How Do You Modernize a Health Service? A Realist Evaluation of Whole-Scale Transformation in London

TRISHA GREENHALGH, CHARLOTTE HUMPHREY, JANE HUGHES, FRASER MACFARLANE, CERI BUTLER, and RAY PAWSON

University College London; King's College London; University of Surrey; University of Leeds

Context: Large-scale, whole-systems interventions in health care require imaginative approaches to evaluation that go beyond assessing progress against predefined goals and milestones. This project evaluated a major change effort in inner London, funded by a charitable donation of approximately $21 million, which spanned four large health care organizations, covered three services (stroke, kidney, and sexual health), and sought to "modernize" these services with a view to making health care more efficient, effective, and patient centered.

Methods: This organizational case study draws on the principles of realist evaluation, a largely qualitative approach that is centrally concerned with testing and refining program theories by exploring the complex and dynamic interaction among context, mechanism, and outcome. This approach used multiple data sources and methods in a pragmatic and reflexive manner to build a picture of the case and follow its fortunes over the three-year study period. The methods included ethnographic observation, semistructured interviews, and scrutiny of documents and other contemporaneous materials. As well as providing ongoing formative feedback to the change teams in specific areas of activity, we undertook a more abstract, interpretive analysis, which explored the context-mechanism-outcome relationship using the guiding question "what works, for whom, under what circumstances?"

Address correspondence to: Trisha Greenhalgh, 206 Holborn Union Building, Highgate Hill, London N19 5LW, United Kingdom (email: p.greenhalgh@pcps.ucl.ac.uk).

The Milbank Quarterly, Vol. 87, No. 2, 2009 (pp. 391–416)

Findings: In this example of large-scale service transformation, numerous projects and subprojects emerged, fed into one another, and evolved over time. Six broad mechanisms appeared to be driving the efforts of change agents: integrating services across providers, finding and using evidence, involving service users in the modernization effort, supporting self-care, developing the workforce, and extending the range of services. Within each of these mechanisms, different teams chose widely differing approaches and met with differing success. The realist analysis of the fortunes of different subprojects identified aspects of context and mechanism that accounted for observed outcomes (both intended and unintended).

Conclusions: This study was one of the first applications of realist evaluation to a large-scale change effort in health care. Even when an ambitious change program shifts from its original goals and meets unforeseen challenges (indeed, precisely because the program morphs and adapts over time), realist evaluation can draw useful lessons about how particular preconditions make particular outcomes more likely, even though it cannot produce predictive guidance or a simple recipe for success. Noting recent calls by others for the greater use of realist evaluation in health care, this article considers some of the challenges and limitations of this method in the light of this experience and suggests that its use will require some fundamental changes in the worldview of some health services researchers.

Keywords: Modernization, change, transformation, realist evaluation.

Almost ten years ago, the British government launched an ambitious ten-year plan to "modernize" the National Health Service (NHS) from one perceived to be inaccessible, disease-oriented, inflexible, disjointed, error-prone, and inconsistent and to be delivered by overworked, unmotivated staff to a health service that was accessible, patient-oriented, flexible, coordinated, safe, evidence-based, and delivered by a professionalized and committed workforce (Department of Health 2000). An arm's-length body of the Department of Health (the Modernisation Agency) was established to promote the development, spread, and sustainability of this innovation using approaches advocated by Don Berwick at the Institute for Health Improvement in the United States (Berwick 2003; Greenhalgh et al. 2005).

The Modernisation Agency, whose annual budget was officially calculated to be more than £200 million ($285 million) (Secretary of State for Health 2007), provided a wealth of change management resources and consultancy support to more than 3,000 health care organizations across the United Kingdom. But the agency was abolished in 2004 after an internal Department of Health report concluded that it was not providing value for money (Herbert 2004). Subsequently, the government gave large-scale modernization initiatives in the NHS lower priority (and much less financial support), although publications and workshops continued via a smaller, leaner body (the NHS Centre for Innovation and Improvement). Whether the Modernisation Agency had actually "failed" or whether the privileging of quantitative and summative over qualitative and formative evaluation had failed to capture genuine and important impacts of its work was disputed (Bate and Robert 2003).

In 2003, when modernization was still a concept close to UK health care policymakers' hearts, a charitable sponsor made £15 million ($21 million) available to health care providers in a deprived inner-London district for what became known as the Modernisation Initiative (MI). Three different services—for stroke, kidney, and sexual health—were selected in a competitive bidding process to receive one-third of the total budget each for a program of "whole-scale transformation." The sponsor also funded an in-depth evaluation of the MI, for which it stipulated using qualitative, formative, and illuminative methods. This provided a unique opportunity to address three key questions: (1) What is the nature of the process by which a "whole-scale transformation" of a health service might be achieved? (2) What factors and preconditions make positive outcomes of such initiatives more likely? and (3) What generalizable lessons can be drawn about allocating funding for large-scale service transformation initiatives in health care?

Methods

Participants and Setting

The MI was focused on two adjacent London boroughs, which had all the challenges of a deprived inner-city area: poverty, poor housing, high burden of disease, low health literacy, high population turnover, linguistic and cultural barriers to effective communication, various socially excluded minority groups, and fragmented services. Moreover, some people seeking NHS care there did not actually live in the area, and conversely, some local residents chose to seek care elsewhere or not at all. The NHS services included two acute care teaching hospitals and two Primary Care Trusts (responsible for general practice and other community-based services such as family planning). Historically, the relationship between the hospitals had been characterized by competition rather than collaboration. Primary care services had limited funds and were of variable quality, with pockets of excellence coexisting with substandard practice. The three service areas chosen for the MI were seen to be particularly in need of improvement.

The MI was externally funded but delivered largely by redeployed NHS staff whose offices were located on NHS premises and hence was both "internal" and "external" to the service. Leadership, management, and governance mechanisms were complex and are described in detail elsewhere (Greenhalgh et al. 2008). They included an overarching board with representatives from participating NHS organizations and chaired by one of their chief executives, as well as numerous operational management groups.

The financial sponsor's aim was to use the generous funding to make a "big difference" in local health services. Significant, tangible improvement was expected in the nature of services (e.g., new services, service options, or modes of delivery), the culture of services (regarding behavior, relationships, and balance of power among health care organizations, staff, and patients), and the quality of care and service provision. These high-level (and very abstract) goals were expected to be achieved across the whole care pathway, to cover all relevant patient populations and risk groups, to be sustained beyond the funding period, and to generate lessons that could be applied elsewhere. Running through the MI's early strategy documents (although, interestingly, not mentioned explicitly) were the modernization goals from the NHS plan proposed in the first sentence of this article. The evaluation was expected to provide formative feedback to the teams as the projects unfolded, to assess the impact of the initiative, to capture wider lessons from the relationship between context and process in transforming the service, and to advise the sponsor on its future investments in the local health economy.

Research Design

In a radical break from traditional health care evaluation in the UK, the MI's financial sponsor acknowledged the unpredictable and iterative nature of transformational change and asked the evaluators not to become tied to predefined milestones or fixed metrics of success. Reflecting this emphasis, we also adopted an interpretive case study design, drawing on the principles of realist evaluation (Pawson and Tilley 1997). The main analytic challenge was not to determine whether or not the modernization effort "worked" but to find out how the MI's fortunes were shaped, enabled, and constrained by interaction between the context of the program and the chosen mechanisms of change.

The evaluation took place between January 2004 and April 2008. A preliminary phase of the MI involving consultation and planning had begun about eighteen months earlier.

The design of the evaluation took into account that each of the three MI projects had numerous objectives and multiple work streams operating across the local health economy (and in some cases beyond it, linking with social services and the voluntary sector). A dynamic local context and wider policy environment influenced their progress in unpredictable ways, and the different subprojects were continually being modified as each one developed and benefited from experience. For all these reasons, any attempts to establish linear, causal relationships between inputs and outputs (e.g., by comparing "before" and "after" data or an "intervention" and a "control" patch) would have been meaningless.

Data Sources and Analysis

To capture the complex and dynamic nature of the modernization effort unfolding over time, we drew on, both pragmatically and reflexively, a wide range of data sources, methods, and materials. These included ethnographic observation at MI management meetings at all levels (the MI board, individual project management groups, and specific workstream leadership groups), at other project activities and events, and within the service itself (e.g., outpatient clinics, dialysis units); approximately 100 semistructured interviews with staff and service users over the three-year study period; group interviews and informal discussions with the MI projects' staff and stakeholders; and scrutiny of minutes, papers, reports, and quantitative and qualitative data collected by and for the projects.

We analyzed this large and heterogeneous data set in different ways for different purposes. We provided formative feedback to the project management groups every one to two months and held numerous informal discussions with project managers and other stakeholders, at which we presented findings relevant to key decisions. Three times a year, we reported our progress to an advisory group set up by the financial sponsor to oversee the evaluation, and periodically we presented our findings to the MI board. These evaluations were in the general format of highlights from an unfinished case report and included qualitative and quantitative data as well as preliminary interpretations. The MI program director and the NHS organizations were represented in our advisory group and were an essential source of respondent validation throughout the study. Our final report for the sponsor was 100 pages long, detailing the MI's history and context, the main activities undertaken in the different work streams, a preliminary analysis, and recommendations for future funding decisions (Greenhalgh et al. 2008).

With a view to drawing transferable lessons about the process of change, we also undertook a more abstract analysis of our complex data set using the realist approach (Pawson and Tilley 1997). This approach explores the relationship over time among "context" (the study's organizational setting and external constraints, including financial and human resources, prevailing policies, and technologies), "mechanisms" (the stakeholders' ideas about how change will be achieved in an intervention), and "outcomes" (the intended and unintended consequences of the change efforts). In a realist analysis, this relationship is not seen as fixed. Rather, particular preconditions are seen as creating what realists call *generative* or *conditional* causality. This assumes that innovations, programs, and interventions will work only in particular circumstances and that the purpose of the evaluation is to find those conditions: Which mechanisms work, in which contexts, and to produce which outcomes?

A theory-testing strategy is used to unpack this generative causality. The basic ideas, or "theories of change," behind a proposed change program are elicited and then are examined for their utility and efficacy across the evaluation's various contexts (e.g., different services, wards, units). The experiences of the program in these different cases then are used to refine the program theory. The objective is to better understand why and when innovations work.

As with all interpretive case study research, a realist evaluation collects and analyzes multiple data sources and constructs from these a coherent and plausible account of key events and actions and their intended and unintended consequences (George and Bennett 2005). The realist methodology cannot be expressed simply in technical or sequential terms (first do X, like this, then move on and do Y, like this). Rather, it uses all the following approaches judiciously and in combination:

- Organizing and collating primary data and producing preliminary thematic summaries of these (e.g., collecting all the heterogeneous data from field notes, interviews, and quantitative audits on a topic such as a waiting-list initiative, and summarizing key themes in a single interim analysis document).
- Repeating the above after an appropriate time interval to capture qualitative and quantitative change (not as a simplistic before-and-after analysis but as a contribution to the emerging picture of change in context).
- Repeated writing and rewriting of fragments of the case study (working mainly from interim analysis documents and using the narrative form as a progressive synthesizing device).
- Presenting, defending, and negotiating particular interpretations of actions and events both within the research team and also to the stakeholders themselves (interpretations that in turn require reflection on why key stakeholders contested, or were disappointed with, emerging findings).
- Testing these interpretations by explicitly seeking disconfirming or contradictory data (e.g., if our findings appear to show that teams do not draw on research evidence in their decision making, we will redouble our efforts to find examples of situations in which they do seek and use such evidence).
- Considering other interpretations that might account for the same findings (e.g., when feedback from mystery shoppers indicates that the service is getting worse, is this because the service is getting worse or because as they gain experience, mystery shoppers become more able and confident in identifying and articulating its flaws?).
- Using cross-case comparisons to determine how the same mechanism (such as "integrating services across providers") or submechanism (such as "introducing boundary-spanning roles") plays out in different contexts and produces different outcomes, thereby allowing inferences about the generative causality of different contexts.

The pursuit of rigor in realist evaluation embraces the principles of interpretive case study in general. Much rests on carefully defining and justifying the "case," achieving immersion (i.e., spending enough time at the field site to understand what is going on), collecting information meticulously and analyzing it systematically, encouraging reflexivity in both researchers and research participants, developing theory iteratively as emerging data are analyzed, seeking disconfirming cases and alternative explanations, and defending one's interpretations to both the research participants and one's academic peers (Stake 1995). Later we discuss the challenges associated with the realist evaluation method.

Main Findings

Overview

The MI was a large and heterogeneous program of work involving scores of staff and extending over a period of almost five years (including an eighteen-month development phase, three years of formal funding, and up to six months' overrun to give late-starting work streams a "soft landing"). The program was heavily influenced by the concepts of quality improvement and service redesign that pervaded thinking about health care organization in both the United Kingdom and United States in the early 2000s (IOM 2003). The project teams drew on an eclectic mix of approaches, including reengineering, total quality management, and lean thinking, which share a focus on the patient or customer and the quality of each one's experience, an emphasis on the patient pathway, and a commitment to improving efficiency and clinical excellence. Because of the program's scale and scope, the mechanisms of change espoused or assumed by different work streams were difficult to unpack. As the program unfolded, new projects and subprojects emerged and some existing initiatives died, changed direction, or were rebranded as something else. Our approach of separating the change process into discrete "mechanisms," while useful at an analytic level, is undoubtedly artificial and fails to reflect their interdependent nature.

With these caveats, we found that six broad mechanisms of change, each of which comprised a number of submechanisms, were evident in the modernization effort. Although the project teams rarely made these mechanisms explicit, they were clear from both their written strategies and their actions. The three MI projects drew on all these mechanisms, but the "same" approach unfolded differently because of the organizational structure and culture of existing services, the nature of the conditions being dealt with and their trajectories over time, the characteristics and circumstances of the patient groups involved, and the particular aspirations of both patients and staff. Taking these (and other) contextual elements into account and using three of the six mechanisms (integrating services across providers, finding and using evidence, and involving service users in the modernization effort) as illustrative examples, we next discuss the key enabling and constraining factors that appeared to make each mechanism more or less likely to produce a desired outcome in any particular set of circumstances. Our findings on three additional mechanisms (supporting self-care, developing the workforce, and extending the range of services) have been omitted for length reasons but are available from the authors.

Mechanism 1: Integrating Services across Providers

Poor coordination of care across organizational boundaries is a leading cause of quality failure and poor patient experience worldwide (Hoffmarcher, Oxley, and Rusticelli 2007; IOM 2003). The need to deal with duplication, fragmentation, and inconsistency of services was a recurring theme in the MI's strategy documents. Early work undertaken in each of the three projects suggested that patients' experiences typically were disjointed and confusing, and the stakeholders were unanimous in seeking a more streamlined, consistent, and "seamless" experience for the patients, oriented to an integrated patient pathway. Pathways were seen as both an end in themselves (because care would be more patient centered) and a way of improving quality within and across organizations. Ways of promoting integration included the following:

Establishing Boundary-Spanning Roles. The MI created a number of new cross-boundary roles, including the MI director, senior project managers, service improvement facilitators, clinical champions (senior doctors whose time was "bought out" by the MI for one or two sessions per week), and a network manager. Because they were employed by the MI, these individuals were neutrally positioned in relation to the various service providers and therefore were well placed to work with all of them. One of the principal objectives was to break down organizational silos, and the boundary spanners appeared to be particularly successful in bridging the interface between primary care and acute care hospitals. But as outsiders, the MI-funded staff had less power to make changes within particular organizations. They did try to create NHS positions shared by hospital trusts, but their efforts were often stymied by the rigidities of human resources policy.

Developing and Implementing Shared Guidelines, Protocols, and Pathways. Even though there was much enthusiasm for a shared approach to care with common guidelines and protocols, this often proved difficult and time-consuming to achieve in practice. The dissent among the organizations sometimes centered on the evidence itself but more often focused on practical considerations, procedural or presentational issues, or the tension between simple, "minimum standard" guidelines that could be readily implemented across the board and more detailed and ambitious "gold standards" defining best practices for particular groups. Once a protocol or model of care had been approved, its implementation depended on numerous practical factors, especially the capability and capacity of the different organizations to deliver on their agreed component and the "unbundling" of funding so that money could follow the patients across organizational boundaries (e.g., a shorter stay in hospital after a stroke was linked to funds being transferred to the community for supported discharge activity).

Introducing Shared IT Systems and Common Data Sets. Initially, all the MI project teams had high hopes that the new IT systems would improve information sharing and help achieve the goal of "seamless care." To some extent, however, all were disappointed. Immaturity of technical solutions, lack of interoperability with legacy systems, scope creep, escalating costs, limited staff time and skills, variable data quality across organizations, and emerging large-scale IT strategies both nationally and hospitalwide all contributed to a climate of uncertainty and risk. The well-recognized tension between "technology push" and "sociotechnical change" also was evident in some areas. While some significant progress was made with developing common data sets, no major new IT system was launched successfully during the MI's lifetime.

Developing Networks and Supporting Networking. An online sexual health network established by the Sexual Health MI helped break down barriers among different services, facilitating agreement of joint protocols and guidelines, and sharing learning and best practices (Baraitser, Alessio, and Brady 2007). This success was partly due to personal input by the network development manager to engage the different stakeholders and listen to their views and expectations of the network. Early work in this area exposed a wide range of stakeholders with different views and

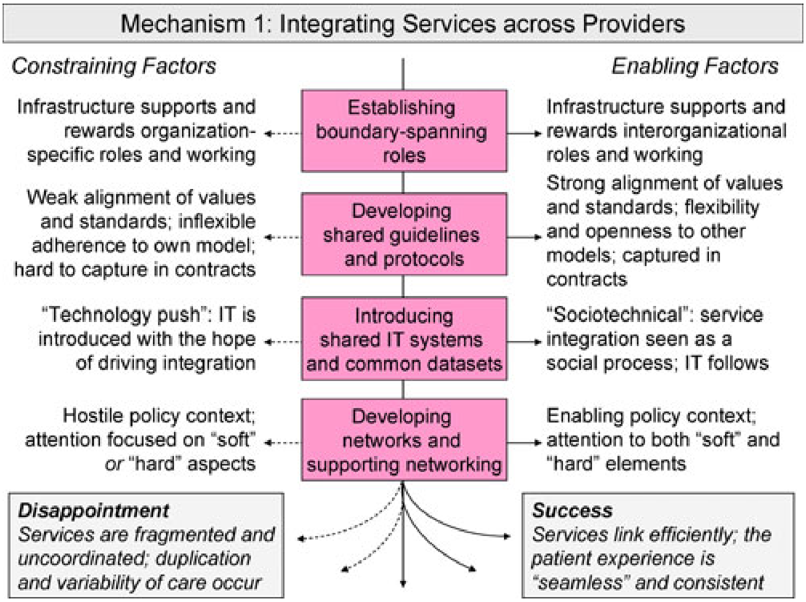
Mechanism 1: Integrating Services across Providers
Constraining Factors
Enabling Factors
Infrastructure supports and rewards organization-specific roles and working
Establishing boundary-spanning roles
Infrastructure supports and rewards interorganizational roles and working
Weak alignment of values and standards; inflexible adherence to own model; hard to capture in contracts
Developing shared guidelines and protocols
Strong alignment of values and standards; flexibility and openness to other models; captured in contracts
"Technology push": IT is introduced with the hope of driving integration
Introducing shared IT systems and common datasets
"Sociotechnical": service integration seen as a social process; IT follows
Hostile policy context; attention focused on "soft" or "hard" aspects
Developing networks and supporting networking
Enabling policy context; attention to both "soft" and "hard" elements
Disappointment
Services are fragmented and uncoordinated; duplication and variability of care occur
Success
Services link efficiently; the patient experience is "seamless" and consistent

FIGURE 1. Realist Analysis of Attempts to Modernize by Integrating Services across Providers

expectations, but the very process of capturing these views and feeding them back both individually and collectively was itself a step toward collective sense making (Weick 1995). Although the field of sexual health will probably never have a unified perspective (because some decisions are value driven), the careful groundwork helped achieve buy-in to the concept of a network and allowed different members to tolerate some diversity within it. No formal networks were created for stroke and kidney care, although there already was a regional clinical network for renal services), but much work was undertaken by the Kidney MI to bring people from different organizations together in "networking" meetings and events. This reinforced relationships and stimulated collaboration, goodwill, and trust.

Summary. Given the long history of fragmented services and intertrust rivalries, it is not surprising that this mechanism for modernizing services was not easy to implement (see figure 1). Overall, the experience of the MI suggests that efforts to achieve integration across providers are more likely to succeed when

- Relationships among organizations are characterized by mutual trust, a history of collaboration, and compatibility of values rather than mutual suspicion, a history of competition, or a mismatch of values.
- Approaches to integration are imaginative, locally responsive, negotiable, and supported by technology rather than rigid, formulaic, nonnegotiable, and driven by technology.
- External incentives (e.g., policies) are designed to reward collaborative performance and do not pit organizations against one another.
- The strategy for integration includes both “soft” and “hard” approaches rather than focusing exclusively on one or the other of these.
- Solutions are participatory (negotiated and owned by all stakeholders) rather than developed by one party and imposed on others.

Mechanism 2: Finding and Using Evidence

The MI developed against a backdrop of widespread commitment to the concept of evidence-based policymaking (Department of Health 2000; Parsons 2002). The MI teams were strongly encouraged to collect “evidence” and feed it into the design and development of new services. Besides published research findings from epidemiological studies and clinical trials, the sources of evidence that were regularly used in commissioning and policymaking included assessments of local burden of need, surveys of patients’ attitudes and experiences, models of good practice from elsewhere, economic evaluation and modeling, consultations with staff and other stakeholders, and dynamic, ongoing feedback on performance. We discussed several approaches to finding and using evidence in connection with other mechanisms (e.g., producing shared guidelines as described in the previous section). In addition are the following:

Using Published Research Evidence to Inform Service Development. Published evidence to feed into a service development or evaluation activity was rarely proactively sought, although one or two senior doctors (who held academic roles and had been involved in the primary research themselves) were enthusiastic and skilled in applying this evidence to their own clinical practice. Sometimes a search produced relevant, credible, and timely research evidence for a key decision about development or delivery. But at other times, these searches proved fruitless, were out of step with the decision-making cycle, or produced evidence of questionable relevance or validity. In general, the "best evidence" from the research literature rarely flowed smoothly and uncontested into practice, a finding resonating with the work of many other research teams (Russell et al. 2008).

Generating and Using Data about Local Needs and Services. All three MI projects tried to map the need for and monitor the provision and impact of services. Initially, it was envisaged that for each work stream, a set of "baseline" metrics would be collected on the burden of need and that the initiative's success would be measured partly by changes in these metrics. Some work streams were able to identify metrics that were important, practicable, and valid (e.g., geographical distribution of patients, waiting times, hypertension process, and outcome measures in general practice). But others found it impossible to identify valid metrics because of shifting denominators, changing clinic demographics, changes in tests undertaken or laboratory normal ranges, and the iceberg of unmet need that was revealed as services improved.

Capturing and Utilizing Experiences of Staff and Users. A number of surveys, some commissioned from commercial market research companies, were undertaken to capture the perspectives of staff, service users, and/or potential users. The MI participants generally viewed local surveys as a valid, robust, and relevant way of generating evidence for decision making. Accordingly, they generally were more apt to seek local evidence, as they regarded it as more authentic and more likely to be used in decision making, than was evidence from published research papers. But surveys aimed at patients' attitudes or experience were found to require a high level of skills (and significant resources) to analyze, and the knowledge generated from such instruments was sometimes complex and opaque. For example, although a quality-of-life instrument used by the Kidney MI was "evidence based" in its validity and reproducibility, it turned out to be cumbersome to administer, and the findings were hard to interpret.

In a few cases, the questionnaires created by MI teams to survey attitudes and preferences were little more than "straw polls," with low psychometric validity, a convenience sampling frame, and little or no statistical analysis. But interestingly, this did not necessarily invalidate the exercise. For example, the Sexual Health MI surveyed both staff and service users with a questionnaire to identify which processes and procedures were considered appropriate for "self-management," "primary care management," and "specialist management." Even though the methodology was flawed, participation in the exercise (designing, completing, and interpreting the questionnaire) was itself a powerful engagement experience that appeared to increase many stakeholders' acceptance of the new service model.

Another approach used particularly by the Kidney MI was the "whole systems" event: a structured one-day meeting between service users and providers regarding a particular aspect of care, such as predialysis services. In the morning, service users and staff met separately and talked about their experiences of services relating to the chosen theme. After lunch, the groups came together to compare perspectives, discuss problems, and generate possible solutions. Subsequently, both users and members of staff gathered in groups to plan implementing different aspects of the work.

Visiting Systems in Action Elsewhere. The MI's generous resourcing enabled visits by teams of staff and patients to other centers in the UK and abroad to view innovative service models in action. The Kidney MI, for example, organized visits to Kaiser Permanente in California and a dialysis center in Brussels. Such visits were greatly valued. Watching others at work allowed the transfer of embodied, tacit knowledge and practical advice that would be difficult to convey in a manual or protocol. "Seeing with your own eyes" was a very powerful persuader of what might be possible, and rubbing shoulders with enthusiasts and practitioners helped build motivation to go home and get started (or, occasionally, convinced people that they did not wish to replicate certain service models and practices).

This form of evidence gathering also had a downside. A system that worked "over there" might inspire and motivate but nevertheless would not be workable "back here" because of contextual differences (institutional, professional, cultural, or economic). For example, Kaiser Permanente's system of managing chronic kidney disease depended on having a single health maintenance organization that provided both primary and secondary care for its members. The visit to Kaiser led to a brief preoccupation with technical solutions (Kaiser's success was attributed to a particular shared IT system), but the model was not readily transferable to the UK, where primary care is provided by multiple independent contractors; acute care hospital trusts have historically operated

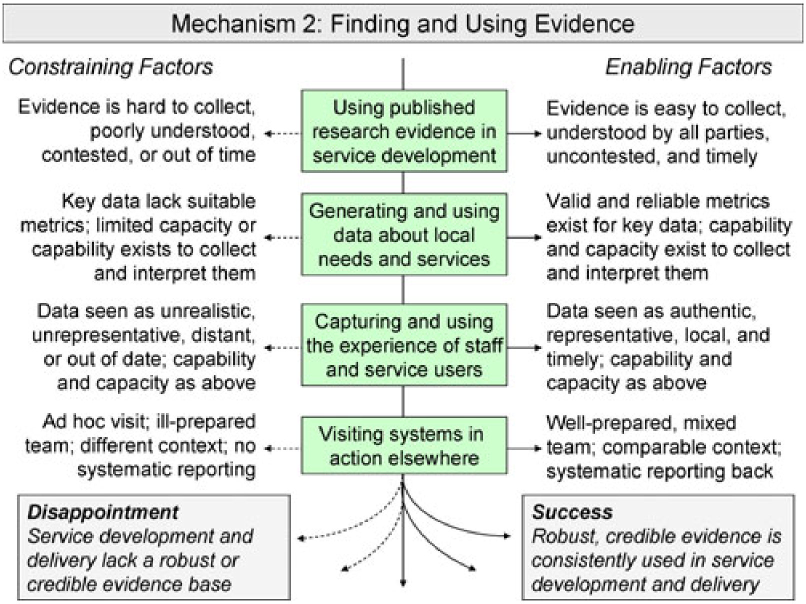
Mechanism 2: Finding and Using Evidence
Constraining Factors
Enabling Factors
Evidence is hard to collect, poorly understood, contested, or out of time
Using published research evidence in service development
Evidence is easy to collect, understood by all parties, uncontested, and timely
Key data lack suitable metrics; limited capacity or capability exists to collect and interpret them
Generating and using data about local needs and services
Valid and reliable metrics exist for key data; capability and capacity exist to collect and interpret them
Data seen as unrealistic, unrepresentative, distant, or out of date; capability and capacity as above
Capturing and using the experience of staff and service users
Data seen as authentic, representative, local, and timely; capability and capacity as above
Ad hoc visit; ill-prepared team; different context; no systematic reporting
Visiting systems in action elsewhere
Well-prepared, mixed team; comparable context; systematic reporting back
Disappointment
Service development and delivery lack a robust or credible evidence base
Success
Robust, credible evidence is consistently used in service development and delivery

FIGURE 2. Realist Analysis of Attempts to Modernize by Finding and Using Evidence

autonomously; and IT strategy is decided at the national or regional level.

The MI teams sent delegations to various quality improvement conferences and events, initially to gather examples of good practice and specific quality improvement tools, and later to share their own findings and offer resources to others. Such encounters were opportunities to exchange stories, especially among staff in comparable positions in different organizations (e.g., change facilitators or data quality leads), in which tacit knowledge was conveyed about "how things are done" and "troubleshooting" moves suggested for common problems, and also to exchange resources such as tools, templates, and protocols. Similar activities applied to local service changes; for example, regular meetings were held among nurses from the different dialysis units to share their experiences of introducing self-care.

Summary. The realist analysis of approaches to finding and using evidence (see figure 2) suggests that these efforts are more likely to be effective when

- The evidence is locally generated, easy to identify and collect, readily understood, seen in action and as timely, and easily applied in practice.
- Stakeholders are open to different types and sources of evidence.
- Staff have the skills, capacity, and commitment to collect, critically interpret, and apply evidence.
- The evidence is widely accepted.
- The evidence is selected and applied flexibly with sensitivity to context.

Mechanism 3: Involving Service Users in the Modernization Effort

A core principle of the MI was that service users should be involved at all points in the modernization effort. Users were involved in a number of ways:

Leading and Managing Projects and Work Streams. In both the Kidney and Stroke MIs, patients were represented on the projects' steering groups and subgroups (and occasionally chaired these groups). Much effort went into involving the users in the steering groups, as a vehicle for user involvement, and this engendered a strong ethos of accountability to users. Over time, however, some groups shifted from "making or approving decisions" to "receiving progress reports from the project team," and senior managers came to view the meetings as too long and not sufficiently focused on strategic issues. It is unclear whether this shift would have occurred anyway as the projects grew in size and complexity. Regular users of sexual health services (referred to colloquially by staff as "frequent fliers") tended to be regarded as more of a problem than an opportunity, and efforts to include users on project management groups were not sustained.

Evaluating and Piloting Services. One way to involve the users that attracted interest outside the MI was the Sexual Health MI's use of "mystery shoppers": patients or ex-patients who were trained to present a standardized clinical scenario and write up their experiences in different parts of the system (Baraitser et al. 2005, 2008). This feedback was seen as credible and was used extensively in designing the new sexual health service model. For example, mystery shopper data led to reducing waiting times and number of visits, building a culture in which service users' privacy and dignity are respected by all staff, and improving the clinics' physical surroundings to make waiting more pleasant. Nevertheless, several challenges were associated with this approach, including the resources required to train, supervise, and support the mystery shoppers and to ensure that they included representatives from different client groups; the risk that as mystery shoppers themselves became "experts" on the service, their assessments would become more critical, thereby possibly masking improvements over time; the complex nature of data generated by mystery shoppers' visits; and the concern of a few staff that the approach was in an ethical "gray zone."

The service users' commitment and cooperation also were essential to the effective piloting of innovative technologies and service models based on self-care. Daily nocturnal home hemodialysis, for example, or remote monitoring technologies in stroke care needed each patient's full involvement in developing the new model.

Providing Peer Support to Other Users. Both the Kidney and Stroke MI teams trained cadres of patients and carers to provide support for other service users. Local evaluation showed that patients particularly valued talking to someone who had "been through it" and that peer support helped them reach decisions and come to terms with starting treatment. These schemes encountered challenges initially. It took time to find and train sufficient keen and able volunteers, and referrals were initially very low and took longer than anticipated to reach what was regarded as a reasonable level. Time, effort, and ingenuity were required to overcome patients' and clinicians' preconceptions and build confidence in peer support, as well as to incorporate offering peer support into standard clinical care. Despite the slow start, however, peer support became established as part of the mainstream kidney services (Hughes et al. 2008), and it continues with voluntary sector support in stroke services.

Producing Information for Patients and Staff. In all three MI projects, service users helped design and develop information and training materials for other users and staff. Their help ranged from suggesting what was needed, to vetting drafts of materials, to undertaking much of the development and piloting. In the Kidney MI, an early video on living donor transplants made by the consultant led team consisted entirely of "talking heads" (clinicians speaking to camera delivering information that they felt patients needed to know). The next DVD, "Living Life to the Full on Dialysis," was radically different and featured patients talking about their experiences of managing their own dialysis. The project manager who developed the DVD spent several months visiting patients in their homes, inviting them to talk about life on dialysis and what they thought the DVD should cover. The project continued past its deadline and over budget but was widely praised for conveying the "real patient experience."

Service users also directly helped train staff. For example, after helping produce good practice guidance and a DVD, a group of people living with stroke helped present training sessions for health and social care staff. The participating staff then used the training sessions to review their professional practice and to identify areas for local service development.

Advocating via the Voluntary Sector. Voluntary sector input was particularly important when individual service users were less enthusiastic or able to be involved and when particular minorities believed that there was a cause to fight for. In sexual health, for example, most people who used the services did so episodically and did not see themselves as potential "representatives" of other service users. The Sexual Health MI project management group did, however, include strong and vocal representatives from a wide range of voluntary sector organizations. Stroke and kidney charities and user organizations also helped represent the perspective of those less able to advocate for themselves (e.g., because of aphasia).

Summary. Figure 3 shows a realist analysis of the enabling and constraining influences on involving users in modernization work. The experience of the MI suggests that efforts to involve users are more likely to succeed when

- There is a strong tradition of user activism.
- The users' identity and motivation are high (i.e., service users identify positively with, and want to help, other users).
- The condition is chronic; management contains a "registration and recall" component; and at least some users are reasonably physically fit.
- The staff value, and try to implement, the user voices.
- Potential users are readily identified, recruited, and managed (i.e., are known to service providers, relatively numerous, and geographically close rather than dispersed).
- The infrastructural support for users' involvement is strong, enabling, and adequately resourced.

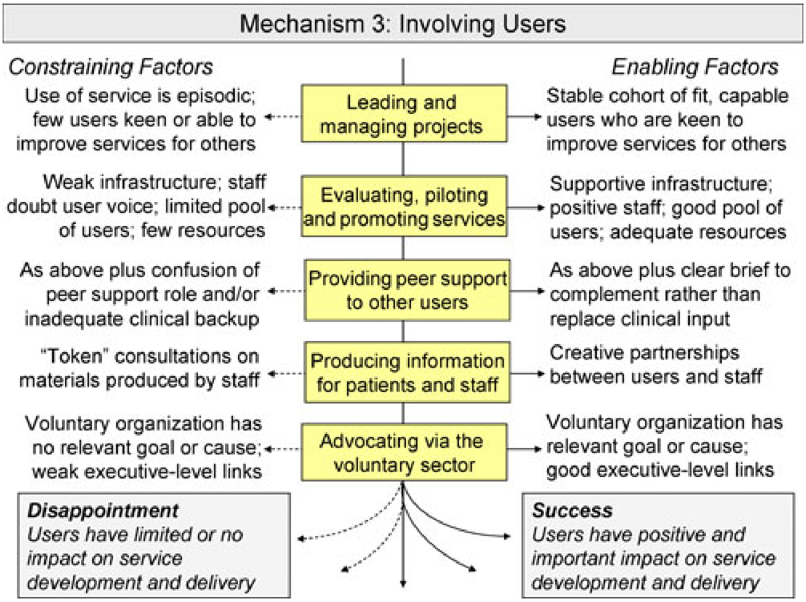
Mechanism 3: Involving Users
Constraining Factors
Enabling Factors
Use of service is episodic; few users keen or able to improve services for others
Leading and managing projects
Stable cohort of fit, capable users who are keen to improve services for others
Weak infrastructure; staff doubt user voice; limited pool of users; few resources
Evaluating, piloting and promoting services
Supportive infrastructure; positive staff; good pool of users; adequate resources
As above plus confusion of peer support role and/or inadequate clinical backup
Providing peer support to other users
As above plus clear brief to complement rather than replace clinical input
"Token" consultations on materials produced by staff
Producing information for patients and staff
Creative partnerships between users and staff
Voluntary organization has no relevant goal or cause; weak executive-level links
Advocating via the voluntary sector
Voluntary organization has relevant goal or cause; good executive-level links
Disappointment
Users have limited or no impact on service development and delivery
Success
Users have positive and important impact on service development and delivery

FIGURE 3. Realist Analysis of Attempts to Modernize by Involving Service Users in the Change Effort

- Networks and connections between users and with staff are strong and sustained (e.g., people know one another from clinic visits or events).

Impact and Sustainability

Our full report describes in detail the tangible benefits, both "hard" (agreed-on, evidence-based protocols; extended opening hours; shorter waiting lists; governance structures for interorganizational working) and "soft" (improved staff attitudes and motivation, greater user satisfaction, and what one senior manager described as a "precious, extraordinary" cultural shift) (Greenhalgh et al. 2008). These benefits were not universal, and not every subproject was an unqualified success. Furthermore, where measurable improvements could be found, the link between them and the MI was not causal in a simple or deterministic way. It was, however, evident that some degree of "whole-scale transformation" was achieved during the study and that most stakeholders attributed this to the funding, vision, ethos, and collective effort of the MI. The MI was characterized by imaginative and sustained efforts to ensure the long-term sustainability of the various gains achieved during the funding period. These are listed in the final report and will be analyzed in a separate article. Briefly, however, they include attention to cultural as well as structural changes; clarification of the resource implications of the new or altered services; the development of strategies for retaining skills and expertise within the local health economy; plans for the continued involvement of users; the maintenance of links with voluntary sector and partner organizations; and a sustained interorganizational structure for governance and formal communication.

Discussion

Our study has shown that even when generous funding is provided, the road to modernizing a health service is neither straight nor smooth. Different actions must be taken at various parts of the care pathway and by many different people (including the patient) in a complex and often rapidly changing context. There is no simple recipe for success. The best laid plans may be stymied by practical, policy, or legal constraints. Positions that seem to be crucial to the project may not attract any suitable applicants. Patients may be too sick or infirm to participate actively in their own care or in the wider change effort. And so on. Conversely, despite these barriers (and others), they do not necessarily predict failure. Creative action by top management and/or front-line practitioners may allow projects in trouble to change direction or feed indirectly into another version of the intervention.

After the program ended, we tried to determine which factors enabled or constrained the approaches to modernization that were taken. We believe that a greater understanding of these underlying mechanisms will help inform similar change programs in the future and so have deliberately not passed judgment on the program's overall "success." As Berg has argued, "The question of whether an implementation has been successful or not is *socially negotiated*" (Berg 2001, p. 144, italics in original). If a team sets out to achieve X but along the way learns things or encounters challenges that convince it that Y is a more appropriate (or practicable) goal, then it will have "succeeded" if it achieves something approaching Y.

In the Sexual Health MI, for example, one initial goal was that all sexual health providers in the locality would work toward common guidelines and protocols, because patients have a right to expect the same standards of care wherever care is delivered. But it soon became apparent that some voluntary sector providers had already produced their own guidelines, which were (for good local reasons) oriented toward particular minority or demographic groups. Other partner organizations held deep religious views and eschewed permissive or "sex-positive" approaches. The Sexual Health MI thus came to redefine success as mutual awareness and tolerance of difference, as well as alignment toward a more abstract goal: the best interests of the individual patient or client. Not only was a single common guideline not achieved (though much progress was made toward it), but a key dimension of success was that the team moved beyond their original rigid goal.

Building the evaluation criteria for expensive, large-scale change programs on such shifting sands is relatively controversial when judged by conventional clinical research criteria, which define "rigor" as the systematic pursuit of well-defined goals, objective measurement of progress, and robust accountability procedures. The *Journal of the American Medical Association* recently felt the need to publish an editorial by Don Berwick arguing that complex, multicomponent policy interventions are essentially a process of social change and that "the effectiveness of these systems is sensitive to an array of influences: leadership, changing environments, details of implementation, organizational history, and much more. In such complex terrain, the RCT [randomized controlled trial] is an impoverished way to learn. Critics who use it as a truth standard in this context are incorrect" (Berwick 2008, p. 1183).

Berwick contends that the evaluation of health care policy must move beyond the RCT and embrace both qualitative (particularly the naturalistic observation of practice) and quantitative methods that originated in engineering and are now widely used in quality improvement (such as statistical process control). Acknowledging the trade-off between "eliminating bias" and "capturing local wisdom," he suggests that it is time for the balance between these competing goals to shift toward the latter. Nevertheless, as reviewers of a previous draft of our article pointed out, much of the health services research community is already convinced of the value of interpretive, context-sensitive research designs and is actively exploring new, interdisciplinary methodologies.

Even though the appetite for approaches like realist evaluation is strong, there currently are very few worked examples of large-scale change programs applied to health care. In a search of the Medline-indexed literature, we found only one empirical study of the use of the realist method (Ssengooba et al. 2007), although a wider search of nursing and social science databases turned up several more (Barnes, Matka, and Sullivan 2003; Byng, Norman, and Redfern 2005; Evans and Killoran 2000; Grocott, Cowley, and Richardson 2002). We therefore were conscious from the outset that our study offered an opportunity to ask both methodological and empirical questions. What lessons did we learn in relation to the former?

First, we found that identifying the mechanisms of change for different activities in a large-scale modernization effort was far more difficult than Pawson's widely cited textbook implies. Neither interviewing frontline practitioners nor applying our own interpretation to their actions produced unambiguous accounts of what they were attempting to achieve or how. Although project management groups were typically very clear about their long-term goals (e.g., to make care more efficient, more evidence based, more patient centered, and more holistic), it was surprisingly rare for them to be able to articulate in real time the medium-term goals of particular subprojects (e.g., to capture the illness experience and feed it into service redesign, to redistribute power between secondary and primary care, and to develop a virtual network in which professionals could share and build tacit knowledge across organizations) that could be studied as candidate change mechanisms. Such goals sometimes (though not always) became clear in retrospect when both practitioners and the research team reflected on the outcomes of particular subprojects or analyzed critical events.

We have tentatively concluded that researchers must anticipate—and learn to tolerate—the mismatch between the realist evaluation's assumption that a set of more or less well-defined "mechanisms of change" can be articulated and tested and the empirical reality in which these mechanisms may prove stubbornly hard to nail. This finding resonates with a realist evaluation of a health-oriented community development program by Barnes and colleagues, who also found that mechanisms of change were hard to detect at the front line and in real time and that when some participants offered candidate mechanisms, others would often contest them (Barnes, Matka, and Sullivan 2003). These authors also found that "surfacing theories of change," albeit useful as a starting point, did not allow a sophisticated analysis of the complex power dynamics among different interest groups, and so they recommended that a realist analysis be supplemented by other theories (such as neoinstitutional theory), in which local politics looms large.

Our second finding was that drawing realist conclusions about the generative causality of particular context-mechanism-outcome alignments is not a logical-deductive exercise. Rather, it is an interpretive task and will be achieved only through much negotiation and contestation. The three figures in this article, for example, though based on a sketch by Pawson on the back of an envelope at an early stage of the analysis, did not "fall out of the data." Instead, each one was the product of a sometimes heated argument among members of the research team that typically involved a three-hour face-to-face meeting as well as lengthy email exchanges and numerous iterations and counteriterations. One bone of contention among us was the level of abstraction at which contextual influences should be expressed. For example, only at a relatively high level of abstraction can "replacing examination rooms with self-management pods in the sexual health clinic" and "early discharge of stroke patients so that the locus of rehabilitation shifts to the patient's own home" be viewed as the "same" mechanism ("redesigning the physical environment for self-care"). We have tentatively concluded that research teams should not only anticipate disputes and deadlocks when producing their realist analyses but that they should also view these as the route to achieving higher-order insights into the change process.

Finally, this study raises the age-old chestnut of the balance that an evaluation team should strike between an emic versus etic position, a formative versus summative analysis, and (interpretive) illumination versus (normative) judgment. A strictly positivist approach would see researchers (and, by implication, evaluators) as separate and separable from the "case" and would define a good evaluation as etic, summative, and normative. In such an approach, engaging with the practitioners, bringing ideas to their meetings, and making suggestions as the projects unfold would count as "bias" and (in the eyes of many purists) thus invalidate the study. But much (though admittedly not all) of the scholarly literature on evaluation recoils from such a perspective, preferring that evaluators engage with the messy reality of the "case" and accompany the practitioners closely on their unfolding journey, hence favoring the emic, formative, and illuminative poles of the duality (Guba and Lincoln 1989; Øvretveit 2002; Patton 1997; Potvin, Haddad, and Frohlich 2001). This article is not the place to open this particular can of worms, but it is worth emphasizing that if you accept a realist evaluation, you (and whoever is sponsoring the evaluation) must also accept its constructionist ontology and interpretivist epistemology. Berwick's view is that this acceptance is still in relatively short supply in the health services research community (Berwick 2008). But there will be blood on the carpet if stakeholders embrace "realist" evaluations but remain wedded to positivist criteria for assessing the rigor of such work.

References

Baraitser, P., G. Alessio, and M. Brady. 2007. Sexual Health Networks: Linking Providers for Improvement. *Journal of Family Planning and Reproductive Health Care* 33(3):193.

Baraitser, P., V. Pearce, G. Blake, K. Collander-Brown, and A. Ridley. 2005. Involving Service Users in Sexual Health Service Development. *Journal of Family Planning and Reproductive Health Care* 31:281–84.

Baraitser, P., V. Pearce, N. Walsh, R. Cooper, K.C. Brown, J. Holmes, L. Smith, and P. Boynton. 2008. Look Who's Taking Notes in Your Clinic: Mystery Shoppers as Evaluators in Sexual Health Services. *Health Expectations* 11(1):54–62.

Barnes, M., E. Matka, and H. Sullivan. 2003. Evidence, Understanding and Complexity: Evaluation in Non-Linear Systems. *Evaluation* 9(3):265–84.

Bate, S.P., and G. Robert. 2003. Where Next for Policy Evaluation? Insights from Researching NHS Modernisation. *Politics and Policy* 31(2):237–51.

Berg, M. 2001. Implementing Information Systems in Health Care Organizations: Myths and Challenges. *International Journal of Medical Informatics* 64:143–56.

Berwick, D.M. 2003. Disseminating Innovations in Health Care. *Journal of the American Medical Association* 289:1969–75.

Berwick, D.M. 2008. The Science of Improvement. *Journal of the American Medical Association* 299(10):1182–84.

Byng, R., I. Norman, and S. Redfern. 2005. Using Realistic Evaluation to Evaluate a Practice-Level Intervention to Improve Primary Healthcare for Patients with Long-Term Mental Illness. *Evaluation* 11(1):69–93.

Department of Health. 2000. *The NHS Plan*. London: NHS Executive.

Evans, D., and A. Killoran. 2000. Tackling Health Inequalities through Partnership Working: Learning from a Realistic Evaluation. *Critical Public Health* 10(2):125–40.

George, A., and A. Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, Mass.: MIT Press.

Greenhalgh, T., C. Humphrey, J. Hughes, F. Macfarlane, C. Butler, P. Connell, and R. Pawson. 2008. *The Modernisation Initiative Independent Evaluation: Final Report*. London: University College London. Available at http://www.ucl.ac.uk/openlearning/research.htm. Ref Type: Report (accessed March 30, 2009).

Greenhalgh, T., G. Robert, P. Bate, O. Kyriakidou, and F. Macfarlane. 2005. *Diffusion of Innovations in Health Service Organisations: A Systematic Literature Review*. Oxford: Blackwell.

Grocott, P., S. Cowley, and A. Richardson. 2002. Solving Methodological Challenges Using a Theory-Driven Evaluation in the Study of Complex Patient Care. *Evaluation* 8(3):306–21.

Guba, E., and Y. Lincoln. 1989. *Fourth Generation Evaluation*. London: Sage.

Herbert, K. 2004. "Arm's Length" NHS Bodies to Be Abolished in Spending Cull. *British Medical Journal* 329:252.

Hoffmarcher, M.M., H. Oxley, and E. Rusticelli. 2007. *Improved Health System Performance through Better Care Coordination. Health Working Paper No. 30*. Paris: OECD.

Hughes, J., E. Wood, S. Cox, L. Silas, and G. Smith. 2008. *"No White Coat between Us." Developing Peer Support Services for Kidney Patients*. Available at http://www.gsttcharity.org.uk/pdfs/whitecoat.pdf (accessed March 10, 2009). London: Modernisation Initiative.

Institute of Medicine (IOM). 2003. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, D.C.: National Academies Press.

Øvretveit, J. 2002. *Action Evaluation of Health Programmes and Changes. A Handbook for a User-Focused Approach*. Oxford: Radcliffe.

Parsons, W. 2002. From Muddling Through to Muddling Up. Evidence Based Policy-Making and the Modernisation of British Government. *Public Policy and Administration* 17:43–60.

Patton, M.Q. 1997. *Utilisation-Focused Evaluation: The New Century*. 3rd ed. London: Sage.

Pawson, R., and N. Tilley. 1997. *Realistic Evaluation*. London: Sage.

Potvin, L., S. Haddad, and K.L. Frohlich. 2001. Beyond Process and Outcome Evaluation: A Comprehensive Approach for Evaluating Health Promotion Programmes. *WHO Regional Publications European Series* (92):45–62.

Russell, J., T. Greenhalgh, E. Byrne, and J. McDonnell. 2008. Recognizing Rhetoric in Health Care Policy Analysis. *Journal of Health Services Research & Policy* 13(1):40–46.

Secretary of State for Health. 2007. Parliamentary Response to Written Question by Rt. Hon. Nicholas Soames MP. In *Hansard*, October 27, London: House of Commons.

Ssengooba, F., S.A. Rahman, C. Hongoro, E. Rutebemberwa, A. Mustafa, T. Kielmann, and B. McPake. 2007. Health Sector Reforms and Human Resources for Health in Uganda and Bangladesh: Mechanisms of Effect. *Human Resources for Health* 5:3.

Stake, R. 1995. *The Art of Case Study Research*. London: Sage.

Weick, K.E. 1995. *Sensemaking in Organizations*. Thousand Oaks, Calif.: Sage.

Acknowledgments: In carrying out the independent evaluation, we received support and cooperation from a wide variety of sources, and many people generously gave their time to discuss ideas and experiences. We are grateful to the staff and service users involved in the Modernisation Initiative, stakeholders in the four trusts, and the trustees and officers of the charity and others who contributed to the Evaluation Advisory Group. We also thank Petra Boynton and Pat Connell for help with data collection. The Modernisation Initiative and the independent evaluation were funded by Guy's and St. Thomas's Charity. The views expressed in this article are those of the authors and do not necessarily reflect those of the MI staff, its partner trusts, or the charity.