

A systematic review and meta-analysis on the effects of young offender treatment programs in Europe

Johann A. Koehler · Friedrich Lösel ·
Thomas D. Akoensi · David K. Humphreys

Published online: 11 August 2012
© Springer Science+Business Media B.V. 2012

Abstract

Objectives To examine the effectiveness of young offender rehabilitation programs in Europe as part of an international project on the transnational transfer of approaches to reducing reoffending.

Methods A literature search of approximately 27,000 titles revealed 25 controlled evaluations that fulfilled eligibility criteria, such as treatment of adjudicated young offenders below the age of 25, equivalence of treatment and control groups, and outcomes on reoffending. In total, the studies contained 7,940 offenders with a mean age of 17.9 years.

Results Outcomes in the primary studies ranged widely from odds ratio (OR)=0.58 to 6.99, and the mean effect was significant and in favor of treatment (OR=1.34). Behavioral and cognitive-behavioral treatment ranked above average (OR=1.73), whereas purely deterrent and supervisory interventions revealed a slightly negative outcome (OR=0.85). Programs that were conducted in accordance with the risk–need–responsivity principles revealed the strongest mean effect (OR=1.90), which indicates a reduction of 16 % in reoffending against a baseline of 50 %. Studies of community treatment, with small samples, high program fidelity, and conducted as part

This study was carried out within the European cooperation project “Strengthening Transnational Approaches to Reducing Reoffending”. The project was funded by the European Union. We thank our partners from the UK Ministry of Justice, London Probation Trust, European Organization for Probation, and the Ministries of Justice of Bulgaria, France and Hungary for their cooperation. We also thank primary study authors for providing us with data to assist in the meta-analysis, and the anonymous peer reviewers for insightful suggestions that contributed to the improvement of this article.

J. A. Koehler (✉) · F. Lösel · T. D. Akoensi
Institute of Criminology, University of Cambridge, Cambridge, UK
e-mail: jak51@cam.ac.uk

F. Lösel
Institute of Psychology, University of Erlangen-Nuremberg, Erlangen-Nuremberg, Germany

D. K. Humphreys
Institute of Public Health, University of Cambridge, Cambridge, UK

of a demonstration project had larger effects; high methodological rigor was related to slightly smaller outcomes. Large effect size differences between evaluations from the UK and continental Europe disappeared when controlling for other study characteristics. *Conclusions* Overall, most findings agreed with North American meta-analyses. However, two-thirds of the studies were British, and in most European countries there was no sound evaluation of young offender treatment at all. This limits the generalization of results and underlines the policy need for systematic evaluation of programs and outcome moderators across different countries.

Keywords Meta-analysis · Systematic review · Offender rehabilitation · Young offenders · Crime prevention · Recidivism

Introduction

Offender rehabilitation research and practice have made considerable progress since the claims of the 1970s that ‘nothing works’. There is now a large body of meta-analyses on ‘what works’ among a diverse array of correctional programs that aim to reduce reoffending beyond conventional criminal justice methods. These meta-analyses report positive effects for general, violent, sexual, and other offender groups (Andrews et al. 1990; Andrews and Bonta 2010; Aos et al. 2006; Hanson et al. 2002; Hollin and Palmer 2009; Lipsey 1992; Lipsey and Cullen 2007; Lipsey and Wilson 1998; Lösel 1995, 2012a; Lösel and Schmucker 2005; McGuire 2002; MacKenzie 2006; Pearson et al. 2002; Tong and Farrington 2006). In particular, these studies show relatively sound effects for cognitive behavioral treatment, structured therapeutic communities, and multimodal systems-oriented programs, whereas pure punishment, deterrence, and supervision-based interventions reveal either negligible or slightly negative outcomes.

In addition to providing support for particular types of programs, meta-analytic evidence has highlighted certain specific characteristics of appropriate treatment. Of these, the popular risk–need–responsivity model (RNR) is supported by robust empirical research (e.g., Andrews et al. 1990; Andrews and Bonta 2010; Andrews et al. 2006; Hanson et al. 2009). According to RNR, treatment should correspond to the offenders’ risk level of reoffending, address their dynamic risk factors, and match their learning styles and capabilities (sometimes divided into ‘general’ and more individualized ‘specific responsivity’). Over time, the RNR concept has been expanded to include organizational, staff, and other characteristics of evidence-supported offender treatment (e.g., Andrews 1995; Andrews et al. 2011; Andrews and Bonta 2010; Lösel 1995).

These criteria are used for program accreditation in European countries such as England and Wales, the Netherlands, Norway, Scotland, and Sweden (e.g., Maguire et al. 2010). Offender treatment programs have proliferated widely throughout Europe over the last decade, but this proliferation has not been rooted in systematic attempts to gather local evidence. When the first findings of accredited programs in England and Wales showed less positive results than North American reviews, controversies arose surrounding the optimism of ‘what works’ and the design of local evaluations (Harper and Chitty 2005; Hollin 2008; Maguire et al. 2010). Those results indicate

that offender treatment might not be easily transferrable across cultural contexts, national legal regulations, criminal justice organizations, staff resources, offender populations, and so forth. This issue is already suggested by the meta-analytic findings that the program contents explain only a limited amount of outcome variance and that many other factors play a role (Lipsey and Wilson 1998; Lösel 2012a).

Given that the vast majority of the above-mentioned findings on ‘what works’ stem from North America and use only English language material, it is unclear how easily they can be generalized to other cultures and countries. Very few systematic reviews addressed offender treatment in Europe. They focused either on specific treatments in the German-speaking world (Lösel et al. 1987; Lösel and Köferl 1989; Lösel 2000) or on rather heterogeneous types of programs and offender groups in Europe (Redondo et al. 1999; 2001). In addition, to reach substantial study samples, these reviews had to include evaluations of moderate design quality. Because of their publication date, they also could not cover more recent research.

Against this background, the present article reports an up-to-date systematic review and meta-analysis of correctional programs for young offenders in Europe. We concentrate on this target group for four reasons. Firstly, youth crime and violence has elicited heightened and immediate concern in many countries. Secondly, in light of the pervasive observation of a peak of offending among adolescents (Farrington 1986), a potential positive intervention effect in young offenders is of particular value because it may cut off long-term criminal careers. Thirdly, the focus on a specific offender population leads to a more homogeneous sample of studies and programs, thus increasing the potential for generalization. Finally, young offender treatment is the most promising field of research with regard to the number of primary studies that are suitable for a meta-analysis.

Method

Study inclusion

Eligibility criteria

1. *Region.* Our literature search targeted evaluations of correctional interventions conducted in Europe.
2. *Target population.* The samples had to comprise young offenders, defined as previously adjudicated youths up to the age of 25. In instances where the evaluation was composed of subjects falling outside of these criteria (for example, where the sample contained at-risk youth with conduct disorders who had not offended), the study was included only if at least two-thirds of the participants met the eligibility criteria. We excluded studies if the sample of participants was deemed too idiosyncratic to allow the findings to be generalized. For example, we did not include treatments that were applied exclusively to sex offenders, psychopaths, or treatments that were uniquely applicable to a dedicated sub-type of offender. Programs that were applied to other offender types, such as substance abusing youths, were eligible only if the treatments were theoretically applicable to a general young offender population.

3. *Intervention.* The intervention had to contain a circumscribed correctional program that aimed to reduce reoffending. For example, we included structured therapeutic programs such as Reasoning and Rehabilitation and Multisystemic Therapy, manualized courses that acted as a supplement to or substitute for traditional punishment methods such as restorative justice, and specially intensified punitive intervention regimes such as boot camps. Broader analyses of national-level policy changes were not included.
4. *Outcomes.* Data had to concern reoffending as an outcome, which was understood either as a formal legal measure (e.g., re-arrest, re-conviction, reincarceration, revocation of probation or parole), or as self-reported data pertaining to crime. Lower-level offenses that did not constitute crime, such as anti-social behavior, were not included.
5. *Evaluation design.* The evaluation had to compare the effect of the intervention in a treatment group with the level of reoffending in a control group. The latter could be defined as the application of no treatment, treatment as usual, or an alternative treatment. We excluded studies if the effectiveness of treatment was compared to national statistics of the general youth offender population. The control group had to show a clear indication of equivalence to the treatment group. This could have been achieved by randomization, matching procedures or statistical comparisons of equivalence. Primary study eligibility was subjected to a group discussion among the authors when the between-group equivalence was questionable, for example due to noticeable differential attrition rates or when study participants reported differences along important characteristics such as prior levels of offending.
6. *Report of effects.* Any common statistics or raw data that allowed calculation of effect sizes were eligible. When primary studies reported data in a manner that was insufficient for the purposes of calculating an effect size, we made every effort to contact the author to acquire the necessary information.
7. *Publication.* We included published and unpublished evaluations appearing between 1980 and 2009.
8. *Language.* Reports in any commonly used European language were eligible. If no member of our research team was able to read a specific language, criminology students with relevant linguistic expertise translated the document.

Literature search

In order to locate unpublished and published studies, we searched online computerized databases. The search terms and databases are shown in Appendices i and ii. In addition, we searched meta-analytic and systematic review publications dealing with juvenile offending and reoffending treatment programs as well as the references of primary studies. In an effort to acquire unpublished evaluations, we conducted a survey of European treatment programs that aim to reduce juvenile reoffending, and requested the data when those programs had been evaluated. We distributed questionnaires to experts in all 27 EU countries, as well as to academics in European countries who were not EU members. We used snowball sampling methods to locate more inaccessible respondents.

Our bibliographic database search yielded a total of 26,989 titles,¹ which, upon deletion of duplicates, yielded 21,223 discrete studies. Preliminary screening of titles and abstracts for *prima facie* relevance to our eligibility criteria yielded 14,001 studies. The abstracts were then screened in more detail according to method, location of study, sample population, and outcome of interest, in order to arrive at 88 studies (for details regarding the eligibility criteria, see above). We retrieved all of these documents in full and 22 studies met most of our inclusion criteria. However, 5 of these studies had to be excluded because they had no control group or insufficient methodological rigor, 3 did not provide adequate data to compute an effect size, and 3 addressed exceptional offender populations, thus resulting in 12 studies. A further 9 studies that had not appeared in our electronic database search were added as a result of contact with other academics and experts and hand-searching through relevant journals and bibliographies.

Our search uncovered no unpublished data. Although our European survey aimed to reduce the potential bias of simple document search (Wilson 2009), this strategy revealed clear limits. Of $n=112$ received questionnaires, 44 % reported some kind of outcome evaluation (Hamilton et al. 2011). However, of those that had been conducted, none met our criteria of substantive or methodological rigor. As a result, this strategy yielded no further contributions to our study sample. We therefore proceeded with statistical analysis to examine a potential publication bias in our data set (see “Results”).

Coding of variables

Outcome measure

According to our eligibility criteria, we used official indicators as well as self-reported reoffending. Although these measures suffer from their own unique threats to construct validity (Lloyd et al. 1994; Lösel 1995), they are nonetheless more relevant for policy than psychometric and other criteria that have also been used in a number of studies. We collected all the available outcome data in a given study, but used only the most relevant, general recidivism measure for computing an effect size (as opposed to multiple effects). Where various indicators of reoffending were available, we combined the data in a composite score. This also applied to different levels of seriousness of recidivism ($k=1$; Lösel and Pomplun 1998).

Effect size

Where studies reported the prevalence of reoffending, we calculated an odds ratio (OR), whereas mean scores were used to calculate Cohen’s d , which were then converted into OR using the formula found in Hasselblad and Hedges (1995). When necessary values were omitted in study reports, we calculated the effect size based on available p , t , or F values. Otherwise, every effort was taken to contact the author. In a few cases, we imputed effect sizes on the basis of other available data (see Ttofi et al. 2008).

¹ The full number of studies yielded within each database is catalogued in Appendix ii.

Where possible, we sought to avoid using effect size data that were adjusted for pseudo-reconvictions (that is, when court data that refers to offenses committed before the first reference offense indicate a ‘reoffense’). Pseudo-reconvictions seem to decrease the observed reoffending rate by up to 7 % (Lloyd et al. 1994). Therefore, this issue would introduce a conservative bias favoring the control condition.

Intent-to-treat data was reported in five studies (20 % of the sample). We therefore had to use treatment-as-treated outcome data in our computation of effect sizes. This could lead to biased results in favor of treatment (Lösel 1995), although there was no substantial difference in the few studies where intent-to-treat data were available (see “Results”).

Where multiple comparison groups were presented ($k=1$; Raynor and Vanstone 1997), we used the data from the group that most closely approximated the normal criminal justice course without the respective intervention. Consequently, all the study comparisons in our meta-analysis applied treatment as usual in the control condition.

Follow-up

We included data on all the follow-up periods provided within each study. Where data were incomplete, we contacted authors to acquire the necessary information. Only two studies (Curran et al. 1995; Lösel and Pomplun 1998) reported follow-up data for more than 2 years after the end of the intervention. The preponderance of relatively short follow-up periods may lead to a positive bias; however, a recent study in England and Wales revealed that approximately two-thirds of offenders discharged from custody reoffended within 24 months (Ministry of Justice 2010; see also Tournier and Barre 1990 for other European countries).

Intervention type

Due to the considerable diversity of treatments, we had to compromise between creating treatment subcategories that were perhaps too broad on one hand, against the competing concern of low statistical power because of too few studies in each category on the other. We ultimately settled on three categories of treatment type: (1) ‘Cognitive–Behavioral and Behavioral’ treatment (e.g., thinking skills programs, social skills and problem solving approaches, reinforcement of behavioral change); (2) ‘Intensive Supervision and Deterrence-Based’ interventions (e.g., amplified sanctions, boot camps without educational/therapeutic elements, and purely control-based supervision), and (3) ‘Non-Behavioral’ treatment. This last category included a mix of program types, ranging from educational and vocational skills training to programs such as mentoring, restorative justice, and intensive probation support. We coded these interventions separately to explore potential heterogeneity within that treatment category.

Studies were also coded based on whether the intervention was conducted in a community or a secure setting, whether the evaluation was conducted in the United Kingdom or elsewhere, and whether participation was voluntary or obligatory.

Risk, needs, and specific responsivity

Because the studies did not contain detailed within-sample data on risk, we applied the aggregate sample approach to coding risk (e.g., Dowden and Andrews 2003). The coding was based on data concerning prior contact with the criminal justice system and the primary study authors' own descriptions of sample risk. As no 'Low risk' study was observed, risk was coded on a three-level item, indicating 'Medium', 'Medium-High', and 'High' risk.

The studies were also coded according to two three-level items: the first indicated the degree to which treatments were designed to address specific criminogenic needs, as established in the rehabilitation program literature (Andrews and Bonta 2010; Hollin 2002), and the second item measured adherence to specific responsivity, or the degree to which program implementers adapted the treatment delivery to the unique learning styles and capabilities of the offender. Both items were coded using the categories 'Low', 'Moderate', and 'High'. All three RNR variables were judged and scored after each study was subjected to intensive discussion among the authors.

Methodological rigor

Common methodological features such as the type of research design, the total sample size, the length of follow-up measurement, and the level of attrition were coded for each study. To avoid too many subcategories with only few studies, we created dichotomous variables. For sample size, a total n of 100 provided the cut-off between 'Large' and 'Small'. Follow-up measurement was dichotomized into 12 months and less as 'Short' and anything above as 'Long'.

Research design was coded as a three-level variable, in which 'Moderate Control' denoted a quasi-experimental post-hoc formation of the comparison group that contained data suggesting overall similarity between treated and untreated groups. 'Strong Statistical Control' denoted that the two groups were matched according to key variables, and 'Randomized Experiment' denoted the randomization of participant allocation to experimental groups (without obvious selective dropouts or other threats to validity).

In order to compare categories with similar numbers of studies, we designated 15 % as the cutoff threshold for 'High' and 'Low' rates of attrition. Where insufficient data was reported on participant levels of dropout, we coded the study as 'Not Reported'. We also collected data regarding whether the evaluation had been conducted as part of a demonstration or a routine practice project. Finally, we coded the level of fidelity to the program objective on a two-level scale, using the categories 'Low' and 'High'. This data was gleaned from primary study author's reports of how closely the treatment delivery approximated the proposed treatment design or whether problems of integrity became apparent. All qualitative variables were independently coded by the first and third author and inter-rater reliability was high (average $kappa=.775$, $p<.005$). In cases of discrepancies, the final coding was made after a thorough review including all the authors.

Results

Description of the studies

Our sample of primary studies comprised 21 evaluations with 25 discrete comparisons between treatment and control groups (in the following, named “studies”). Table 1 outlines key descriptive sample features. 87 % of the studies have been published since 1990 and 52 % appeared since 2000, thus indicating relatively timely interventions. The majority of evaluations (64 %) were conducted in the United Kingdom and the rest came from only four other European countries.

The sample consisted of a total n of 7,940 young offenders, with 3,883 in the treatment groups, and 4,057 subjects in the control groups. In two-thirds of the studies, more than 90 % of the participants were male. The average ages ranged from 14 to 23.7, with a mean of 17.9 years. The average recidivism rate across all the studies’ comparison groups was 50.5 %.

The intervention types were diverse: they included cognitive–behavioral methods (e.g., Cann et al. 2005; Ogden and Hagen 2006; Ogden et al. 2007), mentoring programs (Newburn and Shiner 2005; St. James-Roberts et al. 2005), traditional ‘boot camp’-style programs (Farrington et al. 2002), self-help manuals (McMurrin and Boyle 1990), restorative justice (Shapland et al. 2008), and intensive supervision (Bottoms 1995), among others.

Overall effect size

Figure 1 contains the effects and confidence intervals in the primary studies. The ORs ranged from 0.58 to 6.98, with 15 studies showing an effect in favor of the treatment group, and 9 pointed in the opposite direction. Six effects were statistically significant ($p < .05$) and all in favor of treatment.

The overall mean effect size for the sample was $OR = 1.34$ ($p < .05$; see Fig. 1), which translates to a Pearson’s r of .08 or Cohen’s d of .16. This effect size corresponds to 43 % rate of recidivism in the treatment groups, assuming a base rate of 50 % recidivism in the control groups.² The average effect displayed considerable heterogeneity [$Q_{\text{Total}}(24) = 64.48$; $p < .001$]. Therefore, the meta-analytic results that follow are all derived from random-effects models. Although Borenstein et al. (2009:84–5) caution against choosing between fixed- and random-effects models purely on the basis of the Q statistic (largely due to the rarity with which it attains adequate power), the use of a random-effects model is further justified because our distribution of effect sizes from a few European countries is not attributable to mere sampling error alone. Although small studies may be over-weighted in the random effects model, the mean effect was not much lower and still highly significant when

² Assuming a recidivism rate of 50 % in the control groups, the recidivism rate in the treatment groups can be calculated using the following formula for OR effect sizes:

$$\text{Treatment group recidivism rate} = \frac{OR}{(1 + OR)}$$

We gratefully acknowledge the editors for providing this helpful formula.

Table 1 Study features

Study characteristic		<i>k</i>	Percentage
General			
Publication year	1980–1989	4	16
	1990–1999	8	32
	2000–2009	13	52
Country	Germany	2	8
	Netherlands	4	16
	Norway	2	8
	Sweden	1	4
	United Kingdom	16	64
Sample			
Mean age (years)	≤15	8	32
	16 - 20	14	56
	21 - 25	3	12
Male (proportion)	≤90 %	8	32
	>90 %	11	44
	Not reported	6	24
Treatment			
Treatment type	Behavioral/cognitive-behavioral	11	44
	Non-behavioral	10	40
	Intensive supervision/ deterrence	4	16
Setting	Community	17	68
	Custody	8	32
Participation type	Voluntary	14	56
	Obligatory	11	44
Risk	Medium	12	48
	Medium-high	3	12
	High	10	40
Needs	Low	9	36
	Moderate	6	24
	High	10	40
Specific responsivity	Low	7	28
	Moderate	8	32
	High	10	40
Methodology			
Total sample size	≤100	10	40
	>100	15	60
Group allocation	Moderate control	13	52
	Strong statistical control	8	32
	Randomized experiment	4	16
Follow-Up ^a	≤12 months	17	68

Table 1 (continued)

Study characteristic		<i>k</i>	Percentage
Dropout rate	>12 months	13	52
	≤15 %	8	32
	>15 %	7	28
	Not reported	11	44
Research type	Demonstration	7	28
	Practice	18	72
Program fidelity	Low	14	56
	High	11	44

^a The percentages do not sum to 100, as studies may report multiple follow-up periods

the fixed-effects model was applied ($OR=1.19$, $p<.001$). The heterogeneity of outcomes justified further investigation in a moderator analysis.

Moderator analyses

Treatment characteristics

Table 2 contains the results on the differential effectiveness of the various treatment types and other intervention characteristics. Due to the small number of studies in the

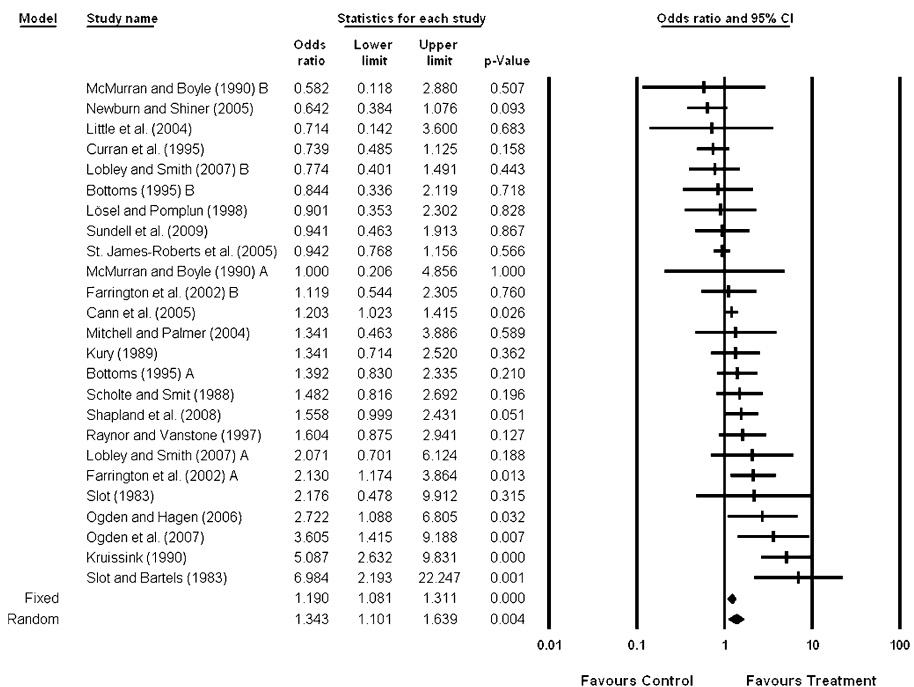
**Fig. 1** Effectiveness of young offender treatment programs

Table 2 Relationship of effect size to treatment moderators

Moderator		<i>k</i>	OR	CI ₉₅ %		<i>Q</i> _{Between}	% Reoffending reduction
				Lower	Upper		
Treatment							
Treatment type						5.55	
	Behavioral/Cognitive-behavioral	11	1.73***	1.26	2.36		13
	Intensive supervision/deterrence	4	0.85	0.50	1.46		−4
	Non-behavioral	10	1.23	0.89	1.69		5
						2.00 ^a	
	Vocational training/education	4	1.69	0.81	3.52		13
	Restorative justice	1	1.56	0.54	4.50		11
	Guidance/counseling	3	1.24	0.62	2.49		5
	Mentoring	2	0.79	0.38	1.65		−6
Fidelity						1.87	
	Low	14	1.21	0.94	1.54		5
	High	11	1.60**	1.16	2.20		12
Setting						1.11	
	Community	17	1.48**	1.13	1.93		10
	Custody	8	1.16	0.80	1.67		4
Recruitment						0.01	
	Obligatory	14	1.37*	1.01	1.85		8
	Voluntary	11	1.35*	1.00	1.81		7
RNR principle							
Risk						2.55	
	Medium	12	1.14	0.84	1.54		3
	Medium-high	3	1.58	0.88	2.82		11
	High	10	1.63**	1.14	2.34		12
Needs						2.00	
	Low	9	1.09	0.74	1.61		2
	Moderate	6	1.41	0.93	2.14		9
	High	10	1.59*	1.11	2.27		11
Specific responsivity						4.63	
	Low	7	0.93	0.62	1.40		−2
	Moderate	8	1.37	0.99	1.88		8
	High	10	1.64**	1.20	2.24		12
Composite							
Adherence to RNR						4.12	
	Low	11	1.13	0.84	1.52		3
	Medium	7	1.34	0.90	1.99		7
	High	7	1.90**	1.27	2.85		16

k Number of comparisons; *OR* odds ratio; *CI*₉₅ % 95 % confidence interval; *Q*_{Between} test of between-group differences (χ^2 with *df*=number of categories -1); estimated % reduction of reoffending indicates treatment group improvement, assuming a 50 % base rate of reoffending in the control groups; a negative sign indicates an *increase* in reoffending

^a Subgroup analysis of the Non-Behavioral programs alone

p*<0.05; *p*<0.01; ****p*<0.001

respective subcategories, most of the differences between them were not statistically significant. However, some heterogeneity tests were only slightly above the significance threshold, in particular for the general treatment type [$Q_{\text{Between}(2)}=5.55$; $p=.06$], adherence to the Responsivity principle [$Q_{\text{Between}(2)}=4.63$, $p=.11$], and the overall RNR model [$Q_{\text{Between}(2)}=4.12$, $p=.13$]. As we used two-sided tests, and most moderators contained theoretically meaningful and significant effects in specific subcategories, we present the respective findings in Table 2.

Behavioral and cognitive-behavioral (CBT) treatments reported the largest OR of 1.73 ($p<.001$), which corresponded to a 13 % reduction in recidivism in the treatment group, compared to the control group. Non-behavioral treatments reported a smaller, statistically non-significant mean effect, although the direction was still in favor of treatment (OR=1.23, $p=.20$). Intensive supervision and deterrence-based treatments reported non-significant *criminogenic* effects, favoring the control condition (OR=0.85, ns).

With the exception of mentoring programs, the effects of the various non-behavioral intervention types were roughly similar. Educational and vocational training programs were the most promising (OR=1.69; $p=.16$); restorative justice programs and guidance and counselling programs also revealed positive, but non-significant effects. The two studies on mentoring programs suggest nonsignificant undesirable effects.

Programs that were delivered at a 'High' level of program fidelity revealed a significant effect in favor of treatment (OR=1.60, $p<.05$), whereas 'Low' fidelity showed only a positive tendency. Programs that were conducted in community settings significantly reduced recidivism (OR=1.48, $p<.005$), and custodial programs also reported a positive effect, although this was non-significant (OR=1.16, $p=.44$). Treatments where participation was voluntary as well as mandatory programs showed significant positive effects. The difference between both framing conditions was negligible.

In comparison to medium risk, treatment of medium-high and high risk participants displayed more favorable effects and the OR of 1.63 for the latter group was significant. A similar systematic trend existed with regard to criminogenic needs. Programs that were 'Moderate' in this category revealed a marginally significant positive effect (OR=1.41, $p=.10$), and those that were 'High' reported a significant positive effect (OR=1.59, $p<.05$). The same pattern was observed for specific responsivity: treatments displaying a 'Low' level of responsivity were not effective, whereas programs that were 'Moderate' or 'High' in this category showed nearly or in fact significantly positive outcomes (for the latter OR=1.64; $p<.05$).

The above findings already showed that programs were most effective when they addressed high-risk offenders, targeted multiple criminogenic needs, and followed the principle of specific responsivity. To assess the *combined* level of adherence to the RNR model, we created a composite variable based on each study's score on the three components. According to Rosenthal (1991), this can be done by simply adding the constituent scores when they have similar standard deviations (maximal ratio <1.5 when there are very few variables). In our dataset this ratio was 1.45; we therefore proceeded with this method. The resulting scores were divided into three levels of treatment adherence to RNR: 'Low' (3–5), 'Medium' (6–7), and 'High' (8–9).

Although the difference between these three categories was slightly above statistical significance there was a clear suggestion of larger effects in programs that adhered more to the RNR model. The effect for 'High' RNR was OR=1.90 ($p<.005$) and indicated a 16 % reduction in recidivism in the treatment group.

Table 3 contains the moderator analyses for various methodological characteristics. Again, the small number of studies in the respective subcategories provided

Moderator		<i>k</i>	OR	CI ₉₅ %		<i>Q</i> _{Between}	% Reoffending reduction
				Lower	Upper		
Research type							
Group allocation						0.10	
	Moderate control	13	1.38*	1.04	1.85		8
	Strong statistical control	8	1.29	0.84	1.84		6
	Randomized experiment	4	1.43	0.70	2.68		9
Dropout rate						2.81	
	≤15 %	8	1.74*	1.09	2.79		14
	>15 %	7	1.51*	1.02	2.25		10
	Not reported	11	1.10	0.77	1.56		2
Research type						1.76	
	Demonstration	7	1.65**	1.15	2.37		12
	Practice	18	1.23	0.98	1.55		5
Total sample size						2.98	
	≤100	10	1.86**	1.22	2.84		15
	>100	15	1.23	0.99	1.52		5
Follow-up						0.10	
	≤12 months	17	1.38*	1.08	1.77		8
	>12 months	13	1.28	0.90	1.84		6
Composite							
Methodological quality						0.48	
	Low	4	1.49	0.86	2.58		10
	Medium	14	1.35	0.98	1.85		7
	High	7	1.17	0.75	1.83		4
Research origin							
Country of evaluation						11.81***	
	United Kingdom	16	1.11	0.91	1.36		3
	Non-United Kingdom	9	2.17***	1.57	3.00		18

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

insufficient power for clear significances. However, there were meaningful effects in some subcategories that could not simply be attributed to fishing for significances.

The amount of control (internal validity) in the design revealed no significant difference [$Q_{\text{Between}(2)}=0.10$, $p=.95$] and also a minor, nonlinear variation in the means (from $OR=1.38$ over 1.29 to 1.43). Overall, the differences in quality of research designs and also the use of an RCT were not related to effect size.

Studies reporting low levels of attrition observed a slightly better outcome (OR of 1.76; $p<.05$) than studies with more dropouts; however, both effects were positive and statistically significant. More positive effects were observed among studies that were conducted as part of a demonstration project ($OR=1.65$, $p<.05$) than studies that were evaluations of extant practice, although the latter were also positive ($OR=1.23$, $p=.07$). There was a nearly significant trend [$Q_{\text{Between}(1)}=2.98$; $p=.08$] of larger effect sizes in smaller samples ($OR=1.86$, $p<.05$), but larger studies also showed positive outcomes. Effect sizes were only slightly larger among studies measuring recidivism at short periods of follow-up ($OR=1.38$, $p<.05$, vs. OR of 1.28, $p=.17$ for longer follow up).

In an effort to synthesize the results from the moderator analyses of methodological variables, we computed a composite score by adding together each study's score on all the methodological moderators. Studies were given a score of 1 for each of the following values: large sample size, longer follow-up measurement, low attrition, high treatment fidelity, and strong statistical control between experimental groups. A randomized experiment was accorded a further point, to signify its particular internal validity. Furthermore, studies that did not report the level of attrition in their sample were given a score of 0.5 on this item. The ratio of the largest to the smallest standard deviation in the variables was 1.92, which is defensible because of the number of variables being combined. The composite score was divided into three categories of overall methodological quality: 'Low' (0 to 1), 'Medium' (1.5 to 3), and 'High' (3.5 and above).

Overall there was no significant difference between the three subcategories [$Q_{\text{Between}(2)}=0.48$, $p=.79$] and also no significant effect for any of them. One can mention only a very slight suggestion of smaller effects in methodologically stronger studies.

National context

Evaluations conducted within the United Kingdom reported smaller positive effects on average ($OR=1.11$, $p>.05$) than evaluations conducted outside of the United Kingdom, which were highly positive ($OR=2.17$, $p<.001$). This difference was highly significant [$Q_{\text{Between}(1)}=11.81$; $p<.001$].

Because of the very strong effect of the national context, we conducted a maximum likelihood random-effects meta-regression to control for the potential effect of confounded moderators in our observation of significant differences between studies conducted within and outside the United Kingdom.³ We controlled sequentially for

³ After examining the properties of the data for potential multicollinearity and the existence of outliers, we determined that none of the standard assumptions had been violated and meta-regression could proceed.

proportions of outcome variance attributable to treatment, methodology, and context factors in order to isolate the influence of the country from which the research originated. The list of included independent variables in the model was parsimonious due to the small number of studies in our sample.

The final model accounted for 60 % of the total outcome variance (see Table 4). Model 1, which tested whether the program was cognitive-behavioral or behavioral, accounted for 34 % of the variance. In Model 2, we added the study's sample size, which accounted for a further 6 % of the variance in outcome effects. The additional variance explained by including variables related to whether the program was delivered in the community or in custody (Model 3), and whether the evaluation was part of a demonstration study or routine practice (Model 4), was 9 and 11 %, respectively. After these other factors were controlled, the country from which the research originated did not explain any further variance.

Sensitivity analysis

Although we took every precaution to ensure that all eligible studies were represented in our meta-analysis, we did not find any unpublished evaluations that met our criteria for inclusion (see above). Therefore, we checked the potential influence of a publication bias by conducting a funnel plot test for bias in the study sample. This displays the logged effect sizes on the x-axis against their respective standard errors on the y-axis (see Fig. 2). Hence, studies with smaller sample sizes will concentrate at the bottom of the graph and large studies at the top. Asymmetry in the distribution of the effect sizes on either side of the vertical line would indicate a publication bias. The relative symmetry found in the funnel plot suggests that our meta-analysis was not heavily biased in any particular direction. The 'Trim and Fill' technique displays the difference in effect

Table 4 Meta-regression results

Moderator	Model 1	Model 2	Model 3	Model 4	Model 5
Cognitive behavioral/ behavioral vs. non-behavioral	.58*** (.18)	.55*** (.17)	.62*** (.15)	.80*** (.12)	.76*** (.15)
Large vs. small sample size		-.27† (.23)	-.27† (.21)	-.21 (.20)	-.19 (.21)
Community vs. custody			.33* (.16)	.50** (.12)	.47** (.14)
Demonstration vs. practice				.34** (.14)	.33* (.15)
UK vs. Non-UK evaluation					-.06 (.19)
R^2	.34	.40	.49	.60	.60
ΔR^2	.34	.06	.09	.11	.00
$Q_{\text{Total}} (df=24)$	35.16†	38.68*	48.54**	64.38***	64.38***

Standardized β (SE)

† $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

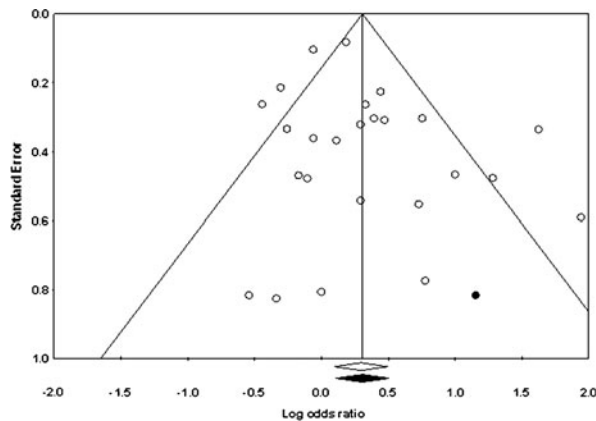


Fig. 2 Funnel plot of standard error by log odds ratio. *Solid points* represent imputed scores from the ‘Trim and Fill’ analysis

sizes that could be attributable to bias (Borenstein et al. 2009:286), by imputing effect sizes until the error distribution more closely approximates normality. The imputed effect sizes are visible on the funnel plot as solid black dots, and the solid black diamond at the bottom renders the shifted summary effect size, had the distribution been normal. In Fig. 2, the random-effects Trim and Fill analysis revealed that an imputed OR would have accounted for an increase of the overall effect size from 1.343 to 1.358. This difference is negligible and even in favor of a *lower* effect on recidivism due to treatment. This suggests that our average is a conservative estimate.

As another test of robustness, we carried out an outlier analysis. Our meta-analysis contained not only a very broad range of effect sizes (ORs from 0.560 to 6.98) but also very different sample sizes (from 26 to 3,068). Although the exclusion of certain evaluations would affect the results of the moderator analysis (where statistical power is much lower), the exclusion of outliers had negligible effects on the total effect size. For example, the exclusion of the studies with the largest sample size slightly *increased* the overall effect size and resulted in $OR=1.37$ ($p=0.008$) when Cann et al. (2005) was excluded or $OR=1.39$ ($p=0.003$) when St. James-Roberts et al. (2005) was excluded. Furthermore, had we used only the most severe reoffending measure alone, rather than a mean of all reported outcome scores, the effect size for Lösel and Pomplun (1998) would have been positive ($OR=1.77$, rather than 0.90); however, even with this analysis, our mean effect size would not have been substantially different ($OR=1.37$, $p<.005$).

Using intent-to-treat outcome data in studies where they were reported changed the summary OR from 1.34 ($p<.05$) to 1.30 ($p<.05$). This difference is small, and suggests that our overall effect may be robust.

Discussion

The mean overall effect size of $OR=1.34$ ($r=.082$; $d=.16$) is noteworthy for five reasons: first, with the exception of a subset of data in Redondo et al. (1999, 2001), it

is the only demonstration of an overall positive effect of European programs for young offenders. Second, according to our sensitivity analyses, it seems to be a relatively robust estimate of effect size. Third, it is in the same range as in other reviews of juvenile offender programs (e.g., $r=.08$ in Cleland et al. 1997; $r=.09$ in Dowden and Andrews 1999; $r=.07$ in Garrido et al. 2006; $r=.09$ in Gensheimer et al. 1986; $r=.12$ in Gottschalk et al. 1987; $r=.09$ in Latimer et al. 2003; $r=.05$ in Lipsey 1992; $r=.12$ in Redondo et al. 2001). Fourth, this small effect equates to an approximately 7 % reduction in recidivism in the treatment group, compared to the control group. Fifth, although we cannot provide sound European data on cost-benefits here, even such moderate reductions in recidivism seem to be practically significant in both a human and a monetary sense (e.g., see Aos et al. 2001; Welsh and Farrington 2001).

As in other meta-analyses on offender treatment, the mean overall effect tells only a small part of the story on the outcomes. Consistent with the literature, cognitive-behavioral and behavioral treatments showed larger effects than other types of programs. Non-behaviorally-oriented programs revealed no significant positive effect, whereas deterrence- and supervision-based interventions even resulted in slightly (although not significantly) *increased* recidivism. These findings are in accordance with North American research (Andrews and Bonta 2010; Lipsey and Cullen 2007; MacKenzie 2006). With regard to non-behavioral programs, our review could not provide differentiated results because this category was very heterogeneous and contained too few studies on specific program types. Therefore, one should only mention the non-significant positive outcome tendency of educational and vocational programs (4 studies).

Applying the RNR model revealed the strongest mean effect ($OR=1.90$) for those programs that fulfilled all three principles. This effect indicated a substantial 16 % difference in recidivism rates between groups, assuming a 50 % baseline rate of recidivism in the control group. In spite of repeated criticism of the RNR model (e.g., Ward and Brown 2004; Ward and Maruna 2007), our review showed that these principles are robust when it comes to empirical evidence. Our results support each of these constituent claims in isolation as well as their effect when in concert. Unfortunately, the sample of available studies did not contain sufficient information to analyze extensions of RNR that contain organizational, staff, relationship, and other issues that are relevant for the outcome and applied in program accreditation (Andrews et al. 2011; Lösel 2012a; Maguire et al. 2010). The moderate number of studies also precluded estimating the relative importance of one or the other principle (e.g., Andrews and Bonta 2010; Lowenkamp et al. 2006). However, one should also bear in mind that, as in most RNR research, our studies addressed group-oriented risks, needs, and responsivity and not individualized approaches.

Although our findings support the particular strength of cognitive/behavioral and RNR-appropriate programs, we need to mention a few other aspects: only 7 out of the 25 evaluated programs adhered closely to the RNR principle; this means that a minority of evaluated treatments in Europe are following these replicated principles. The mean effect size of the most appropriate RNR interventions (equivalent to a reduction in recidivism of roughly 16 %) and of the cognitive/behavioral programs (equivalent to a reduction in recidivism of roughly

13 %) was also consonant with what has been reported in other reviews. For example, North American meta-analyses of cognitive-behavioral juvenile offender treatment found mean effects of 16 % (Wilson et al. 2005), 25 % (Landenberger and Lipsey 2005), and 33 % (Lipsey et al. 2001), and Redondo et al. (1999; 2001) reported a 23 % reduction in reoffending for CBT programs in Europe. The high correlation ($r=.54, p<.001$) between studies measuring the effect of treatments with high RNR and those that applied cognitive-behavioral treatment suggests that these factors may be confounded. This may explain why the mean effect of programs that adhere to RNR is only slightly larger than the mean effect of CBT programs.

The small number of primary studies limited the investigation of confounded moderators and multivariate analyses (see Lipsey 2003). However, we found a number of relevant trends. For example, the tendency of larger effects of programs in the community as opposed to those conducted in secure institutions is in accordance with North American research (e.g., Lipsey and Cullen 2007; Lipsey and Wilson 1998). The advantage of community programs may result in part from criminogenic effects of imprisonment (see Durlauf and Nagin 2011), although the latter findings are not fully consistent (e.g., Villettaz et al. 2006) and mainly based on relatively short incarceration without treatment. One must also take into account that the primary studies in our meta-analysis did not directly compare institutional and community treatment; rather, they examined differences to control groups within the same setting. Nevertheless, from a clinical and educational perspective, it is plausible that community programs show larger effects because they contain more opportunities for real life application and transfer (Lösel and Schmucker 2005). Therefore, practice should aim for community instead of custodial programs as far as criteria of compensation of guilt and public security allow non-custodial sentences. In addition, community programs are cheaper and would help to reduce prison overcrowding.

It is also of practical relevance that we found no outcome difference between voluntary and mandatory program participation. This speaks against older dichotomous assumptions about treatment motivation and supports more process-oriented concepts in which external triggers can lead to intrinsic change over time (McMurran 2002).

Because of the small number of studies in the subcategories of our meta-analysis, the effect sizes were not significantly related to individual components of methodological quality. However, the pattern of outcomes clearly followed similar trends to the literature on correctional treatment. For example, we found larger effects among studies using small samples. This is not only in accordance with research on juvenile offender treatment (Lipsey and Wilson 1998) but also on social skills training in crime prevention (Lösel and Beelmann 2003). This finding may have been influenced by a publication bias (small samples require larger effects to become significant) or a better quality of implementation in smaller studies. As our review contained only published studies, we could not test the first hypothesis. However, the second one is at least partially supported by our result of larger effects in studies with better fidelity, although one cannot exclude the possibility that integrity/fidelity problems were mainly reported when effects turned out lower than expected.

In accordance with the North American research, we also found a tendency of larger effects in studies that contained shorter follow-up periods and evaluated demonstration projects (e.g., Lipsey and Cullen 2007; Lipsey et al. 2006). The latter indicates the difference between ‘efficacy’ in model projects and ‘effectiveness’ in routine practice. Effectiveness evaluations typically reveal smaller effects, not only in correctional treatment. In demonstration projects, the study authors are often involved in the development or delivery of programs that also coincide with larger effect sizes (e.g., Eisner 2009; Lipsey and Wilson 1998; Lösel and Schmucker 2005). However, we could not investigate this aspect in the present review because most primary studies did not contain such combinations.

The type of evaluation design did not show a systematic relationship to effect size. This is in accordance with various reviews of correctional treatment (Lipsey and Cullen 2007; Lösel 1995), although the international findings on this issue are not fully consistent. Broader analyses of crime prevention programs revealed smaller effects in studies of high internal validity such as RCTs (Weisburd et al. 2001), and our results on the integrative score of methodological quality points in the same direction. This supports the view that methodological quality should not only be reduced to one dimension but take various threats to validity into account (Farrington 2003; Lösel and Köferl 1989; Lösel 2012a).

Although most of our findings were consistent with the North American ‘what works’ literature, we are not able to conclude that the latter can simply be generalized to Europe. This is due both to the disparity of findings between studies conducted within and outside the United Kingdom, and to the relatively small number of well-controlled European studies we could retrieve in our intensive search process. Two-thirds of the studies in our sample were carried out in the UK where the culture and legal system is more similar to the United States than that of most continental nations. For the large majority of European countries, we could not find any evaluation with sufficient methodological quality, and for those countries outside the UK that are represented in our review, there were only one or two studies on a specific type of program. Although the country from which the study originated was the only statistically significant difference between levels in our bivariate moderator analyses, the results of our meta-regression revealed that, once various treatment, methodological, and contextual variables were controlled, this difference disappeared. The country in which the treatment was delivered was not dispositive of treatment effectiveness; rather, seven of the nine evaluations that had been conducted outside the UK were cognitive-behavioral programs delivered in the community, and they mostly used very small samples with short follow-up periods. Furthermore, all of the deterrence- and intensive supervision-based treatment with long follow-up periods were evaluated within the UK.

For these reasons, our review does not justify far-reaching conclusions on ‘what works’ in young offender treatment in the whole of Europe. One must bear in mind that juvenile justice systems vary even more than their adult counterparts (Doob and Tonry 2004; Dünkel and Pruin 2010). Differences in the classification of offending, in responses to reoffending, in the age of criminal

responsibility, and in the handling of late adolescent or young adult offenders are only a few aspects of such variation across jurisdictions (Lösel et al. 2012). For example, it is doubtful that the secure custody measure evaluated in Curran et al. (1995), which was applied to Northern Irish youths of roughly 14 years of age, would have been applied to their peers in the Netherlands, where the juvenile justice system is far more welfare-oriented (Weermann 2007:262), or Sweden, where the juvenile justice system is more welfare-oriented still (Janson 2004). Furthermore, the very notion of ‘treatment as usual’ in the control groups may hold a different meaning across different countries. For example, in our meta-analysis, the very same intervention was applied to youths of similar age and psychological profile in Sweden (Sundell et al. 2008) and in Norway (Ogden and Hagen 2006; Ogden et al. 2007), yet only the latter two evaluations found a significantly positive effect. The authors attribute this discrepancy to the more punitive approaches of how youths are usually handled in Norway (Sundell et al. 2008; Andrée Löfholm et al. 2009). Similar aspects of punitivity need to be regarded in countries such as the United Kingdom (Bottoms et al. 2004).

Taking such issues into account, our study suggests four major conclusions. Firstly, the overall effect of young offender treatment in Europe is encouraging, and this is especially so for cognitive/behavioral programs and the RNR model. Secondly, there are various moderators of effect that could only be partially disentangled in our review. Thirdly, our meta-analysis revealed a strong deficit of well-controlled program evaluations in most of the European countries. Fourthly, transnational differences in program effectiveness seem to be confounded with many other moderators that are relevant for the outcome. As a consequence, we need policies to encourage well-controlled program implementation and evaluation in young offender treatment across countries and also broader views on ‘what works’ that go beyond isolated programs (Lösel 2012a). This would lead to an evidence-oriented approach that conserves the solid knowledge that has been developed over the last decades and overcomes current controversies about the ‘right’ paradigm in offender rehabilitation (see Andrews and Bonta 2010; Brayford et al. 2010; Lösel 2012b).

Appendix i: Keywords used to search electronic databases

Target Population	Youth Violent* OR Delinquent* OR Juvenile OR Hooligan* OR Hate
	AND
Intervention	Program* OR Treatment* OR Intervene* OR Correcti* OR Therap* OR Counsel* OR Mentor* OR Rehabilitati* OR Cogniti* OR Relapse OR Boot Camp* OR Wilderness Challenge* OR Intensive OR Incarcerat* OR Court* OR Probation OR Mandated OR Inmate* OR Institution* OR Non-Institution* OR Prison*
	AND
Outcome	Effect* OR Outcome* OR Eval* OR Experiment* OR RCT* OR Quasi* OR Trial* OR Empirical* OR ES* OR Recidiv* OR Reoffen*

Appendix ii: Searched Databases

Database	Number of Studies
Bibliographic Databases	
International Bibliography of the Social Sciences	795
PSYCINFO	4,018
PSYCARTICLES	186
PUBMED	336
COCHRANE Library	67
EMBASE	822
ISI Web of Knowledge	4,018
Cambridge Scientific Abstracts (CSA)	12,105
C2 SPECTR (ERIC)	1,672
Science Direct	2,282
Governmental Publications	
UK Home Office Research Database	5
Brå-Swedish National Council for Crime Prevention	0
Internet Resources	
Google Scholar	N/A
MetaCrawler	N/A
Unpublished Resource Databases	
Dissertation Abstracts International	683
Total	26,989

References

An asterisk (*) denotes the study was included in the meta-analysis.

- *Andrée Löfholm, C., Olsson, T., Sundell, K., & Hansson, K. (2009). Multisystemic Therapy with conduct-disordered young people: Stability of treatment outcomes two years after intake. *Evidence and Policy*, 5(4), 373–397.
- Andrews, D. (1995). The psychology of criminal conduct and effective treatments. In J. McGuire (Ed.), *What works: Reducing reoffending* (pp. 63–78). New York: Wiley.
- Andrews, D., & Bonta, J. (2010). *The Psychology of criminal conduct* (5th ed.). Newark: LexisNexis.
- Andrews, D., Zinger, I., Hoge, R., Bonta, J., Gendreau, P., & Cullen, F. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology*, 28(3), 369–404.
- Andrews, D. A., Bonta, J., & Wormith, S. (2006). The recent past and near future of risk and/or needs assessment. *Crime and Delinquency*, 52(1), 7–27.
- Andrews, D. A., Bonta, J., & Wormith, S. (2011). The Risk-Need-Responsivity (RNR) model: does adding the good lives model contribute to effective crime prevention? *Criminal Justice and Behavior*, 38, 735–755.

- Aos, S., Phipps, P., Barnoski, R., & Lieb, R. (2001). The comparative costs and benefits of programs to reduce crime, 4th ed. Washington State Institute for Public Policy. Retrieved May 23, 2012, from <http://www.wsipp.wa.gov/rptfiles/costbenefit.pdf>.
- Aos, S., Miller, M., & Drake, E. (2006). *Evidence-based adult corrections programs: What works and what does not*. Olympia: Washington State Institute of Public Policy.
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Chichester: Wiley.
- *Bottoms, A. E. (1995). *Intensive community supervision for young offenders: Outcomes, process and cost*. (Cambridge, UK: Institute of Criminology)
- Bottoms, A., Dignan, J., et al. (2004). Youth justice in Great Britain. In M. Tonry & A. Doob (Eds.), *Youth crime and youth justice: Comparative and cross-national perspectives. Crime and justice: A review of research* (Vol. 31, pp. 21–183). Chicago: University of Chicago.
- Brayford, J., Cowe, F., & Deering, J. (2010). What else works – Back to the future? In J. Brayford, F. Cowe, & J. Deering (Eds.), *What else works? Creative work with offenders* (pp. 254–268). Portland: Willan.
- *Cann, J., Falshaw, L., & Friendship, C. (2005). Understanding ‘What Works’: Accredited cognitive skills programmes for young offenders. *Youth Justice*, 5(3), 165–179
- Cleland, C., Pearson, F., Lipton, D., & Yee, D. (1997). *Does age make a difference? A meta-analytic approach to reductions in criminal offending for juveniles and adults*. San Diego: Presented at the Annual Meeting of the American Society of Criminology.
- *Curran, D., Kilpatrick, R., Young, V., & Wilson, D. (1995). Longitudinal aspects of reconviction: Secure and open intervention with juvenile offenders in Northern Ireland. *The Howard Journal of Criminal Justice*, 24(2), 97–123
- Doob, A. N., & Tonry, M. (2004). Varieties of youth justice. In M. Tonry & A. Doob (Eds.), *Youth crime and youth justice: Comparative and cross-national perspectives. Crime and justice: A review of research, vol. 31* (pp. 1–20). Chicago: University of Chicago.
- Dowden, C., & Andrews, D. (1999). What works in young offender treatment: a meta-analysis. *Forum on Corrections Research*, 11(2), 21–24.
- Dowden, C., & Andrews, D. (2003). Does family intervention work for delinquents? Results of a meta-analysis. *Canadian Journal of Criminology and Criminal Justice*, 45(3), 327–342.
- Düinkel, F., & Pruin, I. (2010). Young adult offenders in the criminal justice systems of European countries. In F. Düinkel, J. Grzywa, P. Horsfield, & I. Pruin (Eds.), *Juvenile justice systems in Europe: Current situation and reform developments, vol. 4* (pp. 1557–1580). Mönchengladbach: Forum Verlag Godesberg.
- Durlauf, S. N., & Nagin, D. (2011). Imprisonment and crime: can both be reduced? *Criminology and Public Policy*, 10(1), 13–54.
- Eisner, M. (2009). No effects in independent prevention trials: can we reject the cynical view? *Journal of Experimental Criminology*, 5(2), 163–183.
- Farrington, D. (1986). Age and crime. In M. Tonry (Ed.), *Crime and justice: A review of research, vol. 7* (pp. 189–250). Chicago: University of Chicago.
- Farrington, D. P. (2003). Methodological quality standards for evaluation research. *Annals of the American Academy of Political and Social Science*, 587(1), 49–68.
- Farrington, D., Ditchfield, J., Hancock, G., Howard, P., Jolliffe, D., Livingston, M., et al. (2002). *Evaluation of two intensive regimes for young offenders. Home Office Research Study 239*. London: Home Office.
- Garrido, V., Morales, L. A., & Sánchez-Meca, J. (2006). What works for serious juvenile offenders? A systematic review. *Psicothema*, 18(3), 611–619.
- Gensheimer, L., Mayer, J., Gottschalk, R., & Davidson, W., II. (1986). Diverting youth from the juvenile justice system: A meta-analysis of intervention efficacy. In S. Apter & A. Goldstein (Eds.), *Youth violence: Programmes and prospects* (pp. 39–57). Elmsford: Pergamon.
- Gottschalk, R., Davidson, W., II, Gensheimer, L., & Mayer, J. (1987). Community-based interventions. In H. Quay (Ed.), *Handbook of juvenile delinquency* (pp. 266–289). New York: Wiley.
- Hamilton, L., Koehler, J.A., & Lösel, F. (2011) Programmes to reduce reoffending throughout Europe: Three surveys on current practice. Final report of the project ‘Strengthening transnational approaches to reducing reoffending’, Appendix D. Retrieved May 23, 2012, from https://webmail.springer-sbm.com/owa/redir.aspx?C=23b359279c184631ba3922062a783e9d&URL=http%3a%2f%2fwww.cepprobation.org%2fuploaded_files%2fRep%2520STARR%2520ENG.pdf.
- Hanson, R. K., Gordon, A., Harris, A., Marques, J. K., et al. (2002). First report of the collaborative outcome data project on the effectiveness of psychological treatment for sex offenders. *Sexual Abuse: A Journal of Research and Treatment*, 14(2), 169–194.

- Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). The principles of effective correctional treatment also apply to sexual offenders: a meta-analysis. *Criminal Justice and Behavior*, 36(9), 865–891.
- Harper, G., & Chitty, C. (Eds.). (2005). *The Impact of corrections on reoffending: A review of 'What Works'*, 2nd ed. Home Office Research Study 291. London: Home Office.
- Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, 117(1), 167–168.
- Hollin, C. (2002). Risk–Needs assessment and allocation to offender programmes. In J. McGuire (Ed.), *Offender rehabilitation and treatment: Effective programmes and policies to reduce reoffending* (pp. 309–332). Chichester: Wiley.
- Hollin, C. R. (2008). Evaluating offending behaviour programmes: does only randomization glister? *Criminology and Criminal Justice*, 8, 89–106.
- Hollin, C., & Palmer, E. J. (2009). Cognitive skills programmes for offenders. *Psychology Crime & Law*, 15(2), 147–164.
- Janson, C.-G. (2004). Youth justice in Sweden. In M. Tonry & A. Doob (Eds.), *Youth crime and youth justice: Comparative and cross-national perspectives. Crime and justice: A review of research*, vol. 31 (pp. 391–441). Chicago: University of Chicago.
- *Kruissink, M. (1990). *The Halt program: Diversion of juvenile vandals. Dutch penal law and policy: Notes on criminological research from the research and documentation centre*. The Hague: Ministry of Justice.
- Landenberger, N., & Lipsey, M. (2005). The positive effects of cognitive-behavioural programs for offenders: a meta-analysis of factors associated with effective treatment. *Journal of Experimental Criminology*, 1(4), 451–476.
- Latimer, J., Dowden, C., & Morton-Bourgon, K. (2003). *Treating youth in conflict with the law: A new meta-analysis. Rep. RR03YJ-3e*. Ottawa: Department of Justice.
- Lipsey, M. (1992). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In T. Cook, H. Cooper, D. Cordray, H. Hartmann, L. Hedges, R. Light, T. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook* (pp. 83–127). New York: Russell Sage.
- Lipsey, M. (2003). Those confounded moderators in meta-Analysis: Good, bad, and ugly. *Annals of the American Academy of Political and Social Science*, 587(1), 69–81.
- Lipsey, M., & Cullen, F. (2007). The effectiveness of correctional rehabilitation: a review of systematic reviews. *Annual Review of Law and Social Science*, 3, 297–320.
- Lipsey, M. W., & Wilson, D. B. (1998). Effective intervention for serious juvenile offenders. In R. Loeber & D. P. Farrington (Eds.), *Serious and violent juvenile offenders: Risk factors and successful interventions* (pp. 313–345). London: Sage.
- Lipsey, M., Chapman, G., & Landenberger, N. (2001). Cognitive-behavioral programs for offenders. *Annals of the American Academy of Political and Social Science*, 578(1), 144–157.
- Lipsey, M., Petrie, C., Weisburd, D., & Gottfredson, D. (2006). Improving evaluation of anti-crime programs: summary of a National Research Council report. *Journal of Experimental Criminology*, 2(3), 271–307.
- *Little, M., Kogan, J., Bullock, R., & van der Laan, P. (2004). ISSP: An experiment in multi-systemic responses to persistent young offenders known to children's services. *British Journal of Criminology*, 44(2), 225–240.
- Lloyd, C., Mair, G., & Hough, M. (1994). *Explaining reconviction rates: A critical analysis. Home Office Research Study 136*. London: HMSO.
- *Lobley, D., & Smith, D. (2007). *Persistent young offenders: An evaluation of two projects*. (Aldershot, UK: Ashgate)
- Lösel, F. (1995). The efficacy of correctional treatment: A review and synthesis of meta-evaluations. In J. McGuire (Ed.), *What works: Reducing reoffending. Guidelines from research and practice* (p. 79). Chichester: Wiley.
- Lösel, F. (2000). The efficacy of sexual offender treatment: A review of German and international evaluations. In P. J. van Koppen & N. H. M. Roos (Eds.), *Rationality, information and progress in psychology and law* (pp. 145–170). Maastricht: Metajuridica Publications.
- Lösel, F. (2012a). Offender treatment and rehabilitation: What works? In M. Maguire, R. Morgan, & R. Reiner (Eds.), *The Oxford handbook of criminology* (pp. 986–1016). Oxford: Oxford University Press.
- Lösel, F. (2012b). Towards a third phase of 'what works' in offender rehabilitation. In R. Loeber & B. C. Welsh (Eds.), *The future of criminology* (pp. 196–203). Oxford: Oxford University Press.
- Lösel, F., & Beelmann, A. (2003). Effects of child skills training in preventing antisocial behaviour: a systematic review of randomized evaluations. *Annals of the American Academy of Political and Social Science*, 587(1), 84–109.

- 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: Springer.
- Lösel, F., & Pomplun, O. (1998). *Jugendhilfe statt untersuchungshaft: Eine evaluationsstudie zur heimunterbringung. Studien und materialen zum straf- und massregelvollzug*, vol. 7. Pfaffenweiler: Centaurus.
- Lösel, F., & Schmucker, M. (2005). The effectiveness of treatment for sexual offenders: a comprehensive meta-analysis. *Journal of Experimental Criminology*, 1(1), 117–146.
- Lösel, F., Köferl, P., & Weber, F. (1987). *Meta-evaluation der sozialtherapie*. Stuttgart: Enke.
- Lösel, F., Bottoms, A. E., & Farrington, D. P. (Eds.). (2012). *Young adult offenders. Lost in transition?* Milton Park, UK: Routledge.
- Lowenkamp, C., Latessa, A., & Holsinger, A. (2006). The risk principle in action: what have we learned from 13,676 offenders and 97 correctional programs? *Crime and Delinquency*, 52(1), 77–93.
- MacKenzie, D. (2006). *What works in corrections: Reducing the criminal activities of offenders and delinquents*. New York: Cambridge.
- Maguire, M., Grubin, D., Lösel, F., & Raynor, P. (2010). ‘What works’ and the correctional services accreditation panel: taking stock from an inside perspective. *Criminology and Criminal Justice*, 10, 37–58.
- McGuire, J. (2002). Motivation for what? Effective programmes for motivated offenders. In M. McMurrin (Ed.), *Motivating offenders to change: A guide to enhancing engagement in therapy* (pp. 157–172). Chichester: Wiley.
- McMurrin, M. (2002). Motivation to change: selection criterion or treatment need? In M. McMurrin (Ed.), *Motivating offenders to change* (pp. 3–13). Chichester: Wiley.
- *McMurrin, M., & Boyle, M. (1990). Evaluation of a self-help manual for young offenders who drink: A pilot study. *British Journal of Clinical Psychology*, 29(1), 117–119.
- Ministry of Justice (UK). (2010). *Compendium of reoffending statistics and analysis*. Ministry of Justice Statistics Bulletin. London: Ministry of Justice.
- *Mitchell, J., & Palmer, E. (2004). Evaluating the ‘Reasoning and Rehabilitation’ program for young offenders. *Journal of Offender Rehabilitation*, 39(4), 31–45.
- *Newburn, T., & Shiner, M. (2005). *Dealing with disaffection: Young people, mentoring and social inclusion*. (Cullompton, UK: Willan)
- *Ogden, T., & Hagen, K. (2006). Multisystemic treatment of serious behaviour problems in youth: Sustainability of effectiveness two years after intake. *Child and Adolescent Mental Health*, 11(3), 142–149.
- *Ogden, T., Hagen, K., & Andersen, O. (2007). Sustainability of the effectiveness of a programme of Multisystemic Treatment (MST) across participant groups in the second year of operation. *Journal of Children's Services*, 2(3), 4–14.
- Pearson, F. S., Lipton, D. S., Cleland, C. M., & Yee, D. S. (2002). The effects of behavioral/cognitive-behavioral programs on recidivism. *Crime & Delinquency*, 48, 476–496.
- *Raynor, P., & Vanstone, M. (1997). *Straight Thinking on Probation (STOP)*. The Mid Glamorgan experiment. Probation Studies Unit Report, 4. (Oxford, UK: Centre for Criminological Research)
- Redondo, S., Sánchez-Meca, J., & Garrido, V. (1999). The Influence of treatment programmes on the recidivism of juvenile and adult offenders: An European meta-analytic review. *Psychology, Crime and Law*, 5(3), 251–278.
- Redondo, S., Sánchez-Meca, J., & Garrido, V. (2001). Treatment of offenders and recidivism: assessment of the effectiveness of programmes applied in Europe. *Psychology in Spain*, 5(1), 47–62.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury: Sage.
- *Scholte, E., & Smit, M. (1988). Early social assistance for juveniles at risk. *International Journal of Offender Therapy and Comparative Criminology*, 32(3), 209–218.
- *Shapland, J., Atkinson, A., Atkinson, H., Dignan, J., Edwards, L., Hibbert, J., Howes, M., Johnstone, J., Robinson, G., & A. Sorsby. (2008). *Does restorative justice affect reconviction? The fourth report from the evaluation of three schemes*. Ministry of Justice Research Series, 10/08. (London, UK: Home Office)
- *Slot, N. (1983). The implementation and evaluation of a residential social skills training program for youth in trouble. (In W. Everaerd, C. Hindley, A. Bot, & J. J. van der Werf, (Eds.), *Development in adolescence: Psychological, social, and biological aspects* (pp.192–205). Amsterdam, Netherlands: Marinus Nijhoff.)
- *Slot, N., & Bartels, A. (1983). Outpatient social skills training for youth in trouble; theoretical background, practice and outcome. (In W. Everaerd, C. Hindley, A. Bot, & J. J. van der Werf, (Eds.),

- Development in adolescence: Psychological, social, and biological aspects (pp.176-191). Amsterdam, Netherlands: Marinus Nijhoff.)
- *St. James-Roberts, I., Greenlaw, G., Simon, A., & Hurry, J. (2005). *National evaluation of Youth Justice Board mentoring schemes 2001 to 2004*. (London, UK: Youth Justice Board).
- *Sundell, K., Hansson, K., Andrée Löfholm, C., Olsson, T., Gustle, L.-H., & Kadesjö, K. (2008). The transportability of Multisystemic Therapy to Sweden: Short-term results from a randomized trial of conduct-disordered youths. *Journal of Family Psychology*, 22(4), 550-560.
- Tong, L. S. J., & Farrington, D. P. (2006). How effective is the Reasoning and Rehabilitation programme in reducing offending? A meta-analysis of evaluations in four countries. *Psychology, Crime and Law*, 12, 3-24.
- Tournier, P., & Barre, M. (1990). Enquête sur les systèmes pénitentiaires dans les membres du Conseils de l'Europe: Démographie carcérale comparée. *Bulletin d'Information Pénitentiaire*, 15, 4-44.
- Ttofi, M., Farrington, D., & Baldry, A. (2008). *Effectiveness of programmes to reduce school bullying: a systematic review*. Stockholm: Swedish National Council for Crime Prevention (Brottsförebyggande rådet – Brå).
- Villetaz, P., Killias, M., & Zoder, I. (2006) The effects of custodial vs. non-custodial sentences on re-offending: A systematic review of the state of knowledge. *Campbell Systematic Reviews*, 13
- Ward, T., & Brown, M. (2004). The good lives model and conceptual issues in offender rehabilitation. *Psychology, Crime and Law*, 10(3), 243-257.
- Ward, T., & Maruna, S. (2007). *Rehabilitation*. New York: Routledge.
- Weermann, F. (2007). Juvenile offending. (In M. Tonry, & C. Bijleveld, (Eds.), *Crime and justice in the Netherlands*. *Crime and Justice: A Review of Research*, vol. 35 (pp. 261-318). Chicago: University of Chicago.)
- Weisburd, D., Lum, C., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *Annals of the American Academy of Political and Social Science*, 578(1), 50-70.
- Welsh, B., & Farrington, D. (2001). A review of research on the monetary value of preventing crime. In B. Welsh, D. Farrington, & L. Sherman (Eds.), *Costs and benefits of preventing crime* (pp. 87-122). Oxford: Westview.
- Wilson, D. (2009). Missing a critical piece of the pie: simple document search strategies inadequate for systematic reviews. *Journal of Experimental Criminology*, 5(4), 429-440.
- Wilson, D., Bouffard, L., & MacKenzie, D. (2005). A quantitative review of structured, group-oriented, cognitive-behavioral programs for offenders. *Criminal Justice and Behavior*, 32(2), 172-204.

Johann A. Koehler is a Research Assistant at the Institute of Criminology, University of Cambridge. His interests are in research synthesis, comparative criminology, and evidence-based criminal justice policy.

Friedrich Lösel is Director of the Institute of Criminology at the University of Cambridge, and Professor of Psychology at the University of Erlangen-Nuremberg, Germany. He has carried out research on juvenile delinquency, prisons, offender treatment, football hooliganism, school bullying, personality disordered offenders, resilience, close relationships, child abuse, and family education and evaluation methodology. He is the author or editor of 18 books and over 330 journal articles. In recognition of his scientific work, he has received various honors including the Award for Lifetime Achievement of the European Association of Psychology and Law, the Sellin-Glueck Award of the American Society of Criminology, the Stockholm Prize in Criminology, and the German Psychology Prize.

Thomas D. Akoensi is a doctoral researcher at the Institute of Criminology, University of Cambridge. His research interests include prison officer stress, organizational behavior, legitimacy, comparative criminology, and mixed methods research.

David K. Humphreys is a post-doctoral research fellow at the Institute of Public Health, University of Cambridge. His research interests include the evaluation of social interventions in public health. His work crosses several fields, such as substance abuse, violence, and physical activity.