J Exp Criminol (2015) 11:459–473
DOI 10.1007/s11292-015-9238-7

Introducing EMMIE: an evidence rating scale to encourage mixed-method crime prevention synthesis reviews

Shane D. Johnson[1] · **Nick Tilley**[1] · **Kate J. Bowers**[1]

Published online: 1 July 2015

Abstract
Objectives This paper describes the need for, and the development of, a coding system to distil the quality and coverage of systematic reviews of the evidence relating to crime prevention interventions. The starting point for the coding system concerns the evidence needs of policymakers and practitioners.
Methods The proposed coding scheme (EMMIE) builds on previous scales that have been developed to assess the probity, coverage and utility of evidence both in health and criminal justice. It also draws on the principles of realist synthesis and review.
Results The proposed EMMIE scale identifies five dimensions to which systematic reviews intended to inform crime prevention should speak. These are the Effect of intervention, the identification of the causal Mechanism(s) through which interventions are intended to work, the factors that Moderate their impact, the articulation of practical Implementation issues, and the Economic costs of intervention.
Conclusions Systematic reviews of crime prevention, and the primary studies on which they are based, typically address the question of effect size, but are often silent on the other dimensions of EMMIE. This lacuna of knowledge is unhelpful to practitioners who want to know more than what might work to reduce crime. The EMMIE framework is intended to encourage the collection of primary data regarding these issues and the synthesis of such knowledge in future systematic reviews.

Keywords Systematic reviews · Evidence quality · Evidence standards · Realist review · Meta-analysis

✉ Shane D. Johnson
shane.johnson@ucl.ac.uk

[1] UCL Department of Security and Crime Science, University College London, 35 Tavistock Square, London WC1H 9EZ, UK

Introduction

The volume of research relating to crime prevention is enormous, but of varying quality. Policymakers and practitioners who want to improve their decisions by drawing on evidence thus face a variety of problems. These include, for example, finding the evidence, assessing its quality, working out which evidence is relevant to their issues, and persuading stakeholders that policy and practice should accord with what the evidence suggests.

Systematic reviews (SRs) have emerged as a method for finding, sifting, sorting and synthesizing the findings of primary evaluations relevant to particular interventions. Methods have been developed for the conduct of SRs, including the process of selecting studies for analysis, and the statistical meta-analytic procedures used to summarize the overall impact(s) of intervention. Despite this, just like the primary evaluations on which they are based, SRs vary in quality and do so in ways that should be considered by those involved in evidence based policy.

Building on earlier work concerned with primary studies (Perry et al. 2010; Sidebottom and Tilley 2012), this paper focuses on the assessment of the evidence quality of SRs, and provides guidance for the conduct of future ones. Measures of effect size are discussed, but consideration is also given to other dimensions of importance to practitioners—the intended primary consumers of SRs. These include what an intervention actually comprises and the ease with which it can be implemented. While the work reported is primarily focused on SRs, many of the issues are equally germane to primary studies.

In what follows, we first consider existing efforts that have provided the means to assess the quality of evaluation evidence. We draw on research from public health and medicine as well as crime prevention. Next, we consider what practitioners need to know. SRs that most adequately attend to all of the issues of importance will be more valuable to practitioners and so we present a rating scale[1] designed to enable the systematic assessment of the quality of SRs of crime prevention initiatives, and to inform future ones.

Existing scales for assessing the evidence base

Scholars have noted that evaluations and clinical trials vary in quality, and that their reporting is often incomplete (e.g., Adetugbo and Williams 2000; Perry et al. 2010). In response, efforts have been undertaken to produce guidance regarding the conduct of primary studies (e.g., the CONSORT statement: Schulz et al. 2010; STROBE: von Elm et al. 2007), and SRs of them (e.g., AMSTAR, GRADE, PRISMA, RAMESES). In criminology, the Maryland Scale (Sherman et al. 1997) was developed to gauge the strength of individual studies according to their methodological rigor. It represents a 5-level hierarchy of evaluation evidence intended to indicate the extent to which an evaluation is able to rule out forms of bias as alternative explanations to a program effect. That is, it speaks to the issue of internal validity (see Campbell and Stanley

[1] Developed as part of an ongoing project, joint funded by the ESRC and UK College of Policing, to identify and rate existing SRs of what works to reduce crime.

1963). It says little, however, about the level of detail that authors should report about primary evaluations conducted (i.e., their 'descriptive validity', see Gill 2011).

In other disciplines, more effort has been invested in the provision of such guidance. In the case of primary studies, Moher et al. (2010) report the most recent incarnation of the CONSORT instrument. CONSORT 2010 is a 25-item checklist that focuses on the reporting of randomized controlled trials (RCTs). It primarily focuses on the extent to which study conclusions can reasonably be attributed to the treatment investigated (i.e., internal validity). The Cochrane Risk of Bias Scale (Higgins et al. 2011) also considers such issues, paying particular attention to the blinding of treatment providers, recipients and analysts, and problems with placebos. While systematic, these scales are silent on other types of validity (Sidebottom and Tilley 2012).

Apropos SRs, the AMSTAR (Shea et al. 2007), GRADE (Guyatt et al. 2008) and PRISMA (Moher et al. 2009) guidelines were developed to facilitate the assessment of the methodological quality of conducted studies (see also Higgins and Green 2011). Like the checklists for primary studies, however, they tend to focus on issues of internal validity.

Beyond internal validity

In their review of 302 meta-analyses of evaluations of diverse psychological, educational and behavioral treatments, Lipsey and Wilson (1993) concluded that:

> The proper agenda for the next generation of treatment effectiveness research, for both primary and meta-analytic studies, is investigation into which treatment variants are most effective, the mediating causal processes through which they work, and the characteristics of recipients, providers, and settings that most influence their results. (Lipsey and Wilson 1993: 1201)

Others have made similar suggestions (e.g., Cartwright and Hardie 2012; Rosenbaum 1988). The process of 'realist evaluation' and review has attempted to address such issues more directly (Pawson 2006; Pawson and Tilley 1997) and speaks to this agenda. In particular, realist studies explicitly focus on the causal 'mechanisms' through which interventions bring about their effects, the 'contexts' or conditions needed for treatments to activate potential causal mechanisms, and the 'outcomes' realized by the activation of causal mechanisms in the conditions in which they are introduced. What are produced in realist evaluations and reviews are Context, Mechanisms, Outcome pattern Configurations (CMOCs). This provides a framework for thinking about things other than effect size and factors that SRs might address.

To illustrate the importance of this, consider that interventions may bring about their effects in various ways. One example is the variation in mechanisms through which CCTV might reduce crime in car parks. These include, for example, the 'caught in the act' mechanism which leads to specific deterrence and incapacitation of the offender; 'you've been framed', where the offender perceives an increased risk; and 'memory jogging', where the presence of cameras reminds users to take precautions (Pawson and Tilley 1997: 55–82).

Crucially, the mechanisms being activated will depend on the particular conditions of the car park. For example, 'memory jogging' can only occur when the cameras are

positioned in observable places, and the 'you've been framed' mechanism will only be activated if offenders can see the cameras or are aware of them. SRs and primary evaluations alike can only tease out the possible mechanisms through which interventions work by articulating 'logic models' of how they might do so and collecting the necessary data to test them. SRs will, of course, be limited by what can be found in primary studies, but they should explicitly seek to locate such information, and indicate the absence of information as well as synthesize what is available.

In the case of SRs, the nearest counterpart to realists' mechanisms and contexts are meta-analysts' 'mediators' and 'moderators'. Mediators describe the chains of events (or intermediate outcomes) that occur between a treatment and the ultimate outcomes produced. In our CCTV example, mediator variables that might be used to test for activation of the 'caught in the act' mechanism include the volume of offenders identified on CCTV footage, and the number subsequently prosecuted. In the absence of evidence that offenders had at least been identified on CCTV footage, this mechanism would not represent a plausible explanation for any impact observed. Such data should not be difficult to obtain in primary evaluations, and systematic reviewers should have no difficulty in determining whether chains of causality have been explored in primary studies.

While the checklists discussed above are silent on these issues, the SQUIRE guidelines (Ogrinc et al. 2008), developed to inform primary studies of quality improvement in healthcare, are not. The authors draw on the realist approach (see also RAMESES: Wong et al. 2013) suggesting (for example) that primary studies should "describe the mechanisms by which intervention components were expected to cause changes, and plans for testing whether those mechanisms were effective" (p. 65). The SQUIRE guidelines thus represent a useful complement to those that focus on issues of internal validity. However, such guidance has yet to be incorporated into advice for the conduct or rating of SRs—the focus of this paper.

Moderators are equally important. They refer to variables that may explain variation in outcomes across different studies. They can include circumstances associated with differences in the efficacy of the intervention, such as the type of location. For example, CCTV may work more effectively in contained environments (e.g., car parks) than in open spaces (e.g., town centers). They can also include the study methods employed. For example, weaker effect sizes may be reported for RCTs than quasi-experimental studies (Weisburd 2010). While SRs typically consider the latter type of moderator, more attention could arguably be given to the former.

As suggested by Lipsey and Wilson (see also Cartwright and Hardie 2012; Weisburd et al. 2015), to better inform policy, the evidence base needs to speak to how interventions work and where and when they might do so most effectively. Consequently, when assessing the quality of the available evidence, in addition to considering the extent to which evaluations manage to rule out biases that might distort estimates of effect size, we also need to gauge the extent to which they contribute to understanding of the contexts/moderators relevant to the activation of the mechanisms/mediators that produce variations in outcome across differing sub-groups.

Despite their focus on internal validity, the CONSORT and SQUIRE guidelines for primary studies include items on the implementation of interventions, asking whether they provide sufficient detail to allow replication elsewhere or to determine whether they will be suited to particular situations. In a clinical trial, this would include the dose

of drug, and how and when it was administered. This is encouraging as implementation is rarely straight forward, but we suggest more is required.

Finally, because practitioners have limited budgets, resourcing one intervention means that something else must be forgone. Moreover, the most effective intervention tested will be of little practical value if it is prohibitively expensive to implement or maintain. Thus, to make good decisions, policymakers and practitioners need information on the overall costs and benefits of particular interventions and their alternatives. Current guidelines are typically silent on these issues.

The EMMIE framework

The preceding discussion suggests that the adequately evidence-equipped policymaker and practitioner need to know the following about interventions they might want to implement:

E the overall *effect* direction and size (alongside major unintended effects) of an intervention and the confidence that should be placed on that estimate
M the *mechanisms/mediators* activated by the policy, practice or program in question
M the *moderators/contexts* relevant to the production/non-production of intended and major unintended effects of different sizes
I the key sources of success and failure in *implementing* the policy, practice or program
E the *economic* costs (and benefits) associated with the policy, practice or program.

Both primary evaluations and SRs may attend to each of these more or less adequately. In assessing the evidence, it is thus important to differentiate between what the evidence suggests (e.g., an estimate of effect size) and the quality of that evidence (e.g., the methodological adequacy of the studies on which the estimate is based). With respect to assessing evidence quality, a key question concerns how meticulous the reviewers were in attending to each dimension. In the next sections, we discuss each in turn. As noted, we focus on SRs. We do so as their intended purpose is to synthesize evidence on treatments—an exercise which can provide practitioners with a good starting point in selecting interventions.

E - Effects: overall effect direction and size

The importance of producing unbiased estimates of mean effect sizes in SRs has been discussed elsewhere. For brevity, Table 1 summarizes the features of SRs that should be attended to in high-quality studies. To these we add (in the final row of Table 1) the assessment of unanticipated outcomes (e.g., quantification of crime displacement or a diffusion of crime control benefit, see Johnson et al. 2014).

Table 2 lists the types of evidence (referred to as 'EMMIE-E') that should be included in an SR to inform understanding of an intervention and on which assessments of quality should be based. In terms of assessing the *quality* of an SR on effect size, we suggest that the issues identified should inform a five-point scale as shown in the third column of Table 2 ('EMMIE-Q'). Table 3 lists the individual items that inform the EMMIE-Q summary rating.

Table 1 Factors that should inform the assessment of the methodological adequacy of an SR in terms of estimating effect sizes

Theme	Components (where appropriate)	Example sources
A transparent and well-designed search strategy*		Higgins and Green (2011)
High statistical conclusion validity (at least four of the following are necessary for a study to be considered sufficient)*	–Calculation of appropriate effect sizes –The analysis of heterogeneity –Use of a random effects model where appropriate –Attention to the issue of dependency –Appropriate weighting of individual effect sizes in the calculation of mean effect sizes	Lipsey and Wilson (2001) Borenstein et al. (2009) Borenstein et al. (2010) Hedges et al. (2010) Lipsey and Wilson (2001)
Sufficient assessment of the risk of bias (at least two necessary for sufficient consideration)*	–Assessment of potential publication bias –Consideration of inter-rater reliability –Consideration of the influence of statistical outliers	Hedges and Vevea (1996) Stock et al. (1982) Huffcutt and Arthur (1995)
Attention to the validity of the constructs, with only comparable outcomes combined and/or exploration of the implications of combining outcome constructs*		Petticrew and Roberts (2006)
Assessment of the influence of study design (e.g., separate overall effect sizes for experimental and quasi-experimental design)		Bryant and Wortman (1984)
Assessment of the influence of unanticipated outcomes or spin-offs on the size of the effect (e.g., quantification of displacement or diffusion of benefit)		

Items highlighted with an (*) symbol are particularly important for the EMMIE-Q rating (see Table 2)

Table 2 EMMIE evidence and five-point scales for assessing quality on each dimension

EMMIE component	EMMIE-E (evidence itself)	EMMIE-Q scoring
Effect	Effect size Moderator analysis Measurement/consideration of unanticipated effects	0. Insufficient consideration of validity elements listed above (in Table 1) 1: Sufficient consideration of one *element of validity 2: Sufficient consideration of two *elements of validity 3: Sufficient consideration of three or four *elements of validity 4: Sufficient consideration of five or six elements of validity (including all of those marked with an '*')
Mechanism/ mediator	Map of possible mechanisms/logic maps A priori mediator or mechanism-based moderator analysis Post hoc mediator or mechanism-based moderator analysis Assessment/statements of most likely mechanisms and any contextual conditions (these can be narratives)	0. No reference to theory; simple black box 1: Broad statement of assumed program theory stated (mechanisms and/or processes) 2: Detailed articulation of theory, based on interrogation of relevant literature and/or elicited from practice. 3: Formalization of theory and derivation of precise predictions from it 4: Test, corroboration, falsification and refinement of theories, using data assembled for the purpose.
Moderator/context	A priori context-based moderator analysis/ subgroup analysis (analysis testing the differences that context makes to outcome; theoretically driven) Post hoc context-based moderator analysis/ subgroup analysis (analysis testing the difference context makes to outcome; conducted due to data availability/not theoretically driven/ not mentioned prior to analysis) Statements qualifying contextual variations (these can be narratives)	0: No reference to condition contexts or moderators that may be significant for activation of mediators or mechanisms 1: Ad hoc description of possible relevant moderators or contexts 2: Tests of the effects of moderators or mechanisms defined post hoc using variables that are at hand 3: Theory-based pre-specification of expected moderators and mediators relevant to the activation of mediators or mechanisms 4. Collection and analysis of relevant data relating to the pre-specified expected moderators and contexts.
Implementation	A list/statement of key components necessary for implementation of reviewed interventions A list/statement of key components deemed necessary for replication elsewhere	0: No account of implementation or implementation challenges 1: Ad hoc comments on implementation 2: Systematic efforts to document implementation issues 3: Detailed evidence-based account of expected levels of fidelity to program, policy or treatment plans 4: Complete evidence-based account of expected levels of fidelity to program, expected obstacles and specification of elements necessary for replication elsewhere.

Table 2 (continued)

EMMIE component	EMMIE-E (evidence itself)	EMMIE-Q scoring
Economic	Quantification of inputs to the intervention Quantification intervention outputs Quantification of intensity (e.g., spend per head) Estimate of cost of implementation Estimate of cost of implementation by subgroup Estimate of cost-effectiveness per unit output Estimate of cost-effectiveness by subgroup Estimate of cost-benefit Estimate of cost-benefit by subgroup	0: No mention of costs (and/or benefits) 1: Only direct or explicit costs (and/or benefits) estimated 2: Direct or explicit and indirect and implicit costs (and/or benefits) estimated 3: Marginal or total or opportunity costs (and/or benefits) estimated 4: Marginal or total or opportunity costs (and/or benefits) by bearer (or recipient) estimated

M - Mechanisms/mediators: how the policy, practice or program produces its effects

In pharmaceutical medicine, prior to clinical trials, much laboratory work is undertaken to test and refine understanding of the chemical and physiological processes through which a drug produces its effects. Such background work is rarely undertaken in crime prevention, and hence the mechanism(s) through which an intervention might impact upon crime are often poorly understood prior to implementation.

Moreover, social interventions are generally complex. What is delivered may differ from one site and time to another and there can be long causal chains between the intervention implemented and effects realized. Working out what it is about an intervention that brings about its intended (and unintended) outcomes is thus of practical importance. A strong primary evaluation will explicate the underlying theory or theories of an intervention, and assemble the relevant data to test it. A strong SR will summarize these theories, and synthesize the available evidence to test them.

To do this, authors of an SR may need to engage with a wider literature than is necessary to estimate the effect size of an intervention. Such studies might more explicitly articulate the mechanism(s) through which an intervention is expected to work, or provide a test of this.

An example of such a review is provided by Weisburd et al. (2015), who conducted a SR of broken windows policing (Wilson and Kelling 1982). To test for evidence of the broken windows mechanism (that intervention reduces residents' fear of crime, and this in turn increases their willingness to act collectively to deter crime: p. 6), the authors searched not for studies that examined the impact of intervention on crime but for those that examined the impact on fear of crime and/or collective efficacy. They found no evidence to support this mechanism, but also concluded that "[t]here have simply been too few studies of the mechanisms underlying crime control in the broken windows policing model" (p. 11). We agree, and suggest that this is a more general issue in primary evaluations and SRs of them.

Table 3 EMMIE-Q individual elements for scoring existing SRs and checklist for new SRs

Effect	1a. Transparent well-designed search strategy
	1b. Calculation of appropriate mean effect size (ES)
	1c. Analysis of heterogeneity in ES
	1d. Attention to statistical dependency in effect sizes if appropriate
	1e. Appropriate weighting of individual ESs
	2a. Assessment of publication bias
	2b. Consideration of inter-coder reliability
	2c. Consideration of the influence of statistical outliers
	3. Attention to construct validity
	4. Moderator analysis of study design
	5. Assessment of unintended outcomes
Mechanisms	1. Mention of mechanisms and/or mediators
	2. Search of literature relating to mechanisms and/or mediators
	3. Discussion with stakeholders about mechanisms and reporting of these in the review
	4. Articulation of the theory of change, including sets of testable intermediate (or proxy) variables and patterns that would be observed in the data
	5. Collection and analysis of data that test whether mechansims and mediators are operating as expected
	6. Conclusions that corroborate, falsify or suggest refinements to the mediator or mechanisms theories
Moderators	1. Mention of moderators/contexts
	2. Search of literature for causally significant moderators or contexts
	3. Consultation with others (including practitioners) about the contextual factors that might matter
	4. Pre-specification of theoretically or empirically derived moderators or contexts
	5. Collection of data to allow the effects of context/moderators to be tested
	6. Subgroup/moderator analysis undertaken
	7. A priori moderator analysis undertaken theoretically
Implementation	1. Description of what was delivered in practice
	2. Identification (by review authors) of enablers and obstacles encountered when attempting to implement an intervention
	3. Specification of what is crucial to the successful implementation of the intervention
	4. Specification of what would represent a replication of the intervention
Economic	1. Collection of data to assess the *direct* financial revenue and set-up costs of the intervention as incurred by the provider
	2. Collection of data to assess the *indirect* financial revenue and set-up costs of the intervention
	3. Evidence-based quantification of the direct, financial costs of the intervention to the provider per unit cost of output (marginal costs)
	4. Evidenced-based quantified estimate of the direct financial provider costs per unit of (positive or negative) outcome
	5. Evidence-based estimate of the monetized financial and non-financial costs and benefits per unit of monetized financial and non-financial unit of intended and un-intended outcome
	6. Evidence-based estimate of the distribution by stakeholder of direct and indirect costs and benefits

Table 2 lists the types of evidence that could be included in a SR that seeks to explain *how an intervention works*. As with the rating of effect size, we propose a 5-point scale for assessing the quality of an SR on this dimension.

M - Moderators/contexts: conditions for the activation of the mediator or mechanism

Interventions rarely work unconditionally or equally effectively each time they are applied. The location (geographic or otherwise) and time they are implemented can affect the outcomes observed, as can the characteristics of those who receive or implement them. In deciding if, when, where and on whom to target a specific intervention, policymakers need evidence on which settings and subgroups are most likely to benefit from the intervention, which will most likely be unaffected, and which may have possibly negative outcomes. For this, estimates of mean effect sizes will be insufficient.

Most SRs include statistical moderator analyses to examine effect size variation across subgroups. However, the selection of subgroups varies, as does the rationale for choosing them. Subgroups should not be chosen using standard variables of convenience. Instead, the moderators selected should ideally be those for which the theory (mechanisms and mediators) suggests variations are to be expected. Of course, in the case of SRs, a moderator analysis requires a statistical approach, and so the study authors will be constrained by the data available in the primary studies. However, where the relevant data are unavailable this should be explicitly stated in the review with a view to mobilizing its collection in subsequent primary studies.

Table 2 lists the types of evidence that should be included in a SR to document the *contexts* in which an intervention works, and how the quality of the evidence and the thoroughness with which it was sought out may be assessed.

I - Implementation: how the policy, practice, treatment or intervention is applied

For both successful and unsuccessful initiatives, it is important for the practitioner to know what was done, what was crucial to the intervention and what difficulties might be experienced if it were to be replicated elsewhere. For example, SRs of hot spots policing (e.g., Braga and Weisburd 2012) suggest that this approach to crime reduction is successful. However, those intending to replicate previous efforts need to know more than this. They need to know what to do. This would include an indication of how a spatial hot spot is defined—what density of crime defines a hot spot, or what should inform the selection of such an area. In the case of police patrols, they need to know how many officers might need to be deployed, how frequently and for how long. They need to know whether the effect on crime depends on patrol dosage (e.g., the number of officer patrol hours per day per unit area). And so on. This is problematic for crime prevention because without such information, attempts at replication may vary considerably in terms of what is actually done (e.g., Tilley 1996). It is also problematic for evidence synthesis, as evaluations of interventions that prima facie appear to be the same thing, might actually be rather different, and in some cases it may be that nothing was implemented at all. In this case, the primary evaluator and the systematic reviewer should take account of this.

Finally, even simple interventions can be fraught with difficulties (e.g., Johnson and Loxley 2001; Knutsson and Tilley 2009). Thus, practitioners need to know if particular

interventions are easy or difficult to implement, if successful implementation is contingent upon particular conditions, and what is liable to impede or facilitate the process. We suggest that a strong review will focus on the issues listed in Table 2.

E - Economic analysis: the cost-effectiveness of the policy, practice, program, treatment or intervention

In policy terms, it is necessary but not sufficient that a given measure is capable of producing an intended outcome. In addition to the issues already discussed, the cost of intervention will ideally be known.

Estimating costs is complex. Comprehensive costing will include not only costs incurred by those responsible for the policy but also those falling on any third parties implicated in the delivery of interventions, the program participants themselves and those bearing any negative side effects ('indirect costs'). As programs expand, there are often diminishing marginal costs on those delivering interventions, as set-up and capital costs ('fixed costs') are spread over an increasing volume of activity, and so only those *variable costs* that are explicitly associated with increased output (e.g., police time) will increase.

Various forms of economic analysis exist, two of which will be briefly discussed. Cost effectiveness is *relatively* straightforward. It can speak either to the unit of output (e.g., cost of treatment per day per offender imprisoned) or the unit of outcome (e.g., cost per crime prevented). Such analysis helps to inform practitioners of what it may cost to deliver a given level of intervention, or crime reduction, and enables comparisons across interventions.

Cost–benefit calculations are more difficult as they require monetization of both the costs of intervention and (say) crimes prevented. This is particularly complicated as the range of those implicated expands, as unintended side effects are incorporated and as emotional as well as direct financial costs and benefits are swept into the calculations (see Farrell et al. 2004).

We will not discuss these forms of analysis further (but refer the interested reader to Farrell et al. 2004; McDougal et al. 2008), except to emphasize the fact that the estimation of costs should ideally enumerate the complete portfolio of costs that are necessary to implement an intervention. McDougal et al. (2008) suggest a rating scale to assess the methodological adequacy of SRs that includes a cost–benefit analysis, but this has no provision for rating SRs that include only a cost-effectiveness analysis. Since the latter are helpful to practitioners, Table 2 shows the forms of evidence that could be reported in a review and our proposed quality rating scale for this dimension of EMMIE.

Using 'EMMIE'

We have proposed five dimensions for rating the quality of SRs, described by the acronym EMMIE. Each dimension speaks to a different element of an SR, and may inform the decision-making or activity of different practitioners, or different stages of the policy-making process. Consequently, when rating reviews, we suggest that an EMMIE *profile* be produced rather than a single overall score. While our focus here has been on the rating of SRs, as noted above, with slight adaptation EMMIE scores can and should also be awarded to primary studies.

The use of EMMIE to rate existing SRs can help practitioners to assess the confidence they should place on the conclusions of a review. Applying the framework

on an ad hoc basis will be helpful, but efforts by a consortium of universities led by UCL, in collaboration with the UK College of Policing, are also underway to systematically rate existing SRs using EMMIE (see Bowers et al. 2014; note that future publications will discuss the practicalities of operationalizing the approach and provide empirical examples). The ultimate aim of the exercise is to provide practitioners with an online tool (hosted by the UK College of Policing) to assist their engagement with, and understanding of, the available evidence.

As well as being used to rate existing studies, it is our hope that the EMMIE framework will inform the conduct of future primary studies and SRs. At this point in time, we expect (and have started to find that) existing SRs achieve relatively lower ratings on the MMIE dimensions than they do for effect size (E). However, by encouraging researchers to explicitly focus on these issues in future primary studies and reviews of them, we hope that this will soon change.

With this in mind, three points are worthy of discussion. First, the research methods required to score high on each dimension are liable to differ, some depending heavily on quantitative methods, others on more qualitative approaches, such as realist synthesis (e.g., Pawson 2002). Thus, as is hinted in the title of this article, we encourage the use of mixed-method SRs. Second, to score high on all dimensions of EMMIE, future SRs will ideally employ broader inclusion criteria during the search stage of the review than is traditional, searching for research that addresses dimensions of EMMIE other than effect size. SRs are, of course, time consuming to conduct and hence some pragmatism will be required. Where an extended search proves to be impractical, we suggest that the review authors note this and synthesize what evidence *is* uncovered as it speaks to each dimension of EMMIE. Moreover, to set an agenda for primary studies, one role of future SRs will be to explicitly note the absence of evidence for each dimension of EMMIE (see also Gill 2011; Perry et al. 2010).

It is unlikely that any single primary study will or could score full marks on all dimensions of EMMIE. One reason for synthesizing diverse studies is to draw together what is known across all dimensions. Confining attention to the methodological adequacy with which effect sizes are estimated can establish with some certainty what has worked and hence what can work. Limiting attention in this way, however, is less useful in working out what will work, particularly in new conditions, and what needs to be present and what needs to be done to make something work as efficiently and as effectively as possible. Yet the latter are crucial for policy decisions. Consequently, the EMMIE framework is intended to catalyze both primary and secondary research that speaks to this agenda.

Acknowledgements This was undertaken as part of the Commissioned Partnership Programme: the What Works Centre for Crime Reduction. As this is delivered by a large consortium partnership in which ideas are freely exchanged, we would like to thank all of our colleagues who have assisted with this research. The opinions stated in this article are solely those of the named authors, and are not necessarily shared by other academics or organizations (such as the College of Policing). The research reported in this article was funded by the Economic and Social Research Council (ESRC) grant ES/L007223/1, and the College of Policing.

References

Adetugbo, K., & Williams, H. (2000). How well are randomized controlled trials reported in the dermatology literature? *Archives of Dermatology, 136*(3), 381–385.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*, 97–111.

Bowers, K., Tompson, L., & Johnson, S. D. (2014). Implementing Information Science in Policing: Mapping the Evidence Base. *Policing*, advanced online access.

Braga, A. A., & Weisburd, D. (2012). The effects of focused deterrence strategies on crime a systematic review and meta-analysis of the empirical evidence. *Journal of Research in Crime and Delinquency, 49*, 323–358.

Bryant, F. B., & Wortman, P. M. (1984). Methodological issues in the meta-analysis of quasi-experiments. *New Directions for Program Evaluation, 24*, 5–24.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin.

Cartwright, N., & Hardie, J. (2012). *Evidence-Based Policy: A Practical Way of Doing it Better*. Oxford: Oxford University Press.

Farrell, G., Bowers, K., & Johnson, S. D. (2004). Cost-benefit analysis for crime science: making cost-benefit analysis useful through a portfolio of outcomes. In M. Smith & N. Tilley (Eds.), *Launching Crime Science*. London: Willan.

Gill, C. E. (2011). Missing links: how descriptive validity impacts the policy relevance of randomized controlled trials in criminology. *Journal of Experimental Criminology, 7*(3), 201–224.

Guyatt, G. H., Oxman, A. D., Kunz, R., Vist, G. E., Falck-Ytter, Y., & Schünemann, H. J. (2008). What is "quality of evidence" and why is it important to clinicians? *British Medical Journal, 336*(7651), 995–998.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*(4), 299–332.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39–65.

Higgins J. P. T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Higgins, J., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savovic, J., Schulz, K., Weeks, L., & Sterne, J. A. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal, 343*, d5928.

Huffcutt, A. I., & Arthur, W. (1995). Development of a new outlier statistic for meta-analytic data. *Journal of Applied Psychology, 80*(2), 327–334.

Johnson, S. and Loxley, C. (2001) 'Installing Alley-Gates: Practical Lessons from Burglary Prevention Projects'. *Home Office Briefing Note 2/0*1. London. Home Office.

Johnson, S. D., Guerette, R. T., & Bowers, K. (2014) Crime displacement: what we know, what we don't know, and what it means for crime reduction. *Journal of Experimental Criminology, 10*(4), 549–571.

Knutsson, J., & Tilley, N. (2009). Introduction. In J. Knutsson & N. Tilley (Eds.), *Evaluating crime reduction initiatives* (pp. 1–6). New Jersey: Prentice Hall.

Lipsey, M., & Wilson, D. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *American Psychologist, 48*(12), 1181–1209.

Lipsey, M., & Wilson, D. (2001). *Practical Meta-Analysis*. London: Sage.

McDougal, C., Cohen, M. A., Perry, A., & Swaray, R. (2008). *Benefit-Cost Analyses of Sentencing: A Systematic Review*. Norway: Campbell Collaboration.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine, 151*(4), 264–269.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomized trials. *Journal of Clinical Epidemiology, 63*, e1–e27.

Ogrinc, G., Mooney, S. E., Estrada, C., Foster, T., Goldmann, D., Hall, L. W., Huizinga, M. M., Liu, S. K., Mills, P., Neily, J., Nelson, W., Pronovost, P. J., Provost, L., Rubenstein, L. V., Speroff, T., Splaine, M., Thomson, R., Tomolo, A. M., & Watts, B. (2008). The SQUIRE (standards for QUality improvement reporting excellence) guidelines for quality improvement reporting: explanation and elaboration. *Quality and Safety in Health Care, 17*(Suppl 1), i13–i32.

Pawson, R. (2002). Evidence-based policy and the promise of 'realist synthesis'. *Evaluation, 8*, 340–358.

Pawson, R. (2006). *Evidence-Based Policy*. London: Sage.

Pawson, R., & Tilley, N. (1997). *Realistic Evaluation*. London: Sage.

Perry, A. E., Weisburd, D., & Hewitt, C. (2010). Are criminologists describing randomized controlled trials in ways that allow us to assess them? Findings from a sample of crime and justice trials. *Journal of Experimental Criminology, 6*(3), 245–262.

Petticrew, M., & Roberts, H. (2006). *Systematic Reviews in the Social Sciences* (Chapter 1- Why do we need systematic reviews?). Oxford: Blackwell.

Rosenbaum, D. (1988). Community crime prevention: a review and synthesis of the literature. *Justice Quarterly, 5*(3), 323–395.

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine, 8*(1), 18. **1–9**.

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., Porter, A. C., Tugwell, P., Moher, D., & Bouter, L. M. (2007). Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology, 7*(10), 1–7.

Sherman, L., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing Crime: What Works, What Doesn't, What's Promising*. Washington DC: US Department of Justice Office of Justice Programs.

Sidebottom, A., & Tilley, N. (2012). Further improving reporting in crime and justice: an addendum to Perry, Weisburd and Hewitt (2010). *Journal of Experimental Criminology, 8*(1), 49–69.

Stock, W. A., Okun, M. A., Haring, M. J., Miller, W., Kinney, C., & Ceurvorst, R. W. (1982). Rigor in data synthesis: a case study of reliability in meta-analysis. *Educational Researcher, 11*(6), 10–14. **20**.

Tilley, N. (1996). Demonstration, exemplification, duplication and replication in evaluation research. *Evaluation: The International Journal of Theory, Research and Practice, 2*(1), 35–50.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., et al. (2007). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Medicine, 4*(10), e296. doi:10.1371/journal.pmed.0040296.

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*(2), 209–227.

Weisburd, D., Hinkle, J., Braga, A., and Wooditch, A. (2015). Understanding the Mechanisms Underlying Broken Windows Policing: The Need for Evaluation Evidence, *Journal of Research in Crime and Delinquency* (in press).

Wilson, J. Q., & Kelling, G. L. (1982). Broken windows: the police and neighborhood safety. *Atlantic Monthly, 211*, 29–38.

Wong, G., Greenhalgh, T., Westhorp, G., Buckingham, J., & Pawson, R. (2013). RAMESES publication standards: realist syntheses. *BMC Medicine, 11*(21), 1–14.

Shane D. Johnson is a professor in the Department of Security and Crime Science at University College London. He has particular interests in exploring how methods from other disciplines (e.g., complexity science) can inform understanding of crime and security issues, and the extent to which theories developed to explain everyday crimes can explain more extreme events such as riots, maritime piracy, and insurgency.

Nick Tilley is a professor in the Department of Security and Crime Science at University College London and an adjunct professor at Griffith University, Brisbane. His research interests lie in policing, crime prevention, the international crime drop and program evaluation methodology.

Kate Bowers is a Professor in Security and Crime Science at the UCL Department of Security and Crime Science. Kate has worked in the field of crime science for 20 years, with research interests focusing on the use of quantitative methods in crime analysis and crime prevention.