January 2004

Programs and Policies in Education, Crime and Justice and Social Welfare: Practical Recommendations Based on 14 Test-bed Reviews

Phoebe Cottingham
University of Pennsylvania

Rebecca A. Maynard
University of Pennsylvania, rmaynard@gse.upenn.edu

Matthew Stagner
Urban Institute

Reprinted from *Evaluation and Research in Education*, Volume 18, Issues 1 & 2, 2004, pages 28-53.
Publisher URL: http://www.multilingual-matters.net/

Programs and Policies in Education, Crime and Justice and Social Welfare: Practical Recommendations Based on 14 Test-bed Reviews

Abstract
Review teams tested the systematic review procedures and principles developed under the Campbell Collaboration. Fourteen review teams selected topics for intervention reviews in social policy, education, and criminal justice. Review protocols gave criteria for the extensive research literature search. Randomised Controlled Trials were selected. Systematic reviewers should give careful attention to defining the review topic, setting study inclusion and exclusion criteria, handling variability in outcome measurement and study reporting, appropriate uses of statistical meta-analysis, and reporting review results. Significant differences in review results were observed based on review criteria and procedures.

Keywords
systematic review, campbell collaboration, cochrane collaboration, test-bed, social policy, education

Comments
Reprinted from *Evaluation and Research in Education*, Volume 18, Issues 1 & 2, 2004, pages 28-53.
Publisher URL: http://www.multilingual-matters.net/

Synthesising Evidence on the Impacts of Programmes and Policies in Education, Crime and Justice, and Social Welfare: Practical Recommendations Based on 14 Test-bed Reviews[1]

Phoebe Cottingham
Institute of Education Sciences, US Department of Education, Washington, DC, USA

Rebecca Maynard
University of Pennsylvania, Philadelphia, USA

Matthew Stagner
Center on Labor, Human Services and Population, The Urban Institute, Washington, DC, USA

Review teams tested the systematic review procedures and principles developed under the Campbell Collaboration Fourteen review teams selected topics for intervention reviews in social policy, education, and criminal justice. Review protocols gave criteria for the extensive research literature search. Randomised Controlled Trials were selected. Systematic reviewers should give careful attention to defining the review topic, Setting study inclusion and exclusion criteria, handling variability in outcome measurement and study reporting, appropriate uses of statistical meta-analysis, and reporting review results. Significant differences in review results were observed based on review criteria and procedures.

Keywords: systematic review, Campbell Collaboration, Cochrane Collaboration, test-bed, social policy, education

Background

In 1999, a group of international scholars and policy makers conceived the 'Campbell Collaboration' for the purpose of fostering and disseminating systematic reviews of research evidence on questions of 'what works' in education, crime and justice, and social welfare (Table 1). The 'Campbell Collaboration' is modelled after the 'Cochrane Collaboration', which prepares and maintains systematic reviews of the effects of interventions in health care based primarily on evidence from randomised controlled trials (RCTs).[2] The 'Cochrane Collaboration' was among the first of many review groups that share the goal of facilitating evidence-based policy making. In the past four years, these groups have focused considerable attention and effort on sifting and sorting through research to develop lists of 'effective' programmes or policies.[3] Yet, across these groups, their standards of evidence and methods

0950-0790/04/01 028-26 $20.00/0
EVALUATION AND RESEARCH IN EDUCATION

Vol. 18, No. 1&2, 2004

Table 1 Partners in the Campbell Collaboration test-bed review project

Methods and quality of evidence	
Daniel Levy	Mathematica Policy Research, Inc., Washington, DC
David Myers	Mathematica Policy Research, Inc., Washington, DC
Steve Glazerman	Mathematica Policy Research, Inc., Washington, DC
Improving employment and family well-being	
Friedrich Losel	University of Erlangen-Nuremberg, Erlangen, German
Jane Reardon-Anderson	Urban Institute, Washington, DC
Jennifer Ehrle	Urban Institute, Washington, DC
Laura Winterfield	Urban Institute, Washington, DC
Matthew Stagner	Urban Institute, Washington, DC
Promoting socially desirable adolescent behaviours	
Lauren Scher	University of Pennsylvania, Philadelphia, PA
M. Cay Bradley	University of Pennsylvania, Philadelphia, PA
Psycho-social interventions addressing behaviour problems of children and families	
Julia Littell	Bryn Mawr College, Bryn Mawr, PA
Mark Lipsey	Vanderbilt University, Nashville, TN
Nana Landenberger	Vanderbilt University, Nashville, TN
Community interventions focused on crime and delinquency	
Heather Strang	Australian National University
Anthony Braga	Kennedy School, Harvard University, Cambridge, MA
Larry Sherman	Fels Center for Government, University of Pennsylvania, Philadelphia, PA
Interventions for children struggling in school	
Gary Ritter	University of Arkansas, Fayetteville, AK
Sherri Lauver	University of Pennsylvania, Philadelphia, PA
Susan Zief	University of Pennsylvania, Philadelphia, PA
Interventions to improve marriage and relationships and to reduce domestic violence (Tab 10)	
Katherine Kortenkamp	Urban Institute, Washington, DC
Lynette Feder	Portland State University, Portland, OR

of review tend not to yield consistent conclusions on similar questions. Moreover, the various methods require very different levels of investment and have different levels of transparency.

The Campbell Collaboration is the most prescriptive of these new systematic review initiatives. Key principles of the Campbell Collaboration's guidelines for conducting reviews include the following:

(1) Review topics should be chosen to avoid unnecessary duplication of effort.
(2) Review methods should minimise bias, primarily by employing high standards of scientific evidence, broad-based search strategies and avoiding conflicts of interest.
(3) Reviewers should commit to and facilitate routine updating to incorporate new evidence.
(4) Reviews should address policies that have current relevance and report findings for outcomes that matter to people.

Reviewers adhere to a review protocol that has been approved by experts in the subject area and in systematic review methods.[4]

As early participants in Campbell Collaboration reviews, we endorsed these principles for reviews. Yet, it was not clear how to design and conduct reviews that met the standards and guiding principles. We shared concerns that the underlying conditions for reviews conducted in medicine for the Cochrane Collaboration differed in important ways from those we confronted in education, crime and justice, and social welfare. For example, intervention studies in the social sciences most often rely on non-experimental study designs, whereas those in medicine most often use RCTs. Furthermore, it is rare in the social sciences to find replications of highly standardised intervention models, whereas, in medicine, similarly designed clinical trials often are conducted in multiple locations.[5] At the same time, experimental designs, standardisation and replication are now being adopted or promoted in social and education evaluations because of the clarity and higher confidence in findings obtained through well designed and implemented RCTs, as compared with findings from quasi-experiments.

Our commitment to the Campbell Collaboration principles led us to embark on a project aimed at testing the principles through direct application – asking specialists with subject matter and evaluation knowledge to do systematic reviews and collaborate in the 'test-bed' project. There are 14 review teams currently working with us. A substantial amount of time has been spent on designing, testing and then redesigning the standards that seem most appropriate for systematic reviews of programme and policy impact findings, and determining strategies for most effectively pulling together and reporting the findings from a systematic review. The lessons learned from the 'test-bed' experiment with 'Campbell Collaboration' principles should help others attempting similar efforts.

The Test-bed Experiences

The test-bed consists of 14 ongoing reviews following the 'Campbell Collaboration' review model, together with a process analysis of the findings from this set of pilot reviews. The ultimate goal of the test-bed project is to produce a practitioners' guide to conducting meaningful reviews of programme impact evaluation studies in the social sciences. Table 2 summarises the early products produced thus far from the test-bed reviews. These include protocols, briefs on particular methodological issues or review tools, and some published articles that have either drawn on or are feeding into the test-bed effort.

There have been two phases of the test-bed effort. During the first year, reviewers selected topics and began to develop review protocols following the Campbell Collaboration guidelines and format (http://www.campbellcollaboration.org/C2EditingProcess%20doc.pdf).[6] However, as each review team progressed through the protocol development stage to the actual implementation of the review, a number of critical design issues emerged. Some of the issues, if left unresolved, would have 'sunk' reviews, due to the resources required to comply with particular 'guidelines' – for example, the expectation that the review team will search the published literature worldwide or aggressively seek all fugitive literature.

One of the most significant design issues the teams grappled with related to the considerable ambivalence about whether reviews should exclude or include nonexperimental studies, and uncertainty as to what constituted appropriate guidelines for reporting out findings based on nonexperimental studies, if they were to be included in reviews.[7] Early findings from a review of tests of nonexperimental methods that were commonly judged to have reasonable prospects of producing unbiased impact estimates, similar to those already found in experiments raised serious concerns about the credibility of nonexperimental estimates of programme impacts (Glazerman *et al*., 2003). As a result, it was decided to restrict the test-bed reviews to RCT studies.[8] Similarly, to maintain a focus on only highly credible findings in the reviews, test-bed reviewers were asked by the test-bed leaders to report *intent-to-treat (ITT)* findings and to exclude findings for subgroups formed after random assignment on the basis of criteria that potentially could have been affected by the intervention under study. This latter condition meant, for example, that in a study focused on the impacts of after school-programmes, researchers would exclude any estimates of impacts that pertain to only those youth who participated in the programme for a minimum period of time, as such estimates would not have the credibility of the original randomisation of students to the after-school programme or the control group. Together these second phase guidelines considerably reduced the ambiguity and uncertainty confronting reviewers.

Table 2 Campbell Collaboration test-bed products (May 2003)

Methods	
1	Glazerman, Steven, Dan Levy, and David Myers. Nonexperimental Versus Experimental Estimates of Earnings Impacts. Princeton, NJ: Mathematica Policy Research, Inc. May 2003. (Tab 2)
2	Myers, David. Systematic reviews and the use of random assignment and quasi-experimental designs. Memo to Phoebe Cottingham, May 2003. (Tab 2)
3	Maynard, Rebecca, and Matthew Stagner. Reasons to Include a Statistical Meta-Analysis in a Systematic Review of Program Effectiveness Research. Philadelphia, PA: University of Pennsylvania, May 2003. (Tab 6)
	Maynard, Rebecca, and Matthew Stagner. Suggestions for Facilitating the Review Process. Memo to the C2 Steering Committee, February 23, 2003. (Tab 6)
4	Zief, Susan, and Sherri Lauver. Checklist for Assessing Study Quality. Philadelphia, PA: University of Pennsylvania, May 2003.
Improving employment and family well-being	
5	Visher, Christy A., and Laura Winterfield. A Systematic Review of the Effects of Non-Custodial Employment Programs on the Recidivism Rates of Ex-Offenders. Washington, DC: The Urban Institute, April 2003 (Protocol)
	Visher, Christy A., Laura Winterfield, Mark B. Coggeshall, and William Turner. Systematic Review of Non-Custodial Employment Programs: Impact on Recidivism Rates of Ex-Offenders. Washington, DC: The Urban Institute, May 2003.
6	Losel, Friedrich, and Andreas Beelmann. Efficacy of Child Skills Training in Preventing Antisocial Behavior and Crime. Erlangen, Germany: University of Erlangen-Nuremberg, April 2003. (Protocol)
	Losel, Friedrich and Andreas Beelmann. 'Effects of Child Skills Training in Preventing Antisocial Behavior: A Systematic Review of Randomized Evaluations.' Annals, AAPSS, May 2003 (pp. 84-109).
7	Stagner, Matthew, Jennifer Ehrle, and Jane Reardon-Anderson. Systematic Review of the Impact of Mandatory Work Policies on Family Structure. Washington, DC: Urban Institute, February 24, 2003. (Protocol)
	Stagner, Matthew, Jennifer Ehrle, Jane Reardon-Anderson, and Katherine Kortenkamp. Systematic Review of the Impact of Mandatory Work Policies on Family Structure. Washington, DC: Urban Institute, March 31, 2003.
Promoting socially desirable adolescent behaviours	
8	Scher, Lauren and Matthew Stagner. A Systematic Review of Teen Pregnancy Prevention Interventions. Philadelphia, PA: University of Pennsylvania, April 28, 2003 (Protocol, Version 3)

Table 2 (*Continued*)

	Scher, Lauren and Matthew Stagner. A Systematic Review of Teen Pregnancy Prevention Programs. Philadelphia, PA: University of Pennsylvania, April 29, 2003. (Working Draft Version 1)
9	Bradley, M.C. A Systematic Review of Interventions for Disruptive Behavior Disorders and the School. Philadelphia, PA: University of Pennsylvania, April 29, 2003. (Protocol)
Psycho-social interventions addressing behaviour problems of children and families	
10	Lipsey, Mark, and Nana Landenberger. Cognitive-Behavioral Programs for Juvenile and Adult Offenders: A Meta-Analysis of Randomized Controlled Intervention Studies. Nashville, TN: Vanderbilt University, January 2003. (Protocol)
	Lipsey, Mark, and Nana Landenberger. Preliminary Summaries of RCTs. Nashville, TN: Vanderbilt University, May 2003.
11	Littell, Julia, Burnee Forsythe, and Melania Popa. Impacts of Multisystemic Treatment on Youth Outcomes. Bryn Mawr, PA: Bryn Mawr College, February 24, 2003. (Protocol)
	Littell, Julia, Burnee Forsythe, and Melania Popa. Preliminary Summary of Randomized and Possibly-randomized Studies of Outcomes of Multisystemic Treatment (MST). Bryn Mawr, PA: Bryn Mawr College, May 23, 2003.
Community interventions focused on crime and delinquency	
12	Strang, Heather, and Lawrence Sherman. Effects of Face-to-Face Restorative Justice on Repeat Offending and Victim Satisfaction. Canberra, Australia: Australian National University. March 2003. (Protocol, Version 2)
	Strang, Heather, and Lawrence Sherman. Effects of Face-to-Face Restorative Justice on Repeat Offending and Victim Satisfaction. Canberra, Australia: Australian National University, May 16, 2003. (Working Draft)*
13	Braga, Anthony. Systematic Review of the Effects of Hot Spots Policing on Crime. Cambridge, MA: Kennedy School of Government, Harvard University, March 29, 2003. (Protocol).
	Braga, Anthony. 'Effects of Hot Spots Policing on Crime.' Annals, AAPSS, 578, November 2001, (pp. 104-125).
	Braga, Anthony. Hot Spots Policing and Crime Prevention: Evidence from Five Randomized Controlled Trials. Cambridge, MA: Kennedy School of Government, Harvard University, May 23, 2003. (Working Draft)

Table 2 (*Continued*)

Interventions for children struggling in school	
14	Ritter, Gary, and Rebecca Maynard. Review of the Evidence on the Effectiveness of Volunteer Tutoring Programs. Fayetteville, AR: University of Arkansas, February 22, 2002. (Protocol)
	Ritter, Gary, and Rebecca Maynard. Evidence on the Effectiveness of Volunteer Tutoring. Fayetteville, AR: University of Arkansas, April 2003. (Working Draft)
15	Zief, Susan, Sherri Lauver, and Rebecca Maynard. Impacts of After-School Programs on Student Outcomes. Philadelphia, PA: University of Pennsylvania, December 9, 2002. (Protocol, Version 1).
	Zief, Susan, Sherri Lauver, and Rebecca Maynard. A Review of 11 Reviews of After-School Programs: Documentation of the Literature and Search Process. Philadelphia, PA: University of Pennsylvania, April 16, 2003.
	Zief, Susan, Sherri Lauver, and Rebecca Maynard. Impacts of After-School Programs on Student Outcomes: Interim Report on Progress and Findings. Philadelphia, PA: University of Pennsylvania, April 28, 2003.
	Zief, Susan, and Sherri Lauver. Impacts of After-School Programs on Student Outcomes: Details of 6 Experimental Design Evaluations. Philadelphia, PA: University of Pennsylvania, May 2003.
	Zief, Susan, and Sherri Lauver. Impacts of After-School Programs on Student Outcomes: Results from 6 Experimental Design Evaluations. Philadelphia, PA: University of Pennsylvania, May 2003.
Interventions to improve marriage and relationships and to reduce domestic violence	
16	Stagner, Matthew, Jennifer Ehrle, Jane Reardon-Anderson, and Katherine Kortenkamp. Systematic Review of the Impact of Marriage and Relationship Programs. Washington, DC: Urban Institute, March 12, 2003. (Protocol)
17	Feder, Lynette, David Wilson, and Kimber Keplinger. A Systematic Review of Court-Mandated Interventions for Individuals Convicted of Domestic Violence. Portland, OR: Portland State University, February 19, 2003. (Protocol)
	Feder, Lynette, David Wilson, and Kimber Keplinger. A Systematic Review of Court-Mandated Interventions for Individuals Convicted of Domestic Violence. Portland, OR: Portland State University, February 19, 2003. (Working Draft, Version 8)

There remained many issues confronting reviewers, including the following:

- What is a reasonable question/topic for a 'Campbell Collaboration'-type review?

- What is the role of the protocol and the protocol review process? Are there conditions under which new information revealed during the review process might justify changing the review protocol midstream?
- What criteria should be embedded in the protocol for deciding what studies should be included and excluded? Are there any conditions under which evidence of questionable credibility should be included in a review?
- What is a reasonable and efficient study search strategy?
- What is the appropriate context for a review? Are there conditions under which a meaningful review could or should focus on a limited geographic area – for example, a particular nation? When is it important to include research from other nations?
- When is it appropriate to conduct a statistical meta-analysis of the findings extracted from studies meeting the inclusion criteria?
- How does one decide how to group studies for inclusion in a statistical meta-analysis?
- How should one deal with variability in the outcome measures used in the studies being reviewed?
- What is the most useful metric for reporting outcomes?
- How can one efficiently and effectively address differences in summary results between the 'Campbell Collaboration'-type review and other reviews?

In some cases, common decision-rules could be applied across all 14 test-bed projects. In a few cases reviewers believed that decision-rules should be specific to particular reviews. The common decision-rules that were successfully followed covered key principles. For example, all reviewers were able to narrow the focal questions for their review to ones that were directed at particular policy goals – typically particular interventions. Moreover, each review team proposed specific search strategies for identifying relevant studies.

All 14 test-bed reviews included systematic searches of electronic databases. The extent to which reviewers also hand-searched journals, searched conference proceedings and the Internet, and invested in personal networking to identify the more fugitive literature varied and depended on reviewers' knowledge of the state of the emerging research in their particular field. For example, in fields where there is a high volume of new studies, like teen pregnancy prevention, all of the above search strategies were outlined in the review protocol. In other cases, like the review of after-school programmes (Zief *et al*., 2003), the reviewers specified in their protocol that they planned to by-pass hand-searching journals. The reason for the decision related to the high cost and expected null yield. There is no relevant research on this topic prior to the period covered by the electronic databases, as this was not an issue of any public concern. Moreover, research on after-school programmes is being conducted by a relatively small number of individuals and organisations that are well known to the review team. Those studies that meet the review inclusion criteria will be known to the reviewers well before they appear in formal publications. The same arguments do not apply to other test-bed

review areas, such as that focused on teen pregnancy prevention programmes (Scher & Stagner, 2003).

Each test-bed review protocol included very specific study inclusion criteria that reflected the question being asked by the review and the goals of this test-bed project. All review teams proposed to retrieve all potentially relevant studies identified through their specified search strategies and extract from these studies that information relevant to the stated review inclusion criteria. For studies screened and ultimately excluded from the review, the researchers systematically documented the reasons for exclusion. All reviewers have (or plan to) carefully summarised the results of their search efforts in tabular form (see for example, Table 3). Finally, all review teams have set forth in their protocols the reasons why and conditions under which they will conduct meta-analysis of the set of study findings in order to offer more generalisable conclusions than offered through any single study.

As noted above, the test-bed reviewers were encouraged to limit their in-depth review efforts to RCTs. However, even within the often small subset of RCTs identified for each of the reviews, all reviewers confronted high variability in the quality of the evidence they assembled. Thus, one of the more challenging tasks for the review teams has been to sort through the RCTs to select those that are judged likely to yield unbiased estimates of programme impacts, even if the precision of the estimates is low.

One conclusion from the test-bed effort is that this type of systematic review of evidence can be enormously powerful for facilitating evidence-based policy making. However, there also is recognition that meaningful reviews in the social sciences require attention to the nuances of the particular policy area. One needs to be aware of issues that may affect the reasonableness of

Table 3 Summary of search results for after-school programme evaluations (Zief & April, 2003)

Search source	*Number of citations*	*Unduplicated studies retrieved and reviewed*	*Included studies*
Prior reviews	110	46	3
Electronic databases			
ERIC	184	15	0
Education Digest	21	0	0
PsychINFO	144	15	0
Dissertation Abstracts	125	9	0
Journal hand searches	NA	NA	NA
Internet searching			
google.com	0	0	0
Professional networks	2	2	2
Total	586	87	5

combining results across studies. For example, under what conditions would it be acceptable to combine results for studies of residential and nonresidential treatment for oppositional defiant disorder or for academically focused and recreationally focused after-school programmes? And, it also is important to be attentive to the generalisability of the study findings. For example, in so far as the review of intervention effects on marriage and relationships has identified only research focused on low-income populations, the results are not generalisable to the entire adult population.

Specific Recommendations for Future Work

A recent review of the experiences of the 14 test-bed review teams led us to some recommendations for future work in this area. Below, we focus on five specific issues where the experiences of this team may be especially beneficial to others: (1) framing review questions; (2) searching for studies; (3) sifting and sorting the evidence from studies; (4) coding and archiving data; and (5) analysing and reporting the findings.

Framing review questions

A key element of the *Campbell Collaboration* test-bed has been to frame questions suitable for addressing issues facing US policy makers. The researchers involved in the test-bed have struggled with the many forms in which policy questions are framed. The four most common forms of questions are the following:

(1) How big a problem is X? For example, what are the costs to society of teen pregnancy? Is the high school drop-out rate decreasing? Is student achievement improving?
(2) How is the development of problem X related to other factors? For example, are changes in teen pregnancy rates related to economic trends? Are changing levels of school drop-out rates associated with changes in family structure or background? Does the change in student achievement relate to changes in schools, communities or families?
(3) How might policies be designed to affect X? For example, what theories underlie why teens have sex and why they get pregnant? Do high school students drop out because of emotional issues, family problems or poor school management? What types of schools seem to influence student achievement?
(4) What is the impact of programme/policy Y on problem X? For example, what is the impact of teen pregnancy prevention programmes on the teen pregnancy rate? Do drop-out prevention programmes reduce adolescent school problems? What is the impact of a new reading curriculum on student achievement?

Each of these is a legitimate question for a systematic review. The 'Campbell Collaboration' test-bed project focused on the fourth question since this question most clearly reflects the goals of the Campbell Collaboration and the hopes that it will inform policy makers about 'what works'. However, many of

the conclusions from the test-bed initiative focused on evidence of programme or policy impacts can generalise to these other three areas of inquiry.

There are two areas in which the review process and conclusions from the test-bed will be quite specific to the fact that this is focused on issues related to programme and policy impacts. The first relates to decisions about what constitutes credible evidence. While the most reliable information for judging programme and policy impacts generally derives from RCTs, these are not the most credible evidence to address the other questions. For example, questions about the size of a problem would rely on demographic and epidemiological data; those focused on understanding relationships between particular social problems and other factors would rely on correlation analyses or systematic qualitative assessments of relationships; and evidence regarding programme and policy design would likely involve process and operational analysis, as well as correlation studies. The second major difference relates to the likelihood that the review will involve a meta-analysis of the individual study findings. When the goal is to assess programme impact estimates from studies, a meta-analysis can be used to generate mean impact estimates and to increase the statistical power of impact estimates.

Meta-analysis also can be used in such settings to *descriptively* explore issues of programme targeting, design and/or intensity. However, because the programmes under study were not randomly assigned to different targeting strategies, designs or intensities of intervention, these results do not have the credibility of the overall impact findings based on the experimental design evaluation, but rather are a form of quasi-experimental findings.

When one combines multiple studies or pools them in a meta-analysis, the purpose is often to examine relationships that may be important to understanding the scope of the problem and developing hypotheses regarding interventions that may mitigate it. If a meta-analysis is used for this purpose, it is important for the reviewer to draw a boundary between the impact estimate from the meta-analysis and the description of the pattern of findings of 'what worked, for whom, under what context'.

Framing the questions for reviews of evidence on intervention effectiveness is, on the surface, a straightforward process. However, considerable complexity is hidden within.

The definition of 'a programme' is seldom clear

Policy makers may have a broad definition of 'a programme' in mind, akin to a funding stream to address a class of social problems. The review of teen pregnancy prevention programmes illustrates this complexity in programme definition (Scher *et al*., 2003). Teen pregnancy prevention programmes may be defined by their shared goal (as policy makers often define them), or they may be defined by the programmatic inputs or elements (as programme operators are more apt to characterise them). For example, programmes that emphasise condom distribution and programmes that emphasise sexual abstinence share the goal of pregnancy prevention. But, these two groups of programmes aim to reach that goal of pregnancy prevention in different ways – the former through increasing condom use among sexually active youth and the latter through decreasing sexual activity among youth.

Policy makers may be forgiven for defining programmes by their goals. It is not the job of the policy maker to figure out *how* to solve a problem. Rather, it is the job of policy makers to prioritise problems and direct resources accordingly. In many cases, policy makers want to know whether *anything* works to mitigate the problem, or whether the programmes, on average, have an impact. Systematic reviews can run into trouble when they accept a broad definition of the treatment of interest and/or allow inclusion of studies that use correlational rather than random assignment to measuring impacts so as to have some purported evidence base to guide decision-making.

The range of interventions studied may not accurately reflect range of interventions across the USA or around the world

Certain types of interventions or settings may be more amenable to evaluation or may be the target of evaluation funders. Other interventions, populations or setting may be ignored. When 'summing up' the evidence on varied interventions, as noted earlier, the 'average' impact may not relate directly to the overall impact of the range of programmes in the field. Focusing reviews on interventions of a certain type of intervention, implemented within a certain context lessens the possible misrepresentation of average impacts. However, depending on the political context, it may be more or less relevant to report findings for more homogeneous clusters of interventions.

The definition and measurement of outcomes are often unclear

Some test-bed reviews focused on multiple outcomes. For example, one test-bed review focused on sexual debut, pregnancy risk and pregnancy (Scher *et al*., 2003), and another focused on observed crime and disorder, calls for service, and total crime incidents (Braga, 2003). It is common for programmes to have significant estimated impacts on some measured outcomes and not others. Thus, it is important to clearly define in the review protocol the particular outcomes that a review will examine across studies. The test-bed review experience suggests that it is best to analyse distinct outcomes rather than to create composite outcomes measures. If composite measures are to be used in the synthesis (for example, combining into a single outcome measure impacts on reading and math test scores or indicator of crime reduction), it is important to specify and justify this in the protocol.

The test-bed reviewers also were challenged by the fact that results for their focal outcomes were reported in different units across studies – generally not 'natural' units, such as percentage point increase or decrease in an 'event'. Often times, there was sufficient information to convert results to common units. However, in some cases, it was necessary to make some inferences in order to do so. For example, in cases where outcomes are reported in odds ratios, without information on the base level, it might be necessary to use descriptive information on the sample to infer what the 'base' rates were in order to convert the Odds Ratios to the percent of the sample experiencing the outcome. Then, often statistics on null findings for outcomes in question often are not reported in published reports. The report may simply report that the programme did not affect outcome X; neither the point estimate of the impact nor its confidence interval is reported. Omitting this information biases the findings, while including it requires making some untestable assumptions.

Searching for studies

The 'Campbell Collaboration', like the 'Cochrane Collaboration', is committed to a comprehensive, systematic search for evidence. This implies systematically searching in all international contexts and languages, as well as searching beyond the standard electronic databases for studies that may have never been published or may have been published prior to digital storage. Such a broad, comprehensive search strategy is based on legitimate concerns, particularly the nefarious effects of 'publication bias'. However, the test-bed effort led to the conclusion that reviewers should seek *clarity of search boundaries* and implement a *thorough, systematic searching within (and only within) the stated boundaries*. These boundaries can be defined by a variety of parameters, including calendar years, national boundaries and search domains.

There are both practical and methodological reasons for this recommendation. The practical reason relates to enabling reviews to progress in stages that are consistent with available time and resources. Searching in multiple languages is time consuming and expensive, if one is committed to translating all possible studies for inclusion. Similarly, valuable review efforts could be forestalled by an expectation that the reviewer must search the entire published literature (including hand-searching of journals), all possible electronic sources and professional networks – tasks that may well exceed available time and resources. Still, it would be valuable to encourage the completion of at least one well defined module of the ideal comprehensive search domain and review task in a manner that would allow subsequent completion of other search modules within the complete domain.

The methodological reason for emphasising thorough, systematic searching within well defined boundaries relates to the fact that the contexts in which policies operate differ in ways that are important to their effectiveness. This is even truer in the social science arenas that are the focus of the Campbell Collaboration reviews than in the medical arena, which is the focus of most Cochrane reviews. Where the unit of analysis is a person's health, or even an organ or a cell, the national context may not matter as much as when the unit is a population under the influence of particular policy regimes within national boundaries. An example of the critical importance of geographic context for the test-bed review project is found in the review of 'mandatory work policies,' where the differences between the US public welfare system and those in other industrialised nations could be expected to strongly affect the impacts of particular intervention strategies (Stagner *et al*., 2003b). When reviews do include the international literature, it is critical that the analysis pays careful attention to possible impacts of national context. In many cases, this expanded focus will necessitate partitioning the analysis along national or other contextual boundaries.

Treatment of prior reviews

Included within any systematic search of the literature is the identification and retrieval of prior reviews that addressed similar questions. Many test-bed reviewers found that prior reviews drew different boundaries around the criteria for study inclusion. Prior reviews may have included both RCTs and nonexperimental designs, may have more broadly defined the intervention or

may have included a range of intervention contexts. Test-bed reviewers have treated these prior reviews similarly. They have retrieved all of the studies included in the prior reviews and then applied their own inclusion criteria to these studies. In some cases, many studies from prior reviews were excluded in the set of studies for the 'Campbell Collaboration' test-bed review because they did not fit inclusion criteria or they were of poor quality.

Sifting and sorting the evidence from studies

The decision to focus the test-bed effort on evidence only from high-quality RCTs was based on two considerations. First, a systematic review of the evidence of tests of 16 RCTs versus non-RCTs conducted under the test-bed (Glazerman *et al*., 2002) showed that attempts to simulate RCTs using other study designs and analytic methods generally fail to replicate the results of well designed and implemented RCTs. Glazerman at first found that 16 studies attempting to determine programme impacts using non-RCT methods produced no clear knowledge on how and under what circumstances a non-RCT method may be 'just as good as' an RCT. An update of this review which identified 22 methodological studies examining the comparability of conclusions from experimental and quasi-experimental studies upheld these central conclusions (Glazerman *et al*., 2003).

Second, significant resources are required to examine non-RCT methods to determine, *post hoc*, whether they may constitute credible evidence of programmatic impacts. Myers (2003) has suggested that evidence of programme impacts might be considered in three tiers: (1) High-quality RCTs as the top tier in terms of credibility of evidence; (2) lesser-quality RCTs and high quality non-RCTs as a second level that is, at best, able to highlight interventions that may warrant further rigorous testing; and (3) lower quality non-RCTs as the weakest evidence that should mainly be used to identify relationships that can stimulate new theories or identify hypotheses that warrant empirical investigation.[9]

In a world of unlimited resources, it might be desirable to examine studies falling into the second tier proposed by Myers to determine which, if any, appear to provide credible, unbiased evidence. However, this not only would consume significant time and resources, it also opens the door to significant debate about whether the problems with the RCTs were severe enough to eliminate them from tier one, as well as whether the models and methods for non-RCTs were 'almost as good as' RCTs. The decision to focus the test-bed on only the first tier of studies has afforded reviewers the opportunity to focus their effort on judging what we know with confidence about a programme's impact, rather than diverting scarce research resources to examining, assessing and interpreting evidence from studies that quite likely provide biased estimates of programme impacts.

In some cases, test-bed reviewers have plans to expand their reviews to incorporate non-RCT evidence. However, in such cases, they will partition the evidence following the recommendations of Myers (2003).

Coding and archiving data

There appears to be general consensus that three levels of information should be retained from the search process. At the most general level, it is important to maintain a matrix detailing the key words searched within each information source and the results of that source. In addition, for all studies that are reviewed for possible inclusion in the review, it is important to maintain the full citation, the source through which the study was identified, and the study design and implementation information necessary to determine whether the study is likely to yield credible impact estimates of the relevant outcomes.

Minimum data for all studies identified

In addition to a full citation, a typical 'minimum data set' would include the following 14 elements:[10]

(1) type of research design;
(2) unit for assignment to programme and control group;
(3) unit for comparison of programme and control groups;
(4) method for constructing the control group;
(5) use of statistical control variables in the impact analysis;
(6) longitudinal tracking of sample members or repeated cross sections used in the analysis;
(7) percent of sample lost to follow-up and appropriate treatment of high attrition rates;
(8) prospective or retrospective identification of programme and control groups;
(9) focal population;
(10) geographic area where intervention occurred;
(11) outcomes measured;
(12) duration of sample follow-up;
(13) adequate information to calculate standardised effect sizes;
(14) sufficient detail on the intervention.

This type of information is important for documenting decisions regarding whether or not a particular study is appropriate for inclusion in the review synthesis – whether it provides credible evidence relevant to the focal question. There was general agreement among the test-bed reviewers that studies that nominally address the study question, but that do not meet standards for generating credible evidence on the outcomes of interest, should be inventoried in an appendix to the review with notations regarding the primary reason for exclusion.[11]

Optimal data set for included studies

For all studies that contain credible evidence on the focal question, a more detailed set of information needs to be coded. This set includes the following four categories:

(1) the intervention and the counterfactual conditions against which it is being compared;
(2) sample members;

(3) evaluation components and specific research methods employed;
(4) outcomes, including definitions of outcomes, mean values of outcome measures for programme and control group members, and standard deviations of the outcome measures.

Many of the test-bed reviewers maintained a database capturing these study details in *ACCESS*. However, some relied on *RevMan*, which is the database software used by the Cochrane Collaboration. Either works well. However, at this point, *ACCESS*, and support for it, is more widely available and, thus, somewhat more advantageous.

Analysing and reporting the findings

As noted above, the test-bed project has limited its focus to systematic reviews of evidence of programme and policy impacts. It is not addressing the equally valuable arena of reviews of research for purposes of theory development, hypothesis generation or problem definition. Within this focus, the test-bed effort highlighted five recommendations regarding methods for reporting and synthesising estimates of programme impacts.

Stay with randomised control trials (RCTs) as the key inclusion criteria

Policy experts asked to comment on the test-bed interim products welcomed the clarity and consistency in review findings that this review policy produced. Nearly all of the test-bed reviewers uncovered multiple RCTs relevant to their focal question and had something to report from these studies. Furthermore, while RCTs standardise key aspects of the testing of the effects of policies and programmes, they also offer variability of implementation, context and power that challenge the reviewer and build a stronger base for knowledge.

Report impact estimates in 'natural' units

We were strongly advised by policy makers who have collaborated with the test-bed project to report findings in natural units, not standardised effect sizes. Policy makers involved with social and educational interventions tend to think in terms of actual measures of outcomes of interest – for example, percent of the population that is married; percent of teens who initiate sex; or the number of crimes per 100,000 population. Each of these measures has intrinsic meaning, as will an estimate of the change in the outcome due to a programme or policy intervention. In contrast, measures such as standardised effect sizes, odds ratios and percent changes are relative measures that depend on the 'standardising' assumptions. Figure 1 illustrates that the general pattern of findings is generally similar, regardless of whether they are reported in difference in mean values or the percent change from the control group mean. However, the percent change measures are more variable, as they are sensitive to the base level of the outcome measure. (Figure 1 would look similar if the results based on odds ratios were compared with the differences in means, as odds ratios also are sensitive to the base level for the control group.)

Although standardised effect sizes have some appealing qualities – most notably, facilitating the use of outcomes measured in different metrics – they also are not transparent in terms of their interpretation. Moreover, researchers

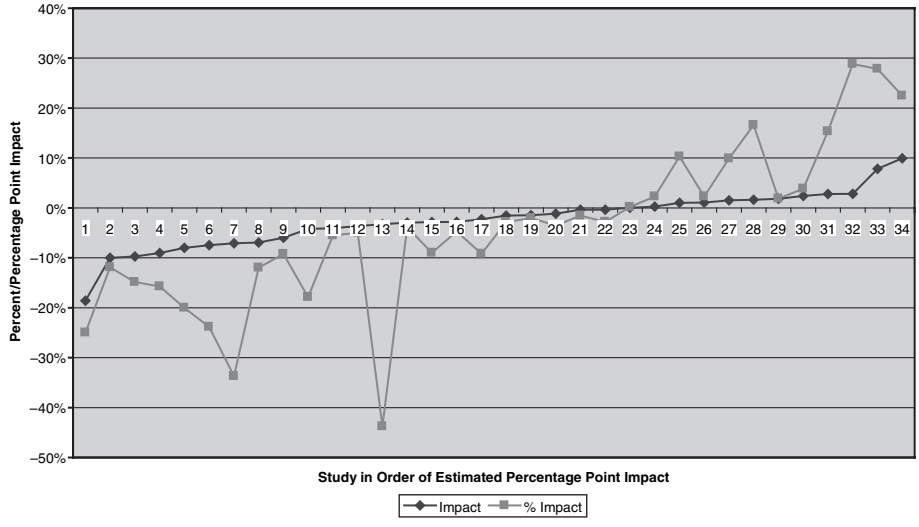

Figure 1 Estimated change in the percent who are sexually experienced and estimated percent change in rate due to the programme
Source: Data from Scher *et al*., 2003

infrequently have the population standard deviation that should be used for standardising outcomes, and instead tend to create effect sizes using the reference-estimated standard deviation for the study sample from which the impact estimate was generated. Thus, a portion of the difference in standardised impact estimates may be attributed to differences in the variance in the outcome measures across the various study samples.

Reviews should include only studies for which intention-to-treat (ITT) estimates are reported or can be reliably generated

The reason is that estimates based on 'treated' subgroups of the programme group or groups selected on the basis of an outcome that could have changed as a result of the intervention (for example, contraceptive use among sexually active youth) are subject to selection, even in cases where the original study design involved a RCT. There are instances when policy makers will be interested in estimates of the effect of treatment on the treated. However, we urge that these estimates be presented only as a supplement to the ITT estimates and that they be accompanied by information on the 'treatment rates' assumed in the conversion.

Reviews should report minimum detectable effects of the included studies and, to the extent possible, study findings that relate to the costs and benefits of the intervention. This recommendation is aimed at helping inform consumers of reviews about what impacts the intervention needed to produce in order for the study to have had a reasonable chance of observing a statistically significant difference in outcomes between the programme and control groups. Often studies have very low statistical power to detect a programme impact of a size that is reasonable given the nature of the intervention. In the same vein, when studies report intervention costs and estimate the value of impacts, this information should be noted, even if not extracted and analysed in the review.

Meta-analysis should be conducted only on homogeneous clusters of intervention studies

Studies included in a meta-analysis should be reasonably homogeneous along a number of dimensions, including the following:

- nature of the intervention;
- nature of the target population;
- setting/context for the intervention;
- outcome measures.

If there is an adequate number of studies within any homogeneous cluster and if the effectiveness of the intervention strategy is not evident from inspection of the study findings, it may be desirable to conduct a meta-analysis. Under such conditions, this will improve the power of the analysis and provide a means of generating a weighted average impact estimate. Even when the overall pattern of results from the various studies provides a clear conclusion regarding programme effectiveness, it may be advantageous to conduct meta-analysis to provide a more parsimonious and statistically reliable measure of programme impacts.

Any meta-analysis subgroup findings for clusters of homogeneous studies or findings based on heterogeneous clusters of studies should be viewed as exploratory or descriptive analysis. This is particularly true of any analysis that compares impact estimates for subgroups of studies defined by characteristics such as the type of intervention, intensity of the treatment or the setting.

Forest plots are valuable, especially when the number of studies is small

Except in cases where there is a large number of studies, Forest plots provide a highly intuitive description of the study findings. An example of the potential power of the Forest plot is illustrated in Figure 2. In this case, it would have been most surprising if the meta-analysis had revealed any finding other than a null-finding, given the pattern of mean impact, which are scattered on either side of zero and which tend to have relatively large standard errors.

Reconcile any different conclusions between the Campbell Collaboration-style review and other recent reviews

Many of the test-bed reviewers found that prior reviews of the literature revealed considerable inconsistency in the conclusions of reviews on the same topic or question. Conclusions seemed especially sensitive to the rules on inclusion and exclusion of studies and on the methods for summarising reputed effects. Most reviewers have begun to deal with these inconsistencies by tabulating *how* a prior review differed from the 'Campbell Collaboration'-style review and *what impact* those differences likely had on the differing conclusions. For example, the test-bed review of teen pregnancy prevention programmes uncovered 13 prior reviews conducted during the past 10 years (Table 4). These reviews included as few as five studies and, in one case, more than 150 studies. Two focused only on RCTs, and five did not distinguish

Citation	EffectName	Year	Treated	Control	Effect	Lower	Upper	NTotal	PValue
Handler (1987)	1		8 / 25	10 / 25	.706	.221	2.252	50	.556
O'Donnell (2002)	1		31 / 42	31 / 37	.545	.179	1.659	79	.282
Jorgensen (1993)	1		29 / 52	29 / 39	.435	.176	1.073	91	.068
O'Donnell (2002)	1		30 / 62	31 / 54	.696	.334	1.449	116	.332
McBride (2000)	1		39 / 69	35 / 59	.891	.441	1.803	128	.749
McBride (2000)	1		65 / 92	44 / 61	.930	.454	1.906	153	.843
Kirby (1995)	1		10 / 87	7 / 71	1.187	.428	3.297	158	.741
Kirby (1995)	1		4 / 96	8 / 108	.543	.158	1.865	204	.326
McBride (2000)	1		112 / 114	108 / 112	2.074	.372	11.558	226	.396
St. Lawrence (1995)	1		63 / 113	74 / 113	.664	.388	1.136	226	.134
Jemmott (1998)	1		50 / 175	54 / 172	.874	.552	1.384	347	.566
Jemmott (1998)	1		44 / 184	54 / 172	.687	.430	1.096	356	.114
Eisen (1990)	1		148 / 212	144 / 196	.835	.542	1.286	408	.413
Philliber (2001)	1		150 / 217	150 / 205	.821	.538	1.252	422	.359
Eisen (1990)	1		163 / 249	145 / 230	1.111	.764	1.615	479	.581
Philliber (2001)	1		158 / 268	163 / 251	.775	.543	1.107	519	.161
Kirby (1997)	1		37 / 340	30 / 324	1.197	.720	1.988	664	.488
Kirby (1995)	1		23 / 329	33 / 384	.799	.459	1.391	713	.428
Kirby (1997)	1		26 / 372	22 / 374	1.202	.669	2.162	746	.538
Blake (2000)	1		199 / 390	313 / 540	.756	.581	.982	930	.036
Kirby (1995)	1		20 / 427	15 / 537	1.710	.865	3.382	964	.119
Kirby (1995)	1		40 / 469	43 / 528	1.052	.671	1.649	997	.826
Grossman (1992)	1		412 / 577	407 / 576	1.037	.804	1.337	1153	.781
Thomas (1992)	1		412 / 762	200 / 453	1.489	1.179	1.881	1215	.001
Coyle (2000)	1		65 / 654	86 / 615	.679	.482	.956	1269	.026
Coyle (2000)	1		41 / 686	35 / 591	1.010	.634	1.608	1277	.967
Moberg (1998)	1		130 / 617	115 / 676	1.302	.985	1.721	1293	.063
Thomas (1992)	1		399 / 831	246 / 524	1.044	.838	1.299	1355	.701
Kirby (1995)	1		138 / 660	128 / 706	1.194	.913	1.561	1366	.195
Kirby (1995)	1		25 / 694	25 / 743	1.073	.610	1.887	1437	.806
Levy (1995)	1		530 / 1001	364 / 668	.940	.772	1.144	1669	.535
Coyle (2001)	1		140 / 1078	144 / 951	.836	.651	1.075	2029	.163
Kirby (1995)	1		175 / 1037	160 / 1042	1.119	.886	1.414	2079	.346
Wight (2002)	1		263 / 1117	298 / 1246	.980	.810	1.185	2363	.832
Wight (2002)	1		423 / 1330	445 / 1350	.948	.807	1.115	2680	.522
Fixed Combined (35)			**4602 / 15428**	**4196 / 14733**	**.982**	**.927**	**1.040**	**30161**	**.537**

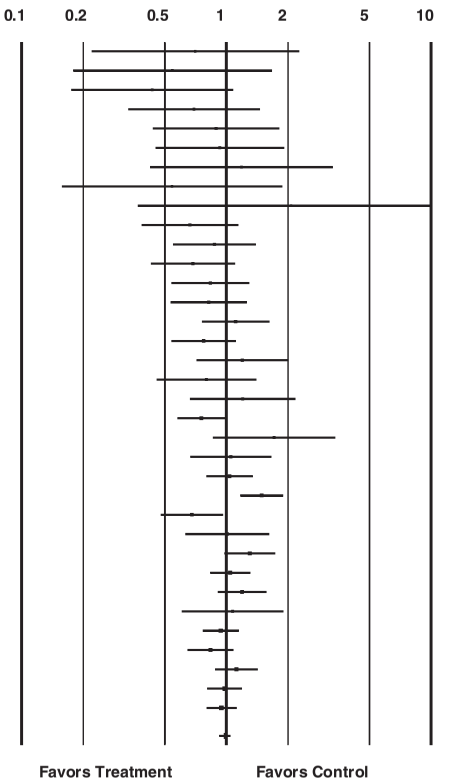

Figure 2 Forest plot of the impact of teen pregnancy prevention programmes on sexual initiation
Source: Scher *et al*., 2003

Table 4 Comparison of methods and conclusions from prior teen pregnancy reviews (Scher, 2003)

Author(s)	*Publication year*	*Systematic Review?*	*Coverage restrictions*	*Methods included*			*Number of studies*	*Summary of conclusions*
				Experiments only	*Experiments and quasi-experiments*	*All methods*		
Dicenso *et al*.	2002	Yes	N. America, New Zealand, Australia, W. Europe	✓			26	No impacts on delay of sex, birth control use or pregnancies.
Franklin *et al*.	1997	No	?			✓	32	Small significant effect sizes for contraceptive use and pregnancy. No significant effect for sexual activity.
Frost and Forrest	1995	No	USA		✓		5	Delayed sexual debut, increased contraceptive use, no change in pregnancy rates.
Grunseit	1997	No	None: international search			✓	47; 11 Exp.	17 studies reported delays in the onset of sexual activity; 3 studies found increases in sexual behaviour.
Kim *et al*.	1997	Yes	USA		✓		40	One-third to half of the studies reported improvements in condom use and higher rates of abstinence. More favourable results for nonexperimental design studies.

Table 4 (*Continued*)

Author(s)	*Publication year*	*Systematic Review?*	*Coverage restrictions*	*Methods included*			*Number of studies*	*Summary of conclusions*
				Experiments only	*Experiments and quasi-experiments*	*All methods*		
Kirbya	2001	Semi-systematic	USA and Canada		✓		74	No impact on sexual debut; some increase in contraceptive use.
Kirbyb	1997	Semi-systematic	USA and Canada		✓		50+	No impact on sexual debut; some increase in knowledge; some affect other behaviours.
Manlove *et al*.	2002	No	USA and Canada			✓	150+	Programmes combining sex ed. with youth dev. will delay sexual debut and pregnancy.
NHS Center for Reviews and Dissemination	1997		English language	✓			42	If linked with contraceptive services, reduced teen pregnancy.
Oakley *et al*.	1995	Yes	English language		✓		12	Mixed results. Three programmes reduced sexual activity; one increased it; others had no effect or ambiguous effects.

Table 4 (*Continued*)

Author(s)	*Publication year*	*Systematic Review?*	*Coverage restrictions*	*Methods included*			*Number of studies*	*Summary of conclusions*
				Experiments only	*Experiments and quasi-experiments*	*All methods*		
Silva	2002	Yes	USA: school based		✓		12	Evidence of very small effects on abstinent behaviour.
Thomas	2000	No	USA: selective			✓	9	Some evidence that well designed programmes will delay sexual debut. Impacts often not observed until 18 months or more after the intervention.
Visser	1994	No	Dutch and English language			✓	21	No impact on sexual activity; some studies showed increases in contraceptive use.

among various methods for estimating programme impacts. The conclusions from the reviews range from fairly encouraging conclusions that there are effective programmes for delaying sexual debut and teen pregnancy (Manlove *et al*., 2002) to strong conclusions that there is no evidence that pregnancy prevention programmes delay sex, increase use of birth control or reduce pregnancies (DiCenso *et al*., 2002). The majority of the reviews offer some positive, some negative and mostly null conclusions (for example, Kirby, 2001).

Concluding Comments

The Campbell Collaboration model for systematic reviews of evidence as adapted for the test-bed project provides a strong, defensible and transparent strategy for assessing: (1) What interventions/policies do we know with reasonable confidence work? (2) What interventions/policies have not been adequately assessed to draw conclusions regarding effectiveness? And, (3) what programmes/policies have been rigorously evaluated, but do not seem to work?[12] It is possible to modularise the approach to a Campbell-type review in a way that permits the efficient, gradual accumulation of evidence. Exercising this option could be important to jump-starting the process of developing review protocols and initiating reviews of the salient bodies of evidence. The procedures for collecting evidence and the standards for 'counting' evidence are transparent under the Campbell Collaboration review model. This feature seems to be highly valued by policy makers. Moreover, this quality embeds within the review explanations for differences in the conclusions based on the review methods and/or search boundaries.

The recommended search and analysis process results in a 'respectful' sorting of evidence on causal relationships and evidence on correlations worthy of further exploration and testing. It recognises the value of multiple research paradigms and study designs, while restricting the evidence for substantiating causal inferences about programme impacts to evidence generated from well designed and implemented RCTs.[13]

The next steps in the test-bed initiative include bringing the 14 reviews to conclusion, and drawing on the infrastructure from this effort to create a practical guide or 'tool-kit' to help others benefit from these early experiences when undertaking similar types of reviews.

Acknowledgements

A number of people were very instrumental in the work underlying this paper. The members of the 14 review teams are listed in Table 1. In addition, Marshall Smith from the Hewlett Foundation; Howard Rolston and Meredith Kelsey from the US Department of Health and Human Services; Grover (Russ) Whitehurst, Alan Ginsberg and David Goodwin from the US Department of Education; Ron Haskins from The Brookings Institution; Rob Hollister from Swarthmore College; Larry Orr from Abt Associates; Steve Bell from Urban Institute; Robert Boruch from University of Pennsylvania; Harris Cooper and Jeff Valentine from University of Missouri; Larry Hedges from University of Chicago; Peter Reuter from University of Maryland; Mark Lipsey from Vanderbilt University; and Dennis Gorman from Texas A & M University.

Correspondence

Any correspondence should be directed to Phoebe Cottingham, Institute of Education Sciences, Commissioner of Education Evaluation, US Department of Education, 555 New Jersey Avenue NW, Fifth Floor, Washington, DC 20037, USA (phoebe.cottingham@ed.gov).

Notes

1. Another versions of this paper was presented at the Conference on Systematic Reviews of Qualitative Evidence in Windermere, UK, January 18, 2003.
2. More information on the Cochrane Collaboration can be found at http://www.cochrane.org.
3. Examples of these groups include: the *What Works Clearinghouse* (US Department of Education, http://www.w-w-c.org); the *Evidence for Policy and Practice Information and Co-ordinating Centre* or EPPI-Centre (Institute of Education, University of London, http://eppi.ioe.ac.uk/EPPIWeb/home.aspx), the *Quan-Qual Evidence Synthesis Group* (University of Leicester, Department of Epidemiology and Public Health, http://www.prw.le.ac.uk/research/qualquan/esrcsummary.htm); *The Future of Children* (David and Lucile Packard Foundation, http://www.futureofchildren.org); the *Alberta Center for Child Health Evidence* (University of Alberta, Edmonton, CA, http://www.ualberta.ca/ARCHE/sysreviews.html); and the *Centers for Disease Control Research Synthesis Project* (Centers for Disease Control, US Department of Health and Human Services, http://www.cdc.gov/hiv/pubs/hivcompendium/hivcompendium.htm).
4. These principles are paraphrased from information on the Campbell Collaboration website: http://www.campbellcollaboration.org/FraAbout.html. See also Petrosino *et al*. (1997), available at http://www.ucl.ac.uk/spp/download/publications/Annexe5.pdf.
5. Other differences include the fact that clinical trials in medicine tend to be clearly directed at a common outcome – for example, lowering blood pressure – whereas those in social sciences often have multiple goals or more broadly defined goals, such as improving economic and social well-being or reducing involvement in crime and increasing legal employment.
6. A systematic review protocol specifies not only the review topic and why a review is warranted. The protocol must also specify how studies will be found (search procedures) and how studies will be screened for inclusion or exclusion from the review. The analytical tasks that the reviewer will undertake must also be prespecified. The goal is to avoid 'fishing' for particular answers or conclusions, and assure review users that scientific procedures were followed in the review process.
7. The review teams recognised that there are similar issues for findings from 'flawed' experiments.
8. This decision reflected comments by external reviewers of the early findings from the first year of Test-bed work as well as a systematic review of research looking at the comparability of programme impact estimates generated from randomised field trials and from various quasi-experimental methods. See the discussion later in this paper of the work of Glazerman *et al*. (2003), which has been expanded as part of this Test-bed effort (see Table 2).
9. If the papers or articles describing RCTs do not provide clear evidence of high quality (for example, if there are low or differential response rates in follow-up data collection), the reviewer must assume such studies cannot be placed in the first tier. Only those that clearly document methods and procedures clearly should be included.
10. The following list was created based on that reported in Scher (April 2003). However, it is very similar to those used by many of the other test-bed reviewers.

11. This recommendation is contrary to that offered by some of the *Campbell Collaboration* protocol reviewers, who preferred that the results for these studies be incorporated into the review and compared with those meeting higher standards of evidence.
12. This conclusion is the most challenging to support empirically, due to the power requirements.
13. There are many good references that detail the characteristics of a well designed and implemented RCT, including Rossi *et al*. (1999); Boruch (1997); and Orr (1999).

References

Bradley, M.C. (2003) *Systematic Review of Interventions for Disruptive Behavior Disorders and the School.* Philadelphia, PA: University of Pennsylvania (Protocol, Version 1).

Braga, A. (2003a) *Systematic Review of the Effects of Hot Spots Policing on Crime*. Cambridge, MA: Kennedy School of Government, Harvard University (Protocol).

Braga, A. (2003b) *Hot Spots Policing and Crime Prevention: Evidence from Five Randomized Controlled Trials.* Cambridge, MA: Kennedy School of Government, Harvard University (Working Draft).

Boruch, R. (1997) *Randomized Experiments for Planning And Evaluation: A Practical Guide*. Thousand Oaks, CA: Sage Publications.

DiCenso, A., Guyatt, G., Willan, A. and Griffith, L. (2002) Interventions to reduce unintended pregnancies among adolescents: Systematic review of randomised controlled trials. *British Medical Journal* 324 (7351), 1426–1430.

Feder, L., Wilson, D. and Keplinger, K. (2003) *A Systematic Review of Court-mandated Interventions for Individuals Convicted of Domestic Violence*. Portland, OR: Portland State University (Protocol and Working Draft).

Franklin, C., Grant, D., Corcoran, J., O'Dell Miller, P. and Bultman, L. (1997) Effectiveness of prevention programs for adolescent pregnancy: A meta-analysis. *Journal of Marriage and the Family* 59, 551–567.

Frost, J.J. and Forrest, J.D. (1995) Understanding the impact of effective teenage pregnancy prevention programs. *Family Planning Perspectives* 27 (5), 188–195.

Glazerman, S., Levy, D. and Myers, D. (2003) *Nonexperimental Versus Experimental Estimates of Earnings Impacts*. Princeton, NJ: Mathematica Policy Research, Inc.

Grunseit, A., Kippax, S., Aggleton, P., Baldo, M. and Slutkin, G. (1997) Sexuality education and young people's behavior: A review of studies. *Journal of Adolescent Research* 12 (4), 421–453.

Guyatt, G.H., DiCenso, A., Farewell, V., Willan, A. and Griffith, L. (2000) Randomized trials versus observational studies in adolescent pregnancy prevention. *Journal of Clinical Epidemiology* 53, 167–174.

Kim, N., Stanton, B., Li, X., Dickersin, K. and Galbraith, J. (1997) Effectiveness of the 40 adolescent AIDS-risk reduction interventions: A quantitative review. *Journal of Adolescent Health* 20 (3), 204–214.

Kirby, D. (1997) *No Easy Answers: Research Findings on Programs to Reduce Teen Pregnancy*. Washington, DC: National Campaign to Prevent Teen Pregnancy.

Kirby, D. (2001) *Emerging Answers: Research Findings on Programs to Reduce Teen Pregnancy*. Washington, DC: National Campaign to Prevent Teen Pregnancy.

Kirby, D., Korpi, M., Adivi, C. and Weissman, J. (1997) An impact evaluation of Project SNAPP: AIDS and pregnancy prevention middle school program. *AIDS Education and Prevention* 9 (A), 44–61.

Lipsey, M. and Landenberger, N. (2003) *Cognitive-behavioral Programs for Juvenile and Adult Offenders: A Meta-analysis of Randomized Controlled Intervention Studies*. Nashville, TN: Vanderbilt University (Protocol).

Littell, J., Forsythe, B. and Popa, M. (2003) *Impacts of Multisystemic Treatment on Youth Outcomes*. Bryn Mawr, PA: Bryn Mawr College (Protocol, Version 2).

Losel, F. and Beelmann, A. (2003a) Effects of child skills training in preventing anti-social behavior: A systematic review of randomized evaluations. *Annals (AAPSS)* 587, 84–109.

Losel, F. and Beelmann, A. (2003b) *Efficacy of Child Skills Training in Preventing Antisocial Behavior and Crime.* Erlangen, Germany: University of Erlangen-Nuremberg (Protocol).

Manlove, J., Terry-Humen, E., Romano Papillo, A., Franzetta, K., Williams, S. and Ryan, S. (2002) Preventing teenage pregnancy, childbearing, and sexually transmitted diseases: What the research shows. *Child Trends Research Brief*. Washington, DC: Child Trends.

Myers, D. (2003) Systematic reviews and the use of random assignment and quasi-experimental designs. Memorandum to P. Cottingham, Smith Richardson Foundation.

NHS Centre for Reviews and Dissemination (1997) Preventing and reducing the adverse effects of unintended teenage pregnancies. *Effective Health Care* 3 (1), 1–12.

Oakley, A., Fullerton, D., Holland, J., Arnold, S., France-Dawson, M., Kelley, P. and McGrellis, S. (1995) Sexual health education interventions for young people: A methodological review. *British Medical Journal* 310 (6973), 58–162.

Orr, L.L. (1999) *Social Experiments: Evaluating Public Programs with Experimental Methods*. Thousand Oaks, CA: Sage Publications.

Petrosino, A., Boruch, R., Rounding, C., McDonald S. and Chalmers, I. (1997) *A Social, Psychological, Educational and Criminological Trials Register (SPECTR) to Facilitate the Preparation and Maintenance of Systematic Reviews of Social and Educational Interventions*. London, England: School of Public Policy, University College London.

Ritter, G. and Maynard, R. (2002) *Evidence on the Effectiveness of Volunteer Tutoring*. Fayetteville, AR: University of Arkansas (Protocol).

Ritter, G. and Maynard, R. (2003) *Evidence on the Effectiveness of Volunteer Tutoring*. Fayetteville, AR: University of Arkansas (Working Draft).

Rossi, P., Freeman, H. and Lipsey, M. (1999) *Evaluation: A Systematic Approach* (6th edn). Newbury Park: Sage Publications International.

Scher, L. with Stagner, M. (2003) *A Systematic Review of Teen Pregnancy Prevention Interventions*. Philadelphia, PA: University of Pennsylvania (Protocol Version 3 and Working Draft).

Silva, M. (2002) The effectiveness of school-based sex education programs in the promotion of abstinent behavior: A meta-analysis. *Health Education Research* 17 (4), 471–481.

Stagner, M., Ehrle, J. and Reardon-Anderson, J. (2003a) *Systematic Review of the Impact of Mandatory Work Policies on Family Structure.* Washington, DC: Urban Institute (Protocol and Working Draft).

Stagner, M., Ehrle, J., Reardon-Anderson, J. and Kortenkamp, K. (2003b) *Systematic Review of the Impact of Marriage and Relationship Programs*. Washington, DC: Urban Institute (Protocol).

Strang, H. and Sherman, L. (2003) *Effects of Face-to-face Restorative Justice on Repeat Offending and Victim Satisfaction*. Canberra, Australia: Australian National University (Protocol, Version 2 and Working Draft).

Thomas, M.H. (2000) Abstinence-based programs for prevention of adolescent pregnancies. *Journal of Adolescent Health* 26 (1), 5–17.

Visher, C.A. and Winterfield, L. (2003) *A Systematic Review of the Effects of Non-custodial Employment Programs on the Recidivism Rates of Ex-offenders*. Washington, DC: The Urban Institute (Protocol).

Visher, C.A., Winterfield, L., Coggeshall, M.B. and Turner, W. (2003) *A Systematic Review of the Effects of Non-custodial Employment Programs on the Recidivism Rates of Ex-offenders*. Washington, DC: The Urban Institute (Working Draft).

Visser, A. and Van Bilsen, P. (1994) Effectiveness of sex education provided to adolescents. *Patient Education and Counseling* 23, 147–160.

Zief, S., Lauver, S. and Maynard, R. (2002) *Impacts of After-school Programs on Student Outcomes*. Philadelphia, PA: University of Pennsylvania (Protocol).

Zief, S., Lauver, S. and Maynard, R. (2003) *Impacts of After-school Programs on Student Outcomes: Interim Report on Progress and Outcomes*. Philadelphia, PA: University of Pennsylvania.