Evaluating Complex Social Interventions
Volume 2: Guidance, Tools and Resources

John Øvretveit, jovret@aol.com
Director of Research, and
Professor of Health Innovation, Implementation, and Evaluation
The Medical Management Centre,
The Karolinska Institutet, Stockholm

Reference citation: Ovretveit, J 2013 Evaluating Complex Social Interventions: Volume 2: Guidance tools and resources, CIPRS, Veterans Health Administration, Sepulveda, Ca.

Contents

Summary

This manual gives guidance, tools and resources for evaluating a complex social intervention (CSI). Examples of CSIs include a series of changes to reduce infections, a national program to enable and encourage health services to apply these changes, or a public health campaign. This guidance has selected those tools and resources which are most useful for evaluating CSIs in the Veterans Health Administration, with a focus on producing research findings which VA clinicians and operations personnel can act on. Other evaluation guides are also recommended which give further details.

This guide will enable a researcher to choose the best research design for the purpose and constraints of the research, and to plan, carry out, and publish the research. Increasingly, researchers are expected to help practitioners to act on the research (e.g., by producing tools). This manual concentrates on guidance for how to make research more useful and used.

Guidance is given in part 2 of the manual for each of eight steps for planning and carrying out a CSI evaluation. Part 3 gives more detailed guidance for choosing and using an evaluation designs. Part 4 shows frameworks for deciding which data to gather about the CSI and its context. Part 5 gives suggestions for different methods for collecting and analyzing data for the evaluation. The appendices give further tools and resources.

This guidance manual is intended to be used with the Volume 1 report, which describes issues and solutions in evaluating complex social interventions (Øvretveit 2013).

1. Introduction

Complex social interventions (CSIs), such as new services or programs to change provider behavior, are expensive. They take time and resources from other activities. Evaluations to discover what their effects are, and the time and costs which they consume, are much needed. Some evaluations may be able to show which parts of the intervention or service have the biggest impact and how and why the intervention achieves its effects. This can help to modify the intervention or the new service to be more effective in different situations.

This manual gives practical guidance, tools and resources for evaluating CSIs. It is intended to be used with a Volume 1 report, which describes why special guidance is needed to evaluate CSIs. Volume 1 considers issues and solutions in evaluating these types of interventions and the difference between complex interventions and complex social interventions (Øvretveit 2013). There are other guidance documents and texts on evaluation – this guide does not repeat these, but refers to them for more details about the methods which are relevant to CSI evaluation.

This guide will help evaluators to answer five key evaluation questions:

- **Aims**: Who is the customer for the evaluation? Which information do they need to inform more effective actions? What are the questions to be addressed?
- **Description**: What are the details of the intervention, the implementation, and its context?
- **Attribution**: How certain can we be that the intervention caused the outcomes reported?
- **Generalization**: In which situations could others copy the intervention and get similar results?
- **Usefulness**: In which situations are the intervention and implementation feasible? How do we enable users to use the findings from the evaluation?

The compromises

Designing and carrying out an evaluation means balancing often-conflicting requirements: between the ideal for scientific rigor and what is feasible, given the time and resource constraints.

This guidance suggests that the way to decide the right balance is to understand the research customer's needs for data, information and actionable knowledge. This is what best enables the customer to make decisions which result in faster and more effective action than would be the case without the evaluation. The focus in this guidance is on evaluation which aims to provide users with actionable information which saves them time and money, which should more than justify cost of the evaluation.

2. Eight steps for choosing design and carrying out the research

This part of the guide presents the purpose of each of eight steps for carrying out a user-focused CSI evaluation.

The eight steps in carrying out a CSI evaluation:

1) User goal-specification
2) Reviewing relevant research
3) Defining practical and scientific questions
4) Listing and choosing designs
5) Preparing for the evaluation
6) Data gathering
7) Data analysis
8) Reporting, publishing and dissemination activities

More detailed guidance is given for some of these steps later in part 3 (designs), part 4 (frameworks), and in part 5 (data gathering and analysis). Others providing general guidance recommended for evaluators of CSI are OBSSR (2013), MRC (2008), Bowen (undated), Kellogg (1998), PHAC (2013), CDC (1999) and Øvretveit (2002).

2.1. User goal-specification

Who is the evaluation for and what should it enable them to do better?
Answering this question defines the purpose of the evaluation, and from this, all else follows.

Key points

- Define the primary user group for the evaluation (the "customer").
- Define the information the user needs to make a key action or decision more effective – this is the difference the evaluation is expected to make for the user's work.
- Do this by working with the users to define the timescale and resources for the evaluation, and the type of information which different constraints will allow different evaluation designs to provide.

Who is it for, and to inform which decisions?

Before planning design and data gathering, the evaluator needs to be clear about who the evaluation is for (the "user" or "customer" of the evaluation), and the user-decisions which the evaluation is intended to inform. The user's goals for the evaluation are the goals of the evaluator. For example: "I need to know if this CSI really does improve physicians' use of clinical decision support, and by how much." Or "I need to know if the strategy used by the service in another state is an effective way for us to implement the chronic care model".

There are two main types of users:

1) Practitioners (clinician, manager or policy-maker) and possibly patients
2) Academic users (researchers and educators)

Researchers are usually oriented to satisfying academics as their primary customer. Other academics make decisions about funding or publications, which decide the researcher's academic career. But funders, publishers, and others are increasingly expecting an evaluation to have a direct practical use to practitioners. In the Veterans Health Administration (VA), some evaluation funding comes from service delivery budgets, where the primary users are practitioners.

Evaluations which focus on one category of users, and the key decisions which they want to be informed by the evaluation, are able to focus their resources, design and data gathering on one purpose and to be more successful for this purpose. No evaluation can answer all users' questions, and few can answer more than one or two. But if performed well, there are usually publications which can follow, so long as the purpose and limitations are described in the publication.

Dialogue for defining the question and constraints

User goal-specification is best carried out by the evaluator discussing and negotiating with representative users or the funder about which decisions the user needs to make that could be informed by the evaluation. Users will have many questions they want answered, but each extra question has a time and a cost attached to it. Researchers need to develop skills and methods to define - with the users - which primary questions are to be answered, because all later decisions about design and the data to collect follow from this. This also involves checking with the user whether an answer to these few questions can really make the user's actions more effective and faster.

Even though this step is the most important, there are few methods and resources to help evaluators in working with users to define the question. Nearly all guidance is for academic research, which is primarily driven by previous research and disciplinary interests. It is certainly important to use previous research as part of the process of defining the question (step 2 below), but this should be only part of the process in a user-directed CSI evaluation.

Some general guidance about defining research questions is useful, especially the section on research questions provided by Robson (1993). One short web resource is provided by Apodaca (2013). For evaluations of some CSIs, the Preskill and Jones (2012) five-step process for engaging stakeholders in developing evaluation questions may be useful. This includes four worksheets for a stakeholder "engagement process."

The MRC guidance (Craig et al, 2008), specifically case study 14, gives an example of involving one type of user (a community) in the design and conduct of an evaluation. It proposes that "involving communities in the design of an evaluation is not just compatible with the use of rigorous methods, but can also improve them," and that,

> *"Memoranda of understanding were signed with community organisations to make explicit the obligations on both sides and the fact that all participants would receive the intervention. These organisations were closely involved in conducting the study, for example by organising community meetings in which the study was explained to possible*

participants, recruiting local interviewers, and organising meetings to disseminate early results. Local health workers were employed to recruit participants into the study."

Summary

An actionable evaluation:

- is designed to give an evaluation "user" or "customer" evidence which helps them to make more informed decisions about what they should do, at the time that they need it.
- is best designed for one user, and utilizes the user's criteria of evaluation to decide which evidence to collect.
- involves collaboration with the evaluation user to clarify why the user wants the evaluation, the user's questions about the intervention and outcomes, which decisions the evaluation is to inform, and which criteria to use to judge the value of the intervention,
- is based on decisions about the design and data gathering that best meet the needs of the user within the timescale and resources available for the evaluation.

Does this mean that a CSI evaluation concentrates on one group of users' questions and ignores the perceptions of the other stakeholders with an interest in a health program or change?

The answer is, "yes" and "no." Yes, a focus on one group of users' questions and decision-information needs is necessary to deliver an evaluation which can be used to inform real decisions. The "no" answer is that the evaluation does not ignore the perceptions which different groups have about the CSI, and does gather data from different perspectives. The reason is that for the user to make informed decisions, she needs to know what different parties think about a CSI. It is true that we can find out some effects of a program or change without looking at what people think and by making objective measures or using statistics. But to value and explain a health program and change we usually need to understand different people's perceptions. What people think of as the outcome is an outcome in itself. What people think about a program or change which they are exposed to can influence what happens.

2.2. Reviewing relevant research

Are the users' questions answered in whole or in part by already-published research? Showing what is known and any gaps in knowledge is necessary to avoid duplication and for getting funded and published.

Key points

- First, do a simple search with Google Scholar and then PubMed. Set a time limit for this of 2 hours and try different search terms.
- Depending on the time and resources available, do a more detailed search and keep a record of the search and findings.
- Use the search findings in step 3, to further define the evaluation question.

Step 2 is reviewing the relevant research to find if the user's questions are answered in whole or in part by already-published research. This step is particularly important if the primary user is other academics, and is also necessary for any publication in a peer-reviewed journal and for preparing a proposal for research funding. The purpose of academic research evaluations is to contribute to empirical, and ideally, theoretical knowledge in the discipline. The review shows what is already known and where the gaps or controversies are in the literature.

Reviews can be simple or long and complex and everything in-between. How much time is spent on this stage and which methods are used depends on the timescale and resources available. Rapid reviews are systematic, but are usually carried out with a six-week to six-month time target – one guide to such is written by Ganann et al (2010). Mays et al (2005) also give guidance.

Another method, where the literature is spread in many different databases, is to use a "management review method." This uses databases of published research, already completed evidence reviews, and the evaluator's existing knowledge of research on the subject in an iterative approach to combine different sources and types of evidence (Greenhalgh *et al* 2004; Greenhalgh and Peacock 2005; Øvretveit 2012). The steps are as follows:

1) **Broad scan.** Define the objectives and search terms for the review. Find and note the literature on the subject.
2) **Narrow the focus on previous reviews.** Identify and select previous reviews. Assess these for answers to the review questions.
3) **Open-out inclusion.** Bring-in high-quality individual studies in order to provide additional evidence to answer the review questions, noting the strength of evidence of the findings and assigning a grade score (e.g., the GRADE scoring system (Guyatt et al 2008))
4) **Open inclusion more widely.** Add other research (of acceptable evidence strength) to fill in the evidence for the questions, noting that the evidence at this level is weaker, and using a snowball approach to identify relevant studies.
5) **Review and synthesize.** Combine the evidence in order to answer the questions, noting the degree of certainty (through the grading system). Identify unanswered questions and priorities for research, and provide any recommendations that are supported by the evidence.

2.3. Defining practical and scientific questions

Step 3 brings the review of previous research together with the user definition of the purpose of the evaluation to define both the practical and scientific questions to be answered by the evaluation.

Key points

- Define the evaluation parameters or resources: the report target-date, the budget and skills available for the research, and how much data you can use which is already collected.
- Be clear about whether the primary user is practitioners or the academic community.

- Define the practical questions as those which will ultimately reduce suffering and costs and the academic questions as those which will build theory or contribute to empirical knowledge shown to be needed by previous research.
- This then allows the choice of the design most suited to answering the questions within the constraints of time and resources available.

2.4. Listing and choosing designs

Step 4 is to list potential designs and make a choice about which will provide acceptable answers within the timescale and resources available.

Key points

- The design is best presented in a time-diagram. It shows which data will be collected and when, and the comparisons to be made to answer the evaluation question.
- Some designs are ruled out by the timescale and resources available for the evaluation, or by other constraints or practical issues.
- Experimental and quasi-experimental designs are suitable for effectiveness-outcome questions.
- Observational designs can be used for both effectiveness and implementation questions.
- Action evaluations may be suitable for questions about how to improve the intervention while it is being implemented.
- When considering which outcome data to collect, consider whether accessible data already exists. Any special primary data collection by evaluators is costly in time and money.

Some designs are ruled out by practical issues such as non-availability of comparison units, difficulties randomizing units, or unavailability of time-series data before the intervention was started. The design shows which types of comparisons can be made so as to allow the value of the intervention to be assessed. The three most common types of comparisons are:

- Between what was planned and which outcomes were achieved (Did the intervention achieve the outcomes intended?)
- "Before" compared to "after" (or "later"), referring to measures of a characteristic of the person or unit receiving the intervention (Does this show the intervention made a difference in this characteristic?)
- Outcomes, compared to what happened in a comparison group (comparative effectiveness)

More details about each type of design are given in section 3 of this guidance. The following references give the best overviews of different designs and their strengths and weaknesses: Mercer et al. (2007), Tunis et al. (2003), Craig et al. (2007), Fan et al. (2010), and Robson (1993), and Øvretveit (2002).

There are different views about whether an evaluation framework or model of the intervention is part of the design stage, or part of "preparing for an evaluation." Such frameworks or models should

certainly be considered in this step, or the next step, as a framework helps to decide which data to collect and the methods to use. A later part of this document gives guidance about frameworks and models.

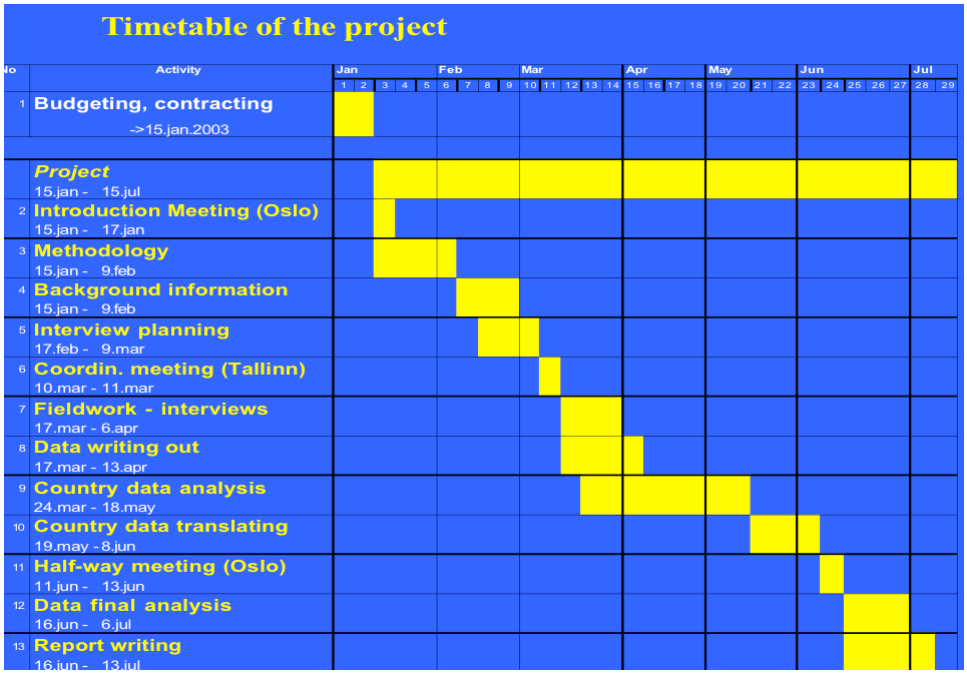
2.5. Preparing for the evaluation - practicalities

Once design is decided, then practical arrangements can be planned, using a time-task map about when and how to gather data. This shows who does what, and the steps for IRB/ethical approval.

Key points

- Decide the date for the last activity in the evaluation (e.g., submission of report, or presentation to users),
- Put this activity and date at the bottom of a Gantt chart,
- List other activities and add them under dates in the Gantt chart,
- Note any critical activities which need to be completed before any next activities can continue,
- Note whether your earlier assumptions about the people, their skills and time for the evaluation still apply, and double check their availability for the times suggested by the plan,
- List events which could slow down the evaluation and possible ways of predicting or mitigating these.

A simple Gantt chart lists the tasks in a vertical column, and the dates across the top row of the chart:



A simple discussion of Gantt charts is given in the wiki outline: http://en.wikipedia.org/wiki/Gantt_chart

With regard to IRB guidance, details for when and how to apply in the VA are provided by DVA (2011). One part of planning is deciding how and where data will be stored and ensuring the security of data.

2.6. Data-gathering

To answer the evaluation questions, data needs to be gathered from different sources, using different methods, and stored in a way to make analysis easy.

Guidance for the data needed to describe the intervention and to report details including context are given by the SQUIRE guidelines (Ogrinc et al, 2008) and Michie et al (2009).

Key points

- Data can be in qualitative or quantitative form and collected using data collection methods which are known to most researchers.
- Qualitative and quantitative data will be needed to describe the intervention as it was actually implemented, and about possible outcomes at different times, as well as about aspects of the context of the intervention.
- Data about immediate and later outcomes are needed, which can be from quantitative before/later measures, or qualitative assessments about outcomes by informed observers.
- Which data to collect is best decided by a model or framework of the intervention, which shows the component parts of the intervention and different outcomes at different times.
- Try to use already-collected data which you can access, if these are of acceptable validity and reliability. Any other data you collect yourself for the study will cost considerable time and money.
- Decide the best balance between concentrating resources on a few data collection methods to increase the reliability and validity of the data and using a number of methods to see if the same findings are to be found in different data (data triangulation).

Many researchers know about most of the data collection methods which need to be used in a CSI evaluation. Nevertheless, more guidance is given in a later section of this document to review, as multiple methods are often needed.

2.7. Data analysis

The aim of this stage is to use the data to describe the CSI changes which were actually implemented, the outcomes which are of interest, and to note the limitations of the findings.

Key points

- Analysis is far easier and quicker if the data, when collected, was stored in a way which was organized with an eye toward how the data would later be analyzed.

- Analysis should start by describing the CSI which was actually implemented, since what was implemented determines which outcomes are worth looking for in the data.
- Computer software for analysis of both quantitative and qualitative data can save time and money, but depends on using the software to explore questions defined earlier in the evaluation.
- If you are not familiar with the software, try to get another researcher experienced with it to guide you through how to use it on your data. Learn by doing - courses are expensive and often less useful.

Analysis should start by describing the CSI which was actually implemented. This is because some planned changes may not have been implemented and some outcomes expected from these changes are therefore not likely. Time is better spent assessing outcomes which are likely from the actually-implemented changes. A later section in this document gives more guidance on different data analysis methods, including multi-level modeling, which can be used to explore variations within groupings at each level.

2.8. Reporting, publishing and enabling use of the evaluation

An evaluation is only effective when it is used to decide what to do, such as whether or how to change a clinical intervention or service or to spread a change. The findings of the evaluation need to be communicated and available to users at a time when they can use it to change what they would otherwise do.

Key points

- Organize the evaluation report around users' questions. To be understood and used, evaluation reports need to be presented with headings and in a style which addresses these questions.
- Describe the limitations, so as not to mislead users who may see findings as more certain than you know them to be.
- Give guidance about the conditions users would need to get similar results from implementing the intervention which was evaluated.
- There is a trend towards the evaluator's role extending beyond only sending a report, and moving into advising users about implementation.

For the evaluation to make a difference for patients or for costs, it needs to be available to decision makers in a way and at a time when it will influence their actions. Evaluators need to "market" their report by identifying the users, discovering where users are likely to look for such information, and finding ways to make summaries and the full report easily available. The internet and databases which can be searched are important "places" where the report needs to be noted and available. Some evaluations establish a web site for the evaluation which is easily discoverable in a search, allows downloading, and provides any supporting materials which can help evaluation users. Resources

describing effective dissemination approaches are given in the reference Research Utilization Support and Health (2009).

With regard to the content of reports and publications, guides have been written about which information needs to be reported for different types of evaluation and audiences (e.g., Boutron et al, 2008 (for RCTs), and Ogrinc et al, 2008 (for quality improvement). The "Reporting Guidelines 2013" list in the references to this document gives more reporting standards guides. Most such guides emphasize describing details of the intervention and its implementation so that others can reproduce it. The detailed descriptions can be in a report appendix, or, for publications, in a Web appendix, including assessments about what can and cannot be modified for the setting, patients or providers if similar results are to be expected.

Some details of the setting for a CSI evaluated by a RCT are recommended to be reported in many reporting guides (Boutron et al, 2008). But for observational or action evaluations, such details of context need to be far greater. The need to report such descriptions shows the importance of pre-study planning and theory to decide which data to collect and how to collect it about the intervention and context during the evaluation. The later section of this guide describes frameworks for deciding which data to collect about context, which can also be used to decide how to present the information in the report.

All of the above brings this guidance back to the first step of the evaluation. It shows the importance of identifying, right at the start, who the evaluation is for and with which of that entity's decisions the evaluation aims to help.

3. Evaluation designs

Volume 1 described three main categories of designs for evaluating CSIs: experimental, observational and action evaluation, and the questions and interventions for which they are most suited.

The section below describes reference papers about each design and other tools which can help to decide whether to use the design, and how to plan the details of applying the design.

For all the designs, useful guidance on how to plan and report the evaluation can be gained by reading studies which have used such designs, especially to evaluate the type of CSI in which you are interested. The appendix to Volume 1 gives eight such studies, and other examples which give good guidance for research are noted below.

The most useful general overviews of relevant designs for CSIs are given by Mercer et al. (2007) and Craig (2007) (for experimental designs), for observational case studies by Yin (1989), and for action evaluations by Øvretveit (2002) or formative evaluations by Stetler (2006).

1) Experimental and quasi-experimental

Comparative experimental (CE): CE designs plan and implement a defined intervention-change to "intervention units" which could be to patients, or to providers, or to service units (e.g., financial incentives and education to reduce hospital acquired infections). The defining feature is that some units receive the intervention-change, and some do not (the "comparison" or "control" group). Outcomes data are from measures selected to show any effects of the intervention which are of interest. These data are collected before exposure ("baseline"), and then at one or more times after exposure.

Different designs in this category use different strategies to design-out other possible influences, apart from the CSI, on the data collected to assess outcomes:

Randomized controlled trial (RCT): Guidance for applying RCT designs to CSIs and for reporting the findings from RCTs is provided by Stephenson (1998), Campbell (2000), Tunis et al (2009) (pragmatic trials) and EPOC (2013). Perhaps the most trenchant and incisive criticism of the RCT design for evaluating many types of CSIs has been provided by Kessler and Glasgow (2011). In contrast, good arguments supporting the use of RCTs for evaluating behavioral health interventions are given by Stephenson and Imrie (1998) and for information technology interventions, by Liu and Wyatt (2011).

Glasziou et al (2007) propose that RCTs are not necessary when the effect size is large, which, unfortunately, is rare for many CSIs. Rothwell (2005) gives guidance as to how to increase the external validity of a RCT, which is one of the limitations often cited. Hawe et al (2004) suggest that RCTs can be appropriate if a different approach to standardization is used:

> *"The issue is to allow the form to be adapted while standardizing the process and function. …For example, "workshops for general practitioners" are better regarded as mechanisms to engage general practitioners in organisational change or train them in a particular skill. These mechanisms could then take on different forms according to local context, while achieving the same objective"*

This idea is similar to the idea underlying the thinking about trigger mechanisms in realist evaluations. For Hawe et al (2004), it leads to proposing more relevant RCT designs, rather than to a realist evaluation.

Treweek and Zwarenstein (2009) describe differences between the design of most randomized trials (which have "an explanatory attitude") and the design of trials more able to inform decision making (which have "a pragmatic attitude") and discuss approaches used to show the applicability of trial results. The example studies 1 and 2 in the appendix to Volume 1 illustrate how RCT design was used to evaluate two different CSIs.

With regard to cluster RCTs, Eldridge et al (2004) summarize lessons for future evaluations of CSI from their review, and Campbell et al (2007) describe recent developments in statistics which are relevant to evaluating CSIs. On the question of a theory-based intervention in an RCT, Hardeman et al (2005) describe a causal modeling approach, and Eldridge et al (2005) note that such theory was important in their trial to reduce falls in older people.

Non-randomized comparative trial: Craig et al (2008) describe how to choose between randomized and non-randomized designs: by considering the size and timing of effects, the likelihood of selection bias, feasibility and the acceptability of experimentation, as well as cost. References which compare this design to the randomized variant include: Glasziou et al (2007), Concato el at (2010) and Concato and Horwitz (2004).

Cross-over comparative trial: In this variation, one group gets the intervention, while the other group gets none or another intervention. Then the intervention is stopped for the intervention group and started for the other group (an example is Devon et al. (2005)). This may be suitable for CSIs where it is thought the effects of the CSI on units decays rapidly.

Stepped wedge trial: In this design the units are often in the same organization (e.g., hospital units or wards). First, one unit is exposed to the intervention; then, after a period, an additional unit is exposed so that the two units are now receiving the intervention. Then another unit is added, and so on, until all are exposed. Quite sophisticated statistical analysis is needed (Brown and Lilford (2006)). Useful guidance for this design is given by Brown and Lilford (2006), Hussey and Hughes (2007) and Mdege et al (2011), and to look at how specific studies applied the design.

Non-comparative quasi-experimental

These designs do not use a comparison group, which makes it difficult to rule out some explanations for any observed before-later differences in outcomes.

Some guidance for planning an evaluation using this sub-category of designs is given in summary overviews by Fan et al. (2010) and Eccles et al. (2003), as well as by Craig et al. (2008). Black (1996) presents the arguments for using these designs to evaluate CSIs in healthcare. When planning the design for such studies and deciding which data to collect, it is useful to look ahead to ensure that data are collected which are recommended by reporting standards for these designs, such as SQUIRE (Ogrinc et al (2008).

Much of the most recent and useful guidance for these designs as applied to CSIs is given in the quality improvement research literature. The most relevant and useful discussions are by Grol et al. (2003), Fan et al. (2010) and Speroff and O'Connor (2004).

Before-After design: The design involves listing other possible explanations for the before/after (or later) data differences and assessing their likely influence, but without comparison groups these explanations cannot be excluded.

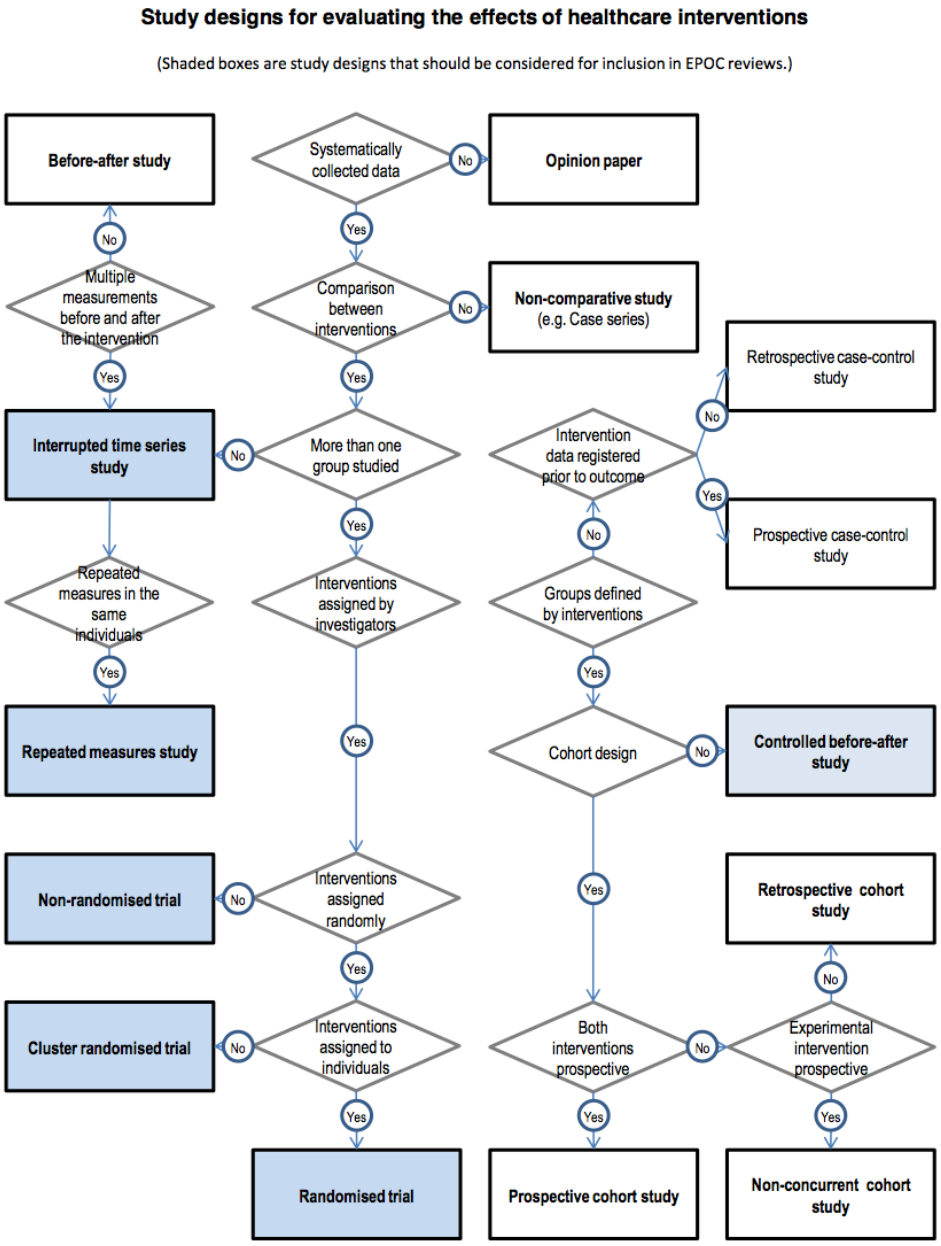
Simple time series, or interrupted time series design: The simple time-series design is like a before-after design but there are many data-collection time points for the outcome of interest (at least 3 before and 3 after are needed).

If enough time data-points are collected in the right way, statistical process control (SPC) methods can be used to define upper and lower control limits and to identify any special causes (such as the intervention) which significantly change the process outcome. Thor et al. (2007) show how this design can be applied and Carey and Lloyd (2001) and Wheeler (1993) give details to guide the statistical calculations which need to be used in different situations. Interrupted time series designs are described by Ramsay et al. (2003) and in the Cochrane centre guidance (EPOC 2013). The study example 3 in the Volume 1 appendix uses this design.

Quality improvement testing (PDSA): The plan – do – study – act cycle (PDSA) is an evaluation technique used in quality improvement to study whether a planned change, when implemented, has an effect on the outcomes of interest (Langley et al. (1996)). If this testing cycle is applied with a certain rigor, and possibly also using a time series design, then some researchers view it as allowing a reasonable degree of certainty about whether possible outcomes can be attributed to the intervention in real-world settings, even without a control group.

Certainty of attribution for evaluation can be enhanced by using the approach described by Speroff and O'Connor (2004): formation of a hypothesis for improvement (Plan), a study protocol with collection of data (Do), analysis and interpretation of the results (Study), and the iteration for what to do next (Act). Needham et al. (2009) give guidance for improving data collection in such designs to increase validity and reliability.

The following figure is useful for showing some of the differences between experimental designs (from EPOC 2013):



2) *Observational*

In this design, the intervention-change is not planned and introduced as an experiment with careful controls, but it and the outcomes are observed. This is often done retrospectively or concurrently when there is little time to plan, but sometimes the evaluation is planned and prospective before the intervention-change starts. These designs are used when controlled implementation is difficult or unethical, when little time and resources are available, or for other practical reasons. Sometimes these designs are called “naturalistic evaluations” where the intervention or service to be evaluated is studied “in the wild,” often as it “evolves” in its “environment”.

The first sub-category are observational evaluations which collect quantitative data - the terms "cohort", "case-control" and "cross-sectional" are usually applied to such evaluations, but these terms can also be applied to describe evaluations using qualitative data, as the terms describe the design, not the data collection method.

Quantitative: Cohort-, case control- and cross sectional- evaluation designs: Guidance and details are provided by Mann (2003) and Dreyer et al. (2010) (on observational studies for comparative effectiveness research).

Qualitative or mixed methods observational evaluation designs: The second sub-category of observational designs are those which collect qualitative data or use mixed data collection methods, sometimes called "naturalistic approaches" to evaluation. The designs use specific techniques to maximize internal and external validity. This group of designs and these techniques are perhaps less familiar to medical- and health service researchers, but have a long history with social scientists, international health researchers and in health promotion/education and public health research, as well as for educational, social work, mental health and welfare program evaluators (WHO (1981), Shadish, et al. (1991), Greene (1993), JCSEE (1994), Owen and Rodgers (1999)).

The main sub-categories are: single case evaluation, case-comparison evaluation, and the more recent realist evaluation, each with different designs within these sub-categories.

Case comparison evaluation: This is like the above-noted "cohort" design, but uses the validity-enhancing strategies for qualitative data and mixed methods of the single case evaluation mentioned above, such as triangulation and program theory. Example 5 in Volume 1 shows the methods in an evaluation of a large-scale safety program, which compared each of four hospitals receiving the program with other hospitals not receiving the program (Benning et al. (2011)). Another example is a comparison of two case hospitals, each of which received a CSI to implement an electronic medical record (Øvretveit et al (2007)).

Realist evaluation: These designs identify context-mechanism-outcome (CMO) configurations in complex interventions in different settings, and aim to establish "what works for whom in which settings." The designs and methods are less standardized than other designs and depend more on the skills and knowledge of the evaluators, using iteration and a number of stages. Guidance and ideas for applying this design can be found by Redfern et al. (2002); Blaise and Kegels (2004); Byng et al. (2005 and 2008), Greenhalgh (2008) and Pawson and Tilley (1997). Some limitations are described by Davis (2005).

3) Action evaluation

Action evaluations aim to provide early feedback from the evaluation to enable CSI implementers to improve the CSI and its implementation. Many "formative evaluations" would be classified as "action evaluations." One assumption is that, if the evaluation is useful to

different parties during, rather than after, the evaluation, evaluators can gain information and insights which they may not otherwise gain. By collaborating and participating in the shaping of the CSI, they may be better able to document how it is changed and why, and may be better able to explain later findings.

One example which gives some guidance is an action evaluation of a continuous quality improvement CSI in a hospital by Potter et al. (1994).

One variant of an action evaluation is being used in the VA to develop and evaluate the VA version of the patient centred medical home. Related to an earlier approach, termed "evidence based quality improvement" (Rubenstein et al. (2006, 2010)), this involves the researchers assisting primary health care personnel in a number of ways to design and implement changes (providing evidence of effective practices, training in quality improvement methods), and reporting findings from the evaluation to the implementers. The best simple overview of action evaluation is by Robson (1993). More comprehensive overviews are provided by Waterman et al. (2001), Hart and Bond (1996), Morton-Cooper (2000) and Øvretveit (2002).

4. Frameworks, Models and Theory in Evaluating CSIs

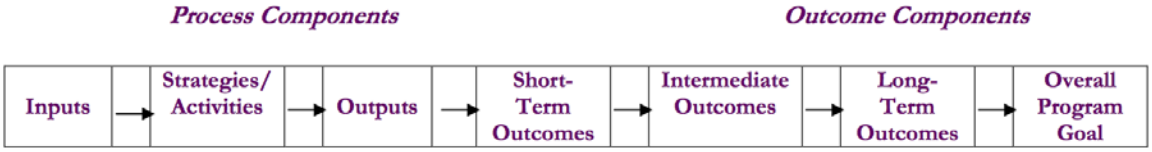
It is becoming good practice for most types of CSI evaluation to use a model, theory, or framework to guide the study. These help to define which data to collect and can be used to develop explanations for how or why the intervention has the effects discovered in different settings. Such explanations can help users to apply or adapt the intervention to their setting or to judge if they are likely to be able to implement it at all.

These models are theories or assumptions about the actions and conditions needed to produce certain outcomes. They are diagrams of either the intervention or of the main parts of the evaluation. The theory may be those of the evaluators, as they try to conceptualize or map the intervention, its context, and their expectations of outcomes. Or it may be the assumptions of the implementers or designers, which the evaluators discover by interviewing them or by studying plans. Or it may be a combination of the evaluator's and implementer's ideas.

Different terms are used, but this guidance refers to the two types of models in this way:

1) Program theory or logic model: A diagram showing the most important ingredients which are thought to be needed to carry out the intervention and the intermediate and later outcomes expected. (Sometimes this is referred to as the "model" or "framework" for the intervention).

Example of a Program Theory or Logic Model

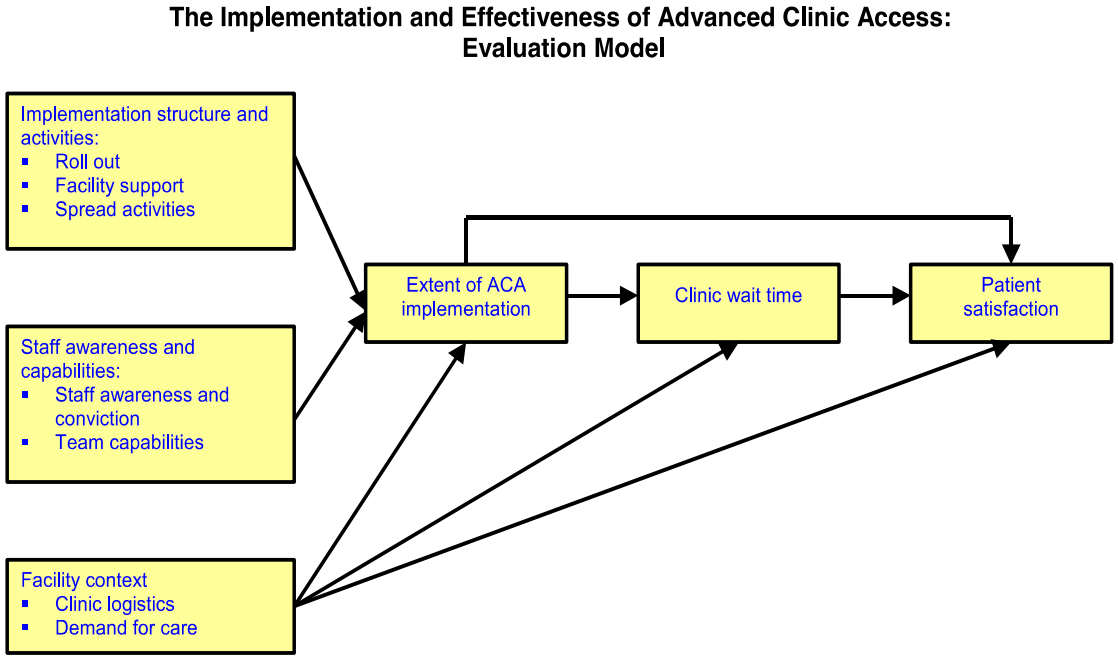


(CDC 2008)

This can serve as a framework that maps which types of data to collect, so as to assess whether the items that are assumed to be needed for the outcomes were implemented. The framework can then be used to decide which outcomes to study.

2) Evaluation model or framework: This is a diagram of the evaluation which shows the items about which data will be collected, which are needed to describe and assess the outcomes of the intervention. (The design diagram sometimes serves as an evaluation model or framework.)

Example of an Evaluation Model or Framework



There are three key points:

- Implementation or evaluation models show: the intervention (its component parts, or which activities are undertaken), the context of the intervention, and the expected outcomes.
- The boxes summarize ideas about the activities and ingredients which are expected to produce different outcomes. This then lays the basis for the evaluator to specify further which data or

measures need to be collected to assess if these were implemented or present, and the methods and times for data collection.

- One way in which models differ is in the extent to which they list "context factors" which may be needed for the intervention to be implemented.

What about context?

The first example with the logic model above does not have a box describing the "context" of the intervention. This may be because the proposers of the model do not think context will have any influence over whether the intervention can be implemented. It may be because the users are only interested in effectiveness in one setting.

However, research shows that many CSIs have different outcomes in different settings, primarily because the setting affects how much of the intervention is implemented. Understanding which context factors helped and hindered implementation allows the evaluator to give some guidance about the settings in which the intervention can be implemented, and about where and when similar results could be expected – it enables an assessment of generalizability.

The second model - the evaluation model - has a box describing "facility context," which includes "clinic logistics" and "demand for care." These are two things the evaluators want to examine to find out how much these affect implementation, as well as clinic wait times and patient satisfaction. These context factors are not the intervention (which is "advanced clinic access"), but they are thought to affect its implementation and the outcomes.

The model also has two other "context" boxes for "staff awareness and capabilities" and "implementation structure and activities." The evaluators are theorizing that these affect the extent to which the advance clinic access intervention will be implemented.

Further details of the evaluation show which data was collected to document and quantify these context factors, as well as to document the intervention implemented and outcomes.

More detailed context frameworks

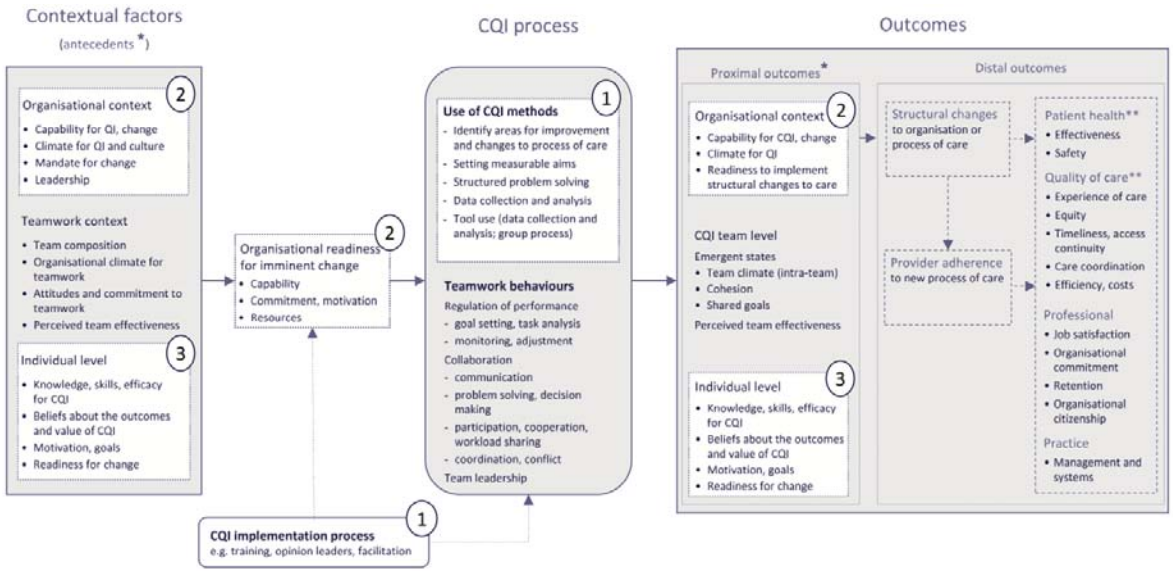
There are more detailed and sophisticated models and frameworks about the aspects of context which are thought to influence implementation of different classes of complex social intervention. These can give a starting point for evaluators to decide which aspects of context to study and which data and measures to use to collect data about these aspects.

For interventions of evidence based clinical changes, a well-developed framework of different elements of context is given in the PARiHS model (Stetler et al. (2011)). The following summarizes these elements:

Context elements in the PARiHS model (Stetler et al 2011).

"Readiness for change" is one aspect of context which can be measured and which can help evaluations of CSIs in natural settings to explain findings and to give guidance to users. Again, which factors to examine depends on the type of CSI being evaluated. Different factors may be important for implementing a computer decision support system to those for implementing a resident fall prevention program in a nursing home. A general framework and measure for collecting data about readiness for change is the ORCA instrument (Helfrich et al. (2009)).

As regards context for quality improvement CSIs in primary care (which includes a wide range of interventions), Brennan et al.'s (2012) review provides one model and a more detailed list of context factors (51 measurement instruments):



Another framework which can serve as a starting point for the evaluation of some quality improvement or similar CSIs is the Model for Understanding Success in Quality (MUSIQ) (Kaplan et al (2012)).

5. Data collection and analysis

Most researchers are aware of different methods needed to collect data about the intervention, its context and outcomes, and to analyze these data. But an evaluation of a CSI often involves some methods which the evaluator has not used. The guidance here will not duplicate excellent guidance about different data-collection and analysis methods, but it gives a selection of the guides which are useful to CSI evaluators for more information about each method.

Guidance for the data needed to describe the intervention and to report details including context are given in different reporting guidelines such as those provided by SQUIRE (Ogrinc et al. (2008)) and Michie et al. (2009).

5.1. Data collection

"Data-gathering" describes the stage of the evaluation where the evaluator identifies sources of data, gets access to these sources, and collects those data which are needed. It describes a range of collection methods within the following five categories:

- **Already-collected data:** data collected for other purposes, by a service, government departments, other researchers, opinion polls, and other people (e.g., journalists) ("secondary data"), as well as diary records, minutes of meetings, patient case records, legal documents, etc. (sometimes called "primary sources")
- **Observation**: unobtrusive, participant, or self-observation
- **Interviews**: structured (e.g., questions), semi-structured, open, guided by a critical incident or vignette stimulus, or focus group interviews

- **Questionnaire or survey**: small- or large-scale survey, with or without rating scales
- **Measurement methods**: biophysical, subjective response, or a pre-formulated measurement instrument such as disease-specific or quality of life composite measures

The most useful general guides about data collection relevant to CSI evaluations are by OBSSR (2013) and Robson (1993), for quantitative data collection and analysis by St Leger et al. (1992) (Chapter 11), and Edwards and Talbot (1994) (Chapter 6).

Already-collected data

The most important general rule which applies especially to evaluations of CSIs is, "never collect data until you are sure no one else has data which you can use." "You can use" means, "that you can easily access," and it means "valid and reliable" for the purposes of the evaluation – the latter means checking how valid and reliable the data are, either by asking those who collected and analyzed the data, or by looking for research or publications which have made these assessments. OBSSR (2013) gives useful guidance about data from administrative data systems. For some VA databases, specialist guides are available via HSR&D cyber seminars and from other VA sources – the people running the database are the best guides.

Observation

A simple but useful guide is provided by Western Michigan University (TEC (2013)). Problems and details of observational methods are described by Pope and Mays (1995a, b) and in more detail by Sapsford and Abbott (1992). Practical general accounts are given in Edwards and Talbot (1994) and description of how to use pre-coded observation is given by Breakwell and Millward (1995). More detailed discussions can be found in general texts on social science methods such as those by Adams and Shvaneveldt (1991).

Interviews

Both observation and interviewing can be used to collect data in a quantitative or qualitative form. Interviewing gives the evaluator access to people's views, their recollected experiences, feelings and their theories about causation. This method can be used to collect data in a quantitative form when the interviewer uses pre-structured categories and questions (e.g., a pre-coded questionnaire administered in an interview). Interviews more often are used to collect qualitative data by using open-ended questions or a set of topics for open exploration and probing by the interviewer.

When choosing a data-gathering method, the researcher needs to consider how she will analyze the data and present it to users. One of the greatest weaknesses of qualitative observation and interviewing is the difficulty in analyzing and presenting the data, especially to users who are unfamiliar with or skeptical of these methods.

Summaries of focus group technique in health services are provided by Fitzpatrick and Boulton (1994), and Kitzinger (1995). More details are given in books on the subject by Morgan (1993) (e.g., when to use focus groups and why) and by Kreuger (1988).

For more details of qualitative open interviewing in health settings, CSI evaluators would be helped by a summaries by Britten (1995), Sapsford and Abbott (1992), and Fitzpatrick and Boulton (1994). Practical summaries are given by Edwards and Talbot (1994) and Breakwell and Millward (1995). One interesting example of the use of semi-structured interviewing is given in a study which sought older people's perceptions of care and problems (Powell et al (1994)). The in-depth interview method in organizational studies is described by Ghauir et al. (1995). General social science methodology texts give extensive practical and theoretical discussion of the method. Kvale (1994) gives a very readable and concise discussion of "ten standard objections to qualitative research interviews."

Questionnaire or survey

Questionnaires are used in CSI evaluations to collect data about specific topics and where the topics have the same meaning and are well understood by people in different settings or social groups. Questionnaires are less expensive than interviews – the latter are unnecessary where simple factual data are required, or where people can easily and authentically express their ideas in terms of the categories used by the researcher in the questionnaire. Questionnaires can gather qualitative data by asking people to write descriptive accounts. More often, questionnaires use one or more of different measurement scales which require subjects to express their views in the terms of a scale and thus provide quantitative data.

The most well-known scale is the Likert 5-item scale or a Semantic Differential scale (pairs of opposites, e.g., painful/not painful, usually with a 7-point scale (see Breakwell and Millward (1995) for a simple summary)). Again, this can be a source of biased data, for example where people from different cultures used the extremes of rating scales in a different way (van de Vijver and Leung (1997)). The most well-known problem is that different responses are gained with different question-phrasing. For example, 44% would allow a terminally ill person to choose a "lethal injection," but 50% would approve a "medical procedure." This rate increased to 65% when the question was phrased, "Would you support the right of the terminally ill to choose 'death with dignity' over prolonging life?"

These and other issues are discussed in detail in general texts by Frankfort-Nachmias and Nachmias (1992) and for research by Breakwell and Millward (1995). Questionnaire design is summarized in overviews of the method in health care by Sapsford and Abbott(1992) and Edwards and Talbot (1994). McKinlay (1992) gives an excellent discussion of methods used for surveying older people. Surveys and questionnaires for organizational research are discussed by Ghauri et al. (1995). Hawe et al (1990) describe surveying for evaluating health promotion programs. OBSSR (2013) gives simple guidance for sample surveys.

There is a fine dividing line between a questionnaire survey and standard measurement instruments such as the General Health Questionnaire (Bowling (1992)). The difference is that the latter are usually constructed on the basis of an explicit conceptual model and have been extensively tested and often validated, whereas questionnaires and surveys are usually developed for the specific purpose of the research and might have little pilot testing or no validation.

Measurement methods

> *"The design of controlled experimentation has been refined to a science that is within the grasp of any researcher who owns a table of random digits and recognises the difference between blind and sighted assessments.*
> *However, the measurement of outcome seems to have been abandoned at a primitive stage of development.....A superfluity of instruments exists, and too little is known about them to prefer one to another."*
>
> Smith et al (1980)

The above critical view of outcome measures is an extreme one and measures have advanced considerably since 1980. However, it is still true that some CSI evaluators do not choose the most appropriate measure for the purpose of the evaluation and resources available. Often, intermediate outcome measures are the only ones which can be linked by the design to the intervention with any degree of certainty. Later potential outcomes such as mortality are often not attributable to the CSI because of other confounders.

When used as a general term, "measurement" describes any method of data collection. However, questionnaires are sometimes described as measures. Here, the term is used in a specific sense to mean, "only methods for collecting data in a numerical or 'quantified' form." More precisely, measurement is assigning numerical values to objects, events or empirical facts according to specified rules. In this sense, we may measure a person's attitude by asking him to express his views in terms of a number on a rating scale (an ordinal scale), or measure his temperature using a thermometer (a ratio scale). We gather data not about the entity or the concept, but about the properties of a concept. This involves using indicators which are observable events that are inferred measures of concepts.

Different measures are well described in general research texts such as Bowling's texts on measuring disease (Bowling (1995)) and her review of quality of life measures (Bowling (1992)) as well as in research texts such as those by Fink (1993), Rossi and Freeman (1993), St Leger (1992), and Breakwell and Millward (1995).

Multi-level modeling is used when the observations that are being analyzed are correlated or clustered along spatial, non-spatial, and/or temporal dimensions, or where the causal processes are thought to operate simultaneously at more than one level. A guide to the techniques is given by OBSSR (2013).

Some common measurement terms

- Sample: A smaller number of a larger population.
- Prevalence: At a particular time, the number of existing cases identified or arising in a population.

- Incidence: Over a period of time, the number of new cases or events identified or arising in a population.
- Rate: A ratio of two measures, such as the proportion of a population with a particular problem or characteristic, often expressed by age or by sex (e.g., cases out of 100,000). Rates require data from interval or ratio scales.
- Prevalence rate: The proportion of cases in a population at a particular time (e.g., 26 in 100,000).
- Incidence rate: The proportion of new cases which arise over a period of time. Death- or mortality-rate is the proportion of a population who die - but who die during a defined time period.

5.2. Data analysis

Quantitative data analysis

In many CSI evaluations, there are two sets of numbers (e.g., a "before" and an "after" set, or outcomes from two services in different places). Statistical significance testing helps to show whether or not any differences between the two sets really represents true differences in the populations from which the samples were drawn. This is based on the idea that any difference between the two sets is caused by a real difference as well as by differences arising from random and systematic error introduced by the measurement method. This involves proposing a null-hypothesis - that there is no difference between the sets, and examining whether any difference shown is greater than that expected by chance. The significance level is the level of probability at which we decide to reject the null hypothesis.

Most CSI evaluators know what "statistically significant" means in evaluations – that the event did not occur by chance alone and that there is probably some external cause. It does not prove that the variables being investigated caused the difference. Black (1992) comments that, "It is up to the evaluator to prove that the variables under consideration are the actual cause and to eliminate the possibility of any other variable(s) contributing to the results found."

Phillips et al (1994) give a useful and simple summary of the main statistical methods for analysis by distinguishing different stages of analysis. The first is to describe and summarize the data by representing each numerical value in a pie chart or bar chart, by calculating the averages (the mean, median and mode), the range (the difference between the smallest and largest value in a data set), and the standard deviation (which is how much the data values deviate from the average). The second stage is to define the generalizability of the data by stating how much confidence we would have of finding the results from the sample in the general population. This is done by calculating the "confidence interval." A third "hypothesis testing" stage involves using data to confirm or reject a hypothesis. A type I error is to reject a null hypothesis when it is in fact true: the analysis calculates the probability of having a type 1 error (called the "significance level"). A fourth stage is to calculate the strength of the association between two variables using chi-squared tests, which calculate a correlation coefficient, or by carrying out a regression analysis.

There are a number of texts which give simple summaries with examples. Techniques for deciding significance levels and other details of measurement, sampling and statistical analysis in health research are described in summary by St Leger et al (1992), Edwards and Talbot (1994), and McConway (1994). A simple general practical overview of "describing and summarizing data" and of drawing inferences in evaluation is given by Breakwell and Millward (1995). Wilkin et al (1992) describe measurement of need and outcome, as does Bowling (1992, 1995). A more detailed and comprehensive text for clinicians is by Gardner and Altman (1989). The appendix to this document gives a summary of key points about confidence intervals and p-values.

Terms used in quantitative data analysis

- Internal validity: The validity of the conclusions in relation to the specific sample of the study. For example, in an evaluation experiment being able to show whether or not the intervention has an effect or the size of the effect.
- External validity: The ability of a study to show that the findings would also apply to similar populations, organizations or situations, for example, when an intervention is applied in another setting.
- Dependent variable: The outcome variable or end result of a treatment, service or policy which is the subject of the study (e.g., cancer mortality, patient satisfaction, resources consumed by a service), and which might be associated with or even caused by other (independent) variables. (The data analysis tests for associations between the dependent (outcome) variable and the independent variables. Establishing causation is more complex.)
- Independent variable(s): A variable whose possible effect on the dependent variable is examined. (Something which may cause the outcome and which is tested in the research. (Note: Many independent variables may be associated with a dependent variable, but only a few have a causal influence, and even fewer can be shown unambiguously to have a causal effect. A dependent variable cannot influence an independent variable: e.g., genetic make-up can predispose to cancer, but cancer, as far as we know, cannot affect genes).
- Mediating variable(s): Other variables which could affect the dependent variable or outcome, which the research tries to control for in design or in statistical analysis.
- Extraneous variable(s): Variables not considered in the theory or model used in the study.
- Confounding variable(s): Any variable which influences the dependent variable or outcome but was not considered or controlled for in the study. Alternative definition: "confounding arises when an observed association between two variables is due to the action of a third factor" (Crombie (1996)).

Analysing qualitative data

The greatest challenge to using qualitative data in evaluation is analyzing the data. The challenge does not stop there - there is another related problem, which is how to display qualitative data and to convince users and scientists that the conclusions are justified by the data. There are two issues: (1) How to use the techniques of analysis (which are generally agreed upon by qualitative researchers in

order to reach conclusions which other scientists using these methods would accept) and (2) How to present the conclusions and analysis to those who are not familiar with these techniques.

A common approach to qualitative data analysis is through the following steps:

1) Interview or observation
2) Text (a write-up of the interview or field notes or transcript of a tape)
3) Code or classify (according to "emergent" themes or patterns)
4) Further analysis (re-coding or hypothesis testing, often by returning to original text or other texts to compare views or settings for similarities and differences)
5) Conclusions/results (categories of experience or feelings of the subjects, meanings subjects give to events, explanatory models and concepts, or generalizations)

Qualitative analysis is inductive, building and testing concepts in interaction with the data or the subjects. It is also usually iterative: the analyst forms categories from the data and then returns to the data to test their generalizability.

These techniques of data analysis are complex and are not easy to describe in research reports for readers unfamiliar with the techniques, but this is also true for methods for analyzing quantitative data. However, examples from the original data give vivid illustrations, and also "ring true" with users.
A comprehensive and detailed account of qualitative data analysis is given by Miles and Huberman (1984), but more simple and shorter summaries are provided by Fitzpatrick and Boulton (1994), Edwards and Talbot (1994), and Sapsford and Abbot (1995). A discussion specifically for evaluation is given by Patton (1987). Other good general texts on qualitative data collection methods and philosophy include those by Denzin and Lincoln (1993), Glaser and Strauss (1968), Greene (1994), Lincoln and Guba (1985), Miles and Huberman (1994), and for reliability and validity tests, by Strauss and Corbin J (1990). Guidance for software for analyzing qualitative data is described by OBSSR (2013).

5.3. Mixed methods analysis

Evaluations of CSIs often need to collect data from many different sources and sometimes to use different data sets to explore one or more questions. The Benning et al. (2011) study in the Appendix to Volume 1 used mixed methods to collect data about the UK safer patients initiative (SPI) large scale program:

1) Semi-structured interviews to discover knowledge and enthusiasm for the initiative among 60 senior members of staff in the four SPI hospitals.
2) Before and after surveys of staff attitudes in the control and SPI hospitals
3) Qualitative studies - ethnographic observations on acute medical wards, interviews, and focus groups in SPI hospitals - staff behavior and views
4) Impact on processes of clinical care: error rates - case notes
5) Improving outcomes: case notes to identify adverse events and mortality and assessment for any improvement in patients' experiences (the NHS patient survey).

The example shows how the researchers combined data sources about and from different levels of the health system to reach their conclusions. The evaluators comment that:

> *"This type of evaluation is particularly suitable for service delivery/management interventions…that are not likely to yield the type of conclusive results characteristic of evaluations of treatments based on measurement of outcomes on patients…Mixed method evaluation draws on the idea of "triangulation," where confidence in the findings increases when observations of one type are corroborated by other types of evidence."*
>
> Benning et al (2011)

Mixed methods can increase objectivity and reduce bias by combining multiple sources of data. One example is cross-checking data when a head quality officer says 70% of projects exceeded targets - a cross check with reports from the 21 projects, and possibly with statistical or other data related to targets (e.g., reported adverse events) may find a different percentage.

Burke Johnson and Onwuegbuzie 2004 distinguish between:

- *Mixed-model designs*: data collected using different methods at the same time (e.g., mixing qualitative and quantitative approaches within or across the stages of the research process)
- *Mixed-method design*: One type of method used after another e.g., quantitative then qualitative data.

		Time Order Decision: Concurrent	Time Order Decision: Sequential
Paradigm Emphasis Decision	Equal Status	QUAL + QUAN	QUAL → QUAN QUAN → QUAL
	Dominant Status	QUAL + quan QUAN + qual	QUAL → quan qual → QUAN QUAN → qual quan → QUAL

Burke Johnson and Onwuegbuzie (2004)

Resources for integrating or combining data collected using different methods are described in the "User Friendly Handbook for Mixed Methods Evaluations" by NSF (2013), by Bridges (2013), and in a web tutorial at SRM (2013). The *Journal of Mixed Methods Research* gives examples of different techniques and issues in planning and using mixed methods in research generally (JMMR (2013)).

Useful guidance can also be found by Sale and Brazil (2004) (critical appraisal of mixed methods studies), Tashakkori and Teddlie (2010), Creswell et al (2004) (primary care), Creswell and Plano Clark (2011), Sandelowski (2000), and for publishing mixed methods studies by Creswell and Tashakkori (2008).

5.4. Summary

Because of the broad range of subjects and user questions they are faced with, evaluators of CSIs need to be aware of a wide range of data gathering methods. Users of evaluations also need to have some understanding of the methods used to gather data in order to judge the validity of the conclusions.

Data for an evaluation can be collected by methods within the five categories of already-collected data, observation, interviewing, questionnaires and surveys, and measurement methods. The choice of data-gathering method should follow from the evaluation design and questions to be answered, rather than design and the questions answered following from the data gathering method with which the evaluator is most familiar.

The ten golden rules for data collection in CSI evaluations

- Don't collect data unless you are sure no one else has,
- Don't invent a new measure when a proven one will do,
- When the person or documents you need to see are not there, don't be blind to what is there which could help – be opportunistic,
- Measure what's important, not what's easy to measure,
- Don't collect data where confounders make interpretation impossible,
- Spend twice as much time on planning and designing the evaluation than you spend on data collection,
- Always do a small pilot to test the method on a small sample,
- As you collect the data, save them to a database which is designed with thought to how to carry out the analysis
- Analyzing the data takes twice as long as collecting it, if you have not defined clearly which data you need and why,
- Data collection will take twice long as you expect

Øvretveit (2002)

6. Appendix: Guidance for data needed to describe intervention and context and for reporting

See also, for quality improvement interventions, SQUIRE guidance (Ogrinc et al 2008).

6.1. Journal "Implementation Science" guidance

From Michie at al (2009):

> "*Implementation Science* editorial policy on describing the content of complex interventions… Authors submitting to *Implementation Science* will be required to provide detailed descriptions of the interventions delivered in their studies. These are the WIDER Recommendations to Improve Reporting of the Content of Behaviour Change Interventions http://interventiondesign.co.uk/
>
> *1. Detailed description of interventions in published papers*
> Authors describing behaviour change intervention (BCI) evaluations should describe: 1) characteristics of those delivering the intervention, 2) characteristics of the recipients (and see Noguchi *et al*., 2007, for unusual but importantly informative detail on participants before and after attrition), 3) the setting (*e.g*., worksite, time, and place of intervention), 4) the mode of delivery (*e.g*., face-to-face), 5) the intensity (*e.g*., contact time), 6) the duration (*e.g*., number of sessions and their spacing over a given period), 7) adherence/fidelity to delivery protocols, and 8) a detailed description of the intervention content provided for each study group.
>
> *2. Clarification of assumed change process and design principles*
> Authors describing BCI evaluations should describe: 1) the intervention development, 2) the change techniques used in the intervention, and 3) the causal processes targeted by these change techniques; all in as much detail as is possible, unless these details are already readily available (*e.g*., in a prior publication).
>
> *3. Access to intervention manuals/protocols*
> At the time of publishing a BCI evaluation report, editors will ask authors to submit protocols or manuals describing BCI evaluations or, alternatively, specify where manuals can be easily and reliably accessed by readers. Such supplementary materials can be made accessible online.
>
> *4. Detailed description of active control conditions*
> Authors describing BCI evaluations should describe the content of active control groups in as much detail as is possible (*e.g*., the techniques used) in a similar manner to the description of the content of the intervention itself."
>
> Michie at al (2009)

7. Appendix: What are confidence intervals and p-values?

- A confidence interval calculated for a measure of treatment effect shows the range within which the true treatment effect is likely to lie (subject to a number of assumptions).
- A p-value is calculated to assess whether trial results are likely to have occurred simply through chance (assuming that there is no real difference between new treatment and old, and assuming, of course, that the study was well conducted).
- Confidence intervals are preferable to p-values, as they tell us the range of possible effect sizes compatible with the data.
- P-values simply provide a cut-off beyond which we assert that the findings are 'statistically significant' (by convention, this is $p<0.05$).
- A confidence interval that embraces the value of no difference between treatments indicates that the treatment under investigation is not significantly different from the control.
- Confidence intervals aid interpretation of clinical trial data by putting upper and lower bounds on the likely size of any true effect.
- Bias must be assessed before confidence intervals can be interpreted. Even very large samples and very narrow confidence intervals can mislead if they come from biased studies.
- Non-significance does not mean 'no effect'. Small studies will often report non-significance even when there are important, real effects which a large study would have detected.
- Statistical significance does not necessarily mean that the effect is real: by chance alone about one in 20 significant findings will be spurious.
- Statistically significant does not necessarily mean clinically important. It is the size of the effect that determines the importance, not the presence of statistical significance.

Davies, H Crombie, I 2009 "What is…? series": www.whatisseries.co.uk

8. Appendix: Useful web sites for CSI Evaluation resources

- VA HSR&D Cyberseminars, Veterans Health Administration Health Service Research and Development: http://www.hsrd.research.va.gov/Cyberseminars
- Using Evaluation to Improve Our Work: A Resource Guide (Mittman, B and Salem-Schatz, S). Veterans Health Administration QUERI program: http://www.queri.research.va.gov/ciprs/projects/ResourceGuideV1-1.cfm
- OBSSR 2013 "e-source book on methods in behavioral and social science research", US Office of Behavioral and Social Science Research, NIH: http://www.esourceresearch.org/tabid/380/default.aspx

Mixed Methods Research web resources

- JMMR 2013 Journal of Mixed Methods Research http://intl-mmr.sagepub.com
- NSF 2013 User Friendly Handbook for Mixed Methods Evaluations http://www.nsf.gov/pubs/1997/nsf97153/start.htm
- Bridges 2013 - Mixed Methods Network for Behavioral, Social, and Health Sciences http://www.fiu.edu/~bridges/
- SRM 2013 Research Design and Mixed-Method Approach: A Hands-on Experience (tutorial website) http://www.socialresearchmethods.net/tutorial/Sydenstricker/ bolsa.html

9. References

- Adams, G and Shvaneveldt, J (1991) Understanding Research Methods, Longman, New York.
- Apodaca, A 2013 Defining Your Research Questions and Hypotheses, http://science.dodlive.mil/2010/10/04/defining-the-beginning-importance-of-research-questions-hypotheses/ accessed 21 January 2013
- Bailie RS, Si D, Robinson GW, Togni SJ, D'Abbs PH. 2004 A multifaceted health-service intervention in remote Aboriginal communities: 3-year follow-up of the impact on diabetes care. Med J Aust. 2004;181 (4):195-200.
- Benning A, Ghaleb M, Suokas A, Dixon-Woods M, Dawson J, Barber N, et al. Large scale organisational intervention to improve patient safety in four UK hospitals: mixed method evaluation. BMJ 2011;doi:10.1136/bmj.d195.
- Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;312:1215-8.
- Black, N (1992) "Research, audit and education", British Medical Journal, (304); pp 698-700.
- Bowen, S (undated) A Guide to Evaluation in Health Research, Canadian Institutes of Health Research and Univeristy of Alberta, Canada.
- Bowling, A (1992) Measuring Health: A review of quality of life measures, Open University Press, Milton Keynes.
- Bowling, A (1995) Measuring Disease: A reveiw of disease-specific quality of life measurement scales, Open University Press, Milton Keynes.
- Breakwell, G and Millward, L (1995) Basic Evaluation Methods, British Psychological Society Books, Leicester.
- Brennan S Bosch M Buchan H Green S 2012 Measuring organizational and individual factors thought to influence the success of quality improvement in primary care: a systematic review of instruments, Implementation Science 2012, 7:121 doi:10.1186/1748-5908-7-121
- Bridges 2013 Bridges - Mixed Methods Network for Behavioral, Social, and Health Sciences http://www.fiu.edu/~bridges/
- Britten, N (1995)"Qualitative interviews in medical research", British Medical Journal, 311, pp 251-253.
- Brown C, T Hofer, A Johal, R Thomson, J Nicholl, B D Franklin, and R J Lilford 2008a An epistemology of patient safety research: a framework for study design and interpretation. Part 1. Conceptualising and developing interventions Qual Saf Health Care 2008;17 158-162
- Brown C, T Hofer, A Johal, R Thomson, J Nicholl, B D Franklin, and R J Lilford 2008b

An epistemology of patient safety research: a framework for study design and interpretation. Part 2. Study design C Qual Saf Health Care 2008;17 163-169

- Brown C, T Hofer, A Johal, R Thomson, J Nicholl, B D Franklin, and R J Lilford 2008c An epistemology of patient safety research: a framework for study design and interpretation. Part 3. End points and measurement Qual Saf Health Care 2008;17 170-177
- Brown C, T Hofer, A Johal, R Thomson, J Nicholl, B D Franklin, and R J Lilford 2008d An epistemology of patient safety research: a framework for study design and interpretation. Part 4. One size does not fit all Qual Saf Health Care 2008;17 178-181
- Brown, C. A. and R. J. Lilford (2006) The stepped wedge trial design: a systematic review BMC Med Res Methodol 6.
- Burke Johnson R Onwuegbuzie A 2004 Mixed Methods Research: A Research Paradigm Whose Time Has Come EDUCATIONAL RESEARCHER Educational Researcher, Vol. 33, No. 7, pp. 14–26 DOI: 10.3102/0013189X033007014
- Campbell M, Donner A, Klar N. 2007 Developments in cluster randomised trials and statistics in medicine. Stat Med 2007;26:2-19.
- Campbell NC, Murray E, Darbyshire J, Emery J, Farmer A, Griffiths F, et al. Designing and evaluating complex interventions to improve health care. BMJ 2007;334:455-9.
- Campbell, M., R. Fitzpatrick, et al. (2000). "Framework for design and evaluation of
- CDC 1999 Framework for Program Evaluation in Public Health, http://www.cdc.gov/mmwr/preview/mmwrhtml/rr4811a1.htm, Centers for Disease Control and Prevention
- CDC 2008 Logic Model Basics, Evaluation Briefs, No. 2 December 2008ETA, CDC washington: available from http://www.cdc.gov/healthyyouth/evaluation/index.htm. Accessed 24th Jan 2013
- CIPRS 2013 Using Evaluation to Improve Our Work: A Resource Guide (Mittman, B and Salem-Schatz, S). Veterans Health Administration QUERI program) http://www.queri.research.va.gov/ciprs/projects/ResourceGuideV1-1.cfm accessed 27th January 2013
- complex interventions to improve health." BMJ 321(7262): 694-6. ID Number
- Concato J, Horwitz RI. Beyond randomised versus observational studies. Lancet. 2004;363(9422):1660-1661.
- Concato J, Lawler EV, Lew RA, Gaziano JM, Aslan M, Huang GD. Observational methods in comparative effectiveness research. Am J Med. 2010;123 (12)(suppl 1):e16-e23.
- Craig, P Macintyre, S Michie, S Nazareth, I Petticrew, M 2006 Developing and evaluating complex interventions: new guidance, MRC London From www.mrc.ac.uk/complexinterventionsguidance Accessed 24jan2013
- Creswell JW, Plano-Clark V. Designing and Conducting Mixed Methods Research. Thousand Oaks, CA: Sage Publications, 2007.
- Creswell, J 2009 Research Design: Qualitative, Quantitative, and Mixed Methods Approaches 3rd ed Sage, Los Angeles
- Denzin, N and Lincoln, Y (eds) (1993) Handbook of Qualitative Research, Sage, London.
- Devon E. Hinton D Chhean D Pich V et al 2005 A randomized controlled trial of cognitive-behavior therapy for Cambodian refugees with treatment-resistant PTSD and panic attacks: A cross-over design Journal of Traumatic Stress Volume 18, Issue 6, pages 617–629, December 2005.
- Devon E. Hinton D Chhean D Pich V et al 2005 A randomized controlled trial of cognitive-behavior therapy for Cambodian

refugees with treatment-resistant PTSD and panic attacks: A cross-over design Journal of Traumatic Stress Volume 18, Issue 6, pages 617–629, December 2005.
- DVA 2011 Department of Veterans Affairs. VHA Handbook 1058.05: VHA Operations that May Constitute Research [Internet]. Veterans Health Administration; 2011. http://www1.va.gov/vhapublications/ViewPublication.asp?pub_ID=2456 accessed 21jan2013
- Eccles M, Grimshaw J, Campbell M, Ramsay C. Research designs for studies evaluating the effectiveness of change and improvement strategies. Qual Saf Health Care. 2003;12(1):47-52.
- Edwards, A and Talbot, R (1994), The Hard-Pressed Researcher, Longman, London.
- Eldridge S, Ashby D, Feder G, Rudnicka AR, Ukoumunne OC. 2004 Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. Clin Trials 2004;1:80-90.
- Eldridge S, Spencer A, Cryer C, Pearsons S, Underwood M, Feder G. 2005 Why modelling a complex intervention is an important precursor to trial design: lessons from studying an intervention to reduce falls-related injuries in elderly people. J Health Services Res Policy 2005;10:133-42.
- EPOC 2013 Methods web site (Cochrane Effective Practice and Organisation of Care Group) http://epoc.cochrane.org/epoc-methods
- Fan E Laupacis A Pronovost P et al. 2010 How to Use an Article About Quality Improvement JAMA. 2010;304(20):2279-2287.
- Fink, A (1993) Evaluation Fundamentals, Sage, London.
- Fitzpatrick, R and Boulton, M (1994) "Qualitative methods for assessing health care," Quality in Health Care, 3, 107-113.
- Frankfort-Nachmias, C and Nachmias, D (1992) Research Methods in Social Sceinces, Edward Arnold, London (4th edn).
- Ganann, R Ciliska, D Thomas, H 2010 Expediting systematic reviews: methods and implications of rapid reviews Implementation Science 2010, 5:56 http://www.implementationscience.com/content/5/1/56 Accessed 21jan 2013
- Gardner, M and Altman, D (1989) Statistics with Confidence, British Medical Journal, London.
- Ghauri, P Grønhaug, K and Kristianslund, I (1995) Research Methods in Business Studies, Prentice Hall, London.
- Glaser, B. G. and Strauss, A. L. (1968), The Discovery of Grounded Theory: Strategies for Qualitative Research, Weidenfeld and Nicolson, London.
- Greene, J (1994) "Qualitative program evaluation", in Denzin, N and Lincoln, Y (eds) (1994) opp cit, pp 530-544.
- Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ 2005. doi:10.1136/bmj.38636.593461.68.
- Greenhalgh T, Robert G, Bate P, Kyriakidou O, Macfarlane F, Peacock R. A systematic review of the literature on diffusion, dissemination and sustainability of innovations in health service delivery and organisation. London: NHSSDO Programme; 2004. www.sdo.lshtm.ac.uk.
- Grol R, Baker R, Moss F, editors. Quality improvement research: understanding the science of change in health care. London: BMJ Books, 2003,
- Grol, R. P., Bosch, M. C., Hulscher, M. E., Eccles, M. P., and Wensing, M. (2007). Planning and studying improvement in patient care: the use of theoretical perspectives. *Milbank Q,* 85(1), 93-138.
- Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an

emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924-6.
- Hardeman W, Sutton S, Griffin S, Johnston M, White A, Wareham NJ, et al. A causal modelling approach to the development of theory-based behaviour change programmes for trial evaluation. Health Educ Res 2005;20:676-87.
- Hawe, P Degeling, D and Hull, J (1990) Evaluating Health Promotion, Maclennan and Petty, London.
- Helfrich C Li Y Sharp N Sales, A 2009 Organizational readiness to change assessment (ORCA): Development of an instrument based on the Promoting Action on Research in Health Services (PARIHS) framework Implementation Science 2009, 4:38 doi:10.1186/1748-5908-4-38:
- Holden, D.J., Zimmerman, M. 2009. A practical guide to program evaluation planning. Sage Publications, Thousand Oaks.
- Hussey M Hughes J 2007 Design and analysis of steppedwedge cluster randomized trials Contemporary Clinical Trials Volume 28, Issue 2, February 2007, Pages 182–191
- JMMR 2013 Journal of Mixed Methods Research http://intl-mmr.sagepub.com
- Kaplan HC, Provost LP, Froehle CM, et al. The Model for Understanding Success in Quality (MUSIQ): building a theory of context in healthcare quality improvement. BMJ Qual Saf. 2012 Jan;21(1):13-20.
- Kellogg 1998 Evaluation Handbook, Battle Creek, MI available at www.wkkf.org. accessed 27 Jan 2013
- Kessler, R Glasgow, R 2011 A Proposal to Speed Translation of Healthcare Research Into Practice Dramatic Change Is Needed Am J Prev Med 2011;40(6):637– 644.
- Kitzinger, J (1995) "Introducing focus groups", British Medical Journal, 311, pp 299-302.
- Kreuger, R (1988) Focus Groups: A practical guide for applied research, Sage, London.
- Kvale, S (1994) "Ten standard objections to qualitative research interviews" Journal of Phenomenological Psychology, pp 1-28.
- Lincoln, Y and Guba, E (1985) Naturalistic Inquiry, Sage Publications, Newbury Park.
- Liu JL, Wyatt JC. The case for randomized controlled trials to assess the impact of clinical information systems. J Am Med Inform Assoc 2011;18:173e80.
- Mays N; Pope C; Popay J (2005) Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. Journal of Health Services Research and Policy 10 (1):6-20).
- Mays, N Pope C Popay J Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field, Journal of Health Services Research and Policy, 2005 Volume: 10 Number: 3 Supplement: 1 Page: 6 -20.
- McKinlay, J (1992) "Advantages and limitations of the survey approach - understanding older people", in Daly et al (1992) op cit., pp 114-137.
- Mdege, N Man, MTorgerson, D 2011 Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation Journal of Clinical Epidemiology Volume 64, Issue 9 , Pages 936-948, September 2011
- Mercer, S, DeVinney, B Fine, L Green, L Dougherty, D Study Designs for Effectiveness and Translation Research Identifying Trade-offs Am J Prev Med 2007;33(2):139–154)
- Michie, S Fixsen, D Grimshaw, J Eccles, M 2009 Specifying and reporting complex behaviour change interventions: the need for a scientific method, Implementation

Science 2009, 4:40 doi:10.1186/1748-5908-4-40.
- Miles and Huberman (1994), Qualitative Data Analysis: A source book of new methods, Sage, Beverly Hills, Ca. 2nd edin
- Morgan, D (ed) (1993) Successful Focus Groups, Sage, London.
- Needham DM, Sinopoli DJ, Dinglas VD, et al. Improving data quality control in quality improvement projects. Int J Qual Health Care. 2009;21(2):145- 150.
- NSF 2013 User Friendly Handbook for Mixed Methods Evaluations http://www.nsf.gov/pubs/1997/nsf97153/start.htm
- OBSSR 2013 "e-source book on methods in behavioral and social science research", US Office of Behavioral and Social Science Research, NiHwashington.:http://www.esourceresearch.org/eSourceBook/SoftwareandQualitativeAnalysis/1LearningObjectives/tabid/380/Default.aspx Accessed 27 January 2013.
- OBSSR 2013 "e-source book on methods in behavioral and social science research", US Office of Behavioral and Social Science Research, NiH washington.:http://www.esourceresearch.org/eSourceBook/SoftwareandQualitativeAnalysis/1LearningObjectives/tabid/380/Default.aspx Accessed 27 January 2013.
- Ogrinc, G Mooney, S Estrada, Cet al. 2008 The SQUIRE (Standards for QUality Improvement Reporting Excellence) guidelines for quality improvement reporting: explanation and elaboration Qual Saf Health Care 2008 17: i13-i32 doi: 10.1136/qshc.2008.029058
- Øvretveit, J (2002) Action Evaluation of Health Programmes and Change, Radcliffe Medical Press, Oxford.
- Øvretveit J (2012). Do changes to patient–provider relationships improve quality and save money? A review of evidence about value improvements made by changing communication, collaboration and support for self-care. London: The Health Foundation; 2012. www.health. org.uk
- Ovretveit, J 2013 Evaluating Complex Social Interventions: Volume 1: challenges and choices, CIPRS, Veterans Health Administration, Sepulveda, Ca.
- Patton, M (1987) How to Use Qualitative Methods in Evaluation, Sage, London.
- PHAC 2013 Public Health Agency of Canada. Program evaluation toolkit at http://www.phac-aspc.gc.ca/about_apropos/evaluation/resources-eng.php accessed 27th January 2013
- Phillips, C Palfry, C Thomas, P (1994) Evaluating Health and Social Care, Macmillan, London.
- Pope, C and Mays, N (1995a)"Rigour and qualitative research" British Medical Journal, 311, pp 109-112.
- Pope, C and Mays, N (1995b)"Observational methods in health care settings", British Medical Journal, 311, pp 182-184.
- Powell, J Lovelock, R Bray, J and Philp, I (1994) "Involving consumers in assessing service quality using a qualitative approach," Quality in Health Care, 3, pp 199-202.
- Preskill, H Jones, N 2012 A Practical Guide for Engaging Stakeholders in Developing Evaluation Questions FSG Foundation Strategy Group http://trasi.foundationcenter.org/browse.php (Tools and resources for assessing social impact)
- Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. Int J Technol Assess Health Care. 2003;19(4):613-623.
- Reporting guidelines 2013: CONSORT Statement for the transparent reporting of clinical trials EQUATOR – promotes transparent and accurate reporting of

health research http://www.consort-statement.org/?o=1001 http://www.equator-network.org/?o=1001; STROBE Statement for strengthening the reporting of observational studies in epidemiology http://www.strobe-statement.org/; TREND Statement for the transparent reporting of evaluations with non-randomised designs http://www.trend-statement.org/asp/trend.asp; Ogrinc, G Mooney, S Estrada, Cet al. 2008 The SQUIRE (Standards for QUality Improvement Reporting Excellence) guidelines for quality improvement reporting: explanation and elaboration Qual Saf Health Care 2008 17: i13-i32 doi: 10.1136/qshc.2008.029058

- Research Utilization Support and Health, 2009. Resources for dissemnination nhttp://www.researchutilization.org/matrix/resources/index.html, accessed 28th January 2013.
- Robson, C (1993) Real World Research, Blackwell, Oxford
- Rossi, P and Freeman, H (1993) Evaluation - A systematic approach, Sage, London.
- Rothwell PM (2005) External validity of randomised controlled trials: To whom do the results of this trial apply? Lancet 365: 82–93.
- Rothwell PM (2005) External validity of randomised controlled trials: To whom do the results of this trial apply? Lancet 365: 82–93.
- Sale JEM, Brazil K. A strategy to identify critical appraisal criteria for primary mixed method studies. Quality and Quantity 2004;38:351–65
- Sandelowski M. Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies. Res Nurs Health 2000;23:246–55.
- Sapsford, R and Abbott, P (1992) Research Methods for Nurses and the Caring Professions, Open University Press, Milton Keynes.
- Shadish WR, Cook TD, Leviton LC. Foundations of Program Evaluation: Theories of Practice. Newbury Park: Sage Publications; 1991.
- Shiell A, Hawe P, Gold L. Complex interventions or complex systems? Implications for health economic evaluation. BMJ 2008;336:1281-3.
- Smith, M Glass, G and Miller, T (1980) The Benefits of Psychotherapy, Johns Hopkins University Press, Baltimore.
- Speroff T, O'Connor GT. Study designs for PDSA quality improvement research. Qual Manag Health Care. 2004;13(1):17-32.
- SRM 2013 Research Design and Mixed-Method Approach: A Hands-on Experience (tutorial website) http://www.socialresearchmethods.net/tutorial/Sydenstricker/ bolsa.html
- St Leger, A Schienden, H Walsworth-Bell, J (1992) Evaluating Health Service Effectiveness, Open Univeristy Press, Milton Keynes.
- Stephenson, J Imrie J 1998 Why do we need randomised controlled trials to assess behavioural interventions? BMJ 1998;316:611–3
- Stetler CB, Damschroder LJ, Helfrich CD, et al. A Guide for applying a revised version of the PARIHS framework for implementation. Implement Sci. 2011;6:99. PMID: 21878092.
- Strauss, A Corbin, J (1990) Basics of Qualitative Research, Sage, London.
- Tashakkori A, Teddlie C, (Eds): Handbook of mixed methods in social and behavioural research London: Sage; 2003. 11.
- TEC 2013 Checklists. Western Michigan University. The Evaluation Centre. http://www.wmich.edu/evalctr/checklists/. accessed 27th January 2013
- Thor J, Lundberg J, Ask J, et al. Application of statistical process control in healthcare improvement: systematic review. Qual Saf Health Care. 2007;16 (5):387-399.

- Treweek S and Zwarenstein M 2009 Making trials matter: pragmatic and explanatory trials and the problem of applicability Trials 2009, 10:37 doi:10.1186/1745-6215-10-37
- Tunis SR, Stryer DB, Clancy CM (2003) Practical clinical trials: Increasing the value of clinical research for decision making in clinical and health policy. JAMA 290: 1624–1632.
- van de Vijver, F and Leung, K (1997) Methods and data analysis for cross-cultural research, Sage, London.
- VHA 2013 Cyber seminars, Veterans Health Administration Health Service Research and Development and QUERI web site http://www.hsrd.research.va.gov/Cyberseminars/#.UQVTNppYtPw accessed 27th January 2013
- Wells KB, Tang L, Miranda J, Benjamin B, Duan N, Sherbourne CD. The effects of quality improvement for depression in primary care at nine years: results from a randomized, controlled group-level trial. Health Serv Res. 2008;43(6):1952-1974.
- Wiltkin, D Hallan, L and Dogget, M (1992) Measures of Need and Outcome for Primary Health Care, Oxford Medical Publications, Oxford.
- Yin, R (1989) Case Study Research: Design and methods, Sage, Beverly Hills. (1994 edition).