for different reasons: 18 persons had died in the meanwhile, 114 had been released into another country, and four had severe health conditions. Accordingly, the final sample contained $n = 693$ sex offenders.

The offenders had been released from 15 prisons in Bavaria. The mean age at the time of release was 43.36 years ($SD = 11.98$). Approximately half of the sample (52.7%, $n = 365$) had received a specific treatment for sex offenders during their imprisonment, whereas 47.3% ($n = 328$) were in regular prisons and may have only received nonspecific interventions such as drug counseling. The vast majority of the treated participants ($n = 231$, 63.3%) participated in the programs of social–therapeutic prisons. These institutions aim for an overall therapeutic climate and deliver individual and group therapy as well as basic education, professional training, and a stepwise opening with work outside, near the end of the sentence (for more information, see Carl, Schmucker, & Lösel, 2018; Lösel & Egg, 1997). In principle, the social–therapeutic units in Germany are more similar to hierarchic than democratic therapeutic communities in the Anglo-American world (e.g., Lipton, 2010; Shuker & Sullivan,

2010). In addition to the inmates from social–therapeutic units, we had data on offenders from other institutions who participated in programs for sexual offenders. About one quarter ($n = 94$, 25.8%) received group treatment and 11% ($n = 40$) individual therapy.

Because our focus is on the evaluation method and not primarily on the treatment content, we included all three treatment approaches in our analysis to enlarge the sample size in the present study. The integrative analysis was also justified because some offenders participated in more than one approach, and there were no significant differences between the three subgroups in recidivism when the risk level was controlled for. There were differences in recidivism between some social–therapeutic prisons, but these were also minimal after controlling for risk in the Static-99 (Carl, Lauchs, Schmucker, & Lösel, 2019). The mean duration of social therapy was $M = 25.36$ ($SD = 10.36$) months; for individual therapy, $M = 12.33$ ($SD = 6.44$) months; and for group therapy, $M = 15.90$ ($SD = 8.06$) months. Drop-out rates were 19.2% ($n = 44$) for social therapy, 17.6% ($n = 16$) for group therapy, and 20.0% ($n = 8$) for individual therapy. Because the social–therapeutic facilities use the first 3 months of treatment as a trial period, participants who had dropped out during this phase were considered untreated, but in line with an intent-to-treat analysis, later dropouts were included in our analysis as treated.

According to the Bavarian Prison Act (Article 11), all sexual offenders with a prison sentence of more than 2 years ought to be transferred to social–therapeutic facilities. Thus, all participants in our sample had the chance of receiving treatment. However, due to restricted treatment capacities, selection processes occurred and almost a half did not receive any kind of specific sex offender treatment during incarceration. The vast majority (80%) of the untreated offenders were not sufficiently motivated or refused to participate in treatment. Others did not meet therapeutic requirements due to language deficits, cognitive impairment, or other reasons (12%). In a few cases, the duration of the remaining sentence was too short when they were ready for therapy. For 31%, multiple reasons for nonparticipation were stated. Even though volunteerism might not affect treatment outcome (Grady, Edwards, Pettus-Davis, & Abramson, 2013), it is likely that treated and untreated offenders differed in various aspects. The incidental assignment in the present study is a typical example for a quasi-experimental design in which matching procedures are implemented in the evaluation of routine practice to increase comparability.

Instruments and Data Collection

Sex offender documentation system. Data about the offenders were collected before or at the time of their release from prison. A comprehensive questionnaire of the CRU of the Bavarian MoJ contained 72 variables that were rated by prison staff (mostly psychologists). These variables included biographical data (e.g., age at release, family background, relationship problems), offense-related variables (e.g., type of actual and prior sex offenses, number and characteristics of the victims, denial of the offense), other criminological variables (e.g., number and type of previous convictions, juvenile

delinquency, placement in shelters), psychiatric and clinical data (e.g., psychiatric diagnoses, substance abuse, psychopathy), items on conduct in prison (e.g., type and duration of treatment participation, behavior during imprisonment, ratings of therapeutic success), and variables about expected living conditions after release (family support, social integration, accommodation).

The items were mainly taken from meta-analyses on risk factors for reoffending (Hanson & Morton-Bourgon, 2005) and structured risk assessment instruments (Static-99, Sexual Violence Risk-20 [SVR-20], Historical-Clinical-Risk Management-20 Violence Risk Assessment Scheme [HCR-20], Rückfallrisiko für Sexualstraftäter [RRS]) but also included additional items that are relevant for risk assessment and treatment evaluation (e.g., release to a foreign country, further treatment after release). A detailed manual contained coding rules for all items. Whereas for treated offenders, information could be gained during the therapeutic process, ratings for untreated offenders were mainly based on information from their prison files. A study on interrater agreement has been conducted for a subsample of 80 offenders (Haas, 2014). The reliability indices varied between variables and domains. The kappa coefficients for categorical data varied between $\kappa = .08$ and .95. The intraclass coefficients (ICCs) for quantitative variables ranged between ICC = .50 and .99. The reliability coefficients were higher for criminological variables such as age at release from prison, previous sex offenses, type of the index offense, and victim characteristics. They were lower for psychological variables such as social integration, work–life problems, romantic relationships, planning, and minimization of the index offense.

Overall, the variables with lower interrater agreements relate to characteristics that are not routinely documented in prison files. Because the analysis of Haas (2014) compared the routine ratings by prison staff (mostly psychologists) with ratings that were based on file information, the resulting coefficients should be considered with caution. It can be reasonably assumed that staff ratings are based on knowledge about the offenders that exceeds the information included in prison files. Thus, differences in the amount of available information might have affected the results on reliability. As a consequence, we recommend to assess a reduced number of variables that can most reliably be assessed in routine practice. This is similar to the experience with the originally very comprehensive Offender Assessment System (OASys) in the United Kingdom.

Actuarial risk assessment. The Static-99 (Hanson & Thornton, 2000) was used for our single variable matching approach. In 2012, the Static-99 had been revised, changing the coding rules for age at release to better account for its relationship with sexual recidivism (Helmus, Thornton, Hanson, & Babchishin, 2012). However, cultural population differences seem to be relevant. For example, a study in Austria showed that the previous version of the Static-99 performed better in predicting sexual recidivism of prisoners than the age-corrected version (Rettenberger, Haubner-MacLean, & Eher, 2013). In our study, the revised coding rules resulted in almost half of the offenders being in the low-risk group. Therefore, we used the original version of the Static-99. Due to some missing data, Static-99 scores could only be calculated for 670 of the 693 offenders.

Table 1. Number of Offenders in Each Risk Category by Treatment Group.

	Treatment group ($n = 354$)		Comparison group ($n = 316$)	
Static-99 risk category	*n*	%	*n*	%
Low (scores 0-1)	142	40.1	101	32.0
Medium–low (2-3)	120	33.9	110	34.8
Medium–high (4-5)	60	16.9	63	19.9
High (6 or more)	32	9.0	42	13.3

Recidivism data. Information on reconvictions after release was obtained from the German Federal Central Crime Register. The follow-up period for sex offenders in our sample ranged from about 3 to 8 years ($M = 5.74$, $SD = 1.47$). Due to a varying time delay between the date of an offense, the conviction, and the official registration, the exact time at risk might be somewhat shorter. The criminal records were coded according to four different recidivism criteria as outcome measures. *General recidivism* was defined as any conviction for a new offense after release. *Sexual recidivism* was defined as any new sex offense (e.g., rape, sexual assault, sexual child abuse, indecent images). *Violent recidivism* contained any new violent offense excluding sexual offenses (e.g., homicide, assault, robbery). In addition, we coded *severe recidivism* in cases for any new convictions of at least 2 years of imprisonment, custody in a forensic clinic, or preventive detention. The respective recidivism rates in the whole sample were 41.1% for general recidivism, 6.9% for sexual recidivism, 13.7% for violent recidivism, and 8.1% for severe recidivism.

Matching Approaches

Single Variable Matching by the Static-99 Sum Score

In our first approach, we matched participants in the CG to the participants in the TG using the Static-99 risk score as a single matching variable (although it contains several items). Static-99 scores were available for $n = 354$ treated and $n = 316$ untreated offenders. Table 1 shows the number of offenders in the four risk bands for the yet unmatched sample. The untreated offenders were underrepresented in the lowest risk band and overrepresented in the two high-risk bands.

Accordingly, the mean total Static-99 sum score was significantly higher in the CG ($M = 2.80$, $SD = 2.09$) than in the TG ($M = 2.43$, $SD = 2.03$); $t(668) = 2.33$, $p < .05$. The matching was based on the sum scores ranging from 0 to 9. Because the number of individuals in our treatment group exceeded the number of untreated participants, we chose weighting over a 1:1 matching procedure to preserve the sample size (for an overview on matching methods, see Stuart, 2010). To equate Static-99 scores in CG and TG, weights were assigned to individuals in the CG. This also results in a fully matched sample regarding the matching variable but avoids dropping cases when there

are not enough matches and having to choose individual cases over others when there are more matches available, respectively. Nonetheless, we had to exclude one treated participant with a score of 9 because none of the untreated offenders had an equivalent score. The weights were calculated separately for controls within each risk score group. Each case in the CG received a frequency weight reflecting the number of treated individuals with the same Static-99 score divided by the number of untreated individuals with the same risk score. For example, in the initial sample, there were 23 individuals in the TG and 25 individuals in the CG with a Static-99 sum score of 5. Thus, each of the 25 individuals in the CG received a weight of 0.92 (23/25).

Individuals in the TG received a weight of 1. This resulted in a total number of 706 cases and an equal number in each group in the weighted sample. Due to the EM procedure, the Static-99 mean scores ($M = 2.41$, $SD = 2.00$) and the distributions of the scores were exactly equal in both groups. In the matched sample, the mean time at risk (years) was very similar for the treatment ($M = 5.78$, $SD = 1.48$) and the control groups ($M = 5.70$, $SD = 1.46$), $t(704) = –0.79$, $p = .43$. There were also no differences in follow-up times across offenders in the four strata of the Static-99, $F(3, 702) = 0.73$, $p = .53$.

PSM

Propensity score methods have gained much popularity in evaluation research as they allow to control for a large number of possible confounders. Originally proposed by Rosenbaum and Rubin (1983), the propensity score is defined as the probability of being assigned to a treatment group, given a certain set of pretreatment covariates. In randomized experiments, the true propensity score is predetermined by the study design, with each participant having a propensity score of 0.50, that is, a 50% chance of being in the treatment group (Luellen et al., 2005). In quasi-experiments, propensity scores can be calculated via logistic regression given a set of pretreatment characteristics (confounders). In the regression model, treatment assignment is the dependent variable (dummy coded 0, 1), and potential confounders are the predictors (Austin, 2011). For balancing the covariates in the TG and the CG, different techniques, such as 1:1 matching or weighting can be applied (Austin, 2011). Matching on the propensity score requires a large number of participants especially in the CG. Because our TG ($n = 365$) was larger than our CG ($n = 328$), propensity score weighting was applied instead of pairwise matching.

The propensity score model. The first step in specifying the propensity score model is the selection of relevant covariates. In general, relevant covariates are variables that affect the treatment assignment as well as the outcome. As there is a lack of empirical evidence concerning the treatment selection process, all outcome-related variables (potential confounders; see Austin, 2011) were considered as covariates in our model. Based on simulation studies and theoretical considerations, this approach is supported in the literature on PSM (Austin, 2011; Austin & Stuart, 2015; Brookhart et al., 2006; Stuart, 2010). Accordingly, all variables from risk assessment instruments (Static-99,

SVR-20, HCR-20) were included in the PSM model. In addition, the abovementioned sex offender questionnaire was screened for a further selection of empirically relevant risk variables (Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2005). Of the remaining variables, those that revealed a relationship at $\alpha \leq .20$ with at least one recidivism criteria were also selected as covariates. Four variables could not be included due to missing or unreliable information, especially in the CG (diagnosis of a personality disorder, diagnosis of paraphilia, having committed an offense under the influence of drugs, and psychopathy). Finally, a total number of 37 variables (see Table A1 in the Supplemental Material) were selected as covariates and ordinal and categorical variables were dummy coded for subsequent analyses. Only 69.3% of participants had complete covariate data. Thus, in a complete case analysis, one third of the original sample would have been lost for further analyses. As the overall proportion of missing data was only 2.3% for the variables in the PSM model, simple imputation of missing values was performed before propensity scores were calculated. Imputation via EM algorithm was conducted using the statistic software NORM 2.03 (Graham, 2012). Propensity scores were then calculated in SPSS via logistic regression as explained above. Participants with propensity scores outside the area of common support (i.e., the overlapping range of propensity scores in the treatment group and the CG) were excluded from subsequent analyses. Thus, weights were applied to a reduced sample of $N = 644$ participants (TG: $n = 352$, CG: $n = 292$), and balance diagnostics were inspected (Austin & Stuart, 2015).

Weighting procedure. Cases in the TG and the CG were weighted by the inverse probability of treatment weighting (IPTW; see Austin, 2011). When e_i denotes the propensity score for the ith participant, weights are defined as $w_i = 1/e_i$ for participants in the TG and as $w_i = 1/(1 - e_i)$ for participants in the CG. Referring to Xu et al. (2010), we calculated stabilized weights to preserve the original sample size.

Balance diagnostics. As shown in Figure 1, the distributions of propensity scores showed a substantial overlap between the TG and the CG in the original sample (area of common support). Nonetheless, they were different, indicating imbalance in the measured covariates. However, an approximation of distributions in the TG and the CG could be reached by IPTW. Besides the overall propensity score distribution, standardized differences for each covariate were calculated (Austin, Grootendorst, & Anderson, 2007) and are reported in Table A1 in the Supplemental Material. Higher absolute standardized differences indicate a lower balance between the treatment group and the CG for a specific covariate, and a threshold of 20% (Rosenbaum & Rubin, 1985) or 10% (Austin & Stuart, 2015) is usually applied. In the unweighted sample, 33 (59.9%) of the absolute standardized differences exceeded the value of 10% and 13 (22.4%) were larger than 20% ($M = 14.80$, $SD = 12.04$). In the weighted sample, balance has improved substantially with only two (3.4%) values exceeding 10% and no value larger than 20% ($M = 3.15$, $SD = 2.98$).

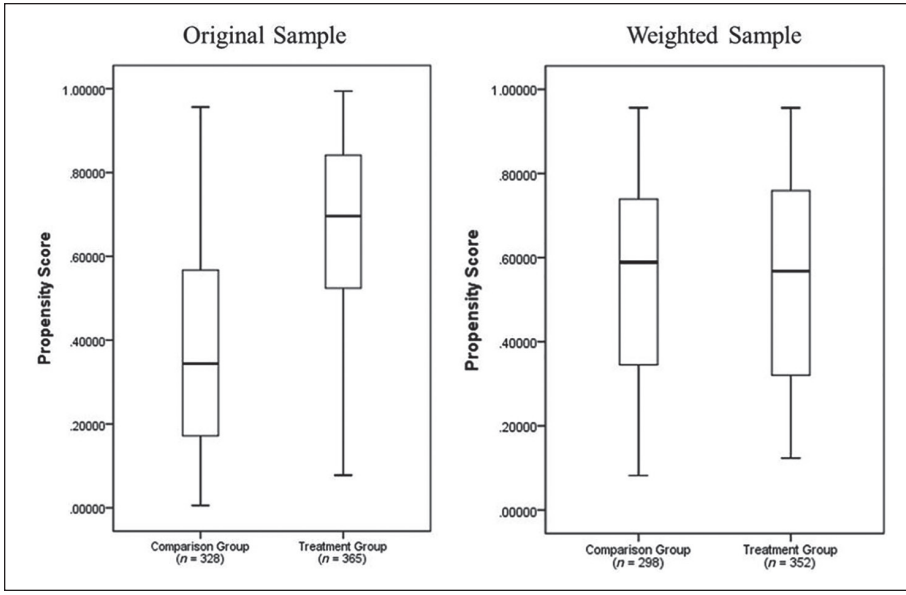

Figure 1. Propensity score distributions in the original sample and the weighted sample by treatment assignment.

Results on Recidivism

Single Variable Matching by the Static-99 Sum Score

Recidivism rates for the matched TG and CG are displayed in Figure 2. The pattern of findings was somewhat different for the various outcome criteria. Recidivism rates were slightly higher in the CG for general (odds ratio [OR] = 0.72), violent (OR = 0.84), and severe recidivism (OR = 0.76) with a significant result only for general recidivism, $\chi^2(1, n = 706) = 4.49$, $p < .05$. For sexual recidivism, the percentage of reoffenders was higher in the treatment group than in the CG (OR = 1.22), but the difference was not statistically significant, $\chi^2(1, n = 706) = 0.42$, $p = .52$.

PSM

The balance diagnostics indicated that all relevant confounders were sufficiently balanced in the TG and the CG after weighting. As shown in Figure 3, recidivism rates were slightly higher in the CG for all recidivism criteria (general recidivism: OR = 0.78, sexual recidivism: OR = 0.83, violent recidivism: OR = 0.83, and severe recidivism: OR = 0.60). No difference was statistically significant (all *p*s larger than .05), but for severe recidivism, there was a statistical trend, $\chi^2(1, N = 651) = 3.56$, $p = .06$, OR = 0.60.

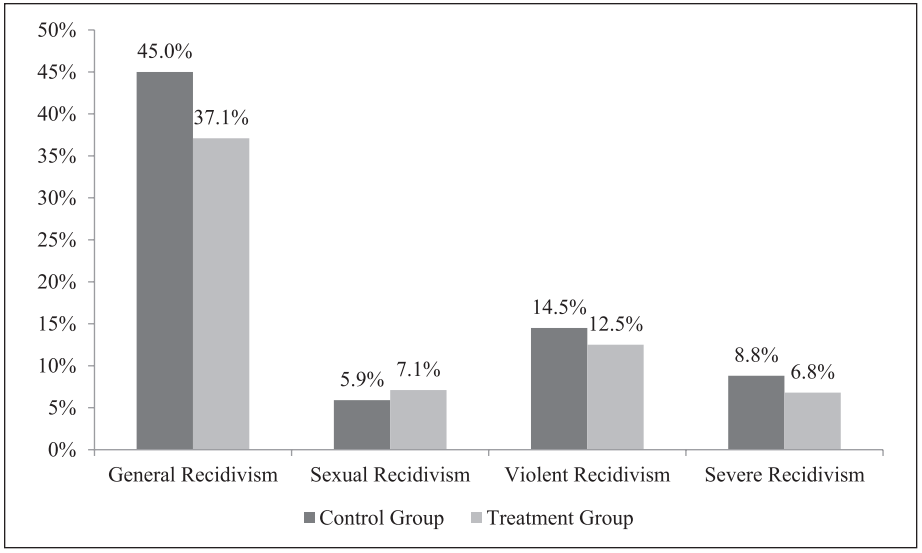

Figure 2. Recidivism rates in the treatment group and the control group matched on the Static-99 sum score.
Note. Sample sizes: $N_{control} = 353$, $N_{treatment} = 353$.

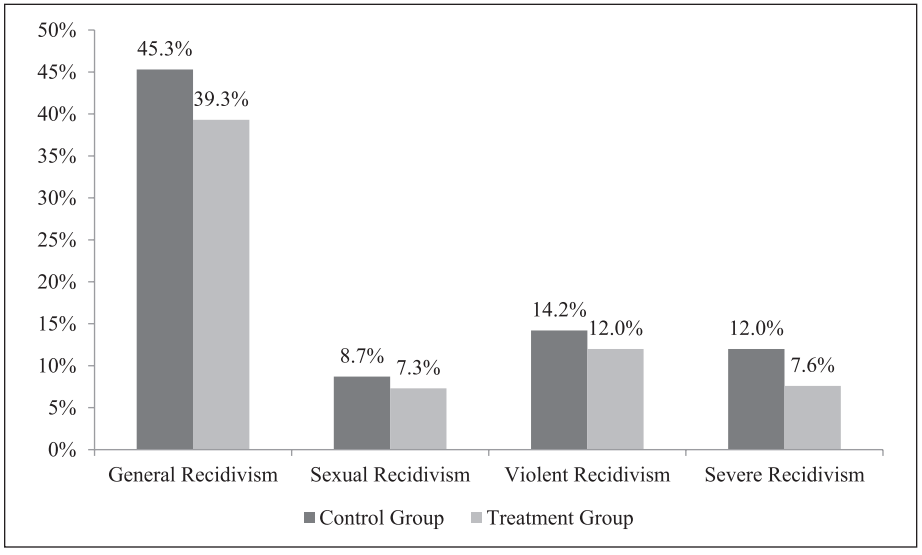

Figure 3. Recidivism rates in the treatment group and the control group matched via propensity score weighting.
Note. Sample sizes in the weighted sample: $N_{control} = 298$, $N_{treatment} = 352$.

Discussion

The aim of this study was a comparison of two widely used matching methods in the evaluation of sex offender treatment. This was triggered by a recent politically very influential study in Britain that used PSM as the best design under given practical circumstances (Mews et al., 2017). In the meanwhile, the U.K. MoJ decided to prioritize PSM as the most feasible and adequate design for evaluations of offending behavior programs. Within this context, we first discuss the specific findings in our study and then address broader issues of sex offender treatment, research, and policy.

Discussion of Our Findings

In our study, the more traditional risk-based matching and PSM showed mainly similar, but also partially different, results on recidivism. In both analyses, there were no significant effects of treatment in prison on the rate of sexual reoffending. This is in accordance with recent meta-analyses (Schmucker & Lösel, 2015, 2017) and several quasi-experimental primary studies (Abracen, Looman, Ferguson, Harkins, & Mailloux, 2011; Grady et al., 2017; Olver et al., 2012; Smallbone & McHugh, 2010; Smid, Kamphuis, Wever, & van Beek, 2016). Some of the latter studies had smaller samples and applied matching methods that may have impaired equivalence of TG and CG. However, there were also studies with relatively large samples and sophisticated PSM methods (e.g., Grady et al., 2017) that found no significant treatment effects on sexual reoffending. Some results were nonetheless encouraging as they showed less harmful or delayed sexual reoffending of treated offenders (Olver et al., 2012). The latter may lead to less repeat recidivism according to the age–crime curve. In our ongoing project, we are gathering such differentiated data and will apply a more complex metric harm index on a much larger sample for which we just received recidivism data.

Although there is widespread agreement among researchers about the need of more differentiated outcome measures, they need to acknowledge political reality. As the British evaluation of Mews et al. (2017) has shown, dichotomous sexual recidivism still plays a key role in public discussions and political decision making. This single study, that had various strengths but also some weaknesses, led to a "seismic event" in policy and practice. We cannot exclude the possibility of similar events in the future when the new programs will be thoroughly evaluated.

Our above findings did not reveal significant prison-based treatment effects on sexual reoffending, but the negative trend in the Static-99 matching analysis requires further consideration. The difference of 1.2 percentage points in favor of the CG was similar as in the larger British study of Mews et al. (2017). In our smaller study, the difference may have been due to chance, but it is not negligible when the low base rate in the CG is taken into account. Although, from a scientific perspective, this nonsignificant difference cannot be interpreted as a negative treatment effect, translated into percentages, it would suggest a difference of 20% that may alert policy and practice. However, the absolute numbers of recidivism in the TG and CG were only 25 versus 21. Accordingly, very few cases influenced the whole picture, what is similar in other

studies on sexual offender treatment. We must also assume a "floor effect" due to the low base rate. Researchers are familiar with these issues, but the recent experience in Britain suggests that we should inform policy makers, practice and broader audiences about these details. Leading associations such as the Association for the Treatment of Sexual Abusers (ATSA) and the National Organization for the Treatment of Abusers (NOTA) play an important role in this process. Perhaps plain language documents could help to avoid potential misinterpretations of small and nonsignificant differences between TGs and CGs and underline the need for large and sound experiments.

Our study also showed a somewhat different trend on sexual reoffending in the PSM analysis. In contrast to Mews et al. (2017), on a descriptive level, our PSM findings were slightly more favorable (1.6 percentage points, 18%) in the TG than in the CG. This small and nonsignificant difference should not trigger far-reaching speculation. However, the different trend in comparison with the Static-99 analysis suggests that the more sophisticated PSM approach does not necessarily lead to negative results. Therefore, we emphasize the urgent need of replications not only across different studies (Farrington et al., 2019; Lösel, 2018) but also with regard to different methods within one and the same study. Policy and practice should be more informed about the fact that evidence comes by replication, and single results need to be embedded in a broader framework.

Another important message from our study is the consistency between both analyses on the other criteria of recidivism. In the Static-99 analysis there was significantly less general recidivism (18%) in the TG, and the PSM analysis pointed in the same direction with a (nonsignificant) difference of 13%. Both analyses on violent recidivism showed also lower rates in the TG. The differences were 2.0 percentage points in the Static-99 analysis and 2.2 in the PSM analysis. With regard to the criterion of severe recidivism, there was also consistency between the two analyses. PSM revealed a nearly significant difference at $p = .06$ (two-sided test) with a difference of 4.4 percentage points (37%). The Static-99 matching showed a nonsignificant difference of 2.0 percentage points (23%) in favor of the TG.

Of course, we do not wish to over-interpret nonsignificant trends, but evaluations of sex offender treatment should be realistic about small absolute differences and low statistical power in clinical studies. Over decades, thresholds of statistical significance have been discussed controversially (e.g., Morrison & Henkel, 1972; Savitz, 1993). Basically, statistical significance is most appropriate in testing a specific theoretical hypothesis, and in these cases, a one-sided approach is often appropriate. However, in applied fields such as sex offender treatment, significance testing became a rarely reflected routine that did not always fit to realistic outcome expectations and questions of practical significance (e.g., with regard to small samples and small to moderate effect sizes). Therefore, more homogeneous results of meta-analytic integrations are very valuable.

Limitations

Comparing our results with other studies is limited by differences such as treatment parameters (e.g., the intensity and duration of treatment), characteristics of the sample

(e.g., the severity of the index–offense), the source of recidivism data, or the length of the follow-up time. Furthermore, methodological limitations of our study have to be taken into account when interpreting the results. One limitation relates to the variables that we could use for the PSM. These were assessed by practitioners in the daily practice in prisons. As a reliability study on interrater agreement on the basis of prison files showed, not all of these variables seem to be sufficiently reliable (and as a consequence more valid; Haas 2014). We do not exactly know how individual expertise may have influenced assessments beyond the respective instructions. This is a general matter of evaluations of routine practice (in contrast to closely monitored model projects), but there is space for improvement by principles of implementation science (e.g., Fixsen, Blase, Naoom, & Wallace, 2009). Another limitation is our inclusion of three types of treatment to increase sample size. These approaches varied in content and intensity, but there was also overlap and their recidivism outcomes were similar when the risk level was controlled. Perhaps, these findings indicate that beyond the specific method of treatment, the therapeutic alliance and other interpersonal factors are highly relevant (a well-known finding in other fields of psychotherapy; for example, Orlinsky, Grawe, & Parks, 1994). This may also explain why the different social–therapeutic prisons varied in the use of specific elements of their therapy, but had similar outcomes in recidivism after control for the participants' risk level (Carl et al., 2019). It needs also to be mentioned that the offenders in our sample were released in 2008 at the latest. Although a long follow-up period is particularly important in sex offender treatment evaluations, this implies that the treatment content may have improved in the meanwhile. We have not yet data on this issue, but we will test it in our currently enlarged data set. In meta-analyses, more recent studies seem not generally reveal better effects (Schmucker & Lösel, 2017), what may be partially due to more rigorous evaluation designs.

Conclusions on Broader Political and Practical Issues

Our consistently positive, although only partially significant, finding on nonsexual recidivism criteria in both of our analyses seems to be a robust and encouraging finding. It agrees with studies that used only one matching method (Grady et al., 2017; Olver et al., 2012; Smid et al., 2016). It is also supported by meta-analyses (Schmucker & Lösel, 2015, 2017). Based on these findings, we draw the following conclusions: Reports for policy and practice should emphasize that sex offender treatment is often effective in nonsexual reoffending that may also cause serious harm for victims. We need more research on prison-based programs and how much they address risk factors for general/violent reoffending (e.g., impulsivity, social problem solving, interpersonal skills) versus more specific risks of sexual offending. Although the views on this issue are controversial, it is worthwhile to explore the potential of more specific programs for different groups of sex offenders as it is currently the case in routine practice. For practical reasons, different types of offenders are allocated on the same programs, although there may be more specific risk–need profiles.

Another more general issue that arises from our study is the partial missing of detailed clinical information on data that are practically relevant for treatment and outcome evaluation. There is sound research on risk factors for recidivism (e.g., Hanson & Morton-Bourgon, 2005), but this correlational knowledge is only partially transferred into the design of treatment and outcome evaluation. At best, subgroups are differentiated according to index offenses, but not with regard to more relevant psychological and clinical categories. For example, in our study, there were not sufficient data on sexual deviance/paraphilia and personality disorders as matching variables in the PSM analysis. This was similar in the larger Mews et al. (2017) study. Although sexual deviance and personality disorders are difficult constructs and often not validly assessed in routine practice (Seto, 2018), they are relatively valid predictors of recidivism (Hanson & Morton-Bourgon, 2005). If such core risk factors cannot be included in matching due to missing data, even the most sophisticated PSM method may have limited validity.

As a consequence, routine practice and evaluation studies should aim to collect not too many more or less relevant data, but a small number of core risk/need variables as validly as possible. This was also an experience with the originally very comprehensive OASys that had to cope with many missing or not very reliable data.

Another more general conclusion refers to treatment practice. The mostly small, often nonsignificant, and sometimes even negative results on sex offender treatment in prisons require more and differentiated analyses. The not yet clearly proven effectiveness of treatment in prisons should lead to practical consequences. For example, even a sound prison-based program may induce limited change in a child molester who cannot test his coping with risk situations in custody. The transfer of acquired new knowledge, attitudes, and skills from custody to the real world outside is a key challenge, not only, but in particular, for sex offenders. A comparison of meta-analytic results on the treatment of sex offenders in prisons and in the community showed generally larger effects in the latter context (Lösel & Koehler, 2014). Of course, these findings did not contain direct comparisons of both settings because these would have been biased due to different populations. The data compared treatment in custody and in the community versus untreated control groups in the respective setting. The more promising results of sex offender treatment in the community fit to those articles in the present special issue that found low reoffending rates in prospective longitudinal studies on sexual offenders. These studies revealed "natural" individual and social protective factors that help sex offenders to increase self-control, make use of social resources, and desist from sexual offending. The findings should be translated into approaches of how to enrich and improve sexual offender treatment programs. Longitudinal and treatment studies should also investigate more intensively the processes that lead to negative outcomes (reoffending in "natural" development or after treatment). Unsuccessful psychosocial interventions can learn from engineering, where failure is not only seen as negative, but used in analyses to improve techniques (Boruch & Ruby, 2015). This implies more investment into widely emphasized relapse prevention, including intensive aftercare and booster programs for sex offenders after

release from prison. Further analyses in our project will analyze the potential add-on value of such programs that have been implemented by the Bavarian MoJ.

A last conclusion of our study refers back to the evaluation designs. As other international researchers, we could only make a virtue out of necessity when we applied two different matching approaches. As mentioned, both designs have specific problems. A large RCT with nonselective attrition would have been preferable, although there are also plausible objections to view it generally as the gold standard for evidence (Nagin & Sampson, 2019). The major advantage of RCTs is that randomized treatment allocation allows to control for all known as well as unknown confounders and thus to maximize internal validity. However, RCTs also contain threats to internal validity, for example, in studies with small to moderate sample sizes, selective dropout, demoralization of control groups, and diffusion of treatment (e.g., Lösel, 2007). External validity is also relevant in evaluations of offender treatment (Lösel & Köferl, 1989).

Here, we cannot repeat the intensive discussion of RCTs on the treatment of sexual offenders (Hanson, 2010; Marshall & Marshall, 2010; Rice, 2010). However, as our study refers to routine practice, we need to mention some issues: The often raised argument that RCTs are not appropriate for ethical reasons is not justified in those situations where treatment is a limited resource and cannot be delivered to all who are in need. In those cases, a random selection is even the fairest allocation procedure. Like in parts of medicine, randomization is also justified when no successful treatment is yet known. When it is questionable whether sex offender treatment in prison has a positive effect or may sometimes even lead to undesirable outcomes such as in the Mews et al. (2017) study, the widely emphasized ethical arguments against RCTs are not justified. RCTs are also not necessarily more costly than quasi-experiments, but they need a thorough planning before treatment programs are implemented.

The main obstacles to RCTs in sexual offender treatment are different: In various countries (as in Germany), the justice systems now require mandatory treatment of serious sexual offenders, so that judges, advocates, and parole boards may object to an allocation to an untreated control group. They and politicians may also fear the risk of a released untreated offender who seriously reoffends. Waiting control groups are much less feasible than in other fields of psychotherapy because reoffending data require long follow-ups. Intermediate psychological or psychometric data are not sufficiently valid because improvement in these indicators shows only very small correlations to later recidivism (e.g., Schwedler & Schmucker, 2012). For these and other reasons, most evaluations of sexual offender treatment use quasi-experimental designs. Of course, we should aim for more large and unbiased RCTs where possible, but if these are not feasible, we should also try to increase sound knowledge by quasi-experiments. For example, countries may implement sound quasi-experiments in a multicenter approach with identical treatment concepts and evaluation designs at various sites. We should also carry out more systematic comparisons of different kinds of treatment and related characteristics of implementation. Particularly important is a detailed documentation of treatment, participant, context, and outcome characteristics. As in other fields of intervention, we need more replication as well as differentiation in our evaluations (Lösel, 2018).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was financially supported by the Bavarian State Ministry of Justice and independently carried out at the Friedrich-Alexander-University Erlangen-Nürnberg (Germany).

Statistical Significance Statement

The authors take responsibility for the integrity of the data and the accuracy of the data analyses, and have made every effort to avoid inflating statistically significant results.

ORCID iD

Eva Link https://orcid.org/0000-0002-0257-6420

Supplemental Material

Supplemental material for this article is available online.

References

Abracen, J., Looman, J., Ferguson, M., Harkins, L., & Mailloux, D. (2011). Recidivism among treated sexual offenders and comparison subjects: Recent outcome data from the Regional Treatment Centre (Ontario) high-intensity Sex Offender Treatment Programme. *Journal of Sexual Aggression*, *17*, 142-152.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399-424. doi:10.1080/00273171.2011.568786

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*, *26*, 734-753. doi:10.1002/sim.2580

Austin, P. C., & Stuart, E. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*, 3661-3679. doi:10.1002/sim.6607

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*, 452-454.

Boruch, R., & Ruby, A. (2015). To flop is human: Inventing better scientific approaches to anticipating failure. In R. Scott & S. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences* (pp. 1-15). New York, NY: Wiley.

Breuer, M., & Endres, J. (2016). Recidivism of treated and untreated sexual offenders released from Bavarian prisons between 2004 and 2009. Unpublished manuscript.

Brookhart, M. A., Schneeweiss, S., Rothmann, K., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*, 1149-1156.

Carl, L., Lauchs, L., Schmucker, M., & Lösel, F. (2019). *"Benchmarking" in der Sozialtherapie: Vergleiche zwischen sozialtherapeutischen Abteilungen für Sexualstraftäter in Bayern*. [Benchmarking in social therapy: Comparison between sociotherapeutic facilities for sex offenders in Bavaria]. *Forensische Psychiatrie, Psychologie, Kriminologie*, 1-9. doi:10.1007/s11757-019-00553-4.

Carl, L., Schmucker, M., & Lösel, F. (2019). *Predicting attrition and engagement in the treatment of young offenders*. Manuscript submitted for publication.

Duwe, G., & Goldman, R. A. (2009). The impact of prison-based treatment on sex offender recidivism: Evidence from Minnesota. *Sexual Abuse: A Journal of Research and Treatment*, *21*, 279-307. doi:10.1177/1079063209338490

Farrington, D. P., Lösel, F., Boruch, R., Gottfredson, D. C., Mazerolle, L., Sherman, L. W., & Weisburd, D. (2019). Advancing knowledge about replication in criminology. *Journal of Experimental Criminology*, *15*, 373-396. doi:10.1007/s11292-018-9337-3

Fixsen, D. L., Blase, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice*, *19*, 531-540.

Friendship, C., Mann, R. E., & Beech, A. R. (2003). Evaluation of a national prison-based treatment program for sexual offenders in England and Wales. *Journal of Interpersonal Violence*, *18*, 744-759. doi:10.1177/0886260503253236

Grady, M. D., Edwards, D., & Pettus-Davis, C. (2017). A longitudinal outcome evaluation of a prison-based sex offender treatment program. *Sexual Abuse: A Journal of Research and Treatment*, *29*, 239-266. doi:10.1177/1079063215585731

Grady, M. D., Edwards, D., Pettus-Davis, C., & Abramson, J. (2013). Does volunteering for sex offender treatment matter? Using propensity score analysis to understand the effects of volunteerism and treatment on recidivism. *Sexual Abuse: A Journal of Research and Treatment*, *25*, 319-346. doi:10.1177/1079063212459085

Graham, J. W. (2012). *Missing data: Analysis and design*. Heidelberg, Germany: Springer.

Haas, S. (2014). *Beurteilerübereinstimmung und -reliabilität des vom Kriminologischen Dienst des Bayerischen Justizvollzugs erstellten Erhebungsbogens für Sexualstraftäter. Ein Untersuchungsinstrument zur Analyse von Rückfalldeterminanten* [Interrater-agreement and -reliability of a data entry form for sexual offenders developed by the Bavarian Prison Service Research Center]. (Unpublished master's thesis). Leopold-Franzens-Universität, Innsbruck, Austria.

Hanson, R. K. (2010, September). *Meta-analysis of treatment outcome in sexual offenders*. Paper presented at the 11th Conference of the International Association for the Treatment of Sexual Offenders, Oslo, Norway. Retrieved from http://www.iatso.org/images/stories/pdfs/hanson_r.k._iatso_2010_meta-analysis_of_treatment_outcome_in_sexual_offenders.pdf

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*, 865-891. doi:10.1177/0093854809338545

Hanson, R. K., & Bussière, M. T. (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology*, *66*, 348-362.

Hanson, R. K., & Morton-Bourgon, K. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology*, *73*, 1154-1163.

Hanson, R. K., & Morton-Bourgon, K. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment*, *21*, 1-21. doi:10.1037/a0014421

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior*, *24*, 119-136.

Helmus, L., Thornton, D., Hanson, R. K., & Babchishin, K. M. (2012). Improving the predictive accuracy of Static-99 and Static-2002 with older sex offenders: Revised age weights. *Sexual Abuse: A Journal of Research and Treatment*, *24*, 64-101.

Ho, D. K. (2015). Ineffective treatment of sex offenders fails victims. *British Medical Journal*, *350*, h199. Retrieved from https://www.bmj.com/content/350/bmj.h199

Ho, D. K., & Ross, C. (2012). Cognitive behaviour therapy for sex offenders: Too good to be true? *Criminal Behaviour and Mental Health*, *22*, 1-6. doi:10.1002/cbm.1818

Jann, B. (2017). *Why propensity scores should be used for matching*. Paper presented at the German Stata Users Group Meeting, Berlin, Germany. Retrieved from https://www.stata.com/meeting/germany17/slides/Germany17_Jann.pdf

Jennings, W. G. (2015). *Innovations and advancements in sex offender research*. New York, NY: Routledge.

King, G., & Nielsen, R. (2016). *Why propensity scores should not be used for matching* (Working Paper). Retrieved from http://j.mp/2ovYGsW

Koehler, J. A., & Lösel, F. (2015). A differentiated view on the effects of sex offender treatment. *British Medical Journal (eLetter)*. Retrieved from https://www.bmj.com/content/350/bmj.h199/rr-0

Lipsey, M. W. (2018). Effective use of the large body of research on the effectiveness of programs for juvenile offenders and the failure of the model programs approach. *Criminology & Public Policy*, *17*, 189-198. doi:10.1111/1745-9133.12345

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

Lipton, D. S. (2010). A therapeutic distinction with difference: Comparing American concept-based therapeutic communities and British democratic therapeutic community treatment for prison inmates. In R. Shuker & E. Sullivan (Eds.), *Grendon and the emergence of forensic therapeutic communities* (pp. 61-78). Chichester, UK: John Wiley.

Lösel, F. (2007). Doing evaluation in criminology: Balancing scientific and practical demands. In R. D. King & E. Wincup (Eds.), *Doing research on crime and justice* (pp. 141-170). Oxford, UK: Oxford University Press.

Lösel, F. (2012). Offender treatment and rehabilitation: What works? In M. Maguire & R. Morgan (Eds.), *The Oxford handbook of criminology* (5th ed., pp. 986-1016). Oxford, UK: Oxford University Press.

Lösel, F. (2018). Evidence comes by replication, but needs differentiation: The reproducibility issue in science and its relevance for criminology. *Journal of Experimental Criminology*, *14*, 257-278.

Lösel, F., & Egg, R. (1997). Social-therapeutic institutions in Germany: Description and evaluation. In E. Cullen, L. Jones, & R. Woodwards (Eds.), *Therapeutic communities for offenders* (pp. 181-203). Chichester, UK: John Wiley.

Lösel, F., & Koehler, J. (2014, November 19-22). *Can prisons reduce reoffending? A meta-evaluation of custodial and community treatment programs*. Paper Presented at the 70th Annual Conference of the American Society of Criminology, *San Francisco, CA*.

Lösel, F., & Köferl, P. (1989). Evaluation research on correctional treatment in West Germany: A meta-analysis. In H. Wegener, F. Lösel, & J. Haisch (Eds.), *Criminal behavior and the justice system* (pp. 334-355). New York, NY: Springer.

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

Lösel, F., & Schmucker, M. (2017). Treatment of sexual offenders: Concepts and empirical evaluations. In T. Sanders (Ed.), *The Oxford handbook on sex offences and sex offenders* (pp. 392-414). New York, NY: Oxford University Press.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review*, *29*, 530-558. doi:10.1177/0193841X05275596

Mann, R. E., Carter, A. J., & Wakeling, H. C. (2012). In defence of NOMS' view about sex offending treatment effectiveness: A reply to Ho and Ross. *Criminal Behaviour and Mental Health*, *22*, 7-10. doi:10.1002/cbm.1821

Mann, R. E., & Thornton, D. (1998). The evolution of a multisite sexual offender program. In W. L. Marshall, Y. M. Fernandez, S. M. Hudson, & T. Ward (Eds.), *Sourcebook of treatment programs for sexual offenders* (pp. 47-57). New York, NY: Plenum.

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: A Journal of Research and Treatment*, *17*, 79-107.

Marshall, W., & Marshall, L. (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*, 175-191.

Marshall, W., & Marshall, L. (2010). Can treatment be effective with sexual offenders or does it do harm? A response to Hanson (2010) and Rice (2010). Retrieved from http://www.sexual-offender-treatment.org/87.html

McCann, K., & Lussier, P. (2008). Antisociality, sexual deviance, and sexual reoffending in juvenile sex offenders: A meta-analytical investigation. *Youth Violence and Juvenile Justice*, *6*, 363-385.

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

Mews, A., Di Bella, L., & Purver, M. (2017). *Impact evaluation of the prison-based core sex offender treatment programme* (Ministry of Justice Analytical Series). London, England. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/623876/sotp-report-web-.pdf

Morrison, D. E., & Henkel, R. E. (1972). The significance test controversy. *The British Journal for the Philosophy of Science*, *23*, 170-181.

Nagin, D. S., & Sampson, R. J. (2019). The real gold standard: Measuring counterfactual worlds that matter most to social science and policy. *Annual Review of Criminology*, *2*, 123-145.

Nisbet, I. A., Wilson, P., & Smallbone, S. W. (2004). A prospective longitudinal study of sexual recidivism among adolescent sex offenders. *Sexual Abuse: A Journal of Research and Treatment*, *16*, 223-234.

Olver, M. E., Nicholaichuk, T. P., Gu, D., & Wong, S. C. P. (2012). Sex offender treatment outcome, actuarial risk, and the aging sex offender in Canadian corrections: A long-term follow-up. *Sexual Abuse: A Journal of Research and Treatment*, *25*, 396-422. doi:10.1177/1079063212464399

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi:10.1126/science.aac4716

Orlinsky, D. E., Grawe, K., & Parks, B. K. (1994). Process and outcome in psychotherapy: Noch einmal. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (pp. 270-376). Oxford, UK: John Wiley.

Procter, E. (1996). A five-year outcome evaluation of a community-based treatment program for convicted sexual offenders run by the probation service. *Journal of Sexual Aggression*, *2*, 3-16.

Rettenberger, M., Haubner-MacLean, T., & Eher, R. (2013). The contribution of age to the Static-99 risk assessment in a population-based prison sample of sexual offenders. *Criminal Justice and Behavior*, *40*, 1413-1433.

Rice, M. E. (2010, September). *Treatment for adult sex offenders: May we reject the null hypothesis?* Paper Presented at the 11th Conference of the International Association for the Treatment of Sexual Offenders, Oslo, Norway. Retrieved from http://www.iatso.org/images/stories/pdfs/rice_m.e._iatso_2010_treatment_for_adult_sex_offenders_may_we_reject_the_null-hypothesis.pdf

Rice, M. E., & Harris, G. T. (2003). The size and sign of treatment effects in sex offender therapy. In R. A. Prentky, E. S. Janus, & M. C. Seto (Eds.), *Sexually coercive behavior: Understanding and management* (pp. 428-440). New York: New York Academy of Sciences.

Rice, M. E., Harris, G. T., & Quinsey, V. L. (1991). Evaluation of an institution-based treatment program for child molesters. *Canadian Journal of Program Evaluation*, *6*, 111-129.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*, 33-38.

Savitz, D. A. (1993). Is statistical significance testing useful in interpreting data? *Reproductive Toxicology*, *7*, 95-100.

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*, 597-630. doi:10.1007/s11292-015-9241-z

Schmucker, M., & Lösel, F. (2017). Sexual offender treatment for reducing recidivism among convicted sex offenders: A systematic review and meta-analysis. *Campbell Systematic Reviews*, *13*. Retrieved from https://dx.doi.org/10.4073/csr.2017.8

Schwedler, A., & Schmucker, M. (2012). Verlaufsmessung im sozialtherapeutischen Behandlungsvollzug - Wie sinnvoll sind allgemeine Persönlichkeitsmaße? [The suitability of personality questionnaires to measure within-treatment change in offenders]. *Monatsschrift Für Kriminologie und Strafrechtsreform*, *95*, 269-280.

Seto, M. C. (2018). *Pedophilia and sexual offending against children: Theory, assessment, and intervention* (2nd ed.). Washington, DC: American Psychological Association.

Seto, M. C., Marques, J. K., Harris, G. T., Chaffin, M., Lalumière, M. L., Miner, M. H., . . . Quinsey, V. L. (2008). Good science and progress in sex offender treatment are intertwined. *Sexual Abuse: A Journal of Research and Treatment*, *20*, 247-255. doi:10.1177/1079063208317733

Shuker, R., & Sullivan, E. (Eds.). (2010). *Grendon and the emergence of forensic therapeutic communities*. Chichester, UK: John Wiley.

Smallbone, S., & McHugh, M. (2010). *Outcomes of Queensland corrective services sexual offender treatment programs*. Final Report. Brisbane, Queensland: Queensland Government.

Smid, W. J., Kamphuis, J. H., Wever, E. C., & van Beek, D. J. (2016). A quasi-experimental evaluation of high-intensity inpatient sex offender treatment in the Netherlands. *Sexual Abuse: A Journal of Research and Treatment*, *28*, 469-485.

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1-21.

Ware, J., Mann, R. E., & Wakeling, H. C. (2009). Group versus individual treatment: What is the best modality for treating sexual offenders? *Sexual Abuse in Australia and New Zealand*, *1*, 70-78.

Weisburd, D., Lum, C. M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *The Annals of the American Academy of Political and Social Sciences*, *578*, 50-70.

Wilson, D. (2017). Correctional programs. In D. Weisburd, D. P. Farrington, & C. Gill (Eds.), *What works in crime prevention and rehabilitation: Lessons from systematic reviews* (pp. 193-217). New York, NY: Springer.

Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C., & Smith, D. (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health*, *13*, 273-277. doi:10.1111/j.1524-4733.2009.00671.x