

International Initiative for Impact Evaluation



WORKING PAPER 2

Better evidence for a better world

Edited by Mark W. Lipsey and Eamonn Noonan

April 2009



THE CAMPBELL COLLABORATION

About 3ie

The International Initiative for Impact Evaluation (3ie) works to improve the lives of people in the developing world by supporting the production and use of evidence on what works, when, why and for how much. 3ie is a new initiative that responds to demands for better evidence, and will enhance development effectiveness by promoting better informed policies. 3ie finances high-quality impact evaluations and campaigns to inform better program and policy design in developing countries.

3ie Working Paper series covers both conceptual issues related to impact evaluation and findings from specific studies or synthetic reviews.

This Working Paper was edited by Mark W. Lipsey (Vanderbilt University, USA) and Eamonn Noonan (CEO, The Campbell Collaboration).

© 3ie, 2009

Contacts

International Initiative for Impact Evaluation
c/o Global Development Network
Post Box No. 7510
Vasant Kunj P.O.
New Delhi – 110070, India
Tel: +91-11-2613-9494/6885
www.3ieimpact.org

Table of Contents

The difficult path toward better evidence and better decisions

Eamonn Noonan, CEO, The Campbell Collaboration

*Elizabeth Kristjansson, Associate Professor, School of Psychology and
Institute of Population Health, University of Ottawa*

Toward evidence-informed policy and practice in child welfare

Julia H. Littell, Bryn Mawr College

Aron Shlonsky, University of Toronto

Mobilizing knowledge to support better learning

Ben Levin, Ontario Institute for Studies in Education, University of Toronto

We all agree we need better evidence.

But what is it and will it be used?

*Howard White, Executive Director, International Initiative for Impact
Evaluation (3ie)*

Better Evidence for a Better World

Mark W. Lipsey, Vanderbilt University, USA

The difficult path toward better evidence and better decisions

Eamonn Noonan, CEO, The Campbell Collaboration

Elizabeth Kristjansson, Associate Professor, School of Psychology and Institute of Population Health, University of Ottawa

This year's impact evaluation conference in Cairo and the forthcoming Annual Colloquium of the Campbell Collaboration serve as an opportunity to highlight the importance of better evidence in the effort to bring genuine improvements in people's living conditions. The rationale for an evidence-based approach remains simple, solid, and relevant. But what is better evidence? We argue here that better evidence is drawn from high quality primary research and rigorously synthesized in a way that shows not only what works, but why and how it works. Better evidence also means evidence that is relevant to and used by policy makers. This makes for better decisions and better programme implementation; this in turn helps towards improving lives. Effective responses to juvenile delinquency, for example, would mean a reduced rate of recidivism, and this gives a double dividend: fewer criminals – and fewer victims. Evidence-based decisions also improve the allocation of resources, diverting funding from ineffective to effective interventions. In an economic downturn, this is pertinent whether funding is public or private.

Despite this, a model whereby the policy is fixed around the evidence is far from the norm. A growing interest in the interaction of research and policy is often hampered by under-utilisation of scientific methods of evaluation. In education, there is failure to overcome the "knowledge to action gap"; as Levin puts it, "we would get large gains in student outcomes if we used universally what we already know about effective policy and practice." (1) In policing, innovations linked to scientific evidence are the exception rather than the rule. As Weisburd and Neyroud point out, "often the introduction of research develops serendipitously from a "bright idea" of police practitioners or researchers, rather than through systematic development of knowledge about practice." (2) In social welfare, lists of effective interventions are in vogue; but many of these are based on "unsystematic, partial, and potentially biased summaries of research evidence." (3) The development policy debate has moved towards a greater focus on results, and the establishment of Millennium Development Goals reflects this. Yet scientific rigour is largely absent from outcome monitoring. (4)

Can we bring the worlds of research and practice closer together? This essay focuses on four barriers to progress: (i) the dearth of primary empirical studies; (ii) the complexities of synthesising evidence from disparate sources; (iii) the need to consider both internal and external validity; and (iv) the funding and decision making framework. The scale of the challenge can be illustrated by a consideration of a recent review of school feeding programmes. (5) This review, co-registered with the Campbell and Cochrane Collaborations, suggests that such programmes, when properly implemented, improve attendance, improve math performance, enhance short term cognition, and produce small gains in weight.

(i) A tiny proportion of primary research is suitable for use in systematic reviews. Kristjansson's experience of proceeding from 5921 titles and abstracts (with duplicates) to 139 potentially useful articles, and then to 18 studies deemed fit for inclusion, is a familiar one. Furthermore, the 18 included studies were of uneven quality. Experiments, and particularly randomised experiments, should be built into the rollout of interventions whenever possible, provided that they meet criteria of avoiding harm and informed consent. Properly conducted RCTs allow greater degree of certainty about conclusions on effectiveness than other forms of evaluation. (6) Although there are questions too broad to be addressed by systematic reviews ("does education work?"), the scientific method can and should be applied to more focussed questions: "does one educational intervention work better than another?" Experiment, involving structured comparison, is often the best way to get to the truth of the matter.

(ii) The methodology of research synthesis needs continual attention. The Kristjansson review comprised a range of studies of different designs, from different time periods (1928 to 2003), in different countries, with great variation in implementation, and a wide range of outcome measures. Synthesis and meta-analysis were essential to meaningful recommendations. The study combined that which should have been combined (e.g. weight results from RCTs from LMIC in different studies) and was able to identify and precisely describe effects. "In... RCTs from LMIC, children who were fed at school gained an average of 0.39 kg more than controls over 19 months; in lower quality studies (CBAs), the difference in gain was 0.71 kg over 11.3 months. Children who were fed at school attended school more frequently than those in control groups; this translated into an average of 4 to 6 days a year per child. For educational and cognitive outcomes, children who were fed at school gained more than controls on math achievement, and on some short-term cognitive tasks." (5)

(iii) There is growing recognition that both internal and external validity need to be considered in rigorous systematic reviews. Evidence on the theory underlying the intervention and of the quality of the intervention itself (obtained from process evaluation or realist review) is as important as information on the quality of the study designed to assess effectiveness. Without evidence on intervention quality, a systematic reviewer might falsely conclude that an intervention does not work when, in fact, it was merely poorly implemented. There is huge scope for inadequate implementation. Examples include poor compliance, lack of trust between service deliverers and recipients, and a level of intensity of the intervention which is too low to make a difference. Recently, van der Knaap et al (7) advocated an approach that involves conducting a systematic review according to strict Campbell guidelines, followed by realist review on the included studies. This approach has merit as attention remains focused on rigorous studies, while it also articulates the theory and the mechanism behind the success or lack of success of the intervention.

Kristjansson et al used a similar approach. (5) The review team comprised experts in nutrition, psychology, internal medicine, statistics, and systematic and realist reviewing; all played important roles in judging not only the study design quality but also the quality of implementation. Process elements were extracted and a realist review of the included studies performed. (8) This revealed factors that may impact on the effectiveness of school meal programmes: energy given, substitution, extent of compliance, and benevolent attention. Finally, in a subsequent study by Galloway, a costing of school meals in four African countries was performed; this was combined with review results to provide evidence of the cost per outcome of school feeding programmes. (9) Hence the authors were able to make policy recommendations not only on what works, but on how programmes might be better implemented so that children receive the maximum benefit for lower cost.

(iv) Rigorous systematic reviews are a first step in building evidence based policy, but they are not the only step. Building an effective, continuing dialogue between researchers and end-users is vital in bridging the 'knowledge to action gap.' Knowledge translation is not a one-way street but "a complex system of social interaction among stakeholders." (10) In this case, the lead reviewer's e-mail contacts with officials of the World Food Program had several positive consequences, including an invitation to present findings to a WFP meeting, and the commissioning of two "Ten Minutes to Learn" policy briefs. (11; 12) This also made a 'central contribution' to the cost-outcome exercise described above, and "to the evidence-base used in a joint analysis by WFP and the World Bank Group, initiated in response to demand from countries for expansion of school feeding programs in response to the global food and financial crises." (13) The impact on policy remains to be seen.

The opportunity to directly reach senior decision making levels remains limited. Funding decisions reflect a web of strategic and political contingencies, almost always in the context of a gap between available resources and the demands on those resources. For

example, under-nutrition in utero and in early life (under the age of two) may cause permanent damage (14); early nutritional intervention is essential to maximize growth (15) and cognition. (16) Thus, increasing emphasis has been placed on feeding programmes for under-twos (17; 18). The problem arises when funding streams and budget structures force hard choices; more funding for pre-school programmes may mean less is available for school-age programmes.

The challenge in development assistance as in other areas is to shift the budgetary parameters, so that funding moves from harmful or ineffective programmes to effective programmes rather than from one effective program to another. Effectiveness should not be a matter of speculation but of solid evidence. It is hard to see this being remedied without a high level political consensus to mainstream a review of evidence in the funding cycle, and indeed to underpin that consensus in the relevant financial regulations. Who dares to lead such an initiative?

In his inaugural address President Obama stated "The question we ask today is not whether our government is too big or too small, but whether it works - whether it helps families find jobs at a decent wage, care they can afford, a retirement that is dignified. Where the answer is yes, we intend to move forward. Where the answer is no, programs will end." The call to arms is the easy part; the hard part is to create a regulatory and decision making frameworks that make it easier to channel scarce resources towards effective interventions, and away from ineffective ones. No one magic bullet will bring us to where we should be; the way forward is to refocus and refine our collaborative effort on a variety of fronts.

References

1. Levin B. Mobilizing knowledge to support better learning. [This collection.]
2. Weisburd D, Neyroud P. Policing and Scientific Evidence: Towards a New Paradigm. [This collection.]
3. Littell, J., Shlonsky, A. Toward evidence-informed policy and practice in child welfare. [This collection.]
4. White H. We all agree we need better evidence, But what is it and will it be used? [This collection.]
5. Kristjansson, E. School feeding for improving the physical and psychosocial health of disadvantaged students. *Campbell Systematic Reviews*, 2007.
6. Campbell DT, Cook TD. *Quasi-experimentation. Design and analysis issues for field settings*. Boston: Houghton Mifflin; 1979.
7. Van der Knaap L, Leeuw FL, Bogaerts S, Nijsson L. Combining Campbell Standards and the Realist Evaluation Approach. *American Journal of Evaluation* 2008; 9 (1):58-67.
8. Greenhalgh T, Kristjansson E, Robinson V. Realist review to understand the efficacy of school feeding programmes. *BMJ* 2007; 335:858-61.
9. Galloway, R., Kristjansson, E., Gelli, A., Meir, A., Espejo, F., and Bundy, D. School Feeding. Costs and Cost-outcomes. *Food and Nutrition Bulletin*. In press.
10. Graham ID, Logan J, Harrison MB, Straus SE, Tetroe J, Caswell W et al. Lost in Knowledge Translation: Time for a map? *Journal of Continuing Education in the Health Professions* 2006; 26: 13-24.

11. Kristjansson, E. Ten Minutes to Learn About... The effects of school feeding on cognition, school performance, and behaviour. Volume II, 2. 2007. World Food Program.
12. Kristjansson, E. Ten Minutes to Learn About... The effects of school feeding on growth. Volume II, 1. 2007. World Food Program.
13. Bundy, D. 13-3-2009, Personal Communication
14. Repositioning Nutrition as Central to Development. A strategy for large-scale action. World Bank, 2006.
15. Shrimpton R, Victora CG, de Onis M, Lima RC, and Clugston G. Worldwide timing of growth faltering: implications for nutritional interventions. Pediatrics 2001:107.
16. Martorell R. Undernutrition During Pregnancy and Early Childhood and its Consequences for Behavioral Development. 1996.
17. Grantham-McGregor S, Cheung YB, Cueto S, Glewwe P, Richter L, Strupp B. Developmental potential in the first 5 years for children in developing countries. The Lancet; 369 (9555):60-70.
18. Walker SP, Wachs TD, Meeks Gardner J, Lozoff B, Wasserman GA, Pollitt E et al. Child development: risk factors for adverse outcomes in developing countries. The Lancet; 369 (9556):145-57.

Corresponding author:

Eamonn Noonan, CEO
The Campbell Collaboration
PO Box 7004 St. Olavs plass
N-0130 Oslo
Norway
Email: eno@nokc.no

Toward evidence-informed policy and practice in child welfare

Julia H. Littell, Bryn Mawr College

Aron Shlonsky, University of Toronto

All societies care about the welfare of children, though childcare practices and definitions of maltreatment vary across cultures. Child welfare policies are shaped by local values, beliefs, and resources. These include convictions about the nature and scope of public (or community) responsibilities for dependent children, beliefs about what is good and bad for children, and competing claims for public and charitable funds. In some countries, children receive psychosocial services, material assistance, and/or alternative living arrangements in attempts to protect them from harm and promote healthy development. In wealthy nations, child welfare and child protection services may be seen as an integral component of social care (e.g., in Norway) or a set of residual programs for children and families whose needs are not met elsewhere (e.g., in the US). In some low- and middle-income countries, child welfare services are virtually nonexistent, as are organized efforts to identify child maltreatment.

Important advances in research have increased our ability to identify vulnerable children, assess their needs, track their whereabouts, and measure the impacts of social and behavioral interventions on children's safety and well-being.¹ Thus, we must find ways to use research evidence judiciously and in concert with other concerns if we are to succeed in protecting and enhancing the welfare of children.

We know that child welfare programs can have unintended, negative consequences and hidden effects, yet we have the means to detect such effects. For example, the earliest observational studies of "intensive family preservation programs" (IFPS) in the USA showed that abused and neglected children tended to remain with their families after brief, intensive, in-home services. It was not until these programs were subjected to randomized controlled trials that it became clear that most of these children would have remained at home even in the absence of IFPS. Further, it was discovered that IFPS could actually increase the detection of subsequent child maltreatment and, thus, increase the likelihood that children would be removed from their families.² While this result may (or may not) be desirable for children, it was clearly not the intended outcome.

Increasingly decision makers have demanded evidence about the effects of child welfare interventions. For example, when the US Congress approved \$1 billion in funding for IFPS in 1993, it directed the US Department of Health and Human Services to conduct a multi-site, randomized experiment to test the effects of these programs on subsequent child maltreatment and out-of-home placements.³

On the other hand, many innovations in child welfare have not been closely linked to evidence. The child welfare field seems to embrace one reform movement after another, even if the new reform is just an old wine in a new bottle. Child welfare program administrators have been drawn to branded interventions, which have sometimes been adopted on the basis of scant evidence.

Widespread implementation of ineffective programs can have serious financial, human, and opportunity costs. The costs of being wrong can be every bit as devastating in child

¹ Lindsey D, Shlonsky A, eds. Child welfare research: Advances for practice and policy. New York: Oxford University Press, 2008.

² Littell JH, Schuerman JR. A synthesis of research on family preservation and family reunification programs. Rockville, MD: Westat, 1995. <http://aspe.hhs.gov/hsp/cyp/fplitrev.htm>.

³ Westat. Evaluation of Family Preservation and Reunification Programs: Final report. Rockville, MD: Westat, 2002. <http://aspe.hhs.gov/hsp/evalfampres94/Final/index.htm>.

welfare as in health care. In child protection services, for instance, children can be severely harmed by their parents or wrongfully taken from their families. Ineffective treatment of behavioral problems in childhood can lead to extraordinary painful and costly problems in adulthood.

A little evidence goes a long way

Government and professional organizations have developed criteria to determine which interventions are effective for problems related to child maltreatment. Many organizations produce lists of “effective” programs and practices. These lists are important because they affect funding and policy decisions that will determine the future of child welfare services.

Several prominent groups use consensus-based standards of evidence to identify “evidence-based” programs that are implemented in child welfare settings. Examples include the California Evidence-based Clearinghouse (CEBC) for child welfare,⁴ Blueprints for Violence Prevention,⁵ the US National Registry of Evidence-based Programs and Practices,⁶ Coalition for Evidence Based Policy,⁷ and the American Psychological Association's Clinical Psychology Division.⁸ Most of these groups require two controlled trials showing some evidence of positive effects for a program to reach the “top tier” or “model program” status. These criteria allow programs with little evidence to achieve the highest rating. A comprehensive review of all of the relevant evidence (including grey literature) is not required, careful assessments of study methodology and implementation issues are not required, conflicts of interest are not always considered, nor is it necessary to consider whether results may be generalized to other populations and other settings. Such disregard for the basic principles of research synthesis can result in endorsements of programs that have little effect or even prove harmful.

Indeed much of what passes for empirical knowledge about the effects of child welfare programs is not based on sound principles of research synthesis. Instead, most sources of information on “programs that work” are derived from unsystematic, partial, and potentially biased summaries of research evidence.

For instance, to our knowledge there is no systematic review of research on the effects of IFPS in cases of child maltreatment. Some reviewers of this literature have expressed preferences for certain studies based on the outcomes of those studies with little attention to their methodological rigor.⁹ Thus, some reviews merely reflect proponents' opinions.

⁴ <http://www.cachildwelfareclearinghouse.org/>

⁵ Mihalic S, Fagan A, Irwin K, Ballard D, Elliott D. Blueprints for violence Prevention. Washington, DC: U.S. Department of Justice Office of Juvenile Justice and Delinquency Prevention, 2004.

⁶ US Substance Abuse and Mental Health Services Administration. National Registry of Evidence-based Programs and Practices. <http://www.nrepp.samhsa.gov/>

⁷ <http://www.evidencebasedprograms.org/>

⁸ The American Psychological Association (APA) Clinical Psychology Division standards are described in: Chambless DL, Baker MJ, Baucom DH, et al. Update on empirically validated therapies, II. *The Clinical Psychologist* 1998; **51**: 3-16.

⁹ Littell JH. Evidence or assertions? The outcomes of family preservation services. *Social Service Review* 1995; **69**:338-351.

Campbell and Cochrane reviews related to child welfare

Systematic reviews have been generated by key questions in child welfare policy and practice. Here we describe three such reviews (other examples are: ^{10, 11, 12, 13, 14}).

In response to concerns about the perceived failings of IFPS in child welfare, some observers suggested that child welfare programs should adopt Multisystemic Therapy (MST), a “model program” that was originally developed in juvenile justice settings. Indeed, MST has been widely replicated in diverse settings on the basis of nonsystematic reviews that claim that MST is effective across problems and populations.¹⁵ A joint Campbell and Cochrane review found that MST was not consistently better or worse than any of the alternatives to which it had been compared.¹⁶

An award-winning Campbell/Cochrane review compared outcomes of kinship foster care with those of traditional, non-relative foster care.¹⁷ Results suggest that children placed with relatives demonstrated better developmental and mental health outcomes, and had more stable living arrangements. Further, there were no differences between kinship care and regular foster care in terms of rates of reunification of children with birth parents. Children placed in non-kin foster homes were more likely to be adopted and more likely to use mental health services. Although methodological limitations of the original studies necessitate caution in interpreting results, this review added much-needed information to a longstanding debate about the relative merits of kinship care and foster care.

Less attention has been paid to diagnostic and prognostic questions in the fields of social care than in medicine, yet the implications of incorrect assessments are every bit as far-reaching. In child welfare, families investigated for child maltreatment are assessed for the likelihood that they will injure their children in the future. An incorrect prognosis can lead to the wrongful removal of children from their parents or, likewise, leave children in harm’s way. Campbell’s first systematic review of prognostic tools will ascertain the

¹⁰ Donkoh C, Underhill K, Montgomery P. Independent living programmes for improving outcomes for young people leaving the care system. *Cochrane Database of Systematic Reviews* 2006, Issue 3. Art. No.: CD005558. *Campbell Systematic Reviews* 2006.

¹¹ Macdonald G, Ramchandani P, Higgins J. Cognitive-behavioural interventions for children who have been sexually abused. *Cochrane Database of Systematic Reviews* 2006, Issue 4. Art. No.: CD001930. *Campbell Systematic Reviews* 2006.

¹² Macdonald G, Turner W. Treatment Foster Care for improving outcomes in children and young people. *Cochrane Database of Systematic Reviews* 2008, Issue 1. Art. No.: CD005649. *Campbell Systematic Reviews* 2008.

¹³ Turner W, Macdonald G, Dennis J. Behavioural and cognitive behavioural training interventions for assisting foster carers in the management of difficult behaviour. *Cochrane Database of Systematic Reviews* 2007, Issue 1. Art. No.: CD003760. *Campbell Systematic Reviews* 2007.

¹⁴ Zwi K, Woolfenden S, Wheeler D, O’Brien T, Tait P, Williams K. School-based education programmes for the prevention of child sexual abuse. *Cochrane Database of Systematic Reviews* 2007, Issue 3. Art. No.: CD004380. *Campbell Systematic Reviews* 2007.

¹⁵ Littell JH. Evidence-based or biased? The quality of published reviews of evidence-based practices. *Children and Youth Services Review* 2008; **30**:1299-1317.

¹⁶ Littell JH, Popa M, Forsythe B. Multisystemic therapy for social, emotional, and behavioral problems in youth aged 10-17. *Cochrane Database of Systematic Reviews* 2005, Issue 4. Art. No.: CD004797. *Campbell Systematic Reviews* 2005:1 DOI: 10.4073/csr.2005.1.

¹⁷ Winokur M, Holtan A, Valentine D. Kinship care for the safety, permanency, and well-being of children removed from the home for maltreatment. *Cochrane Database of Systematic Reviews* 2009, Issue 1. Art. No.: CD006546. *Campbell Systematic Reviews* 2009:1. DOI: 10.4073/csr.2009.1.

psychometric properties of several widely used risk assessment instruments in an effort to maximize the use of reliable and valid predictors of further maltreatment.¹⁸

Putting it all together: Evidence-informed decisions

Although rigorous evidence about the impacts of child welfare programs and policies is needed to inform policy and practice, this evidence cannot tell us what to do. Even the best evidence must be combined with other considerations to formulate wise decisions.

For example, if intensive, in-home services do not prevent (and might increase) the removal of maltreated children from their homes, what should policy makers and practitioners do? The answer depends, in part, on their goals: If protecting children is paramount, then intensive services offer some advantages; if preserving families is paramount, other approaches should be tried.

If MST is no more or less effective than its alternatives, then our choices can be based on other considerations. Following the publication of the Campbell/Cochrane MST review, MST was adopted in some jurisdictions because decision makers liked the structure and documentation that it provides. Elsewhere MST was abandoned because it was seen as too costly or inconsistent with local cultural norms. All of these decisions are legitimate, in light of the current best evidence of the program's impact.

The widespread adoption of "model" programs can squelch innovation and adaptations necessary to meet individual needs, respond to local conditions, and respect cultural traditions. As in medicine, evidence-informed policy and practice in child welfare should increase our options, not restrict them.¹⁹

Conclusions

Decision makers need comprehensive, reliable, and unbiased syntheses of credible evidence to make well-informed choices. They need to know about the accuracy of the decision-making tools and the impacts of child welfare services for various problems, populations, and settings. Systematic reviews can provide such evidence, thus they are essential for decision-making in child welfare. Decision makers must, however, use this evidence judiciously and in concert with other concerns.

Corresponding author:

Julia H. Littell, Professor
Graduate School of Social Work and Social Research
Bryn Mawr College
300 Airdale Rd.
Bryn Mawr, PA, 19010 USA
email: jlittell@brynmawr.edu

¹⁸ Shlonsky A, Saini M, Wu M-J. The recurrence of child maltreatment: Predictive validity of risk assessment instruments. Protocol. *The Campbell Library* 2007.
http://www.campbellcollaboration.org/campbell_library/index.php.

¹⁹ Dickersin K, Straus SE, Bero LA. Evidence based medicine: increasing, not dictating, choice. *British Medical Journal* 2007; **334**(1):s10.

Mobilizing knowledge to support better learning

Ben Levin, Ontario Institute for Studies in Education, University of Toronto

School systems, like health systems, faces the challenge of finding ways to connect evidence to the decisions of governments and to the daily practices of large numbers of practitioners (Grimshaw et al., 2006). In both fields, early optimism that research could guide policy and practice in a direct way has been replaced by an awareness of how difficult these connections are.

First, the good news. There has been a surge of interest in how education research affects policy and practice. There are more studies, more publications and many more practical efforts to assess impact and improve the situation. Governments and professionals are more inclined to search for and use research in shaping their work, and there is much more research to draw on than used to be the case. For example, we know a great deal more than we used to in such diverse areas as teaching reading, engaging parents, motivating students, and working with various disabilities.

The interest is not just theoretical. Many organizations, such as universities or school systems, have made more effort to build research into their work. Many countries have taken steps to strengthen these connections by creating new structures, processes or institutions (e.g. Alton-Lee, 2007; Nutley et al., 2007; OECD, 2007).

The widespread view that research is largely ignored in education is not supported by the evidence (Cooper, Levin, & Campbell, in press; Biddle & Saha, 2002; Rickinson, 2005). Educators at all levels are much more interested and aware than they used to be of the potential value of research. Most educators have a range of connections to research, through professional reading, professional development, graduate work, or other means. Most published material for schools, including the materials produced by professional groups and school systems, schools does make use of relevant research findings. Increasingly, professional development for teachers and others draws extensively on research. Schools and school systems are investing more effort in analyzing data on student outcomes in order to guide practice. There can be no doubt that research has a more prominent place in every area of education than was the case ten or twenty years ago (Levin, 2004).

The key role of third parties in knowledge mobilization in education has become increasingly evident (Honig, 2004; Levin, 2008). Much of the connecting of research to practice for schools is done by groups such as teacher organizations, agencies that provide professional development, lobby groups and the media. Recognizing this connection, new organizations and activities have arisen with a focus on connecting research to policy and practice in education, including new local partnership bodies, national organizations such as research centres, and international bodies such as the Campbell Collaboration (www.campbellcollaboration.org). Various research and development organizations have been created at national and sub-national levels with a specific mandate to promote and share research about schooling.

These developments, however, do not mean one can be complacent about the current situation; much room for improvement remains even if one is not naïve about the potential contributions of research in a political world (Levin, 2008). The situation is also much more difficult in developing countries, where the teaching workforce is less educated, mediating agencies are fewer, and policy processes are even more subject to short-term political pressures. Developing countries also face the challenge of a lack of research that is relevant to their contexts.

The barriers to more effective use of research in education are multiple and real (Nutley et al., 2007). They include insufficient evidence, poor availability of evidence even when it does exist, lack of skill in finding and interpreting evidence, lack of infrastructure to

support research use, strong inertial forces around existing practices, and various pressures in directions contrary to the evidence.

Despite increasing effort, our knowledge about how research affects education practice is still quite limited. In many important areas the empirical evidence is still thin, and research methods to provide better evidence also need development. The education sector still spends much less of its overall budget on research than is the case in health. . Not enough studies are done and studies tend to be small scale and oriented towards cases or interviews. There is not enough replication or cumulative work. Too many studies construct new frameworks instead of building on the work of others. Much remains to be done to improve our research capacity.

Even where there is strong research evidence, teaching as a profession still does not have the same tradition of using research as seems to be the case in the health professions (Hargreaves, 1999; Slavin, 2002). Teacher training gives less attention to the importance of research than do nursing or medical education. Although much remains to be learned about effective education policy and practice, we would get large gains in student outcomes if we used universally what we already know about effective policy and practice. For example, significant resources are used for students who are repeating courses or grades instead of being used to help them succeed the first time. Similarly, there is substantial evidence on way to increase students' intrinsic motivation, which is closely linked to achievement, yet these practices are not as widespread as they should be and there is still much reliance on extrinsic motivators.

Although teaching is often thought of as an individual activity, in education the use of knowledge is mediated through other social and political processes. Teaching is deeply shaped by current and past practices even where these are quite contrary to the evidence. Simply telling someone there is a better way to do it rarely changes their behaviour – just as most people do not stop smoking or start exercising because they are told that would be good for them.

Changing practices in complex systems is a matter of changing the social contingencies around practice. The most powerful vehicle for moving evidence into practice in schools lies in changing the organization of daily work so that evidence is more deeply embedded. This will require a closer connection from research to ongoing school activities such as leadership development or professional development.

At present very few schools and school systems have much infrastructure to support using research. By 'infrastructure' I mean people and systems that are designed to find, share, promote and apply relevant research to the daily problems of policy and practice. A look at the websites of schools, school districts or ministries of education all around the world shows how little attention research receives in the way most of these organizations communicate their work (this analysis is currently being done by the OISE KM team; papers will be posted at www.oise.utoronto.ca/rspe). Most education organizations do not have systems that ensure that relevant research is available to, let alone regularly used by their members. Since we know that teaching practice is shaped more by the practices of peers than by reading research, it is important to find ways in which research findings can be translated into real practices that people find meaningful and practical.

It is worrying that universities and faculties of education appear to give little attention to organized knowledge mobilization in education, notably so in comparison to technology transfer efforts in areas such as science or medicine which are much better organized and supported. For example, very few universities provide good access to the research produced by their education faculty. At best there might be lists of projects or reports, but typically little beyond that.

A neglected area in knowledge mobilization in education is the role of graduate or advanced study. Large numbers of teachers and school administrators participate in

graduate study or advanced continuing education, where they have extensive contact with research and researchers (Hemsley-Brown, 2003). However typically neither universities, who provides these programs, nor the organizations in which the students work take much advantage of this experience to build ongoing relationships with researchers, or to strengthen their internal capacity to share and use research findings. This represents a promising area for progress – if, for example, graduate students received both specific advice and internal support for playing a mediating or brokering role around research in their home organizations.

What Next?

Next steps follow from the preceding diagnosis.

- Schools, school systems and government ministries need to increase their capacity and infrastructure for knowledge mobilization. Steps such as assigning someone responsibility for locating and sharing relevant research, or putting research results on the agenda of regular staff meetings, would be simple yet very helpful.
- The vital role of third parties as primary mobilizers of knowledge in education needs to be recognized and accommodated; more could be done to take advantage of the work of existing third parties such as those who provide professional development for teachers or principals.
- Current efforts to share and use research should themselves be the subject of research and evaluation. We simply need to learn more about the effects of various efforts.

Conclusion

This review suggests that action for knowledge mobilization in schooling is needed on several fronts simultaneously in order to improve the way that research supports and influences policy and practice across the education sector.

References

- Alton-Lee, A. (2007). Making a bigger difference for diverse learners: The iterative best evidence synthesis programme in New Zealand. Paper presented to the American Educational Research Association, Chicago, April.
- Biddle, B. & Saha, L. (2002). The untested accusation: Principals, research Knowledge, and policy making in schools. Westport, CT: Ablex.
- Cooper, A., Levin, B and Campbell, C. (In press). The growing (but still limited) importance of evidence in education policy and practice. *Journal of Educational Change*.
- Grimshaw, J., Eccles, M., Thomas, R., MacLennan, G., Ramsay, C., Fraser, C., & Vale, L. (2006). Toward evidence-based quality improvement: Evidence (and its limitations) of the effectiveness of guideline dissemination and implementation strategies 1966-1998. *Journal of General Internal Medicine*, 21, S14-20.
- Hargreaves, D. (1999). Revitalizing educational research: lessons from the past and proposals for the future. *Cambridge Journal of Education*, 29(2), 239-249.
- Hemsley-Brown, J. (2004). Facilitating research utilization: A cross-sector review of research evidence. *The International Journal of Public Sector Management*, 17(6), 534-552.
- Honig, M. (2004). The new middle management: Intermediary organizations in education policy implementation. *Educational Evaluation and Policy Analysis*, 26(1), 65-87.

Levin, B. (2004). Making research matter more. *Education Policy Analysis Archives*, 12 (56). Retrieved November 15, 2008 from <http://epaa.asu.edu/epaa/v12n56/>

Levin, B. (2008). Thinking About Knowledge Mobilization. *Paper prepared for an invitational symposium sponsored by the Canadian Council on Learning and the Social Sciences and Humanities research Council of Canada*. Vancouver.

Nutley, S. M. Walter, I., & Davies, H. T. O. (2007). *Using evidence: How research can inform public services*. Bristol: Policy Press.

OECD. (2007). *Evidence in education: Linking research and policy*. Paris: OECD.

Rickinson, M. (2005). *Practitioners' use of research*. Retrieved September 21, 2006, from <http://www.nerf-uk.org/word/WP7.5-PracuseofR.doc?version=1>

Slavin, R. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7), 15–21.

We all agree we need better evidence. But what is it and will it be used?

Howard White, Executive Director, International Initiative for Impact Evaluation (3ie)

The challenge of development is to improve the quality of life across the world in a sustainable manner. What we mean by the quality of life is captured in the widely-accepted Millennium Development Goals (MDGs), which include the target to reduce infant and under-five mortality by two-thirds between 1990 and 2015, and to reduce HIV/AIDS prevalence.²⁰ Other goals include halving income-poverty and ensuring that all children are in school. Good health is recognized not just as an end in itself, but as a necessary condition for achieving the other goals.

Despite broad acceptance of the MDGs, each year two million African children die before reaching their first birthday. Another two million die before reaching their fifth birthday. These numbers have not fallen in the last three decades. In 2005, 1.4 billion people lived below the poverty line of US\$1.25 a day²¹ That is, one quarter of the population of the developing world live in poverty, a proportion which has not declined over the last 25 years. In many countries poverty is now higher than at independence forty to fifty years ago.

So whilst some countries, notably those in East Asia, have made rapid progress, many others have not. This lack of progress cannot be blamed on inadequate resources alone. Billions of dollars are spent annually on development programs. In recent years development aid has topped US\$100 billion a year, and far more than this is spent by developing country governments from their own revenues. In fact, as is being documented in an increasing number of places,^{22,23} there is little evidence on the impact of development programs. But this very lack of evidence is reason to doubt that the spending has been put to best use.

The call for stronger evidence has come from a number of sources. Over the last fifteen years governments around the world have adopted the 'Results Agenda'. No longer was the performance of government programs to be judged by attaining spending targets, or even delivering agreed outputs. Rather, there had to be a demonstrable impact on welfare outcomes such as poverty, mortality and empowerment. In the development field the focus on results has coalesced around the Millennium Development Goals (MGDs). Agencies, such as USAID and the United Kingdom Department for International Development (DFID), adopted aggregate-level outcome monitoring at global, regional and national level, for example tracking trends in under-five mortality. However, it was soon apparent that outcome monitoring said nothing about attribution: were development programs behind the observed changes or not?^{24,25}

Existing evaluations were found to be wanting when it came to answering this question. Most evaluations had focused more on management and implementation, what are usually called process evaluations, than on outcomes and impact. Studies with outcome measures were unable to answer the attribution question with any confidence. Some groups, notably the Poverty Action Lab at MIT, have been active in promoting

²⁰ <http://www.un.org/millenniumgoals/>

²¹ Chen S and Ravallion M, *The Developing World Is Poorer Than We Thought, But No Less Successful in the Fight against Poverty*, *Policy Research Working Paper 4703* 2008, World Bank, Washington D.C., USA.

²² Centre for Global Development, *When will we ever learn? Report of the Evaluation Gap Working Group* 2006, Washington D.C., USA

²³ Jerve AM and Villanger E, *The Challenge of Assessing Aid Impact : A Review of Norwegian Evaluation Practice*, *Evaluation Study 1/2008*, May 2008, NORAD, Oslo (revised version forthcoming in the *Journal of Development Effectiveness*).

²⁴ National Audit Office, *Performance Management - Helping to Reduce World Poverty*, Report by the Comptroller and Auditor General 2002, HC 739 session 2001-2002: 12 April 2002.

White H, *Using Targets to Measure Development Performance*, *Targeting Development 2004*, Edited by Black R and White H, Routledge, London, UK.

randomized control trials (RCTs) to fill this 'evaluation gap'.²⁶ RCTs are of course the norm for medical researchers, and have been common in developing countries for public health interventions including nutrition and water treatment. But they have been virtually unknown for other social and economic development programs.

A vigorous debate has ensued over the use of RCTs to assess development interventions.^{27,28,29,30} Some criticisms come from those skeptical about the application of any quantitative methods in assessing development interventions. But to date these critics have not shown how qualitative approaches alone can capture outcomes for large-scale interventions, let alone can adequately assess causation. Others point to practical, ethical and political constraints to randomization. There is some merit in these arguments. Randomization is most applicable for discrete, homogenous interventions, analogous to drug treatments for which the approach is commonly used. There are such cases, for example the impact of deworming on school attendance,³¹ and, most famously, 'conditional cash transfers' by which families receive money for adopting certain behavior, such as keeping girls in school of which the best known is the Progres program in Mexico (later renamed Oportunidades).³² RCTs have become much more common in the last three to four years, incorporating both health and nutrition components. For example, transfers are now being conditioned upon pregnant women attending ante-natal clinic. But there are many cases when randomization is not possible, not least because the evaluators are only called in at the end of the program, so statistical matching procedures, called quasi-experimental methods, have to be used to create an untreated control group. A few years ago there were very few evaluations applying either experimental or quasi-experimental approaches. But now there are many more such studies. The question is, how can they translate into better policy and so better development outcomes?

The International Initiative for Impact Evaluation (3ie) was created to promote evidence-based policy making in development programs. 3ie stresses the need for 'quality impact evaluation', of which technical rigour is a necessary, but not sufficient, part. Policy influence is partly a matter of communication. Policy makers are not interested in the technique of instrumental variable double difference estimation. They want to know how to get better development outcomes and what cost. But policy influence can also be increased by using well-contextualized evaluation designs, drawing on the principles of theory-based evaluation. Theory-based evaluation may be contrasted to 'black box' evaluation designs. The black box approach simply reports the mean difference in outcomes between treated and untreated groups. For a medical treatment, such an approach can suffice, though it would be well to allow for impact heterogeneity between treatment groups – is the treatment more or less effective for people of particular age, sex or physical condition? But for more complex interventions a more elaborated approach examining the theory behind the intervention – how the inputs should result in the intended outcomes – is of use. Some examples illustrate this point.

A nutrition project in Bangladesh was intended to identify growth faltering infants and young children through regular growth monitoring by community nutrition workers.³³ Targeted children received supplementary feeding and their mothers got nutritional counseling. However, the program was found to have little overall impact, being most beneficial to the most malnourished children (and so having its greatest impact in the lean season). So why low impact? A variety of reasons were identified. In the first

²⁶ Poverty action Lab, <http://www.povertyactionlab.com/>

²⁷ Banerjee A, *Making Aid Work* 2007, MIT Press, Cambridge, USA.

²⁸ Deaton A, Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development, *NBER Working Paper 14690*, 2009 (<http://papers.nber.org/papers/w14690>)

²⁹ Ravallion M, Should the Randomistas Rule?, *Economists' Voice*, February 2009, Berkeley Electronic Press.

³⁰ White H and Bose R, *Journal of Development Effectiveness* 2, 2009.

³¹ Miguel E and Kremer M, Worms: Identifying Impacts On Education And Health In The Presence Of Treatment Externalities, *Econometrica*, 2004, v72(1), 159-217.

³² IFPRI, Mexico: Progres – Breaking the cycle of poverty, 2002, IFPRI, Washington D.C. (<http://www.ifpri.org/pubs/ib/ib6.pdf>)

³³ White H and Masset E, The Bangladesh Integrated Nutrition Program: findings from an impact evaluation, *Journal of International Development* 19: 627-652, 2006.

instance, mothers are not the decision makers in rural households especially if they live with their mother-in-law. Participation was lower for such women, and, even if they participated, it was not they who made the nutritional decisions in the home. Supplementary feeding was frequently not supplementary but displaced existing meals. Low impact also resulted as, whilst many malnourished children were missed by the program, children who were not growth faltering were admitted. The reason for this mis-targeting was that the majority of community nutrition workers proved unable to correctly interpret the growth charts, and so decide who should be in the program. And for women who were included, the behavior change communication (BCC) frequently did not result in behavior change, partly because of the role of other decision-makers (mothers-in-law again), and partly because of resource constraints. This example shows how unpacking the causal chain provides concrete policy advice, such as including fathers and mothers-in-law in the nutritional counseling, and better training of community nutrition workers.

Another reason for low impact was that the nutrition supplement suffered from both leakage (it was given to someone else) and substitution (it replaced existing foods rather than being additional). In a nutrition program in Nicaragua low impact was traced to the fact that mothers were not giving the supplement to their children as they reported that the children did not like it, and the mothers also thought it damaged the children's teeth and upset their stomach.³⁴

Taking an example outside of the health sector, it was found that the poorest rural households did not connect to the grid even though they spent more on lower-grade energy from kerosene lamps than the cost of grid electricity. The reason was that they could not afford the initial connection charge, and despite the many proven benefits from these connections.³⁵ Analysis of the data showed that the majority of households connected in the first two to three years, but around a fifth did not do so even after fifteen years. The policy implication is that connection subsidies could be successfully targeted to the poorest by introducing them three years after electricity arrives in a village.

But we need be flexible and look at evidence even when we lack a theory. Recall that Semmelweis was criticized as his empirical findings had no theoretical foundation at the time. The rural electrification study provides an example of when a black box approach may suffice. Women in households having electricity connections have significantly lower fertility (having of course controlled for the fact that these women are in better off, more educated households, which are also correlates of fertility). There are many channels which may explain this link, access to television appearing to be the strongest, a study from Brazil finds that soap operas have played an important role.³⁶ Whilst these channels cannot be provide conclusively, it is still worthwhile reporting the electricity-fertility link. The same World Bank study showed how electrification helps preserve the cold chain, but does not affect overall vaccination coverage since National Immunization Days reach outlying areas – but building routine vaccination into the health system is more cost effective.

Looking at such policy-relevant issues means moving beyond pure impact analysis, but also beyond purely quantitative approaches. The importance of mothers-in-law in Bangladesh was gleaned from the anthropological literature, which led us to unpack the household roster in the household survey to identify women in this position. Focus group discussions regarding the 'knowledge-practice gap' reinforced the role of mothers-in-law in preserving traditional practices (as one group said "we'll do these new things when our mothers-in-law have passed"). Participatory 'poverty assessments' in the nineties

³⁴ Ruel M and Leroy J, The Impact of Conditional Cash Transfers on Nutrition: a review of evidence using a program theory framework, *Journal of Development Effectiveness* (forthcoming).

³⁵ IEG, The Welfare Impact of Rural Electrification: a reassessment of the costs and benefits, 2007, World Bank, Washington D.C., USA.

³⁶ La Ferrara E, Chong A and Duryea S, Soap Operas and Fertility: Evidence from Brazil, *BREAD Working Paper No. 172*, March 2008

shed light on constraints to utilizing health services, including the ill-treatment the rural poor received at the hands of some health centre workers.³⁷ An anthropological study of rural electrification in Zanzibar³⁸ showed how poor information meant households being charged a fixed tariff unnecessarily reduced consumption, so that the potential benefits went unrealized.

Policy relevance also comes from ensuring that studies can address the cost-effectiveness of alternative means of achieving the same outcome. Studies show that point-of-use treatment has a greater health impact (child diarrhea incidence) than the provision of community water supply.^{39,40} The reason for this result is well known. Community water is stored in the household in open containers and becomes contaminated before use. Household connections avoid this problem but are very expensive. The conclusion appears to be that the global development target for clean water should be replaced with a more effective strategy of promoting point-of-use treatment, which would have higher impact at lower cost. However, this is not so straightforward. Installing community water supply delivers time-saving benefits, typically the time spent collecting water falls from 45-90 minutes per household a day to just 15 minutes. But point-of-use treatment entails costs to the household for chemicals and equipment, as well as time. Unless households place sufficient weight on the health benefits they will not want to incur these additional costs. A study simply of the impact of the treatment alone does not give all the policy information needed – but hardly any studies of the impact of point-of-use treatments addresses their sustainability.

It could be argued that what is needed in such cases is better health information. So an evaluation design could include three treatment groups, what I call, an A, B and A+B design. This design examines the impacts of the water treatment (A) and health education (B) alone, and the two combined (A+B). Existing evidence, though it is slight, suggests no greater impact from A+B, than either A or B separately. Evaluations which allow for variations in intervention design in this way yield better policy information, though an untreated control should also be included for calibration of impact compared to no intervention, and so the cost effectiveness of the different interventions.

In summary, there is growing momentum behind better evidence on the effectiveness of development spending. There is still a need for a much broader evidence base, but several recent initiatives, including 3ie, are expanding that base. But in addition to conducting more studies, we also need better studies which result in a bigger impact from development spending. Achieving this means engaging policy makers and adopting policy-relevant evaluation designs, which entails engaging a broader range of techniques than rigorous quantitative impact evaluation.

³⁷ Chambers R, Poverty and livelihoods: whose reality counts?, *Environment and Urbanization*, April 1995, 7, 173-204

³⁸ Winther T, [The Impact of Electricity: Development, Desires and Dilemmas](#), 2008, Berghahn Books, Oxford.

³⁹ Fewtrell L and Colford J., Water, Sanitation, and Hygiene: Interventions and Diarrhoea: A Systematic Review and Meta-Analysis, *Health Nutrition and Population Discussion Paper No. 34960*, 2004, World Bank, Washington D.C.

⁴⁰ IEG, *What Works in Water Supply and Sanitation? Lessons from Impact Evaluations*, 2008, World Bank, Washington D.C.

Better Evidence for a Better World

Mark W. Lipsey, Vanderbilt University, USA

Medical practitioners in bygone eras no doubt were convinced by the evidence of their own eyes, the wisdom of their clinical judgment, and the prevailing understanding of disease that such treatments as bloodletting, mercury therapy, and lobotomy were effective. What psychologists now recognize as confirmation bias¹—the tendency to interpret evidence in ways favoring preexisting beliefs—is a powerful force, especially for those who must act upon that evidence and justify those actions to clients, patrons, and critics. While not immune to such impulses, modern medicine is now far more skeptical of subjective forms of evidence and demands a higher standard of proof for claims that a practice is effective.

In the domain of social interventions we see far less skepticism. As the previous essays in this series attest, few of the programs routinely used in education, criminal justice, child welfare, and social development have been tested for effectiveness against a credible standard of evidence. Nor do the practitioners, clients, and advocates of these programs show much inclination to ask for such evidence. The problems at issue are largely ones of human behavior, and the programs that address them are typically based on concepts that are not nearly so different from everyday understanding as matters of blood chemistry or cell growth. Every voter, city council member, and legislator has well-established ideas about how to change human behavior, whether the topic is educating children, reforming criminal offenders, or alleviating poverty. Professionals responsible for providing the corresponding services have especially strong opinions about what works. For them, the validity of their understanding of the problems addressed and the effectiveness of their practice is amply confirmed by a wealth of evidence from their direct observations and experience.

It would be a simpler world, and one in which improving the human condition would be far easier, if the observations, experiences, and intuitions on which social programs and practices are readily based were indeed reliable guides for effectively attaining the intended benefits. We might then safely assume that most such endeavors were effective without demanding further evidence and be wary only of the occasional outlier that needed attention. When social programs and practices are put to rigorous test, however, we frequently find no measureable benefits. Peter Rossi, one of the giants in the field of social program evaluation, once referred to research on the effectiveness of social programs as “a parade of null results.”² Even more troubling are the instances where adequate evidence has shown the effects to be harmful. Prison visitation programs of the “Scared Straight” genre, which expose juvenile offenders to prison conditions and adult inmates who warn them in graphic terms of the consequences of continued criminal behavior, for example, have enormous intuitive appeal and were widely adopted during the 1980s. As controlled studies slowly accumulated, the results of a systematic review showed almost without exception that juveniles subjected to these programs committed more crimes afterwards, not fewer.³ Faith in personal intuition and confirmation bias are powerful forces, however, and some of these programs continue to have committed advocates.

The interjection of unproven social programs into people’s lives under the guise of helping is little more than quackery. The antidote is the nascent, but growing movement toward evidence-based practice. A significant challenge, however, is determining what constitutes adequate evidence. Few social programs bring about such dramatic, distinct, and extraordinary effects that objective observers can be confident of their benefits without a more probing evaluation. Unlike the spoof of randomized controlled trials widely circulated among critics,⁴ social programs do not function as parachutes that so clearly avert otherwise certain death among those jumping out of airplanes that their effectiveness is obvious. With rare exceptions, the best that social programs can do with the difficult problems they address is produce incremental improvement against a

background of great natural variability in outcomes. Under these circumstances, it is well-executed randomized controlled trials that produce the most convincing evidence of effectiveness. Though there are occasionally insurmountable practical barriers to implementing RCTs with social interventions, the thousands of instances in which it has been accomplished attest to its feasibility.

Those thousands, however, fall well short of covering the significant social interventions currently in actual or potential use. The other essays in this series are clear on the need for better evidence about what works to guide practice and policy for social programming. Better evidence, however, is not simply a matter of more evidence, or more rigorous evidence, though both are required. Important developments in the nature of that evidence must be accelerated for it to be truly useful. Most social programs are not well-defined in the sense of having a detailed protocol that specifies what they are and how they are to be delivered. Rather, they are more like what mathematicians call a fuzzy set, groupings characterized by key family resemblances that identify them as instances of a given program, but with multiple dimensions of variability from one instance to another. Among the more general dimensions of variability are different mixes of component elements, as when a science curriculum does or does not include the supplemental instructional computer program. Quantity of service often varies widely as well, e.g., group counseling for adolescent substance abusers may range from a few sessions in one implementation to dozens in another. Service quality, in turn, can vary for different elements within a program as well as between programs. Moreover, even when generally effective, social programs often have different effects for different subgroups of recipients.

Adequately characterizing the effects of such programs with RCTs must attend to this multidimensionality by incorporating thoughtfully selected moderator variables with potential for accounting for differential effects. The question for social programs is not, "what works," but "what variant works, for whom, under what circumstances." Though this is not so different from the analogous questions for interventions in medicine and public health, they are especially critical for social programs because of the generally greater variability they display in both the interventions and the responses of the recipients.

A further implication of this situation is that a handful of RCTs, each conducted in its distinctive circumstances, is not likely to provide a sufficient breadth of evidence to provide a sound basis for practice and policy. It is especially necessary for social programs to be tested via many studies that encompass a diversity of program variants, recipient characteristics, organizational and service delivery arrangements, and settings. A raft of such studies allows a better estimate of the robustness of the overall effects, the variables associated with differential effects, and the recipient and setting characteristics over which effects do and do not generalize.

The results of those studies, in turn, must be integrated, analyzed, and interpreted in ways that reveal what, when, and how positive effects are produced, and do so in a way that will be informative to practitioners and policymakers. This, in turn, presents a challenge to research synthesis. Reviews of the standard set by the Campbell Collaboration and Cochrane Collaboration must be conducted to ensure systematic and objective treatment of the evidence. However, such reviews must pay more attention than is represented by current practice to the heterogeneity of effects and the moderator variables that may be associated with differential effects, both those associated with program variants and those associated with recipient characteristics. In meta-analysis of effectiveness studies of social interventions, we must be more concerned with the variance of the distribution of effect sizes than with the mean. Around that mean we find variants and circumstances where there are large effects and others with little or no effect, and sometimes negative effects. It is critical to the interpretation of evidence for evidence-based practice is to know which is which.

But, as my colleagues have observed repeatedly in the other essays in this series, better evidence has little practical value if it is not used by practitioners and policymakers. On this point, we must reexamine our assumptions about the relationship between research and practice, and researchers and practitioners. The current framework is one in which researchers develop evidence and evidence-based programs and offer them up as gifts to those responsible for actually implementing interventions. Like the birthday tie that doesn't quite match the rest of the wardrobe, those gifts are not often appreciated or used. Weisburd and Neyroud⁵ propose, in the context of policing practice, that research must move from the outside to the inside to become a natural and organic part of the functioning and mission of service agencies. Similarly, we can imagine research as a more integral component of the infrastructure of policymaking bodies. Such integration would not only facilitate the use of research evidence to shape practice, but would shape the research to produce evidence more useful for practice. Finding ways to attain this integration of research and practice may be the greatest of the many challenges to the ideal of evidence-based practice. These are challenges that must be confronted with vigor, ingenuity, and persistence if indeed we are to find and use better evidence to create a better world.

References

¹ Nickerson, RS. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 1998; 2: 175-220.

² Rossi PH, Wright JD. Evaluation research: An assessment. *Annual Review of Sociology* 1984; 10: 331-52.

³ Petrosino A, Turpin-Petrosino C, Buehler J. 'Scared Straight' and other juvenile awareness programs for preventing juvenile delinquency (updated C2 review). *Campbell Systematic Reviews* 2003.
http://www.campbellcollaboration.org/campbell_library/index.php.

⁴ Smith GCS, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomized controlled trials. *British Medical Journal* 2003; 327: 1459-61.

⁵ Weisburd D, Neyroud P. Policing and scientific evidence: Toward a new paradigm. [this collection].