J Exp Criminol
DOI 10.1007/s11292-017-9297-z

Evidence comes by replication, but needs differentiation: the reproducibility issue in science and its relevance for criminology

Friedrich Lösel[1,2]

Abstract Recent publications in *Nature*, *Science*, and other journals raised concerns about the reproducibility of empirical findings in psychology and other scientific disciplines. This article summarizes some of these arguments and results that led to discussions about a "replication crisis" in research. In criminology, there is not yet a similar discussion, although the need for more replications has been emphasized in the past. The present article addresses this topic with special consideration of program evaluations in early developmental crime prevention and offender treatment. In both fields, there has been substantial progress in research and practice. Most systematic reviews showed mean positive effects; however, nearly all of them demonstrated very heterogeneous findings that could not be attributed to the content of programs. This does not allow simple recommendations of "what works" for policy-making and practice. In addition, there is a serious lack of long-term follow-ups and independent evaluations. The article shows remarkable similarity of the findings and problems in both fields of intervention. Problems of reproducibility prove to be highly relevant for criminology, although there is no need for using the term "crisis". The article proposes various strategies that can enhance the reproducibility of findings, i.e., more systematic investigation of those differentiated conditions under which interventions are most effective. An integrative model of relevant characteristics is briefly presented. It refers to factors of the programs, contexts, participants, and evaluation methods. Confirmatory meta-analyses can play an important role on the path toward more differentiated and replicated knowledge

Updated version of the 2015 Joan McCord Award Lecture of the Academy of Experimental Criminology at Washington DC, USA.

✉ Friedrich Lösel
fal23@cam.ac.uk; friedrich.loesel@fau.de

[1] Institute of Criminology, University of Cambridge, Sidwick Avenue, Cambridge CB3 9DA, UK

[2] Institute of Psychology, University of Erlangen-Nuremberg, Bismarckstr. 1, 91054 Erlangen, Germany

Published online: 18 August 2017

Keywords Replication in science · Program evaluation · Developmental crime prevention · Offender treatment · Meta-analysis

The Joan McCord Award of the Academy of Experimental Criminology has been a great honor for me. Joan inspired much of my own work on criminological topics. She was a role model for me not only in experimental criminology but also in other areas, such as family relationships and juvenile delinquency (McCord 1991), child abuse (McCord 1983), psychopathy (McCord 2001), and resilience (McCord 1994). Joan was always curious, in science as well as in arts, history, politics, and all aspects of human life. When we once walked through a poor and perhaps dangerous neighborhood in Brazil, she emphasized how important it is to get an own impression. And she was always critical and precise in her evaluations. When we visited a new museum of modern art and design in Germany, she dryly commented that the building is wonderful, but the exhibition needs better objects. Her realistic and evidence-based attitude was particularly obvious in her work on the Cambridge–Somerville Youth Study (McCord 1992). Joan frankly reported that the long-term outcomes of this landmark prevention project were not positive and warned that programs can harm, in spite of best intentions (McCord 1978, 2003). This topic and attitude led me to choose the issue of replication as the theme of my Joan McCord Lecture.

The issue of replication in science

Replication of findings is a key issue of any empirical discipline (Popper 1959). Most recently, it became a hot topic when the Reproducibility Project in Psychology published its findings (Open Science Collaboration 2015). This large project, funded by the Laura and John Arnold Foundation, investigated whether the results of empirical studies in psychology are robust when tested in replications. The rationale of the study derived from widespread concerns in the discipline, such as selective data analysis, selective reporting, and insufficient specification of the necessary conditions to obtain a specific result. Numerous international collaborators carried out exact replications of 100 experimental and correlational studies that had been published in 2008 in three prestigious psychological journals. The results were sobering. Less than half of the effects in the original study could be replicated in quantitative terms and approximately one-quarter of effects went in the opposite direction. The mean effect size dropped from $r = 0.40$ in the original studies to 0.20 in the replications.

Some variation in psychological findings on a specific topic is normal due to sampling, situational, and other conditions. Although I held a chair of psychology over many years, I was always skeptical about studies that tested general hypotheses on human behavior in small student samples and artificial scenarios. However, the reproducibility issue is not only a problem of psychology. Ioannidis (2005) investigated replications of 49 highly cited studies ($n > 1000$) in medicine. Forty-five studies reported "effective" results, 44% could be replicated (but often with smaller effects), 16% were contradicted by subsequent studies, 16% got stronger results, and 24% remained unchallenged. A survey of Baker (2016), published in *Nature*, received answers of 1576 scholars from hard sciences (chemistry, biology, physics, engineering,

medicine, earth and environment, and others). Fifty-two percent of respondents said that there is a significant reproducibility crisis, 38% stated a "slight crisis", and only 7% denied a crisis. Across all disciplines, 62–87% of the respondents said that they could not replicate an experiment of somebody else and a slightly smaller proportion agreed that they could not replicate own findings (51–74%). When asked about how much published work in their respective field is reproducible, most answers ranged between 50% and 80%, but more than a quarter assumed lower rates.

The word "crisis" should not be used inflationary, but the reproducibility issue has been repeatedly emphasized in the social sciences before the recent alerting articles in *Nature* and *Science*. For example, already, Farrington (2000) noted that pure replications are too rare in criminological research. Flay et al. (2005), Valentine et al. (2011), Gottfredson et al. (2015), and others addressed standards of evidence that should reduce replication problems in prevention science. The strong need for more replication has been emphasized from a statistical perspective (e.g., Hunter 2001), but there are many social factors in research that form obstacles against a culture of replication. In criminology and other disciplines, the academic world reinforces mass publication ("publish or perish"). Researchers seem to avoid replications because they want to demonstrate their own creativity. Journals require "novelty" so that pure replications are hard to publish. Scholars assume that replications would get less academic recognition, although this may not be the case for falsifications of prominent hypotheses. Journal impact factors are often seen as more important than the real content of a paper. Research foundations tend to promote large collaborative projects, but these make replication more difficult. Although policies encourage open data access, scholars often hesitate to offer hardly gathered own data to others. Randomized experiments play a key role in the establishment of sound knowledge (Boruch et al. 2000), but they are not widely implemented in criminology (Farrington 2003). The Campbell Collaboration aims to provide best evidence by promoting measures of transparency in systematic reviews (Farrington and Petrosino 2001; Petrosino et al. 2001); however, in primary studies, such safeguards are still rare. In studies with many variables, selective reporting and fishing for significance is a widespread danger. In research areas with financial incentives, selective data analysis and reporting can be a serious problem (Eisner et al. 2015) and more neutral, independent evaluations are needed (Petrosino and Soydan 2005). Beyond financial issues, scientific networks may implicitly influence what is analyzed and published. Last but not least, there are time and resource issues that hinder replications in complex field experiments that require years of follow-up. Joan McCord's Cambridge–Somerville Youth Study is an example for that, but there are many shorter criminological projects that would also be very difficult to replicate.

These and other influences on the reproducibility of research in social sciences are not new. For example, already, Rossi (1978) formulated the *Iron Law* of mean zero effects. Although he conceded that there were examples of positive results, he concluded that most social programs, when properly evaluated, are ineffective or only marginally accomplishing their aims. Rossi's *Iron Law* focused on the mean, but the variance was likewise important because only consistent zero effects would advance the knowledge about what does *not* work. Crime prevention was a typical example at that time. Large systematic reviews of correctional treatment (Lipton et al. 1975; Sechrest et al. 1979) found many methodologically weak studies and inconsistent results that contributed to the impression of "nothing works". Later, Rossi (1987)

differentiated three "metallic rules" of program evaluation. The *Stainless Steel* rule meant that the better designed the evaluation of a social program, the more likely is a net impact of zero. The *Zinc* rule denoted that mainly programs that are likely to fail are evaluated. And the *Brass* rule said that the more social programs are designed to change individuals, the more their net impact will be zero.

In connection with the latter rule, the present article will address the replication issue by focusing on person-oriented criminological interventions, in particular on developmental prevention and offender rehabilitation/treatment. I selected these two topics because they are important policy areas and parts of my own research. My discussion will mainly focus on examples of criminological research in these fields. For more general issues of statistical, internal, construct, and external validity, see Shadish et al. (2002).

Replication in developmental prevention

Since Rossi's critical view of the impact of social programs, there has been progress in the evaluation of criminological and related interventions. In developmental and life course criminology, early prevention has strongly expanded (Farrington et al. 2016; Farrington and Welsh 2007). Numerous universal or risk-based programs have been implemented in families, kindergartens, (pre)schools, family education centers, child guidance clinics, and other services. Although most programs that are implemented in practice are not evidence-based (Lösel et al. 2006; Mihalic and Elliott 2015), many sound studies have been carried out and integrated in systematic reviews. An overview of meta-analyses showed that the findings varied widely (Lösel 2012a), for example between a mean of $d = 0.10$ in a meta-analysis of school-based programs (Gottfredson et al. 2002) and $d = 0.65$ in a meta-analysis of parent trainings (Serketich and Dumas 1996), but all means were not zero as Rossi suggested 30 years ago. Most recently, Farrington et al. (2017) analyzed 50 systematic reviews of developmental and social programs that investigated outcomes of delinquency, offending, violence, aggression, or school bullying. Twenty-five reviews contained school-based programs, eleven individually focused programs, nine family-based programs, and five general prevention programs. Mean effect sizes were available from 33 syntheses and, with the exception of four, these were all statistically significant. The mean effects varied widely, that is, from an odds ratio (OR) of 1.08 ($d = 0.04$) in a meta-analysis of school programs (Wilson et al. 2001) to an OR of 3.19 ($d = 0.64$) in a meta-analysis of child-focused programs (Robinson et al. 1999). The average effect was significant for all four types of programs and the overall effect was OR = 1.46. According to Cohen (1992), this is a small effect ($d = 0.21$, $r = 0.10$), but it is, insofar, realistic, as most correlations between single early risk factors and later delinquency are significant, but low to moderate (Hawkins et al. 1998; Lösel 2002; Murray et al. 2010). An OR of 1.46 is also practically relevant: depending on the prevalence of behavior problems in a cohort, it could indicate a reduction from 20% to 15%, that is, of one-quarter (Wilson and Lipsey 2007). Since long criminal careers of young people are very costly (Cohen and Piquero 2009; Piquero et al. 2013), even small effects of prevention programs can be cost-effective (Aos et al. 2004; Welsh and Farrington 2015).

In contrast to the overall encouraging results, it is less clear how far the above findings are reproducible in daily practice. Valentine et al. (2011) thoroughly analyzed various scenarios (cases) of differences in the outcomes of two (or more) implementations of a specific prevention program. They addressed issues of the evaluation design, statistical assessment strategies, investigator independence, and other aspects of inconsistent results. In practice, such factors are often combined and difficult to disentangle. In addition, broader context issues have to be taken into account. For example, the majority of studies on developmental crime prevention stem from North America and often from demonstration projects. Replication within and across different countries cannot simply be taken for granted. Although some research suggests that basic characteristics of interventions can be generalized (Knerr et al. 2013; Koehler et al. 2013), other examples cast doubt on this assumption.

For example, various randomized controlled trials (RCTs) have been carried out on indicated prevention or early treatment by Multisystemic Therapy (MST; Henggeler et al. 2009). Most evaluations came from the United States and, often, the program developers were involved. They showed desirable and sometimes extremely strong effects (e.g., Borduin et al. 2009). Some independent evaluations in other countries found less or no positive effects, for example, Leschied and Cunningham (2002) in Canada or Sundell et al. (2008) in Sweden. Other independent evaluations outside the United States showed desirable effects of MST, for example, Ogden and Amlund Hagen (2006) in Norway or Asscher et al. (2014) in the Netherlands (although the latter not on official delinquency). Sundell et al. (2008) discussed potential reasons in the social welfare system that may have been relevant for the inconsistency in MST evaluation in Norway versus Sweden.

Beyond the cultural/social context, evaluation methods and selective reports seem to be relevant for different results on MST: a meta-analysis of the MST group found a substantial mean effect (Curtis et al. 2004); however, a systematic review by Littell (2006) raised concerns about the validity of various MST evaluations, particularly those by the program developers themselves. Littell objected that most positive effects reported in the articles from the Henggeler group were from post-hoc analyses of subgroups and/or on secondary outcome criteria. The mean effects were rather small and statistically not significant for a priori analyses of full sample results on primary outcomes. Henggeler et al. (2006) defended their findings; however, a more recent independent meta-analysis found a mean effect of MST that was lower than that of Curtis et al. (2004), although a little more positive than Littell (2006) reported (van der Stouwe et al. 2014). There was a significant effect on the primary outcome of delinquency, but numerous moderators played a role (e.g., country of origin, efficacy versus effectiveness, study quality, treatment duration, sample, and outcome characteristics). The above findings clearly show a substantial amount of variance between single evaluations that may not allow a general conclusion about the effectiveness of MST.

This situation is not rare in developmental prevention. For example, whereas Sanders et al. (2000) reported desirable effects of their Triple-P parenting program in Australia and a meta-analysis of Triple-P researchers showed mean positive outcomes (Nowak and Heinrichs 2008), independent research found no effect in Switzerland (Eisner et al. 2012). Eisner (2014) also questioned the results of a large-scale implementation of Triple-P in the United States and Sanders (2015) published a paper on

how to deal with conflicts of interest. As for MST, details cannot be discussed here; however, obviously, there are, again, controversial findings on a widespread program.

In addition to replication across different studies, there are questions of generalizability when one takes a closer look at single evaluations. Even most studies using RCTs or sound quasi-experimental designs have rather short follow-up periods and do not address the issue of sustainability (Lösel and Beelmann 2003; Mihalic and Elliott 2015). Only a handful of evaluations worldwide have long follow-ups of about ten years or more (Farrington and Welsh 2013). Insofar, it remains unclear whether programs that intend to prevent a criminal development really reach this aim. There are a few exceptional studies with positive effects from childhood to adulthood (e.g., Schweinhart 2013; for some other studies, see below), but McCord's (2003) study showed the other side of the coin.

Deficits in well-replicated, long-term findings are also reported from the Blueprints for Healthy Youth Development. This important registry established standards for evidence-based prevention, for example, at least two RCTs or sound quasi-experiments with positive results. Taking stock of the Blueprints, Mihalic and Elliott (2015) reported that more than 1300 prevention programs have been analyzed over time, but only 54 could be certified as model programs that fulfilled the criteria of solid evidence. Although the authors stated an overall progress, they emphasized that the number of model programs would be less than a handful if independent evaluation would be required as a criterion. They also noted that the Blueprint's criterion of "sustained impact" is only at least 12 months. Many programs would not have been certified if a longer period had been demanded. In addition, the quality of model programs often deteriorates in practice (Gandhi et al. 2007) and effectiveness is typically lower than efficacy in demonstration studies (e.g., Weisz et al. 1995). Since evidence-based registries on what works are highly important (Gottfredson 2016), self-critical comments of pioneers in this field must be taken seriously. One should also be aware that various registries apply different criteria, so there is inconsistency with regard to what works or what is best practice (Fagan and Buchanan 2016; Gandhi et al. 2007).

Although researchers are aware of replications *across* studies, it is less recognized that there is a similar issue of outcome replication or consistency *within* single evaluations. These can be illustrated by findings from our own *Erlangen-Nuremberg Development and Prevention Study*. This project combined a prospective longitudinal and experimental study on kindergarten children and their families in Bavaria. In the prevention part, the universal program EFFEKT has been evaluated. It contains a program on positive parenting, child training on social problem solving, and a combination of both. The controlled design showed positive effects on externalizing behavior problems after 2–3 months, 2–3 years, and 4–5 years (Lösel et al. 2009; Lösel and Stemmler 2012). After about 10 years, there were still some significant desirable outcomes, that is, in boys' self-reported property offending (Lösel et al. 2013). We also found various positive effects in shorter evaluations of the program in samples from deprived migrant backgrounds (Runkel et al. 2016) and families with emotional problems (Bühler et al. 2011). Overall, the project showed replicated effects, but the findings varied across different follow-up periods, outcome measures, and sub-programs. In some analyses, the child training had significant effects, while in others, the parent training, and more often the combined program, had better outcomes. We found desirable effects when the kindergarten nurses or school teachers assessed the child

behavior, but not when the mothers were the informants. Some results also varied with regard to the kind of behavior problems. We could provide plausible explanations for these variations, but we are aware of the risks of post-hoc plausibility and fishing for significance.

Perhaps inconsistency in our findings may be partially due to the implementation of a relatively short universal prevention program. However, evaluations of more intensive programs also showed positive effects, as well as different findings across follow-up times, outcome measures, and subgroups (e.g., Asscher et al. 2014; Kellam et al. 2008). Similar observations have been made in some of the most prominent studies of intensive risk-based prevention, for example:

> The *Nurse Family Partnership* program (Olds et al. 1998) supports at-risk mothers during pregnancy and the first two years after birth. A sound evaluation after 12 years showed significant desirable effects on partner relationships, health behavior, need for social care, and other outcome measures, but not on alcohol use and arrest (Olds et al. 2010). In a follow-up at age 19 years, only females had significantly less delinquency than the control group (Eckenrode et al. 2010), although various previous findings were significant for males.
>
> The *Family and School Together (FAST)* prevention trial started at child age 6–7 years and lasted over 5 years. The program contained parent training, home visits, child social skills training, parent–child sessions, academic tutoring, peer coaching, and classroom management (Conduct Problems Prevention Research Group, CPPRG 2002). The RCT showed desirable short- and long-term effects on measures of children's problem-solving, cognitive skills, and social behavior in various follow-ups, but there were also several nonsignificant and some negative outcomes (CPPRG 2004, 2010). At age 19 years, the program group had fewer official offenses (particularly in the highest risk group), but there were no significant effects on self-reported delinquency, which is normally more sensitive to change and had shown positive effects before.
>
> The *Montréal Prevention Experiment* addressed high-risk 7- to 9-year-old boys from low socioeconomic families (Tremblay et al. 1995). It lasted about two years and included a program on adequate parenting and child social skills training. The RCT evaluation with two control groups showed no clear short-term effects, but significant effects after three years and later (e.g., less aggression and gang membership). After 15 years, more program participants had completed high school and fewer had a criminal record than in the control group (Boisjoli et al. 2007).

I referred to these three examples because their research quality is beyond any doubt. Many other prevention studies also found significant effects in some variables, at some times, and in some sub-groups (but not in others). Sometimes, there are decreasing effects over time, but, occasionally, also increasing effects ("sleeper effects"). Researchers provide sound reasons for the inconsistency in some of their results. However, as in the reproducibility discussion in psychology, these post-hoc interpretations are more based on plausibility than on prior hypotheses. In philosophy of science, this is known as "exhaustion", that is, further conditions are added to the deductive-nomological model of explanation (Hempel and Oppenheim 1948). Perhaps practice

may not be interested in philosophy of science; however, recommendations of a program without specified conditions may lead to disappointment when a model program is re-implemented without success. Specification of relevant conditions is essential in practice and its lack may be one reason why scholars argue against randomized experiments (Cook 2003).

Outcome variation is normal in social interventions such as developmental crime prevention. Therefore, meta-analyses are important to estimate reproducibility. As mentioned, they show overall positive, but very heterogeneous results. The mean effect sizes vary substantially and this is also the case for moderator variables. The variation may be due to different types of prevention (e.g., universal, selective, indicated), targets (e.g., child, family, school, or neighborhood), selection of primary studies, coding of variables, outcome measures, follow-up periods, methods of effect size calculation, fixed or random effect models of integration, and so forth. Meta-analyses revealed a broad range of significant moderators, but these are not identical in different syntheses. Some could be replicated more often than others; for example, larger effects in indicated prevention (at-risk groups), multimodal approaches, good program integrity, small samples, short follow-ups, and studies where the evaluators have been involved in the program development or implementation (e.g., Lösel 2012a; Lösel and Bender 2012). More specific moderators have been found for programs against school bullying (Farrington and Ttofi 2009; Ttofi and Farrington 2011). Effective programs include more positive modules, such as parent information, school meetings, schoolyard supervision, clear classroom rules, and disciplinary measures. However, in all these meta-analyses, one must bear in mind that the moderators are derived from different primary studies whose results are mainly short term and not yet well replicated.

Replication in correctional treatment

In the 1980s, we carried out a first meta-analysis on the treatment of adult offenders in German prisons (Lösel and Köferl 1989). Around the same time, Lipsey (1992a) published a much larger meta-analysis on the treatment of juvenile delinquents in North America. Both meta-analyses found an overall desirable effect, but the mean effect sizes were small (between about $d = 0.10$ and 0.20, depending on the method of analysis). There was much variation between the outcomes of different primary studies and both reviews showed various moderators of effect size.

Since the 1980s, there has been clear progress in correctional treatment (Bonta and Andrews 2017; Cullen 2013; Lösel 2012b; MacKenzie 2006). More sound evaluations have been carried out, the majority in North America and English-speaking countries. Systematic reviews and meta-analyses confirmed a mean desirable treatment effect on recidivism (Cullen 2013; Lipsey and Cullen 2007; Lösel 2012b; Wilson 2016). The mean effect sizes in most meta-analyses were positive (Wilson 2016). Compared to a recidivism rate of 50% in the control groups, Wilson (2016) estimated a mean reduction of about 10 percentage points due to treatment. Such moderate effects reduce victimization and can pay off in financial terms (Welsh and Farrington 2000). For the treatment of general and violent offenders, the typical mean effect sizes seem to be relatively homogenous (between $d = 0.20 \pm 0.10$; Lösel 2012b). In sexual offender treatment, they are more heterogeneous, that is, ranging from $d = 0.08$ to 0.54 (Lösel and Schmucker

2017), with meta-analyses on the treatment of young offenders at the upper end (Reitzel and Carbonell 2006; Walker et al. 2004). In spite of such encouraging results, there is still controversy about the effectiveness of sex offender treatment. This is due to often not well-controlled studies, small samples, different treatments, heterogeneous offender types, various comorbidities, variation in outcome measurement, handling of dropouts, and a wide range of follow-up periods (Lösel and Schmucker 2017).

As in developmental prevention, some heterogeneity of findings is normal in correctional treatment. Accordingly, the "what works" literature aims to show what is most effective and what has weak or no effects. The Maryland Report on Crime Prevention required at least two studies with positive findings and designs that were at least at level 3 of the scale of methodological rigor (Sherman et al. 2002). This was a plausible criterion, but the pattern of results is often complicated. For example, the widely used "Reasoning & Rehabilitation" program showed a desirable effect in several studies, but no effect in various others (Tong and Farrington 2006). Similarly, sound evaluations of cognitive-behavioral programs for sexual offenders revealed positive effects in some studies, but zero effects and even negative tendencies in others (Lösel and Schmucker 2005; Schmucker and Lösel 2015). These and other examples suggest that information about a mean effect has very limited value for practice.

To increase effectiveness and reproducibility, Andrews et al. (1990) proposed the risk–need–responsivity (RNR) model of appropriate treatment that became widely used in practice. Treatment showed positive mean effects when all three RNR criteria were fulfilled (Bonta and Andrews 2017). The effect sizes decreased when fewer principles were met and became even slightly negative when no criterion was fulfilled. This pattern has been replicated in meta-analyses on general offender treatment (Bonta and Andrews 2017), sexual offender treatment (Hanson et al. 2009), and young offender treatment (Koehler et al. 2013). In addition to RNR, many recent offending behavior programs integrate research on desistance (Farrall and Calverley 2006; Shapland et al. 2012), natural protective factors (Lösel and Bender 2003; Lösel and Farrington 2012), and the Good Lives Model (Ward and Brown 2004; Ward and Maruna 2007). The impact of such enrichments on reoffending is not yet well evaluated, but they are in accordance with broader RNR models of "what works" (Andrews et al. 2011; Lösel 1995).

Replicated moderators in meta-analyses play a key role in the explanation of heterogeneous treatment outcomes (Lipsey and Cullen 2007; Lösel 2012b). For young offender treatment, effects were larger in programs with a cognitive-behavioral concept, adherence to RNR, fidelity in implementation, ambulatory treatment, good descriptive validity, smaller samples, and demonstration projects (Koehler et al. 2013). Although there were more moderators by trend, the number of primary studies was too small for an adequate analysis. The same problem appeared in our recent meta-analysis of sexual offender treatment (Schmucker and Lösel 2015). The mean finding of 10.1% sexual recidivism in the program groups and 13.7% in the control groups was moderated by various factors. Studies with cognitive-behavioral treatment, small samples, medium- or high-risk offenders, more individualized program delivery, and good descriptive validity revealed better effects. In contrast to treatment in the community, prison programs showed no significant mean effect. These findings suggest that general statements about the effect or failure of sex offender treatment are inappropriate. It is plausible that sexual offender treatment in prisons is less effective (as compared to the

respective control groups) because there is no reality testing for child molesters or internet offenders in custody. However, this is not a sufficient explanation because treatment in forensic hospitals had a slightly better and significant effect (Schmucker and Lösel 2015). General criminogenic effects of incarceration (Durlauf and Nagin 2011; Nagin 2013) must also be considered. However, this explanation may not be sufficient because drug-addicted offenders seem to benefit from a closed institution (Lösel and Koehler 2014).

As in developmental prevention, evaluations of correctional treatment often contain some inconsistency within one and the same study. For example, Lösel and Pomplun (1998) carried out a matched-pairs evaluation of an educational program as an alternative to remand incarceration of young offenders. The findings were mixed. For example, we found nonsignificantly lower rates of any recidivism in the control group, but significantly lower rates of serious recidivism in the treatment group. We felt that this result was plausible and it now fits to current knowledge on larger treatment effects at medium to high risk than at low risk (e.g., Travers et al. 2013). However, did we really have a hypothesis on this differentiated result at the time of our study?

Studies on sex offender treatment also vary in their findings and this cannot simply be attributed to program or design differences (Lösel and Schmucker 2017; Schmucker and Lösel 2015). As mentioned above, custodial programs had no significant effect on the rate of sexual recidivism, but various studies suggest that there may be an impact on other outcomes, such as a lower rate of nonsexual reoffending, or more delayed or less harmful sexual reoffending (e.g., Olver et al. 2012; Smid et al. 2016). Evaluations of sexual offender treatment often raise more questions than answers (Grady et al. 2015). There are plausible theoretical, statistical, or practical explanations for the mixed pattern of results, but these should not only be provided post-hoc, but also form differentiated models of conditions under which treatment is successful.

Some scholars may argue that inconsistent results in offender treatment studies are due to a lack of theoretical foundation. This may only be partially true. Many programs are based on sound social learning or criminological theories (Bonta and Andrews 2017). Others apply more differentiated, eclectic, and case-oriented approaches to treatment that are supported by general research on psychotherapy (e.g., Beutler et al. 2016). The processes of individual change are more complex than the typical 3–5 group trajectories of correlational studies in developmental criminology (Jennings and Reingle 2012). Research is complicated by low correlations between theoretically meaningful proximal measures of therapeutic impact and their relation to later recidivism (Lipsey 1992b; McDougall et al. 2009; Woessner and Schwedler 2014). There are issues of social desirability and impression management in psychometrics, low base rates of reoffending (e.g., for sexual offenses), poor sensitivity of dichotomous recidivism criteria in official crime data, and other methodological factors. Sometimes, a theoretically meaningful explanation of heterogeneous findings can be as challenging as nailing a pudding on the wall.

Discussion and perspectives

It is the fundamental role (and privilege) of scientists to be neutral and to tell the truth as far as they know it. This includes being self-critical. Following the legacy of Joan

McCord, my lecture aimed to raise some problems of reproducibility in criminology. To avoid misunderstanding, it should be stated that criminology has made substantial progress in the fields of developmental prevention and correctional treatment. However, a realistic evaluation suggests that more differentiated and well-replicated findings are necessary. Would any criminologist drive over a bridge when s/he has been told that "on average" such bridges are solid, but 10% collapsed in a certain time period? Of course, it is not fair to compare criminology with engineering or the natural sciences, and the above introduction has shown that reproducibility is even a problem in these disciplines. The topics of this article are more similar to medicine, where many cures have limited effects, but no better alternatives are yet available. In the bridge analogy, people would perhaps drive over the risky construction if they have an urgent reason and know that nearly all collapses happened at times when there were overloaded trucks, heavy storms, and extreme temperatures. This would be an example for asking about the conditions under which a scientific explanation is more or less valid or an intervention is more or less justified.

The above sections have shown that there is much similarity in the findings on developmental crime prevention and offender rehabilitation. Not only the typical mean effect sizes but also large outcome variations are similar and suggest that the topic of reproducibility is relevant for criminology. Replication problems may be partly due to the complex longitudinal field experiments on both topics. Since there are rather consistent as well as inconsistent findings, I would not speak of a "reproducibility crisis" as it is discussed in psychology. However, obviously, there are problems of replication in criminology and these may not be limited to the two areas that are addressed in this article.

It would be worthwhile to analyze problems of reproducibility in other fields of criminology, for example, in the research on the origins of crime. For example, research on prominent theories like that on self-control has shown overall supportive but very heterogeneous results (e.g., Lösel 2017; Pratt and Cullen 2000; Walters 2016). More generally, Weisburd and Piquero (2008) found that the explanatory power of criminological theories is often low and leaves 80–90% of variance unexplained. More crime-specific theories showed somewhat stronger explanatory power than individual-based models. Accordingly, some research suggests that there may be superior effects of situational crime prevention (Clarke 1997) or place-based hot spots policing (Weisburd et al. 2008). However, situational and police-based crime prevention contain rather different programs. Although there are overall positive effects, systematic reviews vary substantially in their outcomes (Bowers and Johnson 2016; Telep and Weisburd 2016). In principle, situational crime prevention seems to contain similar problems of reproducibility as person-oriented approaches.

One should not polarize too much between both types of prevention, which are heterogeneous in themselves. Since a small group of persistent offenders is responsible for about half of all crimes (e.g., Farrington et al. 2006), person-oriented prevention and treatment of criminality is highly important. It should also be taken into account that situational crime prevention often refers to group/population data, whereas most person-oriented approaches use outcomes of single individual acts (e.g., recidivism) instead of more adequate aggregated behavior (Epstein and O'Brien 1985). Individual propensities and situational factors interact (Wikström et al. 2012) and, often, situation-oriented prevention also requires differentiation. A typical example is prevention

through CCTV, where the outcomes differ between countries, crime types, implementation contexts, and combinations of measures (Welsh and Farrington 2009).

These and other examples suggest that the issue of differentiation in replication is not only relevant for developmental prevention and offender treatment. Rossi (1978, 1987) rightly emphasized the importance of methodologically sound evaluations and criminologists repeatedly have underlined the need for more RCTs (e.g., Farrington 2003; Weisburd 2010). The Academy of Experimental Criminology, the ASC Division of Experimental Criminology, and the Campbell Crime and Justice Collaboration promote this aim. However, although criminology would benefit from more RCTs, the above-mentioned findings on replication in psychology and the natural sciences have shown that more experiments alone will not solve the reproducibility problem. Meta-analyses on person-oriented prevention programs revealed large differences in the outcome of RCTs on the same or very similar programs (Lösel and Beelmann 2003; Schmucker and Lösel 2015). Randomization enhances internal validity, but in comparison to other fields of criminology (Weisburd et al. 2001), it is not consistently correlated with effect sizes in treatment evaluations (Lipsey and Cullen 2007). RCTs are also vulnerable in studies with small samples, selective dropout, experimental rivalry, program diffusion, weak outcome measurement, and other threats to validity (Lösel 2007; Shadish et al. 2002).

Beyond the overall design quality, there are numerous influences on the outcome of program evaluations. In the field of correctional treatment, Lösel (2012b) integrated characteristics of programs, offenders, contexts, and evaluation methods in a model of influences on the effects. A slightly modified version is shown in Fig. 1.

Most of these factors are empirically supported by meta-analyses or single studies. Very similar influences seem to be relevant for the outcome heterogeneity in developmental prevention (Lösel 2012a). Not all of these moderators are yet empirically well founded and equally relevant. For example, the context "custody vs. community" is normally not relevant for developmental prevention, whereas personality traits of the target group are more important in correctional treatment.

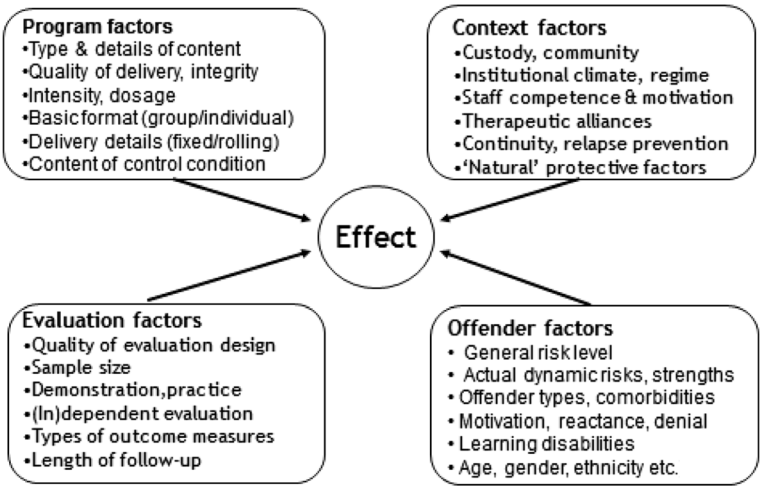

Fig. 1 A model of factors that may influence the effect of offender treatment programs

The model in Fig. 1 contains many empirically relevant factors. Although it is not a theory, it is obviously not in accordance with William of Ockham's (1287–1347) suggestion to keep explanations as parsimonious as possible. If one takes also principles of implementation science (Fixsen et al. 2009) into account, there could be more than 30 factors that are relevant for outcome. Although this would reflect the complexity of intervention, it leads to an information overload for practice, policy, and research designs. Therefore, I propose to select and test only the most relevant moderators in the respective field of intervention. Basic research on the human capacity for information processing found the magical number of seven, plus or minus two (Miller 1956). Perhaps this figure is a proper starting point for differentiations that are robust in replications.

Research on the most important influences on outcome heterogeneity will need fine-tuning. Following Rossi's metallic metaphors, something like a *Tin Can Law* would be a suitable analogy. It assumes some solid material in empirical findings, but one needs to squeeze it into an adequate shape to explain outcomes and guide future interventions. In areas with a substantial number of primary studies, meta-analyses play a key role in this process. They should be systematically reviewed for moderators that are most well replicated, for example, (1) a multimodal concept, (2) sound theoretical foundation, (3) integrity in delivery, (4) staff competence, (5) a favorable social context, (6) medium- to high-risk target groups, and (7) a not too large roll-out that allows proper monitoring. These characteristics should then be included and systematically tested in sound primary studies. As in multicentered treatment research in medicine, these primary studies should be designed as a series of replications to test the reproducibility of findings (e.g., in a meta-analysis). The evaluation of a restorative justice program by Sherman et al. (2015) is a good example of this strategy.

Research on differentiated knowledge about reproducible findings needs to be embedded in the general framework of enhancing replication: empirical studies should adhere to the recommendations and guidelines for sound and replicated evidence that have been made in various contexts; see The Steering Group of the Campbell Collaboration (2016), the standards of the Society of Prevention Research (Gottfredson et al. 2015), the recommendations of the Reproducibility Project in Psychology (Open Science Collaboration 2015), and the CONSORT standards of reporting (Hopewell et al. 2008). Only a few issues can be mentioned here: evaluation studies should not only be carried out by program developers but also by independent researchers. The reason for this is not that one must assume intentional misconduct of program owners; however, there are various decisions in a research process that may provide a more or less "unconscious" influence (e.g., subgroup allocation, definition and coding of variables, aggregation of data, significance testing, selective reporting). Empirical studies should be preceded by research protocols that would enhance transparency and reduce selective post-hoc reports on results. Researchers should mention their main hypotheses about expected findings. There should be replicated outcomes according to explicit criteria. As far as possible, studies should use multiple indicators of a construct, different informants, measurement times, sensitivity tests, and other techniques that allow an estimation of generalizability. This is also necessary with regard to the respective target population. Criteria of "sufficiently" replicated evidence should be explicit and harmonized in different registries. Findings of multiple evaluations should differentiate between efficacy in demonstration projects and effectiveness in routine practice.

These and other guidelines to promote, or at least estimate, reproducibility are not new and well based on evaluation methodology (Shadish et al. 2002). However, one should be aware that such standards are more easily requested than realized in the daily practice of research. Promoting replication studies is a stepwise process that requires adequate funding and dissemination strategies (Valentine et al. 2011). From a realistic perspective, it should also be taken into account that applied research in criminology often has an exploratory character. Most of these studies are not RCTs, but researchers aim to use the best quasi-experiment under given circumstances. Of course, I do not recommend low methodological standards, but too uniform and rigid guidelines may ignore the need for flexible strategies in the real world that Campbell (1969) has so well outlined. However, in any case, the respective research reports should adhere to the above-mentioned standards of reporting; for example, not only highlight positive results, but also provide information on zero or negative findings as well. To ensure transparency, any kind of study should report sufficient details not only on the methods, but also on institutional issues and potential conflicts of interest.

Meta-analyses play a key role in research on replication. Similar to the method of confirmatory factor analysis, there should also be approaches such as confirmatory meta-analyses to validate post-hoc findings on moderators. These would dig deeper into the conditions of program success or failure (Schmucker and Lösel 2011; Shaffer and Pratt 2009). These analyses could establish broader principles of "what works" instead of a too narrow focus on isolated programs (e.g., Beelmann 2012; Lösel 2012b). The extended RNR model (Andrews et al. 2011), the (recently modified) criteria of the Correctional Services Accreditation and Advice Panel of England and Wales (Maguire et al. 2010), and the revised standards for prevention programs (Gottfredson et al. 2015) contain such broader issues of moderating conditions. Unfortunately, research on moderators is very challenging (Lipsey 2003). Many moderators are confounded, interaction effects are difficult to replicate, and, often, there are not enough studies for sound (multivariate) analyses. Statistical criteria for outcome heterogeneity can avoid artifacts (Hunter and Schmidt 2004), but they cannot replace theoretically meaningful hypotheses.

More replicated research on moderators in program evaluations would make an important contribution to validate differential effects, that is, provide answers to the question of what works for whom, under what conditions, with regard to what outcomes, and why. There is also a need for more data on the impact of combinations of programs or of specific program elements or modules (e.g., Hawkins et al. 2008; Lipsey 2009). In demonstration projects, programs are typically evaluated in isolation, and this is most suitable for RCTs or sound quasi-experiments. In practice, however, programs may have different components or are combined with other interventions (e.g., cognitive-behavioral therapy, basic education, employment programs). In custodial offender treatment, for example, programs are more effective when they are combined with adequate measures of aftercare (e.g., Maguire and Raynor 2006). Evaluations of program packages are methodologically more difficult than those of isolated interventions. However, clinical pharmacy shows the need for this type of approach: when patients receive various medications, it is important to know the effect of combinations that may potentiate effectiveness or sometimes lead to negative side effects. Criminological program evaluation can also learn from engineering or climate research, where specific factors often have a minor effect in isolation, but, in

combination, they may show a strong impact. Of course, it is always more easy to say what should be done than carrying this out in research practice. However, I hope that I have shown both challenges and pathways of how developmental prevention, offender rehabilitation, and related areas can produce more well-replicated and differentiated results that are useful for practice and policy-making.

References

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

Andrews, D. A., Bonta, J., & Wormith, J. 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., Lieb, R., Mayfield, J., Miller, M., & Pennucci, A. (2004). *Benefits and costs of prevention and early intervention programs for youth*. Olympia: Washington State Institute for Public Policy.

Asscher, J. J., Deković, M., Manders, W., van der Laan, P. H., Prins, P. J. M., van Arum, S., & Dutch MST Cost-effectiveness Study Group. (2014). Sustainability of the effects of multisystemic therapy for juvenile delinquents in The Netherlands: Effects on delinquency and recidivism. *Journal of Experimental Criminology*, *10*, 227–243.

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*, 452–455.

Beelmann, A. (2012). The scientific foundation of prevention: The status quo and future challenges for developmental crime prevention. In T. Bliesener, A. Beelmann, & M. Stemmler (Eds.), *Antisocial behavior and crime: Contributions of developmental and evaluation research to prevention and intervention* (pp. 137–163). New York: Hogrefe Publishing.

Beutler, L. E., Someah, K., Kimpara, S., & Miller, K. (2016). Selecting the most appropriate treatment for each patient. *International Journal of Clinical and Health Psychology*, *16*, 99–108.

Boisjoli, R., Vitaro, F., Lacourse, E., Barker, E. D., & Tremblay, R. E. (2007). Impact and clinical significance of a preventive intervention for disruptive boys: 15-year follow-up. *British Journal of Psychiatry*, *191*, 415–419.

Bonta, J., & Andrews, D. A. (2017). *The psychology of criminal conduct* (5th ed.). Cincinnati: Anderson.

Borduin, C. M., Schaeffer, C. M., & Heiblum, N. (2009). A randomized clinical trial of multisystemic therapy with juvenile sexual offenders: Effects on youth social ecology and criminal activity. *Journal of Consulting and Clinical Psychology*, *77*, 26–37.

Boruch, R., Snyder, B., & DeMoya, D. (2000). The importance of randomized field trials. *Crime & Delinquency*, *46*, 156–180.

Bowers, K. J., & Johnson, S. D. (2016). Situational prevention. In D. Weisburd, D. P. Farrington, & C. Gill (Eds.), *What works in crime prevention and rehabilitation: Lessons from systematic reviews* (pp. 111–135). New York: Springer.

Bühler, A., Kötter, C., Jaursch, S., & Lösel, F. (2011). Prevention of familial transmission of depression: EFFEKT-E, a selective program for emotionally burdened families. *Journal of Public Health*, *19*, 321–327.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, *24*, 409–429.

Clarke, R. V. (1997). *Situational crime prevention: successful case studies* (2nd ed.). New York: Harrow & Heston.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Cohen, M. A., & Piquero, A. R. (2009). New evidence on the monetary value of saving a high risk youth. *Journal of Quantitative Criminology*, *25*, 25–49.

Conduct Problems Prevention Research Group (2002). Evaluation of the first 3 years of the Fast Track Prevention Trial with children at high risk for adolescent conduct problems. *Journal of Abnormal Child Psychology, 19*, 553–567.

Conduct Problems Prevention Research Group (2004). The effects of the Fast Track program on serious problem outcomes at the end of elementary school. *Journal of Clinical Child and Adolescent Psychology, 33*, 650–661.

Conduct Problems Prevention Research Group (2010). Fast Track intervention effects on youth arrests and delinquency. *Journal of Experimental Criminology, 6*, 131–157.

Cook, T. D. (2003). Why have educational evaluators chosen not to do randomized experiments? *The Annals of the American Academy of Political and Social Science*, *589*, 114–149.

Cullen, F. T. (2013). Rehabilitation: Beyond nothing works. In M. Tonry (Ed.), *Crime and justice in America: 1975–2025. Crime and justice: A review of research (vol. 42)* (pp. 299–376). Chicago: University of Chicago Press.

Curtis, N. M., Ronan, K. R., & Borduin, C. M. (2004). Multisystemic treatment: A meta-analysis of outcome studies. *Journal of Family Psychology*, *18*, 411–419.

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

Eckenrode, J., Campa, M., Luckey, D. W., Henderson Jr., C. R., Cole, R., Kitzman, H., Anson, E., Sidora-Arocleo, K., Powers, J., & Olds, D. (2010). Long-term effects of prenatal and infancy nurse home visitation on the life course of youths: 19-year follow-up of a randomized trial. *Archives of Pediatrics and Adolescent Medicine*, *164*, 9–15.

Eisner, M. P. (2014). *The South Carolina Triple P System Population Trial to prevent child maltreatment: Seven reasons to be sceptical about the study results*. Working paper. Violence Research Centre, Institute of Criminology, University of Cambridge.

Eisner, M., Nagin, D., Ribeaud, D., & Malti, T. (2012). Effects of a universal parenting program for highly adherent parents: A propensity score matching approach. *Prevention Science*, *13*, 252–266.

Eisner, M., Humphreys, D. K., Wilson, P., & Gardner, F. (2015). Disclosure of financial conflicts of interests in interventions to improve child psychosocial health: A cross-sectional study. *PloS One*, *10*(11), e0142803. doi:10.1371/ journal.pone.0142803.

Epstein, S., & O'Brien, E. J. (1985). The person–situation debate in historical and current perspective. *Psychological Bulletin*, *98*(3), 513–537.

Fagan, A. A., & Buchanan, M. (2016). What works in crime prevention? Comparison and critical review of three crime prevention registries. *Criminology & Public Policy*, *15*, 617–649.

Farrall, S., & Calverley, A. (2006). *Understanding desistance from crime: Theoretical directions in resettlement and rehabilitation*. Maidenhead: Open University Press.

Farrington, D. P. (2000). Explaining and preventing crime: The globalization of knowledge. *Criminology*, *38*, 1–24.

Farrington, D. P. (2003). A short history of randomized experiments in criminology: A meager feast. *Evaluation Review*, *27*, 218–227.

Farrington, D. P., & Petrosino, A. (2001). The Campbell collaboration crime and justice group. *The Annals of the American Academy of Political and Social Science*, *578*, 35–49.

Farrington, D. P., & Ttofi, M. M. (2009). School-based programs to reduce bullying and victimization. *Campbell Systematic Reviews*, *2009*, 6. doi:10.4073/csr.2009.6.

Farrington, D. P., & Welsh, B. C. (2007). *Saving children from a life of crime: Early risk factors and effective interventions*. New York: Oxford University Press.

Farrington, D. P., & Welsh, B. C. (2013). Randomized experiments in criminology: What has been learned from long-term follow-ups? In B. C. Welsh, A. A. Braga, & G. J. N. Bruinsma (Eds.), *Experimental criminology: Prospects for advancing science and public policy* (pp. 111–140). New York: Cambridge University Press.

Farrington, D. P., Coid, J. W., Harnett, L., Jolliffe, D., Soteriou, N., Turner, R., & West, D. J. (2006). *Criminal careers and life success: New findings from the Cambridge Study in Delinquent Development*. Research report no. 281. London: Home Office.

Farrington, D. P., Ttofi, M. M., & Lösel, F. A. (2016). Developmental and social prevention. In D. Weisburd, D. P. Farrington, & C. Gill (Eds.), *What works in crime prevention and rehabilitation: Lessons from systematic reviews* (pp. 15–75). New York: Springer.

Farrington, D. P., Gaffney, H., Lösel, F., & Ttofi, M. M. (2017). Systematic reviews of the effectiveness of developmental prevention programs in reducing delinquency, aggression, and bullying. *Aggression and Violent Behavior*, *33*, 91–106. doi:10.1016/j.avb.2016.11.003.

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

Flay, B. R., Biglan, A., Boruch, R. F., González Castro, F., Gottfredson, D., Kellam, S., Mościcki, E. K., Schinke, S., Valentine, J. C., & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151–175.

Gandhi, A. G., Murphy-Graham, E., Petrosino, A., Chrismer, S. S., & Weiss, C. H. (2007). The devil is in the details: Examining the evidence for "proven" school-based drug abuse prevention programs. *Evaluation Review*, *31*, 43–74.

Gottfredson, D. C. (2016). Why registries matter. *Criminology & Public Policy*, *15*, 651–659.

Gottfredson, D. C., Wilson, D. B., & Najaka, S. S. (2002). School-based crime prevention. In L. W. Sherman, D. P. Farrington, B. C. Welsh, & D. L. MacKenzie (Eds.), *Evidence-based crime prevention* (rev ed., pp. 56–164). London: Routledge.

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*, 893–926.

Grady, M. D., Edwards Jr., D., & Pettus-Davis, C. (2015). A longitudinal outcome evaluation of a prison-based sex offender treatment program. *Sexual Abuse: A Journal of Research and Treatment*, *27*, 1–28.

Hanson, R. K., Bourgon, G., Helmus, L., & Hodgson, S. (2009). The principles of effective correctional treatment also apply to sexual offenders: A meta-analysis. *Criminal Justice and Behavior*, *36*, 865–891.

Hawkins, J. D., Herrenkohl, T., Farrington, D. P., Brewer, D., Catalano, R. F., & Harachi, T. W. (1998). A review of predictors of youth violence. In R. Loeber & D. P. Farrington (Eds.), *Serious & violent juvenile offenders* (pp. 106–146). Thousand Oaks: Sage.

Hawkins, J. D., Brown, E. C., Oesterle, S., Arthur, M. W., Abbott, R. D., & Catalano, R. F. (2008). Early effects of communities that care on targeted risks and initiation of delinquent behavior and substance use. *Journal of Adolescent Health*, *43*, 15–22.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*, 135–175.

Henggeler, S. W., Schoenwald, S. K., Borduin, C. M., & Swenson, C. C. (2006). Methodological critique and meta-analysis as Trojan horse. *Children and Youth Services Review*, *28*, 447–457.

Henggeler, S. W., Schoenwald, S. K., Borduin, C. M., Rowland, M. D., & Cunningham, P. B. (2009). *Multisystemic treatment of antisocial behavior in children and adolescents* (2nd ed.). New York: Guilford Press.

Hopewell, S., Clarke, M., Moher, D., Wager, E., Middleton, P., Altman, D. G., Schulz, K. F., & CONSORT Group. (2008). CONSORT for reporting randomized controlled trials in journal and conference abstracts: Explanation and elaboration. *PLoS Medicine*. doi:10.1371/journal.pmed.0050020.

Hunter, J. E. (2001). The desperate need for replications. *Journal of Consumer Research*, *28*, 149–158.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks: Sage.

Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, *294*, 218–228.

Jennings, W. G., & Reingle, J. M. (2012). On the number and shape of developmental/life-course violence, aggression, and delinquency trajectories: A state-of-the-art review. *Journal of Criminal Justice*, *40*, 472–489.

Kellam, S. G., Brown, C. H., Poduska, J. M., Ialongo, N. S., Wang, W., Toyinbo, P., Petras, H., Ford, C., Windham, A., & Wilcox, H. C. (2008). Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. *Drug and Alcohol Dependence*, *95*, S5–S28.

Knerr, W., Gardner, F., & Cluver, L. (2013). Improving positive parenting skills and reducing harsh and abusive parenting in low- and middle-income countries: A systematic review. *Prevention Science*, *14*, 352–363.

Koehler, J. A., Lösel, F., Akoensi, T. D., & Humphreys, D. K. (2013). A systematic review and meta-analysis on the effects of young offender treatment programs in Europe. *Journal of Experimental Criminology*, *9*, 19–43.

Leschied, A., & Cunningham, A. (2002). *Seeking effective interventions for serious young offenders: Interim results of a four-year randomized study of multisystemic therapy in Ontario, Canada*. London: Centre for Children & Families in the Justice System.

Lipsey, M. W. (1992a). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In T. D. Cook, H. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. L. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation* (pp. 83–127). New York: Russell Sage Foundation.

Lipsey, M. W. (1992b). The effect of treatment on juvenile delinquents: Results from meta-analysis. In F. Lösel, D. Bender, & T. Bliesener (Eds.), *Psychology and law: International perspectives* (pp. 131–143). Berlin: de Gruyter.

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *The Annals of the American Academy of Political and Social Science*, *587*, 69–81.

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

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

Lipton, D. S., Martinson, R., & Wilks, J. (1975). *The effectiveness of correctional treatment: A survey of treatment evaluation studies*. New York: Praeger.

Littell, J. H. (2006). The case for multisystemic therapy: Evidence or orthodoxy? *Children and Youth Services Review*, *28*, 458–472.

Lösel, F. (1995). The efficacy of correctional treatment: A review and synthesis of meta-evaluations. In J. McGuire (Ed.), *What works: Reducing reoffending* (pp. 79–111). Chichester: Wiley.

Lösel, F. (2002). Risk/need assessment and prevention of antisocial development in young people. In R. Corrado, R. Roesch, S. D. Hart, & J. Gierowski (Eds.), *Multiproblem violent youth* (pp. 35–57). Amsterdam: IOS/NATO Book Series.

Lösel, F. (2007). Doing evaluation in criminology: Balancing scientific and practical demands. In R. D. King & E. Wincup (Eds.), *Doing research on crime and justice* (2nd ed., pp. 141–170). Oxford: Oxford University Press.

Lösel, F. (2012a). Entwicklungsbezogene Prävention von Gewalt und Kriminalität: Ansätze und Wirkungen [Developmental prevention of violence and crime: Approaches and effects]. *Forensische Psychiatrie, Psychologie, Kriminologie*, *6*, 71–84.

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

Lösel, F. (2017). Self-control as a theory of crime: A brief stocktaking after 27/42 years. In C. Bijleveld & P. van der Laan (Eds.), *Liber amicorum Gerben Bruinsma* (pp. 232–238). The Hague: Boom.

Lösel, F., & Beelmann, A. (2003). Effects of child skills training in preventing antisocial behavior: A systematic review of randomized evaluations. *The Annals of the American Academy of Political and Social Science*, *587*, 84–109.

Lösel, F., & Bender, D. (2003). Protective factors and resilience. In D. P. Farrington & J. Coid (Eds.), *Early prevention of adult antisocial behaviour* (pp. 130–204). Cambridge: Cambridge University Press.

Lösel, F., & Bender, D. (2012). Child social skills training in the prevention of antisocial development and crime. In D. P. Farrington & B. C. Welsh (Eds.), *Handbook of crime prevention* (pp. 102–129). Oxford: Oxford University Press.

Lösel, F., & Farrington, D. P. (2012). Direct protective and buffering protective factors in the development of youth violence. *American Journal of Preventive Medicine*, *43*(2), S8–S23.

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

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 [Residential care instead of pretrial detention of juvenile offenders: An evaluation]*. Pfaffenweiler: Centaurus.

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

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

Lösel, F., & Stemmler, M. (2012). Preventing child behavior problems in the Erlangen-Nuremberg Development and Prevention Study: Results from preschool to secondary school age. *International Journal of Conflict and Violence, 6*, 214–224.

Lösel, F., Schmucker, M., Plankensteiner, B., & Weiss, M. (2006). *Bestandsaufnahme und evaluation der Elternbildung [Survey and evaluation of parent education]*. Berlin: German Federal Ministry for Family Affairs, Senior Citizens, Women and Youth.

Lösel, F., Stemmler, M., Jaursch, S., & Beelmann, A. (2009). Universal prevention of antisocial development: Short- and long-term effects of a child- and parent-oriented program. *Monatsschrift für Kriminologie und Strafrechtsreform/Journal of Criminology and Penal Reform, 92*, 289–308.

Lösel, F., Stemmler, M., & Bender, D. (2013). Long-term evaluation of a bimodal universal prevention program: Effects on antisocial development from kindergarten to adolescence. *Journal of Experimental Criminology, 9*, 429–449.

MacKenzie, D. L. (2006). *What works in corrections: Reducing the criminal activities of offenders and delinquents*. Cambridge: Cambridge University Press.

Maguire, M., & Raynor, P. (2006). How the resettlement of prisoners promotes desistance from crime: Or does it? *Criminology and Criminal Justice*, *6*, 19–38.

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.

McCord, J. (1978). A thirty-year follow-up of treatment effects. *American Psychologist*, *33*, 284–289.

McCord, J. (1983). A forty year perspective on effects of child abuse and neglect. *Child Abuse and Neglect*, *7*, 265–270.

McCord, J. (1991). Family relationships, juvenile delinquency, and adult criminality. *Criminology*, *29*, 397–417.

McCord, J. (1992). The Cambridge–Somerville study: A pioneering longitudinal experimental study of delinquency prevention. In J. McCord & R. E. Tremblay (Eds.), *Preventing antisocial behavior: Interventions from birth through adolescence* (pp. 196–206). New York: Guilford Press.

McCord, J. (1994). Resilience as a dispositional quality: Some methodological points. In M. C. Wang & E. W. Gordon (Eds.), *Educational resilience in inner-city America: Challenges and prospects* (pp. 109–118). Hillsdale: Lawrence Erlbaum.

McCord, J. (2001). Psychosocial contributions to psychopathy and violence. In A. Raine & J. Sanmartin (Eds.), *Violence and psychopathy* (pp. 141–169). New York: Kluwer Academic/Plenum.

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

McDougall, C., Perry, A. E., Clarbour, J., Bowles, R., & Worthy, G. (2009). *Evaluation of HM Prison Service Enhanced Thinking Skills Programme. Ministry of Justice Research Series 3/09*. London: Ministry of Justice.

Mihalic, S. F., & Elliott, D. S. (2015). Evidence-based programs registry: Blueprints for healthy youth development. *Evaluation and Program Planning*, *48*, 124–131.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Murray, J., Irving, B., Farrington, D. P., Colman, I., & Bloxsom, C. A. (2010). Very early predictors of conduct problems and crime: Results from a national cohort study. *Journal of Child Psychology and Psychiatry*, *51*, 1198–1207.

Nagin, D. (2013). Deterrence in the twenty-first century. In M. Tonry (Ed.), *Crime and justice in America: 1975–2025. Crime and justice: A review of research* (vol. 42, pp. 199–263). Chicago, IL: University of Chicago Press.

Nowak, C., & Heinrichs, N. (2008). A comprehensive meta-analysis of Triple P-Positive Parenting Program using hierarchical linear modeling: Effectiveness and moderating variables. *Clinical Child and Family Psychology Review*, *11*, 114–144.

Ogden, T., & Amlund Hagen, K. (2006). Multisystemic treatment of serious behaviour problems in youth: Sustainability of effectiveness two years after intake. *Journal of Child and Adolescent Mental Health*, *11*, 142–149.

Olds, D. L., Henderson, C. R., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., Pettitt, L., Sidora, K., Morris, P., & Powers, J. (1998). Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial. *Journal of the American Medical Association*, *280*, 1238–1244.

Olds, D. L., Kitzman, H. J., Cole, R. E., Hanks, C. A., Arcoleo, K. J., Anson, E. A., Luckey, D. W., Knudtson, M. D., Henderson Jr., C. R., Bondy, J., & Stevenson, A. J. (2010). Enduring effects of prenatal and infancy home visiting by nurses on maternal life course and government spending: Follow-up of a randomized trial among children at age 12 years. *Archives of Pediatrics and Adolescent Medicine*, *164*, 419–424.

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 4716-3–4716-8.

Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology*, *1*, 435–450.

Petrosino, A., Boruch, R. F., Soydan, H., Duggan, L., & Sanchez-Meca, J. (2001). Meeting the challenges of evidence-based policy: The Campbell Collaboration. *The Annals of the American Academy of Political and Social Science*, *578*, 14–34.

Piquero, A. R., Jennings, W. G., & Farrington, D. P. (2013). The monetary costs of crime to middle adulthood: Findings from the Cambridge study in delinquent development. *Journal of Research in Crime and Delinquency*, *50*, 53–74.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Routledge.

Pratt, T. C., & Cullen, F. T. (2000). The empirical status of Gottfredson and Hirschi's general theory of crime: A meta-analysis. *Criminology*, *38*, 931–964.

Reitzel, L. R., & Carbonell, J. L. (2006). The effectiveness of sexual offender treatment for juveniles as measured by recidivism: A meta-analysis. *Sexual Abuse: A Journal of Research and Treatment*, *18*, 401–421.

Robinson, T. R., Smith, S. W., Miller, M. D., & Brownell, M. T. (1999). Cognitive behavior modification of hyperactivity—impulsivity and aggression: A meta-analysis of school-based studies. *Journal of Educational Psychology*, *91*, 195–203.

Rossi, P. H. (1978). Issues in the evaluation of human services delivery. *Evaluation Quarterly*, *2*, 573–599.

Rossi, P. H. (1987). The iron law of evaluation and other metallic rules. In J. Miller & M. Lewis (Eds.), *Research in social problems and public policy* (vol. 4, pp. 3–20). Greenwich: JAI Press.

Runkel, D., Lösel, F., Stemmler, M., & Jaursch, S. (2016). *Preventing social behavior problems in children from deprived migrant families: Evaluation of a child and parent training in Europe*. Paper submitted for journal publication.

Sanders, M. R. (2015). Management of conflict of interest in psychosocial research on parenting and family interventions. *Journal of Child and Family Studies*, *24*, 832–841.

Sanders, M. R., Markie-Dadds, C., Tully, L. A., & Bor, W. (2000). The Triple P-positive parenting program: A comparison of enhanced, standard, and self-directed behavioral family intervention for parents of children with early onset conduct problems. *Journal of Consulting and Clinical Psychology, 68*, 624–640.

Schmucker, M., & Lösel, F. (2011). Meta-analysis as a method of systematic reviews. In D. Gadd, S. Karstedt, & S. F. Messner (Eds.), *The SAGE handbook of criminological research methods* (pp. 425–443). Thousand Oaks: Sage.

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

Schweinhart, L. J. (2013). Long-term follow-up of a preschool experiment. *Journal of Experimental Criminology*, *9*, 389–409.

Sechrest, L. B., White, S. O., & Brown, E. D. (1979). *The rehabilitation of criminal offenders: Problems and prospects*. Washington, DC: National Academy of Sciences.

Serketich, W. J., & Dumas, J. E. (1996). The effectiveness of behavioral parent training to modify antisocial behavior in children: A meta-analysis. *Behavior Therapy*, *27*, 171–186.

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

Shaffer, D. K., & Pratt, T. C. (2009). Meta-analysis, moderators, and treatment effectiveness: The importance of digging deeper for evidence of program integrity. *Journal of Offender Rehabilitation*, *48*, 101–119.

Shapland, J., Bottoms, A., & Muir, G. (2012). Perceptions of the criminal justice system among young adult would-be desisters. In F. Lösel, A. E. Bottoms, & D. P. Farrington (Eds.), *Young adult offenders: Lost in transition?* (pp. 128–145). Milton Park: Routledge.

Sherman, L. W., Farrington, D. P., Welsh, B., & MacKenzie, D. (Eds.). (2002). *Evidence-based crime prevention*. New York: Routledge.

Sherman, L. W., Strang, H., Barnes, G., Woods, D. J., Bennett, S., Inkpen, N., Newbury-Birch, D., Rossner, M., Angel, C., Mearns, M., & Slothower, M. (2015). Twelve experiments in restorative justice: The Jerry Lee program of randomized trials of restorative justice conferences. *Journal of Experimental Criminology*, *11*, 501–540.

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

Sundell, K., Hansson, K., Löfholm, C. A., Olsson, T., Gustle, L. H., & Kadesjö, C. (2008). Multisystemic therapy and traditional services for antisocial adolescents in Sweden: Results from a randomized controlled trial after six months. *Journal of Family Psychology*, *22*, 550–560.

Telep, C. W., & Weisburd, D. (2016). Policing. In D. Weisburd, D. P. Farrington, & C. Gill (Eds.), *What works in crime prevention and rehabilitation: Lessons from systematic reviews* (pp. 137–168). New York: Springer.

The Steering Group of the Campbell Collaboration. (2016). *Campbell Collaboration Systematic Reviews: Policies and Guidelines*. *Campbell Policies and Guidelines Series No. 1*. doi:10.4073/cpg.2016.1.

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

Travers, R., Wakeling, H. C., Mann, R. E., & Hollin, C. R. (2013). Reconviction following a cognitive skills intervention: An alternative quasi-experimental methodology. *Legal and Criminological Psychology*, *18*, 48–65.

Tremblay, R. E., Pagani-Kurtz, L., Mâsse, L. C., Vitaro, F., & Pihl, R. O. (1995). A bimodal preventive intervention for disruptive kindergarten boys: Its impact through mid-adolescence. *Journal of Consulting and Clinical Psychology*, *63*, 560–568.

Ttofi, M. M., & Farrington, D. P. (2011). Effectiveness of school-based programs to reduce bullying: A systematic and meta-analytic review. *Journal of Experimental Criminology*, *7*, 27–56.

Valentine, J. C., Biglan, A., Boruch, R. F., González Castro, F., Collins, L. M., Flay, B. R., Kellam, S., Mościcki, E. K., & Schinke, S.P. (2011). Replication in prevention science. *Prevention Science*, *12*, 103–117.

van der Stouwe, T., Asscher, J. J., Stams, G. J. J., Deković, M., & van der Laan, P. H. (2014). The effectiveness of multisystemic therapy (MST): A meta-analysis. *Clinical Psychology Review*, *34*, 468–481.

Walker, D. F., McGovern, S. K., Poey, E. L., & Otis, K. E. (2004). Treatment effectiveness for male adolescent sexual offenders: A meta-analysis and review. *Journal of Child Sexual Abuse*, *13*, 281–293.

Walters, G. D. (2016). Are behavioral measures of self-control and the Grasmick self-control scale measuring the same construct? A meta-analysis. *American Journal of Criminal Justice*, *41*, 151–167.

Ward, T., & Brown, M. (2004). The good lives model and conceptual issues in offender rehabilitation. *Psychology, Crime and Law*, *10*, 243–257.

Ward, T., & Maruna, S. (2007). *Rehabilitation: Beyond the risk paradigm*. London: Routledge.

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

Weisburd, D., & Piquero, A. R. (2008). How well do criminologists explain crime? Statistical modeling in published studies. *Crime and Justice, 37*, 453-502.

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

Weisburd, D., Eck, J. E., Hinkle, J. C., & , Telep, C. (2008). The effects of problem-oriented policing on crime and disorder. *Campbell Systematic Reviews, 2008*, 14. doi:10.4073/csr.2008.14.

Weisz, J. R., Donenberg, G. R., Han, S. S., & Weiss, B. (1995). Bridging the gap between laboratory and clinic in child and adolescent psychotherapy. *Journal of Consulting and Clinical Psychology*, *63*, 688–701.

Welsh, B. C., & Farrington, D. P. (2000). Correctional intervention programs and cost–benefit analysis. *Criminal Justice and Behavior*, *27*, 115–133.

Welsh, B. C., & Farrington, D. P. (2009). Public area CCTV and crime prevention: An updated systematic review and meta-analysis. *Justice Quarterly*, *26*, 716–745.

Welsh, B. C., & Farrington, D. P. (2015). Monetary value of early developmental crime prevention and its policy significance. *Criminology & Public Policy*, *14*, 673–680.

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

Wilson, D. B. (2016). Correctional programs. In D. Weisburd, D. P. Farrington, & C. Gill (Eds.), *What works in crime prevention and rehabilitation: Lessons from systematic reviews* (pp. 193–217). New York: Springer.

Wilson, S. J., & Lipsey, M. W. (2007). School-based interventions for aggressive and disruptive behavior: Update of a meta-analysis. *American Journal of Preventive Medicine, 33*, S130–S143.

Wilson, D. B., Gottfredson, D. C., & Najaka, S. S. (2001). School-based prevention of problem behaviors: A meta-analysis. *Journal of Quantitative Criminology, 17*, 247–272.

Woessner, G., & Schwedler, A. (2014). Correctional treatment of sexual and violent offenders: Therapeutic change, prison climate, and recidivism. *Criminal Justice and Behavior*, *41*, 862–879.

Friedrich Lösel is an emeritus professor and past director at the Institute of Psychology, University of Erlangen-Nuremberg (Germany), and at the Institute of Criminology, Cambridge University (UK). At both places and at the Psychological University at Berlin he still has honorary functions. His research topics comprise juvenile delinquency, violence, offender treatment, prisoners and their families, football hooliganism, school bullying, psychopathy, resilience, close relationships, prison staff, and developmental prevention. He

has published more than 400 articles in journals and books and about 35 monographs, edited volumes and special journal issues. He is a recipient of the Sellin-Glueck Award of the American Society of Criminology (ASC), the Lifetime-achievement Award of the European Association of Psychology and Law, the Jerry Lee Award of the ASC Division of Experimental Criminology, the Lifetime-achievement Award of the ASC Division of Developmental & Life-Course Criminology (DLC), the Joan McCord Award of the Academy of Experimental Criminology, the German Psychology Prize, and the Stockholm Prize in Criminology. Currently he is president of the Academy of Experimental Criminology and chairman of the DLC Division of ASC.