Title: How Strong is the Evidence-Base for Crime Reduction End Users?

Citation*:* Tompson, L., Belur, J., Thornton, A., Bowers, K. J., Johnson, S. D., Sidebottom, A., Tilley, N., and Laycock, L. (in press). *Justice Evaluation Journal.* DOI: 10.1080/24751979.2020.1818275

* Corresponding author: Lisa Tompson (l.tompson@ucl.ac.uk)

Abstract

To support the development and implementation of evidence-based crime reduction, we systematically identified and appraised 70 systematic reviews of single crime reduction measures published between 1975 and 2015. Using the EMMIE framework, we find that the quality of reporting on the Effectiveness of crime reduction measures is reasonably strong, particularly in systematic reviews published by the Cochrane and Campbell Collaborations. By contrast, evidence concerning the Mechanisms underpinning a crime reduction intervention, the conditions that Moderate effectiveness, Implementation challenges and the Economic costs and benefits of crime reduction was largely absent from the assessed systematic reviews. We conclude that there is a distinct lack of systematic review evidence in crime reduction that currently speaks to the knowledge needs of practitioners (i.e., how to make an intervention 'work' for them).

Key words: crime reduction, EMMIE, evidence-based policing, meta-analysis, systematic review

Word count: 8075

Introduction

In medicine the notion that policy and practice should be evidence-based has come to be widely accepted over the last 30 years. More recently, a similar movement in favor of evidence-based decision-making has been sweeping through other public services, including those relating to crime and its control. The idea, in all domains, is that effectiveness and efficiency will be improved if those making decisions base them on robust research evidence combined with professional judgment, rather than, say, intuition, precedent, popular sentiment or political expediency. Although the evidence-based movement has taken hold at different rates in different policy areas, the direction is consistently towards increased expectations that decisions should be informed by high-quality research evidence and that improvements in outcomes will follow. Such 'high-quality' evidence is discerned through appraisal exercises, which form the bedrock of this movement, and the topic of this paper.

Few disagree with the basic idea of evidence-based decision-making. Any debate which has arisen relates mainly to the nature of 'evidence' in evidence-based decision-making. In policing, as in health and other fields, there have been two general perspectives on this (see Sidebottom and Tilley, 2020). The first maintains that there is an evidence hierarchy which places primacy on experimental methods and internal validity (Welsh, Braga and Bruinsma, 2013). The second perspective contends that all kinds of research evidence might inform decision-making providing it is relevant and subjected to necessary critical assessment (Laycock and Tilley, 2017; Brown et al., 2018). This perspective is associated with a broader view of what is necessary to inform policy and practice, specifically stressing the importance of context and mechanisms on outcome production (see Pawson and Tilley, 1997). However, both perspectives consider evidence syntheses such as systematic reviews to be the most reliable source of evidence to inform policy and practice (Gough et al. 2013; Neyroud, 2018).

The UK has largely embraced the notion of evidence-based policy and practice. In 2013, for example, a network of What Works Centers was established with a view to improving decision-making across various policy areas, making use of the best evidence available (see Gough, Maidment and Sharples, 2018). One network member was the What Works Centre for Crime Reduction (WWCCR), to which the research reported here relates. The WWCCR is intended to promote and facilitate the use of research evidence in decisions about crime reduction. Its objectives involve systematically identifying existing evidence syntheses on the effectiveness of crime reduction measures, assessing the identified review evidence and presenting it in an accessible and relevant format (see Thornton et al., 2019).

Promoting the systematic production and use of research evidence in the service of crime reduction is a complex and challenging task. For example, police receptivity to research evidence has been shown to vary considerably (Telep, 2016). Many police officers are wary of and unfamiliar with using research to inform practice (Tompson et al., 2017). Experience

and the 'craft' of policing tends to trump research evidence (Weisburd and Neyroud, 2011). Moreover, studies show that practitioners oftentimes consider research to be too ambiguous, jargon-heavy, and not attuned to police priorities (Fleming, 2011; Rojek et al., 2015; Hunter et al. 2015).

Initially, the UK What Works Centers largely followed the first perspective on what constituted good evidence. Studies that met the highest standards of evidence on the Maryland Scale (Sherman et. al., 1997) were preferred. However, in a five-year review of the progress of the What Works Centers, it was reported that:

'….the What Works approach, and the more robust methods on which it is founded – such as the use of randomised controlled trials (RCTs) *and* the more systematic analysis of what is working *where, and why* – is rapidly becoming the new normal.' (Halpern, 2018, p.4, emphasis added)

The shift towards greater attention being paid to the 'where' and 'why' of intervention effectiveness is mainly attributed to concerns about what practitioners and policy makers need to know when making use of research evidence to guide decision-making. For example, if practitioners do seek out research evidence on the effectiveness of an intervention, it is considered helpful to make clear how outcome results might vary according to the context in which an intervention is to be implemented (Hunter et al., 2015). Likewise, there is perceived value in articulating *how* a given intervention might produce the sought-after effects, and what needs to be in place to avoid implementation failure. Studies reporting net effects alone are considered to be of limited value for such a translational task. And, although the existing UK What Works Centers vary in their focus of concern, they share the remit of sorting and sifting studies to distil available evidence on interventions, producing new evidence syntheses, and commissioning new experiments to fill knowledge gaps[1], all with the intention of informing (and improving) policy and practice.

In support of the WWCCR, the 'EMMIE' framework was developed as a means through which research evidence could be appraised and presented to practitioners and policy makers in crime reduction (Johnson, Bowers and Tilley, 2015). 'EMMIE' covers five categories of evidence judged necessary for crime reduction. The initial 'E' refers to effects: decision-makers need evidence relating to the effectiveness of an intervention they are contemplating adopting. The initial 'M' refers to mechanisms: decision-makers need evidence relating to the ways in which a given intervention is thought to produce its effects, both intended and unintended. The second 'M' refers to moderators: decision-makers need evidence on the contextual conditions in which interventions are more or less likely to produce the sought-after results. 'I' refers to implementation: decision-makers need evidence on what needs to be done to effectively implement and sustain an intervention

[1] See https://whatworks.blog.gov.uk/about-the-what-works-network/

and what obstacles they are likely to encounter. Finally, the second 'E' refers to economic returns: decision-makers always have limited resources with alternative uses and therefore need to know what an intervention is likely to cost in relation to what returns they can plausibly expect.

In this paper we report the results of an appraisal exercise that assessed the quality of systematic review evidence in crime reduction using the five dimensions of EMMIE. We also provide an assessment of the extent to which the quality of evidence across each dimension has changed over time, and whether it is differentially reported in journals and other publication outlets, the latter being more likely to be accessed by practitioners (Rojek et al., 2012; Telep and Winegar, 2016). Finally, we discuss how the evidence-base might be better tailored to the evidence needs of scholars and crime reduction professionals.

The Current Study

The current study uses the EMMIE framework to assess the quality of evidence reported in 70 systematic reviews of single crime reduction interventions. Two research questions are addressed here. First, has the quality of systematic review evidence in crime reduction improved over time? And if so, are improvements apparent on all or just some of the EMMIE dimensions? Despite growth in the production of systematic reviews in crime reduction (Bowers et al. 2014; Pratt, 2014), presently no studies have assessed whether the quality of systematic review evidence has improved over time. This study therefore begins to fill this research gap.

The quality appraisal of evidence is a complex issue, that relates not only to methodological rigor but also to the relevance of a research design to answer a research question. As Sutcliffe et al. (2017: 131) note in relation to systematic reviews: "a research study judged as having high methodological standards may not necessarily be a suitable or relevant study for answering the review question". Hence, as well as using one of the many evidence appraisal tools that have proliferated in recent years, researchers also need to apply judgement in assessing to what extent a research question has been proficiently answered. In the current study we conceptualized evidence quality as the reliability of codified information that was presented in the reviews relating to the EMMIE dimensions. Such reliability related to internal validity for *effect*, to theoretical plausibility or external validity for *mechanisms, moderators* and *implementation* and to the level of detail on *economics* that would enable reliable judgements to be made based on this information.

Our second research question asks whether the quality of evidence varies across publication status. By this we mean are there differences in the quality of evidence reported in systematic reviews published in journals as compared to other outlets? The latter comprise technical reports, end-of-project reports to funding bodies, government publications, unpublished working papers, conference papers and dissertations. These sources have been traditionally known as the 'gray literature' and were for a long time assumed to be

methodologically inferior to journal papers (Wilson, 2009), since informally published works were not (typically) subjected to peer review. However, gray literature is now easier than ever to locate – indeed many electronic databases now index some forms of gray literature.[2] Moreover, many systematic reviews including those produced by the Cochrane and Campbell collaborations adhere to strict quality standards, and rigorous peer review is built into the publication process, as is also more commonly the case nowadays with government commissioned research. The historic distinctions between 'academic' and 'gray' literature have thus become increasingly blurred and we believe these labels to be inappropriate; we instead chose to use 'journal papers' and 'other publications' to demarcate the differences between publication types reported here.

We are interested in assessing the evidence quality of different publication types because the word limits that are (usually) imposed on journal papers may play to the advantage of other publications where there are simply more words available for authors to speak to the broad evidence encapsulated in the EMMIE model. In addition, other publications are often produced for policymakers and practitioners so they may be systematically more likely to include evidence on mechanisms, moderators, implementation and economics. As policymaking is increasingly being based on evidence, and systematic review evidence is an important (although by no means the only) part of the evidence base for informing policy advisors, it is considered timely to scrutinize all the literature in the field of crime reduction to assess its quality.

Materials and Method

We identified systematic reviews on crime reduction interventions using systematic methods (as reported in our protocol in Bowers et al., 2013). This involved: 1) keyword searches of 14 electronic databases; 2) hand searches of over 20 research and policing organization websites; 3) forward and backward citation analysis, and 4) a review of collections of systematic reviews on crime reduction.

Our search syntax centered around three key themes: crime, reduction, and evidence synthesis (for a detailed description see Tompson and Belur, 2016). These themes also informed the inclusion criteria used in this study. To be eligible for inclusion, a document had to: 1) employ systematic review and/or meta-analytic methods; 2) report an outcome measure that demonstrated a quantifiable impact on crime (*not* related behaviors such as aggression or anti-social behavior); 3) be published between 1975 and the end of the project (2015) and, 4) be written in English.

[2] See Kugley, Wade, Thomas, Mahood, Jørgensen, Hammerstrøm et al. (2016) for examples and one of the best collections of gray literature in criminal justice at https://njlaw.rutgers.edu/cj/gray/search.php

The initial database searches in December 2013 yielded 13,819 returns, with a further 1,130 studies found through the other search tactics (see Figure 4 in the Appendix). Following an extensive screening process, we identified 337 reviews that met our inclusion criteria. To produce a 'map' of the available systematic review evidence, these reviews were then 'light coded' by three researchers to extract basic information about each review (such as type of review and type of outcome data; see Bowers et al., 2014).

The 337 systematic reviews retained following light coding varied considerably. For example, some focused on specific single interventions (e.g. the impact of electronic monitoring on recidivism), whilst others considered a suite of interventions applicable to a particular problem or population (e.g. drink driving or juvenile offending). Since our aim was to support the online toolkit produced by the WWCCR which, like the Cochrane and Campbell Collaboration libraries, presented evidence at the intervention level, the decision was made to concentrate on single-intervention reviews only, rather than those targeting a specific population or crime type. This meant that any reviews that covered multiple interventions were excluded. This extra round of screening reduced the number of systematic reviews from 337 to 44.

In addition to the 44 reviews that were identified in the first stage of the systematic search, a further 15 reviews were identified outside of the formal search. These were identified via journal notification emails and through discussions with topic experts and were often due for publication in 2015 but not yet indexed in the electronic databases. Eleven additional systematic reviews were found via citation analyses of these 59 studies. Consequently 70 systematic reviews are covered in this paper.

A codebook was developed to extract and appraise the evidence reported in the identified systematic reviews and comprised six sections (Tompson et al., 2015).[3] The first covered *descriptive* information, such as study author(s), date, the category of the policy, program or intervention (hereafter referred to as 'intervention'),[4] the target of the intervention (victims, offenders and/or places), the outcome data type and the number of primary studies included in the review. The next five sections corresponded to the dimensions of EMMIE. Two components were distinguished for each dimension; EMMIE-E related to whether 'Evidence' reported in a review was present and EMMIE-Q related to the 'Quality' of that evidence (when present). For example, we extracted the effect size and type as the 'evidence' and then rated the rigor with which that was derived, along with other methodological factors, to generate a 'quality' score. To ensure that overall EMMIE-Q scores were consistent across review coders, logic rules, based on the EMMIE-Q answers to each dimension, were created to automatically suggest an appropriate score (see the coding tool by Tompson et al., 2015 for full details). For each Q-score dimension, the scale used was

[3] Available to be downloaded at: http://discovery.ucl.ac.uk/1462093/
[4] For this we adopted the same classification system as Weisburd et al. (2017).

zero to four, with the lowest score indicating no information or attention, and the highest indicating exemplary information or attention. The full coding procedure involved two researchers blind coding each review, with a moderation session to discuss any disagreements and to collectively agree the EMMIE Q-scores (sometimes involving other members of the research team).

The influence of potential biases on the validity of meta-analytic effect size results in crime policy has been previously acknowledged in the literature (e.g. Rothstein 2008). The effect section of the codebook adapted codes used by Crime Solutions®, an evidence appraisal system to which two of the authors (KJB and SDJ) had contributed. Codes were primarily devised to discern the extent to which threats to internal validity at *each* stage of the evidence synthesis process were addressed by the review authors, with the resulting EMMIE Q-score indicating the adequacy of efforts to mitigate threats to internal validity. For example, effect quality codes covered: 1) the search strategy; 2) statistical conclusion validity, which primarily related to the appropriate calculation of effect sizes; 3) sufficient assessment of various forms of risk of bias; 4) construct validity; 5) assessment of the influence of research design features and 6) assessment of the influence of unanticipated outcomes on the size of the effect. The logic rules for effect simply counted when the criteria thresholds were met for each of these six elements of validity. When review authors expressed an intention to assess the threats to internal validity (for example, to examine construct validity), but were prevented from doing so because the data did not lend themselves to such a task we coded the review as passing the threshold, since the authors were attentive to the issue.

The mechanism and moderator sections of the codebook were devised to account for a variety of reporting styles. For instance, EMMIE-Q codes were created to record if mechanisms were drawn from existing literature and/or elicited from practice. Similarly, we had separate codes to record whether the review authors had searched the literature for causally significant moderators and/or whether they consulted with practitioners or policymakers about the contextual factors which might influence intervention effectiveness. Each of these codebook sections had codes reporting whether the review had empirically tested any statements, or whether they were based solely on secondary information and/or conjecture. Logic rules generated higher quality scores (3 or 4) for the former and lower quality scores (0 – 2) for the latter.

To elaborate, for the mechanism dimension, the logic rules determined the quality score based on: whether the review mentioned mechanisms (not all did); whether the source of statements on mechanisms were provided; the thoroughness of the articulation of the theory of change, the thoroughness of the collection and analysis of data that tested the proposed mechanism/s, and whether the proposed mechanism/s were critically appraised in light of the analytic findings. For the moderator dimension, the logic rules were based on: whether the review mentioned moderators that were causally related to the mechanism/s;

whether the source of statements on moderators were provided; whether data was collected to test moderators; whether sub-group analysis was done; and whether that sub-group analysis was *a priori*, or post-hoc with variables that were at hand.

The implementation section extracted information on the barriers and enablers to putting an intervention in place and scored reviews on the sources of information and the level of detail that was reported. The logic rules that created the implementation Q-score related to: whether implementation features were mentioned in a review; whether source/s of information about implementation were provided; the thoroughness of a description of what was delivered in practice; whether the review authors identified enablers and obstacles to the implementation of the intervention; whether the review authors specified what is crucial to the successful implementation of the intervention; and whether the review specified what would (and would not) comprise a replication of the intervention.

The economics section was generated based on an allied project (see Manning et al., 2016), so that evidence on inputs, outputs, costs by bearer etc. were extracted, where available, and scored for comprehensiveness on the quality scale. The quality scores were derived from logic rules on the level of detail on monetized costs and/or benefits in the review. A review scoring 1 contained the estimates of direct costs, with a score of 2 additionally containing estimates for direct benefits (thus allowing a simple cost-benefit analysis or cost-effectiveness analysis). A score of 3 was given to reviews that estimated direct and indirect costs and benefits *and* marginal costs. Reviews scoring 4 had to additionally provide this information by the bearer (or recipient).

The quality ratings for mechanisms (M), moderators (M), implementation (I) and economics (E) were inspired by ideas from the field of realistic evaluation and proved harder to systematize than the quality ratings for effect. Although logic rules helped to standardize the quality ratings for MMIE, some subjectivity may have crept into this process, and the ratings therefore may not be entirely replicable by others. This is a common weakness in qualitative synthesis exercises and one to which we were alert. A codebook was developed with detailed descriptions of the quality ratings to guide other scholars when undertaking similar exercises (Tompson, et. al., 2015).

The EMMIE Q-scores were subjected to descriptive analysis along two dominant themes relating to the research questions in this study, and Pearson's correlation coefficients and Fisher’s exact test and Cramér’s V test were used to quantify an association between dependent and independent variables. To recapitulate, the central themes chosen for analysis were: (1) an analysis of change in the type and quality of evidence reported in systematic reviews in crime reduction over time and (2) a comparison of evidence from journal papers versus other publications.

Results

Descriptive Statistics

The identified reviews were first categorized by intervention. As shown in Table 1, most interventions were considered in only one review, but some were covered in several. In particular, evidence in relation to policies restricting the sale of alcohol over specific days/hours had been systematically reviewed four times, while those concerned with the effects of alcohol ignition interlocks, boot camps, cognitive behavioral therapy, mentoring and sobriety checkpoints had been subject to systematic review three times.

Intervention topics	N reviews covering intervention topic	N reviews
25	1	25
13	2	26
5	3	15
1	4	4
44		70

Table 1. Reviews broken down by topic

The distribution of quality scores across the five EMMIE dimensions is shown in Table 2 (full coding results can be found in the Appendix). Here we see that the quality of reporting on effect is reasonably strong; over half of the reviews scored a three or four. The quality of systematic review evidence in relation to mechanisms was lower, with the majority scoring two or fewer. The quality of evidence concerning moderators reported had a slightly more promising distribution, with the bulk of the reviews scoring two. Implementation followed a similar distribution to mechanisms, with most reviews scoring on the lower end of the quality scale. Finally, the quality of reporting on evidence relating to economics was strikingly poor, with the majority of reviews scoring zero using our criteria.

Q-score	Effect	Mechanism	Moderator	Implementation	Economics
0	4	9	10	11	64
1	13	36	21	30	5
2	14	15	29	16	1
3	25	8	5	12	0
4	14	2	5	1	0
Total	70	70	70	70	70

Table 2. Quality scores by evidence dimension N.B. a Q-score of 4 denotes the highest quality

The next section is organized according to the five categories of EMMIE. Each section discusses trends pertinent to the research questions set out above, trends over time and publication status.

Effect

As mentioned above, to score highly on the Effect dimension, a review needed to have a transparent and well-executed search strategy, exhibit high statistical conclusion validity, have sufficiently assessed the risk of bias, and demonstrated an awareness of the comparability of the data and the influence of study design features. The 70 reviews were initially differentiated in terms of those that did and did not employ meta-analysis. A competent meta-analysis is attentive to issues of internal validity, and hence it is not unreasonable to expect that those reviews that included a meta-analysis would score higher on the effect Q-scores. Table 3 shows this to be the case, and the difference was statistically significant ($p < 0.001$, two-tailed Fisher's exact test, Cramér's $V = 0.73$). It is interesting that two reviews[5] that did not employ a meta-analysis scored 3 for the effect Q rating. This was because the authors of those reviews anticipated many threats to validity (e.g. publication bias, unanticipated effects of the intervention) and would have performed a meta-analysis had the data been appropriate.

	Effect Q-score					Total
	0	1	2	3	4	
No meta-analysis	3	11	9	2	0	25
Meta-analysis	1	2	5	23	14	45
Total	4	13	14	25	14	70

Table 3. Reviews broken down by analytic strategy and EFFECT Q-score

When variation in the effect Q-scores are analyzed by time - using the periods 1995-2000, 2001-2005, 2006-2010 and 2011-2015 - there is only a marginal statistical relationship between effect quality and time ($p < 0.1$, two-tailed Fisher's exact test, Cramér's $V = 0.30$). These temporal bands were chosen for convenience, in the absence of any theoretical justification for a different banding.

Probing these further reveals that the promulgation of statistical advice and guidelines from organizations such as the Cochrane and Campbell collaborations may have contributed to a trend towards higher quality reviews in recent years. Systematic reviews produced for these organizations have a statistically significant relationship with higher quality effect Q-scores, compared to those produced for other organizations ($p < 0.01$, two-tailed Fisher's exact test, Cramér's $V = 0.50$). However, there is no statistically significant difference in effect Q-scores before and after the Campbell Collaboration was established in 2000 ($p = 0.61$, two-tailed Fisher's exact test, Cramér's $V = 1$), presumably because many other types of review were published during this period.

[5] See Babcock *et al* 2004's review of CBT for domestic violence offenders and Goss *et al* 2008's review on increased police patrols to reduce drink driving.

Analyzing the association between publication status and effect quality reveals a statistically significant relationship ($p < 0.01$, two-tailed Fisher's exact test, Cramér's V = 0.87). As shown in Table 4, proportionately more evidence syntheses published outside of academic journals had a higher effect Q-score, which can be explained by the high proportion of Campbell and Cochrane Collaboration reviews in our sample which are categorized as other publications (although related journal papers may also be published, most frequently in the Journal of Experimental Criminology – see Wilson et al., 2020). When reviews from the Campbell and Cochrane Collaboration organizations were excluded from the sample the statistical difference between journal papers and other publications disappeared.

Effect Q-score	Journal papers	Other publications	Total
0	3	1	4
1	12	1	13
2	9	5	14
3	12	13	25
4	3	11	14
Total	39	31	70

Table 4. Effect quality score by publication status

Mechanism

To achieve a high quality score on the Mechanism dimension, a review needed either to articulate a full description of the theory of change associated with the intervention being reviewed and to offer testable predictions of intermediate steps in that model (scoring 3), or go one step further and empirically test the theory of change proposed (scoring 4). As can be seen from Table 2 and Figure 2, reviews achieving these standards were rare. Most reviews reported a general statement of how an intervention was expected to work, with many (for example) invoking 'general deterrence'. Mechanisms could be derived from the literature or elicited from practitioners but, interestingly, no examples were found of the latter, even within the publications produced by practice-led organizations.

Inspecting Figure 1 by eye there seems to have been a slow shift towards higher quality in the reporting of details on mechanism in recent years, however, the change in Q-scores over time was not statistically significant ($p = 0.27$, two-tailed Fisher's exact test, Cramér's V = 0.26).

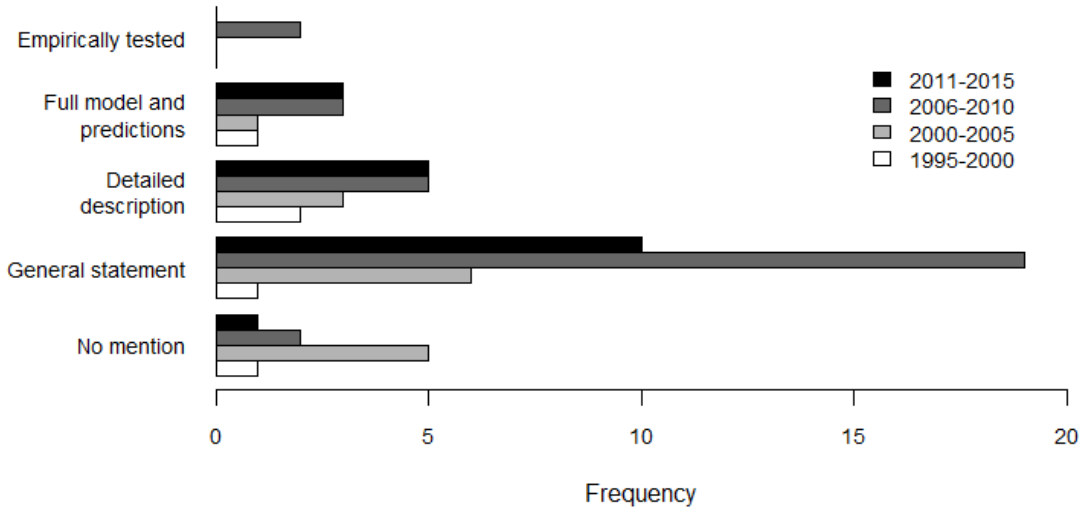

Figure 1 - Mechanism quality score over time

With respect to the (non-significant) relationship between mechanism and publication status, it was interesting that the two reviews that achieved the highest quality mechanism Q-scores were both doctoral theses[6], which are typically in-depth investigations of a topic area. All other scores saw similar proportions of reviews across journals and other publication outlets except the Q-score of zero, which was more common in in journal papers.

Moderator

To be awarded a high-quality score for moderators, a review had to provide a detailed account of contextual factors that might influence the activation of the hypothesized causal mechanism(s) and, by extension, the impact of an intervention (scoring 3). Reviews achieved the highest score if the moderators discussed were robustly tested via analysis (scoring 4). Hence, theory-based descriptions of moderators were afforded a higher score than those derived through the post-hoc analysis of available variables (e.g., gender, country).

Moderators were covered reasonably well across the sample of reviews examined here, potentially because moderator analysis features in most conventional texts on meta-analysis (e.g., Lipsey and Wilson 2001). Hence, reviewers are likely primed to consider that there may be factors that influence the effect dimension (e.g., meta-regression was specifically developed for this purpose).

No statistically reliable relationship was found between moderator Q-scores and time ($p = 0.39$, two-tailed Fisher's exact test, Cramér's $V = 0.24$), yet theory-based descriptions and

[6] These were Shaffer (2006) and Jonson (2010).

testing of moderators only really occurred from 2006 onwards. With regard to the relationship between publication status and moderator Q-scores, the scores were evenly split across journal papers and other publications for the lowest and highest scores, and any differences observed were non-significant (p = 0.41, two-tailed Fisher's exact test, Cramér's V = 1).

Implementation

Implementation is a major challenge in crime reduction (Homel and Homel, 2012). Failure to successfully implement a crime reduction initiative may cause wastage, harm and undermine effectiveness. To score highly on the implementation dimension, reviews had to provide a thorough description of what was delivered in practice, identify enablers and obstacles in the process, and specify what would be necessary to replicate a study. Reviews scoring three provided an evidence-based account of the level of fidelity to the program, policy or treatment plan. Reviews scoring four went one step further and analyzed the impact of structural/legal sources of departure from the intervention plan.

As seen in Table 3 and Figure 2, the most common rating for the implementation Q-score was one, which was awarded to reviews that provided only ad-hoc comments on implementation. The highest quality score of four was awarded to only one review, which was a dissertation on Drug Courts (see Shaffer, 2006). Eleven reviews did not mention implementation at all. These findings chime with Caudy et al's (2017) experiences of reviewing systematic reviews on justice population interventions. Of the 300 systematic reviews that Caudy's team assessed, only 8.7 per cent (26/300) included measures of implementation fidelity.

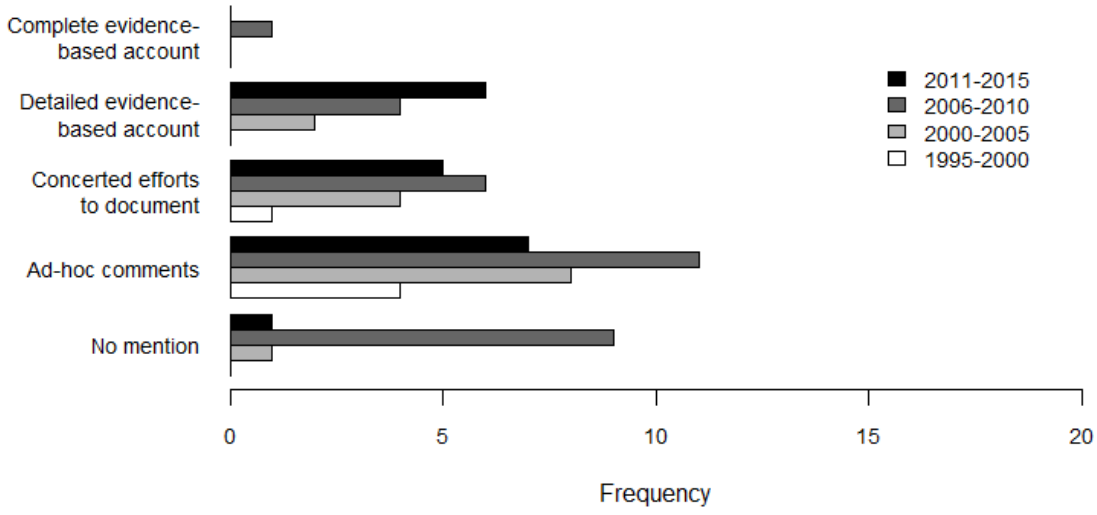

Figure 2 - Implementation quality score by time period

With regard to the different publication status of reviews, the observed differences were unremarkable (p = 0.73, two-tailed Fisher's exact test, Cramér's V = 1). This is perhaps

surprising as one might expect to see implementation covered in more detail in longer documents commissioned by policymakers. Our data only modestly support this assumption (15.4% of journal papers compared to 22.6% of other publications were rated three or four for implementation Q-score). One possible explanation for this is that process evaluations (containing valuable information about implementation) may be reported separately to outcome evaluations (containing outcome data). Since our inclusion criteria specified that crime reduction outcomes had to be reported at the *review* level, this might be biasing our sample towards those reviews synthesizing outcome evaluations to the neglect of process evaluation information.

Economics

Decision-makers always operate with limited resources. The economics of an intervention do not merely concern the cost of implementation, but also the potential benefits that it can provide. Efforts are ideally made to monetize benefits to provide a common currency with which to compare costs. To score highly on the economics dimension, reviews had to go beyond providing the direct and indirect costs and benefits of the intervention, which scored 2 in our economics Q-scores. Reviews scoring 3 had to provide enough information to state, or be able to calculate, the marginal costs, or the cost to particular subgroups and stakeholders. The top score of 4 was an idealized scenario whereby these costs and benefits were disaggregated by stakeholder, or bearer of the costs.

The dearth of economic information included within the reviews is striking (Table 3 and Figure 3). Only five reviews provided sufficient information to be awarded a score of one, with one additional review being awarded a score of two. The majority of reviews provided no information, or only passing reference to certain (usually direct) costs, such as the fact that an intervention cost less per participant than imprisonment. Those reviews that attempted to provide details of direct costs or benefits were on topics such as boot camps (two reviews), curfews, media campaigns to reduce drink driving and restorative justice. The only review that attempted to provide estimates of direct *and* indirect costs/benefits was on alley gating (Sidebottom et al. 2018). There was no discernible trend over time in the number of studies that reported economic data ($p = 0.75$, two-tailed Fisher's exact test, Cramér's $V = 0.17$).

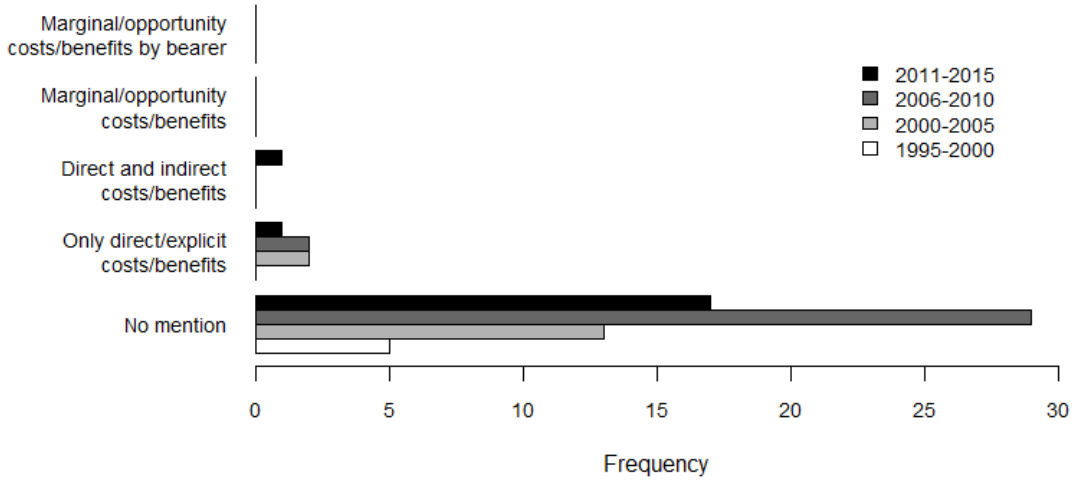

Figure 3 - Economics quality score by time period

Discussion

Systematic reviews are widely considered to be the most reliable source of evidence to inform policy and practice (Gough et al. 2013; Neyroud, 2018). To date, however, little attention has been paid to the coverage, type and quality of evidence reported in systematic reviews of crime reduction interventions. In this paper we used the EMMIE framework to assess the evidence reported in 70 systematic reviews of single crime reduction interventions published between 1975 and 2015.

Across the five dimensions of EMMIE, there was considerable variation in the quality of evidence reported in our sample of systematic reviews. For example, although over half of reviews (55%) achieved a high score (of three or four) for evidence quality, the comparable proportions were only 14.3 per cent for mechanisms and moderators (n=10 apiece), 18.6 per cent for implementation (n=13) and 0 per cent for economics. The latter finding is particularly noteworthy given the obvious importance of intervention costs and benefits. Moreover, the only (borderline) statistically appreciable improvement of evidence over time was seen for effect.

The lack of economic information in crime reduction reviews is conspicuous given the importance of this for decision-making in policy and practice. As such, it was disappointing to find that information about the costs and benefits of interventions were either not collected by those reporting on interventions, or not considered to be important enough to report on in reviews. This is clearly an avenue for future research[7].

[7] To this end, a practitioner-led, user-friendly data collection tool, which allows the direct, indirect and marginal costs of interventions to be recorded and calculated, has been created (Manning et al 2016), in the hope that this will increase the collection and use of such data.

Over half of the sample mentioned receiving some funding for the review. Consultative funding can entail some influence about how a review question is formulated. It may be the case that funding in the period 1975 – 2015 focused more on efficacy (the performance of an intervention under controlled conditions) than effectiveness (the performance of an intervention under 'real-world' conditions). If so, then review authors will have likely been encouraged to emphasize evidence on effect over other more contextual forms of evidence, and thus the dimensions of EMMIE will not have been given equal weighting in the research scope. The historical dearth of EMMIE-compliant studies may be part of the explanation why practitioners rarely use academic research findings to inform their decision-making.

Unsurprisingly, the high proportion of Cochrane and Campbell Collaboration reviews included in the 'other publications' category consistently scored highly for quality of evidence on effect. The reports published by these organizations follow a rigorous research process and are unconstrained by the word limits imposed by journals, so that there was simply more room to report the information on which we were making quality assessments. However, if this latter point holds, we might have expected to find that publication status was also associated with relationships for the other dimensions of EMMIE (mechanisms, moderators, implementation and economics), which was not the case. Nevertheless, the point about journal space representing a constraint remains. This is a problem with a simple solution. Many journals allow authors to provide additional supporting information that is published online alongside typeset journal articles. Commonplace use of this facility would give authors of systematic reviews (drawing on the primary evaluations on which systematic reviews are based) the opportunity to provide more detail on a fuller range of evidence dimensions without exceeding printed journal page limits (an example of this is the Journal of Experimental Criminology, see Wilson et al., 2020).

This study has limitations. Foremost are the inclusion criteria used to identify the sample of 70 systematic reviews. This required a quantifiable crime reduction outcome to be reported in reviews, which meant that reviews reporting other outcomes that might potentially be relevant to policy and practice were excluded. For example, the sample did not include non-crime outcomes such as substance abuse or school exclusions, which means that a considerable swathe of crime reduction evidence was excluded. The exclusive focus on effect in our inclusion criteria may have skewed the sample towards the reporting of evidence on effect to the detriment of the other dimensions. Had we omitted the requirement for a quantifiable crime reduction outcome to be reported we may have been quality assuring reviews which had synthesized process evaluations, and which contained more information on implementation, or reports on costs and benefits. This might have produced a more positive portrait on economic information.

Similarly, the decision to exclude multiple-intervention reviews will have likely skewed the sample of evidence appraised here. Multiple-intervention reviews often focus on interventions pertaining to a particular crime problem or population (e.g., what works to

prevent juvenile recidivism?). These are comparative in nature, for they aim to expose whether certain interventions are more effective, or more cost-effective. Hence, we might have seen economics covered more fully in these sorts of reviews than the ones analyzed herein.

Furthermore, since our evidence appraisal was based on what was reported in the 70 systematic reviews, one potential reason for the lack of information on mechanisms, implementation and economics was that it simply was not present in the primary studies on which each systematic review was based. It could therefore be that the quality of the systematic review level of evidence is indicative of the evidence base more generally on which syntheses rest. In our appraisal tool we were sympathetic to reviews that attempted to report on the evidence dimensions of interest but found in practice that the information was absent in the primary studies.

From the findings presented in this paper, we suggest that there are gains to be made from eliciting and testing evidence about mechanisms, moderators, implementation and economics from practitioners involved in the interventions, in both primary studies and reviews. When research is co-produced with practitioners, the resulting publications can be more in tune with the practical needs of policymakers who are deciding on and leading the implementation of interventions. Practitioners delivering crime reduction interventions are perhaps more likely to have working hypotheses and knowledge about implementation than researchers, that can be profited from in the service of evidence generation and consumption. More complete evidence reviews will also be of direct benefit to scholars by reducing the need to search a fragmented range of sources to gain a comprehensive view of an intervention. Our view is that this knowledge integration ought to be at the heart of evaluations on criminal justice topics.

There is an interconnectedness to the evidence dimensions reported here. For mechanisms to be activated, a thorough understanding of the contextual variations (the moderators) that support their activation is required. Poor implementation of an intervention may undermine both mechanisms and/or moderators, meaning that the desired effect is not achieved. The economic resources available for an intervention may compromise decisions about implementation or limit the rigor of the evaluation research design. These scenarios can undermine the degree to which the effect size of an intervention might be estimated. In contrast, a review that includes a cost benefit analysis is likely to lead to a more nuanced understanding of effect than those that do not. Disaggregating these evidence dimensions means that evaluations are partial, and the substantive interdependencies may be overlooked.

It might be argued that if a review (or primary study) does not adequately address effect size (and thus internal validity) then there is little point in considering the other evidence dimensions for informing practice. We, however, would disagree with this since prevention

interventions that tackle high impact low probability events are unlikely to be evaluated robustly with reference to associated outcomes. For example, terrorism, child sex exploitation and human trafficking are all challenging to evaluate due to low frequencies of events or chronic under-reporting, but that does not mean that practical lessons cannot be learned.

In addition, some interventions are aimed at a complex problem, for which the theory of change through which the intervention works is convoluted and fraught with measurement issues. An example of this is a systematic review focused on asset recovery interventions targeted at organized crime (Atkinson et al., 2017). In this case the effect of the interventions on organized crime rates could not be determined, partly because organized crime is so poorly defined and defuse, but the mechanisms through which it was expected to work were explicated, and it was possible to determine the extent to which they were implemented.

Although EMMIE is a relatively new framework in evaluation science, it is rooted in widely agreed upon evaluation principles that have a strong pedigree. For example, the realist focus on contexts and mechanisms (M and M), first mooted some 25 years ago, are echoed in the now common refrain, 'What works for whom in what circumstances and how'. Thus far, EMMIE has been adopted by other UK What Works Centres (notably in Children's Social Care - see Sheehan et al., 2008), policing scholars (Huey, 2018) and has now been used as an evaluation framework within primary studies as well as systematic reviews (Thornton et al., 2019). Time will tell if EMMIE influences how criminologists evaluate crime reduction interventions, but there is reason to believe that as a framework it has wider applicability than that presented here (Tilley and Westhorp, 2019).

It is perhaps premature to expect an evidence base to conform to the evidence expectations espoused by EMMIE. It is though somewhat damning that the evidence-base in crime reduction lacks critical information on many factors that are important to decision-makers. Thus far, EMMIE has not been subjected to the rigorous testing and critique that is customary in applied science fields. We do not claim that the framework is complete, or flawless, but do suggest that the elements of EMMIE are necessary (even if not sufficient) to the development of local crime reduction initiatives. We look forward to future research contributing to its development.

In our experience producing EMMIE-compliant systematic reviews is a more resource-intensive undertaking than a traditional review focused exclusively on effect. It requires careful consideration of whether to run two parallel sets of inclusion criteria for studies to be considered eligible for the review. It also necessitates the extraction of large amounts of qualitative data, and thoughtful construction of analytic frameworks to synthesize the qualitative findings alongside the quantitative findings. Should EMMIE gain traction in the crime reduction evaluation sphere it is possible that primary studies will increasingly be

reporting on the five evidence dimensions, which will smooth the synthesis process. However, until then systematic reviewers will need to think creatively, with a conceptual grounding in what evidence is required by end-users, to produce fully fit-for-purpose evidence.

References

Atkinson, C., Mackenzie, S., & Hamilton-Smith, N. (2017). A Systematic Review of the Effectiveness of Asset-Focussed Interventions against Organised Crime. Accessed 20 December 2019. Available at: https://dspace.stir.ac.uk/bitstream/1893/26091/1/Organised_crime_SR.pdf

Bowers, K., Johnson, S., Tilley, N., Tompson, L. and Belur, J. (2013). Protocol for Work Package 1: Identifying Existing Systematic Reviews of Crime Reduction. Accessed 20 December 2019. Available at https://discovery.ucl.ac.uk/id/eprint/1462098/

Bowers, K., Tompson, L., & Johnson, S. (2014). Implementing information science in policing: mapping the evidence base. *Policing: A Journal of Policy and Practice*, 8(4), 339-352.

Brown, J., Belur, J., Tompson, L., McDowall, A., Hunter, G., & May, T. (2018). Extending the remit of evidence-based policing. *International journal of police science & management*, 20(1), 38-51.

Caudy, M. S., Taxman, F. S., Tang, L., & Watson, C. (2016). Evidence Mapping to Advance Justice Practice. In D. Weisburd, D. P. Farrington & C. Gill (Eds.) *What Works in Crime Prevention and Rehabilitation* (pp. 261-290). New York, Springer.

Fleming, J. (2011). 'Learning to work together: police and academics', *Australasian Policing*, 3(2): 139–45.

Goss, C. W., Van Bramer, L. D., Gliner, J. A., Porter, T. R., Roberts, I. G., & DiGuiseppi, C. (2008). Increased police patrols for preventing alcohol-impaired driving. *Cochrane Database of Systematic Reviews*, (4).

Gough, D., Oliver, S., & Thomas, J. (Eds.). (2013). *An introduction to systematic reviews*. London, Sage Publications Ltd.

Gough D, Maidment C, Sharples J (2018). UK What Works Centres: Aims, methods and contexts. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London. ISBN: 978-1-911605-03-4. Available from https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3731

Halpern, D. (2018) Foreword in UK Government, (2018) The What Works Network: Five Years On Cabinet Office, HM Treasury. Accessed 20 November 2019. Available at:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/677478/6.4154_What_works_report_Final.pdf

Homel, R., & Homel, P. (2012). Implementing crime prevention: Good governance and a science of implementation. In B. C. Welsh & D. P. Farrington (Eds.) *The Oxford Handbook of Crime Prevention*, pp. 423-445. New York, Oxford University Press.

Hunter, G., Wigzell, A., May, T. and McSweeney, T. (2015). An Evaluation of the What Works Centre for Crime Reduction Year 1: Baseline, London: ICPR. Available at: https://eprints.bbk.ac.uk/21767/1/ICPR%20Evaluation%20of%20the%20WWCCR%20Year%201%20report%20(final)%2026th%20Feb%202015.pdf

Huey, L. (2018). What Do We Know About In-service Police Training? Results of a Failed Systematic Review. *Sociology Publications*. 40. https://ir.lib.uwo.ca/sociologypub/40

Johnson, S. D., Tilley, N., & Bowers, K. J. (2015). Introducing EMMIE: An evidence rating scale to encourage mixed-method crime prevention synthesis reviews. *Journal of Experimental Criminology*, 11(3), 459-473.

Jonson, C. L. (2010). The impact of imprisonment on reoffending: A meta-analysis (Doctoral dissertation, University of Cincinnati).

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A. M. K., Hammerstrøm, K., & Sathe, N. (2016). *Searching for studies: a guide to information retrieval for Campbell Systematic Reviews*. Campbell Methods Guides 2016: 1. https://doi.org/10.4073/cmg.2016.1.

Laycock, G., & Tilley, N. (2017). The why, what, when and how of evidence-based policing. In J. Knutsson & L. Tompson (Eds.) *Advances in Evidence-Based Policing* (pp. 26-42). Abingdon, Oxon, UK, Routledge.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications, Inc.

Manning, M., Johnson, S. D., Tilley, N., Wong, G. T., & Vorsina, M. (2016). *Economic analysis and efficiency in policing, criminal justice and crime reduction: What works?*. Basingstoke, Hampshire, UK, Palgrave Macmillan.

Nelson, M. S., Wooditch, A., & Dario, L. M. (2015). Sample size, effect size, and statistical power: A replication study of Weisburd's paradox. *Journal of Experimental Criminology*, 11(1), 141-163.

Neyroud, P. (2018). ,Systematic Reviews: "Better Evidence for a Better World. In R. Mitchell and L. Huey, (2018), *Evidence-Based Policing: An Introduction*. Bristol: Policy Press.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London, Sage Publications Ltd.

Pratt, T. C. (2014). Meta-analysis in criminal justice and criminology: What it is, when it's useful, and what to watch out for. In T. C. Pratt (Ed.) *Advancing Quantitative Methods in Criminology and Criminal Justice* (pp. 58-74). Abingdon, Oxon, UK, Routledge.

Rojek, J., Alpert, G. and Smith, H. (2012). 'The Utilization of Research by the Police.' *Police Practice and Research: An International Journal* 13(4): 329–341.

Rojek, J., Martin, P. and Alpert, G.P. (2015). Developing and Maintaining police-researcher Partnerships to Facilitate Research Use: A comparative analysis, New York, USA: Springer-Verlag.

Rothstein, H. R. (2008). Publication bias as a threat to the validity of meta-analytic results. *Journal of Experimental Criminology*, 4(1), 61-81.

Shaffer, D. K. (2006). Reconsidering drug court effectiveness: A meta-analytic review (Doctoral dissertation, University of Cincinnati).

Sheehan, L., O'Donnell, C., Brand, S.L., Forrester, D., Addis, S., El-Banna, A., Kemp, A. and Nurmatov, U. (2008) About the What Works Centre for Children's Social Care.Available at: https://assets.ctfassets.net/7swdj0fkojyi/2d9bU5LbiYQIUkiMy4MkMC/aae6df31dd8b86fdd69f62a0d5bfc568/Signs_of_Safety_a_mixed_methods_systematic_review.pdf

Sherman, L. W., Gottfredson, D. C., MacKenzie, D. L., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing crime: What works, what doesn't, what's promising: A report to the United States Congress*. Washington, DC: US Department of Justice, Office of Justice Programs.

Sidebottom, A. & Tilley, N. (2020). Evaluation evidence for evidence-based policing: Randomistas and realists. In Fielding, N., Bullock, K. and Holdaway, S. (eds) *Critical reflections on evidence-based policing*. Routledge.

Sidebottom, A., Tompson, L., Thornton, A., Bullock, K., Tilley, N., Bowers, K., & Johnson, S. D. (2018). Gating alleys to reduce crime: A meta-analysis and realist synthesis. *Justice Quarterly*, 35(1), 55-86.

Sutcliffe, K., Oliver, S., & Richardson, M. (2017). Describing and analysing studies. *An introduction to systematic reviews*, 123-144.

Telep, C. (2016). 'Police Officer Receptivity to Research and Evidence-Based Policing: Examining Variability Within and Across Agencies', *Crime and Delinquency*, doi: 10.1177/0011128716642253.

Telep, C. W., & Winegar, S. (2016). Police executive receptivity to research: A survey of chiefs and sheriffs in Oregon. *Policing: a journal of policy and practice*, 10(3), 241-249.

Tilley, N., & Westhorp, G. (2019). 'Evaluation Research, Quantitative.' In P. Atkinson, S. Delamont, A. Cernat, J.W. Sakshaug, & R.A. Williams (Eds.), SAGE Research Methods Foundations. doi: 10.4135/9781526421036872988).

Thornton, A., Sidebottom, A., Belur, J., Tompson, L. and Bowers, K. J. (2019). On the development and application of EMMIE: Insights from the What Works Centre for Crime Reduction. *Policing and Society, 29*(3), 266-282.

Tompson, L. A., Bowers, K. J., Johnson, S. D., & Belur, J. B. (2015). EMMIE evidence appraisal coding tool [Dataset]. London, UK: UCL Department of Security and Crime Science. Accessed 20 December 2019. Available at https://discovery.ucl.ac.uk/id/eprint/1462098/

Tompson, L., & Belur, J. (2016). Information retrieval in systematic reviews: a case study of the crime prevention literature. *Journal of Experimental Criminology*, *12*(2), 187-207.

Tompson, L., Belur, J., Morris, J., & Tuffin, R. (2017). How to make police–researcher partnerships mutually effective. In *Advances in evidence-based policing* (pp. 175-194). Routledge.

Weisburd, D. and Neyroud, P. (2011). 'New perspectives in policing: Police science—Toward a new paradigm', Harvard Executive Session on Policing and Public Safety, Washington, DC: National Institute of Justice.

Welsh, B. C., Braga, A. A., & Bruinsma, G. J. (Eds.). (2013). *Experimental criminology: Prospects for advancing science and public policy*. Cambridge University Press.

Wilson, D. B. (2009). Missing a critical piece of the pie: simple document search strategies inadequate for systematic reviews. *Journal of Experimental Criminology*, 5(4), pp. 429–440.

Wilson, D. B., Mazerolle, L., & Neyroud, P. (2020). Campbell Collaboration systematic reviews and the Journal of Experimental Criminology: Reflections on the last 20 years. *Journal of Experimental Criminology*. https://doi.org/10.1007/s11292-020-09433-y