ATTRIBUTION AND AID EVALUATION IN INTERNATIONAL DEVELOPMENT: A LITERATURE REVIEW

For

Evaluation Unit
International Development Research Centre

May 2003

Prepared by:

Alex Iverson
9 Humewood Dr., 4(37)
Toronto, Ontario
M6C 1C9
Phone: 416.651.1781
Email: alex.iverson@utoronto.ca

EXECUTIVE SUMMARY

Several objectives guide this literature review. Principally, it aims to shed light on the problems involved in attributing results within aid evaluation research. In doing so, it synthesizes the perspectives of 'frontline' and academic experts working within the field. Drawing on prominent past and contemporary writing, the first part of the review provides a brief, 'purposive' history of evaluation research generally, and 'aid evaluation' specifically. This helps to illustrate the dynamic nature of the discipline and exposes some of evaluations internal frictions, particularly in relation to epistemological and methodological issues. Emphasis is given to demonstrating the evolving character of evaluation; the 'theme of transition' that runs through evaluation's history also runs through the literature. Accordingly, attribution is best understood by looking at how its meaning and significance within evaluation has changed over time.

Following this, an exploration of the links between evaluation research and social scientific research is presented, revealing evaluation's epistemological and methodological dependence on the social sciences. This helps to demonstrate how evaluation has at times learned from the shortcomings of the social sciences, and at time repeated them. A number of dominant research methods and model for attempting to establish causation are reviewed, with emphasis on the experimental and quasi-experimental approaches. In particular, the notion of causation is carefully considered and its problematic nature within social research exposed. It becomes clear that, ‘cause’ and ‘effect’ relationships with social research are, in fact, always correlational, ‘probabilistic’ relationships. Awareness of this fundamental misrepresentation places the attribution question on immediate unstable ground.

Next, a portrait of evaluation research as a dynamic field comprised of diverse practices is presented. To understand the unique issues surrounding the attribution problem, evaluation is disaggregated in terms of sector, and level of intervention and analysis. This reveals the conditions in which determining attribution becomes most complicated, and those conditions in which it is most feasible. It shows how, as sectors become more 'complex' in nature, attribution becomes more difficult. Similarly, as the level of intervention moves from the 'simple' project level to the 'comprehensive' program level, attributing results typically becomes less feasible. Also, attribution is more likely to be established when analyzing at the output, or even outcome level, and, less likely at the impact level. Because of its prevalence within evaluation research, and its significance with respect to the attribution question, the Logical Framework Analysis approach is explored in substantial detail.

Finally, attribution is examined in relation to the 'paradigm shift' that evaluation experienced during the 1980s and into the 1990s. Often referred to as the quantitative/ qualitative debate, this shift represents an epistemological schism within the discipline, and is followed by the emergence of alternative evaluation approaches and research methodologies. Within the field of evaluation, emphasis turned away from measuring and 'proving', and toward understanding and 'improving'. The emergence of ‘participatory’ and ‘action’ oriented approaches, and of evaluation that stresses ‘good

governance' and 'capacity building' reflect a shift in the purpose and practice of evaluation research; and ultimately, in the nature and significance of the attribution question.

Evaluator Thomas Cook reflects on what he has learned during his 25 years practicing evaluation research, and presents the following itemized summary: 'After years of debate, qualitative methods have become accepted'; 'knowledge claims are based on a synthesis of research, not just on one study'; 'contrary to what one might expect, results do not often inform policy decisions, but rather 'enlighten' the situation'; 'evaluation is fragmented by discipline and methodology – this can, and should be, remedied' (Cook, 1997). Consequently, there is no single or simple answer to the attribution question, and the literature on the topic represents a 'work in progress'. To be sure, the ongoing 'burden to demonstrate proof' has helped to secure an important place for attribution within aid evaluation. However, recent innovations in thinking have altered both the meaning and significance of attribution. No longer is determining attribution necessarily dependent on empirical measurement, and no longer is attribution necessarily the principal aim for evaluators interested in 'understanding' how people are changed by complex processes.

TABLE OF CONTENTS

Because they are typically not self-supporting and beholden unto third parties for financial assistance, voluntary organizations face constant pressure to show that they are achieving what they said they would do. They are required continually to justify their existence and to provide a rationale for their work. They carry a heavy and ongoing burden: proof of legitimacy.

– Sherri Torjman, 1999

INTRODUCTION

The field of professional evaluation has undergone remarkable expansion and transformation since its inception in the United States over fifty years ago. Today, international interventions ranging from economic and technical initiatives to social and cultural programs incorporate Monitoring and Evaluation (M&E) into their design.[1] And, while the growth of evaluation has been accompanied by a diversification of its practices and applications, one of its 'primary objectives' has remained essentially the same: *To assure that organizations are accountable to their stakeholders by assessing the degree of success (and/or failure) of a specific intervention whose implementation was sponsored by that organization.* More precisely, evaluation is defined in terms of its systematic approach to assessing the particular effects of a given social intervention:

> Evaluation research is the systematic application of social research procedures in assessing the conceptualization, implementation and utility of social intervention programs (Rossi and Freeman, 1993:5).[2]

From the onset, perhaps the single most important question for policy researchers and evaluators was not *whether* or *what* to assess, but rather *how* to assess the change resultant from specific interventions. How can evaluation research determine if a given intervention 'caused' a particular outcome? How can a specific 'effect' be attributed to a distinct 'cause'? The issue of 'attribution' provides a particularly troublesome thorn in the side of evaluation research, and it has, at times, divided the field and threatened to undermine the discipline (Elzinga, 1981; Patton, 1997; Mark, 2001). Historically, the conflict has proceeded at two interrelated levels: At the epistemological level in which issues of 'causation' are central; and at the methodological level where practicing evaluators dispute the strengths and limitations of different evaluation designs and

[1] Within the literature, *monitoring* and *evaluation* often are used interchangeably; however they are distinctive by purpose and design. Monitoring is usually seen as an ongoing process of data collection, carried out 'in-house', in order to track inputs and outputs, serving the interest and need of the management staff. Evaluations, on the other hand, are typically periodic or single studies, conducted by teams external to the project, which attempt to measure intermediate results and longer-term impacts (Bennendijk, 1990:166).

[2] Innumerable definitions of evaluation exist today. The Human Resources Development Canada (HRDC) defines evaluation as: "[A] collection of methods, skills and sensitivities necessary to determine whether a human service is needed and likely to be used, whether it is conducted as planned, and whether the human service actually does help people" (Posavac and Carey, 1980:6).

approaches. The epistemological and methodological positions, which have provided the basis and rationale for the disunity within evaluation research, are drawn primarily from the social and natural sciences (Rebien, 1996). And, while many of these divisions have endured and are visible within the current 'state of the discipline', it is from this dialectic that important advances have been made. That is, it might be said that evaluation's internal struggle for epistemological verity and methodological rigor has brought about important advancements in its professional knowledge and practices.

The Development Assistance Committee (DAC) Working Group on Aid Evaluation has defined attribution as:

> The ascription of a causal link between observed (or expected to be observed) changes and a specific intervention. Attribution refers to that which is to be credited for the observed changes or results achieved. It represents the extent to which observed development effects can be attributed to a specific intervention or to performance of one or more partner taking account of other interventions, (anticipated or unanticipated) confounding factors, or external shocks (OECD, 2002:17).

Today most evaluators are familiar with the general concerns and considerations involved in attributing change, but arguably fewer possess a comprehensive understanding of the epistemological foundations on which causal claims are made within evaluation research. Satisfied with the soundness of their methods and designs, practicing evaluators may not question the standards of evidence for establishing attribution. But within certain branches of evaluation research, methodological considerations have proven to be more acute – notably, those that deal with comprehensive interventions that are embedded in complex social systems. Perhaps nowhere is this more evident than in the field of international development research wherein the socio-economic, environmental, political and cultural dynamics of 'aid' efforts provide highly unique challenges for evaluators; where change is seldom attributable to any single factor, and can be extremely unpredictable. Moreover, faced with the external pressure to demonstrate results, development organizations and stakeholders are increasingly 'burdened' by having to prove the value of their initiatives to legitimize their work. It is from this burden – from these challenges – that creative alternatives to the traditional modes of conducting evaluation research have evolved.

ORGANIZATION OF THE TEXT

Organized into three (3) sections, the following is a review of the prominent past and contemporary literature related to *attributing results within evaluation research*. The first provides some historical background related to evaluation research generally and international development evaluation specifically, as well as definitions of relevant terms and concepts. This section also establishes the context for the ensuing discussion and is meant to 'ground' the literature review. Part two presents a sketch of the historical issues concerning the idea of causality – from the legacy of early social scientific epistemological discourse, to the more prevalent methodological dialogue generally referred to as the *quantitative-qualitative debate*. Again, this section is meant to provide a sketch of the etiology of the current 'attribution question', and is therefore only an introductory treatment of the topic. The third section directly examines the 'attribution problem' as it has played out in the substantive and theoretical evaluation literature. Perspectives are drawn from the different frontline and academic sources and matched with evaluation research specific to *development assistance.*[3] Specifically, the attribution question is explored in terms of different evaluation research sectors (i.e., 'simple' non-human/non-social systems versus 'complex' human/social systems), and in terms of different levels of intervention and analysis (e.g., 'simple' project versus 'comprehensive' program). Moreover, an account of the evolution of evaluation models and approaches illustrates how evaluators have responded to the problems associated with attempting to determine attribution, exposing the changing shape of evaluation research. To achieve these aims, literature from three interconnected areas is reviewed:

- **Theoretical literature** – predominantly academic writings;
- **Substantive literature** – project results from international development evaluations;
- **'Gray' literature** – government and non-government organizational literature.

[3] *Development assistance* is "a social intervention measure, whether it be aid to a particular project, a sector in a given country or an entire program covering several sectors. Aid interventions are deliberate and intentional attempt on the part of a public or private body - the aid agency to introduce development to a recipient organization, whether the latter be public, private, or a group of individuals" (Rebien, 1996:2).

EVALUATION RESEARCH – BACKGROUND & TERMS

Program evaluation as a distinct field of professional practice was borne of two lessons… First, the realization 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.

– Michael Patton, 1997

While the seeds of evaluation date back to pre-World War I, its professional roots are entrenched in the mid twentieth-century American experiences of social and economic transformation and reform.[4] Within the literature, the portrait of the advent and rise of professional evaluation describes a series of deleterious social and economic circumstances intersecting, and the political resolve to develop programs to address these conditions efficiently and effectively. The social and economic experiences and reverberations of the 1930s Depression saw the commitment of the U.S. federal government to confront and reduce social problems such as poverty, hunger and unemployment. But, early programs were typically without systematic assessment of their effectiveness and efficiency. It was not until the establishment of the 'welfare state' under the Kennedy and Johnson administration that evaluation began to flourish (Bennendijk, 1990; Chelimsky and Shadish, 1997; Patton, 1997; Rebien, 1996); and it was the demand for economic accountability that provided the basis for systematic evaluation:

> It was not until the massive federal expenditures on an awesome assortment of programs during the 1960s and 1970s that accountability began to mean more than assessing staff sincerity or political head counts of opponents and proponents (Patton, 1997:10).

Through the 1970s and 1980s evaluation developed steadily into a professional field, diversifying its practices and expanding its scope, and seeing a miscellany of project and program interventions being evaluated. Correspondingly, the demand for practicing evaluators – and for the professionalization of training and skills – expanded, witnessing the rise of organizations such as the Evaluation Network and the Evaluation Research Society in the 1970s. These two associations amalgamated in 1985 to become the American Evaluation Society (AES), which many considered "the leading international forum for exchange of evaluation theory and methodology between academics, consultants and civil servants…" (Rebien, 1996:12). Throughout the 1980's and 1990s, national and international professional evaluation associations were instituted

[4] It should be noted that it is difficult to discern, from the literature alone, whether the strong emphasis on the American origins and history of evaluation is a product of historical veracity, or due to bias in the literature – i.e., a disproportion of U.S. literature on evaluation.

in countries as far reaching as the United Kingdom, Israel, Ghana and Malaysia.[5] In Canada, the Canadian Evaluation Society (CES) exists as a "non-profit bilingual association dedicated to the advancement of evaluation theory and practice" (CES website, 2002), and is comprised of regional chapters representing over 1,600 members. In addition to regional, national, and international associations, almost all government branches, as well as non-government organizations (such as, the IDRC, United Nations, the World Bank, and the Red Cross) have all incorporated evaluation into their research practices. The legitimization of the discipline is also evidenced in the number of universities that currently offer graduate programs in professional evaluation – 77 masters and doctoral programs offered in American universities alone (Rebien, 1996).

The evolution of evaluation practice has been accompanied by a transformation in its meaning and an expansion of its purpose. Today, hundreds of different kinds of evaluation can be found in the literature (see Patton, 1982), which tend to be grouped into one of two types: "[T]hose that aim to determine if the program has been implemented as planned, and those that measure its success in achieving its objectives (i.e., its impact)" (HRDC, 1998). Respectively, these two types are often labeled 'formative' or process and 'summative' or impact evaluation. However this dichotomy does not reveal the complex nature of the term, therefore evaluation will be disaggregated based on *purpose* and *type*. According to Claus Rebien (1996), the three broad purposes of evaluation are accountability, implementation, and strategy/policy. The importance of **accountability** in evaluation is habitually linked to the cost and expected effects of a given social intervention. Accountability tends to be more critical in evaluations of social interventions that are publicly funded by scarce taxpayers' dollars. More often, accountability "refers to the requirement to show that funds for social interventions have been spent as intended and in ways that produce desirable results" (Rebien, 1996:13-14). At the same time, the concentration and growth of social interventions over the last 50 years has meant increased focus on the evaluation of **implementation**; understanding and improving the implementation process has become a chief goal for evaluation. In addition, by providing critical information to decision makers about the performance or 'effects' of a given social intervention, evaluation serves the purpose of informing **strategic planning and decision-making** (Rebien, 1996:13-14). As Patton explains, "the purpose of applied research and evaluation is to inform action, enhance decision making, and apply knowledge to solve human and societal problems" (Patton, 1990:12).

Rebien also distinguishes between five (5) categories of evaluation each linked to the particular stage of social intervention. They are: Process, effectiveness, monitoring, evaluation synthesis, and meta evaluation. The first type, **process,** often referred to as 'formative evaluation', places less emphasis on determining outcomes or effects, and more on the process involved in generating such outcomes and effects. The second, **effectiveness**, often called 'summative evaluation' or 'impact evaluation', is one of the

[5] For a list of several leading national evaluation associations, see the Canadian Evaluation Society's webpage – http://www.evaluationcanada.ca/site.cgi?section=6&ssection=3&_lang=an

more recognized types of evaluation.[6] Its aim is to "measure the effects of a given intervention and to answer the crucial question of whether the inputs have led to the desired outputs" (Rebien, 1996:14). Third, **monitoring** typically serves the purpose of measuring change by looking at 'performance indicators'. It is ongoing and normally carried out by people directly involved with the intervention. The fourth category is **evaluation synthesis**, which compiles and integrates the results of many related evaluations to draw general conclusions about change. And the fifth, **meta evaluation**, refers to the evaluation of evaluation practices. It is typically carried out by academics and involves the intricate examination of the methods and approaches employed in evaluation so as to determine and improve their methodological and theoretical soundness (Rebien, 1996:14-15). It is important to note that this categorization is meant to illustrate the variety of evaluation types, and that oftentimes a single evaluation may involve more than one emphasis. Therefore, building on other definitions, evaluation is:

> [A]pplied research intentionally designed to assess social interventions, which can serve three distinct purposes: accountability, implementation, and strategy/policy. There are five evaluation subcategories: process-, impact-, monitoring, evaluation synthesis and meta evaluation. Each category is linked to a different phase of an intervention (Rebien, 1996:15).

It is worth mentioning that other conceptualizations of evaluation will be explored in detail in the ensuing sections. Nonetheless, the above definition serves as a useful conceptual guide for the following examination of *attribution within aid evaluation*.

Summary

Emerging out of the need to systematically account for the dramatic increase in public spending on social programs during the early welfare years in the United States, evaluation has undergone an extremely rapid transition toward become a widespread professional discipline. The following historical sketch, as well as the broad definition, disaggregated by evaluation type and purpose, provides an overview of the dynamic character of the discipline. Additionally, it sets a theme that permeates this review – specifically, evaluation as theory and practice continues to evolve in meaning and purpose.

[6] Chris Roche offers the following 'common' definition of impact assessment: "*Impact assessment is the systematic analysis of the lasting or significant changes - positive or negative, intended or not - in people's lives brought about by a given action or series of actions"* (Roche, 1999:21).

AID EVALUATION[7]

Understanding aid evaluation in its present state requires an explanation of its distinct nature, as well as an outline of its historical lineage. Although no single definition will capture all the complexities of aid evaluation, the following developed by the Development Assistance Committee's (DAC) Expert Group on Evaluation (a unit within the OECD) provides a comprehensive characterization. Aid evaluation is defined as:

> [A]n examination as systematic and objective as possible of an on-going or completed project or programme, its design, implementation and results. The aim is to determine the relevance and fulfillment of objectives, developmental efficiency, effectiveness, impact and sustainability. An evaluation should provide information that is credible and useful, enabling the incorporation of lessons learned into the decision-making process of both recipients and donors (OECD, 1992:132).

Aid evaluation typically differs from other types of evaluation in terms of the nature of the 'project or program, its design, implementation and results'. Consequently, the methodologies employed in aid evaluation, as well as the standards of evidence for assessing results, are often distinct.

For organizations involved in development assistance and relying on limited public funds to sponsor and implement research in foreign countries, the incorporation of evaluation has been essential. Aid evaluation can be traced back to the 1950s, however it was not until the 1970s and 80s that systematic evaluation became an integrated component in almost all bilateral and multilateral government and non-government aid agencies (Rebien, 1996). One time chairman of the OECD's Expert Group on Evaluation, Basil Cracknell (2000) offers a brief, four-phase history of the development of aid evaluation:[8]

The **first phase – early developments (from the late 1960s to 1979) –** is characterized by the early implementation of evaluation by US and UN aid organizations. During this phase, two events were of particular significance in the development of aid evaluation: The adoption of evaluation by USAID, and the development of an evaluation guidance manual for the Organization for Economic Cooperation and Development (OECD) by the leading evaluators of the day – Rossi and Freeman. Both helped to

[7] The term 'aid evaluation' will be used throughout this text to refer to all international development initiatives, projects, and programs undergoing evaluation. The Organization for Economic Cooperation and Development's (OECD) Expert Group on Aid Evaluation identifies two specific aims of aid evaluation: "To improve future aid policy, programmes and projects through feedback of lessons learned;" and, "[to] provide a basis for accountability, including the provision of information to the public" (Rebien, 1996:47). Additionally, Michael Bamberger offers the following definition of development programs: "[A]ll social and economic programs in developing countries funded by multilateral and bilateral development agencies or by international non-government organizations (NGOs)" (2000:96).

[8] Cracknell's original history of aid evaluation consisted of three phases; the fourth phase is the result of input from Claus Rebien (1996).

legitimize aid evaluation and provide direction to its professional course. Early aid evaluation was generally carried out by academics working out of universities; it commonly adopted the practices and techniques of 'conventional' evaluation whose ‘tool of choice’ was the Logical Framework Approach (LFA) (Cracknell, 2000).[9] From the onset evaluators were concerned about how to evaluate development assistance given that most donor agencies were only one part of a confluence of aid initiatives. And, questions about whether to study at the project or program level troubled (and continue to trouble) aid evaluators. Additionally, Cracknell explains that during phase one, “evaluation took very much a second place to economic project appraisal, which was seen as crucial to good project selection and formulation” (Cracknell, 2000:43).

The demand for evaluations that accompanied the widespread funding cuts during the late 1970s marks the beginning of the **second phase – explosion of interest (from 1979 to 1984)**. This phase is characterized by a surge of interest in the theory and methods of evaluation, as well as a number of significant transformations in evaluation practices. Lead by the World Bank, aid evaluation began to explore such avenues as, synthesizing evaluations within a given sector, multidonor evaluations, and longitudinal evaluations. Additionally, the World Bank established internal operating guidelines (IBRD *Operations Evaluation – World Bank Standards and Procedure*, 1979) that became the model for other development agencies; at the same time, monitoring and evaluation agencies were being set up in developing countries (Cracknell, 2000).

Phase three – coming of age (from 1984 to 1988) – is characterized by a critical questioning of traditional methods for conducting evaluation. It was during this period that evaluators and stakeholders were beginning to see ‘top-down’ approaches of administering aid as problematic. That is, ‘top-down’ approaches tended to designate a disproportionate authority and control over the evaluation process to the evaluators and their representative organizations, neglecting stakeholders’ input and participation.[10] Consequently, ‘bottom-up’ approaches (such as participatory evaluations) began to emerge during this phase (Patton, 2001). During this period, two landmark studies (Cassen et al., 1986; and, Riddell, 1987) were instrumental in promoting the field of aid evaluation, and “still stand today as the most comprehensive studies of long-term aid effects” (Rebien, 19996:49).

The **fourth phase – aid evaluation at the crossroads (1988 to the present)** – is marked by a shift away from evaluation that emphasized project management and logical frameworks, to evaluation that focused on participatory approaches, accentuating partnership, learning and capacity building. Moreover, drawing on the contributions of Rebien (1996), Cracknell explains the changing character of aid evaluation during this phase:

[9] A full discussion of Logical Frameworks will follow.

[10] Pasteur, for instance, draws attention to the dis-empowering character of the Logical Framework Approach: “[I]t is an imposed procedure, thus maintaining a relationship of control and domination, that does not reflect the… principles of participation and partnership” (Pasteur, 2001:2)

> The whole aid business is changing in significant ways: there are fewer discrete projects now and more emphasis on sectors and programmes and on types of aid that are intrinsically difficult to evaluate such as good governance, community empowerment, poverty alleviation, human rights, etc. (Cracknell, 2000:48).

Notably, this 'fourth phase' may also be understood as a reflection of the professional changes taking place throughout this period within the field of aid evaluation. It represents a departure from the kinds of evaluation practices that stress 'attribution' – whose goals are to determine 'cause' and 'effect' – and which focus on single projects and sectors. To appreciate the implications of these changes, one need only look to the World Bank's report entitled *Assessing Aid: What Works, What Doesn't, and Why?* This comprehensive study by senior research economist David Dollar is grounded in the experiences of 'successful' and 'unsuccessful' developing countries, and provides a penetrating analysis of the effectiveness of international development assistance. Emphasizing institutional 'health' within recipient countries as a precondition for aid to be able to successfully stimulate development, the report lists broad policy reforms that are essential to making aid more effective in reducing poverty. By illuminating the complex relationship between aid and development, the report also exposes the multifarious nature of aid evaluation.

> The effectiveness of finance depends on the quality of all public investments and expenditures, not simply on aid-financed sectors and projects. This finding has important implications for the evaluation and management of aid. Agencies often hone in on the success rate of individual projects as one measure of their effectiveness. At first glance, this appears to be a focus on "quality." But it can lead to distorted incentives, depending on the criteria for judging success. Since money is often fungible, the return to any particular project financed by aid does not reveal the true effect of assistance. Moreover, if agencies are evaluated mainly on the success rate of projects (defined narrowly, without accounting for spillover benefits), managers will avoid risky, innovative projects in favor of things that are known to work. With fungibility, the impact of aid is not the same as the impact of the aid-financed project. The return on the finance depends on the overall effectiveness of public expenditures (World Bank, 1998:20).

Thus, the 'new' aid evaluation emphasizes principles such as 'partnership', 'empowerment' and 'action', but also understands the need to encompass the broad and complex socio-cultural, economic, and political factors that contribute to the success or failure of aid initiatives. As a result, effectively assessing international aid has meant that, "[t]he focus of evaluation has risen above the level of the project to overall country program reviews (World Bank, 1998:20). Consequently, aid evaluation requires new, compatible methodological techniques and procedures – ones able to appropriately address the emerging objectives of modern aid evaluation.

Today, several key objectives guide the evaluation of aid work. In addition to producing data and knowledge that can be used to improve the implementation process, the design of future activities, as well as inform policy and strategic planning, aid evaluation "provides information on the effectiveness and efficiency of aid activities, and thus accountability towards politicians and the public" (Rebien, 1996:4). The importance of accountability – of monitoring funds – was not only significant in the early days of aid work, it has become a driving force behind the implementation and maintenance of aid evaluations today (Rebien, 1996). Still, according to Rebien over its fifty year history, "relatively little attention has been given to the theoretical and methodological aspects of aid evaluation" (1996:5). As a result, there lacks consensus on the best means of obtaining 'information on the effectiveness and efficiency of aid activities'. And, as will be explicated in the following sections, efforts to establish acceptable methods for acquiring information, as well as standards for assessing its validity, have been plagued with controversy.

Moreover, the uniqueness of aid evaluation presents challenges that may not be comparable to other areas of evaluation research. On the one hand, aid evaluators often find themselves working within highly demanding physical environments: scarcity of funding and limited resources, sometimes unfamiliar settings, and potentially 'high-risk' situations contribute to the challenge. At the same time, the distinctiveness of the socio-cultural contexts of international aid evaluation often translates into obstacles for evaluators. For instance, access to and quality of information, cross-cultural ideological differences, as well as general communications are but a few of the conditions that complicate aid evaluation. As Michael Bamberger explains: "[e]valuators of international development programs normally must operate in a very different environment than one would expect to find when evaluating U.S. programs" (Bamberger, 2000:95). And, although development projects have the benefit of relying on 30 years of evaluation wisdom when selecting and implementing methodologies, "the evaluation issues in many of these new international venues are not always the same as those encountered in the past, so that much reinventing of new wheels is required as well" (Chelimsky, 1997:145).

Summary

Understanding aid evaluation requires awareness of its distinct character and history. Aid evaluation can be dated back to the 1950s where it was chiefly practiced by academics. Extensive public funding cuts during the 1970s and into the 1980s saw an expansion of evaluation of internationals aid, and marked the professionalization of aid evaluation. The growth of aid evaluation during the 1980s was accompanied by a critical re-assessment of the theories and methods that dominated evaluation generally; specifically, 'top-down' approaches were recognized by many to be inappropriate for evaluating programs in culturally unique settings. In recent years, aid evaluation has witnessed a shift away from 'top-down' approaches, and has emphasized such ideals as 'partnership', 'empowerment', 'capacity-building', and 'good governance'. It has also begun to deal with the unique challenges of evaluating complex, comprehensive systems. What is therefore important to keep in mind, and will be illustrated in the ensuing, is how aid evaluation often differs from other types of evaluation in terms of the nature of the 'project or program, its design, implementation and results'. As a result, it often employs distinct methodologies and standards of evidence.

CAUSATION – BACKGROUND & TERMS

All philosophers, of every school, imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word 'cause' never occurs ... The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.

– Bertrand Russell, 1913

With the historical background and relevant concepts employed as a backdrop, it is now possible to discuss the specific issue of attributing causes and effects within evaluation research. In doing so, a brief explication of evaluation's social scientific heritage and a general outline of the principles and preliminaries of causation and causal inferences will be provided. This framework will inform remaining sections of the paper and expose the abstruse, and oftentimes controversial, nature of the concepts 'causation' and 'attribution' as they relate to social research generally and evaluation specifically.

CAUSATION & THE SOCIAL SCIENCES

From its inception, social science has been particularly concerned with developing causal explanations of social phenomena. Although the scope of this paper does not warrant an elaboration of the philosophical bases of early social scientific conceptualizations of causation, a provisional account of the social science's epistemological and methodological foundation is necessary. Historically, one of the most complex and important issues within the social sciences has centered on modes of establishing causal relations among different variables. Informed by the natural sciences, early social scientific accounts of causal relations adopted a *positivist* approach. Essentially they maintained that, cause and effect could only be 'determined' scientifically; that it must be informed by theory and based on empirical observation, systematic experimentation, and quantitative (i.e., statistical) analyses (Rebien, 1996; den Heyer, 2001; Patton, 2001). Accordingly, positivist methods and procedures were appropriated by social science and employed to explain phenomena within the social world. Early sociologists were, thus, primarily interested in quantifiable measurement and analysis that, through experimentation, could expose patterns and regularities, leading to general social laws. The quintessential example of early positivist sociology is Emile Durkheim's classic study, *Suicide*, in which he supports his theory that suicide is causally linked to 'social cohesion' by employing rigorous empirical/positivist data gathering methods and statistical analyses.

Unlike the natural sciences in which the subject matter can be isolated and manipulated under controlled conditions, causation within the social sciences has proven to be far more problematic. The principle of 'cause' and 'effect' typically implies a process involving a relationship between one or more independent variables (the cause) and a dependent variable (the effect), and tends to be represented as:

if X then Y (or, X → Y)
if cause then effect (or, cause → effect)

Rooted in the work of philosopher John Stuart Mill, three conditions are generally considered to be 'necessary but not sufficient' for inferring causation. The first condition is that of covariation which stipulates that cause and effect have to be related – that the effect (Y) will be present when the cause (X) is present, and the effect will be absent when the cause is absent. The second involves the principle of temporality and specifies that the cause (X) must precede the effect (Y) in time. And, the third condition requires that the relationship between X and Y be non-spurious, meaning that there is no other variable determining both X and Y concurrently (Mason, 1991; Lofland and Lofland, 1995).[11] So for example, if one were to hypothesize a causal relationship between strenuous physical exercise and level of perspiration, one would first want to ensure that these three conditions were met. Specifically, if perspiration is present then physical exercise must also be present; the exercise must have preceded the perspiration; and, perspiration resulted from exercise and not some other factor (for example, environmental temperature or nervousness).

Today, however, as the notion of causation has become increasingly elaborate and complex, early 'scientific' methods for determining causation are recognizably overly simplistic.[12] Moreover, the belief in causal explanation in evaluation research may be excessively optimistic. In the social world, there are too many influencing variables to be able to "formulate an evaluation project in terms of a series of hypotheses which state that 'activities A, B, C will produce results X, Y, Z'" (House, 2001:311). Moreover,

> We have huge gaps in our knowledge of social events, gaps we don't know about, and gaps we don't even know we don't know about. We can never fill these gaps in so we can never be certain of all that is involved (House, 2001: 312).

Nonetheless, efforts to draw causal links and determine exact project outcomes continue to be pursued within most branches of evaluation research.

Summary

Establishing causation – explaining the ‘causes’ and ‘effects’ of social phenomena – has been central to the social sciences through its history. Modeled after the natural sciences, social scientific research adopted the philosophical assumptions and analytical techniques of *logical positivism*. Insofar as its aim is to determine empirically grounded social patterns and regularities, generalizations and laws, experimentation and quasi-experimentation have been the historical method of choice for positivist social science.

[11] Note that alternative conceptualizations of 'causation' will be explored in subsequent sections (e.g., 'mutual causality', 'interdependence', et cetera).

[12] See Ragin (1997) for a detailed account of the issues associated with causation in the social sciences.

EXPERIMENTATION & QUASI-EXPERIMENTATION

Having established a causal hypothesis, the next step would be to engineer an experiment to test its soundness. Following the principles of the 'scientific method', experimentation entails the deliberate and systematic manipulation of a given process, and the observation and measurement of change in that process.[13] However, depending on the nature and conditions of the investigation, any one of a number of experimental designs could be employed. The following is a description of some of the more familiar experimental and quasi-experimental designs (see Figure 1).[14] Ordinarily there are two broad types of experimental models, *comparison group* and *single group*, each having several design possibilities: *posttest* only, *pretest-posttest*, or *time series*.

Comparison Group: Posttest Only Design

One of the more common experimental approaches involves the comparison of two groups – an *experimental group* and a *control group*. Ideally, the two are identical in every way except that in one – the experimental group – the independent variable of interest is manipulated (i.e., the group is subject to an intervention, often referred to as 'treatment'), and in the other – the control group – the independent variable is not manipulated. In posttest only designs, an outcome variable of interest is measured in both groups, but *only* after the intervention. A difference between the groups is often used to imply the effect of the intervention. However, since there is not baseline data in posttest only design (information taken prior to the intervention about the outcome variable), there is no way of knowing if the effect was the result of the intervention or of some other external factor(s).

Comparison Group: Pretest-Posttest Design

Perhaps the most widely accepted, and arguably the best model for investigating causal relationships, is the pretest-posttest comparison group experimental design. It is similar to the posttest only design except that it measures the outcome variable of interest in both groups before *and* after the intervention. Therefore the change in the experimental group after the manipulation of the independent variable is compared to the change in the control group to determine if the intervention had an effect (Gould and William, 1964; Rossi and Freeman, 1993). Moreover, "with chance differences largely

[13] "The scientific method is based on hypothetico-deductive methodology. Simply put, this means that researchers/evaluators test hypotheses about the impact of a social initiative using statistical analysis techniques" (W.K Kellogg Foundation, 1998:5)

[14] Quasi-experimental designs are distinguished from 'true' experimental designs in that the latter employs a process of randomly assigning participants into control and treatment groups, while the former does not. Randomly assigning individuals makes pre-testing unnecessary, and helps to ensure that groups are as similar as possible. Also, 'true' experiment is often seen as more 'scientific' than quasi experimentation. For a defense of 'true' experimentation see Friedlander and Robins (1995), Burtless (1995), and see Heckman, Hotz, & Dabos, (1987), Heckman and Smith, (1995) for a defense of quasi-experimentation. "Most sociological research is not - and, for both practical and ethical reasons, cannot be - ['true'] experimental in character" (Goldthorpe, 2001:5).

accounted for through standard statistical techniques" (Greenberg and Wiseman, 1992), it is suggested that this is the superior model for inferring causal relationships. Using the above example, a pretest-posttest comparison experiment of the causal relationship between strenuous physical exercise and perspiration level would first entail assigning individuals into one of two groups – control or experimental. The perspiration levels of two groups would then be measured, following which a strenuous exercise regiment would be introduced to the experimental group only. The perspiration levels of the two groups would once again be measured and compared. Assuming non-spuriousness and controlling for other extraneous variables (i.e., ensuring that no outside factors interfere with the experiment), any difference in the level of perspiration between the two groups would be attributed to the introduction of the independent variable – i.e., strenuous physical exercise causes perspiration.

Comparison Group: Time Series

The fundamental difference between the pretest-posttest comparison and the time series comparison is that the latter incorporates multiple measurements of the outcome variable after the intervention. This procedure is performed to reduce uncertainty about the effect of the intervention – i.e., it diminishes the potentiality that the measured difference between groups was due to some extraneous/intervening variable.

Single Group: Posttest Only Design

While comparison group experimentation is often considered the 'ideal', it is frequently not practically or ethically feasible (Denzin, 1978; Ragin, 1994; den Heyer, 2001:22). In an extreme example, testing the effects of a new emergency medical procedure, it would be both impractical and unethical to randomly select critically ill patients and deny treatment to some in order to establish a control group. In such cases, single-group designs are the more appropriate experimental method.

The most basic of the many single-group experimental designs is the posttest only design. This involves measuring the effect of a given intervention only after it has been introduced to the subjects. Since there is usually no baseline data and no group with which to compare results, this design is often seen as the weakest – i.e., the least valid. Moreover, as suggested by the Evaluation and Data Development Strategic Policy Division of HRDC:

> *This design cannot be used to credibly attribute any effects to the program,* for there is no objective basis to suppose that the program caused any changes. Indeed, because there is no information on the pre-program level of the variable(s) of interest, this design yields no information on change (HRDC, 1998).

Single Group: Pretest-Posttest Design

The more common single group design is the pretest-posttest design. This is generally used when the investigator is interested in knowing something about the change effected by a given intervention, but when ethical and practical consideration restrict the use of control groups. Baseline data is systematically obtained by measuring the variables of interest prior to the introduction of a given intervention. Post-test information is then obtained by measuring those variables after the intervention. So for example, the perspiration level of one group would be measured before and after the introduction of strenuous physical exercise. A difference would suggest that physical exercise 'effected' the change in the perspiration levels.

Single Group: Time Series

Additionally, time series designs can be used to modify single group pretest-posttest experimentation, allowing change to be captured over time. Time series designs require measuring change at different time intervals in order to capture the progression of change, and to increase the confidence in the validity of the findings. Nonetheless, without a comparison group, little definitively can be said about effect of the intervention *per se* (HRDC, 1998).

Figure 1. Single and Comparison Group Experimental Design

	Single-group	**Comparison group**
Posttest only	Measures the outcome variable *only* after the intervention No baseline data Cannot determine change	Measures the outcome variable of two groups – control and treatment - *only* after the intervention No baseline data Comparison of the measured difference in outcome variable between control and treatment groups
Pretest-posttest	Measures the outcome variable before and after the intervention Baseline data available No comparison group means that change cannot be attributed to the intervention	Measures and compares the outcome variable of two groups – control and treatment – before and after the intervention Baseline data is available A difference in the control and treatment groups implies effect of the intervention
Time series	Measures the outcome variable before the intervention and several times after Time series strengthens the validity of the findings Baseline data available No comparison group mean that change cannot be attributed to the intervention	Measures and compares the outcome variable of two groups – control and treatment – before the intervention and several times after Time series strengthens the validity of the findings Baseline data is available A difference in the control and treatment groups implies effect of the intervention

Summary

Social science has traditionally 'borrowed' its research models from the natural sciences – adapting and modifying experimental designs for establish causation. As outline above, the more prevalent models range from single to comparison group, and from pretest, to posttest and time series. A review of these key experimental and quasi-experimental designs reveals important strengths and limitations of each type. Among the most serious concerns for experimentation is internal validity.

INTERNAL VALIDITY & EXPERIMENTAL RESEARCH

Whether comparison or single group, posttest only, pretest-posttest, or time series, the most serious concern for experimentation has to do with *internal validity*.

> A study has internal validity to the extent that the data support conclusions about the hypothesis in the specified instance studied… We make judgments about internal validity by examining the procedural details of the specific study to decide whether the procedures used to measure and manipulate variables faithfully represented those variables (Stern, 1979:62).

Campbell and Stanley (1963) submit seven (7) 'threats' to the internal validity of a study. Two of the threats involve actual changes in the environment or in participants. *Historical changes* in the environment which take place concurrently with the study (e.g., world or local events), and biological or physiological *maturation* of the participants in the study can effect the behaviour of the participants. Three other threats may result when participants are not representative of the population. The process of *selecting* and assigning individuals to 'treatment' or 'control' groups – especially in studies in which the treatment group is comprised of volunteer participants and the control group was asked to participate – can affect internal validity. And, when participants drop out of the study, *mortality,* the internal validity can be weakened.[15] Additionally, studies that employ *statistical regression* may encounter internal validity risks. That is, participants who score extremely high or low on a test may score much lower or higher on the next. This is typically due to the inherent problem of random error within statistical regression. Finally, there are threats that might be provoked by the evaluators. In pretest-posttest designs, *testing* can effect the internal validity of the study; participants may score differently on the posttest simply because they have already taken the test before. And, *instrumentation* refers to the changes in the researchers, scores, and/or tools of measurement from one stage of the research to the next (Campbell and Stanley, 1963).

Insofar as they are typically more at risk of the threats to *internal validity*, many feel that single group designs are inappropriate for measuring change, much less determining 'cause' and 'effect' relationships (HRDC, 1998). Furthermore, it has been argued that through careful design and execution (as well as a skillful application of

[15] Evaluation research tends to be particularly prone to problems associated with selection bias (such as, 'cherry picking' and 'creaming').

statistical procedures), comparison group experimentation can limit the possibility of unanticipated, extraneous influences affecting the outcome of the experiment – minimizing internal validity – and, is therefore better suited for making causal inferences. Nevertheless, the problems associated with internal validity are especially pertinent to the field of aid evaluation where unpredictable and dynamic conditions heightens a program's susceptibility to such threats as historical changes, maturation, selection, and mortality.

Summary

One of the most significant risks involved in employing experimentation to provide 'causal' interpretations is that of internal validity. As outlined above, these risks include: historical change, maturation, sample selection, mortality, statistical regression, testing, and instrumentation. And, it is important to keep in mind that while *all* social research employing experimentation may be susceptible to these dangers, single group designs are considered to be the most at risk of internal validity issues.

EXPERIMENTATION & THE SOCIAL SCIENCES

In addition to the internal validity threats within experimentation generally, there exist several obstacles unique to social research when attempting to establish causal explanations: "To identify the causes and reasons for program failure or success, sophisticated research designs such as experimental designs, time series analysis, or panel studies are necessary. However, these designs are not feasible in many circumstances" (Kuchler, 1981:168). First, the complex nature of the social phenomena makes it highly difficult to isolate and control social variables; often, it is not possible to account for the unintended effects of a confluence of influences. Moreover, phenomena within the social world are dynamic, therefore testing and measurement will always be embedded in a particular temporal context beyond which the findings can not be generalized. Secondly, experimentation on human subjects may be inappropriate, if possible at all. As previously indicated, testing and experimentation on human subjects is often unethical and/or impractical. And insofar as inferring causation relies on experimentation, the 'causes' of many social phenomena will go untested. Finally, experimentation assumes that human social action (behaviour) can be studied in the same way that natural objects are studied – i.e., properties are isolated, manipulated, and observed. It neglects the dynamic character of social action and how the meaning given by social actors to different situations alter those situations (Keat and Urry, 1975; Ragin, 1994; Roche, 1999). Molly den Heyer's review of different evaluation models within the sphere of international development reiterates the problems associated with experimental modeling of social programs:

> Proving causal inferences means that evaluation needs to utilize an experimental design that compares control groups that have not received the programs with groups that have. This type of methodology is limited in social programs by the complex variables that make replication difficult if not impossible. Further, there are ethical issues in providing social services to one group and not the other (den Heyer, 2001:22).

In addition, evaluations that employ control groups face unique problems. Particularly, "…control groups may have even less incentive to co-operate (as they are by definition liable to remain excluded) or a greater incentive to exaggerate their needs (in the hope that someone responds)" (Roche, 1999:79). Also, the basis for excluding one group may pose problems. For instance, people may be excluded on the basis of their gender, ethnicity, or age. Insofar as some groups are 'deliberately' left out of the project, Roche suggests that "as far as attribution is concerned, the existence of marginalized groups may not tell one much about what might have happened if the project in question had not occurred" (Roche, 1999:84).

To be sure, the idea of causation poses a variety of philosophical and practical difficulties. This is especially the case for the social research where the feasibility of making causal claims encounters great skepticism:

> There is always the question of whether, despite covariation and proper time order, you can ever be really *certain* a particular independent variable is the cause (or is among the important causes) of the dependent variable. This is the classic problem of 'correlation not proving causation.' Some other unknown factors, or some known but unmeasured factor, may be the cause or among the causes (Lofland and Lofland, 1995:137).

Still, as practicing researchers develop innovative methodologies and statistical procedures to strengthen the internal validity of their experimental designs, contemporary social science continues to speak in terms of 'causes' and 'effects'. However, insofar as social scientific 'causation' has less to do with 'cause' and 'effect' *per se* than with correlation between variables (i.e., statistically probable association), what is 'erroneously supposed to do no harm' is in fact often quite misleading.

Summary

Several conditions unique to social research make the use of experimentation and quasi-experimentation problematic. In particular, the complexity and dynamic character of the social world means that isolating and controlling variables can be pragmatically difficult, and can raise serious ethical issues. Different techniques are employed to address these concerns, yet the challenges are pronounced when researchers use experimentation to treat humans as 'objects of study'.

CORRELATION, SOCIAL SCIENCE & THE SHIFT TO 'PROBABILISTIC' CAUSATION

As sociologist John Goldthorpe (2001) explains, professional discourse has shifted away from notions of 'deterministic' causation to 'probabilistic' causation – i.e., "rather than causes being seen as necessitating their effects, they might be regarded as simply raising the probability of their occurrence" (Goldthorpe, 2001:1). The complex nature of the social world, as well as the incompleteness of our knowledge about it, limits the possibility of anything more than probabilistic explanations (Goldthorpe, 2001:1). Therefore, causation within the social sciences is seldom more than correlation between variables. Michael Scriven elaborates:

> [Correlation is] neither a necessary nor a sufficient condition for causation; nor necessary because causation can be established by eliminative induction (ruling out all other possible causes), and not sufficient because the correlated variables may both be effects of a third variable and have no direct influence on each other (e.g., yellowing of eye whites is not a cause of yellowing skin; if they correlate, it's because you have jaundice, probably caused by liver disease) (Scriven, 1991:104).

Holding that causation within the social sciences is nothing more than statistically significant correlation, Goldthorpe describes several types of 'probabilistic' causation. Two of these represent the majority of empirical sociological research – *robust dependence* and *consequential manipulation*. Quantitative social research has traditionally relied on *robust dependence* to make causal claims. That is, controlling for all other extraneous variables and ruling out spuriousness, statistical techniques can demonstrate a robust dependence of Y (the effect) on X (the cause). Because robust dependence tends to subordinate theory in favor of sophisticated statistical modeling, there has been growing disfavor of 'causation as robust dependence' among sociologists interested in empirical research and methodology.

> Especially from the standpoint of methodological individualism, sociologists have strongly criticized the supposition that statistical techniques can in themselves provide adequate causal explanations of social phenomena. Such techniques show only the relations between variables, and not how these relations are actually produced - as they can indeed only be produced - through the action and interaction of individuals (Goldthorpe, 2001:3).

Emerging out of robust dependence is the idea of causation as *consequential manipulation*, which essentially adheres to the principals of experimentation. This approach to probabilistic causation currently dominates empirical, quantitative sociology. Controlling for other variables, a manipulation in X will effect a particular change in the units of Y. And, this change is relative compared to the control group. Furthermore, consequential manipulation recognizes that a dependent variable cannot be exposed and not exposed to the treatment in the same experiment; hence, elaborate statistical techniques have been devised to determine 'average causal effects' of control versus experimental groups – i.e., randomized experimental design. Causation as consequential manipulation, if applied to evaluation research, exposed the limits of the attribution question. "A variable X can never be regarded as having causal significance for Y in anything more than a provisional sense; for it is impossible to be sure that all other relevant variables have in fact been controlled" (Goldthorpe, 2001:5).[16] In recent years

[16] Goldthorpe presents an alternative approach to causation for the social sciences based on these three understandings of causation. The alternative involves a sequence of three stages: a. establishing the phenomena that forms the *explananda* - this involves ensuring that the phenomenon is not unique, but occurs with some regularity. This requires statistical work, and is basically a descriptive exercise; b. hypothesizing the generative process at the level of social action - the next step is in determining the causes

there has been serious criticism of causation as *robust dependence* and *consequential manipulation*, and alternative positions suggested. Goldthorpe summarizes Lieberson's (1985) proposed alternative direction for sociology:

> This entails a straight rejection of the attempt to impose the experimental model (or, at any rate, that adopted in medical or agricultural research) onto sociology, on the grounds that it represents and undue 'scientism' – i.e., an undue regard for the form rather than the substance of scientific method – and with the implication, then, that sociologists have to find their own ways of thinking about causation, proper to the kinds of research that they can realistically carry out and the problems that they can realistically address (Goldthorpe, 2001:8).

Nevertheless, much evaluation work continues to employs probabilistic causal modeling to determine the effects of a given intervention.[17] As will be shown, however, probabilistic causal modeling tends to be more likely in certain areas of evaluation research and less likely in others (such as, aid evaluation). Additionally, one of the key determinants of the use of probabilistic causal modeling is choice of methodology. As will be shown, the type of methodology employed in evaluation research plays a critical role in whether causal probability is sought, and vice versa.

THE QUANTITATIVE/QUALITATIVE DEBATE

Early social science tended to model itself after the natural sciences; however, the social sciences are also known for epistemological and methodological diversity. Since the mid-nineteenth century, there has been a division within sociology about the principal aim of the discipline.[18] While early on the *status quo* tended to assert the positivist position underscoring the importance of quantification, experimentation and the establishment of generalized social laws, others proposed more heuristic approaches, suggesting that sociology should be about facilitating an understanding of meaningful social action. Methodologically, these two positions advocated fundamentally different strategies for studying the social world. One endorsed empirically based data gathering and analyzing techniques, while the other adopted more interpretive methodologies (for an early example of this approach, see Max Weber's *verstehen*).[19] This latter approach resisted the positivist perspective by stating that adequate explanation is best obtained through an 'empathetic understanding' of the meaning that people give their behaviour. The positivists responded that an approach based on the subjective interpretation of social action lacked objectivity and was therefore non-scientific. Interestingly, early positivist

of these social *regularities.* Any number of theories of action can be used to hypothesize the cause of specific social regularities; and c. testing the hypothesis - hypothesis must first be adequate, but there may be competing adequate hypotheses, and therefore empirical validity should be established for each (Goldthorpe, 2001:10-14).

[17] See, for example, the *Bayesian* method. As opposed to attempting to determine causation *per se*, this approach essentially seeks to increase the probability of X effecting Y.

[18] Paul Diesing's How Does Social Science Work? provides a broad and in-depth look into the ever-changing practices of social science researchers. It also highlights the dominant philosophical-scientific perspectives that have directed the course and shaped the present configuration that is social science.

[19] For a comprehensive account of *Verstehen*, see the work of the classical sociologist Max Weber.

and interpretive sociologists both tended to maintain that their methods could lead to causal explanations. Paul Diesing (1991) explains social science's positivist tradition and offers several cautions:

> The main dependence for decades has been on one philosophy, logical empiricism or *positivism*, as opponents call it. Too many researchers have learned in methods courses that the aim of science is to discover universal laws, and the method is to deduce causal hypotheses from more general theories and test them against masses of observable data. This teaching has had a dogmatic certitude – that's what science *is*; philosophers of science say so, and they know – that has not been present in many of the logical empiricists themselves. The methods also fail to mention problems that have come up in the philosophy that have led to its continual transformation and virtual abandonment by 1980 (Diesing, 1991:x).

Whether positivism has been 'abandoned' is questionable; however, critical epistemological challenges over the last few decades have resulted today in fewer positivists, and even fewer interpretive sociologists, uncritically endorsing positivism as the most effective means of studying social phenomena, let alone as 'the path to truth'. For, "whatever philosophy [researchers] choose has its own problems or weaknesses" (Diesing, 1991:xi). What is important to note is that the etiology of the contemporary qualitative-quantitative methodological debate can be traced directly to these early divisions within the discipline (Denizin, 1978; Denzin & Lincoln, 1994).

Just as the social sciences adopted many of the methods and techniques of the natural sciences, evaluation research has relied almost exclusively on the techniques of the methodologies of the social sciences. In doing so, evaluation has also inherited the methodological disunity characteristic of the social sciences. In its simplest form, this discord, commonly referred to as the *quantitative-qualitative debate*, represents vehemently defended differences of opinion over the appropriate content and manner of social scientific investigation. Interestingly, insofar as the dispute is characterized as a unilateral attack by qualitative advocates, "it would be more appropriate to describe the war as a long-lasting guerilla skirmish than an all out war" (Cook, 1997:33). Since the issues associated with this disunity are fundamentally the same for both the social sciences and evaluation research, these issues will be discussed in relation to evaluation research primarily.

Their respective goals, as well as the modes of achieving such goals, generally distinguish quantitative research and qualitative research. On the one hand, quantitative research emphasizes the collection and analysis of large amounts of measurable data with the aim of identifying patterns and relationships through the use of statistical procedures. While on the other hand, qualitative research endorses methods of collecting and analyzing detailed, descriptive information on fewer cases with the aim of producing meaningful interpretations of social phenomena (Denzin & Lincoln, 1994). Sociologist Charles Ragin offers the following concise definitions of the two types of research:

> *Qualitative research* is a basic strategy of social research that usually involves in-depth examination of a relatively small number of cases. Cases are examined

> intensively with techniques designed to facilitate the clarification of theoretical concepts and empirical categories (Ragin,1994:190).

And,

> *Quantitative research* is a basic strategy of social research that usually involves analysis of patterns of covariation across a large number of cases. This approach focuses on variables and relationships among variables in an effort to identify general patterns of covariation (Ragin,1994:190).

Restated, qualitative research is generally more interested in the commonalities that exist across a relatively small number of cases, while quantitative research seeks out correlation between variables given numerous cases. Figure 2 presents some of the more widely used quantitative and qualitative techniques for gathering data.

Figure 2 – Quantitative & Qualitative Research Methodologies

Methodology	
Quantitative – identifies patterns and correlation within a large number of cases	**Qualitative** – identifies commonalities within a small number of cases
Surveys	Ethnographic studies
Questionnaires	Participant observant studies unstructured interviews (and focus groups, etc.)
Structured interviews	Case studies
Statistical analysis (regression analysis, analysis of variance, etc.)	Historical analysis (including document review, oral and life histories)
	Textual analysis (including visual/audio analysis)

Historically, the quantitative-qualitative debate has been about the most appropriate ways of studying social phenomena, and at its heart are the notions of *objectivity* and *generalizability*. The experimental/quasi-experimental practices characteristic of quantitative research has in the past justified the perception that it is more objective than qualitative research. However, this conception has been strongly contested both by academics and practicing evaluators. The Canadian International Development Agency acknowledges that:

> It is a popular myth that information collected on quantitative indicators is inherently more objective than that collected on qualitative indicators. Both can be either more or less objective or subjective depending on whether or not principles of social science research have been rigorously applied in the data collection and analysis process (CIDA, 1999:18).

Nonetheless, the prevailing view tends to associate objectivity with measurability, and therefore quantitative methods. Additionally, quantitative research deals with large amounts of data, and is therefore considered to be more appropriate for making generalizations, from the study sample to the larger population. And, since objectivity and generalizability are highly valued scientific ideals, qualitative research has had to struggle for legitimacy.[20] Robert Stake explains that, with their social scientific backgrounds, evaluators are also more likely to value generalizations. But he also warns that, because "the structure of evaluation work usually is different" (Stake, 2001:352), evaluation should not always be equated with the social sciences.

The effects of the perceived superiority of quantitative research – on decision makers responsible for a program's future and on evaluators responsible to decision makers – is explained by Michael Quinn-Patton:

> Methodological rigor meant experimental designs, quantitative data, and sophisticated statistical analysis. Whether decision makers understood such analyses was not the researcher's problem. Validity, reliability, measurability and generalizability were dimensions that received the greatest attention in judging evaluation research proposals and reports. Indeed, evaluators concerned about increasing a study's usefulness often called for ever more methodologically rigorous evaluations to increase the validity of the findings, thereby supposedly compelling decision makers to take findings seriously (Patton, 1997:16).

However, in the wake of certain epistemological challenges, as well as methodological and analytical innovations, the sovereignty of quantitative research has come under serious scrutiny, and a reassessment and reappraisal of qualitative research has resulted. On of the more cogent critical examinations of quantitative social science can be found in Pablo Gonzalez Casanova's The Fallacy of Social Science (1981). Written during the early 1980's – what might be considered the height of the qualitative-quantitative debate – Casanova explores the ideological underpinnings of the quantitative hegemony within social science. The author explains:

> In the social sciences, a quantitative "style," perspective and emphasis are related to many other traits of the researcher. Generally speaking, it can be said that quantitative analyses are especially characteristic of the U.S. as compared to other countries and of younger sociologists as compared to older or impressionist ones. It is a style specifically linked to empiricism and the ideology of progress in the social sciences. And often it is only viewed from this perspective. But as an emphasis and perspective, quantitative style is also associated with political position. The researchers choice of style corresponds to political position regarding the social system and the status quo (Casanova, 1981: 10).

Casanova's observations coincide with the shifting perspectives about the function and value of quantitative and qualitative research practices. On the one hand, the expansion during the 1980's of 'postmodern' reasoning critically questioned many of science's sacred

[20] For example, see the National of Academy of Science's criteria for quality research in Gueron and Pauly, 1991.

ideals (in particular, the notion of objectivity), as well as the legitimacy of conventional avenues of 'knowing'. This movement enabled and embraced new methods of investigation and interpretation. On the other hand, methodological and analytic advances (including the new computer software for qualitative analysis) challenged the idea that qualitative research was unable to produce measurable 'scientific' results (for example, see Lawrence Mohr's *The Qualitative Method of Impact Analysis*, 1999). But even with these advances, it is generally maintained that "[in] the same way that they cannot answer complex questions of frequency and magnitude, qualitative field studies are not designed to provide definitive answers to causal questions" (Lofland and Lofland, 1995:136). Still, even this generally accepted notion has had detractors. Lawrence Mohr (1999) explains that determining causality has traditionally relied on one epistemological procedure, that of the counterfactual method (i.e., designing approaches that enable us to show what would have happened to Y if X did not occur). And, he maintains that this method can not be employed when using qualitative designs. However, he proposes an alternative epistemological approach, involving qualitative data, for determining causation:

> [Q]ualitative research to determine the cause of intentional human behaviors, such as whether or not those behaviors were induced by a program being evaluated, must involve a search for the operative reasons behind the behaviors. In most cases this would undoubtedly involve obtaining information from the subjects whose behavior is at issue. Given that operative reasons are unaware, that is not necessarily all that one must do. It might be necessary to obtain information from many other people, from documents, and from the histories of relevant events. These are methods, however, that are familiar in social science, especially in such areas as history, anthropology, and area studies (Mohr, 1999:75).

Mohr posits that the 'operative reason behind people's behaviour' can be ascertained, cross-referenced for validity, and analyzed in such a way as to provide adequate causal explanations. One obvious weakness in Mohr's conceptualization of causality involves generalizability. Although his 'causal reasoning' may be able to explain what led to what, it is doubtful that such an approach to causation could lead to generalizations about X and Y. To be sure, although attempts are made, qualitative research is generally perceived to be inappropriate for generating causal inferences of social phenomena. Therefore, a more important question might be – *What are qualitative studies designed for?*

As explained, quantitative methods allow large quantities of data to be statistically analyzed, and correlation and covariation between variables demonstrated. But many social researchers are unsatisfied with this approach, especially when the intention of a study is not to measure covariation but to enhance understanding about some facet of social life. Qualitative analysis is more interested in answering *why* some events occur the way they do, than in attempting to show that a given change in a particular independent variable is statistically likely to 'cause' a given change in a specific dependent variable. Thus, for example, if a study is solely interested in measuring the change in a specific outcome variable after the introduction of a given intervention, quantitative methods may be most appropriate. On the other hand, if the study aims to enhance understanding of *why* and *in which ways* the intervention affected the lives of recipients, the preferred method will be qualitative (Patton, 2002). Additionally, Thomas Cook (1997) provides some useful advice on the applications of qualitative methods:

> Few quantitative researchers would disagree with such a maxim as these: Qualitative methods are very useful for making explicit the theory behind a program; for understanding the context in which a program operates; for describing what is actually implemented in a program; for assessing the correspondence between what the program theory promised and what is actually implemented; for helping to elucidate the processes that might have brought about program effects; for identifying some likely unintended consequences of the program; for learning how to get the program results used; or for synthesizing the wisdom learned about the program or a set of programs with somewhat similar characteristics (Cook, 1997:34).

And, according to the Canadian International Development Agency (CIDA), evaluators require qualitative methods because without them they are unable to "properly assess the projects, due to lack of qualitative information on what was actually taking place in the project…" (Kuji-Shikatani, 1995:19). Qualitative methods are often seen as favorable insofar as they can be designed to meet the specifications and needs of evaluation, and can be performed swiftly and efficiently. And, by including stakeholders' subjective viewpoints about the effects of the intervention, qualitative research can substantiate and qualify the findings of impact evaluation. Furthermore, some have argued that qualitative methods are better suited for evaluation in that they are more likely to be sensitive to the conditions associated with social programs (Shaw, 1999). And, in response to the critique that qualitative methods are 'unscientific', Judy Baker maintains that, "[t]he validity and reliability of qualitative data are highly dependent on the methodological skill, sensitivity, and training of the evaluator" (Baker, 2000:8). That is to say, just as different quantitative researchers possess varying degrees of methodological ‘know-how’, the ability of qualitative researchers to appropriately employ sophisticated methods will depend on the knowledge and experience of the researcher; and, will determine the degree of analytical rigor.

While quantitative and qualitative proponents have a history of fervent engagement, Thomas Cook argues that, "[t]he case for qualitative methods does not depend on attacking the foundations of quantitative methods; it rests on their utility for answering important evaluation questions either when used alone or when used together with quantitative methods" (Cook, 1997:35). The recent ascent of qualitative studies has not meant a depreciation of quantitative research. On the contrary, applied and scholarly research continues to be dominated by quantitative methods. (This may be especially true in the case of publicly and privately funded applied research, as opposed to ‘academic’ research.)

> At present, what monitoring that might take place is largely limited to the recording of the numbers of people trained, or the amount of information or training materials produced, in other words, the 'quantity.' The 'quality' of programs - quality of learning, training, or support systems - often remains unknown. This is complicated further by the unfortunate reliance on numbers as a manifestation of effectiveness. Project continuation and job availability often drive a need to achieve an 'impressive' record. Programme personnel may believe that they must

> produce quantitative reports, rather than to merely describe what is taking place in a simple narrative format. The qualitative knowledge that could be accumulated over a program's life remains obscure, and opportunities for effective intervention are often lost. The loss of qualitative knowledge is a serious one… (Kuji-Shikatani, 1995:9).

Matching research methodology with the specifics of the given research subject – as well as with the interests and agenda of the investigator – has helped to generate innovative methodological strategies. In an effort to provide a more comprehensive, detailed and valid explanation of the given subject, projects are increasingly using multiple methodological techniques (Denzin, 1978; Ragin, 1997; McMahon, 2001). This procedure is commonly referred to as *mixed methods* or *methodological pluralism* or *methodological triangulation*. Research specialist Norman Denzin explains that:

> The rationale for this strategy is that the flaws of one method are often the strength of another; and by combining methods, observers can achieve the best of each while overcoming their unique deficiencies… When a hypothesis can survive the confrontation of a series of complementary methods of testing it contains a degree of validity unattainable by one tested within the more constricted framework of a single method… methodological triangulation involves a complex process of playing each method off against the other so as to maximize the validity of field efforts. Assessment cannot be solely derived from principles given in research manuals - it is an emergent process, contingent on the investigator, the research setting, and the investigator's theoretical perspective (Denzin, 1978: 308-310).

Multiple methods have been used in evaluation research for more than three decades; today it is common for evaluations to apply more than one method for a single study. The US development agency, USAID, acknowledges that, "[in] practice, designs may sometimes combine different approaches, either to improve persuasiveness or to answer different questions" (USAID, 1997:4).

Mixed method designs are considered especially useful in answering questions about formative evaluation, such as process and implementation questions. Lois-ellin Datta explains that, "[i]t has become almost standard to look to case studies combined with document analysis in evaluating implementation" (1997:347). However, it is only very recently that mixed methods are applied to summative evaluation in which the outcomes, results, and effects of a given intervention are sought after. Moreover, what one typically finds in evaluation research is, for example, a mix of interview and document data, but seldom is there a "close integration of a full case study with other methods" (Datta, 1997:347). Both within the social sciences generally and in evaluation research particularly, triangulation has proven to effective means of reducing the uncertainty about the finding of any one method alone. In her analysis of numerous studies related to female empowerment within the international development context, Naila Kabeer explains:

> The important methodological point brought out is *the critical need to triangulate or cross-check the evidence provided by an indicator in order to establish that it means what it is believed to mean*. Indicators compress not only a great deal of information

> into a single statistic, but also assumptions about what this information means (Kabeer, 1999:29).

Summary

Over the last several decades, qualitative methods have steadily gained ground as legitimate research – both within the social sciences generally and evaluation particularly – moving away from its place of origin on the margins of what is considered credible research. Today, depending on the interests of the researcher as well as the characteristics of her/his subject, projects may be quantitative or qualitative in the main, or may employ a healthy mix of methods. Nevertheless, as will be illustrated, the belief in the union between quantitative methodology and causal explanation has helped to maintain the separation of and opposition between quantitative and qualitative research – the latter continuing to be subordinate to the former, in practice if not in theory.

ATTRIBUTION IN AID EVALUATION REASERCH

The evaluation subject is indeed hard to define, and what is more, evaluation practice is a complicated endeavour. Not surprisingly evaluation designers, evaluators and end-users are faced with a multitude of problems, for example, problems of establishing causality between the effects observed and the intervention being studied.

– Claus Rebien, 1996

Despite the measurement difficulty, attribution is a problem that cannot be ignored when trying to assess the performance of government programs. Without an answer to this question, little can be said about the worth of the program; nor can advice be provided about future directions.

– John Mayne, 1999

[D]etermining attribution…is typically the most difficult, yet the most important, issues addressed in evaluation.

– Treasury Board of Canada

The preceding sections outline the general conditions and concerns associated with causal attribution within social research. They illustrate that, in practice, causation is always 'probabilistic' and therefore refers to correlation and covariation between variables. They reveal the problematic nature of experimentation within the social sciences, especially in terms of threats to internal validity. And, they show how quantitative and qualitative methods tend to serve different research objectives, but can be complimentary. Additionally, the review exposes the dynamic character of evaluation research – how the purposes and types of evaluation continue to evolve. Despite the difficulties associated with 'proving' attribution, the demand to demonstrate the effects of interventions and to account for actions taken and resources allocated continues to drive professional evaluation, impelling researchers to develop innovative new methods and designs. As Terry Smutylo explains, "donors are increasingly basing funding decisions on their recipients' abilities to demonstrate 'impact'… Methodologically, this requires isolating key factors that cause the desired results and attributing them to particular agency or set of activities" (2001:4).

In order to explain how evaluation research has confronted the attribution question, two interrelated sub-areas are presented:

1. **The heterogeneity of evaluation research** – Attribution is explored in terms of different research sectors, as well as the different levels of intervention and analysis. Specifically, how does the research sector and level of intervention and analysis affect the possibilities of attributing causes and effects within evaluation?

2. **Models & approaches –** Attribution is discussed in relation to historically changing conceptual models and approaches. Three stages are discussed (early evaluation, shifting paradigms, and the road ahead). Professional (frontline and academic) recommendations and cautions are provided in relation to the feasibility of attribution at each stage.

THE HETEROGENEITY OF EVALUATION RESEARCH

By now it should be clear that evaluation research practice is both dynamic and varied. As previously illustrated, evaluation has expanded and transformed over its fifty-year history, diversifying in design and practice. As a result of the field's diversity and dynamism, evaluation remains without a uniform framework or standard set of operating procedures:[21] "There is no single well-developed evaluation methodology which is universally applied. When dealing with similar projects, practice varies among agencies, from sector to sector and within the same agency and the same sector" (Rebien, 1996:55). Instead, evaluation is recognized by its heterogeneity and is defined in terms of its multiple purposes and categories (recall Rebien's definition). Accordingly, attribution questions are typically associated with a distinct category of evaluation research with a particular purpose – those that are focused on measuring 'impact' and demonstrating 'accountability'. This is not to suggest that only impact evaluations are interested in attribution, or that impact evaluations attribute 'causes' and 'effects' without complication. Certainly, "one of the most problematic parts of impact assessment is determining causality, because in real life, a combination of several factors is likely to have caused any observed change" (Roche, 1999:32). Moreover, the feasibility of attribution may be affected more by the particular subject matter under investigation and by the evaluation's unique design and methodology, than by the category or purpose of the evaluation *per se*. One way to understand why some types of evaluations are better structured to answer attribution questions is to organize evaluation by *research sector* and *level of analysis*.

Research Sector

Specialized areas of evaluation research are commonly referred to as sectors. According to the DAC Working Group on Aid Evaluation, "[a] sector includes development activities commonly grouped together for the purpose of public action such as health, education, agriculture, transport etc." (OECD, 2002:35). An evaluation sector typically implies a concentration of experience, theoretical knowledge and skill within a specific substantive research area. To be sure, no one sector is fully separated from other sectors. In practice, for example, health and medicine, science and technology, education, and transportation will all influence each other under certain circumstances. Moreover, some sectors are broadly defined, encompassing multiple sub-sectors. This is the case for the *development assistance sector*. Therefore, for heuristic purposes, a more simplistic delineation is adopted: Evaluation research sectors may be characterized dichotomously as dealing primarily with 'simple' systems or dealing primarily with 'complex' systems.[22] The simple-complex system dichotomy provides the conceptual frame to help interpret which types of sector are more likely to ask attribution questions, and which are more likely to be able to answer them.

[21] Interestingly, insofar as non-uniformity is equated with flexibility, some see this as evaluation's heralding feature.

[22] While this dichotomy is an analytical artifice and in practice no such precise delineation exists, it is often suggested that, "[e]valuation research usually deals with complex social phenomena" (Kuchler, 1981:168).

"Simple" systems within the research sector

Evaluation sectors that deal primarily with non-human/non-social environments within which interventions that can be isolated, manipulated, and measured are often labeled simple systems. These types of sectors tend to be more likely to ask, and more compatible of answering, attribution questions. For example, an impact evaluation within the agricultural sector might be able to attribute change in crop yield after the introduction of a particular intervention (e.g., new technology, agricultural practice, or fertilizer, et cetera). Hypothesis testing methodologies are generally most appropriate for the analysis of the effects of an intervention within the simple system sectors. As previously explained, this approach "stresses quasi-experimental research designs – including large-scale statistical analysis and controlled-case comparisons – that supposedly permit control of confounding variables, allow for variance on selected dependent and independent variables, and permit the disaggregation of the relative causal "weigh" of different independent variables" (Homer-Dixon, 1996:146). While appropriate for simple systems, these methods are seldom sensitive to the dynamics and logic of complex systems.

"Complex" systems within the research sector

Evaluation sectors that deal primarily with complex human/social systems rest at the other end of the continuum. These systems are:

> characterized by an immense number of unknown variables and unknown causal connections between these variables, by interactions, feedbacks, and nonlinear relationships, and by high sensitivity to small perturbations. Such complexities and uncertainties make it virtually impossible to choose cases that control for potentially confounding variables (Homer-Dixon, 1996:134).

Interventions within complex systems are embedded in, and affected by, the uniqueness of time and place. Furthermore, unlike simple systems, the variables comprising and influencing these systems are highly obscure to researchers, as well as being extremely "sensitive to small perturbations – characteristics that can altogether overwhelm both statistical and controlled-comparison methods" (Homer-Dixon, 1996:146). Insofar as multiple and often unknown confounding variables are the norm, complex systems present a serious obstacle for attribution: "These characteristics often render moot questions about the weighing, or relative strength, of specific causal variables" (Homer-Dixon, 1996:146). Evaluating aid development interventions, for example, is notoriously difficult because it ordinarily deals with distinct and rapidly changing conditions, which significantly affect research validity and reliability.

> Often, a program is only one of many influences on an outcome. In fact, deciding how much the outcome is truly attributable to the program, rather than to other influences, may be the most challenging task in evaluation study (Treasury Board of Canada, 24).

Additionally, Anne Whyte compares complex systems to communities, both of which are always in a state of transition: "Projects implemented in these systems are likely to have unexpected and decidedly stochastic outcomes" (Whyte, 2000:6). Moreover, attempts to measure the outcome of interventions within complex systems through causal modeling have been strongly criticized (see W.K. Kellogg Foundation, 1998; Lusthaus, C. et al. 2000; Whyte, 2000; Patton, 2001). And, as Pawson and Tilley explain:

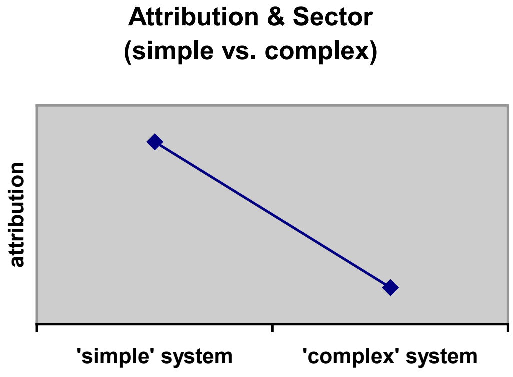
> It is precisely because of this need to explain human actions in terms of their location within different layers of social reality that realists shun the *secessionist* view of causation as a relationship between discrete events (i.e., cause and effect) (1997:406-407).

Summary

Evaluation expert Michael Quinn-Patton offers sage advice on the nature of complex systems and the kind of methods most suited for understanding them:

> [T]he complex world of human beings cannot be fully captured and understood by simply adding up carefully measured and fully analyzed parts. At the *systems* level (the whole program, the whole farm, the whole family, the whole organization, the whole community), there is a qualitative difference in the kind of thinking that is required to make sense of what is happening. Qualitative inquiry facilitates that qualitative difference in understanding human or 'purposeful systems' (2001:122).

Indeed, complex systems present serious challenges when trying to isolate and measure the effects of a given intervention; and in terms of attributing results, they are generally seen as problematic. Of course, in practice one finds that the diverse research activities of any given sector may involve both complex and simple systems. Nonetheless, interventions within some sectors (such as, aid evaluation) are more likely to be embedded in, and deeply affected by, the complexity of human/social systems. The following chart provides an *ideal typical* representation of the relationship between type of sector and feasibility of answering attribution questions.



Level of Intervention & Analysis

Project/program level

In the same way that the sector affects the ability of evaluators to demonstrate the results of an intervention, the level of the intervention and analysis will significantly influence the feasibility of attributing results. At the implementation phase, the complexity of the intervention will partly determine the extent to which attribution questions can be answered. And, "the magnitude of this problem will vary widely with the type of program and result being considered" (Treasury Board of Canada, 6). Here again, it is useful to organize levels of intervention on a continuum from 'simple' project-level interventions to 'comprehensive' program-level interventions. This helps to differentiate between narrower research interventions whose effects tend to be measured over the short-term and broad research interventions whose effects are measured over the long-term.

To clarify, simple project-level interventions refer to single initiatives, with explicit objectives, carried out within a short time frame. The DAC Working Group on Aid Evaluation defines aid evaluation at the project level as: "Evaluation of an individual development intervention designed to achieve specific objectives within specified resources and implementation schedules, often within the framework of a broader program" (OECD, 2002:30-31).[23] Insofar as project level interventions are relatively isolatable and have clearly specified objectives, attributing effects at this level tends to be less problematic. For example, measuring the effects of mosquito nets on the incidence of malaria in a small rural village is likely to be less complicated than trying to determine the effects of early childhood education programs on child poverty throughout a region. (This is not to suggest that attribution of project level interventions is without difficulties.) At the other end of the continuum are comprehensive program-level interventions. Program level interventions are characterized in terms of their extensive range and scope; generally program-level interventions encompass a variety of activities and initiatives, generally over a longer duration.[24] The following definition of evaluation at the program level within the development assistance context helps to clarify:

> a set of interventions, marshaled to attain specific global, regional, country, or sector development objectives. A development program is a time bound intervention involving multiple activities that may cut across sectors, themes and/or geographic areas (OECD, 2002:30).

Therefore, whereas the certainty of attribution is higher for 'simple' project-level interventions, the complex nature of 'comprehensive' program-level interventions makes inferring causation at this level of analysis extremely difficult, if possible at all. As John

[23] "Cost benefit analysis is a major instrument of project evaluation for projects with measurable benefits. When benefits cannot be quantified, cost effectiveness is a suitable approach" (OECD, 2002:30-31).

[24] It is important to recognize that development programs "frequently take many years to bear fruit. The “hothouse”, 2-3 year funding approach, rarely produces honest successes" (Staudt, 1991:114).

Mayne points out, any number of external variables can confound effects at the program level:

> In most cases, there are many other factors at play in addition to the impact of the program's activities. Such things as other government actions and programs, economic factors, social trends, and the like can all have an effect on outcomes (1999:3).

Even the attempt to attribute the outcome of, for example, a 'simple' project to reduce malaria incidents through a mosquito net intervention may be complicated by the influence of uncontrollable, and often unknown, factors such as environmental change, disaster, or conflict. This is equally the case at the 'comprehensive' program level. Again, this is not to imply that ‘project’ level research is free of complexity, or that the depth of complexity at the ‘program’ level renders evaluation of programs impossible. Simply, insofar as the scope is generally narrower (fewer stakeholders, limited resources and timeframe) and the purpose more specific (limited and clear objectives) for project-level interventions, evaluation at the program level will encounter considerable challenges. So, for example, reliably demonstrating the impacts of a broad health promotion program may be far less feasible than measuring the outcomes of a malaria reduction project. The following corporate perspective recognizes the challenges associated with attributing the effects of comprehensive programs:

> Cause and effect can be especially hard to measure in research on comprehensive initiatives. The complexity of these initiatives and the contexts in which they occur often makes it difficult for evaluators and researchers to establish causal relationships between programs inputs and participant outcomes. A control or comparison group, which could show causality, may not be available for every comprehensive initiative. And data collected through qualitative and quantitative methods may indicate different (even contradictory) causal relationships (Anne E. Cassie).

And, the W. K. Kellogg Foundation reiterates the problems and implications of attempting to establish attribution using research practices that are modeled after the natural sciences (i.e., experimentation and quantification):

> [M]any of the criteria necessary to conduct these evaluations limit their usefulness to primarily single intervention programs in fairly controlled environments. The natural science research model is therefore ill equipped to help us understand complex, comprehensive, and collaborative community initiatives (W.K. Kellogg Foundation, 1998:6).

That is to say, where attribution – identifying and measuring the precise effects of an intervention – is the principal goal for evaluation, researchers may be limited to evaluating at the project level.

Output/Outcome/Impact level

In addition to the problems associated with attribution at different intervention levels, the level of evaluation analysis will also generate difficulties in terms of attribution. The evaluation analysis of an intervention may involve different level of analysis. One means of organizing these levels is in terms of *outputs*, *outcomes* and *impacts*. The United Nation (UNDP) explains that **outputs** are the "tangible products (including services) of a program or project that are necessary to achieve its objectives" (UNDP). And, outputs "may also include changes resulting from the intervention which are relevant to the achievement of outcomes" (OECD, 2002:16). Outputs tend to refer to immediate the results of an intervention, making attribution at this level of analysis relatively uncomplicated. Moving from outputs to outcomes increases the level of complexity. **Outcomes** are defined as "[t]he likely or achieved short-term and medium-term effects of an intervention's outputs" (OECD, 2002:16). And, outcome effect refers to "the more immediate tangible and observable change in relation to the initial situation and established objectives, which it is felt has been brought about as a direct result of the project" (Oakley, et al., 1998:35). This level of analysis is generally associated with short-term results, decreasing the certainty of attribution. Finally, at the **impact** level, evaluation is concerned with demonstrating the "[p]ositive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended" (OECD, 2002:22). This level of analysis deals with long-term effects and, insofar as many of these changes may be 'unplanned' or 'unintended', attribution will be least feasible. Impact level analysis is incompatible for addressing attribution in that their precise effect tends to be vague and the changes generally evolve slowly over time (den Heyer, 2001:26). Additionally, as the evaluation moves from analysis at the outputs level, to the outcomes level, to the impact level, the reliability of the program's theory (which hypothesizes how the intervention will effect change) will weaken, and the certainty of attribution will diminish (Luukkonen, 1998:602; den Heyer, 2001).

The emphasis on accountability and the demand to demonstrate the causal relations between interventions and results has profoundly affected the shape of aid evaluation, compelling evaluation to focus on simple project, output or outcome level analyses:

> The need for more accountability, which seems to be a major rationale for the increased attention to impact assessment, encourages a focus on the project level and on being able to attribute impacts at the same level. (James, 2001: 7).

Mayne reiterates this view pointing to the “reluctance to accept accountability for results beyond outputs, i.e., outcomes over which one does not have control" (1999:2)

Summary

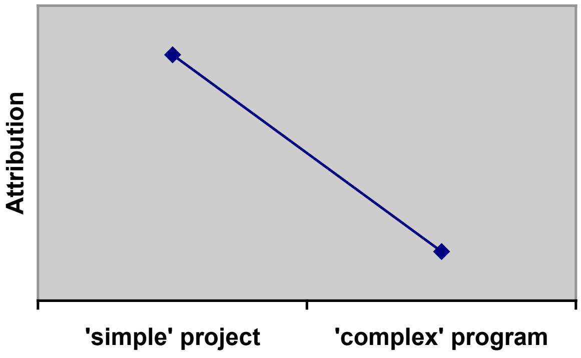
The ability to attribute results is negatively associated with the complexity of the research sector, as well as the comprehensiveness of the intervention and analysis. And, insofar as aid evaluation tends to involves comprehensive program-level analysis, this 'complex' sector is particularly prone to the problems associated with attribution (i.e., attributing the results of comprehensive program-level interventions within a complex system sector). Aware of this problem, evaluators have been critical of the utility of 'conventional' impact evaluation approaches which tend to be more appropriate for evaluating simple projects within simple systems. And, evaluators are beginning to acknowledge and address the gap between 'conventional' evaluation methodologies and the complexity of the intervention (sector and analysis): "[T]raditional evaluation models do no necessarily deal with adaptive, complex systems, which is what human communities and social-information systems are" (Whyte, 2000).[25]

Additionally, in a report on program evaluation methods, the Treasury Board of Canada offers the following warning when attempting to attribute specific effects to a given social intervention:

> It is only possible to generalize from the evaluation-determined results of a program if the program itself can be replicated. If the program is specific to a particular time, place or set of circumstances, then it becomes problematic to draw credible inferences about what would happen if the program were elsewhere under different circumstances (Treasury Board of Canada, 13).

Insofar as development assistance interventions are characteristically 'specific to time, place and set of circumstances', evaluating such interventions indeed will pose difficulties. This is particularly germane for aid evaluations driven by the need to demonstrate accountability and attribution. The following chart provides an *ideal typical* representation of the relationship between level of intervention and ability to answer attribution questions.

Attribution & Level of Intervention (project vs. program)



[25] See also the World Bank report *Assessing* Aid (1998) for an up-to-date, empirically grounded discussion of the far-reaching impact of international aid, as well as the difficulties associated with measuring impact.

MODELS & APPROACHES

The study of alternative evaluation approaches is important for professionalizing program evaluation and for its scientific advancement and operation. Professional, careful study of program evaluation approaches can help evaluators discredit approaches that violate sound principles of evaluation and legitimize and strengthen those that follow the principles.

– Daniel L. Stufflebeam, 2001

A hammer is a wonderful tool. But it is not appropriate for all situations. Similarly, there is no one perfect method which can be used in all situations to document program impact. One of the many contributions of Donald T. Campbell to evaluation was in demonstrating that all methods are flawed. The best way to control for the limitations of any single method is by using a combination of complimentary methods.

– Burt Perrin, 1998

Aware of the general problems involved in attributing results to 'comprehensive' interventions within the 'complex' development assistance sector, it is now possible to explore how different models confront the attribution question. This requires a survey of the different key evaluation designs and approaches that have been employed within the field of aid evaluation during its short history. Rather than an exhaustive typology, the following explores the most prominent and influential approaches, sorting them chronologically by category: *early evaluation*, *shifting paradigms*, and *the road ahead*. Starting with 'conventional' (or 'traditional') evaluation, strengths and weakness are exposed with an emphasis on the exhibiting the incompatibilities of some models for evaluating development assistance interventions. Particular attention is given to the positivist character of 'conventional' evaluation, and to the potentialities and problems associated with the ubiquitous Logical Framework Approach (LFA) specifically. Following, a review of the transitional phase (referred to herein as the 'qualitative turn'), which saw 'conventional' methodologies challenged and new ones embraced, looks at the significance of this shift from the attribution perspective. And finally, the attribution problem will be discussed in terms of the 'future course' of evaluation research. Expert recommendations will shed light on the 'new horizon' of aid evaluation, displaying options for addressing the problems involved in attributing results. As is evident from the literature, the more recent models are generally interpreted as responses to the deficiencies of earlier designs; however, in practice they are often used in conjunction with tradition approaches. Therefore, looking at the current state of the discipline, one finds an expanding assortment of designs and approaches, techniques and methods. Several critical questions will guide this 'purposeful' typology:

- ❑ *Are some evaluation approaches and designs more effective for attributing results?*
- ❑ *Are some evaluation approaches and designs more likely to advocate results attribution?*

These questions will be addressed by looking at the transformations in aid evaluation approaches and designs over the last several decades.

Prior to examining specific evaluation models, it is worth delineating different broad classifications of evaluation research. Basil Cracknell offers a 'taxonomy of aid evaluation' representing different stages in the 'life of a project'. Writing from the development assistance perspective, Cracknell explains that "an evaluation can take place any time after an activity has actually commenced, and many different types of evaluation have come into being, representing different stages in a project's life" (Cracknell, 2000:69). Although a detailed account of each stage will not be provided here, a brief review of each is useful.[26]

Baseline Studies: Though not considered evaluation *per se*, Cracknell acknowledges the importance of "a detailed review of the situation immediately before the development activity starts" (Cracknell, 2000:69). Baseline studies collect critical 'benchmark' information about the project's context that will be needed at later stages in the evaluation. It is worth noting that, while baseline data can provide vital information for measuring subsequent change, critics have pointed out the problematic nature of baseline studies. Particularly, the researcher may collect baseline data on variables that will later prove to be insignificant; and more likely, during the course of the project the researcher may discover that other, unidentified variables for which no baseline data exists, are in fact significant (see Roche, 1999:74-79 for a critic as well as recommendations for reducing the problems).

On-going Evaluation: Because projects may encounter unforeseen obstacles and problems, it is often necessary to carry out an 'interim' evaluation. This is typically conducted by some outside agent, allowing an unbiased, fresh perspective. On-going evaluation is sometimes referred to as 'mid-term review', 'interim evaluation', or 'formative evaluation', and is necessary for the success of later comprehensive impact evaluations (Cracknell, 2000:71-72).

Inter-phase Evaluation: While some agencies sponsor and evaluate long-term programs, "others agencies prefer to split programmes, comprising a number of projects which may be spread out over many years, into a series of phases" (Cracknell, 2000:72). As such, no new project will be funded until the success of the preceding project has been established. This practice is known as 'inter-phase evaluation'.

Built-in Evaluation: This involves setting up the evaluation at the same time as the project is being planned. This will allow personnel an understanding of what is expected from the project, and will facilitate baseline data (if needed). However, built-in

[26] The literature abounds with typologies of evaluation models and approaches with only slight differences in categorization. For a broad and general review, see Daniel Stufflebeam's monograph *Evaluation Models* (2001); and for a typology more specific to impact evaluation, see Chris Roches' *Historical Overview of Impact Assessment* (Roche, 1999:18-20). Also, Molly den Heyer (2001) puts forward a six-part taxonomy based on knowledge construction and knowledge use.

evaluation has been also considered "a waste of time because no one can possibly foretell what will happen in the future since development is a dynamic process" (Cracknell, 2000:73).

Self-Evaluation: This type of evaluation "implies that the operational staff evaluate their own activities" (Cracknell, 2000:73). While self-evaluations are more at risk of being non-objective, they may be the most efficient and effective means of evaluating numerous small projects spread throughout one or more countries.

Ex Post Evaluation: This term came out of the "need to distinguish the process of looking retrospectively at projects from the process of assessing the feasibility of proposed new projects" (Cracknell, 2000:74). Ex post evaluation typically takes place after the project has been fully implemented; it is also known as 'in-operation evaluation' and 'maturity evaluation'.

Impact Evaluation: The importance of impact evaluation emerged, during the early days of aid evaluation, out of a growing awareness that the results of evaluations conducted shortly after implementation were significantly different from the long term results. Consequently, "the emphasis switched more toward evaluations carried out some years after project implementation, that is, impact evaluation" (Cracknell, 2000:74-75).

Internal and External Evaluation: Though much confusion surrounds these terms, internal and external evaluation tends to refer to the personnel engaged in the evaluation research. Personnel from within the funding/donor agency typically conduct internal evaluations; as such, built-in evaluation is often associated with internal evaluation. Evaluators from outside the agency, however, generally conduct external evaluations. (Cracknell, 2000:75).

It has been suggested that, in recent years, development assistance has fallen under increased pressure to evaluate impacts. Scarcity of funds has meant that donors are more interested in seeing the results of projects and programs, discovering what works and what does not for the organization (known as 'institutional learning'), investing in 'sustainable' projects and programs, and being more accountable to the target group (Oakley et al., 1998). Outside of baseline studies, the problems associated with attribution, as outline above, can be found within each of these evaluation categories.

Early Models: Conventional Evaluation

The research practices of early aid evaluators reflect the ‘kind’ of aid that predominate early international assistance; aid initiatives emphasized economic support and growth, and quantitative research approaches were the norm.

> For several decades ‘development’ was understood to be essentially an economic activity; the modernization theorists of the 1950s and 1960s believed it to be synonymous with per capita growth, industrialization and economic indicators. But in 1962 the UN Economic and Social Council argued that ‘development’ is growth plus ‘change’; change in turn is social and cultural and it is both ‘quantitative and qualitative’” (Peter Oakley, et al., 1998:7).

During this period, evaluators attempting to attribute results to development interventions tended to use experimental/quasi-experimental methods; and, in their quest for an effective, systematic means of measuring outcomes, developed and adopted a variety of evaluation models and designs. To the extent that these early approaches are quantitatively orientated, and insofar as the 'logic' of their designs relies on rational linear causal modeling assumptions, these models will be referred to as 'conventional' evaluation.[27] The ensuing looks the quantitative dominion within early evaluation, and at the relation between 'conventional' evaluation and the attribution question with and emphasis on the Logical Framework Approach (LFA).

Positivism Revisited: The Quantitative Dominion

The legacy of professional evaluation is rooted in research designs that attempt to demonstrate causal relationships between variables (i.e., that aim to measure the outcome effect of specific interventions). As has been shown, attributing cause and effect using 'conventional' evaluation typically employed positivist, experimental approaches:

> This methodology required that statistically representative sample surveys be taken as the baseline and periodically over the project's lifetime in order to track changes in outcomes (e.g., living standards, incomes, mortality rates, etc.) among project beneficiaries and control groups, and to prove scientifically that benefits were caused by the project (Binnendijk, 1990:168).

Experimentation and quasi-experimentation had promised "unbiased, precise estimates of the causal consequences of programs or their major constituent parts" (Cook, 1997:32). It is this promise of 'truth and 'objectivity' that has helped to maintain the dominance of quantitative methods within evaluation research today. To be sure, the discipline continues to be guided by a powerful belief in the value of the positivist project. Furthermore, while recent years have witnessed a wave of dissent challenging the status of positivism, the contemporary view from the field tends to reiterate the importance of positivistic, experimentally based evaluation methods:

[27] This description is meant to provide a general distinction between so-called 'hard' models (i.e., quantitative) and 'soft' models (i.e., primarily qualitative).

> experimental designs are especially useful in addressing evaluation questions about the effectiveness and impact of programs… [And that,] experimental designs increase our confidence that observed outcomes are the result of a given program or innovation instead of a function of extraneous variables or events (Gribbon and Herman, 1997).

Today, Canadian, U.S., and European government and non-government agencies involved in evaluation research continue to employ positivistic/quantitative evaluation practices (den Heyer, 2001). And, evaluator Finn Hansson remarks that in Denmark evaluation research practices are "still deeply anchored in positivist (or neopositivist) empiricist data collection in which survey research predominates" (1997:186). But, guided both by the ethos of social scientific 'constructivist' epistemology and the accounts of practical experiences, aid evaluators have been quick to recognize the limits of positivist approaches. Without restating the many epistemological and methodological problems associated with positivism generally, suffice it to say not all evaluators share the same degree of faith in the positivist promise.

Among the many critiques of quantitative methods within evaluation is the view that such methods are only appropriate for answering certain kinds of questions. Specifically, quantitative methods are indeed useful for generating generalizations, from large amounts of empirical data, about multivariate relationships; however, they may have limited utility for answering questions that are central to aid evaluation. Specifically, "[quantitative methods] do not typically shed any light on why a program worked or did not work" (HRDC, 1998). And, although new advances in quantitative research design and data analysis are thought to address a broader range of questions with higher validity and reliability, the sophistication of the procedures needed to meet the prerequisites for establishing causality often make them "quite difficult to implement in field situations" (Lofland and Lofland, 1995:137). In addition, they tend to be very costly and time consuming and, therefore, are ineffectual for evaluating characteristically under funded aid interventions within a short timeframe. Moreover, even if such techniques were to be employed in 'field' situations, they typically require a level of technical skill and expert knowledge that may not be available 'in the field' (Cracknell, 2000).

From early on, evaluators have recognized the limitations of 'conventional' evaluation which has been described as ‘inadequate and misleading’ (Rebien, 1996:55). Annette Bennendijk discusses the 'methodological weaknesses' of comprehensive selection of international development evaluations conducted during the 1970s, the heyday of 'conventional' evaluation. She suggest that:

> attempting to apply experimental design standards to real-life development project situations where ransom assignments of treatments (e.g. project services) is typically infeasible and the alternative quasi-experimental design of carefully matching groups based on important characteristics is difficult to the point of being impractical. Furthermore, extraneous factors are constantly impinging on the project setting and differentially influencing the experimental and control groups. Because of difficulties such as these, the finding of some of these studies were inconclusive in terms of proving impacts and attribution, despite large expenditures on surveys (Bennendijk, 1990:170).

And, Hansson adds:

> It is almost impossible for this kind of social research, with its long tradition in conceiving social relation as just facts to be collected by questionnaires and analyzed using statistical methods, to produce socially usable and relevant knowledge on the growing complexity and changes in social relations in complex modern or postmodern industrial societies. This kind of social research cannot… produce information on social relations that can be interpreted in a historical-cultural context and integrated in a reflexive interpretation of the social processes analyzed (Hansson, 1997:186).]

Nonetheless, quantitatively oriented approaches continue to be widely used throughout evaluation research. One 'positivist' approach that emerged during this early 'conventional' evaluation phase, and which continues to dominate aid evaluation, is known as the Logical Framework Approach (LFA). LFA's notable influence on aid evaluation warrants a more detailed accounting of its utility and limits within the aid evaluation context – particularly in terms of its ability to address the attribution question.

THE LOGICAL FRAMEWORK APPROACH: DEFINED, DELINEATED, AND DISMANTLED

Evaluation work should avoid treating a program as a "black box" that automatically transforms inputs into outputs and impacts. This view leaves a huge gap in our understanding of why programs succeed or fail.

– Treasury Board of Canada, 7

During the 1950s and 1960s, development agencies adopted a variety of methods for predicting the likely impact of an intervention so that they could "approve, adjust, or reject it" (Roche, 1999:18). Emphasis was generally on 'determining the worth of the particular project or program' by weighing and comparing the associated costs and benefits (Bennendijk, 1990; Treasury Board of Canada, 107). The most common of these approaches include Cost-Benefit Analysis (CBA), Social Cost-Benefit Analysis (SCBA) and Environmental Impact Assessment (EIA) (Roche, 1999:18). Even at the end of the 1970s, cost-benefit approaches were widely accepted, though, as will be shown, not without criticism and dissention:

> [Social cost-benefit analysis] and related methodologies are sanctioned before us as scientific techniques globally applicable - here, there, everywhere. Concepts and data are treated as if they possessed cross-cultural generality, as if they were politically and ideologically neutral and theoretically unambiguous (Elzinga, 1981:5).

It was in 1971 that, drawing on the 1950s 'management by objectives' approach, the United States Agency for International Development (USAID) developed and adopted the Logical Framework Approach (LFA) (Wiggins and Shield, 1995). The

Logical Framework Approach was originally defined as, "[a] set of interlocking concepts which must be used together in a dynamic fashion to permit the elaboration of a well-designed, objectively described and evaluable project" (Practical Concepts incorporated, 1979). LFA presents a pictorial description of the 'logical' relationship among interconnected parts of a particular project or program. Represented as a four-by-four matrix, the Logframe summarizes the hierarchical relationship between program inputs ('required resources' and 'activities undertaken'), outputs ('specific result upon successful implementation'), purpose ('intermediate objectives') and goals ('ultimate development impacts'). Additionally, the logframe delineates the assumptions on which the program strategy is constructed, and provides an outline of how the project will be evaluated (what indicators will be measured and how) (Wiggins and Shields, 1995). Basil Cracknell defines three primary 'functions' of logframe approach: it helps to 'clarify objectives', 'establish indicators', and 'provide an account of the program's assumptions' (Cracknell, 2000:108-112). And, Annette Binnendijk explains the significance of the logical framework approach when it was first introduced:

> Logframe solved a major evaluation problem by clarifying at the design stage the specific development objectives of the project, and how the elements of the project were hypothesized to affect those goals" (Binnendijk, 1990:167).

The table below depicts the original logical framework matrix as presented by USAID in the early 1970s.

Figure – The Logical framework

Narrative Summary	**Objectively verifiable indicators (OVI)**	**Means of verification (MOV)**	**Important assumptions**
Goal	Measures of Goal achievement	Sources of information Method used	Assumption affecting Purpose-Goal linkage
Purpose	End of project status	Sources of information Method used	Assumption affecting Output-Purpose linkage
Outputs	Magnitudes of outputs Planned completion date	Sources of information Method used	Assumption affecting Input-Outputs linkage
Inputs	Nature of level of resources Necessary cost Planned starting date	Sources of information	Initial assumptions about the project

Today there exist numerous variants of the original logical framework; Molly den Heyer distinguishes between three 'mainstream structures' of logframe: Logical Framework Analysis, Program Logic Model, and results Chain (see den Heyer, 2001 for a detailed description of each).

The logframe's promise to facilitate standardized evaluation practices within development research was accompanied by rapid expansion in use. Internationally, government and non-government agencies have since adopted variants of the logframe approach; these include the Canadian International Development Agency (CIDA), the World bank, the United Nation, the British Overseas Development Commission (ODC), and the European Commission (GV VIII), to name a few (Wiggins and Shields, 1995; Cracknell, 2000). And, although the logframe continues to be internationally employed as "a tool for conceptualizing the relationships between short term outcomes produced by programs, intermediate system impacts and long-term community goals" (Julian 1997:251), a wave of criticism has brought about a reassessment of the limits of the Logical Framework Approach.

Although logframe remains a dominant model for evaluating development projects and programs, evaluators have been quick to recognize the limits of and problems inherent in this approach (see Gasper, 1998, Carden, 1999; Gasper, 2000). Critiques of logframe range from its inflexibility and the unreality of 'logical' assumptions, to the inappropriateness and misuse of LFA within complex systems, as well as the problems of methodological positivism (Julian, 1997; Cracknell, 2000; Pasteur, 2001).

As stated, the utility of the logframe approach lies in its ability to provide a guiding summary and overview of a project or program's objectives and intended results. This summary is meant to facilitate clarity about what is 'important' and what 'should' result from a project or program. In practice however, what is 'important' may change post-intervention, and what 'should' result may in fact be very different from what does result.

> [L]ogframes are inevitably simplifications, which become dangerous when not seen as such; they can help logical thinking, not substitute for it, yet enforcement of a fixed format tends to produce illogic; and they are prone to rigidification and thus to blocking rather than aiding adaptation (Gasper, 2000:18).

The rigidity (or inflexibility) of the logframe approach is perhaps its most often cited criticism (citations). Kath Pasteur (2001) points out that setting formal indicators is good for accountability, but indicators can become the targets themselves. The pressures of accountability may therefore urge evaluators to focus above all on measurement of indicators. In such cases, the logframe process becomes more important than the project itself. Furthermore, emphasis on attributing results to projects or programs through the strict adherence to the logical framework structure may produce misleading conclusions. That is, the careful measurement of the relations between inputs and outputs, outputs and objectives, and objectives and goals may fail to capture meaningful changes (whether positive or negative) outside of the logframe. Indicators (and in some cases objectives) set prior to the implementation may change in later stages of the project or program. Stakeholders may determine that what was 'important' early on, may not be as the situation changes. Cracknell therefore ask the important question: "What is the point of producing elaborate indices if the project is going to continually be changed by the wishes of the stakeholders?" (2000, 354). The inflexibility of the logframe approach presents obvious problems for confidently attributing project or program results. While inputs may be accurately measured against outputs, the rigidity of the logframe approach is not well suited for attributing development impacts – that is, it does not deal well with the indeterminate nature of the long-term development intervention impacts.

This raises a related criticism, namely the inadequacy of the 'inflexible' Logical Framework Approach for dealing with complex systems. Anne Whyte explains that logframe "has a formal methodology that is sometimes criticized for being too rigid, especially when applied to complex social systems requiring a more flexible, adaptive-systems approach" (2000:12). On the nature of the Logical Framework Approach, Gasper suggests that:

> LFA reflects business and logistics planning of the 1960s, with assumptions of relatively well-understood and controllable change, engineered via a 'project' within or largely controlled by a single organization. It centers attention on outputs and service delivery and on the achievement of intended effects by intended routes (Gasper, 2000:21).

The logframe appears well suited for attributing results of 'simple' interventions within 'simple' systems (see above). However, 'comprehensive' interventions within 'complex' systems may present problems for the logframe approach. During an evaluation of 'a large, urban United Way', David Julian discovers that "[logframe's] simplicity ignores the complex nature of local human services delivery systems and problems" (Julian, 1997:256). One factor that limits the utility of the logframe approach is the likelihood of unintended consequences associated with complex systems. The heightened chance for unintended consequences within complex systems can have enormous effects on the trajectory of a project or program, making attribution a exceedingly difficult task. Therefore, for example, the use of logframe within the complex aid evaluation sector makes a number of precarious assumptions about the evaluator's ability to account for the 'unintended' (citation).

> To adopt a logframe as a central tool in effects and impacts evaluation assumes that we have high powers of foresight, so that neither unforeseen routes nor unintended effects are important… Neglect of unintended effects such as externalities (impact on group other than the targets) could work for a single-mindedly self-concerned organization - but not, for example, for democracy and human rights projects or emergency assistance (Gasper, 1998:24).

Finally, logframe's emphasis on quantitative methods for attributing intervention results has been strongly criticized (Julian, 1997; Gasper, 1998; Cracknell, 2000; Gasper, 2000; Pasteur, 2001). Aid evaluation has a history of employing logframe models "based on the "ideal" of experimental and quasi-experimental research designs, whereby impacts or changes in the living standard and behaviors of the project beneficiaries could be measured and held attributable to project interventions" (Binnendijk, 1990:168). Consequently, logframe within aid evaluation has been susceptible to the various problems associated with positivist research on human behavior within complex social systems. These range from operational issues such as sample selection and control groups, internal validity and accurate measurement, to analytical problems. One of these problems is the lack of subjective accounts from program participants in logframe studies. This is not to say that logframe studies ignore the perspectives of participants/beneficiaries, but rather that subjective accounts play a minor role. Perhaps the most serious problem associated with positivist research (and the logical framework approach particularly) is its potential to neglect the contextual dynamics that are always at play within the 'social'. That is, the logical framework model has a tendency to narrow its field of vision to only those variables that are determined to be 'logically' related to the outcomes. As such, by its very design, logframe studies will neglect other, possibly unknown, intervening variables. The implication being, logical framework studies often provide only a partial representation of the factors that are responsible for observed outcomes.

The utility of the logframe approach within aid evaluation also comes into question in the recent *Review of Evaluation Resources for Non-Profit Organizations*. The authors suggest that some organization may in fact be better off avoiding the technically challenging data that 'logic models' provide: "Less technical data collection methods may, in fact, be more appropriate in the case of many nonprofit sector organizations" (Bozzo and Hall, 1999:2). Nonetheless, the criticisms of logframe within complex systems should not be taken as an outright abnegation of this approach. Rather, some suggest that insofar as the logframe approach is "inappropriate for the study of these systems… an alternative, methodologically 'pluralistic' approach to this research is proposed" (Homer-Dixon, 1996:132). This has entailed incorporating logframe with other approaches that are more sensitive to the distinct characteristics of comprehensive interventions within complex systems. It terms of the potential to demonstrate causal relations between interventions and effects – particularly, long-term impacts – the Logical Framework Approach has been strongly criticized:

> First of all, causality can not be established. Development projects are often based on the assumed causal relationship between inputs, outputs and objectives. Such assumptions are certainly embodied in the Logical Framework Approach. The logical, causal relationship does not exist in real life, development intervention situations, however. In reality, constantly changing conditions and almost unpredictable external and internal factors are the order of the day, significantly affecting intervention results. The assumed casual relationship can therefore hardly be established when evaluating interventions (Rebien, 1996:55-56).

Summary

Given the diverse and changing nature of aid evaluation, no single model is appropriate for *all* evaluation situations; instead, the variety of models reflects the range of research related experiences characteristic of development assistance. Nonetheless, early aid evaluation is dominated by positivist-quantitative models, and, specifically with Logical Framework Analysis (LFA). Designed and most suitable for economic-focused interventions, the appeal of LFA rests on its apparent parsimoniousness and the assumption of scientific accuracy. Critics, however, point out that the LFA model may be too rigid, and that if the aim of the evaluation is understanding (for instance, *why* and *how* things changed), then alone LFA may be inappropriate. It is important to keep in mind that although aid evaluation has a tradition of quantitative research (and, specifically, LFA use), as the nature of the profession has transformed so too has its research techniques and strategies. For example, the increased emphasis on governance and democracy, institutional learning and capacity building, participation and empowerment has meant a shift toward more compatible and appropriate methods of evaluation. So, increasingly one sees LFA models being used in conjunction with more qualitative, case-oriented studies.

Shifting Paradigms: A Qualitative Turn

Our era is characterized by an epistemological transition… The epistemology of the American culture was essentially derived from the Greek-European epistemology based on deductive logic, assumptions of one-way causal flow, and hierarchical social order.

– Muruyama, 1981

From the early 1980's, new methods of inquiry emerged which sought to make people and communities subjects and active participants, rather than objects of impact assessment.

– Chris Roche, 1999

By the late 1970s, the 'critical mass' of problems associated with 'conventional' evaluation helped bring about an epistemological 'turn of tides', accompanied by creative alternative models and approaches for conducting evaluation research. Philosophically, the 'logic of unidirectional causality' embodied in 'traditional science' was seen to be misguiding (Maruyama, 1981:202). Moreover, the 'postmodern' atmosphere of the 1980s engendered a deep skepticism about the positivist 'promise'; and, new ideas about the purpose and limits of evaluation corresponded with new methodologies (see Stake, 1975). The shift is manifest in a debate that has been given many labels – the positivist/constructivist debate, the objectivist/subjectivist debate, and perhaps most often cited, the quantitative/qualitative debate. In *Evaluation: Preview of the Future #2*, M.F. Smith suggests that this debate endures today, and although it is often shrouded in methodological polemics, it "was and is about differences in philosophy and "world view" (Smith, 2001:292). The ensuing discussion explores the attribution question within evaluation research from this new 'world view'. In doing so, it reveals how evaluation has evolved, gradually shifting consensus on the roles and limits of the profession. And, it shows that although attribution continues to remain elusive, this 'qualitative turn' may in fact provide a better means of addressing impact questions. Therefore, the following emphasizes one side of the debate – the qualitative/constructivist/subjectivist side – in order to explicate the relationship between this new 'paradigm' and the specific attribution problem.[28] Furthermore, it assesses attribution with specific questions in mind:

- ❑ *Does this shift provide an alternative means for exploring the relationship between interventions and impacts?*
- ❑ *Does this shift represent a move away from attempting to attribute results?*

[28] A full account of the different models that represent this paradigm shift will not be provided here. For some of the more well-known representative works, see: Feuerstein's *Partners in Evaluation* (1986), Patton's *Qualitative Evaluation Methods*, Guba and Lincoln's *Fourth Generation Evaluation* (1989), Fetterman's *Empowerment Evaluation* (1994), and Pawson and Tilley's *Realistic Evaluation* (1997).

Perhaps the most salient expression of evaluation's shift away from the positivist project is Guba and Lincoln's (1989) *Fourth Generation Evaluation*. Heavily influenced by Robert Stakes' *Responsive Evaluation*,[29] the authors comment that the three preceding 'generations' (*measurement, description,* and *judgement*) represent undue 'managerialism', neglect the importance of 'value plurality', and express baseless reverence for the 'scientific method' (Guba and Lincoln, 1989). They explain that evaluation has moved into a fourth generation. In essence, this generation is characterized by: the rejection of the validity of the scientific method within the human sector; the rejection of the 'ideal' of detached, objective research; an embrace of the plurality of values; and, an embrace of the participation of *all* parties involved in evaluation. Fourth Generation also emphasizes negotiation between agents involved in the evaluation – a negotiation concerning the course and direction of the evaluation, as well as a negotiation of values and meanings (Guba and Lincoln, 1989). Therefore the role of 'fourth generation' evaluators has transformed, from 'detached observer' to 'engaged facilitator'. Cracknell explain: "[T]he evaluator has most of his/her former roles (technician, describer, assessor), but with a difference; now he/she has to take on a number of other roles, such as collaborator, learner/teacher, reality shaper (catalyst), and mediator or change agent" (Cracknell, 2000:318). But perhaps most significantly herein, 'fourth generation' evaluation "was aimed at debunking scientific positivism and evaluation's reliance on quantitative data" (den Heyer, 2001:48).

During the 1980s and 1990s, evaluation gradually adopted many of the 'fourth generation' tenets. Claus Rebien (1996) explains that 'fourth generation' and 'participatory evaluation', (both initially termed ‘participatory evaluation’), tend to embody very similar philosophical and methodological ideals. He defines 'participatory evaluation' as "as an evaluation process where stakeholders are involved in the design, data collection, analysis and use phases of the evaluation. Stakeholders are defined as people working with or being affected by the intervention” (Rebien, 1996:5). Today, evaluation research emphasizing stakeholder involvement has various names including, 'participatory', 'empowerment', 'transformative' and 'inclusive' evaluation. Although distinct, each holds in common the centrality of stakeholder involvement at all levels (see Mark, 2001; Torres and Preskill, 2001)). Furthermore, Elzinga explains the significance of the shift toward more participatory, action-focused evaluation:

> [It represents] a departure from the traditional method which emphasizes non-involvement and justifies it with positivist epistemology. The action-research school for its part criticizes traditional methodology for being extremely unrealistic and maintaining a false isolation of evaluation from the project or program being evaluated (Elzinga, 1981:40).[30]

To be sure, while some of the more extreme views engendered in 'fourth generation' thinking have been received dubiously (such as the outright rejection of the

[29] Among other things, Stakes' 'responsive evaluation' advocated the importance of context, as well as the utility of qualitative methods, within evaluation research.

[30] Bozzo and Hall explain that participatory evaluation does have its problems: "The possible down-side of the participatory approach is that they are time-consuming, since staff need to allocate time to the process and participants may need special guidance to be integrated into the process" (Bozzo and Hall, 1999:4).

validity of the scientific method), the movement has been nonetheless influential. The emergence of numerous qualitatively oriented evaluation approaches, and the intense scrutiny and critique of quantitative methods during the 1970s and 1980s, corresponded with the emergence of participatory focused evaluation (see Stake, 1975; Patton, 1980; Feuerstein, 1986; Pawson and Tilley, 1997), and reached its apogee within the 'fourth generation' (Cook, 1997:33). Today, these 'participatory-focused, qualitative-oriented' evaluation approaches tend to se subsumed under a common rubric: *constructivist evaluation*.[31] As summarized by the W.K. Kellogg Foundation, this new paradigm represents a dramatic shift away from 'conventional' evaluation's emphasis on 'proving' outcomes:

> The primary objective of evaluations based on the assumptions of interpretivism/ constructivism is to understand social programs from many different perspectives. This paradigm focuses on answering questions about process and implementation, and what the experiences have meant to those involved. Therefore, it is well suited to helping us understand contextual factors and the complexities of the programs - and helping us make decisions about improving project management and delivery (W.K Kellogg Foundation, 1998:10).

Oakley, et al. explain, "the search for a more process-oriented, qualitatively sensitive and 'learning' form of evaluation has been intense and, in theory at least, relatively successful" (1998:28). And, in an in-depth survey of aid evaluation, Kuji-Shikatani (1995) finds that "Participatory Evaluation is advocated by the vast majority" (1995:263). From an organizational perspective, participatory methods are widely viewed as crucial within aid evaluation: "Reengineering calls for a more participatory approach to evaluation, involving customers, partners and stakeholders – as appropriate – in all phases of the evaluation process" (USAID, 1997:3).

Methodologies associated with 'constructivist evaluation' are qualitative in the main. They generally emphasize thorough, detailed understanding of a situation, but do not require quantification. Creswell (1994) explains that:

> In a qualitative methodology inductive logic prevail. Categories emerge from informants, rather than are identified *a priori* by the researcher. This emergence provides rich 'context-bound' information leading to patterns or theories that help explain a phenomenon. The question about the accuracy of the information may not surface in a study, or, if it does, the researcher talks about steps for verifying the information with informants or 'triangulating' among different sources of information (Dale, 1998:109).

Common qualitative methods include in-depth and semi-structured interviews, focus groups, open-ended surveys and questionnaires, and textual analysis. A more elaborate and often more fruitful qualitative methods is the *case study* (Stake, 1978; Datta, 1997; House, 2001; Stake, 2001). House (2001) suggest that all 'knowledge is situation, and

[31] See Pawson and Tilley's *Realistic Evaluation* (1997) for view similar to the constructivists, but one that "differs from the idealism held by some constructivists who deny that there is any reality apart from our interpretations" (Henry et. al, 1998:4).

contextually bound', therefore the researcher must become intimate with and sensitive to the program's setting if she/he is to understand it.[32] The case study, which may involve a number of techniques ranging from participant observation to textual analysis, provides a means to this end. Robert Stake offers the following description of the case study:

> "[M]ost case studies feature: descriptions that are complex, holistic, and involving a myriad of not highly isolated variables; data that are likely to be gathered at least partly by personalistic observation; and a writing style that is informal, perhaps narrative, possibly with verbatim quotations, illustration, and even allusion and metaphor. Comparisons are implicit rather than explicit (Stake, 1978:7).

The case study (as well as other qualitative approaches) represents a sharp turn away from the ideals of detached objectivity, quantitative measurement, and causal explanation which is embodied in 'conventional' evaluation. It endorses a kind of evaluation that is interested in generating meaningful interpretations of contextually bound, often highly complex, situations.[33] Stake also reminds us of the "need for explicating special contexts, not neutralizing them as error effects" (2001:352). Consequently, the situational knowledge that emerges through these approaches tends to be perceived as ill suited for the kinds of analyses that can generate causal explication. Moreover, qualitative evaluation has carried the unfavorable reputation of being "an imprecise, ill-focused, descriptive, inductive exercise, strong on vicarious experiences, but chronically at risk of failed credibility in the eye of the people who count [sic]" (Shaw, 1999:123); a reputation that has helped to keep it at the margins of 'mainstream' evaluation. This, however, is changing. In response to the concern that case studies present data validity problems, House offers the following simple answer:

> Since the evaluator relies on impressions and 'personalistic' observation and not on standard data collection and analysis techniques, how does one keep from being wrong? The major way is to try out the ideas on people in the setting. Let them respond to what the evaluator has written and challenge it… allowing others freedom to disagree, even encouraging them to do so. Evaluators should admit their fallibility and make this known to stakeholders and audiences (House, 2001:26).

Moreover, as has been shown, attributing change to a single intervention is particularly difficult within aid evaluation; a confluence of factors (both other initiatives and unknown, uncontrollable conditions) may obfuscate relationship between interventions and effects. Given this characteristic of aid evaluation, the case study may in fact be a better option. Exposing the detailed contexts within which interventions are implemented and carried out may help to disentangle the web of extraneous variables commonly found within 'complex' systems:

> The existence of multiple initiatives in a given community makes it difficult to

[32] "To understand a program one must travel into the program setting in the deeper sense to see where people live and how they think. Their beliefs and judgements should be included in the report even if the evaluator does not agree with them, perhaps especially if the evaluator disagrees" (House, 2001:25).

[33] It should be noted that, when using case studies the researcher must strike "the appropriate balance between random sampling among projects and programs versus purposive sampling, or "cherry-picking"" (Ryan, 2002:8).

> determine which effort is responsible for results. Although the varied role of community factors in comprehensive initiatives makes it hard for evaluators to generalize findings across communities, incorporating community context into an understanding of local interventions helps researchers recognize the idiosyncratic nature of comprehensive services (Anne E. Cassie, 1995).

Other evaluators argue that qualitative methods are in fact quite amenable to causal explanations. Lawrence Mohr's (1999) offers a strategy for inferring causation through qualitative data (see above); and, Ian Shaw's *Qualitative Evaluation* (1999), suggests that "[w]ith regards to outcomes and causality, …qualitative evaluation provides a better fit for evaluation purposes because it attends to micro-processes, local theory, and contextual variables" (Kaminsky, 2000:334). Additionally, the success of qualitative evaluation is also evidenced in its growing practical application and acceptance. Oakley, et al., summarize the findings of a mixed method (combined quantitative and qualitative) study conducted by in Nepal by ACTIONAID:

> [T]he quantitative studies provided information on outcomes but not impact. The impact data was obtained from the qualitative study which showed the felt and observed changes in the quality of life. They concluded that possibly only qualitative data could therefore provide information on impact. (Oakley, et al., 1998).

Summary

Qualitative methods, therefore, offer an alternative means of establishing causal explanations, one in which 'certainty of attribution' – typically derived through quantification and statistical manipulation – is exchanged for a well corroborated, 'deep understanding' of *what* changed and *why*. In effect, the qualitative approach offers a new understanding of attribution within aid evaluation; rather than it implying the effect caused by a specific, isolated variable, attribution is sought through the in-depth study of the relationships between numerous variables that together affect change. In doing so, they also offer a potential remedy to the attribution problem, one that involves redefining the role and goals of the evaluation. Like the disunity within the social sciences, evaluation has struggled over its ultimate role and goals. The movement toward a more interpretivist approach has helped to redefine the roles and goals of evaluation research generally, and aid evaluation specifically (Stufflebeam, 2001).

The Road Ahead: Recommendations & Cautions

The relationship of evaluation research to data-analytic techniques is both complex and opaque: The clarification of this relationship is important not only for 'applied' social scientists but for policy maker's program personnel.

– Manfred Kuchler, 1981

There is no agreement among agencies or academics about the best way of evaluating aid. Different approaches are used by different agencies and a change in approach often occurs from one evaluation to the next.

– Claus Rebien, 1996

There is no one ideal design for an evaluation or research study. All studies involve compromises in the light of on-the-ground circumstances and the realities of resource constraints. To obtain results as accurate as possible, given the available time and funding, many trade-offs are made…

– Anne Whyte, 2000

To review, the problems associated with attributing precise changes to specific aid interventions are many and varied; nonetheless, the growing demand by donors for accountability has meant increased pressure to demonstrate results. For aid evaluators, this often means employing strategies that promise empirically grounded results and relationships. In such situations, the prevalence of logframe approaches and quantitative methods may have less to do with epistemological superiority than with the depth of its roots in tradition. As has been illustrated, a history deeply entrenched in the ideals of positivism has resulted in the perception that attribution can only be 'validly' determined through quantitative measurement and experimental/quasi-experimental designs. This view was critically challenged during the late 1970s and early 1980s, resulting in what has been referred to as a 'paradigm shift' within evaluation research. In addition to an acute skepticism toward the potential and limits of the 'scientific method', emphasis shifted toward 'value plurality' through stakeholder participation and 'empowerment', underscoring the need to comprehend and incorporate 'context' within evaluation. In essence, it marked a move away from the 'box-filling' evaluation that characterized the previous era. This has been particularly important within the field of aid evaluation wherein the 'complex' nature of this sector, as well as the 'comprehensive' character of its interventions, make measuring attribution highly difficult; and, in which accurate explanations of long term impact and change are best achieved through meaningful, contextual understanding.

In terms of attribution, two concurrent sets of circumstances appear to be guiding evaluation's course. On the one hand, aid evaluation has increasingly adopted more qualitative methods of inquiry, emphasizing understanding over measuring. But at the same time, funding concerns associated with development projects have meant increased demand by donors for accountability, renewing an urgency to attribute results to projects.

Therefore, "although evaluation methods have evolved considerably… they are still heavily influenced by the need to measure performance for accountability purposes" (Whyte, 2000). To shed light on this apparent contradiction, several thematically related areas will be discussed. The present status of the quantitative/qualitative debate will be linked to the increasing use of multiple, mixed methodologies within aid evaluation. Secondly, the changing conceptualization of 'causation' within evaluation research will be discussed; and, finally the growing interest in and concern over the role of advocacy within aid evaluation will intimate the familiar 'objectivity/subjectivity' research problem.

Quantitative/Qualitative Debate

The contemporary literature presents a picture of the quantitative/qualitative debate as a 'quiet front' wherein, if not altogether resolved, a 'friendly truce' has been established (Greene, 2001; House, 2001; Pawson and Tilley, 2001). Evidence of qualitative research's achievements can be found in the numerous 'successful' qualitative evaluations within the literature. In Daniel Stufflebeam's monograph, *Evaluation Models*, twenty-two evaluation approaches ranging from 'objective testing programs' and 'performance measurement' to 'program-theory based studies', 'mixed-methods studies', and 'client centered studies' are described and assessed (Stufflebeam, 2001). Most of the approaches that made Stufflebeam's 'best list' are highly amenable to qualitative methods:

> When compared with professional standards for program evaluations, the best approaches are decision/accountability, utilization-based, client centered, consumer -oriented, case study, deliberative democratic, constructivist, accreditation, and outcome/value-added assessment… The worst bets were found to be politically controlled, public relations, accountability (especially payment by results), clarification hearings, and program-theory based approaches (Stufflebeam, 2001:89)

Additionally, he concludes that, "clearly approaches are showing a strong orientation toward stakeholder involvement and use of multiple methods" (Stufflebeam, 2001:89).

As has been explained, the 'qualitative turn' did much to enhance evaluation research. Particularly, it recognized the significance of 'situational, context-bound knowledge', it strengthened the role of stakeholders, and it promoted a deeper understanding of the ways that people and programs interact within unique setting. The adoption of a multiple methodological approach for conducting evaluation research emerged out of this 'paradigm shift'. The trend toward multiple methods seems to have been borne out of the realization that no single method is able to provide the full picture of the relationship between interventions and changes.

> There is no single correct evaluation design for impact evaluations. The goal is to come up with the best design possible under the circumstances. Almost all designs represent a compromise dictated by many practical considerations such as how much money and time are available, what the client considers compelling, how much a design might interfere with the normal operation of the program, and so on (HRDC, 1998).

Accordingly, more and more evaluators have become convinced that the 'best design possible' involves multiple methods. To be sure, the frequency of multiple method use within evaluation has risen dramatically in recent years. This has been particularly important for 'complex' evaluation sectors, and for evaluating at the 'comprehensive' program levels (such as, for aid evaluation). Homer-Dixon's (1996) assessment of the problems associated with evaluating within 'complex ecological-political systems' reiterates the importance of multiple methods. He suggest that many of the commonly adopted methods "are inappropriate for the study of these systems, and an alternative, methodologically “pluralistic” approach to this research is proposed” (Homer-Dixon, 1996:132). To be sure, the call for multiple methods is not new. Within the social sciences generally there has been an emphasis on 'triangulating methods' – on the one hand, to reinforce the validity of qualitative findings, and on the other, to enrich quantitative results. Robert Stake's *Responsive Evaluation* emphasized the importance of involving more than one type of method for any single evaluation. House summarizes:

> The greatest strength of responsive evaluation is that it helped break the intellectual stranglehold that single-method approaches had on evaluation at one time. It legitimated different avenues to conducting evaluations. This influence was liberating and highly beneficial as evaluation evolved into a multimethod professional practice. Stake's responsive evaluation played a major role in expanding the field intellectually (House, 2001:26).

Multiple methods are particularly effective because they are capable of answering a range of different questions within a variety of settings. As such, they help to reconcile the 'inadequacies' of any single method; and, when data from multiple methods are triangulated, findings can be corroborated or contradicted, strengthening validity and credibility. Consequently, evaluation research has been quick to adopt the mixed methods approach. As Basil Cracknell explains:

> Most people now accept that the pluralist approach is the right one that is, the use of a number of different techniques and methods, is the right one, as may seem appropriate rather than just focusing on one (2000:350).

Notably, the widespread use of multiple methods may be evidence that the quantitative/ qualitative debate is all but over – in practice, if not in theory (Mark, 2001).

A Changing Conception of ‘Causation’

The growing acceptance and use of multiple methodologies within evaluation research is linked to the discipline's changing (albeit gradually) conceptualization of 'causation'. On the one hand, there seems to be a change in the 'standard of evidence' by which evaluator's 'measure' the effects of interventions. Notably, where 'proof' of attribution is required, the means of establishing it may not necessarily follow 'conventional' methods. Increasingly, evaluators are adopting mixed methods to 'reduced uncertainty' and generate 'reasonable confidence' as a satisfactory substitute for 'statistical significance' alone. And, they appear to regard the loss in statistical rigor as outweighed by the gain in understanding of 'which programs work', 'what parts of which programs work', 'why they worked', and 'in what contexts'. John Mayne explains:

> Measurement in the public sector is less about precision and more about increasing understanding and knowledge. It is about increasing what we know about what works in an area and thereby reducing uncertainty… We need to include softer and qualitative measurement tools in our concept of measurement in the public sector (Mayne, 1999:5).

Additionally, he suggests that, "[w]e need to accept the fact that what we are doing is measuring with the aim of reducing the uncertainty about the contribution made, not proving the contribution made" (Mayne, 1999: 16).[34] Albeit, this may not represent a change in the conceptualization of causation so much as a new standard for determining the relationship between intervention and impact.

Evaluation's notion of causation is changing in other ways too. Recall that, since causation *per se* is not possible (see 'Causation: Background and Terms'), it has been used within evaluation to refer instead to 'probabilistic causation' (i.e., correlation) determined through quantitative analysis. As such, its strongest feature is its ability to establish 'generalizations' from large quantities of information involving a relatively small, 'controlled' set of variables. Qualitative research, on the other hand, is interested in the 'particularities' of a relatively small number of cases. In this sense, the quantitative/ qualitative debate has represented a struggle of values, between the 'ideal' of generalization and the 'ideal' of particularization.[35] As much as qualitative research has been applauded for recognizing the importance of the particularities of context, the quantitative research ideal of generalization endures. And, insofar as generalizability remains an integral feature of most evaluation research, the question resounds: How does one generalize from the particularities of qualitative data? The response to this question may be found in *evaluation synthesis*. Cracknell explains:

> In then early days of evaluation activity most of the evaluations were of individual projects. But it was soon found that these were of little value…because it was not possible to draw inferences of a general nature from only one project …So it became the practice to 'cluster' projects by sector, and then to produce 'syntheses' of the findings, on the principle that if the same finding recurred in several places it was justifiable to draw a broad conclusion with some confidence (Cracknell, 2000:199)

The emergence of evaluation synthesis therefore helped make single project evaluations more useful, and provided a new means by which evaluators might 'confidently' generalize from the 'particularities' of qualitative data (see also Lipsey, 2001; Pawson and Tilley, 2001; Patton, 2001). Ernest House (2001) suggests three additional avenues for 'reducing uncertainty' when evaluating the specific effects of a particular

[34] Explaining 'contribution analysis', Mayne asks whether "a reasonable person, knowing what has occurred in the program and that the intended outcomes actually occurred, agrees that the program contributed to those outcomes?" (Mayne, 1999:7).

[35] Interestingly, Robert Stake (2001) suggests that "[a]ny evaluation can be thought of as a case-study, given the peculiarities of the program implementation and context" (Mark, 2001:461).

program: the case study, meta-evaluation, and program theory. He explains that each approach "takes account of the more complex social reality by framing the program and the study more precisely, albeit, in different ways" (House, 2001:312). Indeed this marks a shift away from the conception of causation that is based on 'robust dependence' or 'sequential manipulation'.[36] Whether it is more or less valid ultimately depends of the methodological vigilance that the evaluator brings to bear upon the study. But on the issue of 'causation' within evaluation, House is adamant: "[Causation] remains incomplete, unfinished business for the field, except to say that we do understand that social causation is more complex than we thought back in the old days" (House, 2001:312).[37]

A final noteworthy development in aid evaluation is the shift in emphasis from evaluation set on 'proving', to evaluation focused on 'improving'. Astutely aware of the epistemological and methodological limits inherent in *all* research involving complex human/social systems, it appears as though aid evaluators are beginning to redirect the aim of their discipline, as well as their role as evaluators.

> [T]he intended "impact" of the program is its guiding light and directional beacon, not the yardstick against which it is measured. Thus the threat of failing to discover "hidden attribution" is eliminated when feedback on performance concentrates on improving rather than on proving, on understanding rather than on reporting, on creating knowledge rather than on taking credit (Smutylo, 2001:5).

To be sure, evaluation's emphasis on 'improving' programs is not new, but for aid evaluation the ongoing threat of a 'drying well' at home has meant deepening concern with improving programs abroad. Consequently, more organizations appear to be adopting this view, and are directing their research agendas toward 'improving' programs and result over 'proving' cause and effect:

> We also believe that evaluation should not be conducted simply to *prove* that the project worked, but also to *improve* the way it works. Therefore, do not view evaluation only as an accountability measuring stick imposed on projects, but rather as a management and learning tool for projects, for the Foundation, and for practitioners in the field who can benefit from the experiences of other projects (W.K Kellogg Foundation, 1998:3).

But, the emphasis on improving has once again raised the specter 'objectivity-subjectivity' within evaluation research. The 'constructivist' skepticism has corresponded with an increased interest in 'advocacy'. And, although it is not the aim of this paper to 'evaluate the evaluator', it ends with a word of warning from Robert Stake:

[36] Evaluation synthesis may involve examining qualitative, quantitative and mixed methods evaluation, therefore increasing confidence through a kind of 'meta-triangulation'.

[37] The 'old days' to which House is referring was a time when evaluators believed that: "One may formulate an evaluation project in terms of a series of hypotheses which state that 'Activities A, B, C will produce results X, Y, Z'" (Suchman, 1967:93).

> Now, with postmodern insight, validity is methodologically unimportant, epistemology destabilized. Without the backing of the positivist authority, we evaluators are caught in the web of advocacy and have become unwitting, sometimes unwillingly, simply a party to promotionalism" (Stake, 1997:475).

Summary

One of evaluation's strongest assets has been its ability to change and grow in response to perceived limitations and to the evolving 'business' of professional evaluation. The themes of transformation and adaptation throughout its history reflect evaluation's flexibility. While quantitative research dominated early evaluation, the successes of qualitative projects did not go un-noticed or unappreciated, leading to a 'friendly truce' within the quantitative-qualitative debate, and to an increased support of mixed-methods approaches. Perhaps the most significant changes that evaluation has undergone in recent years have been the shifts toward understanding, and the move away from 'proving' and toward 'improving'. The recognition of the problematic nature of attribution has engendered a shift in the conception of causation away from proving relationships between variable, and toward reducing uncertainty about how things relate and change. These changes have been considerable for evaluation generally, and for aid evaluation specifically.

CONCLUSION

The preceding literature review provides a historically grounded account of the issues and concerns associated with attributing results within aid evaluation research. By chronicling the evolution of evaluation research it has explained the context in which the 'attribution question' has emerged and changed over the last several decades. And, it has uncovered and critically examined the assumptions that underlie those dominant evaluation theories and practices that have claimed to satisfy the attribution question. In sum, it reveals the problems associated with traditional conceptions of causation within evaluation research, and it outlines constructive responses. It has shown how early models and approaches tended to employ positivist-quantitative research methods and represented a 'top-down' philosophy for administering and evaluating aid. At the heart of this approach is the Logical Framework Analysis model, which, although ubiquitous within the field, has been subjected to harsh criticism. Particularly, it has a reputation for being 'inflexible' when organizing and analyzing social phenomena, and for neglecting the dynamic and complex character of social life. Essentially, the 'early model' approaches have recognizable limitations. In response, evaluation research began to adopt some of the methods of qualitative social scientific researchers. By doing so, one sees the emphasis on attribution shift, as the 'business' of evaluation increasingly advocates 'understanding'. Still, the utility of quantitative approaches are recognized; thus, recent years have witnessed the rise of mixed-methods research. By employing multiple methods evaluators do more than strengthen reliability, they broaden the depth and range of understanding of the changes associated with interventions. Stimulating this shift toward mix-method evaluation are philosophical changes in the purpose and function of evaluation. This is particularly germane within aid evaluation where understanding complex systems requires 'partnership', 'empowerment', 'good-governance, and 'capacity building' – central tenets for modern aid evaluators. By highlighting the changes and the lessons learned over the short history of evaluation research, this review provides a backdrop for better understanding the 'attribution question', as well as an 'entry-point' for prospecting the future of the discipline.

BIBLIOGRAPHY

Annie E. Casey Foundation. 1995. *Getting Smart, Getting Real: Using Research and Evaluation Information to Improve Programs and Policies* – Report of the Annie E. Casey Foundation's September 1995 Research and Evaluation Conference.

Baker, Judy L. 2000. Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners. The World Bank, Washington, DC.

Bamberger, Michael. 1997. *Understanding the Impacts of Development Projects on Women*. In Chelimsky, Eleanor and William R. Shadish (eds.). Evaluation for the 21st Century: A Handbook. Sage Publications, California, USA.

Bamberger, Michael. 2000. *The Evaluation of International Development Programs: A View from the Front*. American Journal of Evaluation. Vol.21(1):95-102.

Bhola, H.S. 2000. *A Discourse on Impact Evaluation: A Model and its Application to a Literacy Intervention in Ghana*. Evaluation. Vol.6(2):161-178, Sage Publications, London, UK.

Binnendijk, Annette L. 1990. *Donor Agency Experience with the Monitoring and Evaluation of Development Projects*. In Finsterbusch, Kurt, Jasper Ingersoll and Lynn Llewellyn (eds.), Methods for Social Analysis in Developing Countries. Westview Press, Colorado.

Blalock, Ann Bonar, 1999. *Evaluation Research and the Performance Management Movement: From Estrangement to Useful Integration?* Evaluation. Vol.5(2):117-149. Sage Publication, London, UK.

Bozzo, Sandra L. and Michael H. Hall. 1999. *A Review of Evaluation Resources for Nonprofit Organizations*. http://www.ccp.ca/information/documents/gd44.htm. Canadian Centre for Philanthropy.

Campbell, D.T. and J.C. Stanley. 1963.Experimental and Quasi-Experimental Designs for Research. Rand-McNally, Chicago.

Carden, Fred. 1999. *Evaluating Governance Programs: Report of a workshop*. IDRC, Ottawa. http://www.idrc.ca/evaluation/governance.htm

Casanaova, Pablo G. 1981. The Fallacy of Social Science Research: A Critical Examination and New Qualitative Model. Pergamon Press, New York.

Cassen, R. 1986. Does Aid Work? Oxford University Press, Oxford.

Chelimsky, Eleanor and William R. Shadish (eds.). 1997. Evaluation for the 21st Century: A Handbook. Sage Publications, California.

CIDA. January, 1999. *Results-Based Management in CIDA: An Introductory Guide to the Concepts and Principles*. Canadian International Development Agency, Government of Canada.

Clarke, Alan with Ruth Dawson. 1999. Evaluation Research: An Introduction to Principles, Methods and Practice. Sage Publications, London, UK.

Cook, Thomas, Jerry Vansant, Leslie Stewart, and Jamie Adrian. 1995. *Performance Measurement: Lessons Learned for Development Management*. World Development. Vol.23(8):1303-1315. Elsevier Science Ltd., Great Britain.

Cracknell, Basil Edward. 2000. Evaluating Development Aid: Issue, Problems and Solutions. Sage Publications, New Delhi.

Crawford, Gordon, with Iain Kearton. 2002. Evaluating Democracy and Governance Assistance. Centre for Development Studies, University of Leeds.

Creswell, John. 1994. Research Design: Qualitative and Quantitative Approaches. Sage Publications, California.

Dale, Reidar. 1998. Evaluation Frameworks for Development Programmes and Projects. Sage Publications, New Delhi.

Datta, Lois-ellin. 1997. *Multimethod Evaluations: Using Case Studies Together with Other Methods*. In Chelimsky, Eleanor and William R. Shadish (eds.). Evaluation for the 21st Century: A Handbook. Sage Publications, California, USA.

Davidson, E.J. 2000. *Ascertaining causation in theory-based evaluation.* In P.J. Rogers, T.A. Hacsi, A. Petrosino, and T.A. Huebner (eds.), "Program theory in evaluation: challenges and opportunities". New Directions in Evaluation, Number 87:17-26, San Francisco, CA.

Davies, Ian. 1999. *Evaluation and Performance Management in Government*. Evaluation. Vol.5(2): 150-159. Sage Publications, London, UK.

den Heyer, Molly. February, 2001. Literature Review From: The Development of a Temporal Logic Model. Master's Thesis, University of Guelph.

Denzin, Norman K. 1978. The Research Act: A Theoretical Introduction to Sociological Methods. McGraw-Hill Books, New York.

Denzin, N. and Y. Lincoln. 1994. Handbook for Qualitative Research. Sage Publications, California.

Devuyst, Dimitri and Luc Hens. 2000. *Introducing and Measuring Sustainable Development Initiatives by Local Authorities in Canada and Flanders (Belgium): A Comparative Study*. Environment, Development and Sustainability. Vol.2:81-105. Kluwer Academic Publishers, Netherlands.

Diem, Keith G. *Choosing Appropriate Research Methods to Evaluate Educational Programs*. Rutgers Cooperative Extension, Rutgers University.

Diesing, Paul. 1992. How Does Social Science Work? Reflections of Practice. University of Pittsburgh Press, PA.

Dixon, Thomas Homer. 1996. *Strategies for Studying Causation in Complex Ecological-Political Systems*. The Journal of Environment and Development: A Review of International Development Policy. Vol.5(2):132-148.

Durkheim, Emile. 1897/1951. Suicide. Free Press, New York.

Earl, Sarah, Fred Carden and Terry Smutylo. 2001. Outcome Mapping: Building Learning and Reflection into Development Programs. International Development Research Centre, Ottawa.

Earl, Sarah, Fred Carden and Terry Smutylo. 2001. *Outcome Mapping: The Challenges of Assessing Development Impacts*. International Development Research Centre, Ottawa.

Edington, Juliet. 2001. *Logical? Monitoringagainst Logical Frameworks*. IA Exchanges. ActionAid, London.

Estrella, Marisol with Jutta Blauert, Dindo Campilan, John Gaventa, Julian Gonsales, Irene Guijt, Deb Johnson, and Roger Ricafort (eds.). 2000. Learning from Change: Issues and Experiences in Participatory Monitoring and Evaluation. Intermediate Technology Publications Ltd., London, UK.

Elzinga, Aant. 1981. Evaluating the Evaluation Game: On the Methodology of Project Evaluation, with Special Reference to Development Cooperation. SAREC Report.

Fetterman, David M. 2001. Foundations of Empowerment Evaluation. Sage Publications, California.

Feuerstein, 1986. Partners in Evaluation. MacMillan, London, UK.

Finsterbusch, Kurt, Jasper Ingersoll and Lynn Llewellyn (eds.). 1990. Methods for Social Analysis in Developing Countries. Westview Press, Colorado.

Fitz-Gobbon, Carol Taylor. 2002. *Evaluation in an Age of Indicators: Challenges for*

Public Sector Management. Evaluation. Vol.8(1):140-148. Sage Publications, London, UK.

Gasper, Des. 2000. *Evaluating the 'Logical Framework Approach' Towards Learning-Oriented Development Evaluation*. Public Administration and Development. Vol.20:17-28. John Wiley and Sons, Ltd.

Glaser, B.G., and A. Strauss. 1967. The Discovery of Grounded Theory. Aldine, Chicago.

Goldthorpe, John H. 2001. *Causation, Statistics, and Sociology*. European Sociological Review. Vol.17(1):1-20. Oxford University Press, UK.

Greene, Jennifer C. 1999. *The Inequality of Performance Measurements*. Evaluation. Vol.5(2):160-172. Sage Publication, London, UK.

Greene, Jennifer C. 2001. *Evaluation Extrapolations.* American Journal of Evaluation. Vol.22(3):397-402.

Gribbons, Barry and Joan Herman. 1997. *True and Quasi-Experimental Designs*. ERIC Clearinghouse on Assessment and Evaluation, Washington, DC.

Guba, Egon G., Yvonna S. Lincoln. 1981. Effective Evaluation: Improving the Usefulness of Evaluation Results Through Responsive and Naturalistic Approaches. Jossey-Bass Publishers, San Francisco.

Guilmette, Jean-H. 1997. *The case for a new ethic of evaluation*.

Hansson, Finn. 1997. *Critical Comments of Evaluation Research in Denmark.* In Chelimsky, Eleanor and William R. Shadish (eds.). Evaluation for the 21st Century: A Handbook. Sage Publications, California, USA.

Henry, Gary T., George Julnes, and Melvin M. Mark. 1998. *A Realist Theory of Evaluation Practice*. Realist Evaluation: An Emerging Theory in Support of Practice. Jossey-Bass Publishers, San Francisco.

Holland, P. 1986. *Statistics and Causal* Inferenc. Journal of the American Statistical Association. Vol.81(945-960).

Horton, Douglas and Ronald MacKay (eds.). 1999. *Evaluation in Developing Countries: Experience with Agricultural Research and Development*. Knowledeg, Techonology and Policy. Vol.11(4), Rutgers University.

House, Ernest R. 2001. *Unfinished Business: Causes and Values*. American Journal of Evaluation. Vol.22(3):309-315.

HRDC. 1998. Quasi Experimental Evaluation. www11.hrdc-drhc.gc.ca/pls/edd/QEE_

78014.htm

Hughes, Mike, and Tish Traynor. *Reconciling Process and Outcome in Evaluating Community Initiatives*. Evaluation. Vol.6(1):37-49.

Jantsch, Erich ed. 1981. The Evolutionary Vision: Toward a Unifying Paradigm of Physical, Biological, and Sociocultural Evolution. Westview Press, Inc., Colorado.

Julian, David. 1997. *The Utilizing of the Logic Model as a System Level Planning and Evaluation Device*. Evaluation and Program Planning. Vol.20(3):251-257. Elsevier Science Ltd., Great Britain.

Kabeer, Naila. 1999. The Conditions and Consequences of Choice: Reflections on the Measurement of Women's Empowerment. UNRISD.

Kaminsky, A. 2000. Evaluation and Program Planning. Vol.23:333-335, Elsevier Science Ltd., Great Britain.

Keat, R. and J. Urry. 1975. Social Theory as Science. Routledge and Keagan Paul, London.

Kelly, Linda. 2002. Research and Advocacy for Policy Change: Measuring Progress. The Foundation for Development Cooperation.

Kuchler, Manfred. 1981. *Causal Analysis in Nonexperimental Evaluation Research*. Evaluation Research and Practice: Comparative and International Perspectives. Sage Publications, California.

Lipsey, Mark W. 2001 *Re: Unsolved Problems and Unfinished Business*. American Journal of Evaluation. Vol.22(3):325-328.

Lofland, John and Lyn H. Lofland. 1995. Analyzing Social Settings. Wadsworth Publishing Company, University of California, Davis.

Lusthaus, Charles, Marie-Helene Adrien, Gary Anderson, and Fred Carden. 2000. Enhancing Organizational Performance: A Toolbox for Self-Assessment. Vikas Publishing House PVT Ltd., New Delhi.

Luukkonen, T. 1998. *The difficulties in assessing the impact of EU framework programmes*. Research Policy. Vol.27:599-610.

Mark, Melvin M. 2001. *Evaluation's Future: Furor, Futile, Or Fertile?* American Journal of Evaluation. Vol.22(3):457-480.

Mason, Greg. 1991. *Longitudinal Research in Program Evaluation*. In Evaluation

Methods Sourcebook. Canadian Evaluation Society.

Mayne, John. 1999. *Addressing Attribution Through Contribution Analysis: Using Performance Measures Sensibly*. Discussion paper, Office of the Auditor General of Canada.

Mayne, John. 2002. Communication between John Mayne and Burt Perrin, EVALTALK. American Evaluation Association listserv.

Maxwell, J.A. 1996. *Using qualitative research to develop causal explanations.* Working Paper, Harvard Project of Schooling and Children. Cambridege, MA.

McDonald, Diane. 1999. *Developing guidelines to enhance the evaluation of overseas development projects*. Evaluation and Program Planning. Vol.22:163-174.

Mohr, Lawrence B. 1999. *The Qualitative Method of Impact Analysis*. American Journal of Evaluation. Vol.20(1):69-84.

Murphy, Josette. 1997. *Tracing Gender Issues Through Institutional Change and Program Implementation at the World Bank*. In Chelimsky, Eleanor and William R. Shadish (eds.). Evaluation for the 21st Century: A Handbook. Sage Publications, California, USA.

Newcomer, Kathryn E. 2001. *Tracking and Probing Program Performance: Fruitful Path or Blind Alley for Evaluation Professionals?* American Journal of Evaluation. Vol.22(3):337-342.

Oakley, Peter, Brian Pratt and Andrew Clayton. 1998. Outcomes and Impact: Evaluating Change in Social Development. INTRAC Publication, UK.

OECD. 1986. Methods and Procedures in Aid Evaluation, OECD, Paris.

OECD. 1992. DAC Principle for Effective Aid. Paris.

Pasteur, Kath. 2001. *Thinking about Logical Frameworks and Sustainable Livelihoods: A short critique and a possible way forward.*

Patton, Michael Quinn. 1997. Utilization-Focused Evaluation. Sage Publications, California.

Patton, Michael Quinn. 2002. Qualitative Research and Evaluation Methods. Sage Publications, California.

Pawson, Ray, and Nick Tilley. 1997. *An Introduction to Scientific Realist Evaluation*. In Chelimsky, Eleanor and William R. Shadish (eds.). Evaluation for the 21st Century: A Handbook. Sage Publications, California, USA.

Pawson, Ray, and Nick Tilley. 1997. Realistic Evaluation. Sage Publications, London, UK.

Pawson, Ray, and Nick Tilley. 2001. *Realistic Evaluation Bloodlines*. American Journal of Evaluation. Vol.22(3):317-324.

Perrin, Burt. 1998. *Effective Use and Misuse of Performance Measurement*. American Journal of Evaluation. Vol.19(1):367-379.

Perrin, Burt. 1999. *Performance Measurement: Does the Reality Match the Rhetoric? A Rejoinder to Bernstein and Winston*. American Journal of Evaluation. Vol.20(1).

Poole, Denis L., Joan Nelson, Sharon Carnahan, Nancy G. Chepenik, and Christine Tubiak. 2000. *Evaluating Performance Measurement Systems in Nonprofit Agencies: The Program Accountability Quality Scale (PAQS)*. American Journal of Evaluation. Vol.21(1):15-26.

Ragin, Charles C. 1994. Constructing Social Research. Pine Forge Press, California.

Ragin, Charles. 1997. *Turning the Tables: How Case-Oriented Research Challenges Variable-Oriented Research*. Comparative Social Research. Vol.16:27-42.

Riddell, R.C. 1987. Foreign Aid Reconsidered. ODI/James Curry, London.

Rebien, C. 1996. . Evaluating Development Assistance in Theory and Practice. Avebury, UK.

Reynolds, Arthur J. 1998. *Confirmatory Program Evaluation: A Method for Strengthening Causal Inference*. American Journal of Evaluation. Vol.19(2):203-221.

Roche, Chris. 1999. Impact Assessment for Development Agencies: Learning to Value Change. Oxfam, United Kingdom.

Russell, B. 1913, *On the Notion of Cause*. Proceedings of the Aristotelian Society, Vol.13(1-26).

Ryan, James G. 2001. Synthesis Report of Workshop on Assessing the Impact of Policy-Oriented Social Science Research in Sheveningen, The Netherlands – November 12 13, 2001.

Sang, Heng-Kang. 1995. Project Evaluation: Techniques and Practices for Developing Countries. Ashgate Publishing Limited England.

Sanyal, Bishwapriya. 1996. *Intention and Outcome: Formalization and its Consequences*.

Regional Development Dialogue. Vol.17(1):161-183.

Scriven, Michael. 1991. Evaluation Thesaurus. Fourth Edition, Sage Publications, California.

Shaw, Ian. 1999. Qualitative Evaluation. Sage Publication.

Smillie, Ian. 2001. *The Forest and the Trees: Capacity-Building, Results-Based Management and the Pakistan environment Program*. Executive Summary.

Smith, M.F. 2001. *Evaluation: Preview of the Future #2*. American Journal of Evaluation. Vol.22(3):281-300.

Smutylo, Terry. 2001. Crouching Impact, Hidden Attribution: Overcoming Threats to Learning in Development Programs. International Development Research Centre, Ottawa.

Stake, Robert E. 1995. The Art of Case Study Research. Sage Publication, California.

Stake, Robert E. 2001. *A Problematic Heading*. American Journal of Evaluation. Vlo.22(3):349-354.

Stern, Paul C. 1979. Evaluating Social Science Research. Oxford university Press, New York.

Strauss, Anselm L. 1987. Qualitative Analysis for Social Scientists. Press Syndicate, New York.

Suchman, E.A. 1967. Evaluative Research. New York, NY, Russell Sage.

Taylor-Powell, Ellen. 1996. *Analyzing Qualitative Data*. Program Development and Evaluation, University of Wisconsin.

Toffolon-Weiss, Melissa M., Jane T. Bertrand, and Stanley Terrell. 1999. *The results framework-and innovative tool for program planning and evaluation*. Evaluation Review. Vol.23(3):336-359.

Torjam, Sherri. 1999. *Are Outcomes the Best Outcome?* Caledon Institute of Social Policy.

Trochim, William M.K. 1989. *Outcome Pattern Matching and Program Theory*. Evaluation and Program Planning. Vol.12:355-366.

UNDP. *Tracking Human Development Prgress.* www.thdp.unpd.kg/meglogg.html.

USAID. 2000. *Performance Monitoring and Evaluation: TIPS.*

Valadez, Joseph J. and Michael Bamberger. 1994. Monitoring and Evaluating Social Programs in Developing Countries: A Handbook for Policymakers, Managers, and Researchers. Economic Development Institute, Washington.

Vlaenderen, H. Van. 2001. *Evaluating development programs: building joint activity*. Evaluation and Program Planning. Vol.24:343-352.

W.K Kellogg Foundation Evaluation Handbook. 1998. Michigan.

Waldick, Lisa. 2002. *In Conversation: Michael Quinn Patton*. International Development Research Centre, website article.

Wallerstein, Immanuel. 1974. The Modern World-System. Academic Press Inc., New York.

Whyte, Anne. 2000. Assessing Community Telecentres: Guidelines for Researchers. Acadia Initiative of IDRC.

Wiggins, Steve, and Dermot Shields. 1995. *Clarifying the 'logical framework' as a tool for planning and managing development projects*. Project Appraisal. Vol.10(1):2-12. Beech Tree Publishing, Surrey, UK.

Wolfe, Marshall. 1996. Elusive Development. Zed Books Ltd., Geneva.

World Bank. 1998. *Assessing Aid: What Work, What Doesn't, and Why?* Oxford University Press, UK.

Worthen, Blaine R. 2001. *Whither Evaluation? That All Depends*. American journal of Evaluation. Vol.22(3):409-418.