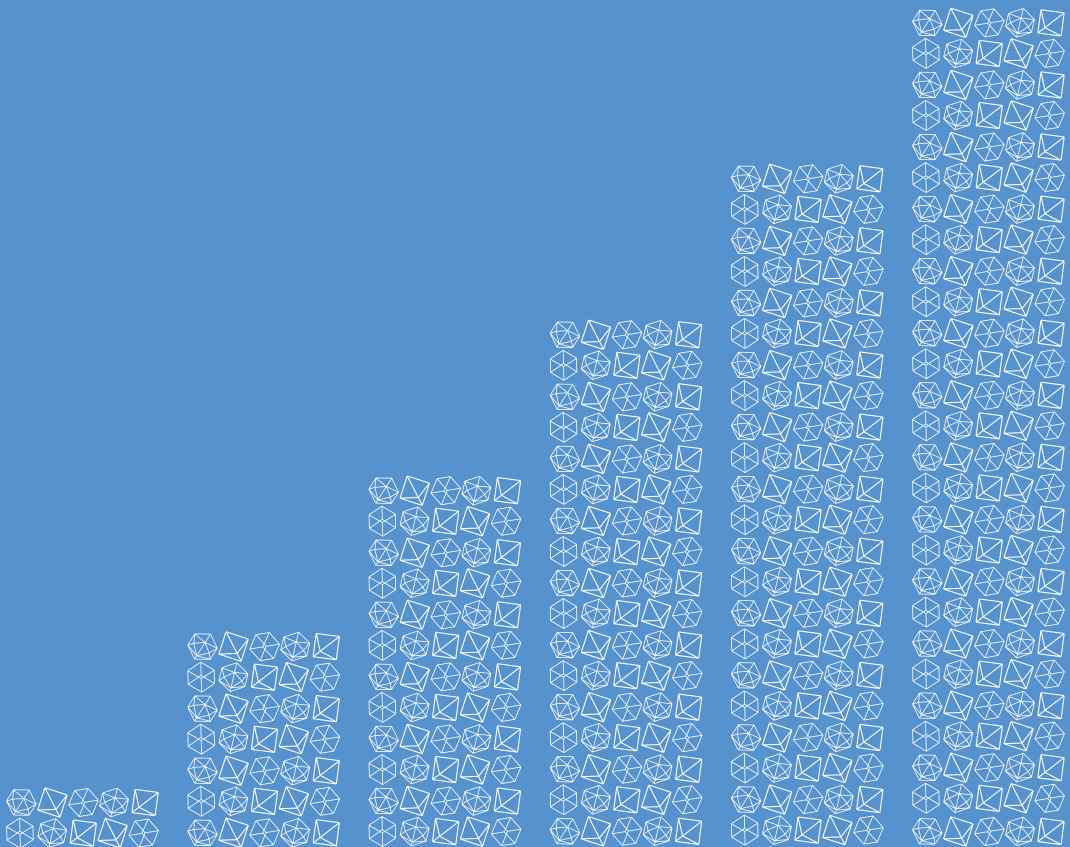
Alliance for Useful Evidence
October 2011

NESTA Making Innovation Flourish

Using Evidence to Improve Social Policy and Practice

Perspectives on how research and evidence can influence decision making

Edited by Ruth Puttick, with an introduction by Geoff Mulgan



Contents

Introduction

The idea that government policies should be guided by evidence is not new. When Florence Nightingale studied the effects of different treatments on death rates in military hospitals in the Crimea, she was not unusual. Many leading figures in the 19th century loved numbers, and believed that the application of scientific principles could largely solve such problems as ill-health and unemployment.

A century and a half later we know that they were both right and wrong. They were right that there are few areas of human activity that can't be enhanced with analysis, data and evidence. But we also know that the development and application of evidence is rarely straightforward, that the experts aren't always right, and that science can't replace values.

So the challenge is to get the right balance between the intelligent supply of evidence and sensitivity to the conditions in which it is used. That is the backdrop to the development of a new Alliance for Useful Evidence. In a context of intense pressure on public resources it's more important than ever to be sure that money is being directed to programmes and services with the best chance of achieving impact. Large sums of money have been invested in the generation of research, not just in the UK but across Europe. But surprisingly little of that evidence has been acted on.

The essays in this book give a good flavour of the work that is underway on the cutting edge of evidence and its uses. Each, in a different way, is trying to bridge the worlds of supply and demand. And each is trying to bring greater rigour, and challenge, to fields that may think they know more than they really do.

So what are the priorities?

First we need better evidence. In some fields, there's a good case for much more use of methods such as randomised control trials, first pioneered in agriculture before becoming mainstream in medicine. RCTs are by no means the panacea they are sometimes presented as. In medicine they are routinely overturned by new evidence, and they are particularly ill-suited to many fields of social policy. But they do bring a rigour and clarity that can be immensely useful, and alongside more effective use of survey evidence, natural experiments and other devices they can help to test assumptions.

Second, we need better mapping of what is already known, and better use of past public funding. In almost every field there is a virtue in more regular and systematic reviews of evidence – but current academic incentives do little to encourage this to happen. The work of figures such as John Hattie in education, who makes complex information useable, without losing nuance, is all too rare.

Third, we need to do more to make evidence quickly and easily available to decision makers. Google has played its part; global projects like The Cochrane Collaboration and The Campbell Collaboration have too; and initiatives like NHS Evidence which provides health evidence in digestible form to every doctor in the country could be replicated in other fields.

The fourth challenge, however, is to make it harder for evidence to be ignored. The UK's National Institute for Health and Clinical Excellence (NICE) is almost unique worldwide in linking formal evidence reviews to purchasing and commissioning decisions. Although many of its features are specific to health, the idea of developing a 'NICE' for other fields such as criminal justice and welfare is worth pursuing. The more that engagement with evidence can become part of the daily life of any department or agency the better. But, as with NICE, that requires a subtle job of intermediation and translation.

We are lucky to live in an era when evidence is easier to access than ever before. We are now able to access and manipulate data in ways that are likely to revolutionise the work of assessors and evaluators. And we are lucky to live in a connected world where, although an RCT or pilot in one country may not easily translate to another, it can provide a vital prompt. For a country like the UK this has meant a widening of horizons. The most impressive results will often be found in Scandinavia, Korea, or increasingly in middle-income countries, rather than just within our own borders or in other English speaking countries.

So what needs to be done? Our aim at NESTA is not to promote any particular method, but rather to work with others as an honest broker, raising the quality of both supply and demand. We need to ensure that evidence is commissioned and carried out in ways that make it more likely to be used and useful; and we need to work with the users of evidence to make it easier for them to act on what's known.

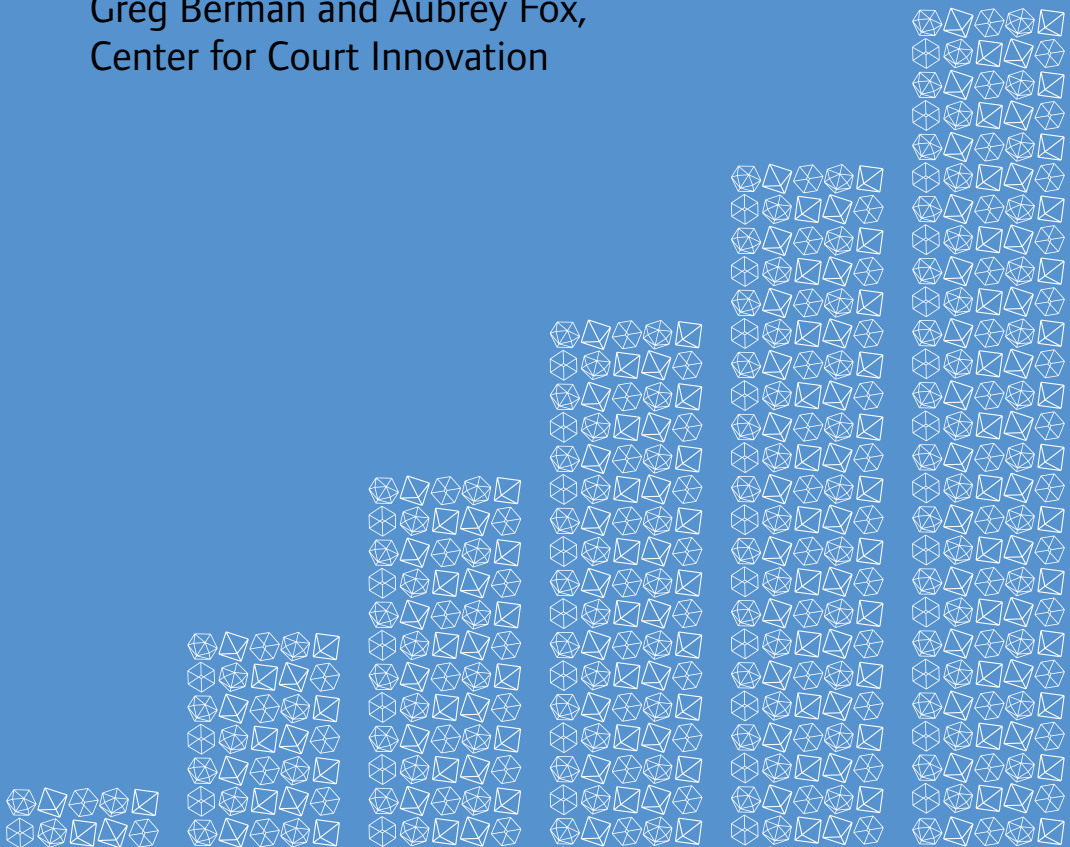
Even the best evidence is incomplete and imperfect, and it will remain legitimate for any government, and any members of the public to ignore evidence. But it's no longer legitimate for them to be ignorant of it.

Geoff Mulgan
Chief Executive, NESTA

Combining Research and Practice

The Center for Court Innovation and the Pursuit of Justice Reform

Greg Berman and Aubrey Fox,
Center for Court Innovation



1. Summary

The Center for Court Innovation is a non-governmental organization dedicated to reforming the justice system through demonstration projects, original research, training and technical assistance. Operated as a public/private partnership between the Fund for the City of New York and the State Court System, the Center for Court Innovation functions as the independent research and development arm of the New York courts – studying chronic problems, devising new solutions, and testing their feasibility. It then looks to aid reformers around the world, using its real world experience implementing concrete reforms to provide them with the tools they need to promote change locally.

In the most recent fiscal year (FY2010), the Center had a budget of $17.6 million, which was underwritten by a range of funders at the city, state, and federal level (87 per cent of the Center's revenues come from government grants, and 13 per cent come from private foundations and fee-for-service contracts). Broadly speaking, the Center's 175 full-time employees work in three principal areas: demonstration projects, research, and technical assistance.

2. Origins and philosophy

As an organization committed to data and analysis, the Center studies problems within communities and government systems. It uses this information to inform the development of demonstration projects that field-test new ideas. Then, based on its hands-on experience implementing real-life reforms, the Center provides assistance to innovators around the world. The Center has aided justice officials and non-government organizations in dozens of countries, helping them assess local problems, implement new solutions and evaluate their effectiveness.

The Center for Court Innovation's three primary areas of work are mutually reinforcing. Research is the foundation upon which demonstration projects are built. In turn, the Center's experience implementing demonstration projects is the basis of its expert assistance to the field. And the Center attempts to apply what it learns from its engagement with the world to its own demonstration projects.

The Center has had a hand in creating 21 different model programs that seek to address specific criminal justice concerns in new and more effective ways. These projects range from court-based projects that focus on domestic violence offenses, drugs and quality-of-life crime, to neighborhood-based programs that aim to reduce teen truancy and halt gun violence (see Box 1). It has also produced original research of international significance, including studies that have examined the effectiveness of drug treatment

Box 1: Demonstration projects

Through demonstration projects, the Center tests new approaches to improving public safety. While the Center's model projects cover a broad range of topics – from juvenile delinquency to the reentry of ex-offenders into society – the Center always relies on rigorous and collaborative planning, with an emphasis on using data to document results and ensure accountability. Evaluations have documented that these demonstration projects contribute to tangible results like safer streets, reduced levels of fear, and improved quality of life.

The Center for Court Innovation has created 21 demonstration projects:

- **Attendance Court**
  A truancy prevention program for students and their families.
- **Bronx Community Solutions**
  An initiative that seeks to apply a problem-solving approach to all non-violent cases in the Bronx Criminal Court.
- **Brooklyn Domestic Violence Court**
  Adjudicates all indicted domestic violence felonies in Brooklyn.
- **Brooklyn Mental Health Court**
  Links defendants with mental illness to long-term treatment in the community.
- **Brooklyn Treatment Court**
  New York City's first drug treatment court.
- **The Child and Adolescent Witness Support Program**
  Provides mental health support to children exposed to violent crime.
- **Crown Heights Community Mediation Center**
  Promotes cohesion in a Brooklyn community known for inter-ethnic conflict.
- **Harlem Community Justice Center**
  Solves neighborhood problems – including youth crime, landlord-tenant disputes and the challenges posed by ex-offenders returning to the community.
- **Integrated Domestic Violence Court**
  A 'one family/one judge' model that addresses related family issues such as child custody and civil protection orders.

- **Manhattan Family Treatment Court**
  Stabilizes families by linking substance-abusing parents or guardians to treatment.

- **Midtown Community Court**
  Targets quality-of-life offenses, such as prostitution, illegal vending, graffiti, shoplifting, farebeating, and vandalism.

- **Newark Community Solutions**
  An effort to re-engineer the local municipal courts' approach to minor crime.

- **New York Juvenile Justice Corps**
  An Americorps service program that puts participants to work offering career, clinical, and educational services to troubled young people.

- **NYC Community Cleanup**
  Presents low-level offenders with meaningful community service work.

- **Parole Reentry Court**
  Helps parolees transition from life in prison to responsible citizenship.

- **Queens Engagement Strategies for Teens (QUEST)**
  Provides after-school supervision and services to young people with delinquency cases.

- **Red Hook Community Justice Center**
  Seeks to improve public safety in Red Hook, Brooklyn through crime prevention initiatives and a problem-solving court.

- **Staten Island Youth Justice Center**
  Offers a peer-led youth court, case management, rigorous compliance monitoring, and after-school programming to troubled young people in Staten Island.

- **Youth Court**
  Trains teenagers to handle real-life cases involving their peers.

- **Youth Domestic Violence Court**
  Addresses misdemeanor domestic violence cases among teenagers.

- **Youth Justice Board**
  Brings together young people to study and propose solutions to public safety challenges.

as an alternative to incarceration, the impact of domestic violence offender intervention programs and the outcomes of intensive supervision of parolees. Additionally, the Center has shaped and aided the work of thousands of practitioners and policymakers around the globe through its consulting work and training initiatives.

The Center's start

The Center's first demonstration project was the Midtown Community Court. When it opened in 1993 in Manhattan, it was the first court of its kind, and it served as a novel response to the cycling of repeat offenders through the justice system who had committed quality-of-life crimes, including drug possession, prostitution, and petty theft.

These activities were wreaking havoc in the area in and around Times Square. Theaters were dark. Tourism was down. The neighborhood was losing population to the suburbs.

The Center began, as it always does, with research. A two-year needs assessment process yielded a wealth of valuable information. Among other things, planners documented that the two local police precincts had the highest volume of misdemeanor cases in the city. They also documented an array of problems with the standard judicial response to these cases, including an over-reliance on both short-term jail and sentences involving no punishment whatsoever (e.g. conditional discharges with no conditions).

The Midtown Community Court was created to respond to these problems. Located in the middle of a busy Midtown block on West 54th Street, the Midtown Community Court shares a building with a local non-profit theater company. Handling misdemeanor cases from the neighborhood, the court seeks to combine punishment and help. The community court judge is provided with an array of alternative sanctions, which include drug rehabilitation, community service, and mental health counseling. An on-site social service clinic provides case management and referrals to local service providers. Accountability is emphasized; failure to follow through with the court's orders results in a warrant for arrest and the prospect of jail time.

Research has confirmed that the community court, in conjunction with aggressive law enforcement and economic development efforts, helped to curb street crime. An independent evaluation by the National Center for State Courts noted that the Midtown Community Court, in conjunction with aggressive law enforcement and economic development efforts, resulted in a drop in prostitution arrests by 56 per cent and a reduction in illegal vending by 24 per cent.[1]

Other community court results include improved compliance with court orders and reductions in case processing time. In addition, approximately two out of three local

residents surveyed in a telephone poll said that they would be willing to pay additional taxes to support the community court.

The next step: Tackling addiction

The Center's next experiment was the Brooklyn Treatment Court. The first such court in New York City, the Brooklyn Treatment Court worked with more serious cases: felony offenders with long histories of addiction. Following a model originally established in Florida, participants were linked to long-term drug treatment in lieu of incarceration. Progress in treatment was regularly monitored by a judge using a system of sanctions and rewards.

The drug court model was a notable departure from the approach of traditional American courts. Instead of using incarceration as a default setting, the drug court sought to address the cause of criminal behavior through community-based treatment.

Researchers from the Center for Court Innovation evaluated the Brooklyn project and found significant reductions in re-offending. Over the course of three years, recidivism among Brooklyn Treatment Court participants was 27 per cent lower than offenders who went through conventional courts.[2] Based in no small part on these findings, the New York State Court System made an institutional commitment to spread the drug court model statewide.

The drug court model also attracted the attention of the executive and legislative branches of government in New York. In April 2009, the Governor of New York signed into law a significant revision of the infamous Rockefeller Drug Laws, long regarded as the toughest in the United States. Enacted in 1973, the Rockefeller laws established stringent mandatory minimum sentences for drug crimes – offenders convicted of possessing at least four ounces of narcotics, for example, automatically received a prison term that ranged from 15 years to a life sentence. By law, New York's judges were required to adhere to the Rockefeller sentence guidelines and had no discretion to propose shorter punishments or alternative sanctions, such as drug treatment.

In contrast, one of the explicit goals of the Rockefeller reforms – which the governor celebrated with an event at the Brooklyn Treatment Court – was to increase the number of defendants who participate in drug court. After nearly 40 years of the Rockefeller sentencing regime, the 2009 reforms signaled a sea change in New York's criminal justice policy. At the press conference announcing the reforms, then-Governor David Paterson noted that: *"drug abuse is an illness. We hope to forever eliminate the regime of the Rockefeller drug laws and replace it with a system that will give addicts and those who commit crimes the treatment that they need."*[3]

The demonstrated success of New York's drug courts provided legislators with concrete proof that a different approach to drug crime would work not just in theory but in practice. As Jonathan Lippman, New York's chief judge, would later recount: *"When Governor Paterson and the legislature reformed the Rockefeller Drug Laws in New York in 2009, they explicitly relied on the success of our drug courts."*[4]

The research findings on the efficacy of drug courts have also had a national impact. On the campaign trail, Barack Obama endorsed drug courts, making reference to the Center for Court Innovation's research in New York: *"Drug courts have proven successful in dealing with non-violent offenders. These courts offer a mix of treatment and sanctions, in lieu of traditional incarceration…The success of these programs has been dramatic: One New York study found that drug court graduates had a re-arrest rate that was on average 29 per cent lower than comparable offenders who had not participated in the drug court program. These programs are also far cheaper than incarceration."*[5] The Obama Administration has backed up this statement with concrete support: working with Congress in the most recent fiscal year, $44 million was appropriated to support drug courts nationally.

Problem-solving justice

While community courts and drug courts are the most prominent of the Center's projects, they are far from the only ones. The Center has also established New York's first mental health court, domestic violence court and reentry court, among other programs. While each of these projects is unique, they have come to be known collectively as 'problem-solving courts,' based on their efforts to address the underlying issues that bring defendants into the justice system (see Box 2).

Over the past 15 years, problem-solving courts have been widely replicated throughout New York State. Most of these projects have either been based on models created by the Center for Court Innovation or they have been created with the help of training and technical assistance from the Center's team of expert consultants. Currently, there is at least one problem-solving court in each of the state's 62 counties, including eight community courts, 192 drug courts, 41 domestic violence courts, and 26 mental health courts (see Box 3).

Underlying this rapid expansion is a critical shift in the perception of crime in the United States. After more than a generation of 'tough on crime' rhetoric and reform (e.g. mandatory minimums, three-strikes-and-you're-out legislation), the national conversation about public safety has shifted noticeably in recent years. Crime is down in many places across the country. Funding shortfalls have limited the ability of state and local governments to build more prisons. And the success of alternative-to-incarceration programs like drug courts have helped fuel a movement dedicated to rethinking the American justice system's reliance on incarceration.

Box 2: What is problem-solving justice?

Problem-solving justice aims to change offender behavior, enhance the safety of victims, and improve the quality of life of communities. Five principles animate problem-solving courts:

- **A tailored approach to justice**
  Problem-solving courts eschew a one-size-fits-all approach in favor of matching the judicial response to the specific needs of each case.

- **Creative partnerships**
  Problem-solving courts find new ways for citizens to get involved in the judicial process; they also integrate social services into the standard operating procedures of the court so that judges and attorneys can access a wider range of sentencing options.

- **Informed decision making**
  Problem-solving courts provide judges and attorneys with psychosocial information about defendants; they also offer legal professionals specialized training so they have a solid understanding of the underlying sociological dynamics of the cases that they handle.

- **Accountability**
  Problem-solving courts aggressively use judicial monitoring to supervise an offender's performance in social service programs and community restitution projects.

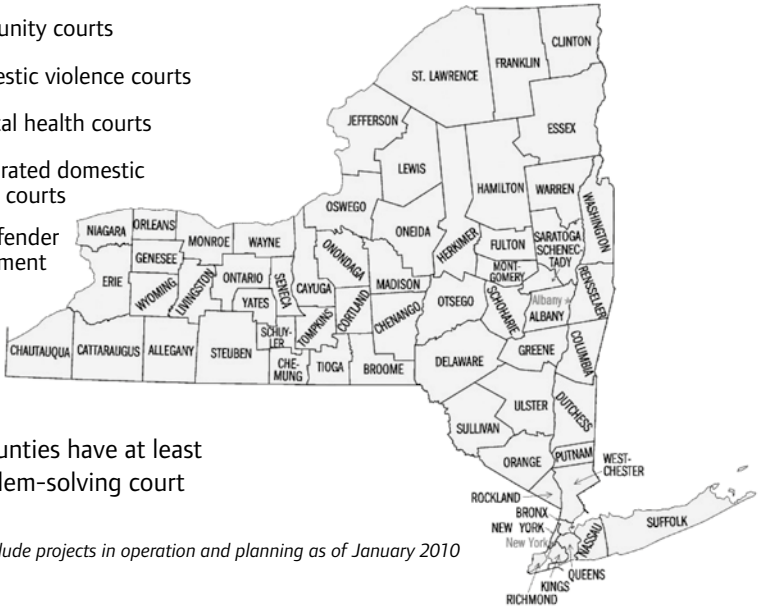
- **A focus on results**
  Problem-solving courts use data to assess their impact on victims, offenders, and communities.

  For more on the history, objectives, and achievements of problem-solving courts, see *Good Courts: The Case for Problem-Solving Justice* (The New Press, 2005).

International interest in problem-solving justice – and the Center for Court Innovation – is growing. Staffers from the Center have worked with criminal justice reformers in 50 countries. For example, the Center has aided the development of new community courts in England, South Africa, New Zealand, Australia and Canada. Australia has recently

Box 3: Problem-solving courts in New York State

- 192 drug courts
- 8 community courts
- 41 domestic violence courts
- 28 mental health courts
- 55 integrated domestic violence courts
- 8 sex offender management courts



All 62 counties have at least one problem-solving court

Numbers include projects in operation and planning as of January 2010

Source: Center for Court Innovation.

launched its own Center for Court Innovation, operating out of Monash University. And several organizations (including the Young Foundation and Policy Exchange) have recently issued calls for a Center for Court Innovation in the UK.

3. An evidence-based approach

Demonstration projects

The Center's primary business is the planning, implementation and operation of demonstration projects. It has been responsible for creating 21 different model projects that vary in size and focus. The Red Hook Community Justice Center, for instance, hears several thousand cases each year including landlord-tenant disputes, family matters

Box 4: Red Hook Community Justice Center

The Center is best known for creating the Red Hook Community Justice Center, a community court that seeks to improve public safety in a Brooklyn neighborhood notorious for high crime rates, urban blight, and social disorder.

Operating out of a refurbished Catholic school, the award-winning project addresses a range of neighborhood problems, including drugs, crime, domestic violence, and landlord-tenant disputes. Cases that would normally go to three different courts – civil, family, and criminal – are instead handled in the same courtroom by the same judge with the ultimate goal of offering a coordinated approach to neighborhood concerns.

The Red Hook Community Justice Center's judge, the Hon. Alex Calabrese, has an array of sanctions and services at his disposal, including community restitution projects, on-site educational workshops and GED classes, drug treatment, and mental health counseling. The Justice Center also serves as the hub for an array of programs aimed at engaging the community in crime prevention, including mediation, community service projects, and a youth court, where teenagers resolve cases involving their peers.

Each year, the court handles approximately 3,000 misdemeanor criminal cases, 11,000 summonses, 500 housing court cases, and 175 juvenile delinquency cases. Since it began operation, more than 75 per cent of Justice Center defendants have completed their community service mandates, compared to 50 per cent at comparable urban courts.

The Justice Center has also had an impact on local attitudes. In 1999, prior to the opening of the Justice Center, 77 per cent of Red Hook residents said they were afraid to go to the local parks or the subway. By 2004, that percentage had dropped to 43 per cent.

and criminal cases (see Box 4). By contrast, the Brooklyn Mental Health Court, which connects mentally-ill offenders with outpatient treatment, serves fewer than 100 defendants annually. Meanwhile, the Brooklyn Domestic Violence Court handles serious felony cases involving intimate abuse, while the Harlem Youth Court deals with minor offenses committed by juveniles.

Regardless of the topic or the size of the project in question, the Center's approach to developing a new program is the same. The Center begins with research, looking at a

given problem from as many different angles as possible. This includes examining both quantitative and qualitative information, using multiple methods (statistical analysis, focus groups, structured interviews, community surveys etc.).

Armed with data, the Center's next step is program planning – devising a solution to address the problem being studied. Almost always, this involves outreach to a wide variety of potential partners, including both traditional criminal justice players (police, judges, prosecutors, defense attorneys, probation officers, etc.) and other relevant agencies (drug treatment providers, health departments, block associations, civic groups, etc.).

In collaboration with its government partners, the Center takes responsibility for all aspects of program design. Depending on the project, this might include defining a target population, creating an implementation plan, training staff, drafting a memorandum of understanding, overseeing architectural design, and raising funds from private and public supporters. After a project opens, the Center helps to oversee operations and evaluate impact.

Research

Researchers from the Center for Court Innovation rigorously measure the work of its demonstration projects (see Box 5). The Center is committed to the idea of 'action research'. It employs a team of in-house researchers that monitor the impact that justice reforms have on street crime, substance abuse, sentencing practice, levels of neighborhood fear and public trust in justice. Action research is designed to provide immediate and useful feedback about everyday program operations so that those in charge can make midcourse adjustments as necessary. For example, the Center has used action research to determine whether the Brooklyn Treatment Court is meeting its volume targets, to learn more about the profile of drug court participants, and to generate a better understanding of the type of participant who tends to succeed under this model.

In addition to monitoring the day-to-day operations of demonstration projects, the Center conducts more formal, long-term evaluations to answer questions about the efficacy of a given reform. These studies are designed to have national and international policy implications. Examples of the Center's research work include:

- The Center has been a national leader in the study of drug courts. The Center's research on drug courts has explored not just whether they work but how and why they work. In 2003, the Center completed a statewide evaluation of New York's adult drug courts that demonstrated consistent and meaningful recidivism impacts across multiple sites. The Center is currently completing a national evaluation of drug courts in the United States funded by the National Institute of Justice (see Box 6).

Box 5: Results

Researchers have documented numerous positive results at the Center's demonstration projects, including:

- **Reducing recidivism**
  Participants in the Brooklyn Treatment Court re-offend at a rate that is 27 per cent lower than offenders who go through conventional courts.

- **Reducing crime**
  Independent evaluators from the National Center for State Courts documented that the Midtown Community Court cut prostitution by 56 per cent and reduced illegal vending by 24 per cent. The local police precinct where the Red Hook Community Justice Center is located is now the safest in Brooklyn.

- **Improving public trust in justice**
  The Red Hook Community Justice Center has a 94 per cent approval rating from local residents. Prior to the Justice Center's opening, only 12 per cent of local residents approved of local courts.

- **Changing sentencing practice**
  Bronx Community Solutions has cut the use of incarceration by one-third and doubled the use of community-based alternatives for misdemeanor offenders in the borough.

- **Repairing disorder**
  Each year, the Midtown Community Court and Red Hook Community Justice Center sentence thousands of low-level offenders to perform community restitution projects such as painting over graffiti, sweeping streets, and cleaning local parks. In total, these two community courts contribute 75,000 hours of community service to the surrounding neighborhoods each year, which adds up to more than $600,000 worth of labor.

- The Center has conducted a wide variety of studies examining how the justice system responds to domestic violence. This includes a randomized trial that tested the effectiveness of batterer programs and ongoing judicial monitoring with convicted domestic violence offenders in the Bronx (no impact was found). Other projects include a national survey of how courts respond to the noncompliance of offenders; statewide

Box 6: National drug court evaluation

The Center for Court Innovation, the Urban Institute and RTI International have jointly conducted the most comprehensive and long-term evaluation of drug courts to date. A five-year study that tracks defendants from 23 drug courts in seven states, the study documents that drug courts reduce recidivism and cut criminal justice expenditures for taxpayers. The study is also notable for examining *how* drug courts work and for *whom*. Among the research findings:

- **Drug use**
  The study found that fewer drug court participants reported and tested positive for drug use 18 months after enrolling in court-mandated treatment compared to traditional court defendants (56 per cent of drug court defendants reported drug use compared to 76 per cent of the comparison group).

- **Recidivism**
  Drug court participants also had lower rates of recidivism: 52 per cent of drug court offenders, compared with 62 per cent of traditional court defendants, were re-arrested within 24 months.

- **Cost benefits**
  A cost-benefit analysis confirmed that drug courts reduce overall criminal justice spending. The drug court model returns an estimated net benefit of $2 for every $1 spent.

Among the key components to a successful drug court identified by the study are the participants' attitudes toward the judge. When defendants feel positively about the judge, they have better outcomes. Furthermore, drug court clients who received higher levels of judicial praise, drug testing, and overall case management reported fewer crimes and fewer days of drug use.

research on integrated domestic violence courts in New York; and an evaluation of an experimental Youthful Offender Domestic Violence Court in Brooklyn.

- The Center examined the Harlem Reentry Court, a program that provides intensive monitoring and community-based services to parolees in their first six months post-release. The goal is to reduce both crime and incarceration. The report, entitled 'Do Reentry Courts Reduce Recidivism?' found mixed results. The Harlem Parole Reentry

Court produced a significant reduction in re-convictions, yet also led to increased technical violations thanks to a 'supervision effect'.

- In an effort to assess the impact of the Red Hook Community Justice Center on defendant perceptions of fairness, the Center conducted a survey of nearly 400 misdemeanor defendants who had their cases handled at either the Justice Center or a traditional, centralized criminal court. Structured courtroom observations supplemented the results of the survey. The Justice Center was considered to be more fair than the traditional court. In addition to offering a wider range of non-custodial sentences (including social and community services), respondents noted that the Red Hook Community Justice Center offered a more transparent and collaborative atmosphere for defendants. At Red Hook, 86 per cent agreed that their case was handled fairly by the court. This was true across the board, regardless of race, socioeconomic status or disposition of the case.

The Center's research findings have been published broadly in the mainstream media, professional periodicals and peer-reviewed academic journals. They have also been compiled in a book entitled *Documenting Results: Research on Problem-Solving Justice*.[6] Other books by Center authors include *Dispensing Justice Locally, Good Courts, Trial & Error in Criminal Justice Reform, Drug Courts: Personal Stories, A Problem-Solving Revolution* and *Daring to Fail*.

The Center's website[7] has become a hub of information for justice reformers around the world, attracting 90,000 unique visitors each month. On its website, the Center offers dozens of how-to manuals and best-practice guides for criminal justice officials written in an accessible, jargon-free style (more than 50,000 publications are downloaded each month). The Center's website is also home to multimedia presentations, including blogs, slideshows, short films and a monthly podcast, 'New Thinking', which features interviews with leading criminal justice thinkers and practitioners.

Technical assistance and training

The Center for Court Innovation provides hands-on, expert assistance to reformers – including judges, attorneys, probation officials, and community organizers – from around the world. Through its training and technical assistance programs, the Center offers guidance on assessing public safety problems and crafting workable, practical solutions.

Based on its first-hand experience implementing demonstration projects, the Center knows the nuts-and-bolts of getting a new project off the ground – from performing a rigorous community needs assessment to figuring out how to measure the impacts of new procedures. The Center is currently working with innovators in the United States and abroad to help create new responses to problems like drugs, domestic violence, delinquency, and neighborhood disorder.

The Center began providing technical assistance in 1996 when it received a grant from the US Department of Justice to assist a number of American cities with community court development. Over time, the Center has also won competitive grants to provide technical assistance on matters such as community prosecutions, domestic violence, drug court, technology, and institutionalizing problem-solving justice.

Each year, more than 650 visitors tour the Center's demonstration projects in New York City. These site visits are structured learning experiences that provide visitors with an opportunity to interact with their peers and see new ideas in action. Notable visitors include US Attorney General Janet Reno, US Supreme Court Justice Stephen Breyer, New York City Mayors Rudy Giuliani and Michael Bloomberg, and the Home Secretary, Lord Chief Justice and Attorney General of England and Wales.

But the typical visitor is not a dignitary or a high-ranking politician; the typical visitor is a local administrative judge or probation official or prosecutor or the head of a community-based organization concerned about a public safety problem in his or her community. By visiting one of the Center's model projects, these officials receive a hands-on education in how to implement new ideas. The goal is not to encourage wholesale replication of the Center's models, although that does happen quite a bit. Rather, the goal is to spark new thinking among visitors, encouraging them to adapt the Center's ideas to their local needs – and to dream up new variations. More than 65 per cent of those practitioners who visit the Center say that they intend to implement something they saw on their tour.

In addition to hosting site visits, the Center for Court Innovation's consulting group provides intensive technical assistance to reformers around the country and across the world. The Center provides intensive one-on-one assistance in the planning, implementation, and enhancement of justice reforms. Assistance is available in six main areas:

- **Needs assessment:** A needs assessment helps pinpoint pressing local problems, providing quantitative and qualitative data to sharpen planners' understanding of the issues at hand.

- **Concept paper:** Once the problems have been clearly defined, the Center helps local planners to create innovative solutions in the form of a detailed concept paper that spells out the scope and goals of a project.

- **Project development:** The Center helps clients identify funding sources, create a start-up budget, and devise new ways to engage the community and potential government partners.

Box 7: Technology

The Center promotes the adoption of innovative technology to support justice reform efforts. The community courts, for example, piloted an award-winning computer program that is used by courtroom staff to generate appropriate sentences, tailor individualized sanctions, and monitor defendant progress in community programs. Additionally, the technology application helps court staff track compliance rates, court appearance rates, and the types of referrals that have been made for the defendant.

The Center also harnesses technology to offer online training opportunities and resources. It has created a web-based learning program specifically for drug court professionals consisting of video presentations, a virtual site visit, and interviews with practitioners.

- **Technology:** Many of the Center's demonstration projects include innovative technology applications designed to improve case management, track participants and share information among partners. The Center's technology team provides help to justice system reformers, helping them analyze their technology team and adapt elements of the Center's management information systems (see Box 7).

- **Evaluation:** The Center helps planners evaluate the project, once it's up and running, by establishing performance measures and tracking the project's ability to meet its stated goals.

- **Troubleshooting:** Finally, the Center assists with troubleshooting, helping program managers analyze operations and make mid-course adjustments.

The Center for Court Innovation also provides customized workshops, panels, and trainings to criminal justice practitioners. For example, experts from the Center have trained hundreds of judges about how to incorporate problem-solving principles and practices on the bench. Other trainings include teaching prosecutors how to reach out to local residents, helping victim advocates work with the justice system, and educating community leaders about novel approaches to local public safety problems.

Speakers from the Center have been invited to participate in conferences, symposia and roundtables across the US and internationally. This includes lectures at leading colleges and graduate schools (e.g. Harvard University, Columbia University, Princeton University, University of Pennsylvania), appearances at gatherings convened by the major criminal

justice institutions in the US (e.g. Conference of Chief Justices, Bureau of Justice Assistance, American Society of Criminology, National Institute of Justice) and speeches at dozens of international events (including visits to China, South Africa, Australia, Japan, Georgia, Afghanistan, Mexico, Argentina and other countries).

4. Lessons learned

Key lessons from the Center for Court Innovation's multi-faceted approach to promoting evidence-based justice sector reform include the following.

Balancing independence and access
The Center for Court Innovation has sought to walk a fine line between working closely with government while remaining formally independent from it. Over the past 20 years, New York has had four governors (both Republican and Democrat), two chief judges, and three mayors of New York City (a Republican, an Independent, and a Democrat). All of these political officials have worked closely with the Center for Court Innovation – authorizing demonstration projects, providing access to crucial data, and making grants.

The Center's commitment to working in concert with – rather than in opposition to – government decision-makers helps to ensure the relevance of the organization's work. Advocacy organizations and academic institutions often run the risk of choosing topics of narrow interest that are unresponsive to the priorities of government.

At the same time, because the Center is not a formal part of the government, it enjoys a measure of insulation from the day-to-day politics of government. No organization is immune to political pressure of course, but the Center's independence grants it the freedom that's necessary to think beyond the next electoral cycle and to pursue a long-term vision of justice reform. The Center's independence from government also means it does not operate under some of the institutional constraints, such as civil service regulations or union rules, that often hamper efforts to create an entrepreneurial culture within government.

Finally, political independence provides the Center with the ability to issue findings that are less than positive. For example, the Center's randomized trial that examined the use of batterer's intervention programs in the Bronx found no evidence of impact on the behavior of offenders. Although this finding called into question a common practice by judges, the Center's study was not suppressed. Rather, it was featured in a front-page story in the New York Law Journal. The Center also conducted numerous behind-the-scenes briefings with court officials about the results, and these meetings ultimately led

the court system to issue a statewide memo with new instructions to local courts about how and when to utilize batterers' programs.

Combining research and practice
Locating researchers side-by-side with criminal justice practitioners within a single organization has had multiple benefits for the Center. First and foremost, it forces practitioners to think more rigorously. In particular, the presence of researchers encourages those who plan and implement the Center's demonstration projects to be more disciplined about articulating measurable goals and objectives for their work. On the other hand, researchers benefit from being co-located among practitioners because they become grounded in the messy realities of day-to-day implementation, which makes their work more nuanced and their writing less esoteric and easier to read. Researchers from the Center understand the challenges and realities of project implementation – and they know how to avoid holding new programs to unrealistic standards of performance.

Bridging the local and the national
The Center for Court Innovation has always had one foot in the world of local practice and one foot in the world of national policy. The Center's sustained engagement on the ground in New York has given it credibility and enabled it to build trust with local practitioners and policymakers. But the Center has a broader worldview than the typical local organization. Its national reach – and connections – means that it can bring ideas culled from across the country back to New York. For example, the Center recently adapted the Ceasefire anti-violence program, which has shown success in reducing gun crime in Chicago, to the Brooklyn neighborhood of Crown Heights.

Using multiple methods of analysis
The Center has conducted several randomized trials, the 'gold standard' in evaluation research, but the Center also understands that these studies are often unfeasible in the real world. Accordingly, most of the Center's program evaluations are quasi-experiments. The Center believes there is much to learn from other types of research, including qualitative studies, process evaluations, and ethnography. The Center is also committed to moving beyond a pass-fail approach to evaluating social programs (see Box 8). In criminal justice, this means that the Center's research tracks more than just a program's impact on crime rates and instead examines a much wider set of program outcomes, including impacts on system efficacy, public confidence in justice, and perceptions of fairness.

Box 8: Trial and error

To encourage criminal justice innovation, the Center is engaged in a multi-faceted policy inquiry designed to examine and capture the lessons that have been learned from criminal justice reform efforts of the past. The inquiry, which is being conducted with support from the US Department of Justice, has included roundtables, site visits, structured interviews, case studies, and literature reviews.

The ultimate goal is to encourage self-reflection and thoughtful risk-taking among criminal justice agencies. In 2010, the Urban Institute Press published *Trial & Error in Criminal Justice Reform: Learning from Failure*, a book based on the Center's study of criminal justice reform.

Acknowledgements

The authors would like to thank Bernice Yeung for her help in writing this article.

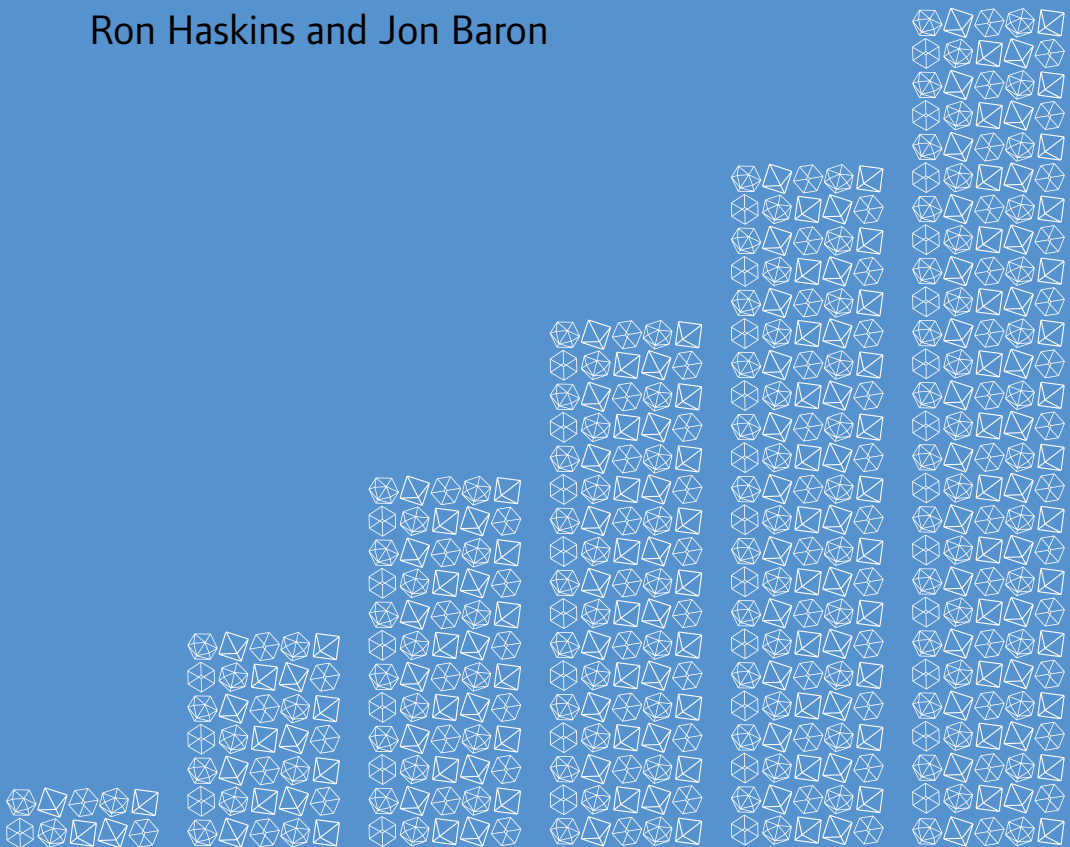
Endnotes

1. See Curtis, R., Ostrom, B., Rottman, D. and Sviridoff, M. (2000) 'Dispensing Justice Locally: The Implementation and Effects of the Midtown Community Court.' Newark: Harwood Academic Publishers.
2. Center for Court Innovation (2003) 'The New York State Adult Drug Court Evaluation: Policies, Participants and Impacts.' New York: Center for Court Innovation. Available at: http://www.courtinnovation.org/sites/default/files/drug_court_eval.pdf
3. See http://www.observer.com/2757/reform-rockefeller-drug-laws-for-real
4. See Lippman, J. (2010) How One State Reduced Both Crime and Incarceration. 'Hofstra Law Review.' Vol. 38, pp.1045-57. Available at: http://law.hofstra.edu/pdf/academics/journals/lawreview/lrv_issues_v38n04_bb1.lippman.final.pdf
5. See http://www.barackobama.com/pdf/issues/Fact_Sheet_Civil_Rights_and_Criminal_Justice_FINAL.pdf
6. Available at: http://www.amazon.com/Documenting-Results-Research-Problem-Solving-Justice/dp/0975950517/sr=11-1/qid=1171485763/ref=sr_11_1 /002-2041942-3400044
7. See www.courtinnovation.org

Building the Connection between Policy and Evidence

The Obama evidence-based initiatives

Ron Haskins and Jon Baron



1. Introduction

There is a growing belief in both the US and the UK that intervention programs addressed to domestic social problems can be greatly improved if policymakers and managers will support programs shown by scientific evidence to produce impacts. Since his inauguration in 2009, President Barack Obama and his administration have developed and are now implementing the most extensive evidence-based initiatives in US history. The purpose of this paper is to trace the evolution of these initiatives and to examine both their promise and problems.

Muddling through vs. rational policymaking

In 1971, Alice Rivlin published a seminal book on decision making entitled *Systematic Thinking for Social Action*.[1] She identified four 'propositions' that can be taken as a reasonable summary of the basic elements of what is often referred to as rational decision making. They are:

- Define the problem.
- Figure out who would be helped by a specific program attacking the problem and by how much.
- Systematically compare the benefits and costs of different possible programs.
- Figure out how to produce more effective social programs.[2]

Rivlin believed that at the time she was writing, economists, statisticians, and other analysts had made good progress on most of the steps in this approach to rational decision making, but that little progress had been made in determining the benefits of particular programs.

A much more skeptical view of the potential for rational, evidence-based policymaking can be seen in the classic 1959 article by Charles Lindblom on making decisions by *"muddling through."*[3] Lindblom argued that no program administrator could actually follow the rational decision making model because the demands on knowledge required to compare all alternative programs are too large, the effects of most programs are not known with any confidence, and not enough time is usually available to perform elaborate analyses before a decision must be made. Thus the choice set faced by managers is limited to incremental adjustments in current policy and practice, and the most important factor in policy choice is usually reaching consensus on a particular alternative. Lindblom argued that this process of what he called *"successive limited comparisons"* among alternatives not radically different from the status quo – or more

famously, *"muddling through"* – was both a better description of how policy actually is made and a more practical guide to action than the rational approach.

Our view is that the dichotomy between the rational decision making approach and the muddling through approach is a false one. Policymaking inevitably involves political constraints on choices as well as limitations on evidence and time. But that does not mean there is no evidence available, or that policymakers should ignore the evidence that does exist or fail to devote resources to obtain better evidence. Indeed, Rivlin argued that the case for *"systematic analysis"* was strong and had been well made, even by 1971, and that *"hardly anyone explicitly favors a return to muddling through."*[4] Rivlin also held that the key challenge is to recognize the limitations of analysis but to nonetheless employ a systematic approach whenever and wherever possible. Rivlin was especially forceful in calling for better evidence of program effects, perhaps the central feature of any systematic approach. Few would disagree that everyone from program managers to senior level policymakers could improve their decisions if they had reliable information about program impacts, or that developing programs with strong positive effects that can be widely replicated should be a fundamental objective in both policymaking and program evaluation.

Rivlin's propositions today

Updated to 2011, the Rivlin view of rational policymaking is still central to improving policy decisions. Ironically, the Rivlin proposition that now provides the strongest basis for expanding evidence-based policies is the dramatic expansion of high-quality evidence on programs that work (or not), the proposition that Rivlin thought the weakest in 1971. The most important contribution of social science to the public good is the use of scientific designs that allow definitive answers about whether specific intervention programs produce their intended impacts. Given this powerful tool, in a perfect world policymakers could follow a simple decision rule on program funding: if the program works, continue or even expand its funding; if it doesn't work, reduce or end its funding or find ways to improve it.

Evidence from scientific designs is now available for a large and growing set of interventions in early childhood education, K-12 reading and math, treatment of families that abuse or neglect their children, preparation of high school students to enter the world of work, community-based programs for juvenile delinquents and their families, several program models that reduce teen pregnancy, 'second chance' programs for children who have dropped out of school, prison release programs, and many others.

Broadening the evidence-based approach to achieve greater impacts in attacking society's social problems, government (and the private sector, especially foundations) can employ two approaches. First, as government provides money to establish new social programs, the money should be accompanied by a requirement that the specific

programs implemented at the local level be supported by strong evidence from scientific evaluations. Indeed, government might even specify a set of evidence-based programs that can be funded in order to avoid conflicts over what constitutes strong evidence. As we will see, the Obama administration has pioneered methods of identifying evidence-based programs and of ensuring that only evidence-based programs are implemented with government dollars.

Of course, anyone who has watched policymakers in action knows that they will rarely allow evidence on program effectiveness to be the sole or even major factor driving the policy process. Politicians focus on costs, the needs and desires of their constituents, the position of their party leaders, public opinion, their own political philosophy, pressure from lobbyists, the position favored by people and groups that finance their campaigns, and a host of other factors in making decisions about how to vote on program proposals. Allowing an adequate range for all these factors however, does not gainsay the possibility that in some circumstances evidence can have (and has had) a major impact on political decisions.

The second approach to employing evidence to improve social programs is to ensure that programs are implemented in a way that reliable information about program impacts is continuously generated. One of the Achilles heels of social programs is diminishing effectiveness as program models are implemented in more locations. A leading example of this problem is Head Start in the US. Over the past four decades, numerous preschool programs have shown that they can have both immediate and lasting impacts on children learning and other behaviors.[5] Yet a recent high-quality evaluation of Head Start, a program specifically designed to spread the benefits of preschool to a very large (enrollment in 2010: 900,000 children) group of disadvantaged children, produced only modest impacts that were barely detectable at the end of the first grade.[6] To combat the problem of diminishing impacts as programs are expanded to new sites, program operators must be vigilant in following the program model, perhaps adapted in some ways to local conditions. The key to replication of effective program models is continuous generation of evidence on program effects on participants and adjustments in implementation if the program is not achieving its expected effects. For this reason, enabling legislation should provide a mandate for continuous evaluation and the funding to make it possible.

2. President Obama's evidence-based initiatives

Origin of the initiatives

A remarkable aspect of the Obama evidence-based initiatives is that the President intended from the beginning of his administration to fund only programs that had solid

evidence of success. In the American system, the president has extraordinary power in agenda-setting. The president can raise the profile of and help to define specific problems, as well as give increased credibility to specific solutions. The president is often the most important single actor in the struggle to enact legislation. He can veto any legislation he doesn't like and because it takes a two-thirds majority in both Houses of Congress to over ride a presidential veto, it is rare that a president's veto, or the threat of a veto, does not spell death for the legislation. The president also has considerable power in enacting legislation. A president whose party controls Congress can almost always force consideration of favored proposals and usually guide them to enactment. Even presidents who face a Congress controlled by the other party can often get his favored issues at least considered, especially if he works to develop bipartisan relations with congressional leaders.

President Obama, encouraged and supported in highly creative fashion by the Office and Management and Budget (OMB) Director, Peter Orszag, intended to place major emphasis on making decisions about social programs based on evidence. OMB, located within the Executive Office of the President, is the lead Executive Branch agency responsible for development and execution of the President's budget and oversight of federal agency performance. As early as his inaugural address, the President made it clear that an important goal of his administration would be to expand programs that work and eliminate programs that don't.[7] From the earliest days of the administration, senior officials at OMB were planning several initiatives to advance the use of evidence-based program models and to generate high-quality evidence on new program models. When President Obama took office, career officials at OMB, who are often the source of ideas for increasing government efficiency, were already involved in a formal attempt to encourage federal agencies to conduct high-quality evaluations of their programs. Building on this effort which began in the Bush administration and even earlier, by the end of the second year of the Obama administration there were five initiatives well underway and one that was well formulated but stuck in the congressional enactment process.[8]

How the Obama administration is building the initiatives

Based on several interviews with members of the Obama administration and others inside and outside Congress knowledgeable about the Obama initiatives, we think the following outline captures the main steps of the President's evidence-based initiatives:

1. Select an important social problem that would make individual citizens and the nation better off if reduced in magnitude.

2. Identify model programs addressed to the problem that have been shown in randomized controlled trials or other rigorous research to significantly reduce the problem.

3. Obtain funds from Congress to scale-up evidence-based programs of this type that addresses the problem in accord with the proven models.

4. Make the funds available to government or private entities with a track record of good performance to replicate the successful model programs.

5. Continuously evaluate the projects as they are implemented to ensure they are faithfully implementing the model program and producing good results.

The Obama administration has now created a sweeping new opportunity for rigorous evidence to influence policy.[9] No president or budget director for a president have ever been so intent on using evidence to shape decisions about the funding of social programs as President Obama, former Budget Director Orszag, and other senior officials at OMB.[10] The Obama plan turns the normal relationship between policy decision making and use of evidence on its head. Instead of evidence being on the outside of the decision making process trying to get in, Obama brings evidence inside from the beginning. The administration must still convince others – especially those who hold the purse strings in Congress – that the use of evidence will improve policymaking and program outcomes. But those arguments are being made by people inside the administration to retain an evidence-based approach as a fundamental part of the President's legislative agenda rather than fighting from the outside to insert evidence-based policies into the process.[11]

Although less emphasized, the Obama plan for basing funding decisions on rigorous evidence can be useful for cutting spending as well as funding new programs. In the current age of fiscal austerity in both the US and the UK, when cuts in social programs are inevitable, it will be far better to cut programs that have minimal or no impacts than successful programs or programs that show promise. Yet experience shows that once enacted, federal programs become, as the economists say, 'sticky.' Thus, social programs are hard to get rid of, no matter how bad they are. The nation's social programs are unlikely to be improved until we learn to enact programs supported by rigorous evidence, to improve existing programs based on evidence, and to shut down failing ones, again based on evidence from high-quality program evaluations. Reliable evidence on program effects can be put to good use both in expanding and cutting programs.

Consider a recent example of how ineffective federal agencies in the US are in separating programs that don't work from those that do. The Deficit Reduction Act of 2005 established an Academic Competitiveness Council (ACC) to, among other goals, identify all federal programs with a science, technology, engineering, or math (STEM) education focus and determine whether these programs were effective. The ACC found that the federal government was operating 105 science, technology, engineering, and mathematics programs that spent well over $3 billion (in 2006). ACC leaders then

asked the federal agencies to send their most rigorous STEM program evaluations to the non-profit, non-partisan Coalition for Evidence-Based Policy, a Washington group that works with federal officials to advance evidence-based program reforms.[12] The Coalition independently reviewed these studies for the ACC, with funding from the W.T. Grant Foundation. The Coalition reviewed a total of 115 evaluations, and found only ten scientifically-rigorous impact evaluations (i.e., well-conducted randomized experiments or well-matched comparison-group studies) that had reported findings, as well as 15 others that were underway but had yet to report results. Six of the ten completed evaluations found that the STEM program had weak or no effects on educational outcomes. Thus, only four of the programs had been found by high-quality evaluations to be effective; most of the remainder had never been rigorously evaluated.

Rigorous evidence can also play a key role in deficit reduction in another important way. The solution to the US's long-term deficit problem is sometimes portrayed as a choice among sharp budget cuts, major tax increases, or a combination of the two. Given the magnitude of the problem, some level of sacrifice is unavoidable. But largely overlooked in the discussion are clear examples, from welfare and health care policy in the 1980s and 1990s, in which rigorous randomized experiments identified program reforms that produced important budgetary savings without adverse effects and, in some cases, with improvements in people's lives. Similar cost-saving opportunities already exist in a few areas, and many more could likely be identified through rigorous research.[13] In the 1980s and 90s, for example, federal officials sponsored many large randomized evaluations of state and local welfare reforms. These studies showed convincingly that certain reform models that emphasized moving participants quickly into the workforce through short-term job search assistance and training – as opposed to providing remedial education – produced large gains in employment and earnings, reductions in welfare, and net entitlement savings (in AFDC and Food Stamps) of $1,700-$6,000 per participant.[14] Such findings helped shape the 1996 federal welfare reform act and the major work-focused reforms in state and local welfare programs that followed.

Similarly, in 1995, federal officials launched a rigorous experimental evaluation to test prospective payment of Medicare home health agencies – i.e., paying such agencies an up-front lump sum per patient – against the usual cost-reimbursement approach. The evaluation found that prospective payment reduced costs to Medicare by 20 per cent over three years, compared to cost reimbursement, with no adverse effects on patient health.[15] This finding helped shape Medicare's nationwide implementation of prospective payment for home health agencies in 2000, producing large cost savings in this $15 billion program.[16]

Such examples illustrate how rigorous experimental studies can build the evidence needed for significant *and* smart spending reductions. But to identify enough of these

cost-saving strategies to produce sizable long-term deficit reduction, many more rigorous experiments testing a wide range of strategies are needed.

Issues raised by the Obama initiatives

Evaluating the impacts of federal spending on social programs encounters a special problem with federal funding that goes to states in the form of block grants or funding streams that provide great flexibility in how state and local governments use the funds. Two examples of such funding streams are the Title I education program that provides funds to states to help increase the educational achievement of poor children and the Temporary Assistance for Needy Families (TANF) program that gives states block grant funding to provide financial support to poor families, to decrease nonmarital births, and to promote work. The advantage of providing states with so much flexibility in their use of federal dollars is that they can better match their use of the money to local needs. But the problem is that so much flexibility makes it difficult to figure out how states are actually spending the money, let alone trying to evaluate the effectiveness of such spending. The Obama administration has addressed the problem of flexible spending by making the money in all their evidence-based initiatives contingent on both a clear statement of goals by grant recipients consistent with the goals stated in the federal notice of funding availability and by requiring exact specification of the model programs they will accept for funding.

Another problem with using an evidence-based approach applies to the case of using evidence to improve or eliminate programs shown to be failing or producing modest benefits. Even when ineffective programs have been identified, it does not follow that the administration or Congress will take action. According to Isabel Sawhill and Jon Baron, since 1990 there have been ten instances in which a large-scale federal social program was evaluated by a scientific research design. In nine of these ten cases (including Job Corps, Upward Bound, Even Start, the Job Training Partnership Act (JTPA), and Head Start), highly popular programs were shown to have modest or no impacts on their participants.[17] So far, these disappointing results have resulted in changes in only a few of the programs. The poor results for Head Start caused the Obama administration to use regulations to propose the most radical changes in Head Start's history.[18] The Even Start family literacy program was eliminated this year and funding for the JTPA youth programs was cut. But the other seven programs have continued without major changes.

These examples show that the federal government needs to find a better way to spend money on social intervention programs – and especially a way to overcome the political support that keeps program funds flowing to ineffective projects and interventions. The Obama initiatives for funding social programs are the most important attempts so far to find this better way and could potentially have a major impact on how social programs are funded by greatly elevating the role of program evaluations in program expansion or

contraction. Moreover, if the Obama initiatives are effective, the average impact of US social intervention programs on the well-being of children and families will increase and the nation will be better off. Then the taxpayer investment in evaluation research will be fully justified.

Development of the initiatives

As unlikely as it might seem, the story of the Obama initiatives begins with Robert Shea, a senior political appointee at OMB in the Bush administration.[19] Republicans have often been harsh critics of spending on social programs in part because they think that most of the programs don't work. Even so, Shea wanted to develop a system that OMB could use to get federal agencies to improve the effectiveness of their programs. After consulting with people inside and outside the administration, Shea and his colleagues at OMB developed the Program Assessment Rating Tool (PART).[20] One important feature of the PART, which OMB ordered agencies to follow in evaluating the effectiveness of their programs, was that it stated that well-conducted random assignment designs were the gold standard for program evaluation, while recognizing that such studies were not always feasible and suggesting second-best alternatives in these cases. This element of PART was based on input from the Coalition for Evidence-Based Policy and other evaluation experts. Not surprisingly, some agencies were not excited about being told to conduct rigorous impact evaluations of their programs, in part because they did not want to use their funds on expensive evaluations. Even so, the significance of the PART episode is that, even before the Obama administration took over at OMB, senior OMB officials – including career officials who would remain in place after President Obama and the new OMB team assumed office – had taken a firm stand about both the importance of program evaluation and the strongest methodology for conducting the evaluations. Moreover, OMB had put the federal administrative agencies that sponsored intervention programs on notice that it expected sustained attention to using evidence to improve programs.

At about the same time that Shea and the OMB professional staff were developing the PART program, David Olds, the designer of the Nurse-Family Partnership and a strong advocate of random-assignment program evaluation, came to OMB for a high-level meeting about his Nurse-Family Partnership program and about the Bush administration's use of social science evidence. The White House website had posted information about youth programs that it believed had been found by good evaluations to be successful, when in fact many of the evaluations listed were inferior. Olds wanted to convince the administration that it was inappropriate to classify well-conducted random-assignment evaluations with evaluations that used inferior methods, not least because the criteria followed in such classifications threw Olds' elegant work in with projects with modest or worse evidence. As a result of the Olds meeting, OMB created an interagency group to review evidence standards and compare the criteria for good evidence being used by various agencies on their websites featuring evidence-based

practices. Again, the agencies were put on notice by OMB that well-conducted random-assignment designs provided the most reliable evidence, although not the only evidence the administration was willing to accept.[21] A number of agencies then began changing their best practice websites to reflect OMB's hierarchy of study designs for assessing program effectiveness. But the most important outcome following the Olds meeting was that Shea and his staff decided the administration should use President Bush's 2008 budget to propose funding for expansion of the Olds program on the grounds that it would demonstrate the administration's (and OMB's) commitment to reward programs that had been shown to be effective in convincing random-assignment studies. Thus, the administration proposed and managed to get enacted a small $10 million pot of funds for states that agreed to use the money, augmented by their own funds, to mount Olds-type home-visiting programs.

During consideration of the Bush home-visiting legislation, the Coalition for Evidence-Based Policy provided input to Congress – as it had previously to the Academic Competitiveness Council and to OMB. Specifically, the Coalition urged the Appropriations Committees in the House and Senate – which are responsible for the annual funding legislation for the federal agencies – to include the administration's $10 million for home visiting in their respective appropriations bills, along with report language directing the implementing agency (HHS) to adhere to a high evidence standard. Coalition leaders had a long-standing relationship with staffers on the Senate Appropriations Committee and a chief staffer there already believed in the importance of rigorous evidence. The Senate Appropriations Committee included the proposed home-visiting funding and evidence standard in their appropriation legislation for 2009.

But the Appropriations Committee on the House side was a more difficult sell. The Coalition engaged extensively with a senior staffer with the Committee, who had heard that many home-visiting programs, in addition to the Olds program, were supported by high-quality evidence. This position had been put forward by advocates for various home-visiting programs that were being supported by state dollars and wanted their favored programs to be included in the $10 million federal funding for home-visiting programs. In a series of phone calls and e-mails, the Coalition provided input to the House Appropriations staffer on the specific studies and, in many cases, identified key flaws that likely biased the studies toward finding that a particular home-visiting program 'worked.' The staffer gradually was convinced that evidence from well-conducted random-assignment studies was necessary to know if a program was truly working. In the end, she recommended to the Appropriations subcommittee chairman Ralph Regula, that the $10 million initiative and the Senate (and OMB) language on giving priority to programs supported by scientific evidence be included in their appropriations bill.[22] The most important outcome of this episode is that the small $10 million appropriation for home visiting was enacted and was ready and waiting to be

Table 1: The Obama plan for expanding evidence-based programs

Stages in Programs	Teen Pregnancy Prevention	i3 (Investing in Innovation Fund)
Administering Agency	Health and Human Services	Education
Review of Literature	Completed by Mathematica	Although there was no formal review of literature, the i3 evidence tiers were based on a process for reviewing evidence developed by IES over several years primarily through its work on What Works Clearinghouse and strengthened
Amount of Awards	$100 million awarded – $75 million to replicate existing programs, $25 million to test new strategies	Up to $650 million across the three types of grants (development, validation, scale-up) **Note:** Applicants must obtain 20 per cent of their award amount in matching funds or in-kind donations
Review Panel Selection	Panels included both expert peer reviewers and federal staff	Over 330 peer reviewers were selected from 1,400 experts in both subject matter and research/evaluation; reviewers assigned to panels of 3-5 people
Selection of Proposals	75 applicants awarded grants to replicate existing programs; 27 grantees awarded grants to test new strategies	49 applications chosen as 'highest-rated.' All secured matching funds (20 per cent of grant amount)

	Home Visiting	Social Innovation Fund	TAA Community College and Career Training Program	Workforce Innovation Fund
	Health and Human Services	Corporation for National and Community Service	Departments of Labor and Education	Departments of Labor and Education
	Completed by Mathematica	None	None	None
	$88 million awarded in year 1, $1.5 billion over 5 years	$49.3 million awarded in FY 2010, $49.9 million appropriated for FY 2011 **Note:** Each federal dollar must be matched 1:1 at grantee and again at sub-grantee level	$500 million a year for 4 years	$125 million appropriated in FY 2011
	Applications were reviewed by grants management officials and program staff	A total of 60 experts were drawn from a pool of experts/ professionals and the CNCS reviewer database of 2,300 people; reviewers assigned to panels of 2-4 people	Not public yet. Technical review panels will evaluate all applications against evaluation criteria provided in application materials	Undetermined
	49 state governments, DC, and 5 territories applied and were awarded funding; second stage of funding to follow	11 grantees chosen for 2010; they have chosen 128 sub-grantees out of over 500 applicants	Not yet	Undetermined

picked up by the Obama administration when it took over and began to inaugurate its own evidence-based work.

Fortunately, the Obama team at OMB, led by Peter Orszag, Robert Gordon, and Jeffrey Liebman, came into office fully apprised of the value of random-assignment evaluations. Indeed, it is doubtful that any team of senior officials in OMB history was as knowledgeable about and committed to scientific program evaluation as the Obama team. With both a head start from the Bush administration and a new team of powerful OMB officials fully committed to the value of experimental evaluations, the Obama administration lost little time in launching its bold initiative to expand evidence-based social programs.[23]

3. Program areas addressed by the initiatives

The administration has now taken action to implement its evidence-based strategy in six specific areas of social intervention. What follows is an overview of the major characteristics and state of play for the administration's six evidence-based initiatives. Table 1 provides an overview and comparison of the major characteristics of the six initiatives.

Home visiting

Home visiting is a service strategy to help families, usually mothers, in one or more of three domains: maternal and child health, early childhood development, and family functioning.[24] Usually conducted by trained social workers or nurses, a number of home-visiting program models have been evaluated in random-assignment studies that estimate impacts on a variety of parenting behaviors and child and maternal outcomes. The home-visiting approach is often advertised as a way to reduce child abuse and neglect, but recent reviews have found that – with a few important exceptions such as the Olds program – many of the leading models produce weak or no lasting effects on these or other important outcomes, such as child cognitive development and family economic well-being.[25] As we have seen, home visiting differs from the other evidence-based initiatives in that federal funds had already been made available to evidence-based home-visiting programs during the last year of the Bush administration. The Bush funds were used to create cooperative agreements with 17 grantees that used the money to conduct planning efforts to coordinate existing federal, state, and local funds that could be used for home visiting.

The Obama home-visiting initiative, which has already been awarded $1.5 billion in guaranteed funding over the 2010-2014 period by Congress, greatly expands the initiative enacted during the Bush administration. In addition, the process of distributing

funds differs from the other Obama evidence-based initiatives. The funds are being distributed in three stages. In the first stage, which has been completed, all states were eligible for a share of funding if they submitted proposals that met administration requirements, primarily that they present a plan for conducting an assessment of the need for home-visiting programs in their state. Forty-nine states, the District of Columbia, and five territories were awarded funds to enter the second stage. In the second stage, states are required to complete their needs assessment and submit the results. States that successfully passed the first two stages were then eligible to update their state plans and receive grant funding to actually scale up their home-visiting program. In preparation for the third stage, and parallel in some respects to the Obama initiative on teen pregnancy (see page 12), the administration commissioned a literature review of home-visiting programs by Mathematica Policy Research.[26] Among other things, Mathematica found that seven home-visiting model programs met the minimum criteria set out in the program's authorizing legislation to achieve the status of evidence-based programs. Thus, states applying for the home-visiting funds in the third stage must spend at least 75 per cent of their funding on one of the seven evidence-based models, leaving up to 25 per cent of the funds to be used for 'promising' model programs. States are free to spend 100 per cent of their funds on one or more of the seven approved evidence-based models.

At this writing, the Department of Health and Human Services (HHS) is expected to soon issue the final solicitation for grants to initiate the third and final stage of the home-visiting program. But HHS outlined its proposed plan for this third stage in a July 2010 Federal Register notice. As noted above, Mathematica's review identified seven program models that meet the minimum evidence threshold set out in the program's authorizing statute. However, Mathematica's review, as well as the Coalition's review and a 2009 review published in the Lancet, found that most of these models – although perhaps meeting the minimum statutory threshold – produced weak or no lasting effects on key outcomes. A few models, such as Olds' program, were found to produce stronger, more durable effects.

Thus, in the third stage of the home-visiting program, HHS plans to allocate program funding that exceeds the first-year (2010) amount through a competition process, in which: *"HHS proposes to give significant weight to the strength of the available evidence of effectiveness of the model or models employed by the State. In this context, the use of program models satisfying the criteria outlined [for the program's initial grants] would be a minimal requirement, but HHS would consider additional criteria that further distinguish models with greater and lesser support in evidence."*

We strongly support this program structure, as it enables the program to evolve toward greater effectiveness over time, based on evidence about impact of the various models.

An interesting aspect of the home-visiting initiative is that the administration has already selected MDRC, one the nation's foremost firms conducting large-scale random-assignment studies, to evaluate the home-visiting programs.[27] The details of the evaluation plan have not yet been released, but the involvement of MDRC makes it clear that the administration intends to follow through on its repeated emphasis on basing their initiatives on evidence. In the case of home visiting, evidence played a central role in the selection of model programs and will again play a central role in the evaluation of the impacts of the state programs.

Teen pregnancy prevention

The teen pregnancy prevention initiative has proceeded mostly in accord with the components of the Obama model outlined above. Teen pregnancy is not only a serious national social problem with demonstrated impacts on the mother, the father, and the child, it is also an area of intervention that has a long track record of creative and diverse programs. A comprehensive review of programs by Douglas Kirby published in 2001 found eight program models that had what Kirby called *"strong evidence of success."*[28] There is also a comprehensive review of the evidence published by the Campbell Collaboration in 2006 that identified several successful evidence-based programs.[29] Thus, the first two components of our outline of the Obama approach to evidence-based initiatives – selecting a serious problem and ensuring that there are evidence-based model programs – have certainly been met in the case of teen pregnancy prevention. As in the home-visiting initiative, the administration commissioned a literature review from Mathematica that was made available to the public. The review identified 28 model programs that were supported by high-quality evidence. However, the review found that only two of these models are backed by well-conducted random-assignment studies showing a sustained effect on the most important measure – the actual reduction of teen pregnancies three to four years after random assignment. The other 26 models are backed by more preliminary evidence – in most cases, random-assignment studies or comparison-group studies showing only short-term effects on intermediate outcomes such as condom use and number of sexual partners, but not the final, most policy-relevant outcomes (pregnancies, births, sexually-transmitted diseases). When programs backed by such preliminary evidence are evaluated in more definitive random-assignment studies with longer-term follow-up, sometimes they are found to produce impacts on the long-term outcomes, but too often they are not. Fortunately, it appears that HHS plans to rigorously evaluate a number of the funded models to determine which are truly effective in preventing teen pregnancies.

Based in part on the Mathematica review, the administration issued its solicitation for grants in April 2010. Because there had been so much previous research in this field, the administration decided to award two tiers of funding. Tier 1, which would receive most of the money, would pay for program models identified in Mathematica's review as having higher-quality evidence of success. Tier 2 would be for programs that had some

evidence of success, but did not reach the higher standard reached by Tier 1 programs. The applications for funding were reviewed by a panel of experts based on review criteria published by the administration. Seventy-five projects were selected for Tier 1 funding of $75 million. In addition, $25 million was awarded to 27 Tier 2 projects that have some, but not strong, evidence of success. A notable feature of the teen pregnancy initiative is how the administration has made so much of the written material available to the public, including a detailed report from Mathematica on how it conducted the literature review, a list of the projects approved for funding that included extensive information about each project, and more.

The congressional journey of the teen pregnancy initiative demonstrates an unfortunate and perhaps fatal political obstacle faced by some or even all of the Obama evidence-based initiatives. Like the UK and many other rich nations, the US has been running a huge budget deficit that threatens to bankrupt the federal government.[30] The UK has recently responded to its deficit crisis with perhaps the most sweeping program of spending cuts and revenue increases in its history.[31] Congress and the executive branch in the US are now in the process of enacting a series of cuts in spending. The first important deficit-cutting action, although little more than a footnote to the extensive spending cuts and tax increases that will be required to make a serious dent in the nation's deficit, was a package of spending reductions as part of congressional action on the 2011 budget. One of the cuts enacted by House Republicans, who hold the majority, was a complete elimination of Obama's teen pregnancy initiative. In fact, House Republicans actually voted to zero out the initiative, but was subsequently prevented from doing so by the Senate. Even so, the $110 million initiative was trimmed to $105 million and Republicans seemed poised to try again to reduce or eliminate the initiative as part of action on the 2012 budget now being considered by Congress. There is little question that several of Obama's evidence-based initiatives will face further attacks as Congress and the President attempt to cut spending to reduce the federal deficit.

Investing in Innovation Fund (i3)

The i3 Fund and the Social Innovation Fund (SIF; see below) are very different from the home-visiting and teen pregnancy reduction initiatives in that both fund a more diffuse set of programs. In the case of i3, virtually any preschool or K-12 intervention with evidence of success or promise could receive funding. The i3 fund, like the teen pregnancy initiative, recognized multiple levels of evidence-based programs, in this case three levels. The top tier, called scale-up funds, was awarded for programs supported by evidence from rigorous evaluations. Funds from the second tier, called validation grants, were awarded to programs with some but less evidence of success. Finally, development grants were awarded to programs with a reasonable hypothesis. In order to qualify for funding, the programs had to aim to improve outcomes for pre-school children, help students qualify for or succeed in college, help students with disabilities or with limited-English proficiency, or serve schools in rural areas. The initiative is funded at $650

million. School systems, consortiums of school systems, or nonprofit agencies partnering with school systems were eligible to apply for funding. Awards were announced in August 2010 for all three categories of evidence-based programs. A total of 49 awards were made; all the projects managed to attract the required matching funds in order to receive the federal dollars and are now in various stages of implementation. Again, the administration made public a large number of documents about their grant-making process and about the projects receiving awards, including the comments of reviewers for the highest-rated i3 applicants and summary information and the applicant narratives for the highest-rated applicants for the scale-up, validation, and development awards. The administration has announced that random-assignment evaluations will be conducted for several of the biggest i3 grants.

Social Innovation Fund (SIF)

The President has said that solutions to America's domestic problems *"are being developed every day at the grass roots"*[32] and that his administration wants to support those grassroots efforts. SIF is one method by which the administration intends to *"identify and grow high-performing nonprofit organizations"* with experience at the local level.[33] The unique feature of the SIF is the mechanism of awarding funds. SIF funds are awarded in a two-stage process, first to intermediary organizations with a track record of funding successful local, community-based organizations; the intermediary organizations then select local organizations for funding. The intermediary and local organizations must raise matching equal to the value of the SIF grant. The intermediaries were organizations that had *"strong track records of identifying and growing high-performing nonprofit organizations."*[34] In July 2010, 11 such intermediaries were awarded $50 million in funding to go with another $74 million they had raised in matching funds to be distributed to nonprofit organizations. The nonprofits in turn were to use the money to conduct evidence-based programs addressed to at least one of three broad areas of social policy: economic opportunity, youth development and school support, and promoting healthy lifestyles and avoiding risky behavior.

The 11 intermediaries selected by the administration are reputable organizations with experience funding local programs, but whether they can make judgments about evidence-based programs is something that needs to be examined. In fact, the entire procedure of awarding funds to some organizations to in turn award funds to other organizations is an issue that bears careful study.

Like the teen pregnancy prevention initiative, the SIF initiative and even its federal administering agency (the Corporation for National and Community Service) seem to be in the bulls' eye for spending cuts by House Republicans. Funding for 2011 survived the budget scalpel, but future attacks should be expected.

Community College Challenge Fund

Several of our sources told us that the US Department of Labor had only a modest commitment to rigorous evaluation at the beginning of the Obama administration. Nonetheless, the administration was eventually successful in getting the Department to sign off on an evidence-based initiative to provide funds for training of displaced and unemployed workers and other young adults by the nation's community colleges. The $2 billion initiative – $500 million a year for four years – was enacted in 2010. On January 20, 2011 the Department, in conjunction with the Department of Education which will play a somewhat unspecified role in the grant program, released an announcement of the availability of funds for *"the development and improvement of postsecondary programs of two years or less that use evidence-based or innovative strategies to prepare students for successful careers in growing and emerging industries."*[35] An important characteristic of the grants is that community colleges and other entities receiving the funds are to experiment with existing employment and training materials in order to adapt them for use with young adults who seek employment. It will be interesting to review the basis for awarding the grant funds because there are few education, employment, or training programs for use at the community college level that have been rigorously tested and found to produce impacts on students. Thus, it appears that this initiative will focus on developing new curriculums and testing them with rigorous evaluation designs. The awards will be for between $2.5 million and $20 million to support projects employing strategies that have been shown to have *"strong or moderate evidence of positive impacts on education and/or employment outcomes."* Evaluation is a central feature of the Challenge Fund: 25 per cent of the assessment of proposals is based on their evaluation plan; all evaluations must include treatment and control groups; and the Department of Labor will select some grantees for rigorous evaluation using random assignment designs.

Workforce Innovation Fund

This initiative is also being run by the Department of Labor in conjunction with the Department of Education. Five per cent of the 2011 budgets of the Workforce Investment Act (WIA) Adult program and the WIA Dislocated Worker program were set aside to create this fund of nearly $108 million. The fund will be used to create competitive grants to states or localities to replicate proven practices in training, employment, and reemployment services, especially for vulnerable groups. Like the other evidence-based initiatives, the fund will also be used to test promising practices. As with community college training programs, there is a paucity of program models for employment and training programs with young adults that have been shown by rigorous evaluation designs to produce impacts on student learning, employment, and earnings. It is anticipated that funds will be focused on 'learn and earn,' apprenticeship, and on-the-job training programs, and that selected programs will be rigorously evaluated to determine their impact on key educational and workforce outcomes.

Federal program evaluation

In addition to these six initiatives, the administration also included money in the 2011 budget for program evaluation. Administration staffers claim that there are enough funds in the 2011 budget to pay for about 20 rigorous evaluations of *"the most promising new programs"* and to build the evaluation capacity of the various administrative departments. Indeed, the budget has well over $60 million for the Department of Labor alone to *"continue to pursue a robust, Department-wide evaluation agenda,"* including rigorous evaluations of WIA performance measures, effects of job counseling, use of administrative data in workforce programs, incentives for dislocated workers, and effects of Occupational Safety and Health Administration inspections.[36] In addition, the White House worked with the Department of Labor to create a Chief Evaluation Office that will manage the new evaluations and work with other components of the Department to assist them in conducting rigorous evaluations of their programs.

4. The promise of the Obama initiatives

These six evidence-based initiatives, plus the new funds for rigorous evaluation across the federal agencies, constitute the most sweeping and potentially groundbreaking emphasis on rigorous program evaluation ever conducted by the federal government. Although normal congressional politics played an important role in the formulation and enactment of the Obama evidence-based initiatives, the role of evidence in all six initiatives was more or less unprecedented. By devising several approaches to bring evidence to the center of policymaking in the federal government, both in obtaining funds to implement programs supported by rigorous evidence and in generating new evidence on program effects, the administration was able to achieve two benefits that are not often enjoyed by legislation enacted through a routine legislative process. The first benefit is that the Obama initiatives focus federal dollars on program models that have at least preliminary – and in some cases moderate or strong – evidence of impacts. It does not necessarily follow that the programs will actually produce impacts because program models backed by preliminary or moderate evidence too often turn out not to work when implemented at scale with a more definitive evaluation. But at least money will be spent on programs that have a good chance of having the desired effects. As compared with money now made available through many federal initiatives, this approach represents a great improvement.

The second advantage of the evidence-based approach is that the initiatives require rigorous evaluation of both program implementation and program impacts. At a minimum, the administration's evidence-based approach requires that a standard set of measures be reported to the federal agency responsible for the initiative. But far beyond reporting a standard set of measures, some of the initiatives require projects, as part of

their application, to submit a plan for evaluating their implementation and their impacts. Moreover, the administration is making it clear that the quality of the evaluation plan will be a major criterion for deciding which applications to fund. Again, as compared with the scores of federal programs that have tepid or no evaluation requirements, the emphasis on rigorous evaluation evidence in the Obama initiatives may set a precedent for future legislation. And even beyond these two requirements on evidence, at least two of the initiatives and possibly more are requiring projects applying for funds to submit to random-assignment evaluations performed by third party evaluators. The administration, for example, has already hired a crack evaluation organization to help develop and then carry out random-assignment evaluations of the home-visiting programs. It is not yet know how many programs will be evaluated or which ones, but the use of rigorous third-party evaluations shows the lengths to which the administration is going to send a message to federal and state agency officials and program operators that a strong evaluation plan is the new normal in federal funding of social programs.

5. Problems lurk and opportunities abound

Proponents of expanding the role of rigorous evidence in policy choice, not least the two of us, are optimistic about the Obama initiatives. However, it would be a serious mistake to accept the initiatives uncritically and to assume that they will greatly improve the quality and impact of the nation's social programs. For this reason, we turn now to a discussion of several potential problems and issues with the Obama approach, each of which is dealt with in separate sections. In addition, we outline several opportunities to strengthen the evidence-based approach to policymaking.

Politics

Those of us who are hopeful that an enhanced role for evidence will greatly improve the quality and impacts of social programs have sometimes been thought to regard the emphasis on evidence as the repeal of normal politics. Let the experts decide, based on their evidence, what programs should be funded and at what level. But that is not at all the way the two of us view evidence-based policymaking in general and the Obama initiatives in particular. For many years, our major goal has been to increase the role of evidence in political decision making. We are both fully aware, however, that constituent views, the positions taken by party leaders, the political philosophy of elected officials, the positions politicians have taken on similar issues in the past, their campaign promises, and the inevitable political compromises necessary to pass legislation will always play a huge role in political decision making. Against this backdrop of politics as normal, which almost always places political considerations above the usually limited role of evidence, expanding the influence of evidence would be useful and could lead to better decisions. Even so we have no doubt and do not regret the fact that politics

will almost always play a determining role in policy choice. We want evidence to be important, not dominant.

Implementation

The field of implementation research is not as advanced as the field of evaluation research. Consider the example of Head Start, outlined earlier. Since at least the late 1950s, rigorous research has shown that high-quality preschool programs can have major impacts on the development of poor children.[37] Random-assignment longitudinal studies show that children who have attended quality preschools, as compared with similar peers who stayed at home or attended regular day care facilities, perform better on standardized tests, have fewer placements in special education, are less likely to be arrested as juveniles or young adults, are more likely to graduate from high school, and so forth.[38] Yet after nearly half a century, a recent rigorous, multi-site study of Head Start, a broad-scale implementation of preschool intended to boost the development and school performance of poor children, showed that its impacts were barely detectable at the end of first grade (i.e., three to four years after random assignment).[39] This seems to be the story of many intervention programs: significant impacts when tested by their designers and on a small scale but modest or no impacts when replicated (often in a much different, more diluted form) on a broader scale.[40]

The lesson here is that implementation is at least as important as program development. If the Obama initiative results in a significant share of federal intervention dollars being spent on programs that have proven track records, that will stand as a great achievement. But experience teaches that poor implementation of good models often fails to produce impacts. Thus, it is to be hoped that the various teams in the agencies implementing the Obama initiatives will focus like a laser on implementation issues. These issues include how to ensure that staff receive adequate training in the intervention approach, how to ensure that staff continue to follow the major features of the intervention program, how to achieve an adequate dosage of the intervention with all program participants, and how to continually monitor and evaluate outcomes. Methods for achieving these and other essential features of quality implementation could be one of the most important outcomes of the Obama initiatives. We are at last focused on evidence of program impacts; now we need to be equally focused on program implementation methods to produce those impacts on a broad scale.

Selecting strong models for scale-up

Some of the Obama initiatives have, by design, selected models for which the supporting evidence is only moderate or preliminary in nature, with the goal of testing them in more rigorous evaluations to determine whether they work. Moderate or preliminary evidence includes nonrandomized comparison-group studies ('quasi-experiments'), or randomized controlled trials with only short-term follow-up, assessment of intermediate rather than final outcomes (e.g., condom use versus reductions in teen pregnancies, abortions,

or births), or other key limitations in study design or implementation. These studies can be valuable for decision making in the absence of stronger evidence. Too often, however, findings from such initial studies are overturned in large, definitive randomized controlled trials. Reviews in medicine, for example, have found that 50 per cent to 80 per cent of promising results from quasi-experiments or preliminary 'efficacy' trials are overturned in subsequent randomized controlled trials.[41] Similarly, in education, nine of the ten major randomized controlled trials sponsored by the Institute of Education Sciences since its creation in 2002 have found weak or no positive effects for the interventions being evaluated – interventions which, in many cases, were based on promising quasi-experiments or preliminary trials (e.g., the LETRS teacher professional development program for reading instruction).[42] Systematic 'design replication' studies comparing large, well-conducted randomized controlled trials with quasi-experiments in welfare, employment, and education policy also have found that many widely-used and accepted quasi-experimental methods produce unreliable estimates of program impact.[43]

That some of the Obama initiatives – including teen pregnancy reductions and home visiting – may fund program models with only moderate evidence of effectiveness leads us to fear that a number of the models, when rigorously evaluated, will be found to produce weak impacts. But at least a few will likely be found to produce meaningful impacts on important policy outcomes. If the administration can use strong evaluation designs to weed out the program models that produce weak impacts in the initial program years, and reinvest the money saved from these models in models with stronger evidence of success, the Obama initiatives may evolve toward increasing effectiveness over time.

Changing the agency culture

As the episode on the Science, Technology, Engineering, and Math (STEM) programs above demonstrates, federal administrative agencies have often failed to emphasize the importance of evidence in determining whether their programs are producing worthwhile results. Indeed, some agencies appear to be conflicted about evaluation because so many programs that are subjected to rigorous evaluation are shown to produce null results. No agency wants to be administering programs that are known to be ineffective. Moreover, a major rule in the federal government is that power and influence are based on big budgets. If agencies want to expand their programs and their budget, they need Congress to believe they are conducting effective programs that are providing benefits to the nation. It is a good bet that if every type of federal program received the scrutiny that OMB and the Coalition for Evidence-Based Policy provided for the STEM programs, many federal programs would be exposed as not better than moderately successful or even failures.

A vital part of the Obama evidence-based initiatives is to change the relationship between OMB and the federal agencies so that OMB is a taskmaster for evaluating

programs by the use of rigorous evidence. Given the history of the Obama evidence-based initiatives reviewed above, it appears that OMB has already become a strong supporter of rigorous evaluation and is now providing strong leadership – not to mention a source of funding – for federal agencies to evaluate their programs using rigorous methods. This is the role of OMB envisioned by the Bush administration's PART initiative. The modest role of PART in the Bush administration has been expanded and, in a different form, PART's approach to rigorous evaluation now seems to be having an impact on a number of federal agencies that administer intervention programs. In the long run, if federal agencies do not become true believers in rigorous evaluation and transparency about the effects of their programs, the impacts of evidence-based policy making will be minimal.

The exploding deficit

Like the UK, the US is now running huge federal deficits – and as the retirement of the baby boom generation gets into full swing in the years ahead and health care costs continue their relentless rise the deficit will get even worse. According to a realistic baseline of US spending over the next decade,[44] the deficit will average a trillion dollars a year. One consequence of such profligacy is that by 2020, interest payments on the federal debt will approach $1 trillion. Although budget hawks have been sounding the alarm for nearly a decade,[45] Congress and the President are just now beginning to get serious about cutting spending and raising revenues. The first major confrontation between the political parties occurred in the context of the 2011 budget. As part of its package of spending cuts, the House enacted legislation that would have zeroed out the Obama teen pregnancy prevention initiative and the agency responsible for administering the Social Innovation Fund initiative. Both were eventually restored, but more of the same lies ahead. Republicans are now talking about cutting trillions in spending – and in truth, even if revenues are part of a compromise deal, it seems likely that spending would be cut by $2 trillion or more over ten years. This level of cutting would pose great danger for the Obama initiatives. Even initiatives that are already being implemented, such as the teen pregnancy prevention initiative, are subject to cutting or even termination. In the face of such a major deficit-reduction effort, the argument that evidence can improve social programs tends to lose its force.

There is not much that can be done about the precarious funding of Obama's six evidence-based initiatives. Miraculously, they all survived the 2011 budget fight. The administration may come to the moment when program triage is necessary. As part of a deal that involves big cuts in spending, perhaps the administration can protect three or four of the initiatives while sacrificing the others. In any case, it seems wise to us for the administration to begin figuring out a fallback position on protecting as many of the evidence-based initiatives as possible from the tsunami-like forces moving Congress toward deep spending cuts. Even when spending is under the knife, as we have seen, evidence can prove useful.

6. Conclusion

The federal government and state governments in the US spend tens of billions of dollars each year on social programs that have been shown to produce modest results or worse. In other cases, billions of dollars have been spent on programs and funding streams for many years, and yet little is known from rigorous evidence about whether the programs are producing good outcomes. Meanwhile, social scientists in the US have developed increasingly sophisticated and reliable methods for evaluating program impacts, and the nation's universities and large and widely respected research companies have achieved significant experience in designing and carrying out large-scale, multi-site evaluations using random assignment. Program administrators, especially in the federal Office of Management and Budget – but in other federal agencies as well – have increasingly emphasized the importance of obtaining rigorous evidence about the impacts of their programs. To a limited but probably growing degree, the federal Congress has even required administering agencies to fund rigorous evaluations of their programs.[46]

But the Obama evidence-based initiatives analyzed here are opening a new chapter in the generation and use of evidence by the federal government. Subsequent administrations may not have the same commitment to evidence that senior officials in the Obama administration have, but to a considerable degree the Obama emphasis on evidence is now being institutionalized in the federal agencies. At least three cabinet-level agencies (HHS, Education, and Labor) as well as the Corporation for National and Community Service are now administering multi-year evidence-based initiatives. Of greatest important, OMB has become a strong advocate for generation and use of rigorous evidence and, beginning at least with the PART initiative during the Bush administration, has been leading all the federal departments that administer social intervention programs to generate and use rigorous evidence.

The existence of the six Obama initiatives, the actions of OMB over at least a four or five year period, the increasing use of evidence by federal agencies, and the approval and funding of evidence-based initiatives by Congress lead us to believe that the role of rigorous evidence in federal policymaking and program implementation is here to stay.

Endnotes

1. Rivlin, A. (1971) 'Systematic Thinking for Social Action.' Washington, DC: Brookings.
2. Rivlin, pp. 6-8.
3. Lindblom, C.E. (1959) The Science of 'Muddling Through'. 'Public Administration Review.' 19, (2): 79-88.
4. Rivlin, p. 3.
5. Ramey, C., Campbell, F. and Blair, C. (1998) Enhancing the Life Course for High-Risk Children: Results from the Abecedarian Project. In 'Social Programs that Work.' Ed., Jonathan Crane. New York: Russell Sage Foundation; Schweinhart, L.J. and others (2005) 'Lifetime Effects: The High/Scope Perry Preschool Study through Age 40.' Ypsilanti, MI: High/Scope Press; Reynolds, A.J. (2000) 'Success in Early Intervention: The Chicago Child-Parent Centers.' Lincoln, NE: University of Nebraska Press; Barnett, W.S. and others (2007) 'Effects of Five State Pre-Kindergarten Programs on Early Learning.' Rutgers University: National Institute for Early Education Research; Gormley, W.T. Jr., Phillips, D. and Gayer, T. (2008) Preschool Programs Can Boost School Readiness. 'Science' 320: 1723-1724.
6. Puma, M. and others (2010) 'Head Start Impact Study: Final Report.' Report prepared for the Office of Planning, Research and Evaluation Administration for Children and Families, US Department of Health and Human Services. Rockville, MD: Westat.
7. Here is what President Obama said in his inaugural address: "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."
8. Funds for an initiative called the Workforce Innovation Fund, which is a mix of systems reforms of the Department of Labor's employment and training programs and funds for high-quality evaluations of employment and training model programs, was in the Obama 2011 budget. However, when Congress was unable to pass the original 2011 appropriations bill, all new discretionary funding (with a few very modest exceptions) was suspended. Then, surprisingly, Republicans in the House allowed the provision to be funded in the final compromise on the 2011 appropriations bill.
9. Carol Weiss has recently called this "imposed" use of research. See Weiss, C.H., Murphy-Graham, E. and Birkeland, S. (2005) An Alternate Route to Policy Influence: How Evaluations affect D.A.R.E. 'American Journal of Evaluation.' 26. (4): 12-30.
10. Budget Director Peter Orszag, after a number of speeches in which he talked about how to identify good and bad programs, spelled out his and the President's thinking on using high-quality program evaluations to improve program quality in a blog post and a memo to the heads of all executive branch departments and agencies. See Orszag, P.R. 'Building Rigorous Evidence to Drive Policy.' Office of Management and Budget. Blog, June 8, 2009. Available at: http://www.whitehouse.gov/omb/blog/09/06/08/BuildingRigorousEvidenceto DrivePolicy and Orszag, P.R. (2009) 'Memorandum for the Heads of Executive Departments and Agencies.' Executive Office of the President, Office of Management and Budget. Available at: http://www.whitehouse.gov/omb/assets/memoranda_2010/m10-01.pdf
11. It would be possible for major players in the legislative branch on either committees with jurisdiction over particular social programs (either the authorizing committee or the appropriations committee) to try to enact evidence-based policies. Even the most powerful committee chairman, however, generally has less power and influence than the president. Nonetheless, powerful Congressional players could use an approach like the Obama administration is using to influence the adoption of evidence-based policies.
12. By way of full disclosure, Baron is the President of the Coalition and Haskins is on their Advisory Board.
13. See editorial by Solow, R. and Baron, J. (2010) Long-Term Deficit Reduction: Less Pain, More Gain. 'The Fiscal Times.' August 12, 2010. Available at http://coalition4evidence.org/wordpress/wp-content/uploads/Solow-Baron-Op-Ed-TheFiscalTimes-8-12-10.pdf
14. These are 2011 dollars. Examples include: (i) the Riverside Greater Avenues for Independence (GAIN) Program (Freedman, S. and others (1996) The GAIN Evaluation: Five-Year Impacts on Employment, Earnings, and AFDC Receipt. Working Paper 96. MDRC; Riccio, R., Friedlander, D. and Freedman S. (1994) GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program, MDRC); (ii) Los Angeles Jobs-First GAIN (Freedman, S. and others (2000) The Los Angeles Jobs-First GAIN Evaluation: Final Report on a Work First Program in a Major Urban Center, MDRC); and (iii) Portland Job Opportunities and Basic Skills (JOBS) Training Program (Scrivener, S. and others. (1998) National Evaluation of Welfare-to-Work Strategies: Implementation, Participation Patterns, Costs, and Two-Year Impacts of the Portland (Oregon) Welfare-to-Work Program, MDRC; Hamilton, G. and others (2001) National Evaluation of Welfare-to-Work Strategies: How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs, MDRC and Child Trends).

15. Cheh, V. (2001) 'The Final Evaluation Report on the National Home Health Prospective Payment Demonstration: Agencies Reduce Visits While Preserving Quality.' Report submitted by Mathematica Policy Research, Inc. to the Health Care Financing Administration.

16. US General Accounting Office. (2000) Medicare Home Health Care: Prospective Payment System Will Need Refinement as Data Become Available. GAO/HEHS-00-9, April.

17. Sawhill, I.V. and Baron, J. (2010) Federal Programs for Youth: More of the Same Won't Work. 'Youth Today.' May 2010, p. 21.

18. Haskins, R. and Barnett, W.S. (2010) A Test for Head Start. 'The Washington Post.' October 11, 2010; Guernsey, L. (2010) 'Proposed Rules Will Shake Up Head Start.' Early Ed Watch. Blog. New America Foundation, Early Education Initiative, September 23, 2010. Available at: http://earlyed.newamerica.net/blogposts/2010/proposed_rules_will_shake_up_head_start-37234

19. This account is based on the documents that will be cited below and interviews with Grover 'Russ' Whitehurst, former head of the Institute of Education Sciences; Robert Shea, the lead for OMB in developing the Program Assessment Rating Tool (PART) during the Bush administration; David Olds, the developer of the Nurse-Family Partnership who played an important role in the Bush initiative on evidence-based policy; Robert Gordon, the OMB senior staffer for Obama who played a central role in the Obama evidence-based initiative; and Kathy Stack, a senior OMB career official, who has been at the center of emphasizing evidence-based policy during both the Bush and Obama administration.

20. ExpectMore.Gov, 'The Program Assessment Rating Tool (PART).' Available at: http://www.whitehouse.gov/omb/expectmore/part.html

21. The memo OMB sent to the executive agencies incorporated key concepts suggested by the Coalition for Evidence-Based Policy, regarding which study designs are most likely to yield valid estimates of a program's impact. The Coalition also did a series of workshops with OMB and agency staff on this and related evidence-based policy concepts.

22. Here is the exact language from the House-Senate conference agreement on funds for home visitation: "Within the amount provided for State grants, the conferees include $10,000,000 for a home visitation initiative to support competitive grants to States to encourage investment of existing funding streams into evidence-based home visitation models. The conferees expect that the Administration for Children and Families will ensure that States use the funds to support models that have been shown, in well-designed randomized controlled trials, to produce sizeable, sustained effects on important child outcomes such as abuse and neglect. The conferees also recommend that the funds support activities to assist a range of home visitation programs to replicate the techniques that have met the high evidentiary standards. In carrying out this new initiative, the conferees instruct the Department to adhere closely to evidence-based models of home visitation and not to incorporate any additional initiatives that have not met these high evidentiary standards or might otherwise dilute the emphasis on home visitation." See Conference Report on H.R.3043, Departments of Labor, Health and Human Services, and Education, and Related Agencies Appropriations Act 2008, p. H12486.

23. Orszag's take on the importance of evidence was succinctly summarized in a now-famous blog post that appeared in the midst of the administration's various evidence-based initiatives; see Orszag, P.R. (2009) 'Building Rigorous Evidence to Drive Policy.' Available at: http://www.whitehouse.gov/omb/blog/09/06/08/BuildingRigorousEvidence toDrivePolicy/

24. Stoltzfus, E. and Lynch, K.E. (2009) 'Home Visitation for Families with Young Children.' (R40705), Washington, DC: Congressional Research Service, October 2009.

25. Howard, K.S. and Brooks-Gunn, J. (2009) The Role of Home-Visiting Programs in Preventing Child Abuse and Neglect. 'The Future of Children.' 19(2): 119-146; see especially Table 2: The Effects of Home-Visiting Programs on Child Abuse, Health, Parenting, and Depression, p. 133; MacMillan, H.I. and others (2009) Interventions to Prevent Child Maltreatment and Associated Impairment. 'Lancet.' 373: 250-266; Paulsell, D. and others (2010) 'Home Visiting Evidence of Effectiveness Review: Executive Summary.' Report prepared for the Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services. Princeton, NJ: Mathematica Policy Research, Inc. Available at: http://homvee.acf.hhs.gov/HomVEE_Executive_Summary.pdf
Additional information is available at: http://homvee.acf.hhs.gov/Default.aspx

26. Paulsell, D. and others (2010) 'Home Visiting Evidence of Effectiveness Review.' Report prepared for the Office of Planning, Research and Evaluation, Administration for Children and Families, Department of Health and Human Services, Washington DC.

27. Haskins is on the Board of MDRC.

28. Kirby, D. (2001) 'Emerging Answers: Research Findings on Programs to Reduce Teen Pregnancy.' Washington, DC: National Campaign to Reduce Teen Pregnancy, p. 179.

29. Scher, L., Maynard, R.A. and Stagner, M. (2006) 'Interventions Intended to Reduce Pregnancy-Related Outcomes among Adolescents.' (Oslo, Norway: Campbell Coalition).

30. National Commission on Fiscal Responsibility and Reform. (2010) The Moment of Truth (Washington DC: White House). Available at: http://www.fiscalcommission.gov/sites/fiscalcommission.gov/files/documents/TheMomentofTruth12_1_2010.pdf

31. The Unlikely Revolutionary. 'The Economist.' August 12, 2010.

32. Obama, B.H. (2010) 'Social Innovation Fund.' Reprinted in Office of Social Innovation and Civic Participation. Available at: http://www.whitehouse.gov/administration/eop/sicp/initiatives/social-innovation-fund

33. Corporation for National and Community Service 'Inaugural Social Innovation Fund Grants Awarded to Experienced Innovators.' Press Release, July 22, 2010. Available at: http://www.nationalservice.gov/about/newsroom/releases_detail.asp?tbl_pr_id=1829

34. Corporation for National and Community Service (2010) 'Social Innovation Fund.' Available at: http://www.nationalservice.gov/about/serveamerica/innovation.asp

35. Department of Labor (2011) 'US Labor Department Encourages Applications for Trade Adjustment Assistance Community College and Career Training Grant Program.' News Release, January 20, 2011.

36. US Department of Labor 'FY 2011 Budget in Brief.' p. 3.

37. Gray, S.W. and others (1966) 'Before First Grade.' New York: Teachers College Press; Gray, S.W. and others (1973) The Early Training Project: A Seventh Year Report. 'Child Development.' 41: 909-924.

38. Pianta, R.C. and others (2009) The Effects of Preschool Education: What We Know, How Public Policy Is or Is Not Aligned with the Evidence Base, and What We Need to Know. 'Psychological Science in the Public Interest.' 10(2): 49-88; Consortium for Longitudinal Studies (1983) 'As the Twig Is Bent: Lasting Effects of Preschool Programs.' Hillsdale, NJ: Lawrence Erlbaum.

39. Puma, M. and others (2010) 'Head Start Impact Study: Final Report.' Washington DC: US Department of Health and Human Services.

40. Head Start is not a test of the Abecedarian or Perry Preschool models. Rather, it is a funding stream that supports a variety of preschool programs that have the same general goals but vary along many dimensions and vary greatly in quality.

41. Ioannidis, J.P.A. (2005) Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. 'Journal of the American Medical Association.' 294 (2): 218-228; Zia, M.I. and others (2005) Comparison of Outcomes of Phase II Studies and Subsequent Randomized Control Studies Using Identical Chemotherapeutic Regimens. 'Journal of Clinical Oncology.' 23 (28): 6982-6991; Chan, J.K. *et al.* (2008) Analysis of Phase II Studies on Targeted Agents and Subsequent Phase III Trials: What Are the Predictors for Success. 'Journal of Clinical Oncology.' 26 (9): 1511-1518.

42. Garet, M.S. and others (2008) 'The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement.' Washington DC: Institute of Education Sciences, US Department of Education. Available at: http://ies.ed.gov/ncee/pubs/20084030/

43. Bloom, H.S., Michalopoulos, C. and Hill, C.J. (2005) Using Experiments to Assess Nonexperimental Comparison-Groups Methods for Measuring Program Effects. In 'Learning More From Social Experiments: Evolving Analytic Approaches.' New York: Russell Sage Foundation, pp. 173-235; Cook, T.D., Shadish, W.R. and Wong, V.C. (2008) Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons. 'Journal of Policy Analysis and Management.' 27(4): 724-50; Glazerman, S., Levy, D.M. and Myers, D. (2003) Nonexperimental versus Experimental Estimates of Earnings Impact. 'The American Annals of Political and Social Science.' 589: 63-93.

44. Auerbach, A. and Gale, W.G. (2010) 'Déjà Vu All Over Again: On the Dismal Prospects for the Federal Budget.' Washington, DC: Brookings.

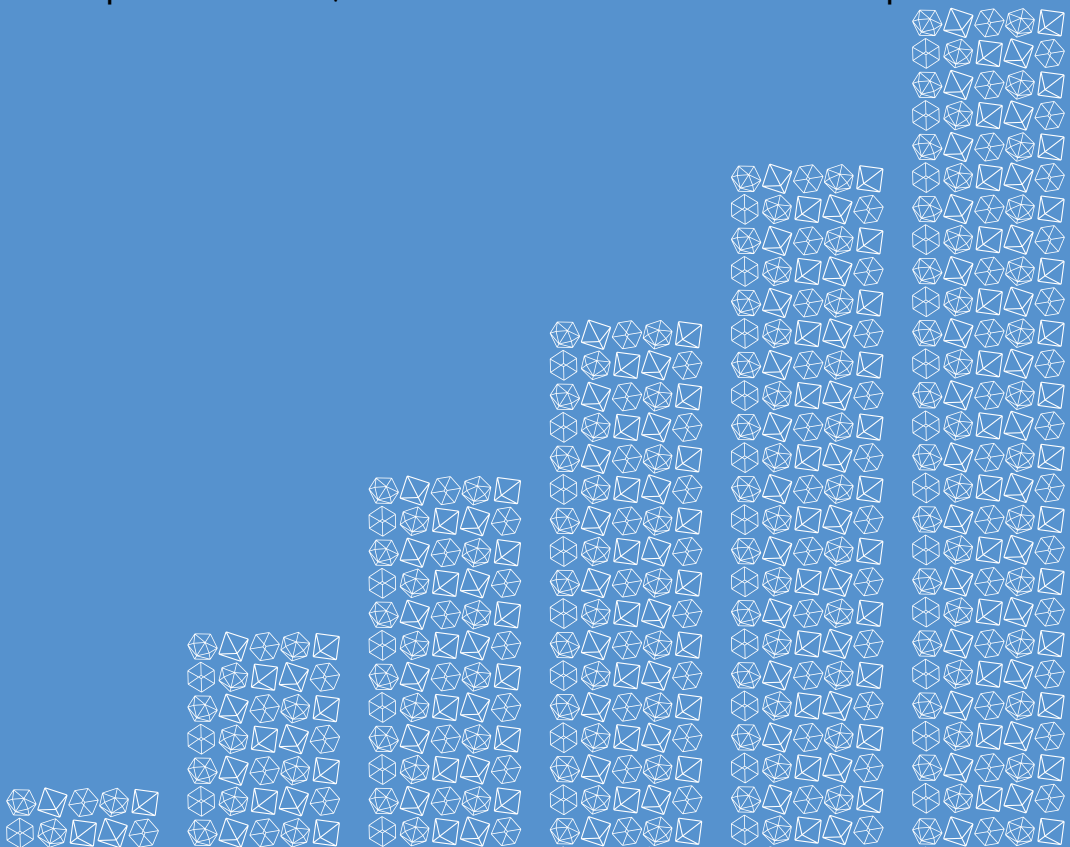
45. Rivlin, A.M. and Sawhill, I.V. eds. (2004) 'Restoring Fiscal Sanity: How to Balance the Budget.' Washington, DC: Brookings.

46. As early as 1996, the Welfare Reform Law gave the Secretary of the Department of Health and Human Services the authority and funding to conduct evaluations of welfare reform and of other provisions in the legislation. The law gave the Secretary wide discretion to select important research topics, but the statute directs the Secretary: "to the maximum extent feasible, [to] use random assignment as an evaluation methodology." See Section 413 of Title IVA of the Social Security Act (42 U.S.C. 613).

Project Oracle

Understanding and sharing what really works

Mat Ilic, Greater London Authority and
Stephen Bediako, The Social Innovation Partnership



Executive summary

What is Project Oracle?
Project Oracle is part of a portfolio of policy interventions that emerged in November 2008 in the Mayor of London's *Time for Action* programme. The programme was the Mayor's call to partners in recognition of the scale and severity of the issues of youth crime and violence in London. *The Time for Action* programme as a whole, as well as *Project Oracle* specifically, focussed on the causes of youth crime and violence – underachievement, lack of opportunity and aspirations, and the varied quality of existent services – in dealing with underlying issues rather than symptoms.

The contribution of *Project Oracle* is to 'understand and share what really works' in improving outcomes for young Londoners. In practice, this goal will be achieved through a fluid evidence base of London-centred interventions that are either highly promising or already proven in delivering positive effects for children and young people in the capital. Centred around an interactive website and improvement process, the product will support commissioning decisions, inform the design of programmes so as to avoid repetition and duplication, and provide an assured listing of projects and programmes that could benefit from additional evaluation and research carried out by academics.

Project Oracle will introduce internationally-recognised Standards of Evidence to be met by providers of services for children and young people – ensuring they focus their attention on interventions that have been proven to work. The Standards offer five levels of evidence, with 'one' being the entry level requiring a sound theory of change or logic model, and 'five' being a programme that is ready for delivery at scale: a 'system-ready' intervention with multiple replication evaluations behind it, and clear manuals for delivering with fidelity and quality.

In contrast to similar initiatives, *Project Oracle* is not based on a pass-fail system, but rather a continuous improvement process that is scalable to the nature and shape of interventions themselves. London is made up of 32 autonomous local authorities plus the City of London Corporation; it also hosts tens of thousands of community and voluntary sector groups, all of varying sizes and levels of organisation. The common factor among these is their apparent eagerness to make a difference to the lives of children, young people, their families and communities.

In general, however, the lack of evidence of effectiveness (or at least lack of access to it) is a well-recognised problem. Consequently, a clear short-term benefit of *Project Oracle* is the accumulation of a knowledge base of available preventative services in London, with the additional advantage being the analysis of their diverse methods and their effectiveness.

Why is Oracle important?
A project like Oracle is worth more than the sum of its parts. Although it has some clear and visible products that have been sponsored and hosted by the Mayor of London and the Greater London Authority, its effect should drive a shift in the way that commissioners, providers and policymakers operate in London. If successful, the project will:

- Challenge funders of services (typically local authorities) and providers of programmes (typically the voluntary sector), as well as academics and policymakers, to begin talking the same language around evidence. Too much time and energy is expended in debates around specific 'toolkits', methods and processes, policy semantics, without a foremost agreement over principles of evidence. These include, for example, applying the same level of scrutiny to statutory provision as to voluntary provision. Several policy drives intuitively emphasise the upcoming role of the voluntary sector in the delivery of public services, without necessary agreement over what the relative evidence requirements are, prior to committing to such radical change.

- Create a progressive dialogue about evidence driven by the Oracle website. The website encompasses a programme registration and assessment platform, where individual programme managers can test their programme design and evaluation quality against the Standards of Evidence. This mechanism, coupled with a brokered social networking function, presents ongoing scope for collaboration, information exchange and consensus-building between teams of funders, providers and academics. Previously, each group would have pursued its own goals without raising concern for the larger issues at stake – namely, the potential improvement in the lives of young Londoners if they brought their expertise together.

- Fundamentally, *Project Oracle* recognises that evidence itself is not an absolute end. Programmes that have failed can potentially be learnt from, and delivered successfully if their mistakes are rectified in the future. Equally, even if a promising programme does not yield strong evidence, such evidence can be cultivated over time. While many of the programmes in question are morally/ethically stable and well-intentioned, they occasionally lack a clear comprehension of what they intend to achieve; consequently they are unable to assess whether they have succeeded in their aims.

- *Project Oracle* is not without its limitations, as it does not carry any guaranteed solutions for funders and policy experts. It does, however, offer informed expert presence and an overarching approach concerning decisions about design, commissioning and implementation of social programmes, by recognising objective evidence as a substantial determinant of these decisions. Through this approach, it enables the local formulation of programmes, while still applying principles exercised by successful precedents, including the Washington State Institute for Public Policy, 'a non-partisan state research institution' based in the US.

Abstract

This document seeks to provide a comprehensive overview of *Project Oracle*. The first section supplies a brief introduction to the context of the current debate around evidence-based policy. The second addresses the origins of *Project Oracle*, exploring the *Time for Action* programme, and the reasons behind its establishment. The third section considers the origins and development of the Standards of Evidence, starting with the work of the Dartington Social Research Unit whose rigorous standards should stand the test of time. Also, the role of Community Links and The Social Innovation Partnership in helping to tailor the Standards to fit the voluntary sector.

The fourth section provides an overview of the technical aspects of the project. The fifth section explores the concept of working with a small range of pilots, with which *Project Oracle* has succeeded to shape the market and elicited interest from a variety of groups. The sixth section explores how the GLA and its partners have enabled Oracle to flourish, by making it practical enough for large and small projects to apply without losing the original academic rigour that underpins it. Section seven considers the different applications for the project, and how it can be utilised in different contexts. Section eight highlights the proposed activities for 2011/2012, recognising the need for the project to expand in a targeted and considered manner so as to remain influential. Finally, section nine provides some conclusions and lessons learnt from Oracle at this early stage in its development.

1. Introduction

Public services are facing a wave of change occurring at an unprecedented speed and scale. The Government has committed to a series of wide-ranging reforms, with a renewed focus on achieving better social outcomes at a reduced cost. Some of the more challenging debates now focus on what constitutes reliable evidence of social change, and how outcomes or results of specific interventions can be verified.

In the UK there is rare cross-party consensus on the benefits of early intervention. For example Iain Duncan Smith MP and Graham Allen MP have progressed the 'what works' agenda.

- The Allen Review[1] recommended a portfolio of evidence-based early intervention programmes aimed at improving outcomes for children, young people and their families – its next phase will look at novel ways of financing these interventions;

- The Centre for Social Justice identified in a recent report *"…a fatal failure at the heart of government spending decisions."*[2] – namely the absence of adequate evidence in the construction of social policies and programmes, in turn leading to poorly designed provision with little or no proof of positive change.

In addition, with the onset of an era of localism (or 'mass localism'[3]) where individuals and communities need to come together to resolve their problems with local solutions, one is entitled to ask how communities and individuals will obtain the knowledge and evidence they need to make sound decisions.

As a response, the Greater London Authority (with support from The Social Innovation Partnership) contributes this paper on *Project Oracle*. The project seeks to further the aims of evidence-based policymaking, by stimulating collaboration between government, academia and the wider social intervention community. *Project Oracle*, having emerged from the Mayor's proposals in *Time for Action* (his programme to tackle serious youth violence), seeks to 'understand and share what really works'[4] in supporting young people and preventing youth violence. It does this by combining, for use by service providers, funders and policymakers:

- A tried and tested programme evaluation methodology.

- A technical platform – the *Project Oracle* website.

- A supportive centralised officer resource.

2. The origins of Project Oracle

Chronology

→ Project initiated, GLA project officer allocated – October 2009

→ Standards of evidence developed – November 2009 – April 2010

→ Standards of evidence: consolidation and review – May 2010

→ Practitioner guidance in development – August 2010 – December 2010

→ First pilot working with ten projects aimed at tackling youth violence – December 2010 – January 2011

→ Consultation period – January 2011 – March 2011

→ Website 'soft launch' – April 2011

→ Second pilot and expansion to GLA group projects – April 2011 onwards

Project Oracle emerged as one of the six work streams from *Time for Action* (November 2008): the Mayor of London's 'proposals and call to partners' aimed at equipping young people for the future and preventing violence. The project asks for a new standard in the quality of evidence we base our social policy decisions on. The Mayor believes that he is *"strategically best placed to initiate joint work to… evaluate programmes that have greatest benefits and to identify those that don't…"*[5] As quality and results cannot be the responsibility of an individual or single group, the Mayor recognises the need for working across a partnership of policymakers, public sector managers, funders, service providers and academics. For this reason, the project's Delivery Board represents cross-sector leadership, encompassing statutory agencies such as the Metropolitan Police, Directors of Children's Services and the Youth Justice Board, as well as membership organisations such as London Funders.

Commissioners and service providers have called for easily accessible information on what really works. The challenge for the project has always been how to generate this content from London-based activities, avoiding the temptation of importing 'proven' programmes from abroad. To achieve its aims, Oracle employs an online, 'refereed' resource to share examples of effective practice from successful projects across the capital. Uniquely perhaps, it will provide a common frame of reference for academics, commissioners and providers.

Such an ambitious mission can only succeed with focus. The Mayor's initial focus will be children and youth projects that are commissioned and often delivered from within the GLA family – specifically those of the Metropolitan Police, Transport for London and the London Fire and Emergency Planning Authority. This already accounts for tens of millions of pounds in investment, and it is imperative that this expenditure demonstrates commensurate returns.

Project Oracle seeks to further the aims of evidence-based policymaking by stimulating collaboration between government, academia and the wider social intervention community. It was established in recognition of four key factors:

- **There is currently no clear understanding of what programmes[6] work**, in what conditions they work, and whether they therefore represent 'value for money', relative to each other or to 'doing nothing'.

- **There needs to be a sustainable body of evidence** so that the knowledge base evolves for future policymaking.

- **Evidence needs to be cultivated from somewhere using a consistent method**, requiring a stimulus and a mechanism for providers to develop continuously.

- **The root causes of social problems** require a deep understanding that should be captured and documented, particularly given the pace of social change.

Out of the six work streams in *Time for Action*, *Project Oracle* has the smallest media presence. It can be difficult to describe without the use of 'policy-speak' or jargon. If successful, however, *Project Oracle* could have far-reaching and dramatic impact.

3. The origins of Oracle's standards of evidence

- Standards of Evidence for London were commissioned by the GLA and delivered by Dartington Social Research Unit in collaboration with international experts such as Delbert Elliot (University of Colorado) and Steve Aos (Washington State Institute for Public Policy).

- The Standards document has been translated into a set of practical activities that can be conducted by providers of social programmes. These activities are part of a continuous improvement process that involves providers going through an on-line self-assessment, which tests the robustness of the evidence in support of their programme(s).

In this context (violence prevention) the current exemplar on 'what works' is attributable to studies begun at the Center for the Study and Prevention of Violence (CSPV), University of Colorado, in the US, where the 'Blueprints' methodology was created. Dartington Social Research Unit (SRU) took forward the academic work on standards of evidence in the UK.

In selecting Blueprint Programmes, CSPV set a very high scientific standard of programme effectiveness. They reviewed more than 900 delinquency, drug and violence prevention programmes. Of these, only 11 met the necessary standards to become future Blueprint Programmes. To pass muster, a sustained effect is required for at least one year after treatment, with no subsequent evidence that this effect is lost. Programme effectiveness is based on an initial review by Blueprint's staff and a final review and recommendation from a distinguished Advisory Board, composed of experts in the field of violence prevention.

When SRU were asked 'Why is there a need for Standards of Evidence?' they countered that standards provide a benchmark for scientific evidence of effectiveness. In the absence of any standards, too many interventions for children and families are being used in ignorance. Worse, some are being implemented in spite of strong indications that they can be harmful.

Better information will mean that resources are more confidently directed towards the most effective projects. Widespread use across the London region of a single set of criteria will bring greater consistency to judgments concerning the reduction of youth violence and promotion of health and development.

The Oracle levels of evidence

Level 1 Sound theory of change or logic model, with clear plans for evaluation.

Level 2 Demonstrating emerging evidence of impact.

Level 3 Effective – comparison group (ideally random controlled trials), statistical analyses, 'effect sizes'.

Level 4 Model – analysis of 'dosage', of 'fidelity of implementation'.

Level 5 System-ready – 'multiple independent replication evaluations', and cost-benefit analysis.

(NB. These standards have been summarised for the purpose of this report.)

It is recognised that London's Standards, while a product of international academic expertise and experience, need to be adjusted to suit the specific experience of the city's landscape. Sections five and six following examine in more detail how the Standards were made practical.

In developing the Standards of Evidence further, the GLA worked with partners to introduce in detail the concept of a Theory of Change. This is an approach to planning and theorising social change. It was developed in the US 25 years ago, and progressed and refined by ActKnowledge and the Aspen Institute. Arguably, in the absence of evidence, a well-composed and theoretically sound programme creates the backbone for well-designed evaluations and studies. As noted in section six, with the use of some well-defined steps within a workshop setting, the articulation of a theory of change is a practical exercise that most practitioners can engage with.

Extract from Oracle practitioner guidance

The Standards Framework: five levels?

The *Project Oracle* self-assessment works against a five-level graded framework. These levels provide organisations with an indication of how they fare against the Standards of Evidence, so that they can improve over time.

This guidance focuses on the details of the five levels, with specific emphasis on Levels 1 and 2, which should be relatively familiar to organisations used to evidencing results. As the levels are cumulative, it is vital to get the foundation levels 1) and 2) right.

Tip: don't worry about the levels. Focus on the instructions.

As the above extract maintains, organisations are asked to focus on the 'foundation levels' of the Standards framework.

There are several reasons for this:

- Levels 1 and 2 are more representative of what organisations understand by 'evaluation', while Levels 3 and above are more representative of social research, and in some cases require an understanding of complex techniques, as well as access to resources and data.

- As such, the focus on Level 1 and 2 places the onus on the provider organisation (irrespective of size) to establish the foundation for solid evaluations and additional research – i.e. Level 1 and 2 should be achievable without the aid of statisticians and social researchers.

- Good programme design, therefore, with reliable data sources and other processes resulting from Level 1 and 2, can provide opportunities for more complex research queries.

- In effect, the Standards become scalable to the type and size of social programme in question. Realistically, small voluntary groups working with a small number of individuals cannot be expected to provide examples of randomised controlled trials, but they should be able to demonstrate how they are taking account of research studies that affect their own work. This can be evidenced by a strong theory/logic underpinning their work, the quality of activities/interventions, and some emerging (ideally comparable) metrics to account for activity.

Given the significance afforded to the Levels 1 and 2 in the Oracle assessment process, it seems essential to elaborate on some of the detail.

Level 1 – testing the soundness of the theory or logic behind the programme

Two recognised methods of explaining programmes in advance of committing to an evaluation are a theory of change or a logic model.

Logic model
A logic model is a standardised framework that is used to provide a cogent description of a project or programme of work. It is most useful when a project is on the edge of advancing into delivery.

A basic logic model includes four components, organised through a linear sequence, describing the logical flow from:

1. Inputs: these are all the resources a group needs to carry out its activities (finance, staff, equipment, facilities).

2. Activities or interventions: which the actions, tasks and work a project or organisation carries out to achieve its aims.

3. Outputs: which are products, services or facilities that result from activities.

4. Outcomes: which are the changes, benefits, learning or other effects that result from what the project or organisation makes, offers, or provides.

The grid below illustrates:

Inputs → Activities → Outputs → Outcomes

Other versions of a logic model distinguish between long-term outcomes and a set of intermediate results.

Many smaller organisations run their projects instinctively. Yet, taking the effort to document their work in a coherent manner that other stakeholders can recognise is recommended to them as a useful investment. It enables them to build models that can be replicated, and allows them to explain their work consistently to potential supporters and funders.

Theory of change
When attempting to challenge a complex social issue, a theory of change can be a helpful way to assess and organise proposed solutions.

A theory of change articulates the long-term goal that any project is aiming to achieve, and systematically considers all of the relevant elements that might affect the desired outcomes. Whilst there is not one standardised format, common core elements include defining:

A Long-term Goal; Outcomes/Preconditions; Strategies/Interventions and Assumptions

A theory of change can be applied to both explain what set of conditions have caused a specific change and to predict what changes may occur if interventions unfold in a particular way.

A specific link is made between conditions created to make a change and the actual outcomes. A theory of change can also support how a project collects meaningful data towards reaching its long-term aim.

Though this activity is neither simple nor instant, it does prove valuable. Practically, a solid theory can be constructed within a half-day facilitated workshop, depending on the complexity of a project. The final product:

- Can be concisely explained.
- Demonstrates causal paths from interventions to outcomes.
- In a way that takes account of assumptions.
- Is targeted at a particular population.
- Can support evaluation.

Level 1 does not express a preference to a logic model or a theory of change. However, the line of questioning asks for a good level of detail, essentially testing the user's understanding of each of the components of their theory of change (see page 63 in italics).

Level 2 – emergence of evidence for the project

Level 2 assumes the successful completion of Level 1 on the assessment. The reason for this is obvious – if a project is not constructed and designed with a sound logic, then there is little reason for it to even consider growing its evidence. As Howard Bloom, the Chief Social Scientist at the US-based research organisation MDRC stresses in his 'Nine Important Lessons', *"The three keys to success are 'design, design, design' (just like 'location, location, location' in real estate.) No form of statistical analysis can fully rescue a weak research design."*[7]

The reasoning behind Level 2, as discussed above, is that it introduces programme providers to standard methods of evaluation, with a view to making organisations self-sufficient in evaluation, and less dependent on experts, who in most cases do little to transfer methodological knowledge to organisations after completing reviews and evaluations.

Components of Level 2 include:
Evaluation planning – progress can only be assessed if the starting point is known. Organisations are encouraged to therefore consider evaluation from the outset of commencing their work, rather than as a by-product or end-result of delivery.

How to measure – this is often a key sticking point that results in organisational paralysis in evaluation. The Oracle guidance, as part of the self-assessment process, points to several terms for consideration, including, but not limited to:

- Distinctions between quantitative and qualitative research, emphasising that both are equally useful, as each project is different, but noting that a mix of both approaches often works best. Once the proposed measurable outcomes have been identified, for each outcome being measured, details of desired results are required – this may come in the form of numerical targets or descriptions.

- Indicators – leading: a signal of future action and lagging: an indicator that follows an event.

- Data capture methods (e.g. questionnaires, observations), pointing in particular to standardised measuring instruments such as the Strengths and Difficulties Questionnaire (SDQ).

- Data source (e.g. GLA London Datastore, participant interviews).

- Data capture intervals (i.e. days/months/years – before, during, after project).

As noted above, Levels 3-5 require advance knowledge of evaluation techniques, or indeed a resource base that can finance external evaluations to this quality. The Oracle practitioner guidance does not focus too much on these elements for exactly these reasons, but the standards are quite clear as to what is required to demonstrate achievement, in summary:

Level 3 – effective-comparison group

- Any documents (e.g. a manual) that detail the interventions and skills and training required.

- At least one replication evaluation with a comparison group, and include a long-term follow-up of the most important outcome issues.

Level 4 – model; as above plus:

- A statement of information on the resources (economic and human) required to deliver the project.

- A statement of evidence to support the causal mechanisms that underlie the change in outcomes being sought.

- At least two rigorous evaluations of the project, including one that does not involve its developer.

Level 5 – system-ready; as above plus:

- Technical documentation to support large-scale implementation.
- Plans for sustaining the project model with quality.
- Multiple independent replication evaluations.
- Data on the economic costs and benefits of achieving the outcomes, arrived at using methods that meet internationally recognised standards.

Level 5 does not feature in the self-assessment since these programmes would already be well-established and led by large professional organisations. Programmes that would qualify would need to demonstrate design, supporting manuals and evaluations on a par with the likes of Triple-P and Family Nurse Partnerships, both very notable international evidence-based programmes. At time of writing, projects originating in London were some way behind these counterparts.

4. Project Oracle: an overview

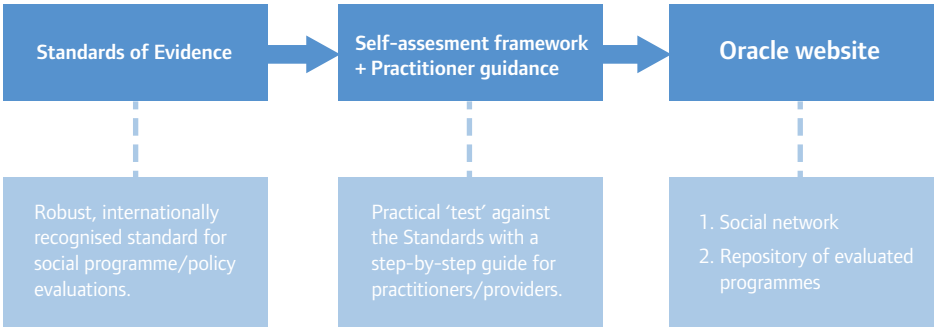
How it works

The Oracle self-assessment is carried out by the provider, alongside a practitioner guidebook. Depending on how their project scores (Levels 1-5, with 1 being entry level), practitioners are expected to undertake an improvement plan aimed at increasing the level of evidence that supports their claim to positive outcomes.

This self-assessment activity requires a degree of human interaction and peer challenge, which in the current pilot stages is facilitated by a dedicated project officer from the GLA, with support from Master's students from social research disciplines. The GLA also runs monthly workshops and will engage in online collaboration once the Oracle website has benefited from further development (April 2011).

The key success factor for *Project Oracle* is the extent to which it can engage a disparate group of social policy organisations into a conversation about evidence – possibly only achievable through incentives, and in the absence of immediate financial incentives, a level of accessibility to expertise. To encourage this accessibility, Oracle must also remain conceptually simple.

The diagram below summarises the mechanics of the project:



In summation, the provider organisation registers its project on the Oracle 'repository'. Apart from the GLA group projects, all other projects are encouraged to participate, but those taking up the offer in the absence of funding and financial incentives tend to have the following characteristics:

- Belief that evidence has inherent value in informing their practice, and in turn leads to better results for their client groups. Most organisations value learning from others and welcome the challenge offered by an independent party.

- Appreciation for the fact that funders want confidence over the likelihood of outcomes being achieved, since this makes their 'business case' easier to approve. Being on the Oracle website means projects have gone some way to demonstrating quality of evidence under the watchful eye of the Greater London Authority/the Mayor of London.

- Strategic commitment to being evidence-based, irrespective of the starting position, understanding that gathering evidence takes time, and resources, and that potential interested parties also appreciate this.

Once it has carried out the self-assessment, the organisation submits evidence to justify its self-assessment at a given level against the Standards of Evidence (as detailed above). Centrally, through desk research and consultation with experts, the

Oracle project officer validates the self-assessment at whichever level is deemed most appropriate based on evidence provided.

The time it takes to complete a validated self-assessment on the Oracle project side is dependent on the complexity and maturity of the project. Typically, four to seven days is not unusual, which includes writing up the findings in a one-page summary paper (content illustrated below). One of the key challenges, therefore, is not only human resource (which in the short-term will be filled through Master's students), but also the implications of this time requirement on any decision making and/or commissioning processes.

The validation process is, of course, performed in consultation with the programme providers, who must not only agree to the moderated level Oracle has marked them at, but also commit to working with the project to improve their evidence base. This is done through a detailed and time-bound joint action plan.

Upon receiving validation at a certain level, the project is then published on the Oracle website. This covers basic project registration details, such as what the project is, where it is based and who it works with, and the current level of evidence. As noted above, the listing also has a downloadable one-page fact-sheet, completed by the Oracle team, encompassing the following:

- Date of most recent review.
- Key priority area (e.g. early years).
- Name of organisation.
- Name of project.
- Borough characteristics – common characteristics of areas where it has been attempted.
- How established it is compared to similar initiatives (or if it can be considered innovation).
- Short summary (100 words).
- Key ingredients to achieve its outcomes.
- Long-term goal.
- Target population.

- Recruitment and referrals.
- Desired outcomes.
- Does it work?
- Level achieved.
- Reference to existent evaluations.
- Key limitations of evidence.
- Key lessons.

5. The pilot approach

In September 2010, the Oracle Challenge Panel (an advisory body that forms part of the project governance) agreed to the commencement of work with an initial group of ten Oracle 'pilot' projects. The GLA selected pilots for their cross representation, but also because of their closeness, so that they would become willing partners on the project in the future.

The pilot exercise was conducted with a view to testing the Oracle self-assessment approach. The sessions examined whether the self-assessment yielded the right information. They also provided an illustration of the relative amount of work required both for the participating organisations and the GLA.

The pilot has consisted of working with predominantly youth crime reduction projects provided by large to medium and small-scale charities, smaller groups, and in one case, a youth offending team. The initial phase of work, formed of intensive one-to-one meetings and joint workshops, has been useful in outlining how *Project Oracle* could work with providers going forward, keeping the burden of work to an absolute minimum so as not to disadvantage smaller voluntary groups.

In the next phase, around half of the initial pilot projects have agreed to go 'under the microscope' to have their evidence appraised for robustness. Clearly, as illustrated in the previous section, not all organisations feel they are ready for this activity.

The following case study summarises the position of one of the Oracle pilot projects against the Standards of Evidence, giving an indication of the application of the Standards in practice. Another case study is included in Appendix 3.

Cricklewood Homeless Concern: YES Project case study

Background
The YES Project is run by the charity Cricklewood Homeless Concern (CHC). CHC provides a comprehensive range of services to vulnerable people. These include housing advice and support; drug an alcohol day treatment services and training and employability. They also have a 'Community Engagement' project that offers support and infrastructure services to local communities. The Youth Engagement Scheme ('YES' Project) is a user-led initiative supported by CHC and local police tasked with creating opportunities for disaffected young people in the local area.

Level One: Project theory/logic
The YES project began in March 2008 when CHC identified a number of young people near its centre that were becoming involved in anti-social behaviour, including criminal damage and drug use. With funding from the Home Office, the project recruited a Youth Engagement Facilitator to work with these young people.

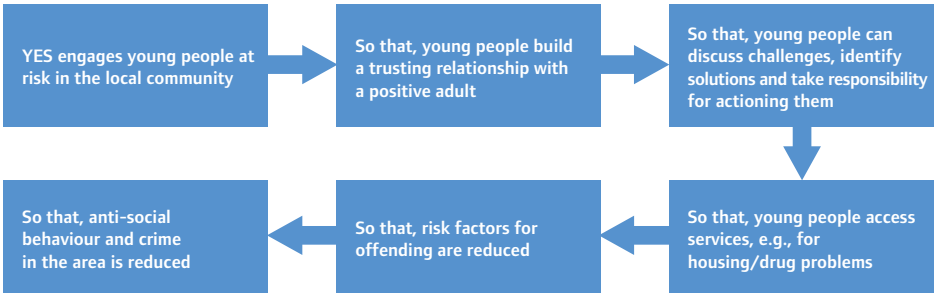
The worker is based at the CHC centre but goes out onto the estates to speak with young people and encourage them to attend the centre. CHC does not have a clearly defined target population; rather it works with young people in the local area that the Facilitator identifies as 'at-risk'. The worker engages with them, conducts an assessment of their needs, and signposts them to services (e.g., for help with housing or drugs). The worker builds the confidence of young people, and helps them to identify opportunities and develop skills (including in youth work).

Ultimately, the YES Project aims to turn young people away from criminal activity and exclusion, and help them to play an active role in the community. The project's aim is to encourage young people to take responsibility for their lives and change the path that they are on. It also aims to give young people a voice and provides positive activities. Its theory is that by engaging young people in their local area, listening to the challenges that they face, and helping them to find solutions, it is possible to re-engage them in community life as well as enabling them to become the solution to challenges faced by their peers.

As the *Project Oracle* guidance explains, a project theory includes a long-term goal, measurable outcomes, and a description of the main interventions/activities. This can be summarised as follows:

Long-term goals	Interventions	Measurable outcomes
Young people are diverted away from crime and antisocial behaviour	Engage young people and assess their needs	Improved confidence, behaviour and relationships
Young people play a proactive leadership role in their community	Provide support and refer to services	Crime and risk factors reduced among cohort
	Identify opportunities and build confidence	Increased take up of positive activities
	Train core group to become youth leaders	Youth leaders gain qualifications

A full theory of change drawing out all the links in the project's approach is too complex to show here. However, it is worth looking at one 'so that' chain to illustrate some aspects of the approach:



As the guidance explains, a theory of change should also identify how a project will collect meaningful data on the progress made towards long-term aims.

Level 2: Project evaluation planning
The project has some anecdotal evidence that indicates the impact of its work. For example, the project reports that levels of youth crime in the area have reduced and local residents report that the area is safer. In addition, the YES Project reports that the local community has embraced the project and views it as an excellent way of turning negative behaviour into something that is beneficial to individuals and the local community.

A possible evaluation framework for one intended outcome – reduced levels of anti-social behaviour and crime – could be illustrated like this:

Desired outcome	Indicators	Data capture
Reduced anti-social behaviour and crime in the area	Fewer phone calls to police re ASB Data capture	Recorded by police annually
	Residents report feeling safer	Interviews with residents annually

The YES Project has information on both of the indicators outlined above, which suggests that there is emerging evidence of its approach. For example, it knows that in 2008 there were 18 calls to the police in relation to anti-social behaviour in the area, but that in 2009, there were zero calls. However, there is not a clear evaluation framework in place – data is collected in an *ad hoc* way, by anecdotal report rather than regular and routine evaluation.

A further challenge is attribution: how does the YES project know that reduced levels of ASB and crime are due to its intervention rather than, say, the work of local police or other factors? Without tracking the specific cohort of young people on the programme, establishing whether they are the main perpetrators of ASB or crime (as opposed to their others in their peer group), it is not possible to determine the extent to which positive benefits in the community can be attributed to the YES project.

The project could therefore, re-frame its outcomes in terms of reducing the offending of young people on the programme (rather than anti-social behaviour in the area), and track levels of offending before and after its intervention.

Alternatively, rather than focusing on reduced ASB in the area, YES could 'positively frame' its outcomes, as the guidance recommends. This means that rather than focusing on the reduction of something negative (e.g. ASB) the project instead looks to capture the associated improvements in young people (e.g. increased resilience).

Next steps

In order to achieve Level 2, the YES Project would need to identify measurable outcomes and create a plan for capturing evidence of its impact systematically. A range of further outcomes could also be captured, as shown in the diagram below. Useful data could be collected by regularly surveying young people on 'soft outcomes' (e.g. self-confidence and relationships) and 'hard outcomes' (e.g. drug use, housing, and qualifications achieved), comparing their responses to the baseline data captured in the initial needs assessment. Some possible indicators are outlined below:

Desired outcome	Indicators	Data capture
Young people have improved attitude and relationships	Positive engagement with YES worker	Reported by CHC staff
Young people address risk factors associated with social exclusion and offending	Young people report improved confidence	Questionnaire every 3 months, e.g. SDQ
Young people play a more active role in the community	YP in contact with services, e.g. DAT	Reported by CHC, services and YP
	Improvements, e.g. reduced drug use	Recorded by services, reported by YP
	Involvement in activities or groups	Database capturing attendance
	Qualifications; progression to EET	Questionnaire to young people

Once a systematic monitoring system is in place and data is regularly being collected, it should be analysed and used to refine the project's approach. Having met these criteria, the YES Project would achieve Level 2 and could begin looking at capturing longer term outcomes, and focusing on replicating its approach to achieve Level 3.

6. Making Oracle practical

As noted in section three, the Standards of Evidence yielded a robust academic framework. Beyond this, *Project Oracle* needed to move into applying it through implementation. Without this, its usefulness to small-scale community organisations would never be realised. This would present a significant limitation.

To progress, the GLA sought a partner to help deliver the following outcomes:

- Key stakeholders are bought into the *Project Oracle* standards, particularly those working with young people most at risk of serious youth violence.
- Practitioners have made a contribution to the further development of *Project Oracle* standards.
- *Project Oracle* understands requirements and concerns of practitioners in the field across London, particularly those working with young people most at risk of serious youth violence.
- User groups and key stakeholders are committed to using the Oracle web portal once available – submitting self-assessments via the portal to enable the Project to build up its evidence base.

The successful providers from this exercise, The Social Innovation Partnership and Community Links worked to invigorate the project through a series of consultation events. Community Links ran ten workshops 'in the community' across London – 530 people viewed the workshop events registration page, 152 individuals registered for the workshops and 109 people (from 78 different organisations) attended the workshops. The participants were self-selecting having been contacted through a range of networks (participating organisations are listed in Appendix 2).

These workshops yielded a number of key lessons

Experiences and benefits of constructing project models (theory of change or logic model)

Broader perspective of the project/programme
The process of developing theory of change or logic model allowed participants to take a broader perspective of their projects or programmes. It also provided an alternative view to challenge their thinking behind the projects. This broader and more systematic examination of their programme also helped to identify and fill the gaps in the programme design (e.g. scope of the project, developed a cross-cutting view of the project activities, helped developed linkages and relationships between programme activities, skills needed, additional set of target audience, etc.)

Better articulated project model
Using theory of change, participants were able to better articulate their project model – how their activities are going to achieve the project goal.

Clearer measurable outcomes and more specific interventions in relation to measurable outcomes

The theory of change became a tool to help participants clarify their measurable outcomes. Instead of simply listing measurable outcomes, participants identified the causal relationships between the outcomes. Participants were also able to explain how their project activities will produce the set of measurable outcomes in a logical manner.

New Element: assumptions

In a theory of change model, projects are required to identify assumptions – existent conditions that affect the project, which project managers can influence but cannot control. Participants generally found this a challenge, but a helpful addition – recognising just how complex their world is. It challenges project thinking, creating a more realistic understanding of the programme and the communities it is meant to help.

Other uses for Theory of Change

Participants thought that there are some potentially interesting uses for theory of change. Some ideas included encouraging commissioners and funders to build a theory of change in to their funding and commissioning plan, so that practitioners can see how their projects/ideas can fit within their commissioners' expectations model.

In addition, it was suggested that funders or commissioners should build the theory of change development process into the funding requirement, requiring projects to develop their project model before delivery. The process of building a theory of change can also be used as a way to build community consensus and partnership, creating a common vision for success.

Discussions around data collection and measurement

Data collection

- Need to highlight the ethics involved in data collection (and possible legal implications e.g. data protection).
- Need to explain difference between quantitative and qualitative – when to use these (when is it appropriate and relevant to use quantitative and/or qualitative methods).
- Include signposting/references for people to learn about other quantitative and qualitative data collection methodologies.
- Give examples of different measuring instruments, case studies, provide links to these (e.g. http://www.outcomesstarsystem.org.uk/).
- Explain how to capture soft outcomes through indicators.

Data references

- Guidance was required around links to national/regional/local data information sources (e.g. census, British Crime Survey).

- Key issues explored around accessing data, such as police and crime statistics regarding individuals.

Community Links also carried out a survey of all the participants and other groups. One of the key findings of this survey was that bigger projects did better than the smaller ones, 62 per cent of the biggest projects (with more than ten staff) still didn't have an evaluation plan, only just behind the smallest projects on 70 per cent. This perhaps indicates that contrary to conventional wisdom good evaluation is not simply a function of organisational size. Full results of the survey are presented in Appendix 1.

The Social Innovation Partnership contributed a significant strategy and policy edge to the project. Working with Lord Victor Adebowale, they ran a series of seminars, with senior figures from academia, funding bodies, membership organisations, provider organisations and statutory bodies.

The events were introduced with the opening premise that the traditional role of evaluation in social intervention and youth projects was weak. There was a clear need to demonstrate results for the resources and effort expended. This was seen as a particular challenge in the current economic environment where all spend has to be justified and prioritised, and the benefits to users made clear.

Summarised below are the key findings that London's senior stakeholders agreed on in the course of these events:

- Oracle must be a portal that brings the youth project community together to share quality information on what really works.

- Oracle must have an easy to use website and methodology.

- If Oracle is to survive it needs to grow, and the GLA must grow it now or miss key spending or investment opportunities.

- The broader remit should be focused on intelligent commissioning and making the market in London work better.

- If successful, Oracle would also mean that genuine innovation could be stimulated – rather than relying on established projects or only inventing projects with each new funding sequence, be it annual, spasmodic, or based on political terms.

- The GLA should offer Oracle as part of the national move towards evidence-based commissioning.

One of the key debates that arose among the project team from these findings was about the challenge of 'individual currency versus evidence'. Commissioners, while reviewing the evidence of tenders/submissions, still make decisions based on track record, brand awareness and personal trust. The GLA recognised that Oracle would never be able to eliminate this fully – these social markets are not based on perfect knowledge – but it would play an important role in reducing investment in 'pet projects', duplication, 'reinvention' and so on.

This work from The Social Innovation Partnership also provided some interesting insights on what behaviour and strategy Oracle needed to follow to be successful in London:

1. Where possible try to remain politically neutral and focused on its core product (to produce a repository of evidence).

2. *Project Oracle* could only be a success if commissioning organisations (both within and beyond the GLA) are involved – and commissioning invested in evidence.

3. The need for a controlled 'go-to-market' strategy, at a time where new policy announcements occurred weekly. It is critical that stakeholders are drawn to Oracle when they appreciate its benefits. In practice, this means working with a small set of projects at first to incrementally build up coverage and remit based on Oracle's own evidence of success.

4. Oracle had to become self-sustainable, from a variety of perspectives. This includes explorations around income streams, looking outwards to other stakeholders to share ownership, and ensuring that 'true' representatives from statutory, VCS, commissioning, academic and youth themselves are involved in development.

7. Oracle's applications

Three scenarios have been identified in which *Project Oracle* and its methodology can make a difference:

1. **When commissioners are buying new services or developing new products 'in-house'** (thus the GLA is working with London Funders to move towards commissioning along Oracle lines).

2. **Helping statutory organisations to reconfigure their operations or programmes** (this has already started with 'GLA family' organisations, including the Metropolitan Police Service).

3. **Improving projects or programmes directly and championing a single standard of evaluation** (where Oracle works directly with providers to improve their evidence base).

Throughout the development of the project, the GLA has focussed on simplicity and clarity, to ensure that all concerned can successfully make the journey to improved outcomes for young Londoners.

In the previous sections, the last of the three applications has been given the most attention. As noted above, this is Oracle's 'core business'. However, in recognition of the fact that evidence is not relevant unless it influences the pathway from idea generation to commissioning and delivery, the project has begun to explore the other two realms. Some scenarios, based on real *Project Oracle* experience to-date, follow.

Influencing commissioning decisions, therefore, is a key objective for the project. This is in recognition of the fact that commissioning bodies are key market shapers for social provision and that failure of social programmes is often attributable to poor commissioning or delivery, rather than the intent behind the original idea or the subsequent design and result of the evaluation.

The underlying issue, it seems, is that despite national efforts through the Commissioning Support Programme (http://www.commissioningsupport.org.uk/), the extent of outcomes and evidence-based commissioning, as well as co-production, certainly in London, is limited.

Scenario 1 – When commissioners are buying new services or developing new products 'in-house'

- A team within the GLA itself was administering grant funding for a number of charity organisations to provide social programmes in London.
- *Project Oracle* supported the bid evaluation process by encouraging co-design of the process between the policy makers/commissioners and potential providers.
- Upon the selection of successful applicants, *Project Oracle* worked with these providers individually to construct their theories of change – they will feature in the second pilot of projects for the Oracle repository.
- The theories of change were fully understood by the commissioners in the GLA and were used to agree the performance-based funding agreement with the providers – ensuring that all parties were fully cognisant of what was expected from the projects.

Scenario 2 – Helping statutory organisations to reconfigure their operations or programmes

- A London local authority was considering establishing a large-scale ('system-level') programme to look at addressing issues with youth 'gangs' over a period of three years.
- The GLA ran a *Project Oracle* theory of change workshop involving local residents, local authority officers and voluntary sector groups in a half-day session involving 25 individuals.
- Critically, it became very clear that the stakeholders were not agreed on what the underlying problems they were trying to solve were, or who the groups they were going to deliver the programme to were.
- *Project Oracle* worked with the local authority to try to clarify these issues in consultation with all of the participants – emphasising that failure to reach a consensus on these issues would result in a poorly designed programme.
- Once the design stage is complete, the local authority could work with *Project Oracle* to embed its commissioning strategy into the programme design (to ensure outcomes are shared) and to appoint an evaluation partner to work alongside the programme for its entire duration.

8. Activities proposed for 2011/12

Influencing the GLA group projects

The GLA expects to embed the principles of this project across organisations that are a part of the GLA group (the GLA itself, Metropolitan Police Service, London Fire and Emergency Planning Authority and Transport for London), thus leading by example. In 2008/09, youth engagement and programme expenditure in the GLA group was estimated to be in the region of £40 million, which makes any effort to 'make it better' a worthwhile exercise in its own right.

The process of impacting GLA projects with the principles of *Project Oracle* has already started. Some examples are listed below:

- PRU advocacy pilot – this project involves small grants being provided to local authorities and pupil referral units to work with parents in order to improve outcomes for children. The contribution from *Project Oracle* includes projects developing logic models through collaboration with providers, and an evaluation commencing at the same time as delivery – a new experience for the schools.

- Expansion of VOYAGE programme – this programme is part of one of *Time for Action*'s other projects (Titan) and offers a specialised BTEC qualification for young BME men aged 14 to 15, with a view to improving attainment and social mobility among these groups. *Project Oracle* is currently working with the programme to establish an evaluation strategy and structure the project using either the Theory of Change or Logic Model approaches.

- Daedalus – this is the Mayor's flagship resettlement programme for young offenders. By drawing on the principles of Project Oracle, the programme it is benefiting from a robust evaluation. It is also the first project to have been published to the Oracle repository.

The need to expand the influence of the project across the GLA group is becoming increasingly important for the Mayor, who is ultimately responsible for the lives of young Londoners. To this end, there is recognition of the importance of being honest where programmes have not gone to plan and something has not worked.

Creating a virtual network to drive Project Oracle

Given the nature of London-based provision of social programmes, the potential scale of the challenge for *Project Oracle* is huge. To ameliorate an element of this challenge, the GLA will be piloting a virtual network for practitioners and providers in London, so that

they can be constantly informed of the latest policy developments, talk to funders and exchange knowledge and experience from practice on-line.

Sustainability and future strategy

The GLA is currently exploring, with Universities and funders in London if Project Oracle can, at scale, provide:

1. A brokerage service between London's university researchers and its youth projects that would carry no direct cost to either party. This will boost the evidence base for projects and act as a source of knowledge for the website.

2. A growing evidence base of initiatives and what circumstances they work in, to inform future policy and funding.

3. A consistent framework or standard (not tool) for evaluation.

In the simplest sense, having a catalogue of London-based youth programmes is only helpful if they have supporting evidence: the opportunity being explored is whether robust evaluations for specific community projects can be delivered in a way that is cost-effective and sustainable to the taxpayer and to voluntary groups. The expensive nature of evaluation, as well as its technical challenge, is a prohibitive factor in doing it more routinely and to a good standard. Based on surveys conducted for Project Oracle, providers cite lack of time, skills, capacity and sometimes a lack of expectation from funders as reasons for not evaluating well or at all. It is apparent that funders can play more of a role in supporting organisations in a sustainable way.

To make the whole service cost effective, there is the prospect of matched funding or *pro bono* support from universities. Further, to make the initative sustainable, young researchers could be 'seconded' into community organisations, thus improving management systems and delivering applied research in residence. There is doubtlessly a win-win to be had since researchers could benefit from professional experience and access to data-sets, providers would gain new evaluations and technical skills, and universities would easily be able to demonstrate the social impact of their research.

9. Conclusions

Achievements so far

In summary, *Project Oracle* was established as part of the Mayor's *Time for Action* programme – against youth crime and violence – but its approach is broader, looking at

causes and problems, not just symptoms, endeavouring to grow and retain evidence for children and young people projects in the capital. The Standards of Evidence utilised in this exercise are international and universally accepted, yet, the application of the standards is evolving to become London-specific, recognising the key drivers of demand and supply for evidence which are quite unique.

The GLA has been striving to make the project practical by moving on from the original focus on 'pure' research and statistical rigour. The 'gold standard' expected by the academic community is an aspiration but the reality is a long way away and we have to work up to it. A major challenge, which appears to have been overcome, was in translating the Standards to a set of practical steps that can be used by providers, who are predominantly practitioners from community and voluntary sector groups. It has been critical not to allow the 'great to be the enemy of the good enough'.

Empowering organisations to get involved

All types of organisations have been involved – not just large ones, both in the pilot and wider developmental work. This transparency and market engagement has proved critical in making this project as credible as it possibly can be. While it is too early to claim outright success, the GLA will carry on working closely with providers, as evidence comes from practice.

NESTA has put forward for consideration *Ten Steps to Transformation*,[8] one of which is 'innovation culture only comes from practice'. This is not about creating an innovative culture for its own sake, but a culture where people feel empowered and supported to affect change and adapt their practice. Empowering involves providing advice and practical support. In this way, *Project Oracle* offers a pathway from practice to evidence, enabling projects to reflect on what made them (or can make them) succeed and why.

Learning the lessons

Project Oracle is a long-term commitment and the GLA is just at the beginning. Below is a summary of key lessons learnt to date:

Success means influencing investment – Ultimately the success of this project is based on whether it can affect outcomes by influencing how spending in London is allocated; however, evidence needs to be cultivated from somewhere. This again points to the indisputable fact that there is a need to get provider organisations on side. Certain projects or organisations might always receive funding through philanthropy whether we like it or not. Government and private funders are equally culpable for this; it's the nature of the system.

The importance of focus – A major threat in the early days of the project is that it ends up spreading itself too thin. There is an enormous wealth of provision, and as a result, a significant degree of repetition. To avoid imploding, the project will have to remain focussed, which is why it will initially face a manageable number of organisations and try to influence others in tranches and to differing degrees.

Balancing flexibility with absolute application – Experience in running this project suggests there is no 'absolute' view of evaluation – just because something 'works' in one place does not mean it will work in another without diligence in application (fidelity); and just because it 'works' to deliver one set of benefits, does not mean it delivers a different set that we may be after. For this reason, the GLA recommends that government is flexible in negotiating on what benefits a certain policy or programme can bring from the design stages – taking on views of providers, experts and service users.

Building genuine accountability – A key focus must be to breed a culture of accountability in which evidence has inherent value, thus increasing the demand for evidence. This is why the GLA will lead by example, with its own projects first. Applying evidence-based approaches cannot be viewed as a process, a compliance requirement nor an interest, but an expectation. To truly achieve this, London needs to have consensus on standards – quality in delivery, evidence, cost/benefit numbers and so on – and this will not be achieved overnight.

Motivating and incentivising action – Organisations and individuals must be mobilised and motivated to do this, from the bottom to the top. The way the GLA managed to motivate voluntary groups is by nesting the need for evidence between two of their key priorities – making a difference to end-users and applying for grants – both need strong evidence of success and of failure.

An important question to conclude with is: 'what is the purpose of knowledge in seeking social change? When demonstration projects are funded as a test for national solutions (as NESTA and others do), how does the learning get retained for future decision making activity? If a promising initiative is identified, what are the chances that it will be sustained?'

The GLA aims for *Project Oracle* to be a source of information, evidence and inspiration for city-wide policymaking and decision making. The model may become one that other cities could adopt. There is no reason why, with time, the wider public cannot be empowered to scrutinise investment in social programmes by being given access to information about 'what works'. If future resources are directed in this way, the outcomes for young Londoners should improve more predictably over time.

Appendix 1: additional findings from the consultation in Autumn/Winter 2011

Considerations about level of support

Quality assurance and challenge function

- Challenge/critical friend – a pool of people that could be identified as fulfilling this role.
- Useful to know whether you're asking the right questions.
- Face-to-face interaction will be an important element to ensure rigour and quality of the evaluation evidence submitted.
- Commission local organisations (e.g. CVS) to provide the support.

Knowledge sharing

- Comments page/forum/drop box on the website – to share and engage with others
- Media – which resources/publications do funders read? In order that organisations can share/demonstrate their impact.

Do you record statistics about your project?

Community Links survey results (number of respondents = 107, of which 85 attended workshops). See Figure 1 on page 85.

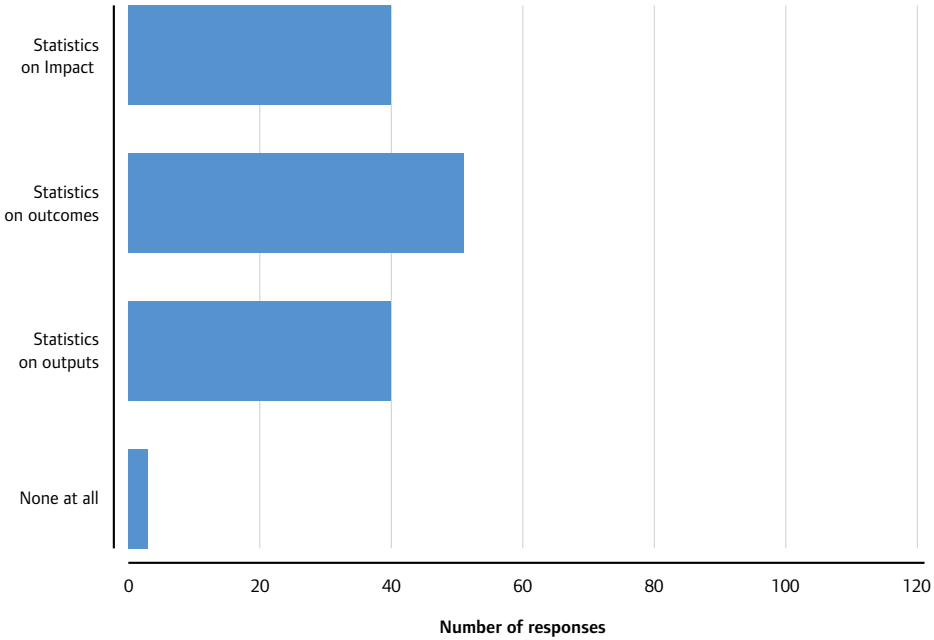
Range of organisations represented

- The largest group (41 per cent) of respondents worked for medium-sized projects employing between four and ten staff, with 19 per cent representing small projects of three staff or less, and 30 per cent from large projects with more than ten staff. 30 per cent worked in just one site, but 25 per cent worked in more than ten sites, with the remaining 37 per cent working in two to ten sites.

Current evaluation efforts

- Fifty-five per cent evaluate their project themselves; while 28 per cent use an external evaluator to complement their own efforts, and 6 per cent just rely on external evaluation. Only 11 per cent don't evaluate their project at all, and they would all like to.
- As would be expected, 95 per cent of organisations record statistics on outputs, dropping to 50 per cent who record outcomes, and only 39 per cent who look at impact – probably reflecting the ease of measuring each one. It is worth remembering

Figure 1: Do you record statistics about your project?
Community Links survey results (number of respondents = 107, of which 85 attended workshops).



these are self-reported, so an organisation that reports measuring impact may not be doing so according to a strict standard, for example.

- When asked about the project itself, 74 per cent said they had a 'clearly defined long-term goal, written down', while only 32 per cent had an evaluation plan. About half the projects had clearly identified their outcomes (64 per cent), indicators (45 per cent), interventions (65 per cent), target populations (63 per cent) and resources (52 per cent). And while the bigger projects did better than the smaller ones, 62 per cent of the biggest projects (with more than ten staff) still didn't have an evaluation plan, only just behind the smallest projects on 70 per cent.

Future evaluation?

- Reassuringly, 96 per cent of respondents would like to get better at evaluation, although there was considerable selection bias, since those who are keen are also more likely to respond to a survey or attend a workshop.

- Lack of money (64 per cent) and time (72 per cent) were identified as the largest barriers to evaluation, compared to lack of expertise and knowledge (49 per cent).

Appendix 2: list of organisations consulted

- Action for Children
- Active Communities
- Arc Theatre
- Blenheim CDP
- Brathay Hall Trust
- Business in the Community
- BVSC
- Capital Conflict Management
- Catch 22
- Chance UK
- Change Makers
- Children and Young People's Service London Borough of Lambeth
- Community Links
- Connexions
- Coram
- Cricket Foundation
- Cricklewood Homeless Concern
- DASL
- Dfuse
- Discovery Initiative
- DKH
- Family-Action
- Fegans Child and Family Care
- FNF
- For Futures Worth Developing
- GLA
- Glypt
- Griffinrc
- Ground Up Development
- Holborn Community
- Hope UK
- Independent
- Inspira consulting ltd
- Inspire!
- Islington Community Safety Board
- Islington Council
- Khulisa
- Kickz (Metropolitan Police)
- Kids Count

- Kidz4mation
- Leap Confronting Conflict
- Lewisham Crime Reduction Service
- London Action Trust
- London Borough of Bexley
- London Borough of Hammersmith and Fulham
- London Borough of Islington
- London Borough of Southwark
- London Borough of Waltham Forest
- London Fire Brigade
- Mayor's Fund for London
- Metropolitan Police various projects
- Music and Change
- National Family Mediation
- NCY Trust
- Non Budget Films
- Novina Research & Consultancy
- One Housing Group
- Poplar HARCA
- Principal
- Prospects Services
- Red Balloon North West London
- Safer London Foundation
- Shelter
- Skill force
- Street League
- Studio 3 Arts
- Substance
- SVI
- The Complete Works
- The Rathbone Centre
- The Weir Link
- Through Unity
- Tottenham Hotspur
- Working with Men
- YES Project
- Youth Justice Board

Appendix 3: Further details – project case studies

London Youth: Positive Change project

London Youth is a vibrant network of 400 community organisations serving young people and their families in every London borough.

Level One: Project theory/logic
Positive Change aims to build the capacity of youth centres to tackle gang problems and youth offending in London. Developed in early 2009, it drew on three existing models: Emmanuel Youth Project's existing initiative 'Identity'; Chance2Change's use of Cognitive Behavioural Therapy (CBT) with young offenders; and the Keyfund model of developing skills and expertise. From the outset it had three explicit aims:

1. To reduce offending, anti-social behaviour and negative engagement in gangs.

2. To increase take-up of positive activities.

3. To re-engage young people with, employment, education or training.

The logic behind Positive Change is that young people at risk of offending need a range of interventions delivered locally by skilled youth workers. They need support to address their offending behaviour, tackle risk factors and build resilience. They also need constructive activities that will build their practical skills and confidence, and divert them from offending. Finally, they need information, advice and guidance to help them progress to education, employment or training opportunities. However, in order to provide all this effectively, it is first necessary to build the capacity of youth clubs, and train youth workers to work in a particular way.

This logical approach is reflected in the programme's structure, which works at five incremental stages:

- Build the capacity of five youth clubs.

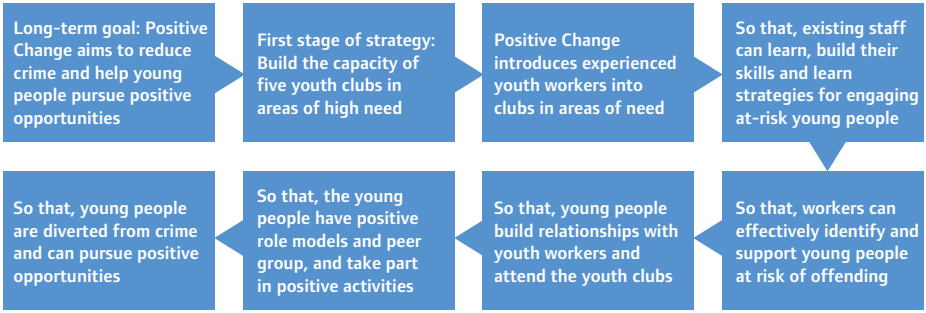
- Engage young people at risk of crime.

- Provide Cognitive Behavioural Therapy (CBT) to young offenders.

- Develop the skills of young people by helping them complete community projects.

- Help young people progress to education, employment or training.

Each stage contains a number of activities and outcomes, linked to the final aims. This is too complex to illustrate in full; an overview of the project's logic model as articulated in the 2009 business plan is included in Appendix 3. However, individual activities can be broken down into a chain linking strategy to activities and outcomes. For example:



Target population
Positive Change had two initial target populations:

1. Five youth clubs in Lambeth, Lewisham, Greenwich, Tower Hamlets and Kensington and Chelsea; sites selected on the basis of evidence of local gang problems, institutional capacity, and proven ability to work in partnership.

2. 2,000 young people aged 13–19 in London over two years. **There were no strict referral criteria.**

After the project developed, a third target population emerged:

3. Offenders in Portland YOI who were being resettled into London. Instead of delivering the CBT locally around the five youth clubs, the project provided the therapy to offenders in the YOI prior to release, and then linked them up with youth clubs in the areas that they were returning to in order to improve their re-integration into the community.

Assumptions
The plan behind Positive Change was extremely ambitious. It aimed to prove the model very quickly and then scale up to support five further youth clubs within two years. This assumed that the necessary structures, processes and funding would be in place very quickly. There was also an assumption that the interventions on which Positive Change is built on were robust, well evaluated and could be replicated in different settings.

Level 2: Project Evaluation Planning
An intention to conduct the evaluation was set out in the business plan at the start of the project: 'The overall approach is to capture outcomes and evaluate young people's journeys at every stage of delivery through a balance of qualitative and quantitative and formal and informal approaches.'

The plan for doing this was also sketched out at a high level: 'Measuring longer-term impact in terms of reduction of young people involved in anti-social behaviour, gangs and offending and increasing employability will be based on local authority baseline data followed by sampling of participants' lifestyles and choices at six monthly intervals.'

However, practical challenges meant that the project has struggled to collect data on its outcomes systematically. Project work began before databases and an evaluation framework had been fully set up.

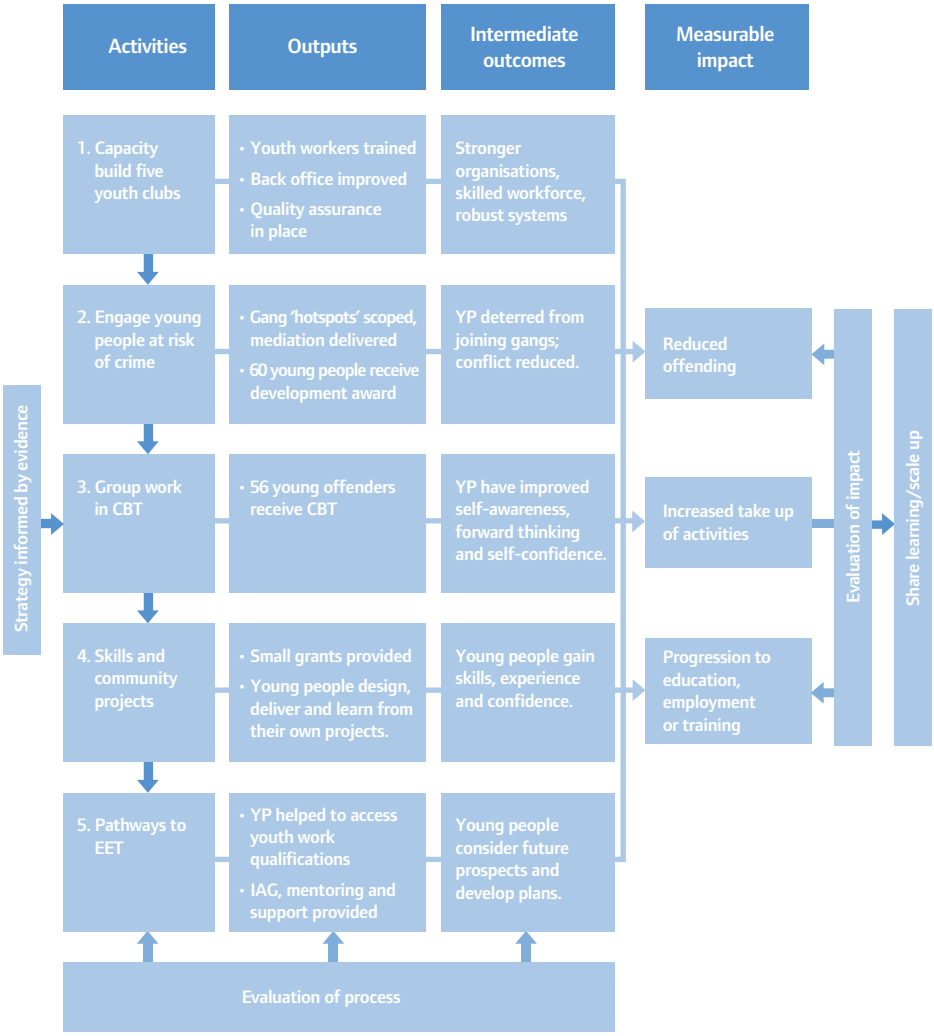
Nonetheless, Positive Change has emerging evidence of its impact. For example, early indications from the CBT work in Portland YOI show that of 37 young people who received CBT and have been released, only three (8 per cent) have re-offended so far (it is seven months since their release at the time of writing). It is too early to determine whether this can be sustained, however, it can be roughly compared with the fact that 75 per cent of young offenders re-offend within one year of release.

One challenge is how such improvements can be attributed to the role of Positive Change, rather than to other factors. For example, without comparing the outcomes of those that received CBT with a control sample that did not, it is not possible to determine the extent to which the benefits can be attributed to this project in particular (rather than other factors, such as YOT interventions, or improved housing etc.).

In addition to focusing on reduced re-offending, the project could 'positively frame' its outcomes, as the *Project Oracle* guidance recommends. This means that rather than focusing on the reduction of something negative (e.g., re-offending), the project could instead capture the improvements in young people that are associated with reduced re-offending (e.g., improved attitude). CBT is typically evaluated through attitudinal questionnaires, observation or learning outcomes, so the project could capture these outcomes to prove its impact.

In order to achieve Level 2, Positive Change would need to create a plan for capturing evidence of its impact systematically. This would involve outlining the aims of the project, and the intermediate and final outcomes that it achieves. The plan should include indicators that show how outcomes would be measured, and a framework for how data would be captured. An abbreviated example is given on page 91.

Desired outcome	Indicators	Data capture
Young offenders leaving Portland YOI who receive CBT from Positive Change have reduced risk of reoffending	Improved attitude and resilience compared with baseline	Attitudinal questionnaire every three months
	Reduced reconviction rate as compared with matched cohort	Data provided by Police National Computer



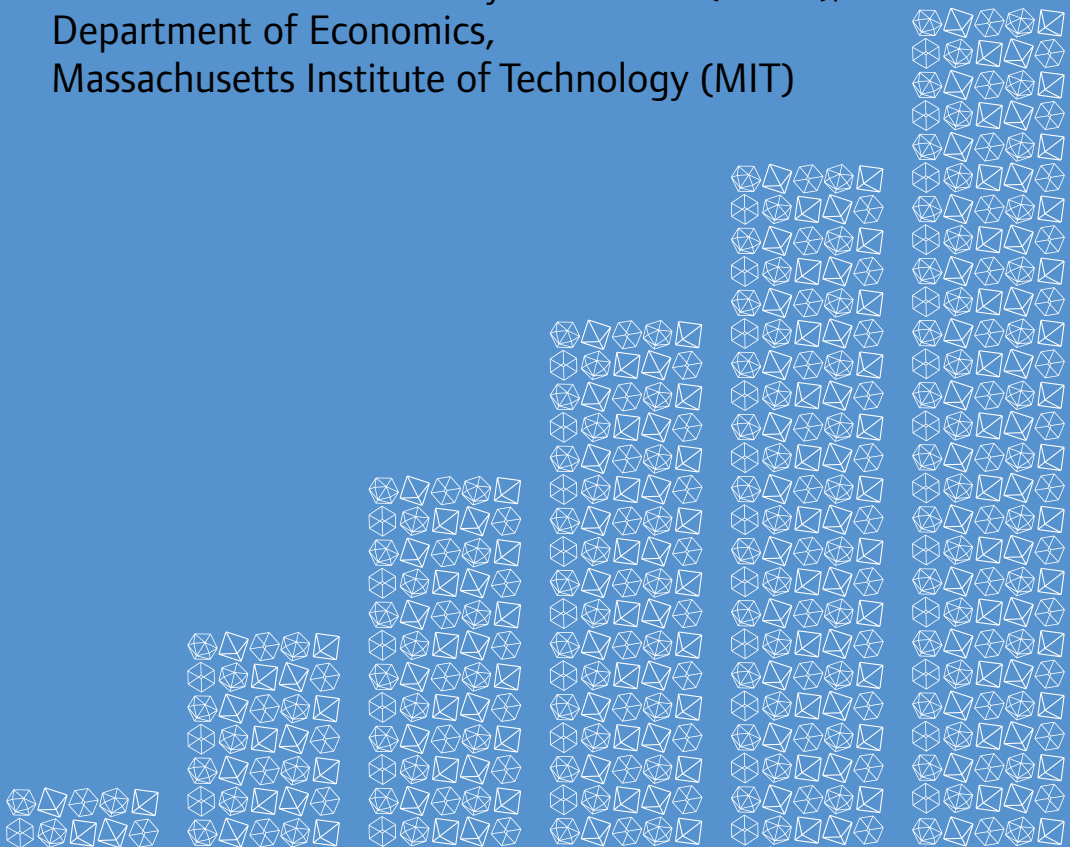
Endnotes

1. Allen, G. (2011) 'Early Intervention: The Next Steps.' London: Crown Copyright.
2. Centre for Social Justice (2011) 'Outcome-Based Government: How to improve spending decisions across government.' London: Centre for Social Justice.
3. Bunt, L. and Harris, M. (2010) 'Mass Localism: a way to help small communities solve big social challenges.' London: NESTA.
4. Mayor of London (2008) 'Time for Action: Equipping Young People for the Future and Preventing Violence – the Mayor's proposals and call to partners.' London: Greater London Authority.
5. Mayor of London (2008) 'Time for Action: Equipping Young People for the Future and Preventing Violence – the Mayor's proposals and call to partners.' London: Greater London Authority.
6. Specifically London-based programmes aimed at benefiting children and young people.
7. Bloom, H. (2010) 'Nine Lessons About Doing Social Research.' New York, NY: MDRC. Available at: http://www.mdrc.org/publications/575/presentation.html
8. Bunt, L. and Puttick, R. (2010) 'Ten Steps to Transformation.' NESTA Blog. Available at: http://www.nesta.org.uk/blogs/ten_steps_to_transformation/

From Research to Policy

Using evidence to inform development policy

Iqbal Dhaliwal and Caitlin Tulloch
Abdul Latif Jameel Poverty Action Lab (J-PAL),
Department of Economics,
Massachusetts Institute of Technology (MIT)



Abstract

Evidence from impact evaluations can help policymakers identify and select the programs that are most effective at achieving policy goals. The last few years have seen a tremendous growth in this body of research as well as an increasing trend among policymakers towards considering rigorous evidence while making key policy decisions and program choices. This paper argues that there is greater scope for incorporating results from research into policy decisions and program design, even in the presence of political and administrative constraints. We draw on our work in the development field, discussions with numerous policymakers around the world, and the experiences of many of our colleagues at the Abdul Latif Jameel Poverty Action Lab (J-PAL) to discuss some of the ways in which evidence is currently incorporated into policy decisions and the constraints that we have observed on wider adoption of evidence-based policy. We then suggest some specific ways that stronger policy-research partnerships can help overcome these constraints. We illustrate many of these issues with examples from J-PAL's policy outreach work.

1. Introduction

Billions of dollars are spent every year on development policies and programs, but until recently there was relatively little rigorous evidence on the true impact these programs have on the lives of the poor. Different programs (e.g. new school buildings or cash incentives for students) targeted at the same policy outcome (e.g. improving student attendance) can have very different results, but without clear evidence on their final impact there is little guidance for policymakers on which program to choose or what the most effective policies are. This is in part because it is very difficult to attribute changes in peoples' lives to a particular program, rather than external factors or other concurrent programs.

In recent years, randomized impact evaluations of social programs have emerged as a robust tool for generating evidence to guide policy, because they measure impact by estimating the difference in outcomes, like student test scores or immunization rates between a randomly selected treatment group that received a program and a comparison group that did not. Random assignment of members into a treatment and comparison group ensures that the two groups do not differ systematically at the start of the program, and thus any differences that subsequently arise in the outcomes between them can be attributed to the program rather than to other factors. Because they allow precise measurement of impact, randomized evaluations can help policymakers identify programs that work and those that do not, so that effective programs can be promoted and ineffective ones can be discontinued. In this way, evidence can improve outcomes for the poor in two ways. First, existing funds can be spent on programs that have a greater impact on the lives of the poor, which helps policymakers get a bigger impact for their development spending. Second, this in turn may lead to greater public support and funding commitments from donors for programs that have been evaluated and found to be effective.

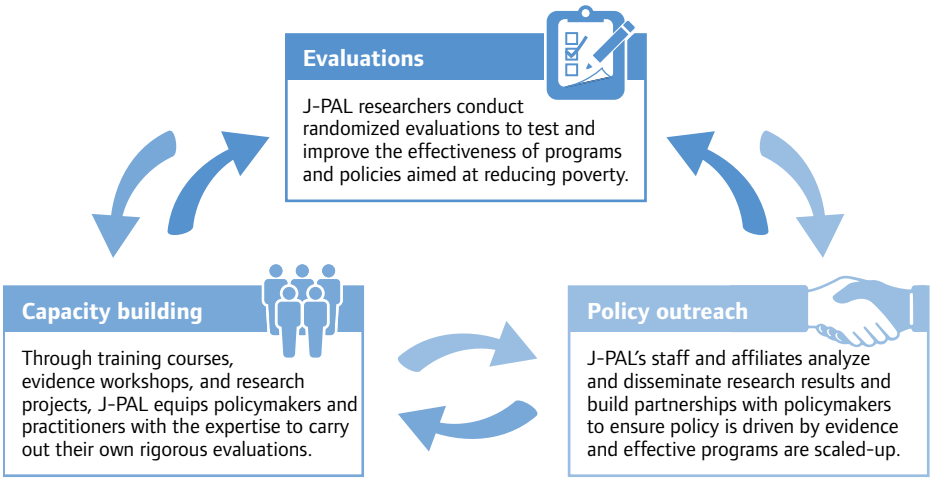
In recent years there has been an increasing trend to incorporate such evidence in policy decisions. But despite the potential benefits from evidence-based policy, we believe that there is still greater scope to incorporate evidence from rigorous evaluations in policy decisions. This potential for increased use of evidence in policy remains for several reasons. First, while the use of rigorous evaluations in assessing development programs has increased exponentially in the past few years, they still cover only a very small subset of the development programs that could benefit from such evaluations. Second, this relative scarcity of rigorous evidence is compounded by the technical language in which it is often presented (typically in academic journals) and the need for greater effort to communicate the results to a policy audience. Third, the growing body of evidence itself poses a dilemma – what weights should policymakers attach to different research studies, especially if there is not clear guidance on how to relate new evidence to the existing body of evidence and draw out general policy lessons. While some development

organizations (broadly defined as governments, NGOs, foundations, and international development organizations working to reduce poverty) have the technical capacity and a mandate to incorporate evidence in their decisions, these constraints can still make evidence hard to access for policymakers, substantial numbers of whom continue to rely on other factors when choosing programs to implement.

When deciding between alternative programs, policymakers have to take into account a number of factors including political constraints, administrative capacity, technical feasibility, time pressures, and limited budgets. We do not believe that these considerations can or should be abandoned in favor of purely evidence-based policy. These considerations will continue to be salient for policy decisions, and addressing them is in fact crucial for the successful implementation of any program. But we argue that there is much greater scope for incorporating evidence in decision making, even in the presence of such constraints, and that closer partnerships between researchers and policymakers can foster more evidence-based policy. Researchers can work to make their results more accessible and understandable to a practitioner audience so that policymakers have the information necessary to make choices informed by evidence. At the same time, policymakers can partner with researchers to measure the impact of their development programs, encouraging researchers to examine the most relevant and pressing problems, and thereby generating more rigorous evidence from the field and creating a virtuous feedback loop. Policymakers can also provide local knowledge, helping researchers understand which policy problems are most pressing and in need of evaluation, thus ensuring that rigorous evaluations address relevant, current issues.

In discussing potential ways to improve the scope of evidence-based policy, we draw on our own experiences as well as the work of the Abdul Latif Jameel Poverty Action Lab (J-PAL). J-PAL was founded in 2003 as a center in the economics department of the Massachusetts Institute of Technology (MIT) with a mission *"to ensure that policy is based on scientific evidence and research is translated into action."*[1] With regional centers in Africa (University of Cape Town), Europe (Paris School of Economics), Latin America (Pontificia Universidad Católica de Chile) and South Asia (Institute for Financial Management and Research in India), J-PAL's network of 59 researchers from 30 universities worldwide has undertaken more than 270 randomized evaluations of development programs in 42 countries. J-PAL has also trained hundreds of researchers and policymakers on how to conduct randomized evaluations, and has been active in disseminating research results about 'what works' to policymakers around the world. The chart on page 97 describes J-PAL's three core activities – research, training, and policy outreach – and how they relate to each other.

Figure 1: J-PAL's mission is to ensure that policy is driven by evidence and research is translated into action



Source: http://www.povertyactionlab.org

As part of our policy outreach activities, J-PAL researchers and staff have worked with hundreds of policymakers in governments, international development organizations, foundations, and NGOs. This paper is based on the collective experiences of a number of our colleagues at J-PAL in working with policymakers, as well as interviews and discussions with policymakers at many of the organizations that J-PAL partners with. Throughout the paper, we will consider the constraints that policymakers appear to face in utilizing research in their decision making and suggest some specific ways that stronger policy-research partnerships can be designed to help overcome these constraints. These suggestions are supported by examples from the work done by J-PAL's Policy Group to promote partnerships between researchers and policymakers and disseminate research results in order to improve policy decisions.

However, as we realize almost daily in our work, policymaking in general, and the use of evidence in policy making in particular, are complex issues. We do not claim or believe that J-PAL's strategy for policy outreach, as described in this paper, is either the only or the best way to inform policy – many organizations have had significant policy influence in their fields and may have followed a completely different path to success. This paper is not a literature review of studies analyzing the role of scientific evidence in policymaking, nor have we done rigorous empirical evaluation of the constraints on

policymaking or tested alternative approaches to influencing policy. Rather, this paper is an attempt to share our current interpretation, based on our experience of the past few years, of the gaps that we have observed between research and policymaking, and to present what are, in our opinion, a few useful strategies for bridging this divide. Our hope is that this can be a useful tool to spur dialogue and identify successful policy outreach strategies among other organizations whose mission is similar to ours: *"Translating research into action."*

Similarly, there are a number of research methodologies that provide scientific evidence to inform development policy. For instance, rigorous cross-country regressions can shed light on the need for changing colonial legacies and institutions to ensure new programs and policies succeed. Qualitative surveys are critical in understanding the reasons why parents do not send their children to school or why the poor do not use free government clinics – essential first steps in designing an effective program to tackle these problems. We do not claim or believe that randomized evaluations are the only way to generate rigorous evidence that can help inform policy – many of these other research methodologies produce reliable and useful evidence. But for reasons discussed in this paper, we believe that randomized evaluations are well-suited to provide rigorous evidence that lends itself particularly well to informing policy.

2. Potential and actual use of evidence in policymaking

In this section we will highlight some of the key considerations that factor in policymaking, describe how policymakers can benefit from the results of rigorous field research, and discuss some of the constraints that prevent evidence from being used in policymaking.

a. Policy decisions are made based on many considerations

Development organizations spend billions of dollars annually on poverty alleviation programs like free health clinics, rural schools, agricultural training, and microfinance. One of the greatest challenges facing these organizations is how to choose programs and policies that will have the biggest possible impact on poverty. There can be many reasons why a particular program is not the best suited to maximize impact – for example, if the needs of the intended beneficiaries were not correctly assessed; if a flawed program design was not based on a well laid out theory of change; if there is inadequate monitoring to ensure that programs are being implemented as designed; or because of an inability to measure the impacts of programs or compare them to other options before choosing which program to implement or continue.

While there are many examples of well-designed and well-executed policies, and many others that need no further evidence (e.g. the benefits of immunization for children are well-established) we believe that there is significant potential for much more of development policy worldwide to be driven by rigorous evidence on program effectiveness, rather than by factors such as ideology, inertia, ignorance, or instincts. For example, suppose a policymaker wants to find the best intervention to increase school attendance. Is the solution better buildings, teacher training, community monitoring, treating children for intestinal worms, free lunches, conditional cash transfers, or something else? Often such decisions get made based on a combination of ideology (*"we should not give parents 'bribes' to do the right thing for their children"*), ignorance (*"treating children for intestinal worms is a health intervention and has nothing to do with school attendance"*), inertia (*"we have already spent millions of dollars in the past ten years on improving school buildings, how can we stop midway?"*), instincts (*"I have a feeling that if teachers were better trained it would lead to increased attendance"*) or incomplete information (*"community participation increases peoples' feeling of empowerment, so let's encourage that in schools as a way to motivate them to increase their children's attendance."*)

All this can lead to programs and policies that are sub-optimal in terms of their impacts, their costs, or both. The problem of 'bad' policy choices is further exacerbated by the fact that once launched, programs are very hard to roll back, even if they are later found to be failures. This is not only because of the political and personal difficulties for the program architects, sponsors, and implementers in accepting failure, but also because of lobbying by entrenched and powerful interest groups like contractors, service providers, bureaucrats, and local politicians who benefit from program funding.

Despite the large potential for more evidence-based policy in development organizations worldwide, we are encouraged by how many policymakers are increasingly willing and often keen to consider evidence in decision making when that evidence is relevant, accessible, and presented in a format that is easy to understand. This trend of increasing interest in evidence has been accelerated by demands for greater accountability and bigger impact from development programs by civil society in many developing countries, as well as increased scrutiny of development spending by international development agencies and foundations in donor countries. But even when policymakers are willing and eager to incorporate evidence in decision making they face the challenge of often not knowing where to find such actionable evidence, since most such results are either presented at research conferences or published in academic journals.[2] When they do find rigorous evidence, policymakers may often have difficulty interpreting it because it is mostly written for academic audiences in technical language. And even if they do access relevant rigorous evidence that they can easily interpret, policymakers may still face the challenge of synthesizing evidence and drawing policy lessons from multiple research studies, which may have been performed in very different contexts and in different years.

Policymakers can also find themselves faced with the dilemma of studies with seemingly conflicting results – can they afford to base their program choices on certain evidence when it is quite possible that another study would find evidence that contradicts the first, and cause them to reverse their original decisions?

Because of these complicating factors, there is currently a wide variation in the use of evidence in policymaking. At one end of the spectrum are some international development organizations and large foundations with full-time technical staff who analyze various research studies to inform their organization's decisions. In recent years, these organizations have shifted to using the results of rigorous and independent impact evaluations from the field in making funding decisions, rather than relying solely on qualitative surveys, anecdotes, or less scientific evidence (e.g. a before-after analysis of beneficiaries, or comparisons of final outcomes of beneficiaries to non-beneficiaries). At the other end of the spectrum are a number of governments, NGOs, and foundations, both in developed and developing countries that may not have access to or the technical resources to analyze evidence from rigorous research studies.

Thus a key challenge for organizations like J-PAL that would like to see closer links between policy and research, is to somehow ensure that the growing body of scientific evidence from around the world can be effectively used to inform policy in three distinct ways. First, the continuation of existing programs can be reviewed in light of evidence on effectiveness from other programs, and changes to program design undertaken based on such evidence. The impact of these changes should be similarly evaluated to increase the body of evidence even further. Second, new programs (including multiple variations of a program's design) can be piloted and their outcomes evaluated before they are launched on a massive scale. Third, innovative programs that have been found to have large impacts on outcomes can be replicated or scaled-up. Later sections of this paper discuss different policy strategies that can help achieve these objectives.

b. Policymaking can benefit from using evidence from randomized evaluations

There are many benefits to policymakers of partnering with academic researchers on randomized evaluations of their programs. Academic researchers whose work is published in peer-reviewed journals in general bring a high level of objectivity to an evaluation, which in turn provides greater credibility for the results. They also have strong technical skills to design rigorous evaluations, and understand the research literature on particular topics. Further, for a number of reasons discussed below, randomized evaluations are a particularly well-suited tool to provide critical insights to policymakers:

- **Randomized evaluations measure impact rigorously:** Randomized evaluations compare the outcome of interest (e.g. test scores in schools) of beneficiaries who

received a program (the treatment group) to another group (the comparison group) that is similar in all respects except that it did not receive the program (e.g. free textbooks for students). Measuring outcomes in this comparison group is as close as one can get to measuring how program participants would have fared without the program (the counterfactual). Randomized evaluations are often considered the gold standard of impact evaluations because they generate a comparison group that is statistically identical to the treatment group, by randomly assigning the targeted population into the treatment and comparison groups. Random assignment prevents possible differences between the treatment and comparison groups, which could arise if, for example, a certain type of person was allocated to or chose to opt into a program, or if either the treatment or comparison groups had specific unobservable characteristics that systematically differentiated them from each other.

This random assignment is similar in methodology to what is used in clinical trials to determine who gets a new medicine versus a placebo when testing the effectiveness of new drugs. Because members of the treatment and comparison groups do not differ systematically at the outset of the experiment in both observable and unobservable characteristics, any difference that subsequently arises between them can be attributed to the program rather than to other factors (i.e. randomized evaluations generate unbiased estimates of program impact). The rigor of randomized evaluations in measuring the impact caused by a program can give policymakers confidence in using this evidence to make important decision like whether to scale-up a program, make design changes, or discontinue it.

- **Rigorous evaluations can provide key insight on why programs succeeded or failed:** Rigorous evaluations can be valuable because they provide information beyond just whether a program worked, shedding light on how and why an impact was seen. Researchers can design evaluations in a way that also yields useful process information, including administrative data and data on intermediate outcomes (e.g. attendance of health staff or number of beneficiaries who attend an immunization camp) via quantitative or qualitative surveys. Such data can provide valuable insights on the processes underlying the program and allow researchers to look beyond estimates of program impact to see why a given impact is being observed or not.

  Take, for example, a project in India that evaluated nurses' attendance at work by installing time clocks and providing incentives based on attendance.[3] After 16 months, there was no discernible impact of the program on nurse's attendance, in contrast to previous studies that had shown that incentives for attendance could have a large and significant impact. Throughout the evaluation, researchers collected administrative data on when exemptions were given to nurses who missed work. These data, as well as qualitative interviews, explained that the program was extremely unpopular with nurses, who over time petitioned their supervisors for

more and more exemptions. Because supervisors could be pressured into relaxing the conditions of the incentives, there was no improvement in behavior. The results of this evaluation suggest that future programs offering incentives for service provider attendance should ensure that incentives are automatically applied without giving blanket discretion to supervisors to relax conditions.

Randomized evaluations often also include evaluations of multiple variations of a large program, allowing policymakers to understand which components are driving program effectiveness. For example, in an evaluation of the Kenyan 'Extra Teacher Program,' researchers randomly assigned some primary school classes to be taught by contract teachers instead of civil service teachers, and also randomly assigned a subset of classes to be tracked by student ability. They found that providing additional contract teachers to create smaller class sizes, and grouping these classes by initial ability caused the greatest increases in student achievement.[4] This evaluation shed light on several policy ideas at once – smaller class sizes, tracking by ability, and contract teachers – allowing policymakers to understand which of the many program components was responsible for the impact.

- **Field trials provide practical information to help facilitate and guide scale-ups:** Because randomized field trials are performed in real-world situations, often with implementing partners who would consider adopting and expanding the program if it were proven effective, they can yield many valuable practical insights beyond simple estimates of program effectiveness:

  First, since many evaluation designs also entail conducting qualitative surveys and measuring intermediate outcomes (e.g. measuring health staff attendance as part of a program that ultimately aims to improve health outcomes like infant mortality), evaluations can provide critical feedback to the implementer at the pilot stage. This feedback can allow implementers to make rapid adjustments to the way the pilot is being implemented to prevent the program from failing for avoidable reasons.

  Second, evaluations can also help shed light on the role of constraints like staff capacity to manage a new program, the enforceability of incentives and penalties on politically powerful service providers, or the inability of local suppliers to provide sufficient inputs like vaccines or food. This provides invaluable information for troubleshooting a large-scale version of a program before it is rolled out.

  Third, because these field evaluations are frequently run on a relatively small scale in a few sub-districts or districts, they entail much lower costs and risk than launching a massive state- or nation-wide program whose effects are still unknown and which, once scaled-up, would be hard to discontinue. The additional costs of implementing an evaluation are small relative to the losses from a wide-scale launch of an

ineffective program and its continuation for years at the expense of other programs that could have had a bigger impact, lower cost, or both.

Fourth, when a proper evaluation is performed, it can provide critical proof-of-concept for an innovative new program, which can be crucial to get political, administrative, and financial support for a larger scale-up. For example, evaluations that measure impact and collect cost data for the underlying program allow for the construction of cost-effectiveness estimates that help compare the program to other policy options, and can also help in the creation of budgets for scale-ups. And because independent evaluations provide an objective and widely accepted 'record' of the true impact of a program or a pilot, they can play an important role in institutionalizing change and ensuring continuity in organizations, such as governments, where civil servants are frequently transferred and new administrations try to reset the policies of their predecessors.

- **Randomized evaluations are easier for non-technical audiences to understand:** The randomized methodology, although less common in the social sciences, is nonetheless much more familiar to policymakers than most other research techniques, due to its use in medical trials. Randomized evaluations are also easier to explain than non-experimental or quasi-experimental methods that rely on statistical and econometric methods that often remain a 'black box' for many policymakers. And because the underlying program is actually implemented in the field, it can be relatively easier for the policymakers, implementers, and program beneficiaries to understand the details of the program design, cost and impacts. This can be particularly useful in building support among various stakeholders for a wide scale-up of the program, and for sustaining the program when its initial champions or creators are no longer in power.

c. But even evidence from randomized evaluations can be difficult for policymakers to interpret and use if it is not disseminated well

Even though evidence from randomized evaluations is often easier to understand for the reasons described above, it may not contribute to the policy debate and policy formulation unless it is disseminated well. This can be difficult for several reasons:

- **Evidence is often presented for an academic audience:** Much of the rigorous evidence on program impact is presented in technical papers and journals that are targeted for an academic, rather than a practitioner, audience. Because they are published in economic journals, many papers focus on the underlying economic intuition, econometric techniques used to arrive at unbiased estimates of program impact, and how the observed impact validates or contradicts economic theories. And because these journals focus on issues of economic theory and analysis, rather than

how economic analysis can guide policy, many researchers (and academic journals) deliberately shy away from drawing strong policy conclusions based on evaluation results. Moreover, since the main focus of most research papers is on the design of the study and the results, many facts that most interest policymakers are not covered in sufficient detail for policymakers to draw conclusions for their context. These can include the context of the problem, nature of location, implication of the results for other studies, implementation details, issues faced by the program implementer, and program costs.

- **It is difficult to distinguish between evidence of different quality:** When evidence is available, it should ideally be judged according to the scientific rigor of the methodology that produced it, so its policy conclusions can be given an appropriate amount of confidence. But in our experience, the distinction between evidence of different quality often requires significant technical expertise to discern. It also requires non-trivial time and effort in researching, things that many policymakers do not have. As a result, many policymakers require assistance to understand which evidence is a reliable and rigorous guide for policy.

  For example, a policymaker working at a large European government department dealing with development issues indicated that because her[5] colleagues were unfamiliar with the various methodologies for generating evidence, they did not perceive any particular advantage to evidence from rigorous evaluation methods as compared to 'homemade evaluations' based on administrative data or unscientific surveys. As a result, rigorous evidence from randomized evaluations sometimes received a 'cold reception' because the policymakers could not appreciate the difference in quality between different types of evidence. Consequently, even when results were explained using non-technical language, it was very hard to convey the strength of a piece of experimental evidence relative to judgments based on ideology or other non-experimental methods.

  Lack of understanding of what constitutes rigorous evidence can have an even more pernicious effect – it can deter the use of rigorous evidence and perpetuate the use of less rigorous evidence that is selected to suit immediate needs. For instance, a large development organization that competes with many other agencies for funding from the US government performs many rigorous evaluations of its programs to calculate its rate of return on program investments and to facilitate organizational learning. But this also means that they find evidence that some of their programs are not actually having the intended impact. Because the government does not require rigorous impact evaluations of all of its grantees, most of the other competing agencies do not conduct such evaluations and end up presenting less rigorous evidence as indications that their programs are effective. This makes the more

evidence-based organization's work look less effective, and can make it harder for them to attract funds from the government.

- **It can be difficult to compare evidence from different contexts:** It can be difficult to draw robust policy conclusions when comparing evidence from different countries. For example, when presented with evidence of the impact of deworming on reducing student attendance in Kenya, policymakers in India are likely to initially discount the policy lessons because of the differences between the two countries in terms of culture, institutions, infrastructure, geography, history, disease patterns, and incomes. In such a case it would be necessary for someone with an understanding of both the Kenyan study and the Indian context to explain to the Indian policymakers how, on the key drivers of impact, the two countries share similar features that make it likely that a similar impact would be observed. In this example, the similarity in the worm load in the relevant regions of both countries, the absence of footwear use among children, the prevalence of anemia linked to worms, low student attendance, easy availability of low-cost deworming pills, and an established school network that can effectively deliver deworming drugs all suggest that deworming could effectively address a similar problem in India.

- **It can be complicated to compare evidence from alternative programs aimed at the same policy outcome:** One of the biggest challenges in development policy is that policymakers are required to make choices between a number of *prima facie* 'good' options. For instance, it is not immediately clear what is the best way to reduce diarrheal diseases in rural settings where it is prohibitively expensive to provide piped water. Is it by distributing chlorine at schools, supplying it directly at home, or making it available at the source of water supply (a spring, well, etc.)? Should it be supplied at full cost, at a subsidy, or totally free? Or is it more effective to run an information campaign on the importance of hand washing and supply free soap to households? All of these options come with robust theories of change, passionate advocates, and anecdotal, qualitative and even rigorous experimental and non-experimental evidence that supports the common-sense logic of these approaches. In such situations, the challenge for a decision maker is to choose the program that will have the biggest impact at the lowest cost – i.e. the most cost-effective program. But there are dozens of complicated assumptions and decisions that need to be made in creating a formal cost-effectiveness analysis, and this deters policymakers from conducting these analyses, even when rigorous evidence is available.[6]

- **Demand for evidence is time-sensitive:** Another factor complicating the use of evidence in policy decisions is the relatively short time span over which policy decisions are commonly made. An evaluation and the associated academic paper can take two or more years to complete, but in most cases policymakers are working

with much shorter time spans, and this can make both using and producing evidence difficult for two main reasons. First, if policymakers in an organization are not intimately familiar with the evidence on a particular subject, it may take a while to gather the facts necessary to inform a particular decision and network with all of the relevant decision makers to get them on board. During this process they may wish to consult with the researchers from the original study, outside experts, or other practitioners who have run similar programs, but the limited turnaround time on decisions can limit their capacity to do so.

Second, if an evaluation is commissioned to deal with a particular question that will inform later decisions, the evaluation will need to be timed so that the results coincide with the decision making process, otherwise the results will not come in time to have any influence on policy. In France, for example, the Ministry of Education is currently evaluating the 'Boarding Schools of Excellence' program, which gives boarding school spots to children from disadvantaged backgrounds. This evaluation was begun in September of 2009, and results are expected in early 2012. But the Ministry of Education is already in discussions with the Ministry of Budgeting over allocations for the public boarding school system, so results may come too late to influence the policy that will be in place for the next two years.

- **Researchers are constrained in their ability to engage policymakers on a continuous basis:** Many of the previously described obstacles could be overcome if there was an active and continuous dialogue between researchers and policymakers, but this happens far less than it needs to for a number of reasons. First, there are relatively few researchers doing rigorous evaluations of development programs compared to the large number of policymakers in governments, NGOs, foundations and international development organizations around the world who could benefit from their findings. Given the specialization of researchers among sectors like education, health or governance, the effective pool of researchers in a particular sector is even smaller. This creates significant demands on the time of these researchers to discuss their findings with policymakers from various development organizations, and puts a constraint on the number of policy outreach requests that they can respond to effectively.

  Second, researchers are not rewarded within their profession for presenting results to practitioners or explicitly trying to inform policy. In particular, those who have yet to get tenure face strong penalties if they take time off from producing papers to invest in dissemination. Surprisingly, this sentiment can even extend to researchers at policy organizations. A full-time professional researcher at an international development organization who has done important work on poverty issues told us that she felt her role was to do her research and it was up to the policymakers to read, understand and interpret it whichever way they liked. This idea that research findings will gradually

percolate into the policy making dialogue, sometimes called the 'Enlightenment Model' of research utilization, is held by many researchers and could help explain why many of them do not undertake active dissemination of their results (Weiss 1979). This can create barriers to the use of evidence to inform decisions, if researchers are not available to discuss the particulars of their research, or interpret their findings for interested practitioners.

Third, the comparative advantage and training of many researchers, especially those in academia, is in doing rigorous research, writing papers for academic journals, and explaining their results to a technical audience. Informing policy requires a number of activities that researchers are not trained in and often do not have an inclination for, including identifying the best organizations to partner with, networking with key policymakers, presenting results in a non-technical manner, and relating the policy implications of their research.

Fourth, informing and influencing policy is an effort-intensive, risky, and long-term commitment for researchers. It takes a long time to build relationships with policymakers and navigate all of the procedural and bureaucratic hurdles to get project approvals. There are numerous risks of such partnerships, including the difficulty in overcoming political hesitation to publish impact results (especially when they show the program in a poor light), and the often small and incremental nature of the policy changes that follows from such efforts. But perhaps one of the biggest risks that researchers face is that of frequent transfers of key officials in governments and the subsequent 'orphaning' or discontinuation of a project when its policy champions are transferred (or administrations are voted out), or the reset of efforts to encourage evidence-based policy. Senior officials in developing countries responsible for key policy decisions are transferred very frequently. For instance, a study found that members of India's top civil service – the Indian Administrative Service (IAS) – have an average tenure of just 16 months, and among those posted as chiefs of district administration (perhaps the most important positions for implementing development programs in the field) almost half had a tenure of less than one year.[7] These frequent transfers and the lack of institutionalized mechanisms for knowledge sharing with successors leads to a significant lack of continuity in decision making and fulfillment of prior commitments. Researchers are often frustrated when, after months of building relationships and finally convincing a policymaker about the relevance of certain research results, they have to start from scratch again with their successor who has not received any information about her organization's previous engagement with the researchers. Even more frustrating is when, despite having signed legally binding Memoranda of Understanding (MoUs), a new official is able to effectively terminate a program and associated research by refusing to cooperate with the researchers or by deliberately delaying key activities essential for the success of the study.

All of these factors can hinder the development of strong linkages between researchers and policymakers, and sometimes even lead to a mutual feeling of 'distrust' between the two groups. Researchers may feel that policymakers are not responding sufficiently or quickly enough to what they believe is convincing evidence, while policymakers feel that researchers are too narrowly focused on the 'theoretical, perfect-world' situation with disregard for the practical 'real-world' constraints confronting the policymaker.

3. Promoting evidence-based policy

Evidence from rigorous evaluations can be particularly useful in choosing between alternate programs and informing the design of new programs, which in turn can be evaluated to grow the body of rigorous evidence. But, as the previous section has shown, there are a number of barriers that prevent the utilization of such evidence in policymaking. In this section we attempt to show how policymakers and researchers can partner to better incorporate evidence in decision making. These observations are based on the work that J-PAL has done over the past few years. As stated earlier, we are still learning how to best inform policy, and the strategies described below are not intended to be a comprehensive list of all of the effective ways to inform policy. They are intended as a starting point for discussion with other organizations that share our mission of informing policy with rigorous evidence.

a. Promote a culture of 'evidence-based decision making' at policy organizations

In the past few years, many organizations have moved towards requiring a review of all relevant evidence as part of the formal process of proposing new programs to fund or implement. But this is not always true, and because of the wide range of evidence of varying quality that exists, it can be hard for decision makers to be certain they are giving more weight to rigorous evidence, and easy for some to cherry-pick the evidence that supports their case. Thus merely requiring that evidence be considered is unlikely in itself to lead to a 'real' move towards evidence-based policy. Therefore J-PAL's approach to encourage more informed utilization of evidence has been to build the capacity of policymakers to become better consumers and producers of evidence.

- **Encourage policymakers to become better consumers of evidence:** J-PAL conducts executive education courses every year during which J-PAL's affiliated professors and senior staff discuss the pros and cons of various research methodologies, use case studies to ask participants to critique various research reports, and encourage attendees to design their own evaluation with feedback

from the trainers. To date, more than a thousand participants have been trained at J-PAL courses around the world, including staff of federal and state governments, international development organizations, NGOs and foundations. Many of the participants at these courses have gone on to design their own evaluations, while others are in key decision making roles in development organizations where they are active consumers of research findings.

One large foundation in Europe that J-PAL has worked with has formalized a process that emphasizes inclusion of evidence in funding decisions. Once potential investments are identified and a concept note prepared, the proposal review requires estimation of the program's anticipated impact based on the best available evidence. This encourages staff to find programs that have been proven to be effective, assess the rigor of the evidence being cited, and to estimate the potential impact of new programs before choosing them.

J-PAL also works with organizations that have potential for significant policy influence to design custom courses that help their senior staff become better consumers of evidence. For instance, the National Academy of Administration in India, which trains all of the senior-most civil servants in India, invited J-PAL's Policy Group to train over 300 of members of the Indian Administrative Service (IAS) in using evidence to design and evaluate programs and policies. The trainers included J-PAL's affiliated professors, senior management, and some key policy and implementing partners from past J-PAL projects. They discussed the various methodologies that can be used to generate evidence, how existing evidence can be used in decisions, and offered examples from the Indian context where a research-policy partnership had generated rigorous results.

- **Encourage policymakers to become better producers of evidence:** While examining evidence from previous evaluations is a valuable exercise, there will always be proposed interventions for which evidence (positive or negative) does not previously exist. This may be because a program is entirely new, or puts an innovative spin on an existing idea. Funding and implementing organizations should encourage rigorous impact evaluations of such interventions wherever feasible (e.g. a phased roll-out that creates an opportunity for randomized allocation of the program) and desirable (e.g. an innovative program that has potential to be scaled-up massively but at a significant cost, where no rigorous evidence exists on program impact, where a number of alternative program designs are being considered, or where there is uncertainty regarding the true impact of the program).

  There are a number of ways to encourage such evaluations. First, including funds for monitoring and evaluation in the pilot program budget can help to facilitate both those processes. Second, J-PAL's affiliates and staff, through presentations at policy

conferences, writings and meetings with policymakers, also try to emphasize the need for evidence-driven policy at various steps in this process.[8] Third, many development organizations are also explicitly encouraging their staff to assess if an evaluation would be useful. For instance, the recent USAID Evaluation Policy outlines the conditions when an evaluation of its programs is required, stating that, *"any activity within a project involving untested hypotheses or demonstrating new approaches that are anticipated to be expanded in scale or scope through US Government foreign assistance or other funding sources will, if feasible, undergo an impact evaluation."*[9] USAID has also founded the Development Innovation Ventures (DIV) initiative, which promotes evaluation and adoption of promising new programs and technologies. DIV provides funding for innovative development ideas and helps the Agency rigorously test these ideas to determine which are most effective at helping the poor. The initiative is also focused on taking proven ideas and scaling them up within the Agency, in partnership with innovators and developing countries.[10]

Fourth, J-PAL also encourages organizations to become better producers of evidence through active partnerships on field evaluations. In Haryana, India, for example, J-PAL was approached to help build the capacity of the state's education department to produce and use evidence for decision making. J-PAL is working with them to create an in-house Monitoring and Evaluation (M&E) division. Similarly, J-PAL is partnering with the education NGO, Pratham in a collaborative evaluation of the Mother and Child Literacy Program that will build the capacity of their state directors to do rigorous evaluations of their programs. While all these capacity-building partnerships and custom training courses aimed at making policymakers better consumers and producers of evidence place a significant demand on J-PAL's affiliates and staff, such interaction can lead to strong research-policy partnerships, especially when targeted at decision makers who make large-scale allocation of development spending and program design decisions.

- **Use field evaluations as an opportunity to build strong long-term relationships between policymakers and researchers, while maintaining objectivity in reporting results:** While nearly all research has the potential to influence policy by adding information to the public debate on social and economic issues, the simple existence of evidence does not necessarily mean that the evidence is answering relevant questions, that policymakers are aware of the results, or that they are correctly interpreting these results. Policymakers can be more easily encouraged to use evidence in their decisions when they have closely partnered with researchers in all steps of the evaluation beginning with the process to identify the most relevant and pressing needs, the design of the program to tackle that need, and the evaluation around it. Further, the continuous feedback that policymakers can receive from evaluators in the field can help tackle unanticipated but avoidable implementation roadblocks. Researchers and policymakers can also work to

disseminate the lessons from the program and its associated evaluation to other policymakers so they can benefit from these dual perspectives. Such a collaborative process can encourage evidence-based decision making at policy organizations.

For example, while all evaluations are different, many evaluations by J-PAL affiliates often begin well before the program implementation starts with intense collaboration between researchers and implementers. They may discuss the underlying problem (e.g. unauthorized absenteeism among government healthcare workers), perform a needs assessment for the different stakeholders, and consider various possible solutions (e.g. incentives linked to attendance or better monitoring) along with the associated theory of change for each of those possible solutions (e.g. role of incentives, penalties and intrinsic motivation in changing behavior). As a part of this process, researchers share the results of previous evaluations in that sector, and work with implementers to help design promising interventions to test. Pre-surveys and early small-scale pilots of such interventions can further improve program design by gathering critical information from the field, even before the program is launched.

Such partnerships can be with governments, international development organizations or NGOs as the implementing partners in the field. (Foundations typically act as donors and rarely have field implementation capacity.) There are advantages and disadvantages in partnering with each of these types of organizations:

- *Governments* in developing countries are often the biggest implementers of social programs, and working with them therefore offers the chance to evaluate programs on a much larger scale than any other development organization across almost all sectors of development. They also have the financial, technical, and personnel capacity to widely scale-up the program if it is proven effective in the pilot form. Work with governments can, however, involve long and cumbersome bureaucratic approval processes. There is also a significant risk of projects being discontinued when the civil servants who championed the program are transferred (and transfers are fairly common, as discussed earlier). Wide variation in the skills and enthusiasm for change among civil servants can lead to very different responses to evidence-based policymaking. In response to political sensitivities, civil servants may also exert influence to effect changes in program design or in the publication of results.

- *Multilateral or bilateral development agencies* offer the chance to implement programs at a large scale, and can bring significant funding for the program.[11] They also have some very skilled and experienced staff, both at the headquarters and regional offices, with experience working in numerous countries. However, there are two key challenges of working with them: their policy priorities can change significantly when their national governments change, and they often have a very specific geographic focus corresponding to their home country's strategic interests.

For instance, in FY 2010, two of the largest recipients of funding from USAID are Pakistan and Afghanistan, whose total allocation dwarfs transfers to other countries.[12] Similarly the UK's Department for International Development (DFID) has identified *"Twenty-seven focus countries on which it will now concentrate its bilateral funding. From DFID's previous list of 'priority' countries, 16 countries have now been removed."*[13]

- *Local NGOs in developing countries* can be faster and more flexible in implementing new approaches and usually have very dedicated staff. Further, if NGOs do not have a guaranteed stream of funding for their activities in the absence of evidence on program effectiveness, they may face better incentives to rigorously evaluate their programs. But if there is already sufficient political and financial support for the program without even rigorous evidence, then new evidence about program effectiveness can actually create risk in case the evaluation reveals that the program has little or no impact (Pritchett 2002). For this reason, NGOs who are truly agnostic about the means of achieving a particular policy goal, or who do not have a guaranteed stream of political or financial support without evidence on program effectiveness may have the best incentives to participate in a rigorous evaluation of their programs.

But the desirability of working with 'agnostic' organizations can also make it harder to work with NGOs as many of them often do have very strong opinions on what interventions to pursue to achieve their policy objective, even though their financial constraints can make them more willing partners due to their dependence on funders for financial support. Moreover, it can be difficult to identify a reputable, reliable and effective NGO from among a very large and heterogeneous field in many developing countries. Further, many NGOs lack the resources or scale of operations to implement an evaluation with sufficient sample size, or to act on evidence via a large scale-up of effective programs. Programs tested with NGOs can sometimes face skepticism when presented to governments, because of the differences in scale and institutional design between government and NGO operations. An example is the difficulty in paying performance-based salaries in governments.

One important lesson that we have taken from many of these efforts is that challenging conventional wisdom (e.g. 'incentives are tantamount to bribes to do the right thing') and changing people's attitude towards the usefulness of evidence in policymaking are often a slow process that requires a fair amount of effort, time, skill, and patience. But with persistence and good evidence, many policies and programs do adapt to incorporate rigorous evidence, even if that happens with a lag. Organizations like J-PAL that can afford to take a long view on the importance of evidence-based policy can play a critical role in this process. After researchers begin the dissemination of their results in academic conferences and publications, J-PAL continues the process via the creation of policy

documents to better communicate those results, networking with policymakers to explain these findings, organizing evidence workshops geared at policymakers and providing technical assistance in using this evidence, including for replications and scale-ups.

b. Facilitate partnerships of researchers with policymakers

Not only do policymakers benefit from close interaction with researchers, but researchers also have much to gain from such partnerships:

First, policymakers have valuable information on the most pressing issues facing their constituents, a great understanding of the context of their region, which programs have already been implemented, and what the primary constraints on program options are, including administrative, political, and technical concerns. This can help guide academics to the most relevant research questions, and also give them a sense of the difficulties that new programs may encounter.

Second, evaluations in the field also provide a valuable opportunity for researchers to empirically test their economic theories, especially those around human behavior and decision making. Testing theories in real-world scenarios not only provides validation (or otherwise) for these theories, but it also gives insight into what other factors may be at work in economic decisions, as well as the relative importance of competing economic forces in different contexts. For example, there has been considerable debate about the merits of charging for preventive health products. Some have argued that charging positive prices for health products may increase usage intensity by screening out those who do not value the products and inducing people to justify their purchase. On the other hand, charging a positive price for such products may reduce demand from those who cannot afford to pay. Knowing whether charging for products elicits such 'screening' or 'sunk cost' effects can not only lead to better policies, but it can also help validate economic theories about how people think about streams of costs and benefits. J-PAL affiliate Pascaline Dupas responded to this debate by setting up randomized evaluations in Kenya, and found that the demand for insecticide treated bednets dropped steeply when a positive price was charged, but usage by those who had purchased the product did not significantly increase.[14] In fact, demand dropped off nearly as steeply when people were offered a 50 per cent subsidy as when they were offered a 90 per cent subsidy, suggesting that demand for preventive health products is sensitive to price rather than subsidy level. This suggests that the screening or sunk costs effects, where they exist, are likely to be far smaller than the simple price effects of charging positive prices. A similar experiment could have been run in a lab environment, but examining actual policies and practices offers the opportunity to examine people's responses to non-hypothetical situations involving crucial matters such as health, consumption, and education.

Third, researchers are reliant upon their implementing partners for the smooth implementation of any of the programs they are testing. If a program is incorrectly implemented, it becomes more difficult to draw strong conclusions about its results. A close feedback loop between researchers and policymakers ensures that such concerns are addressed quickly and effectively so that programs do not fail due to entirely avoidable implementation problems.

Fourth, if policymakers see the researchers contributing positively by providing evidence from various projects and giving feedback on program design, they are more likely to be motivated to ensure that considerations of evaluation design are addressed as well. This is particularly true in randomized evaluations, where a clear separation of treatment and control groups is necessary and often requires the policy partners to be advocates to their colleagues, helping to explain why randomization will yield solid results and what can be gained from rigorous evaluation. An example of this kind of advocacy is visible in an ongoing J-PAL project with the Government of Karnataka. One of the authors, who is also the co-PI on this project (Iqbal Dhaliwal), was actively involved in all stages of program and evaluation design with the government, and in response the health department agreed to change many features of the program design based on learning from previous J-PAL studies. They were also willing to change the rollout plan from 100 per cent coverage in two pilot districts to 40 per cent randomly chosen coverage in five districts, enabling a rigorous impact evaluation.[15]

Fifth, if researchers stay engaged with policymakers, then the policymakers are more likely to approach the researchers for subsequent partnerships for evaluating new and innovative program concepts, designs and evaluations. This can lead to a virtuous cycle of evidence-based policy.

Sixth, while there is currently great interest in funding and promoting evaluations to generate evidence on what works or does not in the fight against poverty, if after a few years funding organizations find no linkage between the results of that research and policy, it is conceivable that the relevance of evaluations can diminish. On the other hand, if the results from rigorous evaluations feed into policy, it can attract other researchers to the field. For instance, recent innovative evaluations in governance (e.g. issues around measuring corruption, using information to strengthen voter control over leaders and the role of community monitoring in strengthening public services) have helped to establish the important role that evaluations can play in research on governance issues.

We now turn to a discussion of some of the strategies that J-PAL employs to facilitate new partnerships between researchers and policymakers:

- **Respond to requests from policymakers for evidence and partnerships:** J-PAL's affiliates and senior management have strong connections with policymakers in

governments, foundations, NGOs, and international development organizations around the world. Because of this, J-PAL is frequently approached with requests from policymakers to discuss evidence on a pressing policy challenge or to help identify researchers who would be willing to work with them to design and evaluate innovative programs to address that challenge. In the past, J-PAL has focused on such opportunistic outreach, making use of existing connections to disseminate policy lessons and encourage the adoption of effective policies. When policy contacts request information on the current J-PAL body of knowledge for a particular policy challenge, J-PAL staff at Cambridge or the regional offices gather the most relevant evaluations and cost-effectiveness analyses, create a presentation or printed materials that meets the particular needs and questions of the organization, and travel to meet with and present to them. There has been a significant increase in such requests that are made to J-PAL the past few years as our affiliate network and their research covers more countries and issues and is featured prominently by the media. The recent publication of *Poor Economics* by two of J-PAL's Directors has further increased such requests.[16]

- **Targeted outreach conferences for policymakers to disseminate evidence and match-make partnerships:** At the same time, the Policy Group has been complementing outreach in response to such requests that originate from policymakers with targeted outreach to organizations that work on a particular development issue, or in a region that could greatly benefit from J-PAL's research findings. When it becomes clear that there is sufficient evidence about a particular region or theme, and indications of 'responsiveness to evidence' among local policymakers to improve outcomes in that area, J-PAL identifies policymakers who have the most potential to impact policy and focuses dissemination efforts on this group. Often, this involves organizing an event like a conference or workshop that features presentations by researchers and their policy partners to discuss the details of tested programs and the results from their evaluations. Such events also provide valuable opportunities to make new contacts within organizations and to get feedback on the most pressing questions that policymakers want answered.

  For instance, in 2010, J-PAL identified the Indian state of Bihar as a region that has some of the lowest indicators in health and education, where evidence from J-PAL studies could make an impact, and where the new political leadership had demonstrated a strong commitment to improving development outcomes. J-PAL reached out to the Government of Bihar and organized a joint regional conference that brought together senior researchers from the J-PAL network, their field partners from different states of India, and top ministers and civil servants in Bihar to discuss pressing social sector issues. The conference addressed problems as diverse as how deworming can reduce endemic health and education problems associated with intestinal worms, and how double-fortified salt may hold promise in the fight against

iron-deficiency anemia.[17] Based on discussions with the Government of Bihar that immediately followed this event, the state government agreed to conduct a massive school-based deworming campaign in partnership with J-PAL's sister organization, Deworm the World.[18] This campaign targeted 17 million school-children aged 6-14 years from February to April 2011.[19] J-PAL affiliates have also begun a pilot field experiment to determine the optimum price and distribution channel for providing double-fortified salt, and to examine its impact on reducing chronic anemia in Bihar.

In addition to such regional conferences, J-PAL also organizes thematic conferences that are primarily aimed at developed country foundations and international development organizations that are responsible for allocating funds to large development programs around the world. For example in 2011, J-PAL organized a conference on agricultural technology adoption with the Center of Evaluation for Global Action (CEGA)[20] and USAID to highlight the role of technology in agricultural production, identify technologies that are appropriate for Africa, enhance understanding of constraints to agricultural technology adoption, and make matches between policymakers and researchers.[21] For such matchmaking conferences, J-PAL staff identify organizations engaged in that sector from around the world, and talk to them about their research priorities and the programs that they need evaluated. Staff spend extensive time screening out organizations that do not have specific research questions or where their program may not be able to support the research design, and then invite the remaining group to a conference. Simultaneously, researchers whose work focuses on related questions are invited to attend and give presentations on their research interests and past partnerships. At the end of the conference, researchers and policymakers are able to share their priorities for new research and find common questions from which new evaluations and partnerships can evolve.

Similarly, in 2010, J-PAL organized a joint conference in New York with our sister organization, Innovations in Poverty Action (IPA),[22] as well as the Financial Access Initiative, Moody's, Deutsche Bank, and CGAP, on Microfinance Impact and Innovation to communicate accumulated knowledge as well as to generate innovations in both microfinance product design and research.[23] And in April 2011, J-PAL organized a Policy Research Colloquium on education jointly with USAID and the World Bank in Washington DC that focused on the use of evidence for education policymaking and reform and where World Bank President Robert Zoellick delivered the opening keynote.[24]

- **Special funding initiatives to identify pressing areas of research need and encourage a coherent research agenda:** J-PAL has recently started creating new funds, termed 'Initiatives,' to promote original research, policy outreach, and capacity building in areas of pressing policy need. These donor-funded Initiatives are led by affiliated professors as co-chairs, and have a full-time senior staff

member who serves as the Initiative manager. The Initiatives begin by conducting a comprehensive literature review to identify what we know and don't know works in achieving the outcome of interest (e.g. technology adoption in agriculture, strengthening democracy and community participation in developing countries). The resulting comprehensive 'review paper' identifies the main issues for future research. Feedback is widely solicited from policymakers around the world who specialize in the particular area, facilitating easier research-policy communications than would take place with multiple one-on-one engagements. This paper then serves as the basis for a Request for Proposals (RFP), in which researchers and their field partners are invited to submit proposals to evaluate innovative programs that can help achieve the outcomes of interest. Because of their institutional continuity in terms of the underlying research paper and leadership, the Initiatives are able to promote a more coherent body of research and evidence. In this way, the review paper acts a medium for dialogue and ideas exchange between a number of researchers and policymakers on pressing issues of common interest, often a more efficient method than numerous individual interactions to arrive at a similar result – a research agenda that is widely accepted by both researchers and policymakers. J-PAL then coordinates conferences, matchmaking between researchers and field partners, and outreach activities around these Initiatives where the focus is on disseminating a comprehensive body of evidence and learning that can translate to policy decisions and programs.

For instance, the Bill and Melinda Gates Foundation helped launch the Agricultural Technology Adoption Initiative (ATAI) with the mission to develop and rigorously test programs that improve adoption and profitable use of the most cost-effective agricultural technologies by small-scale farmers in South Asia and sub-Saharan Africa. The long-term objective of ATAI is to ensure that the poor derive greater benefit from existing and new technologies to help move out of poverty.[25] Similarly, the William and Flora Hewlett Foundation and UK's DFID helped set up the Governance Initiative (GI) that funds impact evaluations of programs designed to improve citizen participation in the political and policy process and reduce corruption in public programs. Through dissemination of findings to policymakers, and by providing support for the scale-up and replication of successful programs, GI will help translate this evidence into concrete policy change.[26]

- **Policy-research collaborations with large development organizations, including national governments to tackle key policy challenges:** J-PAL also responds to requests from large development organizations to set up policy and research collaborations that are aimed at finding answers to particularly challenging policy questions. For example, the federal Government of Chile requested that J-PAL convene a commission to identify the most pressing social problems facing the country, brainstorm innovative programs to tackle these problems, and help evaluate the programs that the Commission recommends and the Government implements.

The 'Compass Commission' consisted of leading academics and policymakers from Chile and around the world, which met several times and submitted their report to the Chilean President in summer of 2011.[27] The Chilean Government has committed to funding the evaluation of at least one of the proposed programs. Similarly, the French Innovation Fund, co-chaired by J-PAL Europe's Scientific Director, Marc Gurgand, was established by the French Government with €200 million to finance 50 per cent of the cost of innovative programs aimed at integrating youth into the mainstream.[28]

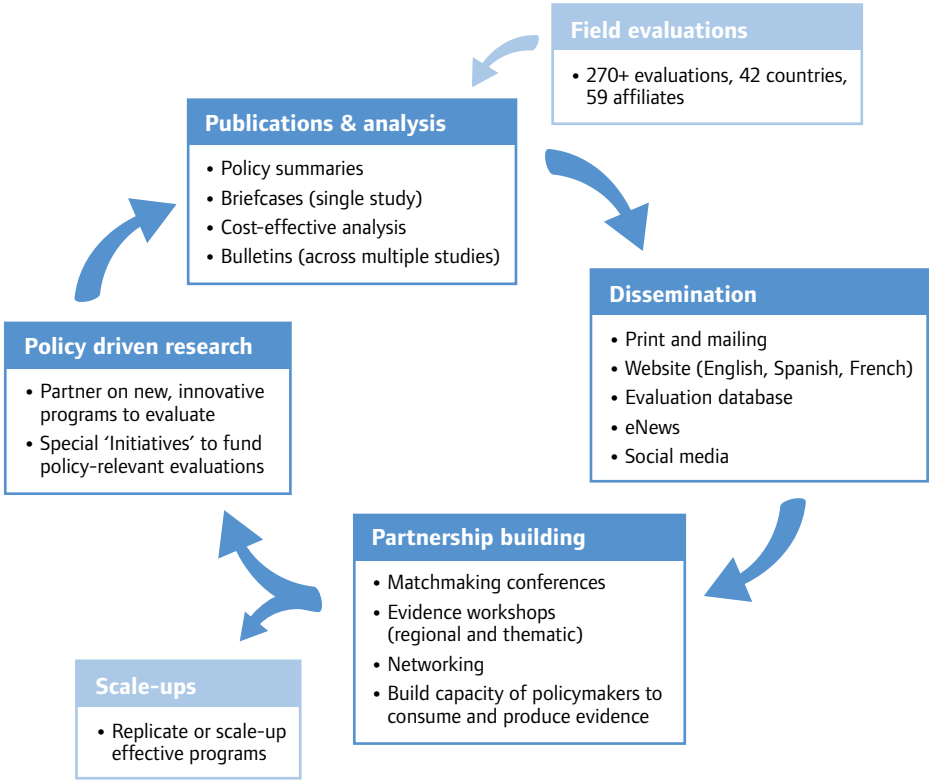
- **Create policy tools that allow researchers to effectively communicate their findings:** As discussed earlier, research results from evaluations are often written in academic journals in a technical format geared towards a research audience. This makes it hard for policymakers to access, interpret, or use these research findings. Further, researchers have few incentives to translate their research results into a format that would be easy for policymakers to follow or to invest the time and effort needed to network with policymakers on a continuous basis. J-PAL created a dedicated Policy Group in 2009 to strengthen the linkages between research and policy. The group's responsibilities include the creation of materials for policy outreach based on the results of the field evaluations, dissemination of knowledge about 'what works' in development to foundations, NGOs, international development organizations and governments, and building partnerships with these organizations to promote evidence-based policy and to scale-up programs that have been found to be effective. Figure 2 on page 119 illustrates the key activities of the Policy Group.

J-PAL creates the following policy tools to help researchers more effectively communicate their findings:

1. **Policy Summaries:** There are more than 270 ongoing and completed evaluations by J-PAL affiliates, and another 100 or more projects are in the process of being launched. For each evaluation conducted by a J-PAL affiliate, the Policy Group creates an 'evaluation summary': a two-page synopsis of the relevant policy questions, context of the study, details of the program being evaluated, and the results of the evaluation. These summaries are targeted at a non-academic audience, and are therefore written in a non-technical format. These evaluation summaries are available online at the J-PAL website in a database that is searchable based on policy theme, sector, region or author.[29]

2. **Policy Briefcases:** For evaluations that address a particularly relevant question for development practitioners, J-PAL creates expanded summaries called 'briefcases.' These briefcases, around four pages in length, provide a longer summary of the project and allow outreach to a larger audience. They are featured on the J-PAL website, printed, and mailed to key contacts of the J-PAL

Figure 2: J-PAL's policy group



Source: http://www.povertyactionlab.org

network around the world. Briefcases expand upon the policy questions that the evaluation addresses, provide more detail on the program being evaluated and how the evaluation was designed. These longer publications also provide additional information on the context in which the program was implemented and discussion of how the results can be extrapolated to other contexts.[30]

3. **Cost-Effectiveness Analyses:** One way to analyze results from multiple evaluations addressing the same policy goal is to combine them in a cost-effectiveness analysis. In simple terms, a cost-effectiveness analysis calculates the ratio of the amount of 'impact' each program achieves to the cost incurred to achieve that impact, or, conversely, the amount of cost required to achieve a

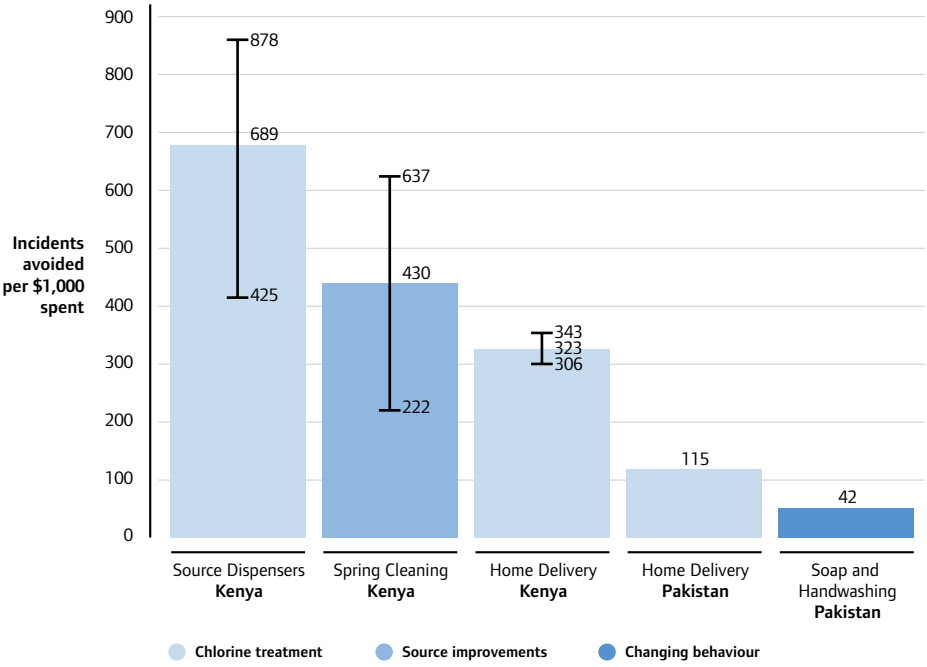
given impact. This ratio, when calculated for a number of alternative programs addressing the same policy goal, conveys the relative impacts and costs of these alternative programs in an easy and intuitive way. There are many advantages of cost-effectiveness analysis. First, it summarizes a complex program in terms of an illustrative ratio of impact to costs that is easy to understand. Second, it uses this common measure to compare multiple programs evaluated in different contexts and different years, allowing for better comparisons. Third, it is relatively simple for policymakers to perform sensitivity analysis by varying key inputs like population density and assumptions like discount rates to see how the programs might translate to their context.

However, relatively few studies include cost data in the published paper, and what data is available is typically presented in a wide variety of formats that make cost-effectiveness analysis very time-consuming and complex. J-PAL's Policy Group collects impact and cost data from programs which aim to achieve the same policy goal (for example, reducing diarrheal disease in children), and calculates the cost-effectiveness of each program, expressed as the cost of achieving one unit of impact (for instance, the cost to avert one incident of diarrhea). Graphs of the comparative cost-effectiveness of these programs are produced, and the underlying calculations are also provided, giving policymakers the opportunity to examine why a particular program looked more or less cost-effective. Figure 3 on page 121 is an example of a J-PAL cost-effectiveness analysis.

The major challenge in this analysis is to strike the right balance in the trade-off between a format that is easy and intuitive to understand, and sufficient details of the underlying programs that provide all relevant information on their context, design and evaluation. For instance, a program piloted in a country with lower costs of labor will look more cost-effective than an identical program piloted in a country with higher costs, and so there is a danger that readers could misinterpret a graph as making absolute judgments about the relative merits of programs piloted in different contexts. For this reason, it is necessary to be extremely careful about how the results of comparative cost-effectiveness analyses are presented, and it is often useful to provide explanatory text that provides interpretation of the graph itself. For a more complete discussion of cost-effectiveness methodology in a policy context, see Dhaliwal *et al.*, 2011.[31]

4. **Policy Bulletins:** The results of these cost-effectiveness analyses, as well as syntheses of evaluations which cannot necessarily be quantified in cost-effectiveness analyses, are presented in longer print publications called 'bulletins,' which are also made available through the J-PAL website. These syntheses are particularly valuable as they offer a unified message about the lessons learned in a particular field. This is more so because the bulletins are developed

Figure 3: Diarrhea incidents avoided per $1,000 spent (intervals)



collaboratively by all of the J-PAL researchers who worked on the underlying evaluations included in the bulletin, and hence present the overall policy lessons from all those evaluations rather than just results from individual studies.

For example, J-PAL recently released a bulletin synthesizing the results from ten evaluations that examined the effects of user fees or cost sharing for preventive health products and education. From the results of these evaluations, J-PAL finds that distributing preventive health products for a nominal fee significantly reduces take-up relative to free distribution, while failing to improve product usage rates.[32]

J-PAL has seen great demand for these bulletins because they address two of the main constraints to evidence-based policy discussed above. First, the bulletins provide a resource for distinguishing high-quality evidence from the multitude of evidence of varying quality, as J-PAL includes only high-quality randomized evaluations in the bulletins. Second, the bulletins address the difficulty in

synthesizing coherent lessons from multiple evaluations, especially when these evaluations appear to have, at face value, conflicting results. Because they tend to contain evaluations of programs from multiple contexts and where multiple variations have been tested, bulletins are also able to provide more information on the sensitivity of this type of program to different contexts and assumptions, such as population density or the cost of program inputs. These bulletins, while longer than an evaluation summary or briefcase, are targeted for a policy audience and written in non-technical language.

J-PAL uses all of these tools to build long-term relationships with policymakers. Sometimes these relationships begin when policymakers approach J-PAL or its affiliates for help identifying evidence regarding a particular policy challenge they are facing. Such policymakers are usually highly motivated to improve outcomes, 'champions' of evidence-based policy and themselves keen to build long-term relationships with researchers. But in other cases, policy-research relationships begin after J-PAL identifies policymakers who are in positions where evidence-based policymaking could have significant benefits. Examples include identifying a program manager in a foundation dealing with large-scale funding of various programs for increasing immunization, or the education secretary of a country or a state who is deciding which programs to implement to increase learning outcomes in schools. In such cases, the Policy Group at J-PAL networks with these policymakers to explain relevant evidence, provides them policy documents, jointly organize conferences or make specific presentations at their organization.

In other cases such relationship building begins even before a policymaker is in a position to directly use J-PAL's evidence, but has the potential to be in that situation in the future. Examples include executive training courses or customized courses for senior civil servants. In this case outreach happens with the hope that when some of these policymakers are in a position to make critical decisions regarding funding or implementing programs, they will be more inclined to incorporate evidence in their decision making.

4. Scaling up successful programs

Effective dissemination of research results to policymakers along with strong policy partnerships between policymakers and researchers can result in an interest from policymakers in implementing similar programs in their regions. This can take the form of a program replication (running a small-scale pilot covering a subset of the population) or a program scale-up (launching a large-scale program expansion covering the entire target population). In such cases, J-PAL works in close partnership with the interested

policymakers to compare the key details of the original program and the context-specific factors like the severity of the problem, the capacity of local implementing agencies to undertake the project, and the service delivery models. If this process reveals that the program has a potential to be successfully replicated in the new context, J-PAL works with the implementing organization to design a 'policy pilot' that introduces the program in a small area. It is our experience that in most cases it is preferable to implement a replication or a 'policy pilot' before scaling-up the program for a number of reasons:

First, because of the smaller size and therefore lower cost of implementing a pilot, policymaker and funders (internal or external) are more likely to approve such pilots to minimize the risk of allocating resources to a new program. This is especially the case when there is some level of uncertainty or debate about the optimum program design or expected outcomes.

Second, such policy pilots replicate the original evaluation on a small scale, but often incorporate important tweaks to the design or implementation in line with either the learning from the final results of the original evaluation or in response to the differences in context between the two locations. As such, these pilots help the implementer understand how well the original impacts and costs generalize to this new context and provide additional data on proper implementation for the larger scale-up.

Third, implementing a program in the field during the replication process helps identify new local implementation challenges that may never have arisen during the original study. This can lead to important changes in the program design or implementation process to make it as successful in the new context as it was when the program was originally evaluated. For instance, J-PAL is working with its sister organization, Innovations for Poverty Action (IPA), as it partners with the Government of Ghana to pilot a program offering remedial education through contract teachers in primary schools in Ghana. This program, which is based on the NGO Pratham's remedial education program in India, is being rigorously evaluated during its pilot phase, and if it is found to be effective, the Government of Ghana may scale it up to the rest of the country. Such a replication will provide useful information to help the Ghanaian policymakers learn what challenges and issues they may need to consider ensuring a successful scale-up.

Fourth, when policymakers are actively engaged in the evaluation or pilot process, they acquire a much better appreciation of the importance of specific features of the successful program design and are less likely to oppose or modify it in a way that weakens the program. An example is the evaluation by J-PAL affiliates on increasing full immunization rates, which found that predictable and regular supply of vaccines coupled with effective demand-side incentives for parents can lead to a large increase in full immunization rates for children in rural India.[33] It has been our experience that many policymakers who have read the results of this study and want to use a similar program

to tackle the problem of low immunization rates in their area, still tend to overestimate the impact of supply-side improvements and underestimate the impact of demand-side incentives, despite very clear evidence from this study that the demand-side incentives caused the larger increase in full immunization rates. Some base this belief on the grounds that 'our context is different', while others are ideologically opposed to incentives on the grounds that they are tantamount to 'bribing' parents to do what is right for their children. Whatever the reason, as a result of this misplaced overemphasis on the supply-side, many of these policymakers may be inclined to design a scale-up that includes only the supply-side intervention, while ignoring the critical findings on the demand-side incentives. In such a situation, it is very useful to have a low-cost pilot study or replication evaluation where different variations of the program are tested. For instance, the immunization program could be tested with the supply-side intervention only, and then with both supply- and demand-side interventions, to demonstrate that the large impact that interested policymakers in the original evaluation can only be achieved by a combination of both demand- and supply-side interventions. Policymakers are far more likely to agree to such tests when they take the form of small-scale pilots.

While not enough programs evaluated by J-PAL affiliates have been scaled-up for us to draw general policy lessons, we provide some examples below of how different types of evidence in different situations led to scale-ups or widespread policy changes:

- **A policy organization may have the capacity to massively scale-up its own program after it is evaluated and found to be effective:** Sometimes an implementing organization may have the capacity to use evidence from an evaluation of its own pilot to massively scale-up that program. A well-known example of this is Mexico's PROGRESA program, which gave cash transfers to poor families conditional upon children's attendance at school and doctor's appointments. Evaluations conducted to estimate the impact of this program found significant effects on school enrollment as well as health outcomes.[34] Shortly after the program was piloted and scaled-up, a new party took power in Mexico. The strength of the evidence from these evaluations, as well as the immense popular support the program enjoyed, likely contributed to make it politically infeasible to discontinue the program, and it continued under the new name of Opportunidades.[35] Another example is the scale-up of police skills training in the Indian state of Rajasthan. After an impact evaluation of a pilot program by the state police found that training in investigation and soft skills had significant positive effects on the quality of police work and public satisfaction. The program was scaled-up to the entire state.[36]

- **A policy organization may leverage evidence to raise money to scale-up its operations:** In 2001, J-PAL affiliates evaluated a pilot program of the education NGO Pratham's Balsakhi program of providing remedial education to school children in public primary schools in Vadodara and Mumbai in India. The evaluation found that

the program had significant impact in raising the learning outcomes of participating students. Pratham used these results to demonstrate the effectiveness of its program and raise funds from the William and Flora Hewlett Foundation and the Bill and Melinda Gates Foundation to massively expand this program to cover more than 300 of the 600 districts of India (305,000 villages).[37]

- **The underlying program found to be effective in an evaluation is so highly cost-effective, generalizable and simple to implement that many other organizations readily replicate or scale it up:** J-PAL affiliates Michael Kremer and Ted Miguel, in their seminal 1997-2001 study in Kenya, showed that school-based deworming was highly cost-effective: at a cost of less than 50 cents per child per year, school-based deworming reduced the incidence of infection by 25 percentage points and reduced school absenteeism by 25 per cent. The relationship between taking a deworming pill and reduction in morbidity associated with worm infections is universally generalizable to areas that have similar worm loads. And the program is inexpensive to implement, as it piggybacks on the existing school infrastructure, requires minimal training of the teacher, and is easy to administer: children need one pill every 6 to 12 months based on the worm load. Following the evaluation, Kremer assisted in the founding of Deworm the World, an NGO dedicated to promoting deworming policies in developing countries worldwide. Using the evidence from the original evaluation in Kenya, as well as similar evaluations in other countries, Deworm the World has helped make deworming a policy priority for both education and health organizations worldwide.[38]

- **An evaluation provides critical and timely evidence on a very salient policy debate:** Sometimes timely and highly relevant evidence can spur organizations to make changes to policies or program designs. For instance, a 2006 study by J-PAL affiliate Pascaline Dupas and Jessica Cohen found that charging even small positive prices considerably decreased demand for insecticide treated bednets, and women who paid positive prices were no more likely to use the bednets than those who received them for free. The study provided critical evidence in the midst of a raging debate on whether such bednets should be given for free or at a positive price. Soon many organizations changed their policy from charging a positive price for preventive health goods to distributing them for free.[39]

As the deworming example shows, having an organization that provides support to policymakers in the scale-up process can be a critical factor for success. While J-PAL does not itself implement the scale-up of programs found to be effective, it provides technical guidance to policymakers including sharing key insights from the original study (both on the implementation and evaluation side), helping the implementing organization plan the logistics for the scale-up process, and providing assistance in monitoring and evaluating during the scale-up process. Such support can be especially

critical for governments, as many of the policymakers there are so busy in implementing the large number of existing programs and reacting to crises that there is a very high valued added of assisting them in program design, implementation planning, and providing continuous objective feedback from the field during a replication or scale-up. As the Secretary of a key social department that implements numerous poverty alleviation programs in one of India's largest states once told us, *"Don't just tell me what the best strategy is, come and help me implement it."*

5. Conclusion

In writing this paper, we have deliberately used the term 'influencing' or 'informing' policy, rather than 'impacting' it, because it is not possible to apply the same rigorous criteria for measuring impact on the policy process. There is no counterfactual for what would have happened in the absence of a particular piece of evidence or a particular outreach strategy, and so we cannot claim to know exactly what the causal impact of J-PAL's outreach is. However, it is still instructive to examine the ways in which the research is disseminated to the policy audience, and look at examples where particular pieces of evidence have moved the debate or contributed to the adoption of a proven effective program. Over time, as J-PAL increases its policy outreach activities, there will be more learning to share. But for now, the purpose of this paper has been to discuss why we believe such policy outreach is critical, both for policymakers and researchers, and to share some of the observations from our experience to date.

In this paper we have tried to provide our perspective on some of the factors that help or hinder the use of evidence in policymaking (see the appendix for a summary), and to explain some of the approaches that J-PAL uses. Still we recognize that there are many different ways to inform policy, and the approach outlined in this paper may not be the best or most appropriate in all circumstances, and there may be other approaches that are more effective.

Even when policy decisions are based on some type of evidence, they must take into account other factors like administrative capacity, political constraints, technical limitations, time pressures, and budget limits. J-PAL's policy outreach works to make it easier for policymakers to include evidence as one more critical input in their decision making process, along with these other factors, while at the same time providing the policy tools and support to researchers for more effective dissemination of their findings to policymakers. Randomized evaluations are particularly well suited for providing evidence to the policy process because they measure impact rigorously, can provide key insight on why programs succeeded or failed, provide practical information to help facilitate and guide scale-ups, and are easier for non-technical audiences to understand.

But even evidence from randomized evaluations can be difficult for policymakers to interpret and use if it is not disseminated well. This is because of a number of factors. Evidence is often presented in technical language, and policy-relevant information about the context of evaluations and program details that may be of most interest to policymakers may not be reported in research papers. Policymakers may also find it hard to distinguish between evidence of different quality, and it can be difficult to compare evidence from different contexts. Demand for evidence is also time-sensitive, and researchers are often constrained in their ability to invest time and effort to engage policymakers on a continuous basis, making it difficult to find the right piece of information at the right time.

J-PAL tries to promote evidence-based policy by encouraging policymakers to become better consumers and producers of evidence and by using field evaluations as an opportunity to build strong relationships between policymakers and researchers while maintaining objectivity in reporting results. Similarly, J-PAL encourages researchers to work with policymakers by responding to requests from policymakers for evidence and partnerships, and performing targeted outreach conferences for policymakers to disseminate evidence and match-make partnerships. In recent years J-PAL has begun creating special funding Initiatives that help overcome some of the obstacles described above, and entering into policy-research collaborations with large development organizations, including national governments, to tackle key policy challenges. To support all of these activities, the J-PAL Policy Group creates policy tools that allow researchers to effectively communicate their findings.

Success in the above efforts can result in the scale-up of effective programs. It may be appropriate for policymakers to first pilot the program in their context before scaling it up extensively. Experience suggests that evidence can be successful in encouraging scale-ups under many different circumstances. Scale-ups may occur when organizations whose programs have been evaluated are also capable of scaling them up, or when organizations leverage the results of evaluations that find their programs to have a strong impact to raise funds for large scale-ups. Outside organizations may also be likely to replicate a program when the underlying intervention is very simple to implement, highly cost-effective and generalizable to other contexts, and when an evaluation provides strong and actionable results in the middle of a particularly vigorous policy debate.

While there is a lot more to learn about what makes particular evidence more likely to be incorporated in policy, our experience has been that policymakers are more likely to use evidence in decision making if that evidence is: *Unbiased* (independent evaluation not driven by any particular agenda); *rigorous*: (used best methodology available and applied it correctly); *substantive* (builds on existing knowledge and provides either insights that are novel, or evidence on issues where there is a robust debate as there is little use for

evidence that reiterates well established facts); *relevant* (fits the policymakers' context, needs and problems); *timely* (available when policymaker needs it to make decisions); *actionable* (comes with a clear policy recommendation); *easy to understand* (links theory of change to empirical evidence and presents results in a manner that is easy to understand); *cumulative* (draws lessons from not just one program or an evaluation, but the larger body of evidence in that area)' and *easy to explain to constituents* (it helps greatly if researchers have been building up a culture of getting the policymakers and other stakeholders on board through conferences, policy papers, opinion pieces, etc.)

Throughout this paper we have given many examples of how J-PAL tries to bring researchers and policymakers together. Our hope is that this can encourage more discussions among organizations whose mission, like J-PAL, is to ensure that *"policy is driven by evidence and research is translated into action."*

Appendix 1: Constraints and strategies for research-policy partnerships

Actions by researchers that deter policymakers	Responses that encourage policymakers to partner with researchers
1. Disconnect of evaluator from program design	Willingness to engage with policymaker in the concept and design of the program to be evaluated *e.g., Rajasthan Police.*
2. Unilateral decisions on design of evaluation	Work with policymakers to understand their main program concerns, and how the evaluation can be structured to answer them *e.g., Immunization.*
3. Inflexibility in evaluation approach	Consider alternate evaluation design to accommodate political constraints and field realities without compromising rigor and objectivity *e.g., Minister's District.*
4. Only measuring 'ideal', but long-term outcomes	Construct additional short-term evaluation outcomes while continuing to design long-term measures *e.g., NRHM attendance vs. health outcomes.*
5. Evaluations that only measure impact, not reasons	Qualitative data collection during impact evaluation to understand if program implemented per plan, and what worked or not *e.g., Nurses breaking machines.*
6. Begin policy engagement only at start of project	Actively participate in policy conferences, meet key policymakers, contribute to civil society debate via op-eds, books etc., in region of interest *e.g., Bihar conference.*
7. End relationship at completion of project	Willingness to stay engaged as a 'technical' resources for policymaker even after publication, especially for scale-ups *e.g., Pratham (they provided expertise for Ghana).*
8. Not report negative results	Maintaining rigor and absence of bias in evaluation and reporting results despite above close relationship *e.g., Flour fortification with Seva Mandir.*
9. Shifting evaluation objective	Register hypothesis ahead of time to avoid allegations of data mining *e.g., JPAL and partners for RCTs.*
10. Discuss only one study (own research) and information overload	Explain 'policy findings' from the entire body of research, not just own narrow research and how this evaluation links to the body of evidence *e.g., DC Education Evidence Workshop, Cost Effectiveness Analysis, 'Poor Economics'.*

Actions by researchers that deter policymakers	Responses that encourage policymakers to partner with researchers
11. Original evaluation vs. replications	Be willing to evaluate the replication of a program found to have succeeded in another context and well suited for this problem, not just 'new and innovative' programs *e.g., Bihar immunization, our partner IPA*.
12. Technical jargon	Frame discussion in easy to understand language, communicate in a style policymakers are familiar with, and customize outreach to audience *e.g., Policy Bulletins, Briefcases, academic papers*.
13. Funding for evaluation may be difficult to raise for policymaker	More and more organizations require impact evaluations in the programs they fund and dedicated funding groups makes this easier *e.g., DIME @ WB, 3ie, IGC, JPAL Initiatives, IPA Funds*.

Actions by policymakers that deter researchers	Responses that encourage researchers to partner with policymakers
1. Political agenda trumps evidence	Target those who are open to evidence, so they use it as an input along with other factors like political agenda, budget constraint and ability of bureaucracy *e.g., TCAI in Ghana*.
2. Low capacity of policymakers to consume, generate or institutionalize evidence	Help train staff, establish M&E divisions, recruit technically competent people and motivate them by giving credence to their research and via formal linkages with leading academics *e.g., Government of Haryana*.
3. Short-term horizon of policymakers	Combine short-term outcome measures with long-term outcomes via phased roll-out *e.g., NRHM*.
4. Risk aversion and failure-avoidance by policymakers	Setup institutions that allow innovation and risk tolerance *e.g., Chile Compass Commission, French Evaluation Fund*.
5. Inability to build coalitions to support new programs	Work with other government agencies that are most receptive to evidence, even if not social development departments *e.g., Finance, Governor's Special Cell*.
6. Change in 'rules of the game' *viz*. evaluation	Sign MoU and stick to the agreement (terms include phased roll-out, control group, sample size, data publication, and scale-up if found to be successful). MoU to survive change of personnel and governments *e.g., Government of Karnataka State in India*.

Actions by policymakers that deter researchers	Responses that encourage researchers to partner with policymakers
7. Lack of institutional continuity and commitment due to frequent transfers of officials, or change in administrations after elections	All of actions listed under #6 plus efforts to build wider culture of evidence-based polic in the organization via capacity building and evidence workshops *e.g., Training of Indian Civil Servants (IAS) irrespective of current assignment or positions.*
8. Lack of pressure from civil society or legislature to conduct evaluations	Convince these institutions to demand evaluations via contribution of civil society debate (Opeds, workshops, legislation) *e.g., Mexican legislature created CONEVAL.*

Appendix 2: Bibliography

Banerjee, A. V., Duflo, E., Glennerster, R. and Kothari, D. (2010) Improving immunisation coverage in rural India: clustered randomised controlled evaluation of immunisation campaigns with and without incentives. 'British Medical Journal.' 2010; 340:c2220.

Banerjee, A. V., Glennerster, R. and Duflo, E. (2008) Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System. 'Journal of the European Economic Association.' 6(2-3): 487-500.

DFID Research and Evidence Division. (2011) 'Operational Plan 2011-2015.' London: Department for International Development.

Dhaliwal, I., Duflo, E., Glennerster, R. and Tulloch, C. (2011) 'Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education.' Cambridge, MA: MIT.

Duflo, E., Dupas, P. and Kremer, M. (2010) 'Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya.' Cambridge, MA: The National Bureau of Economic Research.

Glewwe, P., Kremer, M., Moulin, S. and Zitzewitz, E. (2004) Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. 'Journal of Development Economics.' 2004: 251-268.

Iyer, L., and Mani, A. (forthcoming). Traveling Agents: Political Change and Bureaucratic Turnover in India. 'The Review of Economics and Statistics .'

Kast, F. 'Mind the Gap Conference: From Evidence to Policy Impact.' Cuernavaca, Mexico, 2011.

Office of Evaluation and Oversight (2006) 'Assessment of the 2004 Project Completion Reports (PCRs) Produced Under the Bank's New PRC Guidelines.' Washington DC: Inter-American Development Bank.

Office of Evaluation and Oversight (2006) 'Evaluation of the IDB's Studies.' Washington DC: Inter-American Development Bank.

Pritchett, L. (2002) It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation. 'Journal of Policy Reform.' 2002: 5:4, 251-269.

Trafford, R. A. (2011) 'Case Study of the J-PAL Policy Group: Improving the Translation from Research to Policy.' Capstone Paper, Center for Public Policy & Administration, University of Massachusetts, Amherst.

US Agency for International Development. (2011) 'USAID Evaluation Policy: Learning from Experience.' Washington, DC: USAID.

Weiss, C. H. (1979) The Many Meanings of Research Utilization. 'Public Administration Review.' 1979: 426-431.

In addition to the above references, the authors conducted individual interviews with many senior policymakers at governments, international development organizations, NGOs, and foundations.

Acknowledgements

The authors gratefully acknowledge generous support from NESTA (the National Endowment for Science, Technology and the Arts, United Kingdom), especially from Stian Westlake and Ruth Puttick at NESTA's Policy and Research team. This paper would not have been possible without our numerous colleagues at NGOs, foundations, governments, and international development organizations around the world who have been very generous with their time over the past years in sharing their experiences with us, both while writing this paper and, more importantly, as part of building partnerships with us to ensure that policy is driven by evidence. We also thank our colleagues Rachel Glennerster and Mary Ann Bates who provided very useful feedback to us on many of the ideas we have developed here. All views expressed in this paper are the authors' personal views and do not necessarily reflect J-PAL's position.

Iqbal Dhaliwal is the Global Director of Policy and Caitlin Tulloch is a Policy Associate at the Abdul Latif Jameel Poverty Action Lab (J-PAL) at the Department of Economics, Massachusetts Institute of Technology (MIT). They can be contacted at **iqbald@mit.edu** and **ctulloch@mit.edu** respectively.

Endnotes

1. See: http://www.povertyactionlab.org/
2. Increasingly, many organizations are working to disseminate the results of research for policy audiences. See for example the Coalition for Evidence-Based Policy http://evidencebasedprograms.org/wordpress/ or The World Bank's Poverty Impact Evaluations Database (http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,contentMDK:21534261~pagePK:210058~piPK:210062~theSitePK:384329,00.html)
3. Banerjee, A., Duflo, E. and Glennerster, R. (2008) 'Putting a Band-Aid on a Corpse.' For a summary of this evaluation, see: http://www.povertyactionlab.org/evaluation/incentives-nurses-public-health-care-system-udaipur-india
4. Duflo, E., Dupas, P. and Kremer, M. (2010) 'Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya.' For a summary of this evaluation, see: http://www.povertyactionlab.org/evaluation/peer-effects-pupil-teacher-ratios-and-teacher-incentives-kenya
5. We have kept the identities of all the policymakers we spoke to for this paper anonymous per their request, and use the pronoun 'her' for all of them, irrespective of their true gender.
6. For a detailed description of the issues involved in comparing multiple programs using a cost effectiveness framework, see Dhaliwal, I., Duflo, E., Glennerster, R. and Tulloch, C. (2011) 'Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education.' Available at: http://www.povertyactionlab.org/publication/cost-effectiveness
7. Iyer, L. and Mani, A. (Forthcoming) Traveling Agents: Political Change and Bureaucratic Turnover in India. 'The Review of Economics and Statistics.'
8. For example J-PAL's Director, Esther Duflo presented a Ted Talk in February 2010. Available at: http://www.ted.com/talks/esther_duflo_social_experiments_to_fight_poverty.html and J-PAL's Director of Policy, Iqbal Dhaliwal, presented at the closing plenary of the 3ie 'Mind the Gap' conference, available at: http://www.impactevaluation2011.org/video-conferences_eng.html
9. USAID (2011) 'USAID Evaluation Policy: Learning from Experience.' Washington, DC: United States Agency for International Development.
10. Details at: http://www.usaid.gov/div/
11. Large international NGOs, including those headquartered in developed countries, share many of the characteristics as these international development agencies.
12. See: http://www.usaid.gov/policy/budget/money/
13. See: http://www.gla.ac.uk/media/media_192054_en.pdf and http://www.dfid.gov.uk/aidreviews
14. Dupas, P. and Cohen, J. (2010) Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment. 'Quarterly Journal of Economics.' 125 (1): 1-45. For a summary of this evaluation see: http://www.povertyactionlab.org/evaluation/free-distribution-or-cost-sharing-evidence-malaria-prevention-experiment-kenya. Also see the J-PAL Policy Bulletin (2011) 'The Price is Wrong.' Available at: www.povertyactionlab.org/the-price-is-wrong
15. Dhaliwal, I. and Hanna, R. (2011) 'Integrated Medical Information and Disease Surveillance in Primary Health Centers in India.' For a summary of this evaluation, see: http://www.povertyactionlab.org/evaluation/integrated-medical-information-and-disease-surveillance-primary-health-centers-india
16. See: http://pooreconomics.com/
17. Details at: http://www.povertyactionlab.org/south-asia/bihar-conference
18. http://www.dewormtheworld.org/
19. 'The Hindustan.' Patna Bureau, September 30 2011.
20. See: http://cega.berkeley.edu/
21. Details at: http://atai-research.org/about-atai/news/atai-matchmaking-conference and http://agrilinks.kdid.org/groups/agricultural-technology-adoption-food-security-africa-evidence-summit
22. See: http://www.poverty-action.org/
23. See: http://www.microfinanceimpact2010.org/

24. See: http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTEDUCATION/0,,contentMDK:22896219~menuPK:282428~pagePK:64020865~piPK:51164185~theSitePK:282386,00.html

25. See http://atai-research.org/

26. See http://www.povertyactionlab.org/GI

27. See http://www.povertyactionlab.org/latin-america/compass-commission

28. For details of this fund and related research, see: http://www.ressourcesjeunesse.fr/Fonds-d-experimentation-pour-la.html For an English version, see: http://www.povertyactionlab.org/europe/social-experimentation-in-action

29. The database is available at: http://www.povertyactionlab.org/evaluations

30. J-PAL's print publications can be accessed at: http://www.povertyactionlab.org/policy-lessons/publications

31. Dhaliwal, I., Duflo, E., Glennerster, R. and Tulloch, C. (2011) 'Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education.' Available at: http://www.povertyactionlab.org/publication/cost-effectiveness

32. See: 'The Price is Wrong.' Available at: www.povertyactionlab.org/the-price-is-wrong

33. Banerjee, A. V., Duflo, E., Glennerster, R. and Kothari, D. (2010) Improving immunisation coverage in rural India: clustered randomised controlled evaluation of immunisation campaigns with and without incentives. 'British Medical Journal.' 2010; 340: c2220. For a summary of this evaluation, see: http://www.povertyactionlab.org/evaluation/improving-immunization-rates-through-regular-camps-and-incentives-india

34. Two well-known evaluations of PROGRESA are Gertler, P. J. (2000) 'Final Report: The Impact of PROGRESA on Health.' Washington DC: International Food Policy Research Institute; and Paul Schultz, T. (2004) School subsidies for the poor: evaluating the Mexican PROGRESA poverty program. 'Journal of Development Economics.' Vol. 74(1), pages 199-250, June.

35. For the history and political economy of PROGRESA, see: http://www.brookings.edu/global/progress/pap_total.pdf, or http://www.ifpri.org/sites/default/files/pubs/divs/fcnd/dp/papers/fcndp118.pdf, or http://www.bwpi.manchester.ac.uk/resources/Working-Papers/bwpi-wp-14211.pdf

36. Details at: http://www.povertyactionlab.org/scale-ups/police-training

37. Details at: http://www.povertyactionlab.org/scale-ups/remedial-education

38. Details at: http://www.povertyactionlab.org/scale-ups/school-based-deworming and http://www.dewormtheworld.org/

39. Details at: http://www.povertyactionlab.org/scale-ups/free-insecticidal-bednets

NESTA

1 Plough Place
London EC4A 1DE

research@nesta.org.uk
www.twitter.com/nesta_uk
www.facebook.com/nesta.uk

www.nesta.org.uk

ISBN 978-1-84875-132-3