Advancing Evaluation of Character Building Programs

William M. Trochim

Cornell University

Jennifer Brown Urban

Montclair State University

Advancing Evaluation of Character Building Programs

Perhaps it is best to start out by describing what this paper is not going to do. It will not provide an overview of evaluation, describe major approaches, delineate the broad range of available methodologies or provide examples of evaluations relevant to the topic of character development (CD) programs. Put more succinctly, we do not intend to characterize how evaluation is done. The mainstream literature on evaluation addresses all of these topics far better than we could hope to do here. We assume that either the reader is familiar with that literature or, if interest so motivates, will be able to avail themselves of it at any point.

Instead this paper hopes to make a more general argument about how the endeavor of character development might think about evaluation, how it fits into their work, and what needs to happen in order to sustain it. Along the way we hope to introduce the reader to a broader view of evaluation, not just as something that is applied to a program or at a program level but as an endeavor that needs to be integral to the ecologies and systems within which programs occur. This macro-level view is driven by several tectonic shifts in high-level thinking in evaluation (and many other fields) over the past few decades associated with the rise of ecological and systems thinking, the integration of evolutionary theory into applied social research methodology, and the rise of integrated global computer technology in the form of the internet. These forces are moving us away from the traditional focus of evaluation at the program level and situating traditional program evaluation within a longer-term evolution of program theory and practice. It is at this macro evolutionary level, rather than at the within-program level, that we wish to focus this paper and the field of character development.

In Section I we begin at the most familiar level, briefly describing the landscape of the evaluation of youth-serving and character building programs. Section II introduces a number of

foundational issues that characterize the shifting paradigm for contemporary evaluation. Section III discusses the implications of these issues for character development evaluation both at the program and organizational or systems level.

Our hope for this paper is that it will move the discussion about evaluation, if even slightly, to a higher and more systemic level in the field of character development, move us away from the focus on the evaluation of programs and towards a more evolutionary evaluation perspective focused on evidence-driven program theory development and the assessment of theoretically described portfolios of ideas about positive youth development toward what we hope will be a deeper understanding of character development and how it might be influenced and encouraged.

I. History of Evaluation in PYD and Character Building Programs

When we survey the landscape of youth-serving programs, including both in-school and out-of-school programs, we can readily see that many are designed to promote character attributes. However, the vast majority of these programs have not been formally evaluated and we have little idea of whether and how they work (Roth & Brooks-Gunn, 2003, 2015; Roth, Brooks-Gunn, Murray, & Foster, 1998). Character development for youth is increasingly being recognized as critically important; however, our ability to evaluate and bring effective programs to scale has not kept pace. Early work in youth/character development was focused on shifting the focus from a deficit to a strengths-based perspective. At the turn of the 21st century, the focus shifted to advancing theory in youth/character development and providing preliminary evidence that youth-serving programs are indeed beneficial (Roth & Brooks-Gunn, 2015). We are now ready to embark on the next phase of research and evaluation of youth-serving programs, and this must include detailed and specific articulation of program theory, more careful attention to

definitions and measurement, and most importantly, investment of time and resources in the planning, implementation and utilization of evaluations of such programs.

It is important to recognize that many youth-serving programs are small, local programs (e.g., sports teams, school newspaper) that are locally sponsored, while other youth-serving programs are affiliates of national organizations such as Boy/Girl Scouts or 4-H (Roth & Brooks-Gunn, 2003). Given a relatively recent focus on evaluation and the high turn-over of program staff, many of these programs lack the time and financial resources as well as the skill-set to engage in high quality evaluation and planning. As a result, many youth-serving programs, rely on inappropriate data or lack data to demonstrate the programs' effectiveness at promoting character. In order to increase the quality of character development programs and, in turn, increase the positive impact of these programs on youth, the field of character development must: (1) advance at the organizational level by fully integrating evaluation throughout the organizational culture; and, (2) advance evaluation practice at the programmatic level by adopting an Evolutionary Evaluation perspective.

II. Foundational Issues

In this section, we discuss three broad foundational theoretical perspectives that characterize the shifting paradigm in contemporary evaluation: the incorporation of evolutionary theory into evaluation; the introduction of relational developmental systems theory; and the idea of systems thinking in evaluation. While each represents different traditions and histories, they are most effectively considered in concert as part of the new ways to think about what evaluation is, how it can best be accomplished and how it integrates into the field of CD to advance both knowledge and practice.

Evolutionary Evaluation

In this paper we would like to extend our earlier efforts (J. B. Urban, Hargraves, & Trochim, 2014) to encourage evaluation based on a foundation of evolutionary theory and the idea of natural selection (Darwin, 1859; Mayr, 2001) and especially on the idea of evolutionary epistemology (Michael Bradie & William Harms, 2006; Campbell, 1974, 1988; Cziko & Campbell, 1990; Popper, 1973a, 1985). The central thrust of this line of research is that our knowledge, including our macro-level knowledge of CD interventions and programs, evolves according to the principles of natural selection, the trial-and-error cycle of (blind) variation and selective retention. Over time, program variations are tried and survive or not according to current socially negotiated selection mechanisms. In the modern era, evaluation has increasingly assumed an important role in both the generation of project variations through formative approaches like conceptualization and program theory methods (Caracelli, 1989; Chen & Rossi, 1990; Kane & Trochim, 2006; Linstone & Turoff, 1975; Trochim, 1989a; Trochim & Kane, 2005; Trochim & Linton, 1986) and in the retention of variants that perform well through summative approaches like experiments or outcome assessments (Campbell, 1969; Campbell & Stanley, 1963; T. D. Cook & Campbell, 1979; W. Shadish, Cook, & Campbell, 2002). This macro-level evolutionary theory (phylogeny) is accompanied by program specific ontogeny: like organisms, each program has its own lifecycle and tends to proceed through different phases or stages.

We can see these evolutionary impulses manifested in the model from clinical medicine of phased clinical trials (Chow & Liu, 2004; Pocock, 2004). This model assumes that new interventions progress through different lifecycle phases (ontogeny) of maturation and testing. Typically, Phase I clinical trials are small sample exploratory studies that look at dosage levels,

implementation difficulties and potential side-effects of the treatment. Phase II trials concentrate on whether the treatment yields measurable changes and examine relationships among observed outcomes. It is in the late Phase II or Phase III trials that more formal controlled designs like the randomized experiment are employed. Phase IV trials are concerned with fidelity of transfer and implementation of treatments in uncontrolled real-world contexts. At a macro level (phylogeny) it is assumed following natural selection models of evolution that most interventions do not survive through this lifecycle. In fact, in biomedical research nearly three-fourths of all treatments are abandoned before a Phase III randomized experiment is ever mounted (Mayo Clinic, 2007).

We also see macro-level evolutionary thinking in the literatures on meta-analysis, many of which originated in medicine, (Hasselblad, 1998; Koretz, 2002; Smith & Glass, 1977; Sohn, 1996; Wolf, 1986), research syntheses and systematic reviews (Chalmers & Haynes, 1994; D. J. Cook, Mulrow, & Haynes, 1997; Davies, 2000; Dickersin, 2002; Dickersin, Scherer, & Lefebvre, 1994; Evans & Benefield, 2001; Gallagher, 1999; Juni, Altman, & Egger, 2001; Silagy, Middleton, & Hopewell, 2002; Silagy, Stead, & Lancaster, 2001; Wortman, Smyth, Langenbrunner, & Yeaton, 1998) and evidence-based practice (Antes, 1998; Artelt, 2001; Brownson, Baker, Leet, & Gillespie, 2002; Cronje & Fullan, 2003; Davies, 1999; Dopson, Locock, Gabbay, Ferlie, & Fitzgerald, 2003; Eitel & Steiner, 1999; Gibbs, 2003; McAlister, Graham, Karr, & Laupacis, 1999; Montori & Guyatt, 2002; Pirrie, 2001; Procopis, 2002; Sackett, Richardson, Rosenberg, & Haynes, 1997). In all of these, the assumption is that most programs (organisms) will not survive over the long run and that the trial-and-error process will yield a residual evidence base of those that survive the increasingly stringent natural selection mechanisms of evaluation and experimentation.

This fundamental evolutionary notion and its manifestation in clinical medicine has become a powerful model in contemporary science and has influenced evaluation generally. For instance, in educational evaluation we see it clearly evidenced in the following:

> "There is a critical pathway for the development of successful educational interventions and activities, starting generally with small-scale studies to test new ideas and generate hypotheses, leading to increasingly larger and more rigorous studies to test the effect of a given intervention or activity on a variety of students and in a variety of settings. Different research methodologies are used along the development pathway, and corresponding evaluation strategies must be used to assess their progress." (p. 13) (U.S. Department of Education, 2007)

The Variation and Selection Role of Evaluation

The fact that not all programs will survive is part of the natural cycle-of-life. Over time, programs and their theories evolve. Evaluation can play a key role in that artificial selection, both in encouraging and enhancing variability and in providing feedback and influencing selection. As in evolution generally, it is not clear where program evolution is heading or whether any adaptations can be said to constitute 'progress.' Slight variations and adaptations can survive that subsequently make little apparent sense. Program features may exist today that were adaptive in the past but remain largely as residual, long beyond their original adaptive genesis.

Natural selection involves several concepts, most notably the ideas of variation and selection. Popper maintains that variation in nature and in science is "at least partly random." Campbell describes natural selection using the term "blind" in his phrase "blind variation and selective retention" (BVSR) in order to convey the notion that there is no certain a priori

knowledge that a variation will be adaptive. According to Campbell (1988), BVSR encompasses three essentials: "(a) mechanisms for introducing variation; (b) consistent selection processes; and (c) mechanisms for preserving and/or propagating the selected variations" (pp. 402). Campbell (1988) states: "In going beyond what is already known, one cannot but go blindly. If one can go wisely, this indicates already achieved wisdom of some general sort." However, in the world of programs and program theory, the process of creating new programs or program variations does not feel blind or random. We tend to think of new programs as rational, sometimes obvious, responses to problems or challenges in our environment, not as blind or random endeavors.

The notion of "blind variation" does not mean that our attempts at program construction are irrational; it simply means that when we construct programs we do not yet know whether they will work or survive. Evaluation certainly plays a role in the program variation generation process. We conduct needs assessments (Witkin, 1995) to identify important problems or issues. We construct program theories using methods such as concept mapping (Trochim, 1989a) and logic modeling (L. J. Cooksy, Gill, & Kelly, 2001; Foundation, December 2001; J. A. Frechtling, 2007). We assist planners and program developers in thinking through program implementation challenges (Bammer, 2005; King, Morris, & Fitz-Gibbon, 1987) and in constructing approaches for formative and process evaluation. All of these can be considered activities that help to generate or describe program variation.

The other major component of natural selection is the selective retention of better adapted variants. This is the function that was described by Herbert Spencer with the infamous phrase "survival of the fittest" that conjures up the notion of intense competition often associated with evolution in popular culture. Today, we have a much more nuanced sense of what the term

"fitness" means and how characteristics are selected for over time. Margulis and Sagan describe this well when they say: "…the view of evolution as chronic bloody competition among individuals and species, a popular distortion of Darwin's notion of 'survival of the fittest,' dissolves before a new view of continual cooperation, strong interaction, and mutual dependence among life forms. Life did not take over the globe by combat, but by networking. Life forms multiplied and complexified by co-opting others, not just by killing them" (Margulis & Sagan, 1986, pp. 28-29).

Ontogeny: The Program Evolutionary Lifecourse

An evolutionary systems evaluation perspective leads us to think differently about programs. For instance, the idea of ontogeny in evolutionary theory is concerned with the origin and the development of an organism from inception through maturity. Instead of thinking of our programs as static entities, this notion encourages us to think of each program as continuously evolving through different phases in a lifecycle, much like any organism does. While this lifecycle will manifest itself differently for each program, much as different people develop at different rates at various times in their lives, we can sketch out a hypothetical sequence that would likely fit many programs.

The first stage in this hypothetical program lifecycle might be termed the initiation phase and spans from the initial conceptualization through design and initial piloting. The first time you implement a program – even one that was well established elsewhere – you will likely be dealing with the usual issues of initiation: identifying and training program staff, localizing the program to the immediate context, reacting to the unanticipated problems that arise, and so on. The second lifecycle stage might be labeled the growth or development phase and would encompass the process of successive revisions of the program as it gets implemented repeatedly

over time. With programs in this stage, you are getting used to the program and how it plays out in practice. You have some idea of what to anticipate. You might still be surprised by what occurs, and are still adapting the program as you learn, but you are also increasingly able to anticipate problems before they arise, and you are developing a storehouse of experience in how to deal with them. In the third stage, which we might term the maturity or stability phase, the program has been carried out at least several times with some degree of implementation success. By this point you pretty much know what to expect. You have stable processes for implementation, and the program is routinized and often has a written protocol or process guide. The program is no longer dependent on particular individuals for implementation. If something happens to the initial implementers of the program it can still be carried out with high fidelity. The fourth stage might be termed the translation or dissemination phase. The primary focus of this phase is on extending the program to other settings or populations of interest, pushing the ecological boundaries of the program as originally conceived into new niches or applications.

These stages aren't meant to be a straight-jacket or an inflexible taxonomy. For any given program, the progression may not be sequential. A program may be precocious. It may for instance quickly evolve through the development phase and become stabilized or routinized. Or, a program can revert to an earlier stage, much like the young adult that temporarily reverts to juvenile behavior before resuming more mature development. At any phase, we may decide whether to continue the program or not. Sometimes it is apparent even early in a program's development that it is not able to be implemented well or that it has a fundamental flaw in its conception or structure. In the best of all worlds, probably only a small minority of programs will or should survive to the translation or dissemination phase.

This notion of a program lifecycle has practical implications for evaluation. How should a program be evaluated at each stage of its lifecycle? What role can evaluators play in helping program administrators and organizations assess where their programs are in their development and in encouraging them to think about when and how they will evolve their programs to their next phase?

In many of our program contexts, we become committed to the program as it currently exists. The program evolves up to a point and then we get a type of "lock-in" where we seemingly get stuck in a phase and are unable to move any further. Program decisions turn into a struggle between program preservationists who fear change and the potential loss of their familiar context or even their jobs, and program critics who push for ever-extending demonstrable results and emphasize ever-shrinking funding and resources. An evolutionary perspective on programs and the idea of ontogeny emphasize program change as something to be expected and embraced. Instead of the commitment to preserving the program as it is, it encourages the idea that programs have a limited life-span, that they should not be assumed to live forever, that it is normal to see them as part of an ongoing trial-and-error learning process, and that the abandonment of an older program and the development of new ones is part of the normal cycle-of-life. From the beginning of the program, and throughout its evolution, the focus is on where the program is in its development and how to move it to the next phase. In effect, the idea of a lifecycle creates system pressure to move programs along and not allow them to become static.

Symbiosis and Co-Evolution

The ideas of symbiosis and co-evolution are critically important in evolutionary biology and should be considered equally important in program development and evaluation. One of the

most familiar examples of this phenomenon is the relationship of the flower and the bee. Each provides something to the other. The flower provides nectar that is produced into honey and the bee acts as the vehicle for sexual reproduction by moving pollen from one flower to another. Both benefit from the exchange. Neither participates in this exchange consciously. Flowers did not strategize one day that they needed bees as a vehicle for reproduction. And bees did not decide that flowers would be good vessels for honey production. They co-evolved over millennia in a manner that makes them co-dependent.

There are several ways that symbiosis and co-evolution are important for evaluation. First, if all programs evolve through different stages over time, then we must recognize that the evaluation approaches we use at each stage need to differ throughout the life of the program. That is, the way we would evaluate a program during its initiation stage would not likely be appropriate for evaluating it during its growth stage, and so on. In effect, the evaluation of a program has its own lifecycle and one of the major tasks of Evolutionary Evaluation is to encourage the symbiotic or co-evolutionary relationship between program and evaluation lifecycles. In the initiation phase, an evaluation needs to be dynamic and flexible, providing rapid feedback about implementation and process. In many program evaluations, this is accomplished with simple monitoring or post-only feedback forms, unstructured observation, qualitative methods, informal debriefing and feedback, and through communications systems. In the development phase of an evaluation, the focus tends to shift to the observation and assessment of change and we focus on things like designing observational procedures and measures of key outcomes, assessing the consistency and construct validity of measures, looking at pre-post differences and examining the relationships among different observations using qualitative and/or quantitative methods. The mature phase of an evaluation tends to emphasize

the idea of control. At this point the program is routinized and stable enough to compare performance of participants with some standard expectation of performance or with outcomes of people who participate in alternative programs or none at all. This is the realm of experimental and quasi-experimental designs and of more structured and comparative qualitative approaches. The translation or dissemination phase in evaluation is typically concerned with adaptation, implementation, sustainability, and generalizability or external validity. It examines the consistency of outcomes across different settings, populations or program variations. This is the realm of secondary and meta-analysis and of program review approaches that seek general inferences about the transferability and adaptability of the program. Encouraging a symbiotic relationship between the evaluation approach and the program lifecycle is a critically important Evolutionary Evaluation process.

Second, the ideas of symbiosis and co-evolution also have important practical implications for the level of support people have for evaluation. In many evaluation contexts, one hears a series of laments about how unmotivated people are to evaluate or their resistance to doing evaluation. For instance, the evaluator asks “Why don’t these program people just cooperate when I ask them for data?” Program practitioners ask “Why don’t these evaluations address something that would be useful for us?” Program participants want to know “Why do they keep bugging us for data? We don’t get anything from this.” Ideally, we would want the situation to be co-evolutionary where program participants are providing information naturally as part of their participation, where program administrators are getting what they want from the provided data, and where evaluation happens almost transparently as an integrated aspect of program implementation. That is, the ideal is the flower and the bee. This is a difficult ideal to achieve in practice. It requires that evaluation systems be engineered in such a way that each

stakeholder group's incentive to participate in the evaluation is well understood and integrated into the approach.

Phylogeny: The Program Family Tree

Phylogeny refers to the macro-level evolution of the species. When we look at the phylogeny of a species we are in effect examining its history, the story of the organism variations that survived and those that did not, the species' family tree.

Every program can be viewed as an organism in a population of similar programs that constitutes its species. Program theories, whether stated explicitly or not, make up the essential instructions of the program. Programs have variations within each species of program. Programs have unique characteristics: the people who implement them, the activities that constitute them, the setting and assumptions that guide them, the participants who take part in them. This program variation is essential for their evolution.

Program variations are implemented, have consequences, and are selected for in subsequent program generations. Some programs and their characteristics and theories survive over time; most become extinct. Programs and program theories get selected and survive because of the fitness of their characteristics to a specific environmental or ecological niche. While most of us probably hope or believe that programs are selected for using rational criteria to yield specific desirable characteristics or outcomes, in many situations they probably survive because people like them, get used to them, or because there are institutional, political and economic forces that favor their survival.

There is an integral connection of ontogeny and phylogeny in evolutionary theory and in its evaluation analogue. Just as with other organisms in nature, in addition to their participation in a broader species, each program has its own individual life (ontogeny), a unique life course

that moves through various phases. Individual programs are born or initiated. They grow and change as they are implemented and revised. They mature and may reach a relatively stable state sometimes becoming routinized and standardized. And, they regenerate, die, are translated and disseminated, and so on, starting new cycles of program instances. And it is this churning of many program variations over time that leads to the selective retention of those programs with a fitness or adaptability to their environment that makes up the family tree or history (phylogeny) of that program species.

Relational Developmental Systems Theory

Relational Developmental Systems (RDS) meta-theory, is a leading theory of developmental science and many, if not most, character development programs theoretical underpinnings are consistent with this perspective. RDS can rightfully be considered part of a broader systems perspective that holds that development is the result of bidirectional individual ←→ context coactions across multiple levels (i.e., from cells to society; e.g., Gottlieb, Wahlsten, & Lickliter, 2006; Lerner, 2006; W.F. Overton, 2006; W.F. Overton, 2010). When these coactions are mutually beneficial, they are referred to as adaptive developmental regulations (Brändtstadter, 1999). It is important to note that development is not viewed as unidirectional, but rather as a bidirectional process whereby an organism shapes and is shaped by the environment. RDS rejects a split-reductionist view of development that sees genes as the primary engine of development. Organisms are viewed as autopoetic, active agents that are self-creating, self-organizing, and self-regulating which, in turn, produces greater complexity and the potential for plasticity (Lerner, 2006; W. F. Overton, 2013). Relative plasticity highlights the potential for adaptive development across the life span within certain constraints.

Although RDS meta-theory describes human development, many of the same concepts can be applied to program development. Program theory is akin to DNA; but, just as human development is not the result of genes alone, program theory (program DNA) alone does not account for what actually happens when a program is implemented. The interaction of program theory (what was intended) with the context in which the program is situated produces growth and change in the program. When those changes are positive, this process is akin to adaptive regulations.

Therefore, it should come as no surprise that practitioners often make changes to evidence-based programs in order to increase the chances of a mutually beneficial relationship between the program and the local environment. It should also come as no surprise that a program will likely evolve and change due to the influences of the context in which it is situated. The question for evaluation is whether these adaptations are indeed adaptive. Therefore, the question we really should be asking is what fundamental attributes of which programs (e.g., what programmatic features); addressing the needs of individuals of what status attributes (e.g., people at what portion of the life span, and of what sex, race, ethnic, religious social location); in relation to what characteristics of the context; are likely to be associated with what facets of adaptive functioning (adapted from the Bornstein Specificity Principle; Bornstein, 2006)?

Evaluation is inherently a part of Popper's adaptive behavioral learning. It can be considered one of a class of processes whereby we attempt to learn about the world around us. It is a mechanism for creating the conditions for observing the world and providing feedback about it and about the observation itself. Evaluation is a form of feedback, and as such is part of the selection process. In the view proposed here, programs and their theories are selected for over

time because they have characteristics that promote adaptive regulations (there is a mutually beneficial relation between the program and its environmental context).

System Thinking and Evaluation

One of the most striking characteristics of the evolution of evaluation in the past several decades is the increasing degree to which evaluation encompasses the idea of "systems." At the outset, we need to acknowledge the multiple ways that the term "system" is typically used. Standard dictionaries offer no less than 12 distinct definitions of "system," its principle one being "a regularly interacting or interdependent group of items forming a unified whole." Given the definitional challenges, we take a pragmatic and more direct approach to defining the territory. In our view, "systems" refers to parts, the whole, and their interrelationships. Systems are nested within other systems – systems constitute the parts of still larger systems. In this sense, virtually all evaluation already can be considered "systems" based – the complex endeavor of any evaluation, even the most defined and circumscribed program evaluation – is an interaction of many parts (myriad stakeholders, participants, measures, and organizing structures such as research designs, analytic procedures, and environmental and policy contexts) into an emergent whole that we call an evaluation. While this may be the case, we take the position that the idea of a "system" is a continuum ranging from simpler systems and components to more complex systems of systems. In this paper, when we use the term "system" we are not referring to traditional unitary evaluation projects or programs, but to multiple evaluations considered in concert, to multiple programs or initiatives, or both.

Over the past few decades, the process of evaluation has become increasingly complex. For instance, consider how conducting a traditional one-off evaluation of a program has evolved. In contrast to several decades ago, high-quality program evaluation today relies on the

integration of a complex array of perspectives and methods including, but not limited to: development and integration of program theory (Chen & Rossi, 1990) and program logic (Kellogg Foundation, 2001); use of a critical multiplism (W. R. Shadish, 1986) of samples, measures, program implementation and analyses; integration of qualitative and quantitative mixed methods (Greene & Caracelli, 1997); and the incorporation of a variety of stakeholders and participants (Fetterman, Kaftarian, & Wandersman, 1996; Macaulay, 1999; O'Fallon & Dearry, 2002; Reason & Bradbury, 2001). These demands for complexity in evaluation simply were not present, or were present in more incipient ill-defined form as recently as 1980. Consequently, there is a growing need to address integration of the complex activities and methods of even the simplest evaluation, and for tools and methods that can be used to design and implement evaluation systems.

The growth in complexity of evaluation processes has been matched by a commensurate growth in the complexity of what we are asked to evaluate. The emphasis on program evaluation has been expanded with increased calls for evaluating multiple programs, initiatives, portfolios, or directly, systems. The need to understand what works generally (rather than just in specific contexts or settings) has led to the rise of meta-analysis (Chalmers & Haynes, 1994; Lipsey & Wislson, 2000) and the desire to summarize research at higher levels of aggregation than just the individual program. Consider also the increased reporting and accounting needs of all levels of government - especially at the federal level – and in the not-for-profit sector with corollary increased demands for evidence and evaluation of broad programs and research initiatives (Brainard, 2002). Most of the pressures for accountability existed in at least incipient forms in earlier evaluation eras, but the ways we approach addressing these evaluation demands (e.g., through meta-analysis, research syntheses, performance management) have evolved significantly

in recent decades. All of these developments suggest that we are increasingly being called upon to conduct not just program evaluation, but to construct a more integrated systems evaluation. In this section we characterize how systems thinking relates to evaluation by presenting a number of metaphors that each describe a particular aspect of systems thinking. For each, we provide a short name for the metaphor and indicate parenthetically the systems thinking principle it is meant to represent.

Greater Than The Sum (Part-Whole Relationships)

Systems are by their very nature collections of multiple things, so it's not surprising that one of the most fundamental distinctions in systems theory is that of "part" and "whole." What do we mean by "part" and "whole" in a systems context? A part can be almost anything. For instance, it might be a piece in a machine – a wheel is a part of the larger whole that we call an automobile. Or it might be an organ in an organism – a heart is a part of a body. In most systems, part-whole relationships exist in nested hierarchies. For instance, a hubcap is a part of the whole wheel which is in turn a part of the automobile which might in turn be considered part of a fleet of vehicles. Or, a cell is a part of the heart which is part of the body which is in turn part of a class or group of organisms. But the part-whole distinction is more than just a physical one. We can also talk about part-whole hierarchies in concepts. For instance, our idea of the concept of humanity consists of parts like nationalities and sub-parts like people from different states or towns. Or, we might divide the whole concept of humanity into the parts of those who were born in different years, subgroups who are male and female, and subgroups of those who have brown hair or are right-handed. As these examples show, the part-whole concept is a universal one that can be applied to almost anything.

When thinking of parts and wholes it is also important to keep in mind that in addition to the whole and the parts that it is made up of, we also can think about the relationships between these as something that is distinguishable and meaningful (almost as if the relationships are separate "parts" of the part-whole distinction). For instance, in a car there is the whole of the car, its various parts (e.g., wheels, engine) and the relationships between these. This idea of the importance of relationships is central to systems thinking. It gives rise to the famous saying that "the whole is greater than the sum of its parts." We have to be careful about this saying, however. In mechanical systems, the whole is very often precisely the sum of its parts. You can take a car apart and reassemble it and it will work. However, in dynamic or living systems this is not the case – you cannot take a human body apart and then reassemble it and expect to have a working system. So, the phrase "the whole is greater than the sum" really is meant to refer to dynamic systems, not mechanical ones.

Part-whole relationships are everywhere in evaluation. For instance, an organization (whole) will often operate in multiple program areas (e.g., an educational or outreach organization might have programs for children, teens, adults and the elderly; or programs in health, education, environment, science, etc.); each program area (whole) might have multiple programs (parts); each program (whole) will usually consist of multiple activities (parts); each activity (whole) can typically be broken down into different tasks (parts); and so on. Or we can view different levels of part-whole hierarchies in terms of stakeholder groups. In an educational or outreach program we might think of stakeholders at the program level (participants, their families and program deliverers), the organizational level (program managers and organizational administrators), the local context (local officials or the local public), the funders level (e.g., private, state or national) and even the societal level (Congress or society as a whole). The idea

of part-whole relationships is essential in the development and implementation of programs. It is central to our description of the program, the development of program models, and the analysis of stakeholders and their interests.

The Rock and the Bird (Static and Dynamic Processes)

There is a parable in systems thinking that illustrates well the difference between static and dynamic processes. If you throw a rock into the air, you can predict with some accuracy where it will go. The harder you throw it the farther it will generally go. The higher you aim it, the higher its trajectory. And, if we eliminate the variability of the human thrower and use mechanical devices like a catapult, we can predict even more accurately where the stone will go. A rock is a static object, one that cannot direct itself. On the other hand, if you throw a bird (gently, please!), there is virtually no way to predict which way the bird will go and where it will land. The bird can sense its surroundings and may head off in any direction. A bird is a living dynamic system that gathers and processes input and interacts with its environment.

This distinction between static and dynamic processes is important in systems theory and in evaluations that are done from a systems perspective. Since programs involve people and organizations they are inherently dynamic, it is difficult to predict where they will go and what will happen. As programs unfold, the directions they take are influenced by the surroundings and by the interactions of the participants. In this sense, programs are more like birds than like rocks. On the other hand, the idea of a “program” suggests that we are trying to do something systematic, that we are attempting to follow a pre-determined set of steps in order to achieve some predictable result. In this sense, programs are more static, they are more like the stone in the parable.

So, which is it? Are programs static or dynamic? Should our evaluations be constructed for one or for the other? The short answer is: both are important. Both the rock and the bird can be understood from a systems perspective. Both are parts in a larger whole. Both have relationships to the other parts. Over time, programs are likely to evolve through different phases, some more static and others more dynamic. For instance, when a program is first being developed and piloted it is likely to be very dynamic and unpredictable. In fact, that dynamism is essential for learning and adaptation, for enhancing the focus and quality of the endeavor. Over time, many programs tend to become more static. They become routinized and develop standard operating procedures that can be implemented consistently. They can be transferred to new contexts and settings with some degree of predictability. This standardization is also essential. Over even a longer period of time the program may become too static or rigidified, or it may lead to insights that suggest even better variations that might be tried. In either case, we might be motivated to begin other cycles of dynamic-static program development and evolution. Understanding the interplay of static and dynamic systems is essential for Evolutionary Evaluation. We need to recognize that both have their place in evaluation and be able to identify how evaluation approaches need to evolve over time both to encourage program evolution and to provide feedback and learning about it.

The 'Local' and the 'Global' (Scale)

In systems thinking we are always dealing with hierarchies of part-whole relationships as described above. When we think or talk about different *levels* of this kind of hierarchy we are operating at different levels of *scale* in the system. Physical part-whole hierarchies can exist from the subatomic level to the scale of the universe as a whole. Conceptual hierarchies can exist

from the most general level (programs in general) to the most specific subcategory (a summer science youth camp in Ithaca, New York).

We can look at any system from many different viewpoints. For instance, if we are looking at an organization with multiple programs, each program can be viewed as a "part" in the system that constitutes the organization. When we talk about the relationship between a program and its organization, we can think of the program as "local" and the organization as "global" in relation to each other because they are at different but related levels in the hierarchy. On the other hand, when we compare or contrast two programs within an organization we can think of that as a "local–local" relationship because both are at the same level of scale in the hierarchy. If we shift our perspective to a higher level of scale, we are also shifting what we consider "local" and "global." For instance if we think about an organization as one part in a larger system of similar organizations (e.g., a county office in a state-wide system of such offices), then the organization becomes "local" to the system's "global." When we compare two county level offices, we are looking at a "local-local" relationship. When we look at the county level office in relation to the state office we have a "local-global" relationship because we are looking across different levels of scale.

Why are the ideas of scale and of local and global relationships important in evaluation? Different parts of a system do not exist in isolation. If we do not take them into consideration throughout our evaluation efforts we can run into significant problems that can jeopardize the whole endeavor. For instance, very often something in one part of a system may be in conflict with something at another. A program activity may conflict or compete with the activity of another program (a local – local relationship in a system) or with an organizational policy or effort (a local – global relationship). Or the expectations that stakeholders at one level of scale

have for an evaluation may be very different than those of stakeholders at a different level. Funders may expect that the evaluation will focus on accountability and impact while program implementers may be more interested in how evaluation can contribute more immediate feedback that can be used to improve programs.

The Eye of the Beholder (Multiple Perspectives)

A system can be viewed from many different perspectives. Almost everyone is familiar with the famous drawing from the psychology of perception that shows either faces or a vase depending on how you look at it. When you stare at this picture you can actually experience the shift in perspective that psychologists have described as the "figure-ground" effect. The same system can seem very different when looked at from different viewpoints. We cannot really understand the system and its interdependencies unless we look at it from multiple points of view.

The issue of multiple perspectives is essential in evaluation for a number of reasons (depending on your perspective!). For instance, all program evaluations involve a multiplicity of stakeholders including the participants, program developers, administrators, support staff, families and community, funders, policymakers, politicians, and the general public. One of the most important things an evaluator can do is to help the different stakeholders see the system of a program from the perspectives of other stakeholders. For instance, program deliverers may not perceive why they are being pressured to evaluate their programs "from an outside perspective" or why they need to demonstrate outcomes and impacts. If they understand the system pressures on different stakeholders, in this case the funders, they may gain a greater appreciation of how their view fits into the larger system. Conversely, funders may not understand why the organizations they fund are resisting their calls for evaluation. If they can begin to view the

program through the eyes of those who deliver it or participate in it they are likely to understand their pressures better. In this example, it is easy to see that the issue of perspective is intimately related to the motivations and incentives of different stakeholders. The field of evaluation has long emphasized the values of participatory evaluation approaches, in part because of the critical importance of multiple perspectives.

But multiple perspectives are also critical for understanding the content and meaning of programs. Throughout an evaluation it is valuable to have key stakeholders look at different parts of the program, to share their views, and to consider how others might perceive them. For instance, it is surprising how many times even in simple programs different people will have remarkably different views of what they are trying to do or what they think the program is affecting. We find that when people share their perspectives they can uncover such differences and that this learning is critical for informing the evaluation.

Inside – Outside (Boundaries)

All systems have boundaries that distinguish the system from what is outside it. That sounds simple enough, and for many systems it is a relatively simple thing to define what is in the system and what is not. But, for other systems the boundaries are continuous (not abrupt) and are not easily defined. For instance, in nature how do we define the exact borders of the system that constitutes an organism, or a river, or a cloud? What is the "boundary" between two different breeds of dogs or two different species of animals? In living systems, the boundaries may be different depending on how you look at the system and the level of scale or precision at which you look.

In systems evaluation, defining boundaries is a very challenging endeavor. For instance, where does one draw the boundaries on who is a stakeholder to a program? In an educational

program, for instance, do you limit the stakeholders to the program participants and implementers? Do you include family members? Administrators? Funders? The public? In many programs, we think of the program as a set of activities that we can list. However, when we actually try listing program activities we can often find that even co-workers in the same program may list different items. For instance, one person might say the planning of the program or training of program staff is an essential "part" of the program, while others would say the program just consists of what is done once the program is planned and the staff is trained. Is one right and the other wrong? Even if we take the narrower version, we can run into difficulties. Two trained staff members who try to do exactly the same set of activities will inevitably do things slightly differently. A teacher will adapt the way they are presenting material depending on the reactions of students. A doctor will adapt the way they are treating someone depending on their pain level or initial response to treatment. Is that part of what we call the "program"? What exactly is the boundary of the program? The same problem occurs in relation to outcomes. If we have a character development program that is trying to influence children's sense of purpose, where do we draw the boundaries on what that means? Does that mean that children think more about their future? And what does that mean? What do we include in "sense of purpose?" What do we mean by "think more about?" All of these questions involve determining boundaries, often in circumstances where there simply are no fixed and easily determined borders between what is or is not in the system.

Developing an understanding of boundary issues is an important part of systems evaluation. There are no simple answers and often reasonable people involved in the same program will disagree. In some sense, boundary discussions require that stakeholders negotiate a consensus about what they mean by their "program." For instance, in a teacher-training program,

is the program just the set of activities used to train teachers or does it also include the activities that the teachers subsequently do in teaching their students? Discussions about program boundaries often become important learning events for stakeholders because they lead to discussions about the meaning of what they are doing with their programs and the evaluations of them.

And So On, and So On… (Causal Pathways)

The idea of cause and effect is central to systems thinking. The field of System Dynamics, for instance, develops cause-effect models or "causal" chains and uses them to think about the way causes produce effects throughout the system and the different types of feedback loops that result that can often lead to unanticipated outcomes. In effect, you are describing the chain of thinking in a system: "X leads to Y which leads to Z, and so on, and so on…"

The notion of causality is critically important in Evolutionary Evaluation. It is central to theory of change approaches to evaluation, to path analysis, to theory-driven approaches, and to much more. In program logic models there is a general idea of causality – activities are expected to lead to outputs which are in turn expected to produce short-, medium-, and long-term outcomes and ultimately impacts. However, one problem with traditional logic models is that they are "columnar" in nature. The entire set of program activities, or outputs or each phase of outcomes are typically treated as a whole. That is, in traditional logic modeling while we expect that program activities produce outputs, we do not specify *which* activities are expected to produce *which* outcomes. In other words, traditional logic models do not spell out the specific cause-effect relationships that are expected.

Because systems thinking suggests that distinguishing different cause-effect chains can be important, we prefer program logic models that describe the specific causal pathways

involved in programs. For example, typical programs usually involve multiple activities, outputs and outcomes. In a pathway approach, you would specify each connection that you think might be relevant. You might specify that Activity A effects short-term Outcomes A and C which in turn affect mid-term outcomes E and F, and long-term Outcomes A and D. You might also expect that there will be feedback loops in your model. For instance, changing the results of an outcome could trigger a change in a program activity that would then influence other outcomes.

This kind of causal pathway model is useful in telling the story of the program and is essential in developing a high-quality evaluation of it. A program model is likely to have many pathways from activities to outcomes. Drawing pictures of the pathway model enables you to understand better how you think your program should operate. It is especially useful for tracing the "throughline" of your program, the major causal paths through the model of your program. The throughline points out the key program activities that may lead to any outcomes and helps you to identify key outcomes that should be measured.

III. Advancing Character Development Evaluation

The foundational theories associated with evolutionary theory and systems thinking are interesting in themselves but have important practical implications for real-world application. In this section, we turn our attention to the implications of these theories to the evaluation of CD programs and program theories. We first consider the "global" level, the macro level of the organization and its context, including its multiple niches and cultures, its varied value systems, and how these relate to evaluation. We argue that issues of motivation, the challenges of sustaining evaluation, the need to cultivate an evaluation culture and the integral role of critical evaluative thinking are essential at this level. We consider several systems methodologies that are essential to encouraging evolution of knowledge of character development program theory

and practice. We then turn to the "local" level, the micro level of the program and its planning, implementation and evaluation. We present a concrete systems evaluation "protocol" that has been developed and tested over the past decade that is designed to help any program stakeholders to develop a program evaluation that is systems and context sensitive, can be used to evaluate a program at any stage of its lifecourse (ontogeny) and will contribute to a more general agenda of knowledge evaluation (phylogeny). We also introduce an evolving web-based cyberinfrastructure that is freely available and can be used to manage both local program evaluations and more global portfolios of programs.

Organizational (Macro) Level

The challenges to building and sustaining evaluation at an organizational level are many, particularly within the context of character development programs. In order to truly advance, the field of character development needs to commit to an enhanced focus on evaluation as a valuable, institutionalized, and embedded aspect of organizational practice (Mayne, 2010; Sanders, 2002). This includes adequately investing time and money in evaluation and evaluation planning. Evaluation should not be an afterthought, but rather should be considered early on in program planning and ideally should be fully embedded within program development and practice. Equally important is the need for committed leadership, especially at the funder level in order to build and sustain evaluation systems. Given the considerable turnover in frontline staff for CD programs, the need for sustainable evaluation systems and policies is even more critical. Specifically, the field of character development needs to work toward building an evaluation culture.

Evaluation Culture and the Experimenting Society. Trochim (1991, 2006) originally suggested the idea of an "evaluation culture" as a coherent set of values and norms that are

essential to the development and sustaining of evaluation as an evolutionary feedback mechanism for society. He based this notion loosely on and in deliberate contrast to the earlier notion of an "experimenting society" offered by Campbell (1991) and based on a legendary unpublished and highly circulated earlier paper that dated back to 1971. Both the experimenting society and the evaluation culture shared a number of key value and principles that such societies and cultures might embody. For Campbell, an experimenting society would be active, honest, nondogmatic, scientific, accountable (challengeable, due-process), decentralized, committed to means-idealism as well as ends-idealism, popularly responsive, voluntaristic and equalitarian (Campbell, 1991). For Trochim, an evaluation culture would be: action-oriented; teaching-oriented; diverse, inclusive, participatory, responsive and fundamentally non-hierarchical; humble, self-critical; interdisciplinary; truth-seeking; forward-looking; and, ethical and democratic (Trochim, 2006). Both were idealistic statements of societies that did not then and do not yet exist. Each one characterized its era and incorporated the language and norms that were prominent. Each constituted an "historical indexicality" of its generation and provided a glimpse of the evolution of meta-thinking about evaluation and its role in society. And, each one emphasized the important role of data and methodology, with the experimenting society stressing the importance of experimental and quasi-experimental approaches and the evaluation culture focusing on a critical multiplistic theory-driven mixed methodology approach. In the ongoing evolution of evaluation we might now update these with ideals that emphasize evolutionary and systems thinking, but both of these visions are recognizable to us today and remain worth pursuing and continually revising. At the highest level, we cannot have a true integration of evaluation in our society until we harmonize the values of that society and those of contemporary

evaluation. This section attempts to describe some of the macro-level changes that would move us closer to such visions and some of the methodological approaches that could help get us there.

Systems Change

This section describes the areas where broad-level systems change typically needs to happen in order to develop a sustainable and effective evaluative function in a complex hierarchical system like the CD context.

Evaluation Capacity Building. In the past fifteen years, the field of evaluation has seen a proliferation of Evaluation Capacity Building (ECB) definitions, models, and approaches. Although broad consensus has not been reached on any single definition of ECB, the most commonly cited definition of ECB is "the intentional work to continuously create and sustain overall organizational processes that make quality evaluation and its uses routine" (Stockdill, Baizerman, & Compton, 2002, p. 14). ECB is often associated with collaborative, participatory, and empowerment evaluation (Fetterman & Wandersman, 2005; O'Sullivan, 2004; Rodriguez-Campos, 2005) all of which share a common interest in democratizing and decentralizing evaluation practice. This stands in contrast to more traditional models of program evaluation wherein evaluation was solely the purview of expert researchers. ECB provides opportunities to infuse organizations and programs with evaluation skills, attitudes, and practices that promote a culture of evaluation and ultimately improve program outcomes (Labin, 2014; Suarez-Balcazar & Taylor-Ritzler, 2014; Wandersman, 2014).

Systems that consist of multiple projects and organizations typically require a wide range of evaluation skills, resources and capabilities in order to provide the essential feedback mechanisms for system learning and to meet the accountability reporting needs of their funders and stakeholders. Some of these needs are best met by developing internal organizational or

system capacity as by hiring experienced evaluators, supporting external or in-service training in evaluation, hiring external consultants, and so on. Evaluation capability is sometimes met by funders who require evaluation, through additional funding for this function or through evaluation technical assistance. In many cases, individuals or groups within organizations simply seek out resources and information on evaluation, increasingly these days through technology-based solutions such as evaluation websites. There is no shortage of useful information available for meeting evaluation capacity needs, including websites and materials that present case studies, measures, examples, tutorials, and other resources, available from public sector government agencies such as NSF (J. Frechtling, 2002; J. Frechtling & Sharp, 1997) and the Centers for Disease Control (www.cdc.gov/eval/), individual researchers (www.socialresearchmethods.net), professional associations (www.eval.org) and foundations (www.wkkf.org). The problem is not a lack of resources; rather, there are too many resources of varying and indeterminate quality that are not screened or adapted for use in particular contexts. And, even if program staff can get to high quality resources, their ability to use them effectively is greatly enhanced by appropriate support systems and experienced evaluators. There is little in the way of systematizing the plethora of resources and little support for practitioners who need to navigate this complex terrain. Nor is this problem solved through developing and providing more resources, more websites, and more materials. Despite the ubiquitous nature of evaluation needs, and the abundance of supporting resources, there is surprisingly little attention paid in the evaluation literature to the development of systems for supporting evaluation, and especially a paucity of research on how to coordinate and integrate the varied existing supports of the evaluation function. Much of the evaluation literature in this area emphasizes developing the government's

evaluation capacity (Wye & Sonnichsen, 1992), rather than enhancing the practitioner systems (Compton, Baizerman, & Stockdill, 2002).

Evaluative Thinking. A key recent development in the field of evaluation is the development of the idea of evaluative thinking as an essential skill for evaluation and as a mechanism for building evaluation capacity (Compton et al., 2002). In recent years, several authors have defined evaluative thinking (Chinnock, 2008; Davidson, 2005; Patton, 2005; Preskill, 2008). Most agree that it includes higher order thinking and decision making skills as well as an evaluative attitude and set of beliefs. Patton (2005) defined evaluative thinking as "a willingness to do reality testing, to ask the question: how do we know what we think we know? To use data to inform decisions. Evaluative thinking is not just limited to evaluation projects, it's not even just limited to formal evaluation; it's an analytical way of thinking that infuses everything that goes on." This definition suggests that evaluative thinking is a skill that would be valuable to many different people in many different fields and situations; and therefore its development would have positive impacts beyond evaluation. Program implementers who are also evaluative thinkers will be good evaluators as well as better program planners, managers, organizers and even participants. The supposition is that engaging people in evaluative thinking is also likely to be a personally exciting endeavor that enhances their understanding of and motivation to engage in evaluation, and consequently contributes to evaluation capacity and sustainability in an organization.

Recent efforts to explore evaluative thinking more deeply (Archibald, Buckley, & Trochim, 2011, November) have defined the construct of evaluative thinking from the perspective of evaluation capacity building and education. This work builds upon work in education and cognitive science that defined and explained critical thinking (education) and

evaluativist thinking (cognitive science) and established a research base for developing programs and activities designed to promote evaluative thinking in and across an organization (Beyer, 1987; Bloom, 1956; Brookfield, 1987; Kuhn, 2005). Cognitive scientists, going back to Bloom in 1956, have established that higher order thinking skills, such as evaluative thinking, are non-trivial and require conscious effort to learn and practice to improve (Bloom, 1956; Perkins, 1986).

Evaluation Policy. An evaluation policy is "any rule or principle that a group or organization uses to guide its decisions and actions about evaluation" (Trochim, 2009). Every group and organization that engages in evaluation has evaluation policies. Sometimes these are formal, explicit and written; but often they are implicit and ad hoc principles or norms that have simply evolved over time. This is the essence of the connection between evaluation policy and evaluation culture: implicit evaluation policies reflect evolved organizational norms. By making these implicit policies more explicit – and written – we make the implicit organizational culture more transparent and encourage dialogue within that culture. Organizations and systems responsible for CD programs should move toward an explicit articulation of evaluation policies as part of its overall policymaking efforts. The central contention here is that "evaluation policy" can be a major methodological vehicle for connecting organizational cultural norms about evaluation to the world of practice.

In the past decade, considerable work has been devoted to articulating evaluation policy and its implications for evaluation practice (L.J. Cooksy, Mark, & Trochim, 2009; Mark, Cooksy, & Trochim, 2009; Trochim, Mark, & Cooksy, 2009) and evaluation policy is considered important for a number of reasons. It has a major role in signaling or communicating to an organization what the rules and expectations are regarding evaluation. Explicit policies help

make organizations more democratic and transparent. Evaluation policies provide a sustainable mechanism for teaching and learning about evaluation within an organization. They are also an efficient way to influence practice – a change in a single policy has a distributive effect on all practice affected by that policy. Finally, evaluation policy has an important function as a "lightening rod" because most controversial issues in contemporary evaluation can be framed as policy issues and the debates about those issues/policies consequently help the field address the issues.

Our research team has been central in leading the evaluation field to focus on the role of evaluation policy and, especially, on the need to develop methodologies for identifying and managing existing policies

Systems Methods

Here we consider several methodological approaches that are appropriate at the systems level for supporting and sustaining the evaluation function.

Structured Conceptualization or Concept Mapping. As we shift to more systems level endeavors in evaluation, there is a need for methods and tools that are explicitly designed to handle the complexity of systems work. By systems methods we mean any methods that were designed or can be adapted to operate in systems contexts. Because of the nature of systems work, these methods often are participatory and collaborative. They might be extensions of traditional methods, as in the use of survey research or traditional measurement tools applied to systems contexts. But increasingly these methods are themselves more complex hybrids and mixtures of traditional methods, participatory processes and novel uses of technology. A methodological agenda for CD evaluation should include advancing understanding and encouraging development of appropriate systems methods and approaches, either through

adapting traditional methods more effectively to systems contexts or by suggesting development of new methods.

One example of a systems method that could be useful in the evaluation of CD efforts is the method of structured conceptualization or collaborative concept mapping (Trochim, 1989b; Trochim & Linton, 1986; Trochim, Milstein, Wood, Jackson, & Pressler, 2004) that was explicitly designed to enhance the ability of a system to develop shared frameworks and models, and measures or observational approaches for assessment and evaluation. This approach involves the identification of a focus for the map (e.g., a program), the brainstorming of a large set of ideas to address the focus, individual participant sorting and rating (importance, feasibility) of each idea, the use of a sequence of multivariate statistical analyses (multidimensional scaling, hierarchical cluster analysis), and the interpretation and use of the maps that results. The method is generic, taking the input of a group and resulting in a visual conceptual framework. It has been used widely for strategic and operational planning, for conceptualizing evaluations and measurement systems, and for developing frameworks for analyzing data.

Funder Portfolio Analysis. Throughout most of its history, educational evaluation has tended to focus primarily on the program as the primary unit, almost as though programs exist in a systems vacuum. The literature is replete with discussions that address program evaluation without attending to the systems issues within which programs are situated. Often it is the system that determines how evaluations are done, who is allowed to do them, how they will be resourced, when and to whom results will be reported, and so on. Yet evaluators tend to treat each evaluation as though it can be designed as a unique one-off endeavor based only or primarily on local considerations, proximal needs and immediate concerns about threats to validity.

A challenge for CD programs is to link the multifaceted and disparate set of experiences that occur when a youth participates in multiple programs with the broader goals they are collectively designed to affect. The central evaluation problem is how best to make a connection between the many and varied local activities and more macro longer-term global outcomes. This is a classic systems thinking problem, a part-whole or "local-global" one. Local character development programs are typically most concerned with the experiences of their participants and how service delivery can be improved (local level concerns). Funders are typically more concerned with how their portfolio of programs affect more global questions related to longer-term impact (global level concerns). We need to recognize that this is essentially a hierarchical systems challenge – how to harmonize the local model of a specific program with the hierarchically broader model of programs of that type.

When evaluations do occur, they are typically done in isolation at the local level which results in a portfolio of similarly oriented programs that exist across the same or multiple organizations that often have separate mutually uninformed program models. In addition, there is

also usually at least an implicit model at the next level up in the organization. For instance, in a case where a foundation funds a portfolio of grants that each have character development programs, it is likely that there is some model at the foundation level, although this may not have been formally articulated. In this case, there is typically no linkage or integration of the models either horizontally (local-local) or vertically (local-global). Programs do not identify common activities or how their activities connect with a global model. They do not harmonize their outcomes or even a subset of them, making it difficult if not impossible to aggregate results subsequently across the portfolio of programs. There is seldom formal policy across the portfolio regarding how the programs should be evaluated. One option is to impose a single common model onto the local programs, but this runs the risk of increasing local burden, reducing flexibility and local adaptation, not being responsive to local conditions, and creating program "monocultures" (Trochim, 2007).

Instead, we propose a process whereby the local and global entities work as a system to coordinate and harmonize what they are doing, understand each other's needs and perspectives, seek common connections in their models, and encourage an emergent harmonization that both enables local flexibility and some global synthesis of results. This linked and integrated system model of programs, created together by multiple levels of the system, would provide the foundation for development and analysis of program portfolios. This becomes especially important when funders are trying to understand where they are making programmatic investments, whether their portfolio of funded work is addressing the overarching mission of the agency, where gaps may exist, and in identifying emerging areas that warrant additional investment. Without a methodology for portfolio analysis, funders typically must rely on the disjointed evaluation efforts of their grantees and post-hoc evaluations that try to assess broad

programmatic impact based on limited data. A fundamental issue driving the difficult process of evaluating portfolios is that at the funder level, models are typically not articulated. Without a well-articulated funder-level model, practitioners of programs at the local-level may believe they are addressing the goals and objectives of the funder, but they are unlikely to be able to describe *explicitly* how they are doing so.

By embracing the dynamics of phylogeny, an Evolutionary Evaluation approach addresses this need by building hierarchical nested program models (nested pathway models) that integrate the logic of individual programs with other peer programs and the logic of higher levels of the system that fund and oversee the programs. By developing nested program models, funders can then more effectively assess a "portfolio" or hierarchical aggregation of similar programs across a system. See the discussion on the creation of pathway models in the section below on Program (Micro) level approaches.

Evaluation Policy Methods. Evaluation policy does not just happen on its own; it has to be developed. Trochim (2009) offered a taxonomy of eight evaluation policy domains or categories organized into a visual "policy wheel" for analyzing and managing evaluation policies (Figure 1) that includes policy domains that cover: evaluation goals; participation; capacity building; management; roles; process and methods; use; and, meta-evaluation. The policy wheel organizes evaluation policy into a simple circle diagram, divided into wedges that correspond to the proposed eight types of evaluation policies in the taxonomy. All evaluation policies can be placed somewhere on the wheel. There are different concentric circles on the wheel, with more general policies in each category placed on the outer circles and more specific policies (and ultimately practices or procedures) more central. There are also different layers in the wheel representing different levels of organizational hierarchy. The different levels correspond to

different global-local arrangements in a nested hierarchy. For instance, Level 1 might correspond to a national funder level, Level 2 to a cross-site lead organization level and Level 3 to an individual program site level. Each level inherits policy from above and delegates responsibility for greater policy specificity to those below.

Trochim (2009) also describes a number of principles or rules that guide the organization of evaluation and functioning of policies in this structure and constitute the basis of a methodology for evaluation policy analysis:

- All policies and practices "inherit" their parent characteristics (inheritance)
- A child policy can never be broader than its parent (encapsulation)
- Inner policies are more specific than outer (parent) policies (specificity)
- Policies should cover the entire relevant domain (exhaustiveness)
- There should be no large gaps between levels of policies (continuity) (Discontinuity suggests micromanagement)
- Responsibility is delegated for more detailed policy or practice than specified (delegation)
- "Any reasonable interpretation" of the delegated policies is legitimate (accountability)

Policymaking in this model is a very dynamic and iterative process. It begins with the policymaker(s) describing the most general and highest-level policies in each of the eight domains. For example, for the Evaluation Capacity domain, the broadest level policy might be something like: *The organization will develop and implement sufficient organization-wide capacity to support evaluation activities.* If this were the only policy in this domain it means that the policymakers are delegating to those responsible for enacting the policy the responsibility for defining and achieving it. The practitioners or staff are responsible for reporting on how they operationalized and addressed the policy. The policymakers are responsible for reviewing and evaluating the degree to which the practices that were enacted addressed their intent. Let us assume for the moment that the organization has this policy in place for a year, the staff report

back to policymakers on what they did to address it, and the policymakers do not think it is sufficient. They might then enact one or more child policies that spell out more specifically what the general policy intended. For instance, these might include a policy such as: *Staff will be provided training in the methodology and use of evaluation appropriate to their program roles*. Again, staff would operationalize and implement, presenting their results in a subsequent cycle to policymakers and if the result is still insufficient additional policy specification may be needed. One especially nice feature of this evaluation policy model is that it is inherently delegative, discourages micromanagement, and acknowledges that there may be multiple ways to address any policy and tailor it to the varying needs of organizational or program components. Organizations engaged in CD programming should consider the potential utility of using a policy model like this one for developing, managing and disseminating evaluation policy.

Program (Micro) Level

In this section we take the foundational issues described above and begin to describe how they might be applied in the character development field to enhance the evolution of our knowledge about programs and what works. First, at the program level we consider two ways an evolutionary evaluation approach can enhance our work: through the use of a Systems Evaluation Protocol which integrates evolutionary and systems principles into program evaluations; and through the development of methods that integrate practice and research.

The Systems Evaluation Protocol

The Systems Evaluation Protocol (SEP) is a systems approach to evaluation that incorporates Evolutionary Evaluation principles and considers the complex factors that are inherent in the larger systems within which a program is embedded (Trochim et al., 2012; J.B. Urban, Hargraves, Hebbard, Burgermaster, & Trochim, 2011; J. B. Urban et al., 2014; J.B.

Urban & Trochim, 2009). This approach emphasizes: the importance of creating a causal diagram that illustrates the programmatic theory of change; incorporating the perspectives of both internal and external stakeholders of the program; recognizing how the program is related to other programs either in the same system or other systems, in part by identifying research on similar or related outcomes which can help link the program to more universal long-term goals; and continually assessing and revising the theory of change and evaluation plans based on knowledge gained through evaluation efforts (J. B. Urban et al., 2014; J.B. Urban & Trochim, 2009). The SEP integrates the three primary phases of any evaluation: Planning, Implementation and Utilization. Each stage of each phase has a detailed list of components or steps that can be followed to accomplish an evaluation. This presentation emphasizes the planning phase of evaluation. The complete summary of the steps in the three stages of planning are presented in Table 1.

The SEP incorporates these principles and distills them into a series of steps resulting in the creation and implementation of an evaluation plan. Although the foundations of the approach are complex and rooted in theories of developmental systems (e.g., Lerner, 2006; W.F. Overton, 2006, 2010), evolution (Darwin, 1859; Mayr, 2001), evolutionary epistemology (M. Bradie & W. Harms, 2006; Campbell, 1974, 1988; Cziko & Campbell, 1990; Popper, 1973b, 1985), and systems (Bertalanffy, 1995; Laszlo, 1996; Midgley, 2003; Ragsdell, West, & Wilby, 2002), the process of undertaking a high-quality evaluation becomes manageable, replicable, and teachable.

The idea of an evaluation protocol comes from research in clinical medicine. A protocol is the foundation of any clinical trial. It describes the standardized series of steps taken to accomplish the trial:

> "Every clinical trial must have a protocol, or action plan that describes what will be done in the study, how it will be conducted, and why each part of the study is necessary - including details such as the criteria for patient participation, the schedule of tests, procedures, and medications, and the length of the study" (National Institutes of Health, 2007).

The SEP is a protocol that is designed to generate evaluation protocols or plans. It is a series of repeatable steps that when followed lead to the creation of project logic and pathway models and an evaluation plan that can subsequently be implemented and utilized. In this sense it addresses well the CD environment which needs standardization of evaluation approaches while recognizing the enormous varieties of contexts within which CD programming occurs.

Systems Theory Foundations. While the SEP yields evaluation plans that meet current best practices, it was also explicitly constructed to extend current evaluation practice through deliberate integration of principles from systems theory. The topic of systems thinking and systems theory is extremely complex; the literatures that informed the formulation of the SEP were considerable and included: causal feedback (Richardson, 1991); stock–flow structures and open and closed systems (J. D. Sterman, 2000); centralized, decentralized, heterarchical, hierarchical, and self-organizing systems (Kauffman, 1993, 1995); nonlinear systems and chaos (Strogatz, 1994); complex adaptive systems (Gell-Mann, 1995, 2003; J. Holland, 1995; Waldrop, 1992); boundary conditions and scaling, (Strogatz, 1994; Yeomans, 1992); emergence (J. H. Holland, 1998; Johnson, 2001); cellular automata (Wolfram, 2002); fractal self-similarity (McKelvey, 1999); general systems theory (Bertalanffy, 1995); cybernetics (Francois, 2004; J. F. Young, 1969); control theory (Sontag, 1998); information theory (Shannon, 1948); computational simulation (Gilbert & Troitzsch, 1999; Resnick, 1994); system dynamics

(Forrester, 1961, 1994, 1997; Richardson, 1996; Society, 2002; J. Sterman, 2001); evolutionary theory, biology, and ecology (Capra, 1997, 2002; Capra, Mill Valley School District (Mill Valley Calif.), & Elmwood Institute., 1994; Capra, Steindl-Rast, & New Dimensions Foundation., 1993); small world phenomena (Strogatz, 2003; Watts, 1999a, 1999b); and set, graph, and network theory (Capra, 2002; Capra et al., 1994; Maturana & Varela, 1980; Newman, 2003; Strogatz, 2003; Watts, 1999a, 2003). Because the SEP integrated principles associated with these theories it helps to assure that programs that use it will incorporate such principles when developing program pathway models and identifying key pathways and nodes (outputs and outcomes); determining the boundary conditions for their program model, assessing program lifecycles, and selecting evaluation designs that are appropriate to their program evolution.

Cyberinfrastructure Foundations. The SEP also builds on principles from contemporary cyberinfrastructure research (National Science Foundation, 2007). While the SEP can be implemented as a manual process that does not depend on any specific technology platform, it is designed so that it can be enhanced throughout by using a system developed in our current research called the Netway, a Web 2.0 application consistent with second-generation web-based communities and hosted services such as social-networking sites, wikis, and blogs (Wikipedia, 2007). The Netway[1] is constructed so that when program educators and managers enter program information about activities, outputs and outcomes, the system can immediately identify and suggest other existing programs that have similar or common elements and enable the users to adopt or adapt these elements for their own programs while automatically creating

[1] The Netway is currently available as a free web service at http://www.evaluationnetway.com/ that requires the user to register with a unique Username and Password. The Netway incorporates extensive resources including the complete documentation of both the Systems Evaluation Protocol and the web-based system itself.

networked linkages of their models with others (Asim, Essegaier, & Kohli, 2000; Burke, 2000). Each new program model adds to the online network of such models and can be accessed in turn by others. This also helps ensure that different parts of the system can learn from each other and that even programs with no direct contact with one another, can use the cyberinfrastructure to benefit from each other's experiences (Marathe, 1999). Evaluators who are supporting programs that use the Netway can see their portfolio of programs and what they are adding to the system in real-time, and can communicate with program users about their models and evaluation plans, thus enabling new models of virtual consultation. The system is designed so that researchers will be able to identify clusters of programs that are in their substantive areas of interests and learn about new and emerging programs that are responding to local needs and conditions. Funders (e.g., NIH) can view meta-summaries of programs across program areas, see where they are in their developmental lifecycles, and more effectively manage their portfolios of evaluations. The Netway cyberinfrastructure is a creative incorporation of technology that fundamentally changes the nature of evaluation practice for both the evaluator and the practitioner and has the potential to be a transformative mechanism for CD evaluation particularly and for evaluation generally in the 21st century.

SEP Stages

The Systems Evaluation Protocol includes three stages of evaluation planning: (1) Preparation, (2) Model Development, and (3) Evaluation Plan Development. The primary objectives of the Preparation stage are to acquaint the working group with the SEP process, identify people's key roles, and collect basic information about the program. The Model Development stage is a central and distinguishing component of the SEP, focused on surfacing and articulating deeper understandings of the program through: stakeholder analysis and

mapping (a visual depiction of the stakeholders and their relationship to each other); group discussion and program review including development of a written program description; identification of program and evaluation lifecycle phases; structured program modeling in two forms - the more familiar columnar logic model and a corresponding visual pathway model; linking the pathway model with the evidence-base; and, determining the scope of the evaluation. A stakeholder map, lifecycle determinations, and the logic and pathway models are the products of the Model Development stage. These products form the foundation for strategic decision-making about the evaluation scope, evaluation purpose, specific evaluation questions and other components of the evaluation plan which in turn serve as the basis for Evaluation Plan Development in the third and final stage of the SEP. In the Evaluation Plan Development stage the specific design, sample, measures, and analysis plan are developed with careful consideration of program and evaluation lifecycle alignment and principles of Evolutionary Evaluation. A clear benefit of working through the SEP is that the process: builds evaluation capacity in terms of specific skills and knowledge (modeling, lifecycle analysis, evaluation methodology, etc.); deepens staff understanding of the program and ability to communicate with stakeholders; and, cultivates skills, patterns of thought, and commitment to evaluation that constitute evaluative thinking (J.B. Urban et al., 2015).

Causal Pathway Models for integrating Research and Practice. Local practitioners (e.g., local CD program managers) tend to be most interested in shorter-term outcomes and improving their practice. The practitioners' dilemma is that they operate on a local level yet they are asked to demonstrate effects on long-term, broader outcomes. Those who are situated at a more global level of the system (e.g., CD program funder level) tend to be more interested in longer-term outcomes and focus on broad impact. The program director's dilemma is that they

are beholden to their funders and expected to demonstrate large-scale impact often aggregated across multiple local program sites. This is a classic systems thinking problem, a part–whole or local–global challenge (G. Young, 1999). How do we connect the varied local experiences with the broader global outcomes of interest? At the core of this systems challenge is the central role that program evaluation, planning, the research evidence-base, and particularly detailed and clearly articulated program modeling can play in making these local–global connections.

The heart of the solution to this systems challenge is visual causal diagrams of program theory which are based on work done in theory-driven evaluation (Chen & Rossi, 1983) and logic modeling (Bickman, 1987; Foundation, 2004; McLaughlin & Jordan, 1999). A well-articulated visual causal model (such as the causal pathway model presented in Figure 2) provides the foundation for systematically and dynamically linking program theory with the research evidence-base. Similar to logic models, pathway models provide a conceptual framework for describing programs. However, while logic models rely on columnar representations that link whole sets of activities to sets of outcomes, pathway models make these connections more explicit and precise by graphically depicting a network of causal linkages, primary pathways and nodes.

The Systems Evaluation Protocol facilitates research-practice integration particularly during the process of linking the pathway model with the research evidence-base. Combining the research literature with high-quality evaluation can help build a strong case for the underlying theory of change. We provide an example from the Inspire Aspire character development program in which youth reflect on personal strengths and areas in need of improvement, research an inspirational figure, and contemplate what they can do to bring their vision for a better world to life. The program culminates with the youth creating a poster.

Figure 3 zooms in on the first highlighted throughline from the pathway model in Figure 1. In our evaluation of Inspire Aspire, we knew that we would not be able to follow the youth long enough to actually measure the long-term outcome of interest. However, we were able to measure the short- and medium-term outcomes. We have developed a tool to code student posters for alignment between values and articulation of future ambitions (short-term outcome). We have gone to the research literature and found validated scales for measuring youth goal setting (short-term outcome; Freund & Baltes, 2002; Gestsdóttir, Bowers, von Eye, Napolitano, & Lerner, 2010) and we are measuring whether there is an association between better aligned posters (short-term outcome) and scores on goal setting measures (short-term outcome). We also found scales that measure sense of purpose (Bundick, Andrews, Jones, Moran, Mariano, Bronk, & Damon, 2008; Steger, Frazier, Oishi, & Kaler, 2006) and we are measuring whether there is an association between goal setting (short-term outcome) and sense of purpose (medium-term outcome). However, we do not have the resources to follow-up with these kids in the future to see if they also go on to develop a commitment to a vision for a better world (long-term outcome); but we have found research that shows that increased sense of purpose and goal setting abilities lead to a commitment to a vision for a better world (Moran, 2014; Yeager, Bundick, & Johnson, 2012). The research evidence-base can pick up where we left off in our measurement and demonstrate the logical connection to longer-term outcomes. Ultimately, our goal is to link the evidence that is derived from a "local" evaluation of a program with the more "global" evidence that is generated by research, to identify places where local evaluation efforts and the research literature meet; in other words, to find what we refer to as the "golden spike" (J.B. Urban & Trochim, 2009). The development of the pathway model and the subsequent linkage with the research evidence base provides a framework for identifying and supporting the

connections from short-and medium-term outcomes to long-term ones. A pathway model can provide a compelling rationale for explaining how and why any given program can influence real change even if the only change they are able to demonstrate in their own evaluation is short-term and/or local.

Conclusion

Evaluation is essential to the continued growth of the field of character development and Evolutionary Evaluation which is aligned with Relational Developmental Systems meta-theory is particularly well-suited for the complex challenges associated with evaluation of character development programs and systems. This systems thinking perspective takes into account the dynamic nature of CD programs as well as the dynamic processes CD programs aim to address. The field of character development is at a crossroads and it is time to make significant investments in evaluation in order to reach our shared goal of increasing the number of flourishing people committed to enhancing civil society. This must be done at both the macro, organizational level, as well as at the more micro, program level. Funders must be willing to invest in building evaluation systems. Organizations must be willing to invest time and energy in building evaluation capacity and evaluation policies. Program leaders and practitioners must be committed to careful evaluation planning. And, people at all levels of the system must be willing to understand the broader system, their place within it, and the shared commitment needed to make meaningful change. With these goals in mind, we can collectively advance the evaluation of character development programs.

Table 1. Summary of the Systems Evaluation Protocol (SEP)

Phase I: Evaluation Planning

Stage 1 – Preparation

The Preparation stage is intended to acquaint the participants with this Systems Evaluation Protocol (SEP) and the Evaluation Partnership arrangement, and identify current evaluation resources. The Preparation stage involves the following steps:

1. **Enter the System**: Connect with key decision makers of the organization to discuss commitment to evaluation planning and responsibilities.
2. **Memorandum of Understanding**: Work with key decision-makers in the organization to create a written document that: describes the roles and responsibilities of participants in the evaluation project; details the expectations for the evaluation consulting team, partner site administrators and staff; and provides a timeline of project activities and completion.
3. **Internal Stakeholders**: Identify people in the program/organization who should be involved or consulted in evaluation planning.
4. **Working Group(s)**: Identify those who will play a key role in developing the evaluation plan.
5. **Assess Evaluation Capacity**: Identify the resources available within the organization and within the program - the degree of evaluation training the staff has already received, information technology resources, and evaluation policies.

Stage 2 – Program Model Development

The Modeling stage is intended to enhance participant knowledge of evaluation concepts, and identify how their program "works". The Modeling stage involves the following steps:

1. **Stakeholder Analysis:** The process of identifying, and describing the perspectives of all of the potential people and/or organizations that have a stake in the program and its evaluation
2. **Program Review:** Gain a firm understanding of the components and characteristics of the program including how it operates and whom it serves. The program mission statement is a description of the overarching or long-term goals specific to the program being described.
3. **Program Boundary Analysis:** Program definition and boundary analysis involve defining the structure and elements of a program; specifically, what is considered to be part of the program as opposed to activities or elements that may be just outside the definition of the program.
4. **Lifecycle Analysis:** Lifecycle analysis refers to the process of characterizing the history and current phase of development of a program and its evaluation. The program lifecycle is the individual course a program takes as it evolves, changes, or remains the same over time. Typically, programs earlier in the lifecycle are smaller in scale and more variable in their implementation. Later phase program often, though not always, become more stable and reach a wider audience. Understanding a program's lifecycle history is a critical part of planning for its evaluation. The evaluation lifecycle exists in parallel with the program lifecycle. Early lifecycle evaluations are characterized by rapid feedback, basic designs,

and less of a focus on external validity and reliability. Later lifecycle evaluations are typically more involved, build upon prior knowledge of the program, and allow for broader claims.

5. **Logic Model:** A logic model is a representation of the thinking behind a program. Logic models come in many different formats. Some are more graphical than others. They all include the basic components of program activities and outcomes. Many also include program inputs and outputs. Some include program context and assumptions. Overall, logic models represent the theory of change behind a program - the ways in which the program planners imagine the program activities have an effect on the program's intended outcomes. Generate an initial logic model including the assumptions, context, inputs, activities, outputs, short-, mid-, and long-term outcomes.
6. **Pathway Model:** A pathway model is a type of logic model. It is a graphical representation of the relationships between the activities, outputs, and outcomes that make up a program. Pathway models communicate the "story" or "theory of change" of a program. They are the essential foundation for determining the scope and questions that guide the evaluation of the program being modeled. Use the logic model as a basis for articulating clear and direct linkages between program activities and outcomes.
7. **Evaluation Scope:** Determine the specific components of the pathway model that will be the focus in the upcoming evaluation cycle.
8. **Program-System Links:** Introduce tools and strategies for finding similar programs and shared outcomes, develop research support by drawing on literature and on resources in the systems within which the program exists
9. **Reflection and Synthesis:** Finalize the logic and pathway models including reviewing the program logic model, assessing the model from the perspectives of key stakeholders, reviewing the Program Boundary Analysis, reviewing the Program and Evaluation Lifecycle Analyses, and revising the models as needed. This step also involves integrating relevant research literature as it relates to the causal pathways that have been articulated in the Pathway Model.

Stage 3 - Evaluation Plan Creation

The third stage, "Evaluation Plan Creation," focuses on the creation of an evaluation plan that will guide the implementation of the evaluation. The Evaluation Plan Creation stage involves the following steps:

1. **Evaluation Plan Introduction:** Present and discuss the components of an evaluation plan.
2. **Evaluation Purpose:** The evaluation purpose statement is an introduction to the evaluation plan document that briefly describes the scope of the evaluation being planned, how it fits in with both prior as well as intended future evaluation work, as well as a brief summary of the methodology and intended use of the evaluation. Think of the evaluation purpose statement as an executive summary of your evaluation plan. This purpose statement will be useful for readers of the plan and as a touchstone document for people implementing the plan to make sure that it stays on track.
3. **Evaluation Questions:** Evaluation questions are the broad inquiries about the program that the evaluation will seek to address. The language used in evaluation questions has broad implications for both the methodology of the evaluation as well as the claims that

can be made as a result of completing the evaluation. Develop evaluation questions based on the logic and pathway models, lifecycle analysis, stakeholder analysis, and systems insights. The evaluation questions will function as the core determinants of all the evaluation plan components.

4. **Measurement and Measures:** Measurement is the process of collecting information systematically, using appropriate methods and/or tools, to address an evaluation question. At this stage in the evaluation planning process, this involves selecting a data collection strategy that fits the evaluation question you've identified, and then finding or developing whatever tools (often referred to as "measures") are needed in order to fulfill that strategy. Identify measures already being used in evaluating the program and assess them for quality and feasibility; identify other existing measures that might fit the program evaluation needs; and/or develop any new measures that are needed.
5. **Sampling Plan:** In the context of program evaluation, the sample is the group (of people, objects, etc.) that you will collect data about in order to address the evaluation question(s). For some evaluations, it will be appropriate to collect data about the entire population of interest (for example, all program participants); for other evaluations, the sample may be a sub-group of the population of interest, selected based on the needs of the evaluation and practical considerations (feasibility, resources, etc.).
6. **Evaluation Design:** The evaluation design lays out the sequence of program implementation and data collection. In its most basic form, the evaluation design tells when information is collected and from whom. For example, an evaluation design might specify a "pre-post" data collection strategy, a "post-only", or might include plans for collecting data from a comparison group, etc.
7. **Data Management and Analysis Plan:** Data management is the ongoing process of recording, documenting, tracking, securing, and organizing data that is collected during the course of an evaluation project. It's important to plan this out in advance, in order to ensure that nothing gets lost as data come in, that there is a record of when it came in and through what means, and that data are stored in a way that is retrievable and organized to facilitate the analysis. Analysis is the process of describing and making meaning from data. The appropriate analysis strategy for a given evaluation depends on the evaluation question(s), the type, quantity and quality of the data gathered, as well as the insights and potential claims that emerge from initial review of the data.
8. **Evaluation Reporting and Utilization Plan:** "Reporting" refers to the process of communicating results and recommendations to stakeholders. This includes internal (program staff, participants) and external (funders and administrators) stakeholders. "Utilization" refers, more generally, to the process of using evaluation results to make decisions, if any, about program design, support, staffing, and so on.
9. **Implementation Plan and Timeline:** The evaluation plan timeline outlines the program and evaluation schedule for the evaluation. To ensure smooth progress, overall feasibility, and availability of staff for various tasks, it is important to include as much detail as possible. For example, a good timeline would include tasks such as measure development, pilot-testing of measures, and training of data collectors and/or data entry staff as needed, in addition to the obvious such as program implementation, data collection and analysis.

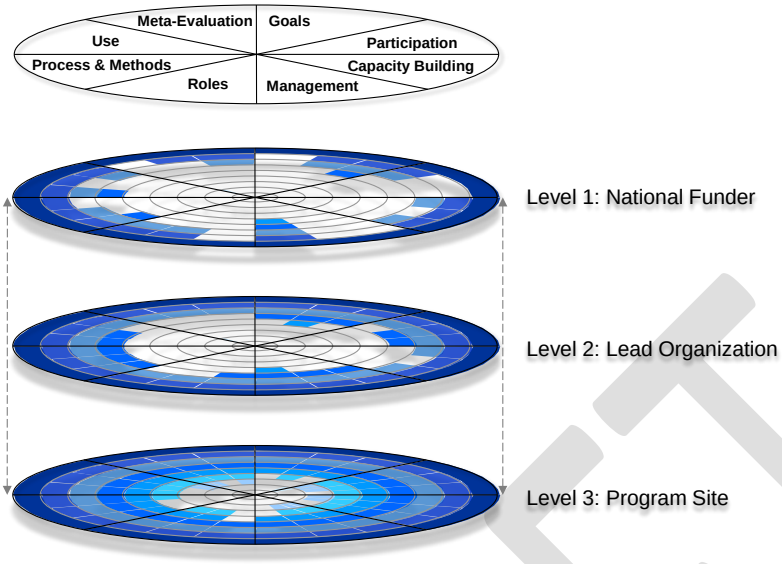

Figure 1. The Evaluation Policy Wheel

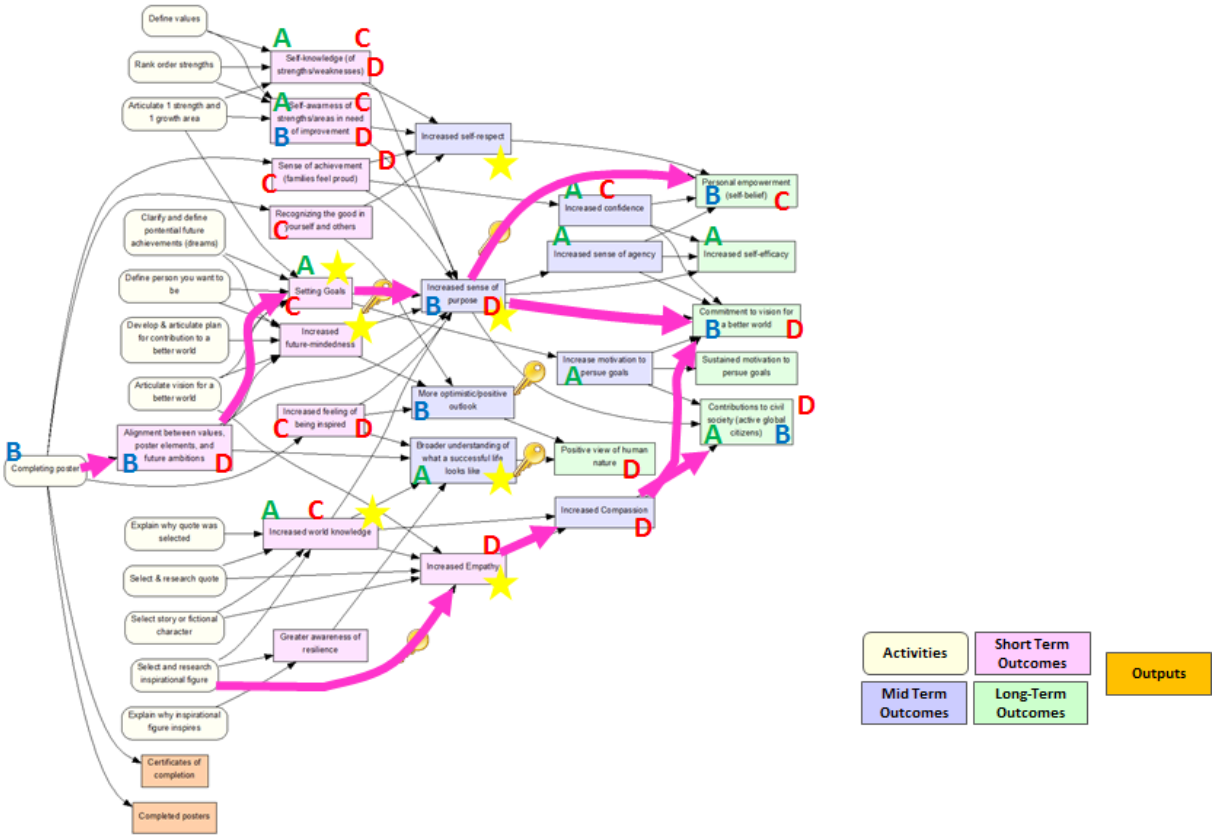

Figure 2. Inspire>Aspire Pathway Model after completing Mining the Model

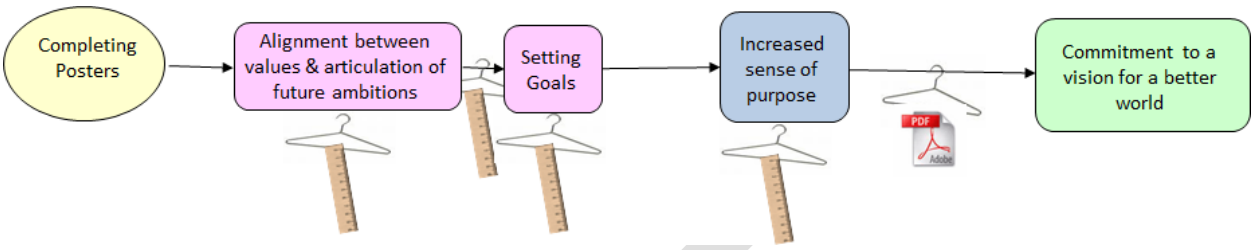

Figure 3. Illustration of the Golden Spike

References

Antes, G. (1998). Evidence-based medicine. *Internist, 39*(9), 899-908.

Archibald, T., Buckley, J., & Trochim, W. (2011, November). *Evaluative Thinking: What is It? Why Does it Matter? How Can We Measure It?* Paper presented at the Annual Conference of the American Evaluation Association, Anaheim, CA.

Artelt, T. A. (2001). Evidence-based social work practice with families: A lifespan approach. *Research on Social Work Practice, 11*(3), 404-405.

Asim, A., Essegaier, S., & Kohli, R. (2000). Internet recommendation systems. *Journal Of Marketing Research, 37*(3), 363-376.

Bammer, G. (2005). Integration and implementation sciences: Building a new specialization. *Ecology and Society, 10*(2).

Bertalanffy, L. V. (1995). *General system theory: Foundations, development, applications* (Rev. ed. ed.). New York: Braziller.

Beyer, B. (1987). *Practical Strategies for the Teaching of Thinking*. Boston, MA: Allyn and Bacon, Inc.

Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation, 33*(5-18).

Bloom, B. (Ed.). (1956). *The Taxonomy of Educational Objectives, The Classification of Educational Goals, Handbook I: Cognitive Domain.* White Plains, NY: Longman.

Bornstein, M. H. (2006). Parenting science and practice. In K. A. Renninger, I. E. Sigel & W. Damon (Eds.), *Handbook of Child Psychology: Child psychology in practice.* (6th ed., Vol. 4, pp. 893–949). Hoboken, NJ: Wiley.

Bradie, M., & Harms, W. (2006). Evolutionary Epistemology *The Stanford Encyclopedia of Philosophy* (Fall 2006 ed.).

Bradie, M., & Harms, W. (2006). Evolutionary Epistemology *The Stanford Encyclopedia of Philosophy* (Fall 2006 Edition ed.).

Brainard, J. (2002). White House Proposes Criteria for Measuring Progress and Results of Federally Financed Research. *Chronicle of Higher Education, 48*.

Brändtstadter, J. (1999). The self in action and development: cultural, biosocial, and ontogenetic bases of intentional self-development. In J. Brändtstadter & R. M. Lerner (Eds.), *Action and self-development: Theory and research through the live span* (pp. 37-65). Thousand Oaks, CA: Sage.

Brookfield, S. D. (1987). *Developing Critical Thinkers*. San Francisco, CA: Jossey-Bass Inc.

Brownson, R. C., Baker, E. A., Leet, T. L., & Gillespie, K. N. (Eds.). (2002). *Evidence-Based Public Health*: Oxford University Press.

Burke, R. (2000). Knowledge-based recommender systems. . In A. Kent (Ed.), *Encyclopedia of library and information systems* (Vol. 69, Supplement 32). New York: Marcel Dekker.

Campbell, D. T. (1969). Reforms As Experiments. *American Psychologist, 24*(4), 409-429.

Campbell, D. T. (1974). Evolutionary Epistemology. In P. A. Schilpp (Ed.), *The Philosophy of Karl Popper*. LaSalle, IL: Open Court Publishing Co.

Campbell, D. T. (1988). Evolutionary Epistemology. In E. S. Overman (Ed.), *Methodology and Epistemology for Social Science: Selected Papers of Donald T. Campbell*. Chicago: University of Chicago Press.

Campbell, D. T. (1991). Methods for the Experimenting Society. *Evaluation Practice, 12*(3), 223-260.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and Quasi-Experimental Designs for Research on Teaching. In N. L. Gage (Ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally.

Capra, F. (1997). *The web of life : a new synthesis of mind and matter*. London: Flamingo.

Capra, F. (2002). *The hidden connections : integrating the hidden connections among the biological, cognitive, and social dimensions of life* (1st ed.). New York: Doubleday.

Capra, F., Mill Valley School District (Mill Valley Calif.), & Elmwood Institute. (1994). *From the parts to the whole : systems thinking in ecology and education*. S.l.: s.n. ;.

Capra, F., Steindl-Rast, D., & New Dimensions Foundation. (1993). New paradigm thinking [sound recording /]. San Francisco, CA: New Dimensions Foundation.

Caracelli, V. (1989). Structured conceptualization: A framework for interpreting evaluation results. *Trochim, W. (ed.), Evaluation and Program Planning, 12 (1) p. 45-52, a special issue on Concept Mapping for Evaluation and Planning*.

Chalmers, I., & Haynes, B. (1994). Systematic Reviews - Reporting, Updating, and Correcting Systematic Reviews of the Effects of Health-Care. *British Medical Journal, 309*(6958), 862-865.

Chen, H., & Rossi, P. (1990). *Theory-Driven Evaluations*. Thousand Oaks, CA: Sage.

Chen, H., & Rossi, P. H. (1983). Evaluating with sense: The theory-driven approach. *Evaluation Review, 7*, 283-302.

Chinnock, P. (2008). Knowledge Translation Toolkit: A Resource for Researchers. Retrieved July 30, 2010, from http://www.tropika.net/svc/specials/KT-Toolkit/pages/KT-Toolkit

Chow, S., & Liu, J. (2004). *Design and Analysis of Clinical Trials: Concepts and Methodologies*. Hoboken, NJ: Wiley.

Compton, D. W., Baizerman, M., & Stockdill, S. H. (Eds.). (2002). *The Art, Craft, and Science of Evaluation Capacity Building* (Vol. 93). San Francisco: Jossey Bass.

Cook, D. J., Mulrow, C. D., & Haynes, R. B. (1997). Systematic reviews: Synthesis of best evidence for clinical decisions. *Annals of Internal Medicine, 126*(5), 376-380.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis for Field Settings*. Boston: Houghton Mifflin Company.

Cooksy, L. J., Gill, P., & Kelly, P. A. (2001). The program logic model as an integrative framework for a multimethod evaluation. *Evaluation And Program Planning, 24*(2), 119-128. doi: 10.1016/s0149-7189(01)00003-9

Cooksy, L. J., Mark, M. M., & Trochim, W. (2009). Evaluation Policy and Evaluation Practice: Where Do We Go From Here? *New Directions for Evaluation, 123*, 103-110.

Cronje, R., & Fullan, A. (2003). Evidence-based medicine: toward a new definition of 'rational' medicine. *Health, 7*(3), 353-369.

Cziko, G. A., & Campbell, D. T. (1990). Comprehensive Evolutionary Epistemology Bibliography. *Journal of Social and Biological Structures, 13*(1), 41-82.

Darwin, C. (1859). *On the Origin of Species by means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Davidson, E. J. (2005). *Evaluative Thinking and Learning-Enabled Organizational Cultures*. Paper presented at the Canadian Evaluation Society & American Evaluation Association Conference, Toronto, Ontario.

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies, 47*(2), 108-121.

Davies, P. (2000). The relevance of systematic reviews to educational policy and practice. *Oxford Review of Education, 26*(3-4), 365-378.

Dickersin, K. (2002). Systematic reviews in epidemiology: why are we so far behind? *International Journal of Epidemiology, 31*(1), 6-12.

Dickersin, K., Scherer, R., & Lefebvre, C. (1994). Systematic Reviews - Identifying Relevant Studies for Systematic Reviews. *British Medical Journal, 309*(6964), 1286-1291.

Dopson, S., Locock, L., Gabbay, J., Ferlie, E., & Fitzgerald, L. (2003). Evidence-based medicine and the implementation gap. *Health, 7*(3), 311-330.

Eitel, F., & Steiner, S. (1999). Evidence-based learning. *Medical Teacher, 21*(5), 506-512.

Evans, J., & Benefield, P. (2001). Systematic reviews of educational research: does the medical model fit? *British Educational Research Journal, 27*(5), 527-541.

Fetterman, D., Kaftarian, S. J., & Wandersman, A. (1996). *Empowerment Evaluation: Knowledge and Tools for Self-Assessment and Accountability*. Thousand Oaks, CA: Sage.

Fetterman, D., & Wandersman, A. (Eds.). (2005). *Empowerment evaluation principles in practice*. New York, NY: Guilford Press.

Forrester, J. W. (1961). *Industrial dynamics*. Cambridge, MA: MIT Press.

Forrester, J. W. (1994). *Learning through system dynamics as preparation for the 21st century.* Paper presented at the Systems Thinking and Dynamic Modeling Conference for K-12 Education, Concord, MA.

Forrester, J. W. (1997). Roadmaps: A guide to learning system dynamics. Retrieved March 18, 2003, 2003, from http://web.mit.edu/sdg/www/roadmaps.html

Foundation, W. K. K. (2004). *W.K. Kellogg Foundation logic model development guide*. Battle Creek, MI: W.K. Kellogg Foundation.

Foundation, W. K. K. (December 2001). Logic Model Development Guide.

Francois, C. (2004). *International Encyclopedia Of Systems And Cybernetics* (2nd edition ed.). Munchen: K G Saur North Amer Research.

Frechtling, J. (2002). *The 2002 User Friendly Handbook for Project Evaluation*. Alexandria, VA.: The National Science Foundation.

Frechtling, J., & Sharp, L. (Eds.). (1997). *User-Friendly Handbook for Mixed-Methods Evaluations*. Alexandria, VA: National Science Foundation.

Frechtling, J. A. (2007). *Logic Modeling Methods in Program Evaluation*. San Francisco, CA: Jossey-Bass.

Freund, A. M., & Baltes, P. B. (2002). Life-management strategies of selection, optimization and compensation: Measurement by self-report and construct validity. *Journal of Personality and Social Psychology, 82*(4), 642-662.

Gallagher, E. J. (1999). Systematic reviews: A logical methodological extension of evidence-based medicine. *Academic Emergency Medicine, 6*(12), 1255-1260.

Gell-Mann, M. (1995). Let's Call It Plectics. *Complexity, 1*(5).

Gell-Mann, M. (2003). *The quark and the jaguar : adventures in the simple and the complex*. London: Abacus.

Gestsdóttir, S., Bowers, E., von Eye, A., Napolitano, C. M., & Lerner, R. M. (2010). Intentional self regulation in middle adolescence: The emerging role of loss-based selection in positive youth development. *Journal of Youth and Adolescence, 39*(7), 764-782.

Gibbs, L. E. (2003). *Evidence-Based Practice for the Helping Professions*. Pacific Grove, CA: Thomson, Brooks-Cole.

Gilbert, G. N., & Troitzsch, K. G. (1999). *Simulation for the social scientist*. Buckingham ; Philadelphia, Pa.: Open University Press.

Gottlieb, G., Wahlsten, D., & Lickliter, R. (2006). The significance of biology for human development: A developmental psychobiological systems view. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology* (Vol. 1, pp. 210-257). Hoboken, NJ: Wiley.

Greene, J. C., & Caracelli, V. J. (Eds.). (1997). *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms.* (Vol. 74). San Francisco, CA: Jossey-Bass.

Hasselblad, V. (1998). Meta-analysis of multitreatment studies. *Medical Decision Making, 18*(1), 37-43.

Holland, J. (1995). *Hidden Order: How Adaptation Builds Complexity*. Reading, MA: Perseus Books.

Holland, J. H. (1998). *Emergence : from chaos to order*. Reading, Mass.: Addison-Wesley.

Johnson, S. (2001). *Emergence : the connected lives of ants, brains, cities, and software*. New York: Scribner.

Juni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care - Assessing the quality of controlled clinical trials. *British Medical Journal, 323*(7303), 42-46.

Kane, M., & Trochim, W. (2006). *Concept Mapping for Planning and Evaluation*. Thousand Oaks, CA: Sage Publications.

Kauffman, S. A. (1993). *The origins of order : self-organization and selection in evolution*. New York: Oxford University Press.

Kauffman, S. A. (1995). *At home in the universe : the search for laws of self-organization and complexity*. New York: Oxford University Press.

Kellogg Foundation. (2001). *Logic Model Development Guide: Using Logic Models to bring together Planning, Evaluation and Action*. Battle Creek, Michigan: W. K. Kellogg Foundation.

King, J. A., Morris, L. L., & Fitz-Gibbon, C. T. (1987). *How to Assess Program Implementation*.

Koretz, R. L. (2002). Methods of meta-analysis: an analysis. *Current Opinion in Clinical Nutrition and Metabolic Care, 5*(5), 467-474.

Kuhn, D. (2005). *Education for Thinking*. Cambridge, MA: Harvard University Press.

Labin, S. N. (2014). Developing common measures in evaluation capacity building an iterative science and practice process. *American Journal of Evaluation, 35*(1), 107-115.

Laszlo, E. (1996). *The Systems View of the World: A Holistic Vision for Our Time (Advances in Systems Theory, Complexity, and the Human Sciences)*. Creskill, NJ: Hampton Press.

Lerner, R. M. (2006). Developmental science, developmental systems, and contemporary theories of human development. In R. M. Lerner & W. Damon (Eds.), *Handbook of Child Psychology: Theoretical models of human development* (6th ed., Vol. 1, pp. 1-17). Hoboken, NJ: John Wiley & Sons.

Linstone, H. A., & Turoff, M. (1975). The Delphi method: Techniques and applications. *Addison-Wesley Publishing Company, Reading, MA*.

Lipsey, M., & Wislson, D. B. (2000). *Practical Meta-Analysis*. Newbury Park, CA: Sage.

Macaulay, A. C. (1999). Participatory research maximizes community and lay involvement. *British Medical Journal, 319*(7212), 774-778.

Marathe, J. (1999). *Creating community online*: Durlacher Research, Ltd.

Mark, M. M., Cooksy, L. J., & Trochim, W. (2009). Evaluation Policy: An Introduction and Overview. *New Directions for Evaluation, 123*, 3-12.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition : the realization of the living*. Dordrecht, Holland ; Boston: D. Reidel Pub. Co.

Mayne, J. (2010). Building an evaluative culture: The key to effective evaluation and results management. *The Canadian Journal of Program Evaluation, 24*(2), 1-30.

Mayo Clinic. (2007). Clinical trials: A chance to try evolving therapies. from http://www.mayoclinic.com/health/clinical-trials/DI00033. Accessed June 26, 2007

Mayr, E. (2001). *What Evolution Is*. New York, NY: Basic Books.

McAlister, F. A., Graham, I., Karr, G. W., & Laupacis, A. (1999). Evidence-based medicine and the practicing clinician. *Journal of General Internal Medicine, 14*(4), 236-242.

McKelvey, B. (1999). Complexity Theory in Organization Science: Seizing the Promise or Becoming a Fad? *Emergence, 1*(1), 5-32.

McLaughlin, J., & Jordan, G. (1999). Logic models: A tool for telling your program's performance story. *Evaluation and Program Planning, 22*(65-72).

Midgley, G. (2003). *Systems thinking*. Thousand Oaks, CA: SAGE.

Montori, V. M., & Guyatt, G. H. (2002). What is evidence-based medicine? *Endocrinology and Metabolism Clinics of North America, 31*(3), 521-526.

Moran, S. (2014). What "purpose" means to youth: Are there cultures of purpose? *Applied Developmental Science, 18*(3), 163-175. doi: 10.1080/10888691.2014.924359

National Institutes of Health. (2007). What is a protocol? Retrieved December, 2007, from http://www.nlm.nih.gov/services/ctprotocol.html

National Science Foundation. (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. Arlington, VA: Retrieved from http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf.

Newman, M. E. J. (2003). The structure and function of complex networks. *arXiv, 1*.

O'Fallon, L. R., & Dearry, A. (2002). Community-based participatory research as a tool to advance environmental health sciences. *Environmental Health Perspectives, 110*, 155-159.

O'Sullivan, R. G. (2004). *Practicing evaluation: A collaborative approach*. Thousand Oaks, CA: Sage.

Overton, W. F. (2006). Developmental psychology: Philosophy, concepts, methodology. In R. M. Lerner (Ed.), *Theoretical models of human development* (6th ed., Vol. 1, pp. 18-88). Hoboken, NJ: Wiley.

Overton, W. F. (2010). Life-span development: Concepts and issues. In W. F. Overton & R. M. Lerner (Eds.), *Handbook of life-span development: Cognition, biology, and methods* (Vol. 1 pp. 1-29). Hoboken, NJ: Wiley.

Overton, W. F. (2013). Relationism and Relational Developmental Systems: A Paradigm for Developmental Science in the Post-Cartesian Era. In R. M. Lerner & J. B. Benson (Eds.), *Advances in Child Development and Behavior* (Vol. 44, pp. 21-64).

Patton, M. (2005). Interview with Lisa Waldick, from the International Development Research Center. Retrieved August 25, 2005, from http://www.idrc.ca/en/ev-30442-201-1-DO_TOPIC.html

Perkins, D. N. (1986). *Knowledge as Design*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pirrie, A. (2001). Evidence-based practice in education: The best medicine? *British Journal of Educational Studies*, *49*(2), 124-136.

Pocock, S. (2004). *Clinical Trials: A Practical Approach*. New York, NY: Wiley.

Popper, K. (1973a). *Evolutionary Epistemology*. Paper presented at the Sections I-VI of "The Rationality of Scientific Revolutions" given at the Herbert Spencer Lecture at the University of Oxford, Oxford.

Popper, K. (1973b). *Evolutionary Epistemology*. Paper presented at the Sections I-VI of "The Rationality of Scientific Revolutions" given at the Herbert Spencer Lecture, University of Oxford.

Popper, K. (1985). Evolutionary Epistemology. In D. M. Miller (Ed.), *Popper selections* (pp. 78-86). Princeton, NJ: Princeton University Press.

Preskill, H. (2008). Evaluation's Second Act: A Spotlight on Learning. *American Journal Of Evaluation, 29*(2), 127-138.

Procopis, P. G. (2002). Evidence-based medicine. *Developmental Medicine and Child Neurology, 44*(7), 435.

Ragsdell, G., West, D., & Wilby, J. (2002). *Systems theory and practice in the knowledge age*. New York: Kluwer Academic/Plenum Publishers.

Reason, P., & Bradbury, H. (Eds.). (2001). *Handbook of Action Research: Participative Inquiry and Practice*. London: Sage.

Resnick, M. (1994). *Turtles, termites, and traffic jams : explorations in massively parallel microworlds*. Cambridge, Mass.: MIT Press.

Richardson, G. P. (1991). *Feedback thought in social science and systems theory*. Philadelphia: University of Pennsylvania Press.

Richardson, G. P. (1996). Problems for the future of system dynamics. *System Dynamics Review, 12*(2), 141-157.

Rodriguez-Campos, L. (2005). *Collaborative evaluations*. Tamarac, FL: Llumina Press.

Roth, J. L., & Brooks-Gunn, J. (2003). Youth development programs: Risk, prevention, and policy. *Journal of Adolescent Health, 32*(3), 170-182.

Roth, J. L., & Brooks-Gunn, J. (2015). Evaluating youth development programs: Progress and promise. *Applied Developmental Science*, 1-15. doi: 10.1080/10888691.2015.1113879

Roth, J. L., Brooks-Gunn, J., Murray, L., & Foster, W. (1998). Promoting healthy adolescents: Synthesis of youth development program evaluations. *Journal of Research on Adolescence, 8*(4), 423-459.

Sackett, D. L., Richardson, S., Rosenberg, W., & Haynes, R. B. (1997). *Evidence-based Medicine: How to Practice and Teach EBM*. London: Churchill Livingstone.

Sanders, J. R. (2002). Presidential address: On mainstreaming evaluation. *American Journal of Evaluation, 23*, 253-259.

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Shadish, W. R. (1986). Planned Critical Multiplism - Some Elaborations. *Behavioral Assessment, 8*(1), 75-103.

Shannon, C. (1948). A mathematical theory of communication" *Bell System Technical Journal, 27*(July and October), 379-423 and 623-656.

Silagy, C. A., Middleton, P., & Hopewell, S. (2002). Publishing Protocols of systematic reviews - Comparing what was done to what was planned. *Jama-Journal Of The American Medical Association, 287*(21), 2831-2834.

Silagy, C. A., Stead, L. F., & Lancaster, T. (2001). Use of systematic reviews in clinical practice guidelines: case study of smoking cessation. *British Medical Journal, 323*(7317), 833-836.

Smith, M. L., & Glass, G. V. (1977). Meta-Analysis of Psychotherapy Outcome Studies. *American Psychologist, 32*(9), 752-760.

Society, S. D. (2002). What is system dynamics? Retrieved December 19, 2002, 2002, from http://www.systemdynamics.org/

Sohn, D. (1996). Meta-analysis and science. *Theory & Psychology, 6*(2), 229-246.

Sontag, E. D. (1998). *Mathematical Control Theory: Deterministic Finite Dimensional Systems. Second Edition*. New York: Springer.

Sterman, J. (2001). System dynamics modeling: tools for learning in a complex world. *California Management Review, 43*(4), 8-25.

Sterman, J. D. (2000). *Business Dynamics: Systems Thinking and Modeling for a Complex World*. New York: McGraw-Hill/Irwin.

Stockdill, S. H., Baizerman, M., & Compton, D. W. (2002). Toward a definition of the ECB process: A conversation with the ECB literature. In D. W. Compton, M. Baizerman & S. H. Stockdill (Eds.), *The Art, Craft and Science of Evaluation Capacity Building: New Directions for Evaluation* (Vol. 93, pp. 7-25).

Strogatz, S. H. (1994). *Nonlinear dynamics and Chaos : with applications to physics, biology, chemistry, and engineering*. Reading, Mass.: Addison-Wesley Pub.

Strogatz, S. H. (2003). *Sync : the emerging science of spontaneous order* (1st ed.). New York: Hyperion.

Suarez-Balcazar, Y., & Taylor-Ritzler, T. (2014). Moving from science to practice in evaluation capacity building. *American Journal of Evaluation, 35*(1), 95-99.

Trochim, W. (1989a). An Introduction to Concept Mapping for Planning and Evaluation. *Evaluation and Program Planning, 12*(1), 1-16.

Trochim, W. (1991, June 16-19, 1991). *Developing an Evaluation Culture in International Agriculture Research.* Paper presented at the Cornell Institute on International Food, Agriculture and Development's (CIIFAD) Workshop on the Assessment of International Agricultural Research Impact for Sustainable Development, Cornell University, Ithaca NY.

Trochim, W. (2006). An Evaluation Culture. Retrieved 1/2/2012, 2012, from http://www.socialresearchmethods.net/kb/evalcult.php

Trochim, W. (2007). *Evolutionary Perspectives in Evaluation: Theoretical and Practical Implications*. Paper presented at the 30th Annual Conference of the Eastern Evaluation Research Society, Absecon, NJ. http://www.socialresearchmethods.net/research/EERS2007/Evolutionary%20Perspectives%20in%20Evaluation%20Theoretical%20and%20Practical%20Implications.pdf

Trochim, W. (2009). Evaluation Policy and Evaluation Practice. *New Directions for Evaluation, 123*, 13-32.

Trochim, W. (Ed.). (1989b). *Concept Mapping for Evaluation and Planning* (Vol. 12, 1). New York, NY: Pergamon.

Trochim, W., & Kane, M. (2005). Concept mapping: An introduction to structured conceptualization in health care. *International Journal for Quality in Health Care, 17*(3), 187-191.

Trochim, W., & Linton, R. (1986). Conceptualization for Planning and Evaluation. *Evaluation and Program Planning*, 289-308.

Trochim, W., Mark, M. M., & Cooksy, L. J. (Eds.). (2009). *Evaluation Policy and Evaluation Practice*. Hoboken, NJ: Jossey Bass.

Trochim, W., Milstein, B., Wood, B., Jackson, S., & Pressler, V. (2004). Setting Objectives for Community and Systems Change: An Application of Concept Mapping for Planning a Statewide Health Improvement Initiative. *Health Promotion Practice, 5*(1), 8-19.

Trochim, W., Urban, J. B., Hargraves, M., Hebbard, C., Buckley, J., Archibald, T., . . . Burgermaster, M. (2012). *The guide to the systems evaluation protocol*. Ithaca, NY: Cornell Digital Print Services.

U.S. Department of Education. (2007). *Report of the Academic Competitiveness Council*. Washington, D.C.: U.S. Department of Education.

Urban, J. B., Archibald, T., Hebbard, C., Burgermaster, M., Barrios, V., & Trochim, W. M. (2015). *Results and implications of using the SEP and the mySEP in evaluation planning and implementation for STEM outreach programs*. Paper presented at the Eastern Evaluation Research Society, Absecon, NJ.

Urban, J. B., Hargraves, M., Hebbard, C., Burgermaster, M., & Trochim, W. M. (2011). *Evaluation in the context of lifecycles: 'A place for everything, everything in its place'*. Paper presented at the American Evaluation Association, Anaheim, CA.

Urban, J. B., Hargraves, M., & Trochim, W. (2014). Evolutionary Evaluation: Implications for evaluators, researchers, practitioners, funders and the evidence-based program mandate. *Evaluation And Program Planning, 45*, 127-139. doi: 10.1016/j.evalprogplan.2014.03.011

Urban, J. B., & Trochim, W. M. (2009). The role of evaluation in research-practice integration: Working toward the "golden spike". *American Journal of Evaluation, 30*(4), 538-553.

Waldrop, M. M. (1992). *Complexity: the emerging science at the edge of order and chaos*. New York: Simon & Schuster.

Wandersman, A. (2014). Moving forward with the science and practice of evaluation capacity building (ECB): The why, how, what, and outcomes of ECB. *American Journal of Evaluation, 35*(1), 87-89.

Watts, D. J. (1999a). *Small worlds : the dynamics of networks between order and randomness*. Princeton, N.J.: Princeton University Press.

Watts, D. J. (1999b). *Small Worlds: The Dynamics of Networks Between Order and Randomness*. Princeton, NJ: Princeton University Press.

Watts, D. J. (2003). *Six degrees : the science of a connected age* (1st ed.). New York: Norton.

Wikipedia. (2007). Web 2.0. Retrieved December 18, 2007, from http://en.wikipedia.org/wiki/Web_2

Witkin, B. R. A., J.W. (1995). *Planning and Conducting Needs Assessments: A Practical Guide*. Thoudans Oaks, CA: Sage.

Wolf, F. M. (1986). *Meta-Analysis: Quantitative Methods for Research Synthesis.* Newbury Park, CA: Sage.

Wolfram, S. (2002). *A new kind of science*. Champaign, IL: Wolfram Media.

Wortman, P. M., Smyth, J. M., Langenbrunner, J. C., & Yeaton, W. H. (1998). Consensus among experts and research synthesis. A comparison of methods. *Int J Technol Assess Health Care, 14*(1), 109-122.

Wye, C. G., & Sonnichsen, R. C. (Eds.). (1992). *Evaluation in the Federal Government: Changes, Trends, and Opportunities* (Vol. 55). San Francisco: Jossey-Bass Publishers.

Yeager, D. S., Bundick, M. I., & Johnson, R. (2012). The role of future work goal motives in adolescent identity development: A longitudinal mixed-methods investigation. *Contemporary Educational Psychology, 27*(3), 206-217.

Yeomans, J. M. (1992). *Statistical mechanics of phase transitions*. New York: Oxford University Press.

Young, G. (1999). A piece of the main: Parts and wholes in ecology and human ecology. *Advances in Human Ecology, 8*, 1-31.

Young, J. F. (1969). *Cybernetics*. New York: American Elsevier Pub. Co.