Journal of Developmental and Life-Course Criminology (2023) 9:433–454
https://doi.org/10.1007/s40865-023-00233-2

ORIGINAL ARTICLE

The Short-Term Impact of Formal Controls on Subsequent Offending and Future Formal Controls in a German and UK City

Klaus Boers[1] · Florian Kaiser[1,2] · Marcus Schaerff[1] · Per-Olof H. Wikström[3]

Received: 11 August 2022 / Revised: 21 March 2023 / Accepted: 3 April 2023 /
Published online: 1 June 2023

Abstract

The study of sanctioning effects has a rich history in deterrence and labeling theory. Most analyses have only used official data to study these effects. Yet, some more recent studies indicate that it is necessary to investigate self-reported as well as official data since it appears that sanctioning has differential effects on self-reported delinquency and formal control interventions. The current study contributes to this small body of research by using propensity score matching to analyze panel data from an ongoing English (Peterborough Adolescent and Young Adult Developmental Study) and a German (Crime in the modern City) study. We estimated average treatment effects of system contacts on both reoffending and subsequent contacts for juveniles living in Peterborough (ENG) and Duisburg (GER). Our findings are that (1) although official contacts have no substantial effects on the prevalence or versatility of reoffending, (2) they substantially increase the risk of a future formal contact. These results were almost identical at both sites, which may indicate a more general finding on the effects of formal control interventions.

Keywords Effects of formal control interventions · Secondary sanctioning · Propensity score matching · Cross-national research

✉ Klaus Boers
boers@uni-muenster.de

[1] Institute of Criminal Law and Criminology, University of Muenster, Bispinghof 24/25, 48143 Münster, Germany

[2] Max-Planck-Institute for the Study of Crime, Security and Law, Freiburg, Germany

[3] Institute of Criminology, University of Cambridge, Cambridge, UK

Introduction

In criminology, the impact of interventions and sanctions by the formal social control agencies, police and criminal courts, is assumed to have two directions: They may prevent offenders from further offending, or they may reinforce subsequent delinquency.

The idea that penal sanctions shall have a crime-preventive impact is mainly a heritage of the Age of Enlightenment. In order to restrict the excessive retributive punishments in feudal regimes, two concepts evolved. First, the concept of treatment: if offenders are treated through a strict working regime— in lieu of corporal or capital punishment—they will become honest and will rehabilitate (Howard, 1777).[1] Second, the concept of deterrence: Philosophers like Beccaria (1986 [1764]) or Bentham (1988 [1776]) proposed that a humane penal harm (which excluded corporal or capital punishment) should be determined in such a way that it should merely deter the offender from further offending by a sanction proportionate to the offense-induced harm (see Bruinsma, 2018).

It took some time until these preventive ideas arrived in the law books and in penal practice. In England and the USA, preventive programs became relevant already in the nineteenth century, while in Germany (with the exception of juvenile penal law),[2] they became influential only in the late 1960s, two decades after the Nazi-Regime had been defeated.

Today, rehabilitative treatment and deterrence form next to retribution (as offense-proportionate and insofar restricted and just punishment) the basic legitimacy of a modern criminal law. The preventive turn also resulted in a further innovation in terms of modern rationality. From then on, the effectiveness of the criminal justice system became an object of empirical investigation. Next to the black letter lawyer, the social and behavioral scientist entered the stage of penal sciences.

However, against the backdrop of empirical observations, penal sanctions appeared to be much less promising in preventing further offending than the rather optimistic modern reformers had expected (for a review, see Sherman et al., 1998). This was apparently one reason for the broad attention given to the alternative theoretical perspective of labeling, which assumes that penal interventions do not prevent but reinforce or even initiate delinquent behavior.

With the methodological progress of panel studies in developmental and life-course criminology, scholars received the appropriate tools to analyze the causal impact of penal sanctions using quasi-experimental designs. Nevertheless, these sophisticated studies did not produce clear support for unidimensional preventive or promoting effects of penal sanctions either. Rather, many estimated effects were statistically insignificant (Barrick, 2014; Huizinga & Henry, 2008; Kleck & Sever, 2017).

[1] Amsterdam working houses (*tuchthuis;* Krause, 1999, pp. 99).

[2] In 1923, a special, education-centered law for dealing with juvenile offenses was enacted as *Reichsjugendgerichtsgesetz* (RJGG). The basic architecture of this law is still in force in the current *Jugendgerichtsgesetz* (JGG).

However, today, the different theories on sanctioning effects assume mainly a mediated causal process (see Bernburg, 2019; Paternoster, 2018; Krohn et al., 2014): Penal sanctions may lead to an at maximum moderate increase in subjectively perceived detection and sanctioning risks (deterrence); or may support programs which combine the (re-)construction of social bonds with the promotion of cognitive agency (rehabilitative treatment); or may disturb prosocial structural resources and support a delinquent self-concept (labeling). Empirical results appear to support these contradicting assumptions about a mediated impact in one way or another, lending somewhat more evidence to delinquency-promoting than delinquency-preventing mechanisms (Bernburg, 2019; Huizinga & Henry, 2008; Paternoster, 2018). Due to limited space, we will not report on the impact of formal controls on mediating factors, which were analyzed separately in Kaiser (2022; see also discussion section).

Furthermore, most studies on the impact of formal controls were limited by using official court and police data as proxy for subsequent delinquency (Barrick, 2014; Kleck & Sever, 2017). However, official crime data result from a mixture of delinquent behavior and the reactions of criminal justice agents and thus do not present a pure measure of juvenile behavioral change. Indeed, some US studies show that the exclusive reliance on official data may be problematic. According to their results, formal controls may increase the risk of further formal controls, independent of changes in delinquent behavior (called "secondary sanctioning" by Liberman et al., 2014).

The current study uses adolescent data from two panel studies that have been conducted in England and Germany from 2002 onwards: the Peterborough Adolescent and Young Adult Developmental Study (PADS+) and the study Crime in the modern City (CrimoC), carried out in Duisburg. The goal of this analysis is to investigate whether the differential impact on official and self-reported data as found in a few US studies can also be seen in European countries. To conduct this investigation, the current study explores both the impact of formal control interventions on (i) subsequent delinquent *behavior*, and (ii) on the risk of subsequent formal controls in a quasi-experimental design (propensity score matching).

After discussing the theoretical framework and reporting on the data as well as on the analytical method, the impact of formal control interventions during adolescence will be analyzed.

Theoretical Framework and Previous Research

In this study, we will investigate whether a criminal justice intervention is associated with changes in young people's future offending. We will also explore whether a criminal justice intervention amplifies the risk for a future criminal justice intervention—independent of changes in delinquent behavior (Fig. 1).

There are two major theories of why criminal justice interventions may affect people's future offending: deterrence theory and labeling theory. In the literature studying the association between criminal justice interventions and future offending, it is common to assume (and sometimes conclude) that increases in future offending

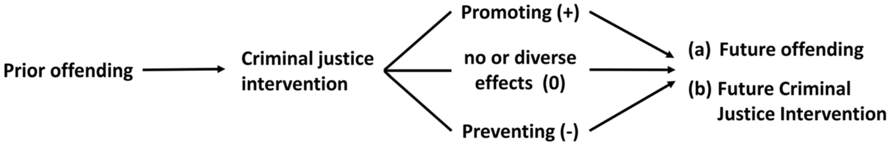

Prior offending
Criminal justice intervention
Promoting (+)
no or diverse effects (0)
Preventing (-)
(a) Future offending
(b) Future Criminal Justice Intervention

Fig. 1 Potential impact of criminal justice interventions on future offending and future interventions

are indicative of a labeling process and that decreases in future offending are indicative of a deterrence process.[3] However, establishing whether young people's future offending is amplified or reduced (or unaffected) by a criminal justice intervention does not in itself answer the question why this happens: "neither increases nor decreases in levels of delinquent involvement following the imposition of sanctions provides unequivocal evidence for either the labeling or deterrence paradigms" (Thomas & Bishop, 1984, p. 1229).[4] However, one could argue that if there are no strong changes (increases or decreases) in future offending, there is no evidence of (no room for) strong unidirectional deterrence or labeling influences.

In line with this, a review of research shows that the most frequent finding is the absence of a statistically significant association between a criminal justice intervention and future offending, although it is somewhat more common with a finding of increased rather than decreased offending among the results that are statistically significant (Barrick, 2014; Kleck & Sever, 2017).[5] A result that does not lend much support to the existence of a strong universal unidirectional labeling or deterrent effect (especially since statistical significance in these studies typically does not equal strong effects; ibid.; Huizinga & Henry, 2008; Pratt & Turanovic, 2018).

Different from the impact on offending behavior is an *institutional effect*: formal control interventions appear to increase subsequent control interventions independent of the level of delinquent behavior (Fig. 1, outcome b). This impact of formal control interventions on further interventions has—although not often investigated—been found to be quite strong.

More precisely, having been arrested (compared to not arrested) tripled the risk of a re-arrest in adolescence or adulthood, and this effect was remarkably stronger than the impact on subsequent self-reported delinquency (Beardslee et al., 2019; Klein, 1986; Liberman et al., 2014; Lopes et al., 2012). Surprisingly, this effect was by and large only slightly stronger in case of a more formal intervention (e.g., court petition) compared to a less intervening informal case processing (e.g., police diversion; Beardslee et al., 2019; Klein, 1986). That this is an institutional effect and not

[3] Furthermore, many studies testing for labeling and/or deterrent effects cannot (like ours) control for rehabilitative treatment due to a lack of data. However, various kinds of treatment are often part of informal and formal sanctioning procedures (see Lipsey, 2009). Therefore, a preventing effect may be assumed to be deterrent although it is (at least in part) rehabilitative.

[4] Few studies have effectively explored the proposed mechanisms theorized as responsible for an association between criminal justice interventions and future offending (i.e., they are typically assumed rather than demonstrated; Huizinga & Henry, 2008).

[5] "A quantitative assessment of studies examining the impact of arrest, conviction, juvenile justice intervention, and incarceration on recidivism provides modest support for the hypothesis that official sanctions, in certain situations,may increase subsequent deviance" (Barrick, 2014, p. 110, *italics added*).

an individual reaction to the application of a formal label (secondary deviance) has already been noticed by Klein (1986): "[L]abelers are somehow responding to their own prior decisions" (p. 63). Liberman et al., (2014), explicitly deliberating on this phenomenon, call it more precisely a "secondary sanctioning" effect which should be reflected in its own right, differently from "secondary deviance" effects as outlined in the early labeling literature (see Lemert, 1951; Becker, 1963): "The effects of secondary deviance and secondary sanctioning are essentially independent" (p. 363).

Hypotheses

Following the current state of research, first, the findings on the overall not strong impact of formal controls on subsequent delinquent behavior are quite mixed (see Barrick, 2014; Huizinga & Henry, 2008; Motz et al., 2020). There is somewhat more support for delinquency-promoting rather than delinquency-preventing effects, while there are also many insignificant findings. Second, regarding an institutional impact, formal controls may increase the risk of subsequent controls (see Liberman et al., 2014).

Our study explores whether the finding that formal controls have a different impact on official re-contact and on subsequent self-reported delinquency can also be observed in European countries. So far, only studies with US data analyzed both official and self-reported outcome data simultaneously (Beardslee et al., 2019; Klein, 1986; Liberman et al., 2014). Their results indicate that the institutional effect of formal controls is (much) larger than the effect on (self-reported) delinquent behavior. These studies should serve as a warning not to use official police or court data as proxies of delinquent behavior, as was done in most previous sanctioning research (see Barrick, 2014; Kleck & Sever, 2017). Their findings imply that formal controls seem to trigger processes that change the risk of re-contact with the criminal justice system beyond changes in delinquent behavior. However, it is unclear whether these processes also operate in other—less punitive—jurisdictions than the USA. To explore whether this might be the case, the current study is, to our knowledge, the first to analyze behavioral and institutional effects simultaneously in a non-US setting. It does so by testing the following hypotheses using data from two European countries:

> Hypothesis 1: Formal controls are more likely to increase rather than decrease later delinquency.
>
> Hypothesis 2: Formal controls increase the risk of subsequent formal controls.

Formal Control Effects in Peterborough and Duisburg

Samples

Our analyses are based on data from the *Peterborough Adolescent and Young Adult Development Study* (PADS+; Wikström et al., 2012, 2023) and the *Crime in the modern City study* (CrimoC; Boers et al., 2010). Both are panel studies that started data collection with 13-year-old school students in Peterborough and Duisburg. Participants were asked to complete standardized questionnaires. In addition, researchers collected the participants' police and court records.

The target population of PADS+ covered all 11-year-old school students who lived in Peterborough and entered year 7 in 2002. After sampling randomly, 710 juveniles (approximately one-third of the population) finally participated in the first wave in 2004. In the follow-up waves—that were conducted annually until age 17 and in 2- and 3-year intervals thereafter—PADS+ achieved retention rates of more than 95% (707 in wave 2, 703 in waves 3 and 4, and 693 in wave 5). *Police National Computer* (PNC) records were collected for 700 students.[6]

In CrimoC, researchers tried to survey all 7th-graders in Duisburg in 2002. Out of 56 schools, 40 (71%) agreed to participate, resulting in 3411 completed questionnaires at wave one (approximately two-thirds of all 7th-graders). The follow-up waves were conducted annually until age 20 and then biennially until age 30. The difference in design resulted in somewhat more unit-non-responses in CrimoC compared to PADS+, although participation was also high (3392 in wave 2, 3339 in wave 3, 3405 in wave 4, and 4548 in wave 5).[7] Official records were available for 2964 respondents (87%).[8]

To establish proper time order for causal inference, three time periods were defined (see Liberman et al., 2014; Wiley & Esbensen, 2016): pretreatment (T1; covariates), treatment (T2, i.e., official contact), and post-treatment (T3; outcomes: self-reported delinquency and official contact). Table 1 shows how the PADS+ and CrimoC data were aligned with these three periods.

In order to be included in the final analyses, participants from both studies had to meet two conditions: (1) participation in waves 3, 4, and 5, as well as (2) access to their official records. All in all, 690 juveniles in PADS+ (97% of 710), and 2117

Table 1 Time order

Time period measures refer to in				
Phase	Ø-age	Pads+	CrimoC	Measures
T1	14	01/–12/2005	01/2003–02/2004	Covariates
T2	15	01/–12/2006	03/2004–12/2004	Official contact
T3	16	01/–12/2007	01/2005–02/2006	SRD, official contact

CrimoC's treatment period (T2) is shorter to take into account that some covariates (e.g., self-reported delinquency, SRD, in T1) refer to the time period from January 2003 to January/February 2004, whereas comparable measures in PADS+ refer only to whole years (e.g., whole year 2005)

[6] For more information on PADS+, see https://www.cac.crim.cam.ac.uk/research/padspres or Wikström et al. (2012).

[7] Wave 5 included also students from vocational schools who participated for the first time.

[8] Official records (received from the *Bundeszentralregister* (BZR) and *Erziehungsregister (ER)*) are based on court and prosecution data and comprise all decisions made after opening an official investigation by police: from dismissal in case of lacking evidence up to convictions. – Number of respondent refers to wave 4 when official records were collected; for more information on CrimoC, see https://www.crimoc.org or Boers et al. (2010).

in CrimoC (62% of 3405)[9] matched these criteria. In CrimoC, the resulting sample consists of somewhat less "high-risk youth" than the original sample.[10]

Measures

Our measurement descriptions follow the division into the three (quasi-)experimental time-periods: treatment, outcomes, and covariates (see Table 2 and Supp. material S2a, S2b for descriptive statistics).

The *treatment* variable is official control, a binary variable distinguishing between those with "no-official contact" (=0) and those with "official contact" (=1). In PADS+, 37 of 690 (5.4%) and in CrimoC 88 of 2117 (4.2%) juveniles had been officially registered for an offense within T2. This low number of juveniles with a system contact is in line with previous literature on the risk of police contact (Kaiser et al., 2022a; Lochner, 2007; Wikström et al., 2012). In both samples, official intervention was generally not very intensive. Usually, juveniles were diverted out of the system or received some form of educational measures (see Table 3).

Outcome variables are self-reported delinquency (SRD) indices and official contact measures (PADS+: PNC; CrimoC: BZR, ER). The pool of SRD indicators consists of nine (PADS+) or 13 (CrimoC) offenses, respectively, committed *in the last year* (PADS+) or *since the start of the last year* (CrimoC). Although the number of offenses varies between PADS+ and CrimoC, they cover the same categories of delinquent behavior. On the one hand, SRD indicators were used to calculate prevalence rates of general, violent, and property offenses as well as vandalism.[11] On the other hand, in

Table 2 Descriptive statistics

Variables	T1		T3	
	PADS+	CrimoC	PADS+	CrimoC
Female	50.4%	56.5%		
Migration background	22.3%	36.2%		
SRD violence prevalence	28.6%	13.4%	20.6%	8.7%
SRD property prevalence	20.1%	17.8%	12.8%	10.0%
SRD vandalism prevalence	18.7%	17.9%	10.9%	9.5%
SRD general prevalence	40.8%	32.1%	29.6%	19.7%
SRD versatility	0.8	0.6	0.5	0.4
Official contact prevalence	3.5%	3.1%	5.9%	5.5%
n	690	2117	690	2117

[9] The sample size of wave 4 was selected because in this wave, respondents were asked to consent to a collection of their official data.

[10] Among those not included in the final CrimoC sample, the reported level of self-reported delinquency, police-related problems, deviant peer group activities, and school performance problems was somewhat higher; for more information, see Supp. material S2b.

[11] In PADS+, offending categories consist of the following offenses: (1) violence = assault, robbery, (2) property offending = shoplifting, theft from car, theft of car, theft from person, residential burglary, non-residential burglary, (3) vandalism; in CrimoC: (1) violence = assault without weapon, assault with weapon, robbery, bag snatching (2) property offending = shoplifting, theft from car, theft of car, bicycle theft, theft from person, burglary (3) vandalism = property damages, graffiti spraying, scratching.

Table 3 Reactions of the juvenile justice system in PADS+ (English sample), CrimoC (German sample)

	Study samples	
Reaction	PADS+	CrimoC
Diversion	73.0%	80.7%
Conviction	27.0%	19.3%
Non-custodial measures	90.0%	82.4%
Short-term juvenile detention	–	17.6%
Juvenile imprisonment	10.0%	–

n(PADS +) = 37; *n*(CrimoC) = 87 (missing data for one participant)

order to measure offending intensity, versatility scores were computed (with a maximum of 9 or 13 different offenses in PADS+ or CrimoC, respectively).[12] In addition, official control (0 = no contact; 1 = contact) within T3 was also considered as an outcome variable in order to analyze effects of "secondary sanctioning" (Liberman et al., 2014).

Covariates For each study site, the selection of more than 50 covariates was guided by theoretical considerations and empirical evidence. Consequently, they cover a wide range of variables known to be related to offending or an official contact: deviant and delinquent behavior, previous formal controls, individual characteristics, peer, family and school bonding, parental education, neighborhood, and demographics. Including multiple indicators is regarded as a promising strategy to tackle selection bias threats effectively (Steiner et al., 2010). SRD and official control measures in T1 are also included as covariates because matching on them assures that the treatment and the control group are balanced on the focal variables of the current study at baseline.[13]

Analytical Procedure

Methodologically, the crucial point in analyzing formal control interventions is to avoid selection bias: to make sure that post-intervention differences between an intervention and a control group are based on the intervention only, both groups should not differ on other characteristics (so called covariates), following ideally the ceteris paribus-rule. This can best be achieved by an experimental research design based on a random selection of both groups. However, for legal reasons, police, prosecutors, or judges are not allowed to decide randomly whether to intervene or not to intervene in delinquent behavior. Therefore, formal control interventions can typically only be investigated within a quasi-experimental setting. Here, one tries to minimize selection bias by controlling statistically for confounding covariates (see Morgan & Winship, 2015; Shadish et al., 2002). It was common

[12] Another common way to measure intensity is to compute frequency rates (number of offenses per offender; Blumstein et al., 1986). We used versatility scores because they have better statistical properties (Sweeten, 2012), and produced more precise estimates than frequency rates with the current data.

[13] For a complete list of covariates, see Supp. material S2a and S2b.

practice to rely on multiple regression analysis to address threats of selection bias (see Nagin et al., 2009). After it turned out, however, that multiple regression is not efficient enough in controlling for confounding effects of covariates (Smith & Paternoster, 1990), propensity score matching (PSM) has been applied as a more appropriate tool of accounting for selection effects (McAra & McVie, 2007; Morris & Piquero, 2013; Ward et al., 2014; Wiley & Esbensen, 2016; Wiley et al., 2013).

To explore how a contact with the juvenile justice system affects the risk of reoffending and further official contact, we use PSM to estimate the *average treatment effect on the treated* (ATT) as our causal estimate of interest. Derived from the potential outcome model (see Morgan & Winship, 2015; Rubin, 1974), the ATT is computed in the following way:

$$\mathrm{ATT} = \mathrm{E}[\delta|\ \mathrm{Tr} = 1] = \mathrm{E}[\mathrm{Y}_\mathrm{i}^1 - \mathrm{Y}_\mathrm{i}^0|\mathrm{Tr} = 1]$$

The ATT refers to officially treated individuals only (Tr = 1). It is defined as the average (E[])[14] difference ($Y_i^1 - Y_i^0 = \delta$) between their observed reoffending (Y_i^1) and "their" hypothetical reoffending, i.e., under the assumption that they would not have been treated (Y_i^0). In reality, a treated individual experienced only the treatment condition (official contact) and not the control condition (*no* contact). Hence, only one (Y_i^1) of the two potential outcomes (Y_i^1, Y_i^0) can be observed. Consequently, causal effects cannot be computed from the observed values of the treated individuals alone. This missingness of the counterfactual outcome value (Y_i^0 as the value not realized in reality) is termed the "fundamental problem of causal inference" (Holland, 1986).

To overcome this problem and estimate ATTs, we applied PSM. Matching (including weighting) procedures generally mimic a randomized experiment by balancing the treatment and control group on an array of covariates selected for matching (Morgan & Winship, 2015; Stuart, 2010). They do so by finding and matching control units that are equal (exact matching) or at least most similar to treated units on all selected pretreatment measures. Individuals from the control group that are too dissimilar to the treated individuals are excluded from analyses. Included individuals from the control group are finally used to infer the counterfactual outcome value, allowing for an ATT estimation. Unlike a randomized experiment, matching, however, does not automatically balance unobserved characteristics of treated and untreated individuals. Furthermore, classical matching procedures were based on exact matchings, i.e., finding individuals for the control and treatment group with exactly the same values. However, the higher the number of covariates the less likely it is to meet such a requirement ("curse of dimensionality," Apel & Sweeten, 2010, p. 544). To face this problem, Rosenbaum and Rubin (1983) introduced the so-called propensity score. It refers to the probability that an individual received the treatment. For this study, the propensity score describes the probability that a juvenile was officially recorded for an offense in T2. A great advantage of this single score is that matching on it (i.e., finding individuals with most similar propensity scores

[14] E[] is the probability theory's expectation operator.

among treated and untreated respondents) may be sufficient to balance the treatment and control group on all pretreatment covariates (Kainz et al., 2017).

Our matching procedure followed four steps (Stuart, 2010): First, we estimated propensity scores for each PADS+ and CrimoC sample member with the help of three different estimation procedures: Bayesian logistic regression (BLR; McElreath, 2016), Bayesian Additive Regression Trees (BART; Chipman et al., 2010), and the covariate balancing propensity score (CBPS; Imai & Ratkovic, 2014).[15] Second, these three propensity scores were applied in 12 different matching (or weighting) algorithms to find the combination that leads to the best distributional balance of all covariates between the treatment and control groups.[16] The application of different combinations of propensity score and matching algorithms is recommended to ensure that selection threats induced by pretreatment differences in observed covariates are minimized (e.g., Kainz et al., 2017; Morgan & Winship, 2015). Third, we selected the best PSM procedure for each sample by assessing the covariate balance achieved by the different method combinations using recommended balance statistics (Kainz et al., 2017; see section *Covariate Balance Assessment*). Fourth, we applied regression models (R's *Zelig* package, Imai et al., 2008) to the best-matched samples to estimate ATTs and simulate their uncertainty. While binary SRD prevalence and official contact outcomes were modeled by logistic regression, SRD versatility indices were predicted by Poisson models.[17]

Because the investigated samples suffered from item-non-response, all analytical steps were applied to multiple imputed data sets. Multiple imputation embraces the estimation uncertainty emerging due to missing information in the data set (van Buuren, 2018). It generates multiple data sets by making multiple predictions for the missing values using observed information from other variables. As recommended by Penning de Vries and Groenwold (2017), we conducted matching, the generation of weights, and also the outcome analyses for each imputed data set. The imputed

[15] For estimation of the propensity score, we included (a subset of) the covariates (36 in PADS+, and all 52 in CrimoC) as predictors in each modeling procedure. All computations were conducted in *R* (version 3.5.2; R Core Team, 2018). A list of all R packages used for the analysis is provided in Supp. material S5.—Potential overdetermination of the propensity score model was checked by a simulation study. As a result, PADS+ models included less covariates.

[16] The 12 different matching (weighting) algorithms were (Stuart, 2010): (1–5) nearest neighbor matching with replacement, a caliper of 0.25, and ratios of 1:1 to 1:5, (6–9) optimal matching with ratios 1:1 to 1:3, (10–11) genetic matching with replacement and ratios 1:1 to 1:2, and (12) weighting by the odds.

[17] As predictors, the models included the treatment variable (official contact in T2), the lagged outcome and their interaction term: $\text{Outcome}_{\text{T3}} = \alpha + \beta_1\ \text{Treatment}_{\text{T2}} + \beta_2\ \text{Outcome}_{\text{T1}} + \beta_3\ (\text{Treatment}_{\text{T2}} * \text{Outcome}_{\text{T1}})$. Zelig applies the following formula to compute the ATTs from the regression models: $\frac{1}{\sum_{i=1}^{n} T_i} \sum_{i:T_i=1}^{n} \left\{ Y_i^1 - E\left[Y_i^0\right] \right\}$. Within this formula, only the counterfactual values (i.e., Y_i^0) are estimated with the help of the regression models because Y_i^1 is observed for all treated individuals and can, therefore, be directly filled into the equation. To check for the robustness of the outcome analyses (see Supp. material S4), we conducted not only the aforementioned regression specification but also estimated mean differences (regression models with only the treatment as predictor), and weighted regressions (including a fuller set of predictors).

information was finally combined by merging the ATT simulations of all imputed data sets together.[18]

In addition, we also conducted robustness analyses to check how sensitive the ATTs were in relation to different missing data, propensity score estimation, matching, and outcome modeling procedures (Young & Holsteen, 2016). We restricted our sensitivity checks to those propensity score and matching procedure combinations that were relatively successful in establishing covariate balance between treated and untreated individuals.

Results

In this section, we first report how the best-working matching methods balanced the covariate distributions before presenting the ATT estimates and robustness checks.

Covariate Balance Assessment

In the following, we assess the covariate balance of the best-balancing matching procedures using standardized bias (SB) and variance ratio (VR) statistics (Kainz et al., 2017). SB is the difference in covariate means between the treated and untreated group divided by the standard deviation of the treated group. VRs, in contrast, inform about the variance differences in continuous covariates across the treated and untreated groups. SB thresholds of larger than 0.1 and VRs larger than 2 or smaller than 0.5 indicate imbalance (Harder et al., 2010; Kainz et al., 2017).[19]

PADS+

In PADS+, treated individuals differed from untreated ones on an array of pretreatment characteristics. The majority of covariates (44 of 57) was imbalanced before matching, showing SB statistics larger than 0.1; for 34 covariates, the bias was larger

[18] The CrimoC sample is more strongly affected by missing values than that of PADS+. Thus, we produced only 12 imputed data sets for PADS+ but 70 for CrimoC. We applied predictive mean matching within a fully conditional specification (van Buuren, 2018) and additionally also other imputation procedures (e.g., random forests). These sensitivity checks showed that our results are quite robust to the imputation technique applied (see Supp. material S2b and S4). Discussing the difficulties of combining multiple imputation and propensity score matching, Penning de Vries and Groenwold (2017) present two approaches: (1) conduct propensity score estimation, the matching, the generation of weights, and the outcome analyses for each imputed data set before merging the results; (2) estimate propensity scores for each imputed data set, calculate an average propensity score for each individual, and then do the matching, generation of weights, and outcome analyses based on only this single data set including the average propensity score. Doing simulations, Penning de Vries and Groenwold find that the first approach produces less biased and more precise estimates. Consequently, they recommend using this approach—an advice that we followed in our analysis.

[19] $SB = (Mean_{Treated} - Mean_{Control})/SD_{Treated}$; $VR = SD^2_{Treated}/SD^2_{Control}$; experts have not yet settled on a SB threshold and some recommend a less strict 0.2-threshold (Harder et al., 2010; Kainz et al., 2017).

than 0.2 (see Table 4 for a selection of focal covariates).[20] Across all covariates, the average absolute SB difference was 0.18 (median: 0.12) and the maximum bias was 1.02. In addition, the average of the VRs of the 19 continuous covariates was 1.75 (median: 1.49). Only 3 of the 20 continuous covariates exceeded the VR threshold of 2, including the SRD versatility index (2.98).

For PADS+, *optimal matching*[21] with a ratio of 1:3 without replacement on the linear propensity score estimated via BART resulted in the best covariate balance. This procedure led to adjusted groups of 37 treated and 111 control cases. For this adjusted sample, mean differences and VRs of the covariates declined strongly. The mean and median of the SB statistics decreased to 0.05, and only 16 covariates exceeded the threshold of 0.1 (only one variable had a bias larger than 0.2). VRs declined to 1.47 on average (median: 1.28) and three covariates had a ratio larger than 2. According to the most stringent thresholds, remaining imbalances indicate that in the adjusted sample treated individuals showed still a slightly different delinquency pattern, were slightly more involved with the legal system (antisocial behavior order, ASBO; youth offending teams, YOT), perceived the risk of consequences when caught somewhat lower, reported less deviant peers, a less supporting family environment, and more informal social control in their neighborhood (see Supp. material S3). Overall, however, the matching procedure decreased the likelihood that differences in pretreatment characteristics confound the ATT estimates.

Table 4 Covariate balance statistics for PADS+

PADS+	Unadjusted sample		Adjusted sample	
	SB	VR	SB	VR
Covariates–lagged (T1) outcomes				
SRD violence prevalence	.24		−.07	
SRD property prevalence	.33		.00	
SRD vandalism prevalence	.37		.04	
SRD general prevalence	.42		.06	
SRD versatility	.78	2.98	.15	1.25
Official contact prevalence	.13		.03	
Global covariate balance statistics				
Mean (absolute)	.18	1.75	.05	1.47
Median (absolute)	.12	1.49	.05	1.28
Maximum (absolute)	1.02	4.98	.29	3.03

SB = standardized bias; *VR* = variance ratio

Note: VRs are standardized in a way that they are always larger than 1, so that only ratios above 2 indicate balance problems; because prevalence covariates are binary, we report raw percentage differences and no VR statistics for them (Kainz et al., 2017)

[20] For balance statistics of all covariates, see Supp. material S3.

[21] The procedure matches treated and untreated individuals by minimizing a global distance measure (Hansen, 2004). A ratio of 1:3 indicates that three individuals without official contact (i.e., from the control group) were matched to one treated individual.

CrimoC

In CrimoC's unadjusted sample, covariate differences between treated and untreated individuals were much less pronounced, though still remarkable. The mean of the SB statistics across covariates was already quite low (0.07; median: 0.04), and only 28 of the 57 covariates had a bias greater than 0.1 (only 8 covariates exceeded a threshold of 0.2); the maximum standardized mean difference was 0.35 (see Table 5 for a selection of focal covariates).[22] VRs were with few exceptions within an acceptable threshold.

Weighting by the odds[23] on the covariate balancing propensity score (CBPS) resulted in the best-balanced distribution of covariates across the treatment and control group. After weighting, the CrimoC sample included an adjusted number of 205.8 control and 88 treated units. For this adjusted sample, imbalances in covariates diminished completely. SB statistics of all variables were below 0.1. Mean and median bias was essentially zero (<0.01). Additionally, VRs of the 23 continuous variables were also all below a value of 2 and their mean (1.19; median: 1.10) was pleasingly low, too. For CrimoC, we can actually assume that it is very likely that our weighting procedure is capable of preventing potential selection bias due to observed covariates.

Table 5 Covariate balance statistics for CrimoC

CrimoC	Unadjusted sample		Adjusted sample	
	SB	VR	SB	VR
Covariates–lagged (T1) outcomes				
SRD violence prevalence	.10		−.05	
SRD property prevalence	.14		.06	
SRD vandalism prevalence	.07		.01	
SRD general prevalence	.14		−.01	
SRD versatility	.24	1.08	.01	1.12
Official contact prevalence	.10		.00	
Global covariate balance statistics				
Mean (absolute)	.07	1.70	.00	1.19
Median (absolute)	.04	1.25	.00	1.10
Maximum (absolute)	.35	7.35	.06	1.89

SB = standardized bias; *VR* = variance ratio

VRs are standardized in a way that they are always larger than 1, so that only ratios above 2 indicate balance problems; because prevalence covariates are binary, we report raw percentage differences and no VR statistics for them (Kainz et al., 2017)

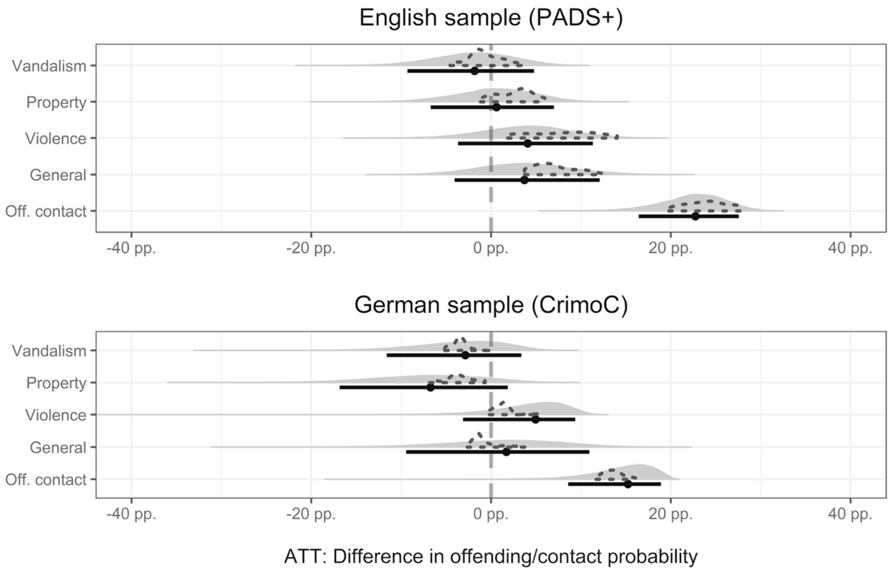
[22] For balance statistics of all covariates, see Supp. material S3.

[23] The procedure weighs the control group up to the treatment group by giving more weight to control individuals who are more similar to treated individuals on the propensity score and less weight to those more dissimilar (Harder et al., 2010).

Average Treatment Effects on the Treated

ATT estimates for the Peterborough and Duisburg samples tell a quite similar story. Most estimates are statistically insignificant, suggesting that a contact with the juvenile justice system had at best weak effects on the prevalence and versatility of reoffending (for prevalence ATT estimates, see black points and lines and gray shaded area in Fig. 2). According to the ATT point estimates, the *prevalence* of reoffending typically would have changed by less than 5 percentage points (pp.) had offenders with a system contact instead had no contact (see section *Analytical Procedure* for a definition of the ATT).

For example, among PADS+ juveniles, an official contact decreased the prevalence of committing vandalism in T3 on average about 2 pp. (ATT = −1.8 pp. [89%-CI[24]: −9.3 pp. 4.8 pp.]), whereas the reduction was estimated to be about 3 pp. (ATT = −2.9 pp. [−11.6 pp. 3.4 pp.]) among the Duisburg youths. The probability of property offending decreased slightly but insignificantly in the German sample (ATT = −6.7 pp. [−16.9 pp.



Note:
pp. = Percentage points
Gray shaded area = Distribution of ATT simulations of best-balancing model;
Black dots = Medians of ATT simulations of best-balancing model;
Black lines = 89% confidence intervals of best-balancing model;
Dotted black lines = Distribution of medians of the ATT simulations of all candidate models (sensitivity checks)

Fig. 2 Average treatment effects on the treated (ATTs), prevalence rates, Peterborough (PADS+), and Duisburg (CrimoC)

[24] Because the typical p value threshold of 0.05 and confidence interval width of 95% is chosen arbitrarily as a cut-off point in declaring certainty/uncertainty in estimation (McElreath, 2016), we decided against their use. We instead show the full estimation uncertainty by displaying density plots of the ATT simulations (Fig. 2, gray shaded areas), supplementing them with 89% confidence intervals (Fig. 2, black lines) to avoid the use of the typical cut-off points. The simulations, hereby, approximate the ATTs full probability distributions (King et al., 2000).

1.9 pp.]), whereas the effect of a system contact on property offending was estimated to be close to null in the English sample (ATT = 0.6 pp. [−6.7 pp. 7.0 pp.]). The likelihood of violent and general offending was somewhat—but again insignificantly—increased due to a system contact in both samples ($\text{ATT}_{\text{PADS+.Violence}}$ = 4.1 pp. [−3.7 pp. 11.3 pp.]; ATT $_{\text{PADS+.General}}$ = 3.7 pp. [−4.1 pp. 12.1 pp.]; $\text{ATT}_{\text{CrimoC.Violence}}$ = 4.9 pp. [−3.1 pp. 9.4 pp.]; $\text{ATT}_{\text{CrimoC.General}}$ = 1.7 pp. [−9.5 pp. 10.9 pp.]). The versatility of offending, finally, was barely affected by an intervention of the juvenile justice system. The insignificant ATT estimates indicate that an official contact had probably negligible or only relatively weak effects on the offending variety of adolescents in Peterborough (ATT = 0.04 [−0.24 0.28]) and Duisburg (ATT = −0.09 [−0.38 0.08]).

Despite these at best rather weak control effects on subsequent delinquency, the ATT estimates suggest that an official contact increased the prevalence of a renewed system contact substantially in the follow-up year. While in PADS+, the prevalence of a repeated contact rose by some 23 pp. (ATT = 22.7 pp. [16.4 pp. 27.6 pp.]) due to a prior official contact, the increase was still about 15 pp. in CrimoC (ATT = 15.2 pp. [8.6 pp. 18.9 pp.]).

Sensitivity of ATT Estimates to Modeling Approach

To compute the ATTs, we applied not only the reported methods (that best balanced the covariates) but tried several different method combinations (varying in the imputation, propensity score, matching, and/or regression procedure). Among these combinations, only those were selected for ATT robustness checks that balanced the covariate distributions well. For each outcome and each of these 36 (PADS+) or 60 (CrimoC) "candidate" method combinations, we computed ATT point estimates. The distribution of all point estimates was then plotted in density plots (see dotted lines in Fig. 2). Overall, the density plots suggest that the ATT estimates are relatively robust to changes in the analytical procedure. However, ATT estimates are somewhat more model sensitive in the English than the German sample, probably because of PADS+'s smaller sample size and stronger imbalance before matching. This is especially true for the general and violent offending prevalences as well as for the SRD versatility index. For these three outcome variables, most alternative method combinations produced ATT estimates that indicated somewhat more substantial (and in some cases significant) system contact effects than those reported above.[25]

Discussion and Conclusion

Do criminal justice interventions promote or prevent young offenders' future offending? And, do criminal justice interventions promote future formal interventions? These are the main questions addressed in this research. Although it is commonly assumed that increases in young people's offending after criminal justice contacts is evidence of some form of labeling and that decreases in their offending after such contacts is evidence of deterrent effects, the interpretation of these relationships is clearly not as simple as that (see *Theoretical Framework and Previous Research*).

What is studied here are short-term effects of (previous year) criminal justice interventions (mainly diversion measures like cautions, community work; some

[25] For more detailed information about how robust the ATTs are, see Supp. material S4.

convictions) on future (next year) offending and criminal justice interventions controlling for selected key background factors through propensity score matching (including previous frequency of delinquent behavior). Most initial criminal justice contacts are first-time criminal justice interventions. The study does not explore (and therefore does not exclude) whether *repeated* official criminal justice contacts (or the extent of such contacts) tend to gradually promote or prevent an offender's future offending. In sum, the study led to three key results:

1. The findings do not support any stronger effect of criminal justice contacts on future (next year) offending and, hence, do not support any consistent unidirectional labeling (amplification) or preventive (deterrent or treatment) effect by criminal justice contacts on the future level of young people's offending.
2. The findings support an increased likelihood of future police contacts for those who already have had a (past) police contact.
3. The findings are remarkably similar in the studied UK and German cities of Peterborough and Duisburg.

The fact that there is no consistent unidirectional association between a criminal justice contact and future offending (*finding 1*) does not exclude the possibility that this finding may mask the existence of deterrent, treatment, and labeling effects canceling each other out (i.e., for some people, criminal justice contacts may promote, and for others, reduce their future offending). What the findings indicate though, is that there is no evidence of (or room for) any strong consistent unidirectional impact of at least deterrence or labeling on the participants' future offending. If there are any effects of criminal justice interventions on future offending among our study populations, they must be differential and, if so, may depend on things such as individual differences in how people react to a specific intervention, for example, based on their personality, their experience of previous criminal justice contacts, or the content of the intervention in itself and its social context. Exploring any potential duality of effects (i.e., the existence of labeling, and deterrent effects), and, if so, what determines *which effect appears for whom in what context* (see Sherman, 1993; Cullen & Jonson, 2014) should be a priority for future studies into the effects of criminal justice interventions.

Fortunately, two previous studies with the data at hand already provide some insights into these questions. In the first, Kaiser (2022) examined whether official contact affected some mediating factors proposed by deterrence and labeling theory (personal morals, deviant peer associations, risk perceptions). Overall, no (in the German study) or at best weak (in the English study) effects on these (intermediate) factors could be found, indicating that criminal justice interventions may trigger only weak crime-relevant processes and providing little support for stronger labeling and deterrent effects canceling each other out. In the second study, Kaiser et al. (2022b) found that the impact of formal controls differs depending on offenders' personal morals, suggesting that the effects of criminal justice interventions may indeed be differential (for self-control as moderating factor, see also Schulz, 2014; Thomas et al., 2013).

The fact that a criminal justice contact has no impact on self-reported offending but increases the risk of a future criminal justice contact (*finding 2*) is highly interesting. This may also be consequential for the interpretation of research findings in this area of study. It is similar to the finding of Liberman et al. who found a "considerably larger effect on arrest than on SRO [self-reported offending]" (2014, p. 363; see also Beardslee et al., 2019). One possible explanation is that those already known to the police are more likely to be apprehended for future crimes because they are on the radar of police (see Beardslee et al., 2019). Liberman et al. call the process that leads to an increased probability of being arrested after having been arrested in the past "secondary sanctioning." They speculate that this may be due to "increased scrutiny of the individual's future behavior, by police as well as other actors such as teachers and school staff, as well as from reduced tolerance by police and other actors of an arrestee's future transgressions" (2014, p. 363). "Being on the radar" of the police could also explain why the "secondary sanctioning" effects in the current study seem to be somewhat larger in the English than in the German sample. In the study period, the English police was mandated to search actively for juvenile offending, while the German police mainly reacted to reported crimes (Bateman, 2015; Eisenberg & Kölbel, 2017; Morgan & Newburn, 2012).

Based on the presumption that control interventions are usually initiated by a specific delinquent behavior, one can conclude from our findings that such secondary control effects are *auto-dynamic*: The posterior event, the second control intervention, is generated by an essentially same anterior event, the first control intervention. Such an institutional-decision-on-institutional-decision impact is different from a *causal* institutional-decision-on-individual-behavior-effect as originally stated in labeling theory. The latter one is a causal (and not auto-dynamic) effect because here the posterior event (the individual's delinquent behavior) is generated by an essentially different (i.e., extrinsic)[26] anterior event (the control intervention). Overall, it may appear that such an auto-dynamic effect might best be understood in the light of an assumption of self-reference as suggested in systems theory (see Luhmann, 1995): A social control system reproduces itself by referring to its own prior control decisions, filed in the institutional memory of police and court registers.

However, since most juvenile crimes are of a less grave nature, and unlikely to engage investigative resources of the police, an increased detection-by-investigation risk may be a less plausible reason (with the exception of drug-related and traffic crimes, for example, police activity is rarely a main source of detection of crime and identification of offenders). Another possible explanation is that there is some unmeasured qualitative difference in the general seriousness of the crimes committed between those who already have an official contact and other offenders (i.e., those apprehended and processed by the police may generally commit more serious crimes). Our data do not differentiate between the seriousness of the crimes of the same kind. For example, some assaults could involve quite minor harms, while others could involve more severe injuries and, therefore, are taken more seriously by

[26] Following Bunge (1959, p. 197), "extrinsic determination" marks the crucial difference between a causal and an auto-dynamic effect.

victims and bystanders (witnesses) and the police, increasing the risk of the crime being reported and that identified offenders are being formally processed. Crimes that become known to the police are overwhelmingly reported by the general public, as is the identification of possible suspects.

The fact that the findings are almost identical in the studied English and German cities (*finding 3*), and that they tally well with other research in Western countries, indicates that the results may reflect a more general phenomenon: criminal justice interventions appear to have some smaller effect on future offending than on future criminal justice contacts (Beardslee et al., 2019; Klein, 1986; Liberman et al., 2014; Lopes et al., 2012).

A *limitation* of our analysis is that we cannot formally test the cross-national differences in the effects of justice contacts. This is because the measures of self-reported delinquency were not initially developed for comparison and thus differ too much between the English and German samples to construct a joint data set and analyze them within a single analysis (see Kaiser et al., 2018). Consequently, our results are affected by two sources of unobserved heterogeneity: first, differences in the measures of delinquent behavior and, second, differences in the experiences of English and German offenders due to (unmodeled or unobserved) differences in the juvenile justice systems (such as being treated differently by police). Against the background of this heterogeneity, the similarity of results across both samples may be seen as even more remarkable and imply that our findings may be quite robust.

Another shortcoming of our study, as is true for all research that cannot rely on random assignment within an experimental design, is that it may be biased by selection effects. Individual differences (e.g., in criminal propensity) may explain both the official contact and subsequent offending (or re-contact). Not accounting for such confounding factors may bias treatment effect estimates. Applying propensity score matching, we tried to counteract confounding by balancing groups of treated and untreated individuals in terms of previously observed characteristics. Mimicking a randomized experiment, this technique is still incapable of balancing unobserved factors. Furthermore, it only prevents bias of observed confounders if these are successfully balanced across groups with and without official contact. Although our matching methods seemed to be quite successful in this respect, some observed covariates remained imbalanced in the English sample. To prevent bias due to these imbalanced factors, we used the matched samples within lagged dependent variable regressions (see footnote 17), which exploit the panel nature of the data to produce (even) more robust causal estimates (Morgan & Winship, 2015). By applying these panel models on groups that were successfully balanced on many potential confounding (observed) factors and conducting various sensitivity analyses (e.g., using different sets of controls in the regression models), we think that our results are quite immune to selection bias.

It is furthermore important to note that, while official contact at T2 (age 15) was the first criminal justice contact for most of our detected juveniles (English sample: 75.7%; German sample: 81.8%), we do not restrict our study to first-time contacts. Being interested in the *average* impact of official intervention, contact could be either a novel experience or a repeated encounter with the criminal justice system. Despite this focus on an overall effect, it is reasonable to assume that the impact of criminal justice intervention depends on an individual's *prior* history with the formal control system. Liberman et al. (2014), for example, who restricted their analysis to first-time arrests, emphasized that a novel experience should have a larger impact than a repeated formal control experience according to both deterrence and labeling theory (see also Anwar & Loughran, 2011; Bernburg, 2019). Unfortunately, due to the relatively low number of participants with official contact at T2, we do not have the statistical power to properly study whether the effect of formal controls indeed depends on juveniles' sanctioning history. Some preliminary regression analyses (including a product term of official contacts at T1 and T2 as predictor) did not provide consistent patterns of whether the effects depend on the sanctioning history (see Supp. material S6), which highlights the need for future research with larger samples.

Against these limitations, we would caution against making firm general *policy recommendations* as a result of our findings. There are no strong directional and clear-cut findings as to potential labeling or deterrent effects from criminal justice interventions on future delinquent *behavior*. Our results rather suggest that if there are such effects, they may operate in different directions (i.e., both promote and prevent future offending), potentially being dependent on the people involved, their life-circumstances, stages in a criminal career and the kind of intervention and its execution. Regarding the risk of secondary control interventions for already registered offenders, it may be important for law enforcement to consider the possibility that increased interventions for those already under formal control may enlarge structural and personal obstacles for a non-delinquent development compared to offenders of a similar delinquent potential who have not been registered.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40865-023-00233-2.

Funding Open Access funding enabled and organized by Projekt DEAL. The CrimoC study was funded by a grant from the German Research Foundation (GRF grant numbers: BO1234/6–1 to 4, 9–1). The PADS+study was funded by grants from the UK Economic and Social Research Council (ESRC grant numbers: L330253002, ES/E021042/1, and ES/K010646/1).

Declarations

Conflict of Interest The authors declare no competing interests.

References

Anwar, S., & Loughran, T. A. (2011). Testing a Bayesian learning theory of deterrence among serious juvenile offenders. *Criminology, 49*(3), 667–698.

Apel, R. J., & Sweeten, G. (2010). Propensity score matching in criminology and criminal justice. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 543–562). New York: Springer.

Barrick, K. (2014). A review of prior tests of labeling theory. In D. P. Farrington & J. Murray (Eds.), *Labeling theory: Empirical tests* (pp. 89–112). New Brunswick: Transaction.

Bateman, T. (2015). Trends in detected youth crime and contemporary state responses. In B. Goldson & J. Muncie (Eds.), *Youth crime & justice* (2nd ed., pp. 67–82). London: Sage.

Beardslee, J., Miltimore, S., Fine, A., Frick, P. J., Steinberg, L., & Cauffman, E. (2019). Under the radar or under arrest: How is adolescent boys' first contact with the juvenile justice system related to future offending and arrests? *Law and Human Behavior, 43*(4), 342–357.

Beccaria, C. (1986). *[1764]).* On crimes and punishments. Indianapolis, IN: Hackett Publishing Company.

Becker, H. S. (1963). *Outsiders*. New York: Free Press.

Bentham, J. (1988). *1776. The principles of morals and legislation. Great books in philosophy series*. Buffalo, NY: Prometheus Books.

Bernburg, J. G. (2019). Labeling theory. In M. D. Krohn, N. Hendrix, G. P. Hall, & A. J. Lizotte (Eds.), *Handbook on crime and deviance* (pp. 179–196). New York: Springer.

Blumstein, A., Cohen, J., Roth, J. A. & Visher, C. A. (1986). *Criminal careers and "career criminals"* (Vol. 1). Washington, DC: National Academy Press.

Boers, K., Reinecke, J., Seddig, D., & Mariotti, L. (2010). Explaining the development of adolescent violent delinquency. *European Journal of Criminology, 7*(6), 499–520.

Bruinsma, G. (2018). Classical theory: The emergence of deterrence theory in the age of enlightenment. In D. S. Nagin, F. T. Cullen, & C. L. Jonson (Eds.), *Deterrence, choice, and crime* (pp. 3–28). New York: Routledge.

Bunge, M. (1959). *Causality: The place of the causal principle in modern science*. Cambridge: Harvard University Press.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics, 4*(1), 266–298.

Cullen, F. T., & Jonson, C. L. (2014). Labeling theory and correctional rehabilitation: Beyond unanticipated consequences. In D. P. Farrington, & J. Murray (Eds.). *Labeling theory. Empirical tests* (pp. 63–88). New York: Routledge.

Eisenberg, U., & Kölbel, R. (2017). *Kriminologie* (7th ed.). Tübingen: Mohr Siebeck.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association, 99*(467), 609–618.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*(3), 234–249.

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

Howard, J. (1777). *The state of the prisons in England and Wales*. Warrington: Eyres.

Huizinga, D., & Henry, K. L. (2008). The effect of arrest and justice system sanctions on subsequent behavior: Findings from longitudinal and other studies. In A. M. Liberman (Ed.), *The long view of crime: A synthesis of longitudinal research* (pp. 220–254). New York: Springer.

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (statistical Methodology), 76*(1), 243–263.

Imai, K., King, G., & Lau, O. (2008). Toward a common framework for statistical analysis and development. *Journal of Computational Graphics and Statistics, 17*(4), 1–22.

Kainz, K., Greifer, N., Givens, A., Swietek, K., Lombardi, B. M., Zietz, S., & Kohn, J. L. (2017). Improving causal inference: Recommendations for covariate selection and balance in propensity score methods. *Journal of the Society for Social Work and Research, 8*(2), 279–303.

Kaiser, F. (2022). *Does contact with the justice system influence Situational Action Theory's causes of crime?* A study of English and German juveniles: International Criminal Justice Review. https://doi.org/10.1177/10575677221082071.

Kaiser, F., Schaerff, M., & Boers, K. (2018). Effekte jugendstrafrechtlicher interventionen in Duisburg und Peterborough. In K. Boers & M. Schaerff (Eds.), *Kriminologische Welt in Bewegung* (pp. 344–368). Mönchengladbach: Forum Verlag Godesberg.

Kaiser, F., Huss, B., & Reinecke, J. (2022a). Revisiting the experiential effect: How criminal offending affects juveniles' perceptions of detection risk. *Journal of Developmental and Life-Course Criminology, 8*(1), 47–74.

Kaiser, F., Huss, B., & Schaerff, M. (2022b). Differential updating and morality: Do people learn differently from police detection depending on their personal morals? *European Journal of Criminology*. https://doi.org/10.1177/14773708221128515.

King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science, 44*(2), 347–361.

Kleck, G., & Sever, B. (2017). *Punishment and crime: The limits of punitive crime control*. New York: Routledge.

Klein, M. W. (1986). Labeling theory and delinquency policy: An experimental test. *Criminal Justice and Behavior, 13*(1), 47–79.

Krause, T. (1999). *Geschichte des Strafvollzugs*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Krohn, M. D., Lopes, G., & Ward, J. T. (2014). Effects of official intervention on later offending in the Rochester Youth Development Study. In D. P. Farrington & J. Murray (Eds.), *Labeling theory* (pp. 179–208). Transaction.

Lemert, E. M. (1951). *Social pathology*. Englewood Cliffs, NJ: McGraw-Hill.

Liberman, A. M., Kirk, D. S., & Kim, K. (2014). Labeling effects of first juvenile arrests: Secondary deviance and secondary sanctioning. *Criminology, 52*(3), 345–370.

Lipsey, M. W. (2009). The primary factors that characterize effective interventions with juvenile offenders: A meta-analytic overview. *Victims and Offenders, 4*(2), 124–147.

Lochner, L. (2007). Individual perceptions of the criminal justice system. *American Economic Review, 97*(1), 444–460.

Lopes, G., Krohn, M. D., Lizotte, A. J., Schmidt, N. M., Vásquez, B. E., & Bernburg, J. G. (2012). Labeling and cumulative disadvantage: The impact of formal police intervention on fife chances and crime during emerging adulthood. *Crime & Delinquency, 58*(3), 456–488.

Luhmann, N. (1995). *Social systems*. Stanford: Stanford University Press.

McAra, L., & McVie, S. (2007). Youth justice?: The impact of system contact on patterns of desistance from offending. *European Journal of Criminology, 4*(3), 315–345.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press Taylor & Francis Group.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). New York: Cambridge University Press.

Morgan, R., & Newburn, T. (2012). Youth crime and justice: Rediscovering devolution, discretion, and diversion. In M. Maguire, R. Morgan, & R. Reiner (Eds.), *The Oxford handbook of criminology* (5th ed., pp. 490–530). Oxford: Oxford University Press.

Morris, R. G., & Piquero, A. R. (2013). For whom do sanctions deter and label? *Justice Quarterly, 30*(5), 837–868.

Motz, R. T., Barnes, J. C., Caspi, A., Arseneault, L., Cullen, F. T., Houts, R., Wertz, J., & Moffitt, T. E. (2020). Does contact with the justice system deter or promote future delinquency? Results from a longitudinal study of British adolescent twins. *Criminology, 58*(2), 307–335.

Nagin, D. S., Cullen, F. T., & Jonson, C. L. (2009). Imprisonment and reoffending. In M. H. Tonry (Ed.), *Crime and justice* (Vol. 38, pp. 115–200). Chicago: University of Chicago Press.

Paternoster, R. (2018). Perceptual deterrence theory. In D. S. Nagin, F. T. Cullen, & C. L. Jonson (Eds.), *Deterrence, choice, and crime* (pp. 81–106). New York: Routledge.

Penning de Vries, B. B. L., & Groenwold, R. H. H. (2017). A comparison of two approaches to implementing propensity score methods following multiple imputation. *Epidemiology Biostatistics and Public Health, 14*(4), 12630–12630-21.

Pratt, T. C., & Turanovic, J. J. (2018). Celerity and deterrence. In D. S. Nagin, F. T. Cullen, & C. L. Jonson (Eds.), *Deterrence, choice, and crime* (pp. 187–210). New York: Routledge.

R Core Team. (2018). R: A language and environment for statistical computing. https://www.R-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.

Schulz, S. (2014). Individual differences in the deterrence process: Which individuals learn (most) from their offending experiences? *Journal of Quantitative Criminology, 30*(2), 215–236.

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

Sherman, L. W. (1993). Defiance, deterrence, and irrelevance: A theory of the criminal sanction. *Journal of Research in Crime and Delinquency, 30*(4), 445–473.

Sherman, L. W., Gottfredson, D. C., MacKenzie, D. L., Eck, J., Reuter, P., & Bushway, S. (1998). *Preventing crime: What works, what doesn't, what's promising*. College Park, MD: University of Maryland.

Smith, D. A., & Paternoster, R. (1990). Formal processing and future delinquency: Deviance amplification as selection artifact. *Law & Society Review, 24*(5), 1109–1132.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250–267.

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

Sweeten, G. (2012). Scaling criminal offending. *Journal of Quantitative Criminology, 28*(3), 533–557.

Thomas, C. W., & Bishop, D. M. (1984). The effect of formal and informal sanctions on delinquency: A longitudinal comparison of labeling and deterrence theories. *The Journal of Criminal Law and Criminology, 75*(4), 1222–1245.

Thomas, K. J., Loughran, T. A., & Piquero, A. R. (2013). Do individual characteristics explain variation in sanction risk updating among serious juvenile offenders? Advancing the logic of differential deterrence. *Law and Human Behavior, 37*(1), 10–21.

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Boca Raton, FL: CRC Press Taylor & Francis Group.

Ward, J. T., Krohn, M. D., & Gibson, C. L. (2014). The effects of police contact on trajectories of violence: A group-based, propensity score matching analysis. *Journal of Interpersonal Violence, 29*(3), 440–475.

Wikström, P.-O.H., Treiber, K., & Roman, G. (2023). *Character, circumstances and criminal careers*. Oxford: Oxford University Press (in press).

Wikström, P.-O. H., Oberwittler, D., Treiber, K., & Hardie, B. (2012). *Breaking rules: The social and situational dynamics of young people's urban crime*. Oxford: Oxford University Press.

Wiley, S. A., & Esbensen, F.-A. (2016). The effect of police contact: Does official intervention result in deviance amplification? *Crime & Delinquency, 62*(3), 283–307.

Wiley, S. A., Slocum, L. A., & Esbensen, F.-A. (2013). The unintended consequences of being stopped or arrested: An exploration of the labeling mechanisms through which police contact leads to subsequent delinquency. *Criminology, 51*(4), 927–966.

Young, C., & Holsteen, K. (2016). Model uncertainty and robustness. *Sociological Methods & Research, 46*(1), 3–40.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.