2

Fundamental Issues in Evaluation and Research in Violently Divided Societies

An Analysis of the Literature

Philip McDermott, Zahbia Yousuf,
Jacqueline Strecker and Ethel Méndez

The evolution of evaluation is closely tied to the rise of publicly funded institutions and social programmes, primarily in North America and Europe. Respected evaluator Donna Mertens notes that evaluation, in its earliest form, appeared in the 19th century when 'the [US] government first asked external inspectors to evaluate public programs, such as prisons, schools, hospitals and orphanages' (Mertens, 2001, p. 367). However, most other commentators date the origin of evaluation (as we know it today) to the 1960s when more standardised forms and approaches to evaluation came to the fore (Georghious and Laredo, 2005, p. 1; McCoy and Hargie, 2001; Scriven, 2003).

Since this period, evaluation has established itself as a discipline in its own right. An important consequence of this has been a move towards *professionalisation* of evaluation as a field of theory and practice (Scriven, 2003; p. 7; Hay, Chapter 9 in this book). Thus, bodies such as the American Evaluation Association (AEA) and the European Evaluation Society have established quality standards and formal principles of practice for members and for the profession more broadly. These or similar standards of technical competency and integrity for evaluators

have since been adopted by numerous national and regional evaluation associations in the Global North and South. Nevertheless, the field of contemporary evaluation theory and practice carries the legacy of the second half of the last century in its underpinning theories and methodologies, derived largely from government-driven desires to assess social programming, particularly in the areas of health and education (Rossi and Lipsey, 2004, p. 8).

Within the sphere of interest of this book, it bears noting that although the evaluation of academic and extra-academic research has also been going on since the 1970s (Luukkonen, 2002, p. 81), the debates and evolution in this field took place separately, albeit in parallel, to those taking place in the field of social policy and programme evaluation. As such, the evaluation of research was not a significant point of reference in the early evolution of the field of evaluation science (McCoy and Hargie, 2001; Rossi and Lipsey, 2004).

Indeed, contemporary evaluation research and practice have continued to be weighted towards social policy and programming, with a more specific focus on the evaluation of research, as interested stakeholders (research funders, users of research and researchers themselves) have come to understand that there is much learning that can be taken from the intersection of the fields of programme or policy evaluation and research evaluation. Fleshing out this learning has become increasingly important as those with an interest in producing actionable, social change research have realised that approaches to and methods for evaluating research uptake, use and impacts are imperfect and underdeveloped. As noted by Bush and Duggan in the introductory chapter to this book, this task is even more difficult in violently divided society (VDS) contexts that are characterised by volatility, complexity and non-linearity. In addition, the trajectories of research in these settings, from the moment of its conceptualisation through to it use for social change objectives, are highly politicised.

Overview

Given that the literature on social programme and policy evaluation and research evaluation is vast, this chapter reviews a selection of this literature in order to explore and ground the conceptual and theoretical foundations of the current book. It considers how the process of evaluation has evolved and developed historically, and highlights the main

trends within the field. This will guide us to the relevant principles and concepts in the evaluation literature for use in subsequent discussions in this book of the evaluation of research in and on VDS. It will also allow us to identify missing components in the existing literature and potential gaps in theory and practice.

The sources included in this chapter were selected for detailed review from a larger set of references generated through searches on databases of peer-reviewed journals, Google and Google Scholar. The principal and secondary researchers involved in the VDS project (authors of this book) also provided articles and chapters which they considered to be of particular importance in influencing debates and forging trends in the fields of evaluation and of peace and conflict research in general.

The literature reviewed was selected on the basis of its potential to illuminate the primary question that forms the basis of this book: How can we improve evaluation practice to better understand the difference that research makes in VDS? Secondary questions that guided the review include: Why is evaluation so much more difficult in these contexts? What can we learn from the current evaluation practice in the Global North and South?

In order to organise learning around these questions, the chapter categorises sources reviewed around four constituent works of literature. The first body of literature is rooted in the evaluation of publicly funded social programmes and policies, a process which draws our analytical gaze back to the very origins of the field of evaluation. The second constituent literature brings to the fore, the important issue of violence and examines the much more recent, and still evolving, work on the evaluation of conflict prevention, peacebuilding and humanitarian assistance programmes. The third constituent literature is rooted in efforts to evaluate academic and extra-academic research. The final constituent literature is the most embryonic of all, and is explored throughout the chapters that make up this book—the evaluation of research in and on VDS. Debates and trends in the evaluation of international development aid programmes tend to cross-cut the four constituent works of literature and for this reason, this chapter also draws from this area of practice as needed.

In looking across these pieces of literature in the context of the VDS research project, it quickly became apparent that there are recurring themes, debates, problems and issues that absorb the attention of stakeholders involved in research and evaluation. We refer to these as fundamental issues, understood as 'underlying concerns, problems or

choices that continually resurface in different guises throughout evaluation work' (Smith and Brandon, 2008, p. vii). By their very nature, these are issues that can never be finally resolved to the satisfaction of all stakeholders; this is especially true in VDS contexts where the ubiquitous presence or threat of violence renders these fundamental issues more extreme. Of the myriad fundamental issues that surface, the following are examined in varying degrees of depth across the works of literature as a means of deepening our understanding of the ways in which particular norms and values have become embedded in the field of evaluation, affecting the conceptual framing of the evaluation of research in VDS:
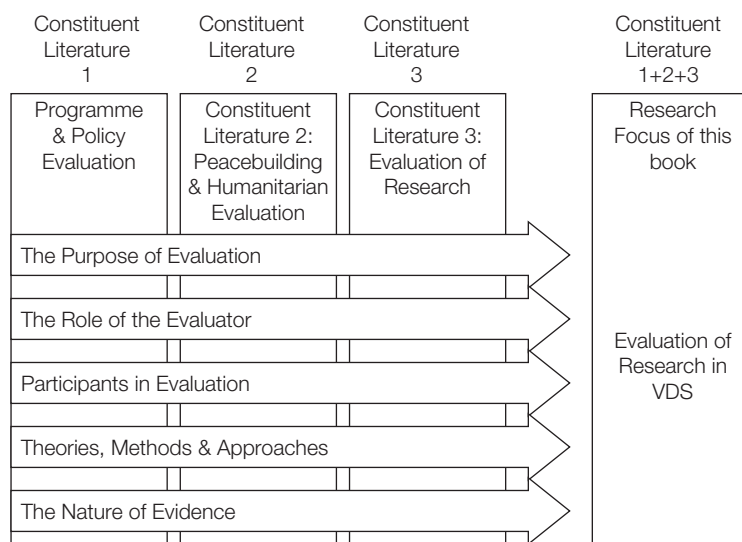
- The purpose of evaluation
- The role of the evaluator
- Who participates in evaluation
- Theoretical and applied approaches and methods
- The nature of evidence

Figure 2.1 illustrates the relationship and interaction in these four focus areas across the constituent pieces of literature, and also the nuanced differences between each area of evaluation. The purpose of the map is to show that, although there are many areas of evaluation in which there are coinciding interests, there are also numerous complexities which need to be borne in mind. Also, the map forms the basis for the structure of the rest of our chapter where we offer a more detailed explanation of the constituent pieces of literature and their fundamental issues.

Constituent Literature 1: The Evaluation of Programmes and Policies

As noted, the field of evaluation arose predominantly in the public administration domain in USA, Europe and Australia. It began in the fields of human, social and welfare services (including healthcare, social deprivation and poverty) with an early focus in the field of education. The intention of such evaluations was to assist in improving the quality of social programmes and policies. Much of the material written to date continues to focus on this particular area and, therefore, acts as the first port of call in our discussion. Box 2.1 provides a brief primer with basic evaluation terminology.

Figure 2.1
Constituent Literatures and Cross-cutting Themes in the Evaluation of Research in VDS

Constituent Literature 1	Constituent Literature 2	Constituent Literature 3		Constituent Literature 1+2+3
Programme & Policy Evaluation	Constituent Literature 2: Peacebuilding & Humanitarian Evaluation	Constituent Literature 3: Evaluation of Research		Research Focus of this book
The Purpose of Evaluation				
The Role of the Evaluator				Evaluation of Research in VDS
Participants in Evaluation				
Theories, Methods & Approaches				
The Nature of Evidence				

Source: Author.

Box 2.1
Key Evaluation Terms—Evaluation, Approach and Method

Evaluation Type: refers to whether the evaluation will be formative, summative or developmental.

Formative Evaluation: is intended to improve performance; most often conducted during the implementation of a programme.

Summative Evaluation: is intended to provide information about the merit or worth of a programme; most often conducted at the end of a programme.

Developmental Evaluation: positions the evaluator as a part of a programme's design and development process. It is done is real-time; has a series of short, rapid feedback loops; and allows the entity being evaluated to quickly adapt its strategies and activities.

Evaluation Approach: commonly called an evaluation theory, an approach describes the philosophy and process of evaluation.

Method: refers to the data collection strategy (qualitative, quantitative, mixed) tools and instruments used in an evaluation.

Source: Fournier (2005); Preskill and Russ-Eft (2005).

The Purpose of Evaluation:
Accountability versus Learning

Programme evaluation was first developed with the primary purpose of improving the quality of social programmes. Within this aim, a larger tension surfaced amongst proponents who viewed evaluation as a tool for accountability for funds disbursed and those who saw it as a means for learning and improvement.

The accountability model of evaluation became popular in the early 1960s. In the USA, Senator Robert Kennedy was concerned with the spending of federal money in education and, thus, fought to include an evaluation clause in the Elementary and Secondary Education Act (ESEA). With the successful passing of this act, evaluation requirements became part of every US federal grant, and evaluation as a means for judging whether or not a programme was successful, was inadvertently linked to the allocation of subsequent funding (Preskill and Russ-Eft, 2005). Similar practices were also occurring in Europe and Australia, raising the interest and need for an expanding evaluation field.

As affirmed by Patton (1997, p. 11):

> Programme evaluation as a distinct field of professional practice was born of two lessons: first, the realisation that there is not enough money to do all the things that need doing; and second, even if there were enough money, it takes more than money to solve complex human and social problems. *As not everything can be done, there must be a basis for deciding which things are worth doing.* Enter evaluation.

Consequently, the agencies and other bodies that fund public programmes have at times used evaluation findings to justify expenditure and future allocation of financial resources. In this respect, evaluation was charged with identifying flaws in a completed initiative in order to ensure that similar projects or programmes would not be funded in the future. Objective-based studies, pioneered by Ralph Tyler, Percy Bridgman and E. L. Thorndike were well tailored for this function, identifying the initial aims and objectives of the programme and then measuring whether or not these objectives had been met.

As the evaluation field developed, debates concerning the aim and purpose of evaluation broadened, and programme evaluation moved beyond a one-dimensional accountability approach, shifting the debate

towards a wider discussion around the need to use evaluation as a learning tool rather than an end unto itself. Evaluation practitioners and scholars problematised the endemic power structures within the funder–client relationship, and identified the limitations of objectives-based approaches. For example, evaluators noted that in the course of a programme the initial goals may alter due to the nature of the programme, and, as a result, evaluating with these initial aims and objectives in mind may be detrimental to a fair and thorough judgement of the programme (Parlett and Hamilton, 1977). Others have similarly suggested that goal formation in many programmes is an ongoing process, with the major aims and objectives only becoming evident to leaders and managers once the actual programme is under way. Some evaluators have also advocated for goal-free approaches where the achievement of the initial aims and objectives is not a major concern with respect to the wider scale of the entire evaluation (Scriven, 1991). Instead, consideration is given to the wider context in which the programme has developed over time and what it has achieved.

In the light of these and other discussions, the field of evaluation has largely embraced evaluation to serve three broad functions:

1. Accountability: measuring and judging the merit and worth of programme results and accounting for use of resources;
2. Knowledge generation: creating new understandings about what works and what does not;
3. Programme planning or improvement: supporting clear, well-designed, feasible and measurable programmes with a view to increasing overall organisational effectiveness (Rutnik and Campbell, 2002).

Proponents of evaluation as a means for learning also contend that an evaluation should not just be viewed as the 'retrospective analysis of programme effects', but rather that it can play an important role in the design, implementation and ongoing monitoring of a programme (McCoy and Hargie, 2001, p. 319). These perspectives have led to increased investments into *formative evaluations*. Formative evaluations are episodic in nature, and are conducted during the development of a programme with the intent to improve or refine it (Scriven, 1991, p. 168). Formative evaluations are often contrasted with *summative evaluations*, which are conducted after stabilisation or completion of a programme, and are used to make a final evaluative judgement to determine the programme's merit, worth or value.

A third evaluation form, *developmental evaluation* (DE), was introduced by Michael Quinn Patton in the early 1980s. DE embeds the evaluator into the programme development process, so that the evaluator can provide regular and informed feedback to the programme team to assist with on-going refinements to design before a programme is tested and while a programme is underway. DE, and in particular the concept of the *embedded evaluator*, has raised further questions about what exactly the role of the evaluator is.

The Role of the Evaluator

The increased professionalisation of the field of evaluation has given rise to practitioner voices and created rich debates and discussions about the role of evaluators. There are some evaluators and commissioners of evaluation who maintain that evaluators should play only one role: providing arm's length, independent judgements about a programme's merit or worth (Scriven, 1991; Stufflebeam, 1994). In contrast, others have advocated for multiple evaluator roles, depending on the evaluation context. These include, but are not limited to: advocate, activist, change agent, collaborator, communicator, diplomat, external expert, group facilitator, information broker, internal colleague, methodologist, organisational analyst, politician, problem solver, trainer and technician (Greene, 1997; Love, 1991; Patton, 2008). The evaluator's role is often influenced by several factors including the epistemological and ontological view of the evaluator and those managing the evaluation; the type of evaluation being conducted; the situation and challenges present in the evaluation; and the type of evaluator (internal or external).

Conley-Tyler (2005) highlights that the internal–external debate is often underemphasised in evaluation practice. 'Given how common this scenario is, it is surprising that the choice between internal and external evaluators has not been the subject of much critical debate. Too often the issue is assumed either one way or the other without discussion of the issues involved' (Conley-Tyler, 2005, p. 3). Conley-Tyler further argues that there is a distinct separation between literature aimed at business and organisational audiences, which favours internal evaluators, and literature aimed at evaluation professionals, which privileges the external evaluator (Conley-Tyler, 2005).

External or *independent* evaluators are those not associated with the evaluated programme, and thus, in theory, have no stake in the evaluation

outcomes (Molund and Schill, 2004, p. 17). Scriven (1991, p. 160) out-lines the benefits of external evaluators stating that

> the external evaluator is less likely to be affected by personal or job-benefit considerations, is often better at evaluation; has often looked closely at comparable programs, can speak more frankly because there is less risk of job loss or personal retribution/dislike, and carries some cachet from externality.

The role of the external evaluator is often dependent on the type of evaluation selected and its methodology, but overall his or her role is to provide an external assessment of the programme as an outside observer. One limitation of external evaluation is that stakeholder participation is often limited out of a fear of jeopardising evaluation independence. As a result, stakeholders can feel removed from the evaluation process, and may dismiss or disregard the evaluation findings. In contrast, internal evaluators are organisationally attached to the programme being evaluated. According to Scriven, the internal evaluator

> knows the programme better and so avoids mistakes due to ignorance, knows the people better and hence can talk to them more easily, will be there after the evaluation is finished and hence can facilitate implementation, probably knows the subject matter better, costs less, and is sure to know of some other comparable projects for comparison. (Scriven, 1991, p. 160)

Internal evaluators are also at an advantage because they are familiar with the political and cultural factors which may be affecting the environment within which the programme is operating (Braskamp et al., 1987). Love (1991) cautions that the role of an internal evaluator is sometimes con-flated with manager responsibilities, resulting in increased tension and poor evaluation quality.

While internal evaluators have been used in both formative and sum-mative evaluations, a particular feature of formative evaluation is that it advocates a position whereby evaluation is something which need not be undertaken solely by a formal external evaluator. Instead, all actors may be encouraged to think about evaluation throughout the implemen-tation of a particular programme. For example, if the programme staff is encouraged to think about evaluation issues throughout, this is viewed by many as having a positive impact on the project or programme as a

whole. In other words, evaluation becomes an *in-built* feature of a programme itself (Scriven, 1991).

One assumption that often pervades the external–internal debate is the notion of objectivity. In 1988, Cummings et al. proposed rationales for why the objectivity of internal evaluators may be compromised by the underlying value systems of organisational policies. This position is not, however, supported by most current evaluation theories (Conley-Tyler, 2005). In contrast, evaluation scholars have continued to challenge the objectivist approach arguing that every evaluator, regardless of whether he or she is internal or external, brings implicit values to an evaluation (House, 1980). As a result, the majority of scholars have since highlighted the importance of impartiality over objectivity. There are, however, a few remaining scholars (such as Scriven) who maintain that objectivity should remain the ideal.

While value can be found in both internal and external evaluator roles, the tendency to privilege one type over the other is representative of the different worldviews at play. 'In one world view, evaluation is something that should be carried out primarily by professionals (external evaluators), while in the other world view, evaluation skills should be spread as widely as possible (such as to management and other staff)' (Conley-Tyler, 2005, p. 10). To help transcend these rigid distinctions, some practitioners and evaluation commissioners have started to advocate for combining internal and external evaluation, and thinking about evaluation relationships as existing along a continuum rather than as polarised extremes (Patton, 2008). Patton suggests, 'there are a good many possible combinations of internal and external evaluations that may be more desirable and more cost-effective than either a purely internal or purely external evaluation' (Patton, 2008, p. 222). The combination of internal and external evaluation reaps the benefits of the diverse types of evaluation roles, but proper execution can be difficult, as care is required to ensure that the responsibilities are appropriately divided.

The role of the evaluator is therefore variable and is often articulated and negotiated during the process of an evaluation. Although the increasing professionalisation of the field of evaluation has provided fuel for fiery and recurring debates about evaluator roles, little agreement has been reached as to whether evaluation should be viewed as a modality or as a profession. Increased desire for professionalisation is leading many evaluation practitioners to call for the strengthening of the field through training and accreditation.

Evaluation Approaches

Another prominent discussion that permeates evaluation is the approaches debate. Chambers defines an approach as, 'an orientation of behaviour, attitudes, and mindset associated with a method, methods or methodology' (Chambers, 2010, p. 10). There are two dominant approaches that have dominated evaluation practice: the positivist approach and constructivist approach. These approaches influence and shape the creation of knowledge paradigms in general, and by extension, the course of public policy discussion, discourse and practice. The duelling evaluation approaches are driven by different epistemologies rooted in contrasting worldviews, which are uniquely constructed from different values, assumptions, habits and beliefs.[1]

Programme evaluation was first governed by a positivist or post-positivist approach, which originated from experimental traditions in agriculture. Patton affirms, '[e]valuation was initially dominated by the natural science paradigm of hypothetico-deductive methodology, which values quantitative measures, experimental design, and statistical analysis as the epitome of "good science"' (Patton, 2008, p. 423).

As evaluation developed, the constructivist approach was established as an alternative, based on anthropological traditions, with a preference for using qualitative data, naturalistic inquiry and rich engagement with participants of the programme (Patton, 2008). Constructivism is premised on an understanding that social reality is rooted in the diversity of human experience, specifically the perceptions of those realities. Social reality is not a collection of unchanging, empirical facts, so much as a collection of subjective and inter-subjective experiences of events and conditions. Constructivists would view a *fact* to be both fluid and contingent upon social structures and processes. As such, the approach is advocated as being more responsive to the various characteristics of the programme as well as to the environment and issues that may be flagged by stakeholders.

Ultimately, the constructivist approach 'places the evaluators and programme stakeholders at the center of the inquiry process, employing all of them as the evaluation's "human instruments"' (Stufflebeam, 1999, p. 57). Constructivist evaluation has also provided the ontological foundations for the creation of additional approaches that emphasise certain facets of evaluation practice in order to ensure increased evaluation relevance and use. Some unique and prominent approaches born

from constructivist ontology include DE, utilisation-focused evaluation (UFE) and participatory evaluation (McCoy and Hargie, 2001; Patton, 1997). These approaches will be further discussed later on in this chapter.

Two other distinct approaches that are rooted in both constructivist and positivist perspectives and which merit discussion here are goal-free evaluation and theory-driven evaluation. Goal-free evaluation was pioneered by Michael Scriven in the early 1970s, as an alternative approach to goal-based evaluations. Scriven argued that since most programmes either fall short or over-achieve the goals defined during initial planning, it does not make sense to restrict an evaluator to making judgements about achievement along these lines. Scriven (1972) noted that 'consideration of evaluation of goals was an unnecessary but also a possibly contaminating step… the less the external evaluator hears about the goals of the project, the less tunnel vision will develop, the more attention will be paid to looking for actual effects' (as quoted in Patton, 2001, p. 170). Goal-free evaluation, therefore, focuses not on preset goals and objectives, but on actual programme effects. Other practitioners have cautioned that this approach runs the risk of replacing the evaluator's goals with those of the programme and assumes the presence of both internal and external evaluators (Alkin, 1972; Patton, 2001).

Theory-driven evaluation assumes no direct ideological bias and can be traced as far back as the 1930s and Tyler's notion of testing programme theory. However, it was not until 1990, with the publication of *Theory-Driven Evaluation* by Chen, that this approach became prominent within evaluation practice (Coryn et al., 2011). Theory-driven evaluation is predicated on the use of an explicit theory or model to understand how a programme has caused the intended or observed programme outcomes (Rogers et al., 2000). This approach is based on a conceptual notion that an evaluation should explain programme theory, while empirically examining how programmes cause observed outcomes (Coryn et al., 2011). 'By developing a plausible model of how the programme is supposed to work, the evaluator can consider social science theories related to the programme as well as programme resources, activities, processes, and outcomes and assumptions' (Preskill and Russ-Eft, 2005, p. 121). As this approach grew in popularity, richer models were developed for understanding a variety of programme contexts. Coryn et al. (2011, p. 202) highlight that:

> In earlier conceptualisations, numerous theorists, including Weiss (1997; 1998) and Wholey (1979), among others, tended to favour

> linear models to describe programme theories. In recent writings, others (e.g., Chen 2005; Rogers, 2008) have advocated for more contextualised, comprehensive, ecological programme theory models … these types of models are intended to integrate systems thinking in postulating programme theory, taking contextual and other factors that sometimes influence and operate on programme processes and outcomes into account.

Despite the creation of these new approaches, Patton (2010) still questions their suitability for complex adaptive situations, which are constantly changing and unpredictable. This observation is of particular importance for evaluation practice in VDS situations which are characterised by fluidity, unpredictability and political complexity, as will be discussed further under constituent literature 2. Another weakness of the theory-driven approach is that it can be very time-intensive and, therefore, not feasible within certain evaluation timelines (Stufflebeam, 1999). Notwithstanding these critiques, theory-driven approaches continue to expand and accumulate their own debates and discourse.

With the creation of each new evaluation approach, longstanding debates on the worth of stakeholder participation, the types of methodologies used and what qualifies as evidence continue to re-surface as fundamental issues. These debates are ultimately influenced by differing ontological views, but many scholars have noted that the differing perceptivities actually help enrich and strengthen evaluation practice.

Who Participates

The merits of stakeholder participation are another important discussion that has permeated the programme and policy evaluation literature. Evaluation stakeholders are defined as individuals who have a vested interest in the evaluation findings (Patton, 2008) (see Box 2.2).

While different stakeholders each have a *stake* in the programme, their individual interests are often divergent, and can at times compete. For this reason, early positivist evaluation argued for limiting stakeholder participation, rationalising that the involvement of stakeholders would undoubtedly compromise the objectivity of the evaluation, and could bias the findings to suit personal interests. In the 1970s, however, studies started to emerge which showed a correlation between stakeholder involvement and evaluation uptake. Patton highlights that evaluations that had the 'presence of an identifiable individual or group of

Box 2.2
Key Evaluation Terms—Categories of Stakeholders

Evaluation Stakeholders: individuals who have a vested interest in the evaluation findings. Stakeholders can be clustered into the following five categories:

1. Individuals with authority to make programme decisions (policy-makers, funders, advisory boards, etc.)
2. Individuals with direct responsibility for the programme (developers, administration, managers, etc.)
3. Intended beneficiaries of the programme (individuals, families, communities, etc.)
4. Individuals who have been disadvantaged by the programme (those who lost in funding opportunities)
5. Individuals with indirect interest in the programme (journalists, taxpayers, etc.)

Source: Patton (2008).

people who care about the evaluation and the findings it generated' had an increased likelihood of long-term follow-through in getting evaluation findings used (Patton, 2008, p. 69).

Over time, evaluation practice has seen an increasing trend towards engaging stakeholders in evaluation across a wide variety of approaches (Christie, 2003). Today, stakeholder involvement has become an accepted practice in the evaluation profession, and has been noted as a hallmark of exemplary practice (Fitzpatrick, 2004; Patton, 2008). It is important to note, however, that there are prominent distinctions between how stakeholders engage. The degree, scope and nature of engagement differ significantly across evaluation practice, and have been theorised within a number of different evaluation approaches.

Participatory evaluation is one such approach which developed in the 1970s and 1980s to advocate for wider stakeholder involvement in the evaluation process. Advocates of participatory evaluation maintain that the process of evaluation needs to move outside of simple scientific frameworks towards approaches in which the target population is afforded a key role in shaping the evaluation plan (Drewett, 1997). Participatory evaluation ultimately presents a bottom-up approach that places engagement as a central focus. Cousins (2011) highlights that there are two streams of participatory evaluation: the practical stream, which is problem-solving and utilisation-oriented; and the transformative stream, which is emancipatory and empowerment-oriented. Both streams are similar in that they involve a partnership between the stakeholders and the evaluator, the parameters of which should always be discussed during the design of an evaluation.

Cousins and Whitmore (1998) highlight three dimensions of participation or collaborative inquiry. These include (a) who controls the technical decision-making in the evaluation process (a researcher–evaluator continuum); (b) stakeholder selection for participation (a continuum from all legitimate groups to just primary intended users); and (c) the depth of stakeholders participation (a continuum from consultation to rich involvement in all aspects of inquiry). How these different dimensions play out within an evaluation context determines the overall nature of participation and the benefits generated. Apart from increasing evaluation use, participatory approaches can assist stakeholders in building evaluation skills and knowledge; creating internal support-networks; and achieving a greater sense of self-efficacy (Whitmore, 1988).

The implementation of participatory evaluation can, however, add an additional level of complexity for the evaluator. Patton notes that 'participatory evaluation partnerships can be particularly challenging in part because of underlying fears, bad past experiences with evaluation, resistance to reality testing, and cultural norms that undercut openness and questioning' (Patton, 2008, p. 176 citing Podems, 2005). In spite of these barriers, the principles of participatory evaluation continue to inspire innovative approaches and methods which are all designed to incorporate stakeholders into the evaluative process.

Michael Quinn Patton's UFE approach is one example of an innovative, practical participatory evaluation approach. UFE is founded on the understanding that evaluations should be judged by their utility and actual use. The approach, therefore, focuses on the needs of the primary intended users, and considers how all aspects of the evaluation, from planning to dissemination, will affect use. Patton maintains that 'since no evaluation can be value-free, UFE answers the question of whose values will frame the evaluation by working with clearly identified, primary intended users who have responsibility to apply evaluation findings and implement recommendations' (Patton, 2012, p. 4). Similar to the majority of participatory approaches, there are no specific methods or theories advocated for within UFE. Instead, this approach suggests a need for situational responsiveness, whereby the evaluator and intended users work together to make decisions and source an appropriate evaluation design.

UFE is only one participatory approach, among many. However, it provides a prolific example of the polarised views regarding the merits of stakeholder participation. Evaluation practice has shifted significantly from the early positivist traditions that denied engagement to a flood of new constructivist approaches, which are built explicitly on stakeholder

involvement. While discussions and debates still circulate on the ideal nature and degree of engagement, it is clear that involving stakeholders is beneficial to evaluation practice.

Methodological Practices

Another of the most prominent and important fundamental issues treated in the programme evaluation literature focuses on the appropriateness of methodological tools. While all evaluation approaches affirm that the key to evaluation is evidence-based research, there has been significant disagreement over whether quantitative or qualitative methods should be employed. In general terms, the collection of quantitative data is often supported by positivist arguments which hold that quantitative data yields more objective and accurate information because they are collected using standardised methods. These methods can be replicated and analysed using sophisticated statistical techniques and are based on hypothetico-deductive and quantitative models where an evaluator relies heavily on the use of statistics in order to measure the impact of a socially grounded programme. Quantitative measures focus on countable data that can be collected from information systems, official indicators, programme records, questionnaires, quasi experiments, rating scales, standardised observation instruments, norm-referenced tests, a posteriori statistical test and significance tests (Patton, 2008; Stufflebeam, 1999). The overall aim is to use statistics to facilitate comparisons and identify relationships between the programme and its outcomes.

In contrast, qualitative methods are generally supported by constructivist approaches and aim to capture personal meaning and participant perspectives on programme experience, by examining the diverse ways in which people articulate and express themselves. Qualitative data is, therefore, open-ended and not predetermined by imposed standardised choices. Qualitative methods include: 'ethnography, document analysis, narrative analysis, purposive samples, participant observers, independent observers, key informants, advisory committees, structured and unstructured interviews, focus groups, case studies, study of outliers, diaries, logic models, grounded theory development, flow charts, decision trees, matrices, and performance assessments' (Stufflebeam, 1999, p. 34). Advocates of qualitative methodologies highlight their ability to deepen understanding on how and why programmes work, and what outcomes mean (sense-making).

While both qualitative and quantitative methods have received support and criticism from their contrasting ontological camps, many scholars and evaluation practitioners have acknowledged that both contribute in important and differing ways to understanding a programme. Consequently, many evaluation approaches have encouraged the use of mixed methods. Stufflebeam suggests, 'by using both quantitative and qualitative methods, the evaluator secures cross-checks on different subsets of findings and thereby instils greater stakeholder confidence in the overall findings' (Stufflebeam, 1999, p. 34). From a practical perspective, it has been noted that much current evaluation practice routinely involves multiple types of data. For example, evaluations often use time series analyses of outcome monitoring or administrative performance data with interviews or surveys, together with case studies including participant observation (Datta, 2001).

Although there are infinite possibilities for applying mixed methods to technical aspects of evaluation, some such as Guba and Lincoln have argued that because the assumptions of different paradigms are incompatible, it is not possible to mix paradigms in the same study (Guba and Lincoln, 1989). Similarly, Patton suggests that even when mixed methods are used, 'one kind of data is often valued over the other' (Patton, 2008, p. 435). In contrast, other practitioners have suggested the possibility of applying mixed methods not only to technical procedures, but also to conceptual aspects of the evaluation process (Greene, 1997).

Methodological debate on evaluation has experienced significant ebbs and flows throughout the development of the field. Qualitative methodologies have gained relative acceptance, which is a significant change from early 1970s, which privileged quantitative methods (Patton, 2008, p. 421). Although disagreement about the relative merits of each has always existed, Mark (2003) suggests that the 'paradigm wars', which occupied the field during the 1970s and 1980s, quieted down at the turn of the century, settling into an uneasy peace. In 2003, however, these debates were refuelled when the US Department of Education's Institute of Education Science publicised their commitment to prioritising funding for quantitative approaches such as experimental and quasi-experimental designs. In response, hot exchanges and debates were reignited within the evaluation community, focusing attention on what counts as credible evidence and what method should be regarded as evaluation's 'gold standard'.

What Counts as Evidence:
Attribution versus Contribution

Although discussions about the merits of methodologies and evalua-tive design are entangled within the disputes about evaluation evidence, Christie and Fleischer speculate that the credible evidence debate is more likely rooted in 'how each "side" conceptualises "impact"' (Christie and Fleischer, 2009, p. 20). Both positivists and constructivists recognise that no method is infallible; however, disagreements arise on how to approach strengthening evaluation evidence. These diverse perspectives are once again supported by their underlying viewpoints about truth and science and perceptions of the world (Christie and Fleischer, 2009).

A positivist or post-positivist approach views the world as some-thing which is fixed and constant and, therefore, suggests that impact can be determined by providing evidence of a causal relationship between the intervention (e.g., the programme or other unit of analysis) and the outcome generated. Davidson (2000) suggests that causation is critical to determine; in its absence, financial and personal resources may be wasted in expanding defective programmes. In order to determine cau-sation, experimental or quasi-experimental approaches call for the use of counterfactuals (a control or comparison group that allows the evalu-ator to establish what would have happened if the programme had not taken place).

With the release of the Center for Global Development's 2006 report *When Will We Ever Learn* (Savedoff et al., 2006), Randomised Control Trials (RCTs) were purported to constitute a fail-safe *gold stan-dard* methodology for measuring the impact that a programme may have had in relation to a particular established goal (Clinton et al., 2006).

Support for RCTs as a gold standard has since fluctuated over the years but continues to be the 'design of choice' for some funders of inter-national aid, who have dubbed this approach 'impact evaluation'. Many evaluators have highlighted the limitations of RCTs noting that

> RCTs are not always best for determining causality and can be mis-leading. RCTs examine a limited number of isolated factors that are neither limited nor isolated in natural settings. The complex nature of causality and the multitude of actual influences on outcomes render RCTs less capable of discovering causality than designs sensitive to local culture and conditions and open to unanticipated causal factors. (AEA, 2003)

The feasibility of using experimental evaluation designs with RCTs for evaluating certain programmes including humanitarian relief, peace-building and democracy strengthening in contexts affected by violence and conflict has also been called into question given that these types of programmes are typically *emergent* and 'their impacts are often difficult to evaluate or measure using established tools' (Stern et al., 2012). It has also been noted that experimental and quasi-experimental approaches generally will have very limited utility because they cannot answer many important impact evaluation questions regarding how a particular intervention might exacerbate conflict drivers or enable structures and processes for peace (Chigas et al., 2014, p. iv). While researchers, evaluators and funders continue to argue the merits and demerits of experimental and quasi-experimental approaches, what is clear is that the debate has pushed actors in the worlds of evaluation and research to strive for improved evidence quality through the exploration of multiple approaches to *impact evaluation* (see Befani et al., 2014; Stern et al., 2012).

While all experimental methods have differing strengths and weaknesses, their reliance on counterfactual evidence is inconsistent with constructivist ontology. Constructivist worldviews suggest that it is impossible to distinguish cause and effect, since relationships are multidirectional and, therefore, everything is impacting everything at once (Christie and Fleischer, 2009). According to this perspective, providing evidence based on mere counterfactual claims is not inherently credible, since any number of factors present within a given context could have created the same result, even if the result was not produced by the intervention (Cook et al., 2010). Furthermore, it is suggested that 'evaluations need to recognise policy and programme interests of sponsors and yield maximally useful information, given the available resources and programme constraints' (Rossi and Freeman, 1989, p. 40). The assumption here is that there is a requirement for an evaluation to be designed in order to satisfy the information needs and particular agenda of stakeholders rather than be stifled through narrow adherence to abstract scientific norms (Stufflebeam and Shinklefield, 1985). Reliance on solely scientific means in the field of evaluation has, thus, been questioned and regarded as being somewhat restrictive in garnering a truly multilayered perspective on the impact of a particular programme.

Constructivists, therefore, maintain that credible evidence can be established in a number of diverse ways. Mathison (2009) suggests that evidence credibility is dependent on experience, perception and social convention. Rallis (2009) proposes that along with methodological

rigour, characteristics of goodness, moral correctness and probity determine evidence credibility. From a similar perspective, Greene (2009) highlights that credible evidence 'needs to account for history, culture, and context; respects differences in perspective and values; and opens the potential for democratic inclusion and the legitimisation of multiple voices' (Donaldson, 2009, p. 15).

The differing perspectives of evidence creditability will be further expanded when this chapter discusses research quality or excellence. However, what is important to note here is that within the programme evaluation literature there no single agreed definition of credible evidence. In the epilogue of *What Counts as Credible Evidence in Applied Research and Evaluation Practice?*, Donaldson (2009) laments that, 'we are a long way from consensus and a universal answer to the question of what counts as credible evidence in contemporary applied research and evaluation' (Donaldson, 2009, p. 249). Instead of moving towards consensus, Donaldson suggests that the different understandings of evidence credibility are dependent on multiple characteristics including: the question(s) of interest; the context; assumptions made by the evaluators and stakeholders; the evaluation theory used to guide practice; and practical time and resource constraints (Donaldson, 2009, p. 250).

The programme evaluation literature has, thus, demonstrated the influence that ontology has on framing perspectives on the purpose of evaluation, the merits of stakeholder participation, what approaches and methodologies should be used and what qualifies as credible evidence. While these fundamental issues have not been resolved, a rich discourse from these opposing viewpoints has played a principle role in shaping practice outside of mainstream public social policy and programme evaluation.

Constituent Literature 2: Evaluation of Conflict Prevention, Peacebuilding and Humanitarian Assistance Programmes

The current section shifts its focus to the second constituent literature within which this chapter, and this book, is rooted: peacebuilding evaluation. This includes both the evaluation of self-labelled *peacebuilding projects* as well as the evaluation of the peacebuilding effects of development and humanitarian initiatives. The influence of fundamental issues in mainstream social programme, including the international

development programme evaluation literature, is also evident in the literature on conflict prevention, peacebuilding and humanitarian programme evaluation. The peacebuilding and humanitarian evaluation sub-field has wrestled with, and learned from, the approaches, methods and debates discussed in the previous section. One of the main points of departure for this subset of literature is the distinctive violent and fragile settings in which these programmes are embedded.

Following the end of the Cold War, development and humanitarian agencies found themselves working in conflict (and *post*-conflict) zones, through levels of violence which would have spurred them to cease programming in the past. Within these contexts, a new kind of programming evolved which sought explicitly to support or build the social, economic and political structures, processes, and capacities for sustainable peace. It quickly became clear that the conventional approaches applied to the evaluation of development programmes would need to be revised, and adapted, before they could be effectively applied to initiatives in VDS settings.

At the same time, a related set of evaluation challenges began to loom large within the same violence or conflict-prone environments: consideration of the ways in which development, humanitarian and even peacebuilding interventions may positively or negatively affect the dynamics and structures of peace and conflict (Anderson, 1999, 2004; Bush, 1998, 2001, 2005). As noted by the Organisation for Economic Cooperation and Development (OECD) in its recently released guidelines for the evaluation of peacebuilding activities in settings of conflict and fragility, evaluation practice in these contexts poses real risks to both evaluators and those being evaluated. Understanding and adapting evaluation approaches and methods to violence, while mitigating the risk that evaluation itself might exacerbate violence or cause harm to those involved, forms the foundation upon which both research and evaluation practice should be built. Our analysis of the second constituent literature acknowledges the profound importance of understanding and making allowances for contexts affected by violence and conflict (OECD, 2012; Bush and Duggan, Chapter 1).

Purpose of Evaluation

Peacebuilding evaluation adopts a similar purpose and definition of evaluation as developed in the public programming literature. Evaluations

are heralded as mechanisms for contributing to learning and account-
ability. However, the peacebuilding evaluation literature tends to high-
light process orientation over goal orientation. Church (2011), however,
maintains that in peacebuilding evaluation, the learning purposes can
remain unrealised, with other drivers relating to public relations, fund-
raising, justifying exiting programmes and/or predetermined decisions
taking their place. In practice, peacebuilding evaluation has failed to live
up to the ascribed purposes of learning and accountability. According
to Church, only a few evaluation processes actually catalyse learning,
and these exist outside of the norm (Church, 2011). She notes that there
are various reasons for these learning gaps, but the most common rea-
sons include: a failure to integrate learning as a central pillar within the
organisational culture of most agencies; poorly designed evaluations
which do not support learning; and peacebuilding processes which rarely
capitalise on *process use* learning (Church, 2011). Duggan notes that one
important characteristic of conflict- and violence-affected settings is that
often the social capital of organisations and evaluation stakeholders has
been eroded. For this reason, more serious consideration should be given
to the development of evaluation approaches that embrace rather than
side-step the messiness of high-risk contexts; the purpose of evaluation
should be to revitalise rather than erode the social capital of those being
evaluated (Duggan, 2010).

The concept of accountability in evaluation has yet to be completely
unpacked by the peacebuilding field (Church, 2011; Whitty, Chapter 3).
Instead, accountability continues to be largely understood as upwards in
nature, responding to funders and governments, rather than to the com-
munities being served (Church, 2011). While movement towards other
types of accountability has been made by humanitarian organisations,
until recently, accountability has tended to be a non-issue within peace-
building programming, resulting in a lack of understanding from prac-
titioners (Church, 2011). Church suggests that this failure to fully grasp
the concept of accountability poses a particular challenge for evaluators
attempting to develop systems to contribute to it, not least because of the
moral vacuum it creates for evaluation undertaken within environments
where ethical challenges are considerably more acute—and potentially
lethal—than in normal (non-conflict) conditions (Bush and Duggan,
Chapter 1).

While these challenges are not unique to peacebuilding evaluation,
they are rendered more extreme by the legacy or imminent threat of

violence to evaluation actors. For example, transparency and the provision of programming information are important dimensions of accountability, but these can present potential risks. For example, 'in highly volatile contexts, misused information could spark unrest or lead to local partners being threatened' (Church, 2011, p. 475). Evaluation in VDS must, therefore, not only seek to contribute to accountability and learning, but must also take particular care to not aggravate existing tensions in the process.

Evaluation within these particular contexts must, therefore, be sensitive to the structures and drivers of conflict and violence, and must ask what effect a programme, or an intervention, may have on actions, structures and processes that can support 'prospects for peaceful coexistence and decrease the likelihood of the outbreak, reoccurrence, or continuation, of violent conflict,' or on 'those structures and processes that increase the likelihood that conflict will be dealt with through violent means (Bush, 1998 as cited in USAID, 2008, p. 1). This focus on assessing drivers of violence and enablers of peace is unique to peacebuilding evaluation, a fact which sets it apart from other forms of international development programme evaluation (Chigas, 2014, p. iii).

This concept of *conflict sensitivity* has come to be understood as 'systematically taking into account both the positive and negative effects or impacts of interventions, in terms of conflict or peace dynamics, on the contexts in which they are undertaken, and, conversely, the implications of these contexts for the design and implementation of interventions' (Conflict Sensitivity Consortium, 2004 as cited in OECD, 2012, p. 11). It encapsulates the intricacies of undertaking any kind of intervention in a fragile and/or divided place, encompassing debates around power, gender, vulnerable groups and accountability, among others. This takes evaluation into murky operational territory.

The Role of the Evaluator

The conflict prevention, peacebuilding and humanitarian field evaluation literature also acknowledges the benefits of both formative and summative evaluations to assess the on-going progress as well as final achievements of programmes (Church, 2011). In response, many agencies have invested in in-house monitoring and evaluation expertise, to provide technical guidance to programme staff in charge of monitoring and formative evaluations. In situations which require quick evidence

and feedback, humanitarian agencies often prefer the use of in-house staff to lead *real-time evaluation* (RTE) (see discussion later in the chapter), noting that it averts the lengthy recruitment procedures and start-up costs often associated with external evaluation (Jamal and Crisp, 2002 as quoted in Cosgrave et al., 2009).

The use of external evaluators, however, tends to be the dominant practice within this field, especially within longer term programmes. Since there is no accreditation process to distinguish *conflict/peacebuilding* evaluators, consultants and academics are often selected as evaluators on the basis of their credentials as subject matter experts who provide evaluative judgements by relying on the transference of their social science research skills and on-the-job learning (Church, 2011). Despite the expertise of these evaluators, they often remain outside of the larger professional programme evaluation community, and at times can conflate the role of the evaluator with that of the researcher, producing mini-research studies that are labelled evaluations (Church, 2011). Church highlights that, 'from the perspective of the evaluation discipline, these studies often fall short of the accuracy standard. Common accuracy gaps include a lack of contextual grounding and inadequate description of the evaluation purposes and procedures' (Church, 2011, p. 466). This is problematic since those who commission external evaluations require clarity of purpose in order to use evaluation findings to inform decision-making. Similarly, external evaluation should ensure credible methodological design that yields robust, contextually applicable evidence. Poor quality evaluations have been one impetus that has fuelled the professionalisation debate within this field, as well as strengthened the arguments for accreditation within the larger evaluation community.

While the variety of possible roles described in the first constituent literature on programme evaluation are applicable here, internal and external evaluators assessing conflict prevention, peacebuilding and humanitarian programmes are faced with myriad unique challenges including the high risk of violence, complex institutional contexts, multiple actors with shifting agendas, operational challenges to data collection and the politicisation and manipulation of evaluating findings (Bush and Duggan, Chapter 1). These evaluators are employed with the difficult task of assessing the effects of a programme both in terms of its stated aims or intentions and within the overall environment of fluidity and unpredictability. For this reason, evaluators are expected to possess deep peace and conflict contextual knowledge in addition to the subject

matter and technical expertise required to assess the programme in question. Evaluators must also exhibit an acute degree of conflict sensitivity, to be able to anticipate potential positive and negative programme impacts, while also ensuring that the evaluation process itself has limited effects on the locality and people in which the evaluation is taking place.

As noted earlier, the responsibilities of evaluators are aligned with the growing discussion of conflict sensitivity within the peacebuilding, development and humanitarian fields which recognise that all assistance can, in effect, contribute to, and in many cases aggravate, the particular circumstances of a conflict and even lead to an upsurge in violence. This realisation has led to an expectation that evaluators have an understanding of the symbiotic relationship between the delivery of aid and its potential to do harm. In concrete terms, this translates into competencies that include conflict risk and analysis skills and a concurrent ability to mitigate negative risks and maximise, if possible, positive impacts (USAID, 2008).

Who Participates

Another important fundamental issue within this sub-field of evaluation is once again the question of participation. One of the particularities of peacebuilding and humanitarian evaluation is that stakeholders often come from vulnerable, traumatised, disenfranchised or otherwise disempowered populations. In these circumstances, who should participate, what form that participation should take, and most importantly, who, ultimately, is in control of participation matters immensely. Each of these questions is fraught with ethical, methodological, logistical and political considerations and implications.

Within this sub-field of evaluation, there are a variety of participatory approaches including practical stream approaches and transformative evaluation approaches, which offer potential for re-building and strengthening the social fabric often damaged by multiple forms of violence. For example, empowerment evaluation, a transformative approach stresses the active engagement of a diverse group of stakeholders, especially during decision-making processes (Weaver and Cousins, 2004 as quoted in Cox et al., 2009). Motivation for these approaches has grown as a response to evaluation's predominantly *Northern* roots, which have fostered push-back from Southern (e.g., developing country) and collectivist approaches. These approaches emphasise the need

for a more active role for programme stakeholders, with local participation, capacity building and stakeholder empowerment at the heart of the evaluation process. Some empowerment approaches have even gone so far as to argue that interventions should be measured on their ability to act as a catalyst for social change (Bush, 2003).

Complementary debates have also focused on the need to expand the range of voices involved, and to acknowledge and address the unbalanced power structures in the evaluation process (Cousins and Whitmore, 1998; Chouinard and Cousins, 2009). Donna Mertens notes the importance of looking towards *Southern* methods as a potential way forward. According to Mertens, 'participatory models of evaluation that evolved in Latin America, India, and Africa provide guidance in ways to legitimately involve important stakeholders in an evaluation, especially those stakeholders who have been traditionally excluded from the corridors of power and decision making' (Mertens, 2001, p. 368). However, others argue that the concept of local ownership poses a dilemma on its own, since in some circumstances it has come to be perceived as a tokenistic gesture within peacebuilding and development initiatives (Schmelzle, 2005).

The question of whose voice counts in peacebuilding and humanitarian evaluations has been interrogated in multiple sources over the years. For example, in the Berghoff series, Paffenholz and Reychler ignore the more overtly political implications of this debate by advocating a technocratic approach that argues for the use of standardised procedures developed by Northern consultancy firms (Barbolet et al., 2005; Bush, 2005). Carl, on the other hand, warns against the potential of romanticising the local and indigenous capacities for peacebuilding. 'While these are vitally important, it is often overlooked that traditional capacities for conflict management have failed' (Carl, 2003 as quoted in Schmelze, 2005, p. 6). In the *Utstein report* Dan Smith also notes 'that in the context of violent conflict, local ownership becomes a more complex concept and needs to be handled with care. Local ownership can unintentionally come to mean ownership by conflict parties, or by the most powerful sectors of society' (Smith, 2004 as quoted in Schmelze, 2005, p. 6).

In evaluation practice, this once again raises the political implications of methodological choice, and whether linear models for programme management and evaluation, such as the logical framework approach, should uncritically remain the dominant tool-of-the-trade in peacebuilding, humanitarian and development practice. NGOs and

practitioners, in particular, have been experimenting with the application of systems thinking to peacebuilding evaluation practice by development of practical tools that enable practitioners 'to exploit the insights of systems thinking, while avoiding the more arcane and complex methodological elaborations of the field' (Woodrow and Chigas, 2011, p. 205). Similarly, scholars have questioned the plausibility and utility of using standardised indicators across violence-affected contexts,[2] recognising that no two are the same and that external validity, thus, remains one of the principal challenges faced by evaluators for measuring results such as outputs, outcomes and impacts.

While each of these debates occupies a prominent and important place within this constituent literature, the realities on the ground often influence the degree to which participatory approaches can be used in VDS contexts. For example, the OECD (2012, p. 46) notes that

> Donors generally carry out their conflict prevention, peacebuilding and statebuilding actions in support of and in partnership with host governments. A logical extension of such cooperation is working together in evaluation. Such partnerships, however, may pose challenges where governments lack legitimacy or are primary actors in an ongoing conflict. The political context and its high stakes not only affect external partners, they are also likely to have very real impacts on how and why partners engage in an evaluation process.

In the light of the above, the OECD guidelines on evaluating peacebuilding activities highlight the critical importance of appropriately discussing and managing stakeholder participation for each evaluation context. Difficulties surrounding collaborative management are also apparent when working with programme beneficiaries or local community members. Church (2011) maintains that 'the current reality is that this constituency [programme beneficiaries] is called upon to be sources of information only. True participatory evaluation, where participants play a key role in every stage of the process, is exceedingly rare' (Church, 2011, p. 465). She notes that evaluation findings in peacebuilding programmes are often not even shared with this stakeholder group (Church, 2011). It would seem, therefore, that in spite of the rich discourse about the potential of participatory evaluation approaches within peacebuilding and humanitarian assistance programming, the participation of beneficiaries remains a largely theoretical discussion, with less practical experience currently contributing to the discussions.

What Counts as Evidence of Influence and Impact: Grappling with Volatility, Unpredictability and Risk

Within this sub-field, the primary question that tends to drive calls for *evidence-based* programming or policy-making (how do we know we are making a difference?) is accompanied by a second, equally pressing question: 'How do we know we are not causing harm or deepening conflicts and divisions?'

While this sub-field of evaluation faces many of the same obstacles faced in the evaluation of more traditional social policy or development programmes, VDS contexts generate several unique challenges. The attribution problem (ascribing a causal link from a particular programme to observed or expected social change on the ground) is more acute due to the fluidity and unpredictability of these settings and the frequently non-linear, multidirectional social change processes that take place within complex conflict systems. In addition, the presence or threat of violence does not only have implications for the safety of evaluation actors, it also poses thorny methodological challenges including the absence of baseline data, erratic access to data and stakeholders, and risks that data may be biased, incomplete and/or (voluntarily or involuntarily) censored (OECD, 2012, pp. 32–33). The often politicised nature of these settings also opens possibilities for strategic misinformation and the instrumentalisation of evaluation findings.

In the light of the above, there has been an increasing recognition that context is much more than a landscape or backdrop for evaluation; it must be the starting point of any evaluation (OECD, 2012, p. 34). In recent years, evaluation actors in this sub-field have been grappling with and reflecting upon the implications that contexts of violence and conflict will have upon evaluation practice. As a result, in its guidelines, the OECD recommends that conflict analysis should be built into the evaluation process noting that it 'may be used as the basis for assessing whether activities have been sufficiently sensitive to conflict settings, determining…what will be the evaluation questions…[and ensuring] that the evaluation itself is conducted in a conflict-sensitive way' (OECD, 2012, p. 35).

How this might be done in practice remains unclear. There is a conceptual and operational gap between what peace and conflict scholars and practitioners call conflict analysis (and the barrage of tools and frameworks articulated to categorise conflict and fragility), what humanitarian actors understand as 'do no harm' and what evaluation actors

understand as contextual or situational analysis. In the world of international assistance, the last two decades has seen the rise of multiple tools and frameworks for analysing conflict dynamics. In general terms, these tools propose different methods to undertake a more 'systematic study of the profile, causes, actors, and dynamics of conflict' (International Alert et al., 2004) and exist to help practitioners better understand how particular causes and drivers of conflict—political, military, cultural, economic, etc.—might affect, either negatively or positively (but most often the former), their particular development, humanitarian or peacebuilding programme.

At the same time, there have also been concrete efforts to operationalise the principal of non-maleficence—or 'do no harm' in international aid. As noted, the Collaborative for Development Action's (CDA) 'do no harm' framework is one of the most widely used tools for conflict sensitivity in development and humanitarian practice and aims to assist organisations in identifying 'the conflict and peace potential of their programmes' (Anderson, 1999). The framework does this by identifying operational programme components which may affect a conflict and also puts forwards several agency behaviours that can reinforce conflictive or non-conflictive relations (as noted in OECD, 2012). The 'do no harm' framework has also played a pivotal role in expounding the nuances of conflict evaluation, serving as the catalyst for further discourse and the creation of diverse conflict evaluation tools.

The underlying concepts and motivations of conflict evaluation or analysis tools and the 'do no harm' framework resonate with the 'peace and conflict impact assessment' (PCIA) approach that was developed by Kenneth Bush (1998). PCIA focuses on assessing the actual or potential effects of initiatives on peace and conflict dynamics before, during and after their implementation. Proponents of PCIA see it as a process of mutual learning that can benefit local people living in conflict zones, and aid agencies who work in these settings (Anderson, 1999).

Moving these frameworks into peacebuilding evaluation and evaluation in contexts of violence and fragility has not been unproblematic. In a learning meeting organised by the OECD in 2011, practitioners, evaluators, commissioners and aid managers discussed the strengths and limitations of the OECD guidance. Evaluators cited a number of challenges in trying to incorporate conflict analysis into evaluations, including: lack of time or expertise to do a full socio-political, military, etc., analysis of the environment in which the programme was taking place;

and difficulty in moving from conflict analysis to evaluation framework questions (OECD, 2011), among others.

The difficulties encountered in efforts to incorporate conflict analysis frameworks into the evaluation process spring from two realities. First, existing conflict analysis tools have been developed to assist the analysis of aid practitioners or programme staff who may also be evaluation commissioners and managers—but not evaluators; and second, conflict analysis, PCIA and 'do no harm' frameworks focus upon how an intervention (project, programme) is affected by conflict, or how the intervention itself might affect peace and conflict dynamics (by either *doing harm* or *doing good*) in a given environment. In other words, the unit of analysis is the project or programme intervention—not the evaluation process which itself is also an intervention. Despite the expectation that evaluators should do or obtain a conflict analysis (OECD, 2012, p. 42), these frameworks are not meant to be used by evaluators whose main concern is how to determine which dimensions of violence or conflict will be relevant in influencing, either positively or negatively, prospects for planning and conducting *an evaluation*, and for articulating evaluative judgements that are underpinned by sound principles of ethical practice. While there is renewed interest among evaluators to incorporate a more explicit focus on context into evaluation inquiry,[3] evaluators disaggregate context in ways that are different from those of evaluation commissioners and managers. While conflict analysis can help evaluators navigate particular challenges that bubble up from conditions of violence, instability and tension, analysis around the drivers and causes of conflict must be part of, but cannot replace, good situational analysis for evaluation.

In reflecting upon how violence embeds unique challenges into the contexts in which evaluation must take place, the discussion that follows of approaches and methods will focus on how evaluation actors in this sub-field have been dealing with operational and logistical constraints and with the implications of dealing with non-linearity and complexity.

Approaches and Methodological Practice: Dealing with Non-linearity, Complexity and Emergence

The debates over qualitative or quantitative methodologies have also found a place in the conflict prevention, peacebuilding and humanitarian evaluation literature. However, due to the complexity and heterogeneity

inherent in these contexts, the field has widely acknowledged that there is no single methodological blueprint for evaluation. Instead, evaluation practitioners reinforce the utility of qualitative and quantitative methods noting that 'the complex nature of interventions in fragile and conflict-affected situations generally makes it necessary to combine different methodologies in order to answer the evaluation questions. Many favour a mixed-method approach, using both qualitative and quantitative methods and data' (OECD, 2012, p. 50).

Church and Rogers' (2006) *Designing for Results* proposes a list of 14 considerations that assist in filtering suitable methodologies for specific environments. While the list is applicable for any programme evaluation situation, it highlights the nuances needed for conflict-affected settings. For example, the first consideration assesses a programme's level of complexity, highlighting that qualitative methods are often better suited to deal with the intricacies of increasingly complex environments (Church and Rogers, 2006). Another consideration examines the operational constraints of accessing data. For example if the respondents (or data sources) are in the bush fighting a war, direct observation or interviews may be required. In contrast, if respondents are located in city centres with access to the Internet, online questionnaires might prove to be the most efficient method (Church and Rogers, 2006).

In this sub-field of evaluation, methodological choice also needs to take into account issues of volatility. Church notes that in certain situations respondents are unable to articulate dissenting opinions without great personal risk. In these cases, methods that ensure the anonymity of sources should be selected over others.

> Any method that requires experiences to be documented, through the use of participant diaries or photographs, for example, deserves extra consideration in these contexts. If discovered by the wrong people, such as a paramilitary group or the army, these participants and sources might be in danger. (Church and Rogers, 2006, p. 213)

The examples put forward by Church allude to the diversity of settings that are all classified as situations of conflict or violence but which require drastically different evaluation methodologies. In order to ensure that appropriate methods are selected, many agencies and organisations operating in these settings have opted to produce their own evaluation publications to assist staff and consultants. As a result, there has been a proliferation of agency handbooks for practitioners and policy-makers responsible for managing or conducting evaluations.

Similar guides can also be found in the mainstream programme evaluation literature. One of the most notable examples that emerges from international development evaluation practice is Bamberger, Rugh and Mabry's *Real World Evaluation* (RWE) guide which was designed to assist evaluators who are confronted by budget, time and data constraints, as well as receiving pressure from government agencies, politicians, funding or regulatory agencies and stakeholders. RWE does not present new data collection or analysis methods, but rather provides a guide that evaluators can use to draw from a wide range of mixed methods and approaches to address the common evaluation constraints of the *real world* (Bamberger et al., 2006).

Although the RWE approach was not specifically designed for VDS settings, most evaluations required in the conflict prevention, peacebuilding and humanitarian fields are confronted by the constraints that RWE addresses. For example, RWE tackles the baseline challenge by providing measures that help reconstruct baseline data by using 'secondary data sources, recall, key informants, focus groups, construct mapping and participatory group techniques' (Bamberger et al., 2006, p. 4). Similarly, RWE suggests several steps for addressing time and budget constraints.

As evaluation actors come to grips with the need to adopt pragmatic approaches, aid agencies such as the UK's Department for International Development (DFID) have noted that evaluators need to be especially attentive to addressing methodological bias and unreliable data, particularly in situations of conflict where disinformation may be a tactic used by the combatants (DFID, 2010, p. 12). DFID's cautionary advice speaks to the fragility of environments affected by conflict and violence and the additional care needed for applying any evaluative approach. The non-linear, dynamic and uncontrollable nature of many of these settings has pushed the evaluation community to look for alternative approaches that offer more flexibility and potential for tailoring and adaptation, as programmes evolve in response to contexts that are in flux.

In spite of the particular contextual complexities that characterise the evaluation of programmes in this sub-field, the dominant practice has been to adopt linear evaluation approaches that are often nested within complex conflict systems and which are subject to disruptions, shifts and change. Reina Neufeldt (2007, p. 8) maintains that the dominance of rigid, linear evaluation approaches is in part due to

> an optimism that is built into the belief that when we identify objective measures, this will lead to universal patterns and lessons to improve

> our work in the future not only in one locale, but in many locales—
> an optimism that, just as we grow crops or markets, we can grow
> more peace.

In her 2011 publication, *'Frameworkers' and 'Circlers': Exploring Assumptions in Impact Assessment,* Neufeldt further examines this inherent optimism, along with the ontological orientations of peacebuilding practitioners, partitioning them into two archetypal camps. The *Logical Frameworkers* group which adheres to a more positivist approach, believing that programme design and evaluation should follow a linear, cause–effect trajectory. The nature of a programme's relationships are frequently visualised and tracked through detailed logical framework matrices. Frameworkers, therefore, believe that evidence should demonstrate 'the degree to which particular activities and their outputs contribute to larger or higher-order objectives and goals' and thus, 'indicators for activities, outputs, results or objectives are to be "SMART"', meaning specific, measurable, achievable, relevant and time-bound (Rouche, 1999 as quoted by Neufeldt, 2007, p. 3).

In contrast, Neufeldt's *Complex Circlers* group adopts a decidedly constructivist angle. Individuals in this group are characterised by their preference to work within a systems or complexity approach. This group approaches peacebuilding through a more elliptical lens; they are relationship focussed and have a desire to be flexible and responsive to each situation. Circlers generally do not believe that events in conflict environments can be predicted because they are part of a larger complex system made up of intermeshed forces over which peacebuilders or development agents have little (if any) influence. The assumptions that underpin this approach include the belief that every situation is unique, lessons are not transferable from one country or setting to another, planning has limitations and flexibility is always an asset.

In line with this thinking, there is a growing movement in the conflict prevention, peacebuilding and humanitarian sub-field towards the use of systems-based approaches in evaluation. While the move away from linear approaches is part of a larger trend in programme evaluation (particularly international development programme evaluation), evaluation practitioners in this sub-field have also recognised the analytical utility of systems thinking and complexity science for understanding developments and relationships in fast changing and unpredictable environments. This awareness is a prerequisite for ensuring that programmes are better able to adapt, and thrive, within conflict contexts (Ramalingam and Jones, 2008).

The discussion below provides a glimpse of some of the evaluation approaches that are anchored in systems and complexity thinking. These approaches are not necessarily unique to peacebuilding and humanitarian evaluation, nor do they serve as a comprehensive summary. Rather, what is presented is an acknowledgment of the range and diversity of approaches available and currently in use in this sub-field.

Real-time Evaluation

RTE was created in response to the particular needs of humanitarian assistance programming and is regarded as one of the most demanding types of evaluative practice. The primary objective of RTE is to 'provide feedback in a participatory way in "real time" (i.e., during the evaluation fieldwork) to those executing or managing a humanitarian response' (Cosgrave et al., 2009, p. 10). RTEs are most effective when they are conducted during the early stages of a humanitarian of conflict programme, as they are designed and executed to construct knowledge that can assist in the ongoing operational decision-making and programme adaptation. The primary audience for RTE is the staff of the implementing agency (at the various field, national, regional and global levels). One unique component of RTE is the rapid timeframe. Evaluation teams, often composed of one to four members, are expected to conduct 'light evaluation exercises' and deliver their report, or at least a substantive draft, before departing the field (Cosgrave et al., 2009). Evaluation practitioner Cosgrave notes that 'RTEs are well suited to the fast pace of decision making within the humanitarian sector, where they can bring a strong strategic focus at a critical stage of the response' (Cosgrave et al., 2009, p. 12).

Developmental Evaluation

While not specifically designed for application in conflict-affected settings, Michael Quinn Patton's DE is an approach that speaks directly to the complexity of conflict environments. The OECD Development Assistance Committee's (DAC) 2012 guidelines state that, 'few would dispute that settings of conflict and fragility are complex, combining high levels of unpredictability, a general lack of information, and potential strategic misinformation' (OECD, 2012, p. 32). Complexity is defined here, as contexts in which the relationships of cause and effect are fundamentally unknown, or only known in retrospect (Patton, 2010).

DE is informed by systems thinking, and addresses non-linear dynamics, enabling innovation and adaptive management (Patton, 2010). Patton specifies five contexts for DE: ongoing development and adaptation; pre-formative evaluation to support exploration and innovation; supporting local adaptation of general principles to navigate top-down and bottom-up forces for change; evaluating major systems change and evaluating in turbulent, disaster situations (Patton, 2011). While each of these situations could be present in conflict contexts, the last speaks directly to this field. Patton highlights that in turbulent and disaster situations, 'planned interventions must adapt and respond as conditions change suddenly' and, as a result, 'planning, execution and evaluation occur simultaneously' (Patton, 2011, p. 12).

DE was, thus, developed as an alternative to formative and summative evaluation, with the evaluator's primary function being to 'infuse team discussions with evaluative questions, thinking, and data, and to facilitate systematic data-based reflection and decision making in the developmental process' (Patton, 2010, p. 1). The DE evaluator is, therefore, embedded within the programme team helping to adapt programme strategies based on the data available and the changing circumstances.

Unlike other forms of evaluation, DE is not focused on accountability, but instead concentrates on adaptive learning. In this way the evaluator plays an active role in shaping the programme's development, while at the same time capturing innovative strategies and ideas. Although DE is a relatively new arrival in the evaluation community, it has been met with great enthusiasm. Patton (2011) is quick to caution, however, that DE is not meant for programmes that have a model, and are attempting to improve it, but rather for initiatives where there is no model in place.

Outcome Mapping

A final approach that has come to be applied to evaluation in complex settings is outcome mapping (OM)—although its origins and orientation were never peacebuilding or conflict-specific.[4] Rooted in contribution analysis thinking, OM is an integrated planning, monitoring and evaluation approach. It takes a learning-based and user-driven view of evaluation and is guided by principles of participation and iterative learning.

At the root of OM, is the notion of identifying strategic actors within a programme's sphere of influence, so that programme actors can credit their *contributions* to social change outcomes, rather than crediting the entire outcome to a single programme in isolation from the rest of

the system. Results are consequently calculated by an analysis of 'the changes in behaviour, actions and relationships of those individuals, groups or organisations with whom the initiative is working directly and seeking to influence' (Smutylo, 2005 as quoted in Jones and Hearn, 2009, p. 1). Within the conflict prevention, peacebuilding and humanitarian sub-field, this means that peacebuilding or conflict prevention programmes can accurately acknowledge if their interventions contributed to resolving or preventing conflicts, rather than claiming that their programmes were solely responsible for the end of all violence.

Although OM has been applied in a multitude of contexts, it was originally born as a response to researchers' frustrations over inadequate ways of measuring the reach of research effects beyond a programme, particularly within development research programmes. The history, approaches and considerations of research evaluation are further explored in the next section, which focuses on this specialised sub-field of evaluation.

Constituent Literature 3:
The Evaluation of Research

Constituent literature 3 moves us away from the field of programme and policy evaluation and delves into questions that relate to the evaluation of research, which itself can be considered a specialised sub-field of evaluation. This section presents an overview of the literature on research evaluation and hones in on some of the unique challenges that evaluators of research face.

Evaluation of research has gained new importance in the last decade due to the increased pressure on governments and other funders of public research to allocate funds to researchers and institutions who produce high quality research. Much of the literature in this section stems from debates in the research council funding community in England, Australia and USA, and the controversies that have developed as they modify their research evaluation schemes. However, research evaluation is neither a new topic nor is it exclusive to funding councils and academia. According to Marjanovic, Hanney and Wooding, research evaluation blossomed during the 1960s and 1970s when several studies, particularly in USA, sought to understand, mostly through case studies, if and how research influenced innovation (Marjanovic et al., 2009).

Similarly, while much of the literature on research evaluation concerns academia or university-based research in the Global North and South, there are extra-academic research centres, think tanks, governmental and non-government research institutions that are also interested in research evaluation. These institutions often fund and/or conduct research that is meant to solve societal problems and the outputs may or may not be publishable articles that can be subjected to the traditional research evaluation mechanisms of peer review and bibliometric analysis (see discussion below for more detail on these mechanisms). A research-derived policy brief, for example, may be high quality but may not be considered under the traditional mechanisms of research evaluation. Similar to researchers conducting applied research within academia, these researchers and institutions have also raised concerns about the effectiveness of current research evaluation practices.

Before delving into the fundamental issues that surround research evaluation, two issues merit clarification and discussion. First, it is important to distinguish between the terms 'research' and 'evaluation'. As noted by Bush and Duggan in the introductory chapter of this book, evaluation is generally understood to refer to 'the process of collecting and analysing information in order to judge value, worth or impact' (Butcher and Yaron, 2006, p. 5). *Research* on the other hand has, as Mertens notes, been defined as, 'a systematic method of knowledge construction' (Mertens, 2009, p. 1). As noted by Patton (n.d.),

> because research is driven by the agenda of knowledge production, the standards for evidence are higher, and the time lines for generating knowledge can be longer. In evaluation, there are very concrete deadlines for when decisions have to get made, for when action has to be taken. It often means that the levels of evidence involve less certainty than they would under a research approach and that the time lines are much shorter.

The boundaries between research and evaluation can easily become ambiguous as evaluation may contribute to knowledge construction, and research may employ evaluative logic or analysis. However, while evaluation employs many of the same research techniques to gather empirical information, it is not accorded the status of research (Scriven, 2003, p. 7).

Second, and of equal importance, if we are to broaden our understanding of the value of research within and beyond the university research community, we need to understand how the intersection

Box 2.3
Research Approaches

> *Multi-disciplinary research:* refers to researchers from a variety of disciplines working together at some point during a project, but having separate questions, separate conclusions, and disseminating in different journals.
>
> *Inter-disciplinary research:* refers to researchers interacting with the goal of transferring knowledge from one discipline to another; allowing researchers to inform each other's work and compare individual findings.
>
> *Trans-disciplinary research:* refers to collaborative research in which exchanging altering discipline-specific approaches, sharing resources and integrating disciplines achieves a common scientific goal.

Source: Singh et al. (2013) citing TREC Centers at the Washington University School of Medicine at St. Louis.

of disciplines—that is multi-, inter- or trans-disciplinary research approaches (see Box 2.3)—influences approaches to research evaluation. The standards to which research will be held are influenced by the disciplinary bent of the researchers involved and the process by which a research interacts with and crosses disciplinary boundaries.

The literature discussed in this section focuses broadly on the evaluation of research, with no particular focus on research in VDS. In fact, review of the literature revealed that there is almost no literature about research on or in VDS informing the debates about research evaluation.

Purpose of Evaluation

Excellence in research is desirable in any type of research. However, if the purpose of social change research is to go beyond mere knowledge generation to the generation of knowledge that can be used to improve social and economic outcomes, one could argue that the stakes are higher when findings are meant to influence decisions that affect people's lives, the environment, governance and other areas of development (Méndez, 2012). Research findings gain credibility and chances for their uptake and use tend to increase if they derive from excellent research. Research evaluation, like programme evaluation, can help answer fundamental questions: Is the research effective? (e.g., is it being used for harm or for good?) Linked to this question is the principal question and rationale for research evaluation: How do we know the good research from the bad?

The purpose for evaluating research can vary depending on who is commissioning the evaluation. A funding agency may have different reasons for evaluating than a university research department or a local policy-maker who uses research findings to inform the creation of new policies and programmes. Roughly speaking, the literature groups the rationale for evaluating research into four main categories. The first, according to Marjanovic et al., is 'to increase the accountability of researchers, policy-makers and funding bodies, by making the research process, its outputs and impacts more transparent' (Marjanovic et al., 2009, p. 6). Similar to the evaluation of non-research projects, an evaluation process of this nature seeks to ensure that the researchers have done what they intended to do and that funds have been adequately spent.

Marjanovic et al. also explain that evaluation can help 'steer the research process towards desired outcomes' (2009, p. 6). This implies the type of learning that is usually associated with formative evaluations, where the research project is evaluated prior to its culmination and where the learning can lead to improvements, perhaps by adjusting research design, methods, mechanisms of analysis or management.

The third rationale also considers learning, albeit the learning occurs once a research project has concluded and lessons can be drawn to inform strategic or managerial decisions about future research. For example, evaluation findings about the relevance of a research topic can inform questions, hypotheses or even new lines of research that the institution may want to pursue in the future. Similarly, findings about the effectiveness or efficiency of the research process can inform managerial decisions.

The fourth and last suggested rationale has to do with evaluations that are conducted to prove research process or research output quality in order to advocate for the research team. This is the case of evaluations that are used to inform funding decisions. The evaluation signals the ability of a team to conduct good quality research, which determines future levels of funding (Marjanovic et al., 2009).

While these are the main recurrent research evaluation purposes, review of the literature also suggests others. Research funders across the board—public and private—are often interested (for different reasons) in evaluating the research they support to understand whether they are getting *value* for the money invested. This line of motivation is frequently pursued through the conduct of socioeconomic impact evaluation, a specific approach to evaluation that focuses on calculating,

through the use of economic valuation methods, the 'social return on investment' (SROI) of interventions (including research). While SROI attempts to quantify and monetise the socioeconomic and environmental benefits of research are in many ways problematic and controversial, funder demands for *value for money* are pushing research evaluators to innovate and constantly question and improve practice.[5]

Review boards at peer-review journals evaluate research with the purpose of determining whether it meets the publication standards of a particular field, while there are others, like Van Raan who holds that the purpose of evaluating research is to 'to promote research quality' (Van Raan, 1996, p. 398). This preoccupation gives rise to an additional purpose for evaluating research, one that harnesses considerations of programme evaluation to research evaluation. As noted by Knox in this book, research is one of the fundamental building blocks of programme theory. Donors of peacebuilding and development projects and programmes use research to form the basis of the change theories that underlie the initiatives they fund. Evaluating the quality of that research—its relevance, methodological soundness, scientific merit, etc.—is essential for challenging assumptions and questioning the foundations upon which international aid initiatives are built.

Evaluation Approaches

Similar to programme evaluation, approaches in research evaluation depend largely on the purpose of the evaluation and the epistemological orientation of the evaluation commissioner and the evaluator. However, unlike programme evaluation, there are particularities about research that make selecting an approach more cumbersome. Deciding what approach to use when evaluating research is perhaps the most contested aspect of research evaluation. One reason for this difficulty is that there are various types of research and each one requires a nuanced approach to evaluation.

In 1997, Donald Stokes put forward the proposition that research can be classified along two dimensions: whether it advances human knowledge by seeking a fundamental understanding of nature; or whether it is primarily motivated by the need to solve immediate problems (see Figure 2.2; Ofir and Schwandt, 2012).

Stokes divides research into three distinct classes:

1.   Pure basic research (exemplified by the work of Niels Bohr, the early 20th century atomic physicist).

Figure 2.2
Research Classification
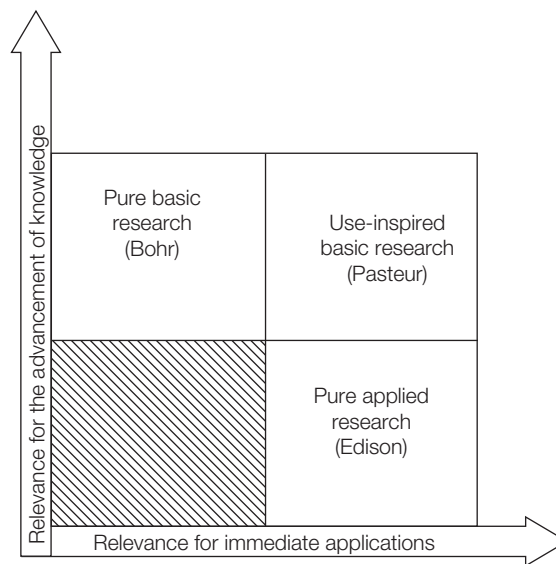
Source: Ofir and Schwandt (2012).

2. Pure applied research (exemplified by the work of Thomas Edison, inventor)
3. Use-inspired research (exemplified by the work of Louis Pasteur, pioneering 19th century chemist and microbiologist) (Ofir and Schwandt, 2012).

When determining the evaluation approach, the evaluator needs to consider whether it is applied or basic research that is under review since the approach used to evaluate applied research may consider the practical applications or use of that research, which may not be a consideration when evaluating basic research (Furlong and Oancea, 2005). Use-inspired applied research may require even further considerations.

The concern about research types extends to revolutionary research, which is generally understood to be research that moves beyond well-established fields and that is trying to break new ground by challenging basic assumptions or involving a paradigm shift (Andras, 2011). A similar concern emerges when evaluating interdisciplinary research. Most disciplines have their own standards for research quality. RCTs, for example, are considered the gold standard in healthcare research.

However, their value in the social sciences is increasingly questioned, among other reasons, because of the assumption that two groups living in a complex social environment can have parallel trends. Boaz and Ashby explain this problem from an epistemological perspective: 'methodological debates in the natural sciences focus on the quest for "truth" and the elimination of bias. In the social sciences the existence of objective truth is often contested, while bias is often an accepted dimension of knowledge, to be acknowledged rather than eliminated' (Boaz and Ashby, 2003, p. 9).

The unit of analysis is another consideration that may influence the evaluation approach. The first two constituent literatures examined included a clear and identifiable unit of analysis: The programme or project. Literature on research evaluation, however, identifies several potential units of analysis, among which are research outputs (paper, journal article, book, policy brief, etc.); individual researchers or research teams; scientific laboratories and institutions such as universities; a scientific discipline; government departments and funding agencies; a country's entire research base, etc.

The unit of analysis is important in determining the evaluation approach because it will determine what should and should not be considered in the evaluation. Take for example the new mechanism for research evaluation that the Higher Education Funding Council for England started to use in 2014, the Research Excellence Framework (REF). The REF considers departments within higher-level education institutions as the unit of analysis. The evaluation framework considers research outputs and impacts but given that the unit of analysis is research departments where researchers also have teaching and grant writing responsibilities, the framework must also consider things like doctoral degrees awarded, research income and research income in kind. Considerations of this nature have been grouped under a third aspect of the framework labelled 'research environment'. In a context where evaluation is conducted to inform funding, such as with the REF, such considerations become critical since omission of the research environment may incentivise researchers to focus strictly on research, therefore potentially compromising their teaching duties.

Methodological Practice

Although there are a variety of methods used to evaluate research, peer-review processes and bibliometric analysis are by far the most used and also the most controversial. Much of the discussion in this section will

focus on those two methods with the recognition that due to their many shortcomings, several research evaluation exercises have used other methods such a self-evaluation, historical ratings, benchmarking, case studies, mixed approaches and even computerised semantic analysis. This wealth of different and sometimes new approaches suggests that research evaluators recognise the flaws of peer review and bibliometrics and are attempting to overcome them by complimenting them with creative and new methods. Be this as it may, peer review and biblometric analysis continue to dominate evaluation research practice.

Wooding and Grant (2003, p. 20) define peer review as

> a system in which experts make a professional judgment on the performance of individuals or groups, over a specified cycle, and/or their likely performance in the future. The groups could be research groups, departments or consortia. Assessment may be undertaken entirely by peers or may incorporate other experts such as representatives of user groups, lay people, and financial experts.

Peer review is often tied to publication since it is peer-review boards at journals and other publishing enterprises that determine what is publishable and what is not. Although peer review is the most commonly used method to evaluate research, a study that involved 142 research stakeholders in England noted that it also has the 'least good features and most bad features' in comparison to the other methods (Wooding and Grant, 2003, p. 25). Since peer review is based on people's opinion, it is criticised for its subjectivity and for the unhealthy competition that it generates among peers. Furthermore, since the review is often solely of research outputs, an evaluation based on peer or expert review often misses the research process, therefore missing potential learning opportunities.

Additional criticism comes from certain types of research, such as interdisciplinary and revolutionary research, which face several challenges including finding the right expertise for the review panels and overcoming the conservative tendencies of influential peers. Peer review can be time and resource intensive; it is sometimes unable to detect fraud and does not guarantee relevance to policy-making or compliance with methodological standards.

Despite the flaws, peer review continues to be the method of choice. After reviewing the benefits and challenges of peer review and other research evaluation methods, the 142 research stakeholders involved in

the study cited earlier were asked to gather in groups to design their ideal method. Despite the many flaws they cited, 22 out of the 29 groups based their ideal system on an expert review process.

Bibliometrics, or arithmetical calculations of various publication rates, are the other dominant method used in research evaluation processes. Bibliometrics consist of

> publication count in a defined list of venues (journals, conference proceedings) by individuals, departments or universities, the citation count of these publications over a defined period of time, indexes calculated using these counts (e.g., h-index), metrics derived from citations and authorship graphs, and market share measures. (Andras, 2011, p. 90)

Bibliometrics offer various advantages over peer review: They are considered more objective, less burdensome and cheaper indicators of research quality.

However, despite the alleged benefits, even supporters agree that bibliometrics have many shortcomings and should not be used as sole indicators of research quality. The main problem with this type of measure is their validity, for they are not intrinsic indicators of quality but proxies that are generated after research outputs have gone through a peer-review mechanism and are then published. As Boaz and Ashby explain, 'it is a faulty assumption that all research that is published in journals or cited by others is accurate, reliable, valid, free of bias, non-fraudulent, or of sufficient quality' (Boaz and Ashby, 2003, p. 2).

Bibliometrics are also criticised because they generate negative incentives for researchers, especially in academia, to focus on publishing rather than on teaching or on producing outputs that may be more relevant to policy or other vehicles for social change. In addition, there is a strong bias favouring research produced in English or dealing with topics that are of interest to review boards in Northern-led, mostly English language journals. Researchers from the Global South are, thus, at a disadvantage because the research questions that are relevant in their context, in their language or the type of research output format they favour (i.e., policy or practice brief) may not be considered in the publication venues that calculate bibliometrics. This may generate the erroneous impression that research is of low quality when in reality, low citation or publication rates may be only a reflection of different languages, research priorities or indeed the purpose of the research in question (Tijssen et al., 2006).

Something similar occurs for researchers who are working on innovative or revolutionary research or science. This group may sometimes

find it hard to get their research published because their ideas or findings may not interest journal review boards since they are considered unorthodox or ahead of their time. Publication or citation rates of these researchers may, therefore, not reflect the novelty, quality or importance of their research. Young researchers must also compete with more experienced researchers whose name may carry more weight in publication spheres, despite the quality of the research under consideration.

Another area of improvement for bibliometrics is in their uneven coverage of different disciplines. Coryn explains how the arts are at particular disadvantage since their work is usually expressed in forms, such as recital or painting, that are not publishable or citable (Coryn, 2006). Negative citations are also an issue with bibliometrics, given that a research paper may be heavily cited because of its many shortcomings. The same occurs with measures of paper downloads (as an indicator) since a paper may be deemed low quality after it is downloaded.

At the heart of the critiques that have been levelled against both peer review and bibliometrics is the concern that both of these mechanisms focus on one specific type of output: scholarly publications in journals with high impact factors. Implicitly, such an approach treats outputs as outcomes. The *success* of research is measured by publication within international peer-reviewed journals. This is taken to constitute evidence of both international esteem and good research practice. Additionally, it is assumed to constitute effective research communication and dissemination. In contrast to the evaluation practice under constituent literatures 1 and 2, there is little discussion of tangible outcomes in the way that programme evaluation would demand. In a nutshell, the main limitation of both of these methods is that they are primarily interested in the contribution that research makes to the advancement of scholarly knowledge—researchers influencing researchers.

This reality is particularly important for research for international development which may not be generated for the purposes of academic or scholarly advancement. This would include research for influencing policy and practice, undertaken by the broad research and policy-shaping communities that include think tanks, not-for-profit firms, and governmental, non-governmental and inter-governmental actors.

What Counts as Evidence?

Evaluating research excellence is vitally important in both academic and non-academic spheres. It is a simple fact that not all research is of

the same quality. Without rigorous examination of research quality, a false equivalence is made between all research endeavours, leading to a poor understanding of the phenomena involved. When it comes to evidence based policy-making or decision-making, the inability to distinguish between good and bad research can lead to counter-productive policies and ineffective solutions.

It is necessary to evaluate research quality for any policy research/academic research organisation. That is because assessment of the quality of research can provide a very useful indication of whether the organisation is fulfilling its purpose or not, and what role is being played by research work of the organisation in this regard.

The two comments above, offered by two researchers located in the Global South under the umbrella of a study on perceptions of research excellence (Singh et al., 2013, pp. 11–12), demonstrate not only why considerations of quality and use are important, but also that *excellence* acquires meaning only in context. Defining quality or excellence in research is perhaps the most contentious fundamental issue in determining a research evaluation approach. As mentioned earlier, different disciplines have different quality standards, so it is a challenge to attempt to come up with a universal definition of research quality or excellence. In fact, these two terms, which are often used interchangeably, mean different things to different people. A useful distinction is provided by Grant, Brutscher, Kirk, Butler and Wooding who, referring to the REF, suggest that the distinction between quality and excellence is that excellence is a broader dimension comprising both quality and research impacts (Grant et al., 2010). However, this distinction in terms does not answer the questions of what is meant by research quality or research impacts. The following sections will address some of those questions.

Most definitions of research quality consider the methodological and technical aspects of research. For example, quality research has been described as comprising

the scientific process encompassing all aspects of study design; in particular, it pertains to the judgment regarding the match between the methods and questions, selection of subjects, measurement of outcomes, and protection against systematic bias, non-systematic bias, and inferential error. (National Center for the Dissemination of Disability Research, 2005, p. 2)

However, a review of over 30 sets of research quality standards suggests that quality criteria can be grouped conceptually under five categories:

scientific merit, ethics, originality, relevance and purposivity. Under this classification, scientific merit comprises different forms for validity, soundness and rigour of methods and a logical interpretation of data (Méndez, 2012).

The ethical dimension of research, which is discussed by Jayawickrama and Strecker in this book and which is an area of particular importance in VDS, stems primarily from the principle of do no harm (as discussed under constituent literature 2). However, it is also related to the tension between the transparency required for accountability purposes and the need to ensure the safety and security of stakeholders. So too is it nested in the political sensitivities around the evaluation process (especially collaborative processes), as well as the volatile environment within which it is undertaken.

In a study with over 250 research stakeholders, it was found that originality as an element of research quality 'revolved around viewing existing issues, ideas or data in a new or different way, more so than generating new data or novel methods. Originality involved, also, the development of new theoretical and practical insights and concepts' (Becker et al., 2006, p. 12).

Relevance in research, as described by Boaz and Ashby refers to 'the extent to which the research addresses the needs of key stakeholders' (Boaz and Ashby, 2003, p. 12). Unlike impact, which occurs once the research is carried out, research relevance can be determined beforehand by looking at the research questions and determining whether it is significant for the discipline or for society.

Finally, purposivity speaks to the quality of research that Aagaard-Hansen and Svedin associate with having a well-formulated research question that is supported by relevant literature and that guides the research to meet a specific purpose (Aagaard-Hansen and Svedin, 2009).

Evidence of Impacts

In the introductory chapter of this book, Bush and Duggan emphasise that one of the key messages of this publication is 'the importance of appropriate approaches to, and methods for, evaluating the effects or "impacts" of research' (Bush and Duggan, Chapter 1). Research evaluation has indeed attempted to address the role of impact in determining research excellence. While there are no definite answers, much of the recent literature on research impacts has attempted to address three main questions: What is meant by research impacts? How do we measure

research impacts? And, should research impacts be considered an element of research excellence?

While there is no clear consensus of what constitutes *research impact*, a number of researchers have explored this area in some depth. In a project launched by the London School of Economics (LSE) in 2011, research impact is understood to be 'an occasion of influence and hence … not the same thing as a change in outputs or activities as a result of that influence, still less a change in social outcomes' (LSE Public Policy Group, 2011, p. 21).

In 2003, Nutley, Percy-Smith and Solesbury conducted a thorough literature review on research impact and distinguished between the conceptual and instrumental use of research, wherein conceptual use consists of 'changes in levels of understanding, knowledge, and attitude' and instrumental use includes 'results in changes in practice and policy making' (Nutley et al., 2003, p. 11). Similarly, the LSE project mentioned above makes a distinction between academic and external impacts. Academic impacts refer to instances when research influences actors in academia or universities and external impacts occur when research influences actors outside of academia, such as businesses, governments or civil society. Donovan refers to these distinctions as academic and extra-academic impacts (Donovan, 2008).

However, even if an evaluation clearly specifies the meaning and parameters of research impact, the path of research influence is typically erratic and unpredictable, making measurement and attribution of impacts a challenge. As discussed above, a few bibliometric measures, such as the journal impact factor and citation counts, have been used as indicators of impact. Non-bibliometric mechanisms have also been used, such as the case study approach used by the UK RAND/Arthritis Research Campaign Impact scoring system (RAISS), the self-evaluation approach used by the US Programme Assessment Rating Tool (Grant et al., 2010), and SROI approaches, as noted earlier. However, these too face the challenge of attributing causation.

Finally, there are authors who challenge the idea that research impacts should be considered an element of research excellence. Such criticism has emerged from circumstances in the UK where research evaluation is conducted with the purpose to inform future funding. Opponents argue that the focus on impacts that is promoted by such an *investment model* overlooks the primary value of research to produce new knowledge. That, in turn, can generate incentives that, according to Hammersley, may 'be at odds with any commitment to research informing public

discussion of policy issues, since it frames inquiry within assumptions about predictable payoff' (Hammersley, 2008, p. 753).

In recent years, evidence of impact, as a fundamental issue in research evaluation, has expanded and come to focus more pointedly on research effectiveness, a notion that includes research use, influence and impact. The contribution that research can make to solving problems at a societal level is increasingly being questioned with a greater importance placed on the role of research within a knowledge-based economy as well as on the more overt demands from government and other funders for scientific contributions to the development and progression of society. As a result, there is increasing emphasis being placed on economic and social impacts, as well as on how funded research meets the needs of policy-makers and the broader population of stakeholders. Nowhere is this more evident than in the evaluation literature on policy-oriented research (see Boaz and Ashby, 2003; Carden, 2009; ODI, 2004).

The introduction of knowledge-for-action theories such as knowledge utilisation, implementation, diffusion, transfer or translation into evaluation practice has assisted research evaluators in developing more nuanced approaches to tracing impact trajectories and understanding how knowledge contributes to social change, including policy change (Ottoson, 2009). This evolution in evaluation practice reflects a growing recognition, among those who produce and use research, that research effectiveness should include some social change intent, highlighting the need for more integration between research uptake, policy and (presumably) practice.[6] Researchers are being asked more explicitly to demonstrate the relevance and utility of their work to solving *real-world problems*.

Towards a Literature on the Evaluation of Research in and on Violently Divided Societies

Figure 2.1, in the introductory section of this chapter, identifies the constituent literatures within which the current book is nested. The chapters which follow represent contributions to the development of a body of literature which is still at the earliest stages of formation. The diversity of themes, types of research, perspectives and cases in the following chapters are an accurate reflection of the heterogeneity and complexity of the sub-field we seek to explore. The current chapter is meant to demonstrate that, while we may be entering uncharted territory (Bush and Duggan,

Chapter 1), we are not without the tools and sensibilities that have
been cultivated over the last 70 years of evaluation research and practice.

Notes

1. John A. and John R. Healy examine in further detail the implications that epistemology
   has on research funder evaluation practices in Chapter 8 of this book.
2. See http://www.berghof-handbook.net, 2005 for a discussion of standardisation.
3. See Conner et al. (2012) framework for context assessment in evaluation.
4. The OM manual is freely available online. A virtual collective of evaluation
   practitioners contributes and document the growth of OM, highlights synergies with
   other evaluation approaches and explores the applicability of OM in new fields;
   www.outcomemapping.ca
5. Unfortunately, a more complete examination of SRIO approaches to research
   performance and evaluation is beyond the scope of the current chapter.
6. Chapter 6 (Kevin Kelly) in this book includes an empiric case study which employs
   the influence frameworks emerging from the policy research literature.

References

Aagaard-Hansen, J. and Svedin, U. (2009). Quality issues in cross-disciplinary research:
    Towards a two-pronged approach to evaluation. *Social Epistemolog*y, *23*(2), 165–176.
Alkin, Marvin C. (1972, December). Wider context goals and goals-based evaluators.
    Evaluation Comment: The Journal of Educational Evaluation, *3*(4), 10–11.
American Evaluation Association. (2003). American Evaluation Association position on
    'Scientifically based evaluation methods'. Retrieved from www.eval.org (accessed
    on 1 June 2015).
Anderson, M. B. (1999). *Do no harm: How aid can support peace—or war*. Boulder, CO:
    Lynne Rienner.
Anderson, M. (2004). *Experiences with impact assessment: Can we know what good we
    do?* Berlin: Berghof Research Centre for Conflict Management.
Andras, P. (2011). Research: Metrics, quality, and management implications. *Research
    Evaluati*on, *20*(2), 90–106.
Bamberger, M., Rugh, J. and Mabry, L. (2006). *Real world evaluation: Working under
    budget, time, data, and political constraints*. Thousand Oaks, CA: SAGE Publications.
Barbolet, A., Goldwyn, R., Groenewald, H. and Sherriff, A. (2005). The utility and
    dilemmas of conflict sensitivity. In B. Bloomfield, M. Fischer and B. Schmelzle
    (Eds), *Dialogue series no 4—New trends in peace and conflict impact assessment
    (PCIA)*. Berlin: Berghof Research Center for Constructive Conflict Management.
Becker, S., Bryman, A. and Sempik, J. (2006). *Defining 'quality' in social policy research:
    Views, perceptions and a framework for discussion*. Lavenham, UK: Social Policy
    Association.
Befani, B. Barnett, C. and Stern, E. (2014). Introduction—Rethinking impact evaluation.
    IDS Bulletin, *45*(6). Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/1759-
    5436.12108/epdf (accessed on 2 February 2015).

Boaz, A. and Ashby, D. (2003). *Fit for purpose? Assessing research quality for evidence based policy and practice*. ESRC UK Centre for Evidence Based Policy and Practice, Working Paper 11. Retrieved from https://www.kcl.ac.uk/sspp/departments/ politicaleconomy/research/cep/pubs/papers/assets/wp11.pdf (accessed on 20 July 2015).

Braskamp, L. A., Brandenburg, D. C. and Ory, J. C. (1987). Lessons about clients' expectations. In J. Nowakowski (Ed.), *The client perspective on evaluation: New directions for program evaluation* (No. 36). San Francisco: Jossey-Bass.

Bush, K. (1998). A measure of peace: Peace and conflict impact assessment of development projects in zones of conflict. *Working Paper No. 1; The Peacebuilding and Reconstruction Program Initiative and The Evaluation Unit O*ttawa: *International Development Research Centre*. Retrieved from http://web.idrc.ca/uploads/ user-S/10533919790A_Measure_of_Peace.pdf (accessed on 22 April 2015).

———. (2001). Peace and conflict impact assessment (PCIA) five years on: The commodi- fication of an idea. In A. Austin., M. Fischer and O. Wils (Eds), *Dialogue Series No. 1—Peace and conflict impact assessment*. Berlin: Berghof Research Center for Constructive Conflict Management. Retrieved from http://www.berghof-handbook. net/documents/publications/dialogue1_bush.pdf (accessed on 27 April 2015).

———. (2003). PCIA five years on: The commodification of an idea. In Alexander Austin, Martina Fischer and Oliver Wils (Eds), *Berghof Handbook Dialogue Series No. 1: Peace and conflict impact assessment. Critical views on theory and practice* (pp. 37–51). Berlin: Berghof Centre for Constructive Conflict Management.

———. (2005). Field notes—Fighting commodification and disempowerment in the development industry: Things I learned about peace and conflict impact assessment (PCIA) in Habarana and Mindanao. In D. Bloomfield, M. Fischer and B. Schmelzle (Eds), *Dialogue series no. 4: New trends in peace and conflict impact assessment (PCIA)*. Berlin: Berghof Research Center for Constructive Conflict Management. Retrieved from http://www.berghof-handbook.net/documents/publications/dialogue4 _bush.pdf (accessed on 22 April 2015).

Butcher, C. and Yaron, G. (2006). *Scoping study: Monitoring and evaluation of research communications*. Background paper produced for the Communications M&E Group Workshop, 5–6 September 2006, London.

Carden, F. (2009). *Knowledge to policy: Making the most of development research.* Ottawa and Delhi: International Development Research Centre and SAGE India. Retrieved from http://web.idrc.ca/openebooks/417-8/ (accessed on 27 April 2015).

Chambers, R. (2010). *Paradigms, poverty and adaptive pluralism* (IDS Working Paper 344). Brighton: The Institute of Development Studies, University of Sussex. Retrieved from http://www.ids.ac.uk/files/dmfile/Wp344.pdf (accessed on 22 April 2015).

Chigas, D, Church, M. and Corlazzoli, V. (2014). *Evaluating impact of peacebuilding interventions: Approaches and methods, challenges and considerations* (DFID Com- missioned Report). London: Department of International Development. Retrieved from http://www.cdacollaborative.org/publications/reflecting-on-peace-practice/rpp- guidance-materials/evaluating-impacts-of-peacebuilding-interventions-approaches- and-methods,-challenges-and-considerations/ (accessed on 20 June 2015).

Chouinard, J. and Cousins, J. B. (2009). A review and synthesis of current research on cross-cultural evaluation. *American Journal of Evaluation*, *30*(4), 457–494.

Christie C. A. (Ed.). (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. In *The practice-theory relationship in evaluation: New directions for evaluation* (pp. 7–36). San Francisco, CA: Jossey-Bass.

Christie, C. A. and Fleischer, D. N. (2009). Social inquiry paradigms as a frame for the debate on credible evidence. In S. I. Donaldson, C. A. Christie and M. M. Mark (Eds), *What counts as credible evidence in applied research and evaluation practice?* (pp. 19–30). Thousand Oaks, CA: SAGE Publications.

Church, C. (2011). Evaluating peacebuilding: Not yet all it could be. *Berghof Series.* Retrieved from http://www.berghof-handbook.net/documents/publications/ scharbatke_church_handbook.pdf (accessed on 23 April 2015).

Church, C. and Rogers, M. M. (Eds). (2006). Ethics in evaluation. *Designing for results: Integrating, monitoring & evaluation in conflict transformation programs* (pp. 188–200). Washington, DC: Search for Common Ground. Retrieved from http:// www.sfcg.org/Documents/dmechapter11.pdf (accessed on 22 April 2015).

Clinton, T., Nunes-Neto, B. and Williams, E. (2006). *Congress and program evaluation: An overview of randomized control trials (RCTs) and related issues*. Congressional Research Service, Library of Congress. Hauppauge, New York: Nova Science Publishers.

Conflict Sensitivity Consortium. (2004). An introduction to conflict-sensitive approaches to development, humanitarian assistance and peacebuilding tools for peace and conflict impact assessment (Chapter 2). *Resource Pack*. Retrieved from http://www. conflictsensitivity.org/publications/conflict-sensitive-approaches-development- humanitarian-assistance-and-peacebuilding-res (accessed on 20 June 2015).

Conley-Tyler, M. (2005, March/April). A fundamental choice: Internal or external evaluation. *Evaluation Journal of Australasia*, *4* (new series, 1 and 2), 3–11.

Cook, T., Scriven, M., Coryn, C. and Evergreen, S. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation*, *31*(1), 105–117.

Coryn, C. L., Westine, C. and Schroter, D. (2011). A systematic review of theory- driven evaluation practice from 1990 to 2009. *American Journal of Evaluation*, *32*(2), 199–226.

Coryn, C. L. S. (2006). The use and abuse of citations as indicators of research quality. *Journal of MultiDisciplinary Evaluation*, *3*(4), 115–121.

Cosgrave, J., Ramalingam, B. and Beck, T. (2009). *Real-time evaluations of humanitarian action. An ALNAP guide*. London: Overseas Development Institute. Retrieved from http://www.alnap.org/pool/files/rteguide.pdf (accessed on 5 June 2015).

Cousins J. B. (2011, November 23). *Participatory evaluation: Theory, research, practice.* Presentation at IDRC, New Delhi.

Cousins J. B. and Whitmore E. (1998). Framing participatory evaluation. *New Directions for Evaluation, 80*(Winter), 5–23.

Cox, P. J., Keener, D., Woodard, T. and Wandersman, A. (2009). *Evaluation for improvement—A seven-step empowerment evaluation approach for violence prevention organisations*. Atlanta, GA: Centres for Disease Control and Prevention. Retrieved from http://www.cdc.gov/violenceprevention/pdf/evaluation_ improvement-a.pdf (accessed on 22 April 2015).

Datta, L. (2001). Mixed methods evaluation: The wheelbarrow, the mosaic and the double helix. *Evaluation Journal of Australasia*, *1*(2), 33–40.

Davidson, J. (2000). Ascertaining causality in theory-based evaluation. *New Directions for Evaluation, 200*(87), 17–26.

Donaldson, S. I. (2009a). A practitioner's guide for gathering credible evidence in the evidence-based global society. In S. I. Donaldson, C. A. Christie and M. M. Mark

(Eds), *What counts as credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: SAGE Publications.

Donaldson, S. I. (2009b). In search of the blueprint for an evidence-based global society. In S. I. Donaldson, C. A. Christie and M. M. Mark (Eds), *What counts as credible evidence in applied research and evaluation practice?* (pp. 2–18). Thousand Oaks, CA: SAGE Publications.

Donovan, C. (2008). The Australian research quality framework: A live experiment in capturing the social, economic, environmental, and cultural returns of publicly funded research. *New Directions for Evaluation*, *2008*(118), 47–60.

Drewett, A. (1997). Evaluation and consultation. learning lessons. *Evaluation 3*(2), 189–204. Retrieved from http://evi.sagepub.com/content/3/2.toc (accessed on 22 April 2015).

Duggan, C. (2010). Guest editorial.Special issue: Transitional justice on trial: Evaluating its impact. *International Journal of Transitional Justice. 4*(3), 315–328.

Fitzpatrick, J. (2004). Exemplars as case studies: Reflections on the links between theory, practice and context. *American Journal of Evaluation*, *25*(4), 541–59.

Furlong, J. and Oancea, A. (2005). *Assessing quality in applied and practice-based ducational research*. Oxford: Oxford University Department of Educational Studies.

Georghiou, L. and Laredo, P. (2005, September). *Evaluation of publicly funded research*: Report on the Berlin workshop, 26–27 September 2005. Paris: OECD.

Grant, J., Brutscher, P. B., Guthrie, S., Butler, L. and Wooding, S. (2010). *Capturing research impacts: A review of international practice.* Santa Monica, CA: RAND Corporation.

Greene, J. (1997). Advancing mixed-method evaluation. *The Evaluation Exchange*, *3*(1). Retrieved from http://www.hfrp.org/evaluation/the-evaluation-exchange/issue-archive/methodology/advancing-mixed-method-evaluation (accessed on 20 June 2015).

Guba, E. G. and Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park: SAGE Publications.

Hammersley, M. (2008). Troubling criteria: A critical commentary on Furlong and Oancea's framework for assessing educational research. *British Educational Research Journal*, *34*(6), 747–762.

House, E. R. (1980). *Evaluating with validity.* Beverly Hills, CA: SAGE Publications.

International Alert et al. (2004). *Conflict-sensitive approaches to development, humanitarian assistance, and peacebuilding*. London: International Alert.

Jones, H. and Hearn, S. (2009). Outcome mapping: a realistic alternative for planning, monitoring and evaluation *ODI Background Notes*. London: Overseas Development Institute.

Love, A. (1991). Internal evaluation: Building organisations from within. *Applied Social Research Methods Series* (Vol. 24). Newbury Park, CA: SAGE Publications.

Luukkonen, T. (2002). Research evaluation in Europe: State of the art. *Research Evaluation*, *11*(2), 81–84.

LSE Public Policy Group. (2011). *Maximizing the impacts of your research: A handbook for social scientists.* London: LSE Public Policy Group.

Marjanovic, S., Hanney, S. and Wooding, S. (2009). *A historical reflection on research evaluation studies, their recurrent themes and challenges* (pp. 7–12)*.* Cambridge: RAND Corporation.

Mark, M. M. (2003). Toward an integrative view of the theory and practice of program and policy evaluation. In S. I. Donaldson and M. Scriven (Eds), *Evaluating social*

programs and problems: Visions for the new millennium (pp. 109–141). Mahwah, NJ: Erlbaum.

Mathison, S. (2009). Seeing is believing: The credibility of image-based research and evaluation. In S. I. Donaldson, C. A. Christie and M. M. Mark (Eds), *What counts as credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: SAGE Publications.

McCoy, M. and Hargie, O. D. W. (2001). Evaluating evaluation: Implications for assessing quality. *International Journal of Health Care Quality Assurance*, *14*(7), 317–327.

Méndez, E. (2012). *What's in 'Good'?* Ottawa: International Development Research Centre. Retrieved from http://www.idrc.ca/EN/Documents/Lit-review-Final-English. pdf (accessed on 7 April 2015).

Mertens, D. (2001). Inclusivity and transformation: Evaluation in 2010. *American Journal of Evaluation*, *22*(3), 367–374.

———. (2009). *Transformative research and evaluation*. New York: Guildford Press.

Molund, S. and Schill, G. (2004). *Looking back, moving forward: Evaluation manual*. Stockholm: SIDA Publications. Retrieved from http://www.oecd.org/development/evaluation/dcdndep/47470919.pdf (accessed on 27 April 2015).

National Center for the Dissemination of Disability Research. (2005). *What are the standards for quality research?* (Focus: Technical Brief Number 9). Retrieved from http://www.ncddr.org/kt/products/focus/focus9/ (accessed on 23 April 2015).

Nutley, S., Percy-Smith, J. and Solesbury, W. (2003). *Models of research impact: A cross-sector review of literature and practice*. London: Learning and Skills Research Centre.

Overseas Development Institute (ODI). (2009). *Research and policy in development (RAPID): Toolkits*. Retrieved from http://www.odi.org.uk/RAPID/Tools/Toolkits/index.html (accessed on 22 April 2015).

OECD. (2011). *Evaluating conflict prevention and peacebuilding activities. Lessons from the application phase. Chairman's summary*.Paris: OECD. Retrieved from http://www.oecd.org/dac/evaluationofdevelopmentprogrammes/dcdndep/48092571.pdf (accessed on 22 April 2015).

———. (2012). *DAC guidelines and references series: Evaluating peacebuilding activities in settings of conflict and fragility—Improving learning for results*. Paris: OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264106802-en (accessed on 23 April 2015).

Ofir, Z. and Schwandt, T. (2012). *Understanding research excellence at IDRC*. Ottawa: International Development Research Centre.

Ottoson, J. (2009). Knowledge-for-action theories in evaluation: Knowledge utilization, diffusion, implementation, transfer and translation. *New Directions in Evaluation 2009* (124), 7–20.

Overseas Development Institute, Research and Policy in Development (RAPID) programme. (2004, October). *Bridging research and policy in international development: An analytical and practical framework* (Briefing Paper). London: ODI. Retrieved from http://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/198.pdf (accessed on 17 April 2015).

Parlett, M. and Hamilton, D. (1977). Evaluation as illumination. In M. Parlett and G. Dearden (Eds), *Introduction to illuminative evaluation: Studies in higher education*. Columbus, Ohio: Ohio State University Press.

Patton, M. Q. (1997). *Utilization focused evaluation: The new century text* (3rd ed.). London: SAGE Publications.

Patton, M. Q. (2001). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

———. (2008). *Utilization focused evaluation* (4th ed.). Thousand Oaks, CA: SAGE Publications.

———. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use.* NY: New York: Guildford Press.

———. (2012). *Essentials of utilization-focused evaluation*. Thousand Oaks, CA: SAGE Publications.

———. (n.d.). In conversation: Michael Quinn Patton. Retrieved from http://www.idrc. ca/en/themes/evaluation/pages/articledetails.aspx?publicationid=771 (accessed on 5 June 2015).

Podems. (2005). *Nonprofit evaluation in South Africa: A study of relationships between the donor and nonprofit organisation in the developing World* (PhD Dissertation). Union Institute at University, Cincinnati, OH.

Preskill, H. and Russ-Eft, D. (2005). *Building evaluation capacity: 72 activities for teaching and training.* Thousand Oaks, CA: SAGE Publications.

Rallis, S. F. (2009). Reasoning with rigor and probity: Ethical premises for credible evidence. In S. I. Donaldson, C. A. Christie and M. M. Mark (Eds), *What counts as credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: SAGE Publications.

Ramalingam B. and Jones, H. (2008). *Exploring the science of complexity: Ideas and implications for development and humanitarian efforts* (Working Paper 285). London: Overseas Development Institute. Retrieved from http://www.odi.org.uk/ sites/odi.org.uk/files/odi-assets/publications-opinion-files/833.pdf (accessed on 22 April 2015).

Rogers, P. J., Petrosino, A., Huebner, T. A. and Hacsi, T. A. (2000). Program theory evaluation: Practice, promise, and problems. *New Directions for Evaluation* (No. 87). San Francisco, CA: Jossey-Bass.

Rossi, P. and Freeman, H. (1989). *Evaluation: A systematic approach* (4th ed.). Newbury Park, CA: SAGE Publications.

Rossi, P. H. and Lipsey, M. W. (2004). *Evaluation: A systematic approach.* London: SAGE Publications.

Rutnik, T. and Campbell, M. (2002). *When and how to use external evaluators* (Briefing Paper I: Monitoring and Evaluation). Baltimore, MD: Association of Baltimore Grantmakers.

Savedoff, William D., Levine, R., Birdsall, N. et al. (2006). *When will we ever learn? Improving lives through impact evaluation.* Washington, DC: Centre for Global Development. Retrieved from http://www.cgdev.org/files/7973_file_ WillWeEverLearn.pdf (accessed on 23 April 2015).

Schmelzle, B. (2005). New trends in peace and conflict impact assessment: introduction. In *Dialogue Series No. 4: Berghof handbook for conflict transformation*. Berlin: Berghof Research Center for Constructive Conflict Management. Retrieved from http://www.berghof-foundation.org/fileadmin/redaktion/Publications/Handbook/ Dialogue_Chapters/dialogue4_pcianew_intro.pdf (accessed on 23 April 2015).

Scriven, M. (1972). Prose and cons about goal-free evaluation. *Evaluation Comment*, *3*(4), 1–7.

———. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: SAGE Publications.

———. (2003). Michael Scriven on the differences between evaluation and social science research. *The Evaluation Exchange*, *10*(Winter), 7.

Singh, S., Dubey, P., Rastogi, A. and Vail, D. (2013). Excellence in the context of use-inspired research: Perspectives of the Global South. Ottawa: International Development Research Centre. Retrieved from http://www.idrc.ca/EN/Documents/Perspectives-of-the-global-south-Full-paper.pdf (accessed on 16 January 2015).

Smith, D. (2004). *Towards a strategic framework for peacebuilding: Getting their act together: Overview report of the joint Utstein study of peacebuilding* (Evaluation Report 1/2004). Oslo: Royal Ministry of Foreign Affairs. Retrieved from https://www.regjeringen.no/globalassets/upload/kilde/ud/rap/2004/0044/ddd/pdfv/210673-rapp104.pdf (accessed on 23 April 2015).

Smith, N. and Brandon, P. (2008). *Fundamental issues in evaluation*. New York, NY: Guilford Press.

Smutylo, T. (2005). *Outcome mapping: A method for tracking behavioural changes in development programs* (ILAC Brief 7). Retrieved from http://issat.dcaf.ch/ser/content/download/29484/414272/file/Smutylo_Outcome%20Mapping.pdf (accessed on 23 April 2015).

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R. and Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations* (DFID Working Paper 38). London: Department for International Development.

Stufflebeam, D. L. (1994). Empowerment evaluation, objectivist evaluation and evaluation standards: Where the future of evaluation should not go and where it needs to go. *Evaluation Practice*, *15*, 321–338.

———. (1999). Foundational models for 21st century program evaluation. In *Social Agenda/Advocacy Approaches. Occasional Paper Series.* The Evaluation Center, Western Michigan University.

Stufflebeam, D. L. and Shinkfield, A. J. (1985). *Systematic evaluation.* New York: Kluwer-Nijhoff Publishing.

Tijssen, R. J. W., Mouton, J., van Leeuwen, T. N. and Boshoff, N. (2006). How relevant are local scholarly journals in global science? A case study of South Africa. *Research Evaluation*, *15*(3), 163–174.

USAID (Gunduz, C. and Klein, D.). (2008). *Conflict-sensitive approaches to value chain development*. Washington, DC: USAID. Retrieved from http://pdf.usaid.gov/pdf_docs/PNADY232.pdf (accessed on 12 April 2015).

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, *36*(3), 397–420.

Weiss, C. (1997). Theory-based evaluation: Past, present and future. *New Directions in Evaluation*, *76*, 41–55.

———. (1998). Have we learned anything new about evaluation? *American Journal of Evaluation 19*(1), 21–31.

Whitmore, E. (1988). Participation, empowerment and welfare. *Canadian Review of Social Policy*, *22*, 51–60.

Wooding, S. and Grant, J. (2003). *Assessing research: The researcher's view*. Leiden and Cambridge: RAND Europe. Retrieved from http://www.ra-review.ac.uk/reports/assess/AssessResearchReport.pdf (accessed on 23 April 2015).

Woodrow, P. and Chigas, D. (2011). Connecting the dots: Evaluating whether and how programmes address conflict systems. In N. Ropers et al. (Eds), *The non-linearity of peace processes: Theory and practice of systemic conflict transformation*. Berlin: Berghof Foundation.