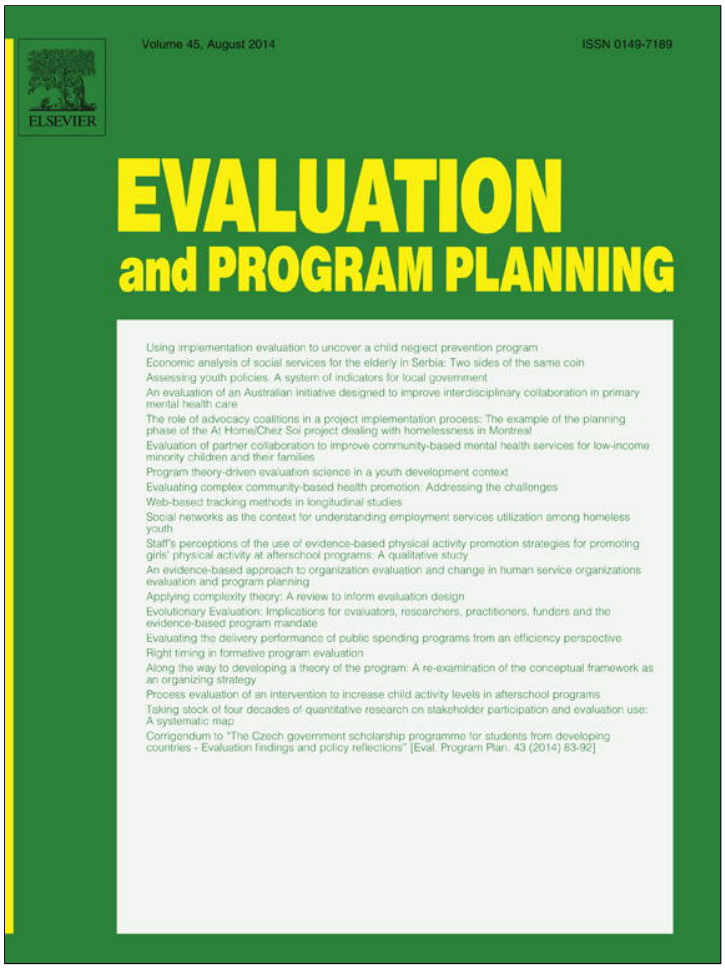

Evaluation and Program Planning 45 (2014) 127–139

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Evolutionary Evaluation: Implications for evaluators, researchers, practitioners, funders and the evidence-based program mandate

Jennifer Brown Urban [a,*], Monica Hargraves [b], William M. Trochim [c]

[a] Department of Family and Child Studies, Montclair State University, 1 Normal Avenue, University Hall 4144, Montclair, NJ 07043, United States
[b] Cornell Office for Research on Evaluation, Cornell University, 2301 Martha Van Rensselaer Hall, Ithaca, NY 14853, United States
[c] Cornell Office for Research on Evaluation, Department of Policy Analysis and Management, Cornell University, 2301 Martha Van Rensselaer Hall, Ithaca, NY 14853, United States

ARTICLE INFO

Article history:
Received 6 March 2013
Received in revised form 20 March 2014
Accepted 20 March 2014
Available online 30 March 2014

Keywords:
Evolutionary Evaluation
Program evolution
Evaluation design
Experimental design
Randomized controlled trial (RCT)
Validity
Evidence-based program (EBP)
Evolutionary theory
Developmental systems theory
Evolutionary epistemology
Lifecycles
Systems evaluation

ABSTRACT

Evolutionary theory, developmental systems theory, and evolutionary epistemology provide deep theoretical foundations for understanding programs, their development over time, and the role of evaluation. This paper relates core concepts from these powerful bodies of theory to program evaluation. Evolutionary Evaluation is operationalized in terms of program and evaluation evolutionary phases, which are in turn aligned with multiple types of validity. The model of Evolutionary Evaluation incorporates Chen's conceptualization of bottom-up versus top-down program development. The resulting framework has important implications for many program management and evaluation issues. The paper illustrates how an Evolutionary Evaluation perspective can illuminate important controversies in evaluation using the example of the appropriate role of randomized controlled trials that encourages a rethinking of "evidence-based programs". From an Evolutionary Evaluation perspective, prevailing interpretations of rigor and mandates for evidence-based programs pose significant challenges to program evolution. This perspective also illuminates the consequences of misalignment between program and evaluation phases; the importance of supporting both researcher-derived and practitioner-derived programs; and the need for variation and evolutionary phase diversity within portfolios of programs.

This paper offers a way of thinking about program development that has deep theoretical foundations and casts new light on some of the major contemporary controversies in evaluation and applied social research. Specifically, Evolutionary Evaluation draws on theories of evolution, developmental systems, and epistemology to articulate a view of program development and evaluation as evolutionary processes with inherent lifecycle qualities. When programs are understood in this way, there are powerful implications for strategic decision making regarding the management and evaluation of existing individual programs and – notably – portfolios of programs; for the imperative of sustaining a large stream of diverse, even emergent programs from varied sources; and ultimately for our investments in knowledge and innovation altogether.

In the sections that follow we: (1) present the theoretical foundations for an evolutionary view of program development and evaluation; (2) operationalize this perspective by defining program and evaluation evolutionary phases and discussing the issue of alignment as a key consideration in ensuring optimal decision-making regarding programs and their evaluation; and (3) link these to the current controversy over evidence-based programming by proposing a more comprehensive definition of what constitutes sufficient evidence. The framework presented here has a number of important implications for program practitioners, researchers, and funders and we explore some of these in a brief conclusion.

Of the many implications of Evolutionary Evaluation, we focus here on the appropriate role for experimental designs and the currently prevailing standards of evidence because these pose the largest contemporary challenge to programming, especially for social and educational programs, and to program evolution. These issues have significant historical roots: one of the major controversies in applied research and evaluation over the past century has centered around randomized controlled trials (RCTs)

* Corresponding author. Tel.: +1 973 655 6884.
E-mail addresses: urbanj@mail.montclair.edu (J.B. Urban), mjh51@cornell.edu (M. Hargraves), wmt1@cornell.edu (W.M. Trochim).

http://dx.doi.org/10.1016/j.evalprogplan.2014.03.011
0149-7189/

and, in its more recent manifestations, the definition of evidence-based programs (EBPs). We argue that the evidence-based label is being applied to programs prematurely and that the definition of EBPs needs to consider multiple types of validity and the importance of methodological pluralism.

We begin with a discussion of the theoretical foundations for Evolutionary Evaluation. First, we present the concept of evolutionary epistemology which applies biological theories of evolution to the development and progression of knowledge and ideas. We extend this reasoning to program development and evaluation, highlighting the critical role that evaluation plays in the variation, selection, and retention of programs. The application of evolutionary reasoning to programs is further supported by the concepts of ontogeny and phylogeny including insights gained from developmental systems science. Ontogeny and phylogeny are typically terms reserved for the evolution of organisms and species, respectively; however we will describe how the concepts can be applied to programs and to portfolios of programs.

1. Theoretical foundations

The foundations for Evolutionary Evaluation can be found in the fields of evolutionary theory, natural selection (Darwin, 1859; Mayr, 2001), evolutionary epistemology (Bradie & Harms, 2006; Campbell, 1974, 1988; Cziko & Campbell, 1990; Popper, 1973, 1985), developmental systems theory (e.g., Lerner, 2002, 2006; Overton, 2006, 2010), ecology (Molles, 2001; Pickett, Kolasa, & Jones, 1994; Richerson, Mulder, & Vila, 1996) and systems theory (Bertalanffy, 1972; Laszlo, 1996; Midgley, 2003; Ragsdell, West, & Wilby, 2002). These are foundational theories in the life and developmental sciences. Here we show that these theories can be applied directly to programs and how they develop, providing a basis for thinking about how programs evolve over time.

1.1. Evolutionary epistemology

Evolutionary epistemology applies the concepts of biological evolution to the growth and development of human knowledge. The term evolutionary epistemology was reportedly coined by one of the leading thinkers in evaluation, Donald T. Campbell, and the field was initially developed by him and the philosopher of science Sir Karl Popper (1973, 1975, 1985). In his essay entitled *Evolutionary Epistemology*, Campbell (1974, 1988) argued that "...evolution – even in its biological aspects – is a knowledge process, and that the natural-selection paradigm for such knowledge increments can be generalized to other epistemic activities, such as learning, thought and science" (Campbell, 1988, p. 393). Campbell is not suggesting evolution as a metaphor for learning, thinking or science; he is asserting that evolution *is* the fundamental process for all of these. Additionally, he is making the argument that biological evolution itself can perhaps most aptly be viewed as a knowledge process. Toulmin makes the same point: "In talking about the development of natural science as 'evolutionary,' I have not been employing a mere *facon de parler*, or analogy, or metaphor. The idea that the historical changes by which scientific thought develops frequently follow an 'evolutionary' pattern needs to be taken quite seriously; and the implications of such a pattern of change can be, not merely suggestive, but explanatory" (Toulmin, 1967, p. 470).

In his identically titled paper *Evolutionary Epistemology*, Popper (1985) describes three levels of evolution: "genetic adaptation, adaptive behavioral learning, and scientific discovery, which is a special case of adaptive behavioral learning" and argues that for all three "the mechanism of adaptation is fundamentally the same" (Popper, 1985, p. 78–79). Of course, that mechanism is the process of natural selection (whereby traits or features that offer the greatest "fitness" to the environment tend to prevail over time as organisms without those advantageous characteristics tend not to survive or reproduce as successfully). Popper notes that all three levels of evolution share an inherited structure. At the genetic level it is obvious that the inherited structure is the genome. However, it may be less obvious at the behavioral level that there is also an inherited structure – "the innate repertoire of the types of behavior which are available to the organism" (Popper, 1985, p. 79). Perhaps most intriguingly, the corresponding 'inherited' structure in science consists of the "dominant scientific conjectures and theories" that get passed down through academia and distributed throughout communities of researchers. For those who are accustomed to thinking of evolution as something that applies only to biology or genetics, it may initially be somewhat disorienting to accept that both Popper and Campbell are saying that ideas and knowledge follow the exact same process as biological species.

The central thrust of this argument is that our knowledge, including our macro-level knowledge of interventions and programs, evolves according to the evolutionary principles of ontogeny (development of an organism over its lifespan), phylogeny (evolution of a species over time), natural selection, and the trial-and-error cycle of (blind) variation and selective retention (for example, genetic mutations that survive and persist, or disappear). Over time, program variations are tried and survive or not according to current socially (usually unconscious) negotiated selection mechanisms. Instead of the commitment to preserving a program as it is, this perspective encourages recognition that individual programs, like organisms, have a finite life-span, that they should not be assumed to have an infinite horizon, that it is normal to see them as part of an ongoing trial-and-error effort, that they should not be expected to function at a mature level when they are first "born" or initiated, and that the abandonment of an older program and the development of new ones is part of the normal cycle-of-life. From a program's inception and throughout its life course, the focus is on where the program is in its development and how it can be moved along to the next phase in development or abandoned for a better program alternative.

1.2. Ontogeny and the evolution of programs

One of the evolutionary concepts that needs to be re-interpreted in terms of programs is the idea of ontogeny. Ontogeny refers to the development of an organism through different stages or phases over its life course (i.e., in humans: infancy, childhood, adolescence, adulthood). Developmental systems theory recognizes that ontogeny describes a change process that is not necessarily anchored in chronological time or associated with age (e.g., Lerner, 2002, 2006; Overton, 2006, 2010). Age typically serves as a proxy variable for change or development, and is used for convenience or ease of measurement rather than because it has a direct link to the developmental phenomenon of interest. This variability can be seen around the acquisition of any new developmental skill. For example, some children will begin talking as early as 12 months-old while others will not talk until they are 24 months-old.

Moreover, the developmental process is not necessarily linear. Stage theories (e.g., Freud's theory of psychosexual development, Erickson's theory of psychosocial development, Sullivan's theory of interpersonal development, Kohlberg's theory of moral development) which dominated the developmental literature in the early to mid-20th century tended to compartmentalize development into distinct circumscribed phases and individuals were expected to transition through the phases in lock-step. More recently, developmental theory has rejected a stage theory approach and recognizes that development is not described well by abrupt

qualitative shifts. Rather, it tends to be gradual and progressive with both large and small shifts and at times may be characterized by the temporary loss of previously acquired skills. For example, Kohlberg (1963, 1984) outlined three broad stages for the development of moral reasoning (pre-conventional, conventional, and post-conventional) with pre-conventional being the least sophisticated stage of moral reasoning and post-conventional being the most sophisticated. Moral reasoning typically follows a pattern whereby reasoning at one level is fairly consolidated (e.g., an individual primarily reasons at a pre-conventional level) followed by periods of transition and variability (e.g., an individual demonstrates reasoning that includes elements of both pre-conventional and conventional reasoning), followed again by a period of consolidation at a higher level (e.g., an individual primarily reasons at a conventional level) (Walker, Gustafson, & Hennig, 2001). This means that at any given point in time, an individual may display characteristics of more than one level of reasoning.

Just as there is variability in the timing and manifestation of developmental milestones, there is also variability in the extent to which any developmental skill is mastered. Empirical research demonstrates that pre-conventional reasoning typically emerges in childhood, conventional reasoning emerges in early adolescence, and post-conventional reasoning emerges in late adolescence or early adulthood *if it emerges at all*. In fact, many people never achieve this most sophisticated level of moral reasoning and remain at the conventional stage all of their lives (Colby, Kohlberg, & Lieberman, 1983).

We also know from developmental systems theory that developmental change is characterized by a bi-directional person↔environment interaction. In the past, developmental science has focused on a dichotomous view of development (e.g., nature vs. nurture, continuity vs. discontinuity). The commonly held view now is that developmental change is driven by the bi-directional interaction between the individual and his/her environment.

Developmental systems theory can also contribute to our understanding of program development. Similar to the development of organisms, programs can also be described in terms of ontogenetic development. Programs are rarely static entities; rather they develop and grow at varying rates over the course of time. Just as we characterize human development into broad phases (e.g., infancy, childhood, adolescence, early adulthood, etc.) we can similarly discuss the development of programs in terms of broad phases (see section on program evolution phase definitions below and Fig. 1). Each program has its own individual life, a unique life course that moves through the various phases. Programs are born or initiated, typically either in practice-based settings or as the product of a formal research and development process. They may grow and change as they are implemented and revised. They may "linger" in a particular phase as program changes are integrated, and they may even cycle back to an earlier phase if the changes in the program or surrounding environment are substantial enough. They may mature and reach a relatively stable state, sometimes becoming routinized and standardized. They may regenerate in signifi-

	Program Evolution	Phase	Evaluation Evolution	
Initiation	Program is in *initial implementation(s)*, either as a brand new program or as an adaptation of an existing program.	I-A	Examines *implementation, participant and facilitator satisfaction*. Uses process and participant *documentation* and assessment and *post-only evaluation of reactions and satisfaction*.	Process & response
	Program still undergoing *rapid or substantial change/adaptation* or revision, after initial trials.	I-B	Focuses on *implementation*, and increasingly on *presence or absence of selected outcomes*. Evaluation is *post-only*; outcome measures may be under development with attention to internal consistency (reliability).	
Development	*Scale and scope of revisions or changes/adaptations are smaller;* most program elements are still evolving while a few may be implemented consistently.	II-A	Examines *program's association with change in group outcomes,* for these participants in this context. Uses *unmatched pre- and post-test of outcomes*, quantitative/qualitative assessment of change, assessment of measure reliability and validity.	Change
	Most program *elements* are *implemented consistently*; minor changes may still take place as some elements may still be evolving.	II-B	Examines *program's association with change in group (and/or individual) outcomes,* for these participants in this context. Uses *matched pre- and post-test of outcomes*, quantitative/qualitative assessment of change, verifying measure reliability and validity.	
Stability	*Program is implemented consistently;* participant experience from one implementation to the next is relatively stable (formal lessons or curricula exist).	III-A	Assesses *effectiveness* using design and statistical controls and comparisons (*control groups, control variables or statistical controls*).	Comparison & Control
	Program has *formal written procedures/protocol* and can be implemented consistently by new well-trained facilitators.	III-B	Assesses *effectiveness* using *controlled experiments or quasi-experiments (randomized experiment; regression-discontinuity).*	
Dissemination	Program is being *implemented in multiple sites*.	IV-A	Examines *outcome effectiveness across wider range of contexts*. Multi-site analysis of integrated large data sets over multiple waves of program implementation.	Generalizability
	Program is *fully protocolized and is being widely distributed*.	IV-B	Formal assessment across multiple program implementations that enable general assertions about this program in a wide variety of contexts (e.g., meta-analysis).	

Fig. 1. Program and evaluation evolutionary phase definitions.

cantly new form, or die out, or be translated and disseminated, and so on.

This notion of a program life course has been considered previously by Cronbach and colleagues in their taxonomy of program maturity (Cronbach et al., 1980), by Chen in his taxonomy for program evaluation means and ends (2005), and by Scheirer (2012) in her life cycle evaluation framework. However, we were the first to ground this perspective in evolutionary theory, the leading theoretical perspective in the life sciences, considerably strengthening the argument (Cornell Office for Research on Evaluation, 2009; Colosi & Brown, 2006; Hebbard et al., 2009; Trochim et al., 2012; Trochim, 2007; Trochim, Hertzog, Kane, & Duttweiller, 2007; Urban, Hargraves, Hebbard, Burgermaster, & Trochim, 2011). In addition, the idea of multiple phases over the program life course is directly analogous to the notion of multiple phases of clinical trials in biomedicine (National Library of Medicine, 2008). The longevity and preeminence of these foundational theoretical perspectives provide deep grounding for thinking about program development and evaluation. Moreover, the evolutionary perspective goes beyond typical lifecycle frameworks in ways that address some practical limitations of those views, and have important implications for both science and evaluation policy (particularly as they relate to the conceptualization of evidence-based programs).

Lifecycle frameworks described by Scheirer and others have much to offer in terms of a clear description of and prescription for alignment of evaluation methodologies according to program situation (e.g., Scheirer, 2012). Contributors to the Forum on *Planning evaluation through the program life cycle* in the American Journal of Evaluation (Scheirer et al., 2012) also point out practical limitations of a basic lifecycle framework and directions for future work, several of which underscore the particular value of the Evolutionary Evaluation approach. As explained above, Evolutionary Evaluation accords with the reality that program development is not a linear process anchored in chronological time, that programs sometimes revert to an "earlier" phase because of program or environmental context, that some components of a program may be more developed than other components at a moment in time depending on how the developmental process has adapted or incorporated particular new or well-understood features, and that it may not always be appropriate to proceed in the sequence of phases as laid out in the life cycle framework (Chapel, 2012; Grob, 2012; Mark, 2012).

In general, the tension inherent in the practical realities of programming under time and funding constraints poses challenges for all evaluation frameworks. However, the Evolutionary Evaluation approach can help clarify the consequences of deviating from the lifecycle prescriptions, so that we can better assess the tradeoffs posed by (for example) a funder's need to make decisions based on relative program effectiveness despite the fact that a particular program might not be "ready" for that type of evaluation. Evolutionary Evaluation can help inform real-world decision-making and offer guidance even when a "stage model is not a good fit to program history and to key information needs" (Mark, 2012). We turn to these costs of misalignment in a later section. Evolutionary Evaluation also opens new lines of evaluation inquiry by shedding light on issues regarding portfolios of programs. This is highly relevant to funders such as the National Science Foundation, the National Institutes of Health, and others that support groups of programs with a common broad goal. We turn to this in the next section.

1.3. Phylogeny and program portfolios

Evolutionary Evaluation allows us to think not only at the level of individual programs, but also in terms of collections or portfolios of programs. Whereas ontogeny refers to the development of an organism over its life course, phylogeny refers to the evolution of species (collections of organisms) over time. New principles become important as we shift our thinking from the evolution of a single organism to the evolution of a collection of organisms that comprise a species. From evolutionary theory, we know that for a species to evolve over time and be more likely to survive in a dynamic environment there must be "variation" – that is, diversity of characteristics amongst the organisms within a population and the emergence of new characteristics – and a "selection mechanism" which preferentially selects organisms within the species that have a more favorable fit with the environment (however defined). Variation and selection contribute to the prevalence of organisms within a species with characteristics that are more advantageous.

As applied to programs and evaluation, ontogeny refers to the evolution of a single program over its life course. Phylogeny refers to the evolution of a portfolio (or collection) of programs. For example, the United Way may fund a portfolio of twenty after-school programs each of which is designed to meet the needs of the local context, but all of which aim to provide constructive activities for adolescents. Evaluation plays an essential role in both the generation of program variations (after-school programs that are responsive to the local community as determined through a needs assessment) and the selection of programs with greatest fitness to their environment.

The process of *consciously* developing and evolving programs can be considered a type of artificial selection. Artificial selection refers to a managed process of selective breeding for particular traits. For example, in agriculture, efforts are often made to develop varieties of vegetables with better resistance to certain blights. Natural selection refers to the non-managed process in which individual organisms tend to survive, or not, based on the extent of their fitness to the environment, resulting over time in changes in the prevailing characteristics of the species (evident in, for example, changes in biodiversity associated with climate change) (UNEP/CMS Secretariat, 2006). Both natural and artificial selection follow the same evolutionary rules of variation and selective retention. That is, both require diversity among organisms and on-going sources of new characteristics, as well as some process that determines which characteristics will come to prevail.

Much of evaluation, particularly in the past decade, has been concerned with the generation of program theory, logic models, structured conceptualizations, and so on (Caracelli, 1989; Chen & Rossi, 1990; Kane & Trochim, 2006; Linstone & Turoff, 1975; Trochim, 1989; Trochim & Kane, 2005; Trochim & Linton, 1986). Each of these can be viewed as a variation generation or exploration methodology that potentially stimulates or describes "blind" variations that may be subsequently developed into programs, implemented and selected for (Campbell, 1969; Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002).

Programs and their theories are selected for over time because they have characteristics that enhance their fitness to the environment. Evaluation is a form of feedback, and as such is a vital part of the selection process. It is possible that a program could continue to survive without feedback. However, evolutionary theory suggests that without feedback, a program or portfolio of programs is more likely to stagnate or fail to achieve desired ends.

From evolutionary theory we also know that developmental diversity is crucial for a species' survival. If all organisms were simultaneously in the same developmental phase, the potential for survival of the species would be reduced. There needs to be a diversity of young, middle-aged, and elderly organisms of a

particular species in order to achieve generational succession. In addition, it is important to have more rather than less variation of organisms within a species. One especially important danger is that of monocultures, in which a single variant of a species dominates in an ecological niche. The problem with monocultures is that they are vulnerable to catastrophic failures. One of the best examples of this is the story of the Irish potato famine of 1845 (Pollan, 2002). The potato had its biological origins in Peru where literally thousands of varieties co-exist. When Western Europeans began traversing the Atlantic, they brought back only a few varieties and in Ireland it was the Lumper potato that almost exclusively got planted and became a major staple of the diet. When a disease known as the potato blight was inadvertently imported to Ireland, probably on a ship from America, it spread through the Lumper monoculture within months, ultimately leading to famine, an estimated 1 million deaths, and a significant migration of people out of the country. This problem could not occur in Peru because the great biodiversity of potato variants helped assure that there would be some varieties with resistance to any potential disease and therefore, the species as a whole would be able to adapt rather rapidly. Evolutionary change can only occur under changing ecological circumstances when there is enough variation from which to select. And, with greater diversity comes reduced risk of a monoculture, more opportunity for rapid selection and adaptation, and greater potential for positive change (Lerner, 2006).

Why is the problem of monocultures a significant evolutionary issue with implications for evaluation? If we have a portfolio of programs that are virtually identical and as a set lack variation, we run the risk of cultivating program monocultures. These program monocultures are susceptible to the same dangers described above. With limited variability, there would be fewer programs from which to select which impedes further evolution or adaptation, especially when circumstances or contexts change. Program monocultures are less likely than more diversified portfolios to promote the evolution of programs that have better fitness to their environment.

What this means for portfolios of programs is that in order for more rapid adaptation to occur it is evolutionarily desirable that there be (preferably more) variations of programs that address any given problem in order to avoid program monocultures and to provide the grist for selective retention of more promising alternatives.

In evolution, selection pressures occur when there are more varieties than can be sustained in a particular context and some have greater survivability because of their fitness to the environment. This is likely as true for programs as for biological organisms. For instance, in the system of phased clinical trials in biomedicine, nearly three-fourths of all medical treatments (programs) are not successful, never reach a patient population (Mayo Clinic, 2007) and consequently do not survive. This suggests that, as in all evolution, it is important that there be a high enough rate of new program generation (variation) in order to account for the inevitable failure rates of early phase programs or treatments. There is always a tension between the cost of developing/implementing new treatments/programs and the willingness to invest in development. In the realm of medicine, we are generally comfortable with the idea of investing resources in the development of many promising potential treatments with the inherent understanding that a high percentage will never make it to use in medical practice. The same rationale should apply when considering program or intervention development in areas other than medicine. Of course, this does not mean promoting all program variations regardless of source and quality. Making selection decisions about programs on such criteria as the quality of their conceptual models, proposed delivery mechanisms, and likely ability to implement are some of the important early evolutionary filtering mechanisms. This underscores the importance of having a diverse pool of programs at any given time, and the value of incubating promising new variants, given the overall survival challenges.

It is worth noting that from the point of view of evolutionary epistemology, failures are or can be beneficial. The goal of science, and indeed all knowledge generation is to advance our understanding of phenomena. From that perspective, it is not important whether the individual entity (i.e., organism, program, scientific study, etc.) succeeds or fails; what is important is what is learned in the process (Green, 2008). Thus, even "failed" experiments (including social experiments) provide opportunities for knowledge acquisition. Risk of failure alone should not be allowed to stifle innovation. The National Institutes of Health (NIH), which has traditionally funded incremental science, recognizes the need to invest in high risk, high reward research in order to generate potential leaps forward in knowledge and programs to treat the most intractable problems. The NIH Director's New Innovator's Award and the NIH Director's Pioneer Award Program (The National Institutes of Health, 2012) are expressly designed to fund such projects with an explicit understanding that the vast majority will fail and the hope that at least some will lead to great leaps forward.

Thinking in terms of portfolios of programs helps us to consider the evolution of not just a single program but of multiple programs that are all working toward a common goal. Evaluation plays a fundamental role in this regard particularly when it is used as a tool for decision making. Multiple federal agencies, including the Department of Education, the National Science Foundation, and the National Institutes of Health all have portfolios of programs geared toward particular long-term goals (e.g., developing the next generation of scientists or improving math or science performance). In addition to considering the evolutionary phase of any given program contained within one of these portfolios, it is advantageous (from an evolutionary perspective) for managers of such program portfolios to consider the evolution of the entire portfolio. Are there enough programs in the portfolio that are "young" or new to provide the variation from which to select? Are the programs within the portfolio distributed across program evolution phases? Are there selection mechanisms in place for identifying particularly promising programs? Evaluation conducted at the portfolio level of the system encourages us to think strategically about the evolutionary phase of all programs contained within a portfolio and make funding and policy decisions that will encourage the continued evolution of such initiatives.

In the following sections, we describe our operationalization of Evolutionary Evaluation in practical terms including a discussion and definition of program evolution phases, evaluation evolution phases, and the importance of their alignment. Next, we discuss how both the theoretical underpinnings described above as well as the practical implementation of these concepts in the real world require us to rethink our definitions of rigor, value, and evidence.

2. Characterizing the evolution of programs

We turn first to considering how Evolutionary Evaluation can be applied to considering the development of a single program. It is not just the passage of time that marks a program's evolution but rather a substantive progression that includes refinement and stabilization of program content and approach (reducing the variability of the program from one round of implementation to the next as a program "settles" into its essential components). In other words, as a program develops, the internal stability of the

program typically increases. This progression also reflects decisions that are made along the way about a program's expansion, continuation, or contraction. A program may be retired or substantially revised at any evolutionary phase. Inherent in this evolution is a bidirectional relationship between the program and its environment. That is, a program is intended to change the environment or community in which it resides, and in turn, changes in the environment affect how a program evolves. To operationalize Evolutionary Evaluation we sketch out a hypothetical sequence of program evolution phases: initiation, development, maturity or stability, and implementation or dissemination (Fig. 1). Each phase is then broken down into two sub-phases (see Fig. 1 for specific sub-phase definitions).

2.1. Program evolution phase definitions

A program in the "Initiation" phase is a relatively new program that is still undergoing substantial changes/revisions or an existing program that is being implemented in a new ecological (e.g., cultural, historical, geographic) context. The first few times a program is implemented in a community the usual issues of initiation are likely to arise: identifying and training program staff, localizing the program to the immediate context, adapting an existing program so that it is culturally responsive, reacting to the unanticipated problems that arise, etc. These issues will arise again whenever a previously established or research tested program is introduced into a new environment or community.

A program in the "Development" phase is still undergoing changes or revisions; however, the scale and scope of those revisions are smaller than what is seen during initiation. A program in the "Development" phase is in the process of successive revisions as it gets implemented repeatedly over time. Implementers are getting accustomed to the program and how it operates in practice. Surprises may still occur and implementers are still adapting the program as they learn, but they are also increasingly able to anticipate problems before they arise, and they are developing a storehouse of experience in how to deal with them. Toward the end of the "Development" phase, most program elements are implemented consistently though minor changes may still be taking place as some elements continue to develop.

A program in the "Stability" phase is being implemented consistently and this typically means that there are formal, written protocols, procedures, or process guides in place. A program in the "Stability" phase has clearly stated expectations and has been carried out at least several times with some degree of implementation success. The program is no longer dependent upon particular individuals for implementation. If the initial implementers are no longer present, the program can still be carried out with high fidelity. Therefore, the experience of participants remains relatively stable from one round of implementation to the next.

A program in the "Dissemination" phase is fully protocolized and is being widely distributed and implemented in multiple sites. The primary focus in the "Dissemination" 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. Programs in the "Dissemination" phase still retain an element of controlled implementation. That is, delivery mechanisms are managed to ensure strong implementation fidelity to the tested program. This is distinct from programs in the "Initiation" phase which are more subject to real-world influences.

Thus far, we have characterized the evolution of a single program and offered concrete definitions for the phases of program evolution. Next, we will characterize the evolution of evaluation and offer concrete definitions for the phases of evaluation evolution.

3. Characterizing the evolution of evaluation

According to Campbell and Popper's views on evolutionary epistemology, research is the mechanism that drives the evolution of knowledge. We extend this reasoning to the realm of programs where the driving mechanism of knowledge evolution is evaluation. Just as a program is never "done evolving" neither is an evaluation ever fully complete. A program's evaluation is not a onetime activity. Rather, it is a continuous, dynamic process. Evaluation is a process involving a sequence of evaluation cycles.

Note, we want to draw a clear distinction between the evaluation of a program over its entire lifetime versus an individual round of evaluation in a particular time period (which we refer to as an evaluation cycle). An individual evaluation cycle can be classified as being in one of the four evaluation phases (defined below). Just as optimal development of an individual is defined as successfully reaching milestones for each phase (e.g., crawling, to standing, to walking, to talking, to healthy adolescence, etc.), so too optimal development of our knowledge about a program involves progressing through multiple evaluation phases over time.

3.1. Evaluation evolution phase definitions

The evaluation phases are distinguished by the kinds of claims one would be interested in making in any given evaluation cycle, the corresponding methodology/design, and the kind of validity addressed. The definitions are not exhaustive and there are many designs/methods that are not discussed. However, the evaluation phase definitions provide a general framework for considering phased approaches. Evaluation evolution can be usefully divided into four phases: (1) Process and Response, (2) Change, (3) Comparison and Control, and (4) Generalizability. Each phase is then broken down into two sub-phases (see Fig. 1 for specific sub-phase definitions). Both qualitative and quantitative approaches can be used at any evaluation phase. Qualitative methods may be particularly useful during earlier phase evaluations when rapid feedback, exploration, and pilot testing are the hallmark.

A "Process and Response" evaluation generally examines initial implementation in a particular context and should therefore be dynamic, flexible, and provide rapid feedback about process. This can be accomplished with simple monitoring (i.e., participant documentation), post-only measurement, and unstructured observations for example. Formal measures may still be under development and are being assessed for reliability. Construct validity is being assessed in this phase of evaluation and refers to "an assessment of how well your actual programs or measures reflect your ideas or theories" (Trochim, 2005, p. 52). Programs in this stage, especially those involving highly complex phenomena, are often well-served by adaptive evaluation approaches such as developmental evaluation (Patton, 2011).

A "Change" evaluation generally examines a program's association with change in outcomes for participants in a limited and specific context (the focus is not yet necessarily on generalizability to other contexts, settings, etc.). Evaluations in this phase are generally correlational studies that use either matched or unmatched pre- and post-tests. This phase also generally includes greater focus on verifying the reliability and validity of measures. Conclusion validity generally corresponds with this phase of evaluation and refers to "the degree to which conclusions you reach about relationships in your data are reasonable" (Trochim, 2005, p. 206). The focus is on whether a relationship or association exists between the program and an outcome. It also assesses the degree to which an inference or conclusion is believable given the available data. It is not concerned with whether or not this relationship is causal in nature (the focus of internal validity).

A "Comparison and Control" evaluation examines the strength of a potential causal relationship between program and outcome(s), that is, the emphasis is on assessing effectiveness. Comparison and control can be achieved through experimental and quasi-experimental designs as well as more structured and comparative qualitative approaches. Internal validity is the focus of this phase of evaluation and is "the approximate truth about inferences regarding cause-effect or causal relationships" (Trochim, 2005, p. 135). The focus is on whether any observed changes in the outcome of interest (the effect) can be attributed to the program (the cause). It is important to note that internal validity is distinct from construct validity and that a program can demonstrate internal validity without having demonstrated construct validity. For example, an evaluation may be looking at the effects of contraceptive availability in high schools on teen pregnancy and STD rates. A positive program effect may be found (establishing internal validity), however, perhaps it was not due to condom distribution but to something else that occurred in the program. Perhaps the teens had to engage in conversations with peer mentors who distributed the condoms and it was those conversations that affected pregnancy and STD rates. Although the evaluation in this case may have internal validity, it lacks construct validity because the label "contraceptive availability program" does not accurately describe the actual cause.

A "Generalizability" evaluation focuses on examining outcome effectiveness across a wider range of contexts and is concerned with translation and/or dissemination. These evaluations examine the consistency of outcomes across different settings, populations, cultural contexts, or program variations and frequently include multi-site analyses. Meta-analysis may be used as well as other program review approaches that seek general inferences about the transferability of the program. External validity most closely corresponds with this phase of evaluation and is "the degree to which the conclusions in your study would hold for other persons in other places and at other times" (Trochim, 2005, p. 27). The focus is on whether the conclusions extend beyond the particular sample used in a study.

Finally, viable validity is an additional type of validity that is relevant and should be considered at all evaluation phases. Viable validity focuses on stakeholder and program implementers' perspectives on whether a program is "practical, affordable, suitable, evaluable, and helpful in the real-world" (Chen, 2010, p. 207). In other words, can the program be implemented without the assistance of research staff, is it suitable for the specific community targeted, is it needed, is it acceptable to those receiving and implementing the program, and does its cost justify its use? Even if a program has established construct validity, conclusion validity, internal validity, and external validity, it can still fail if viable validity is not established. Particularly in funding climates where resources for both programs and evaluation are scarce, it is essential to consider viable validity at all evolutionary phases. At the very least, viable validity should be considered whenever any kind of program change or adaptation is introduced or being considered.

At this point, we have used an Evolutionary Evaluation framework to characterize and define program and evaluation evolutionary phases for the development of a single program (ontogeny) which include specification of validity most centrally addressed at each phase. Fig. 2 presents the program phases, evaluation phases, and validity types in a circle. We chose to use a circle to represent the continuous nature of these evolutionary progressions as well as the notion that there is not a singular distinct beginning and end point. The absence of a singular starting point fits the reality that programs originate in different ways and thus have different needs (Chen, 2010; Mark, 2012). It should also be noted that in the diagram, program phases are aligned with specific evaluation phases (and types of validity). This is not a coincidence. In fact, there are both important practical and theoretical reasons for alignment. It should also be noted that viable validity has been placed outside of the circle to convey its importance at all evaluation phases. We will begin by discussing

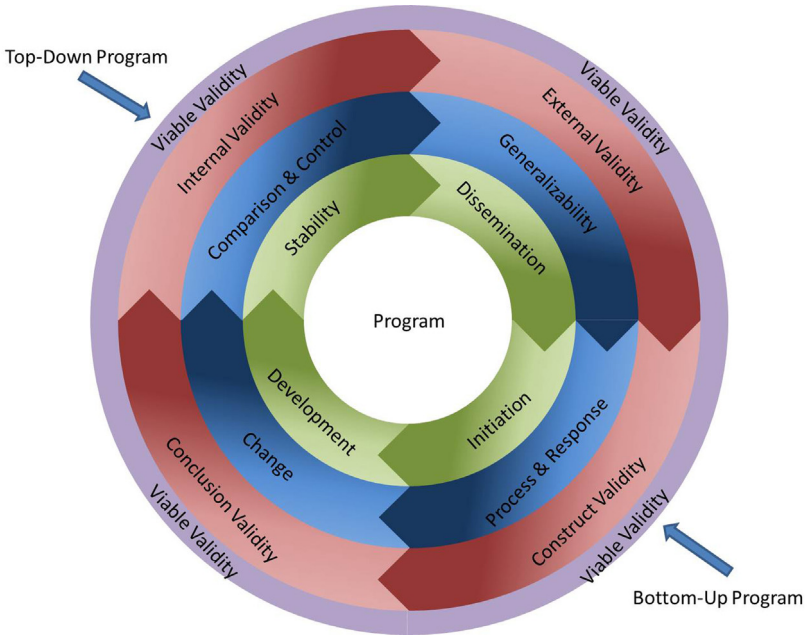

Fig. 2. Evolutionary Evaluation Model. The program is at the center of the circle. The inner ring depicts the program phases, the middle ring depicts the evaluation phases and the outer rings depict the relationship of validity to Evolutionary Evaluation. Top-down programs typically enter the Evolutionary Evaluation Model at Stability/Comparison and Control and address internal validity. They subsequently move toward Dissemination/Generalizability and address external validity. Many top-down programs fail to address the other two phases. Bottom-up programs typically enter the Model at Initiation/Process and Response and address construct validity. They subsequently move toward Development/Change and address conclusion validity. Many bottom-up programs have a harder time addressing the other two phases.

the practical reasons for working toward program and evaluation phase alignment, and the implications of misalignment. These practical arguments have theoretical counterparts which we will discuss in turn.

4. Interaction of program and evaluation evolutionary phases: practical implications for evaluators and program planners

The process of program evolution through phases is driven by evaluation, whether formally done or naturally accomplished (through naturally occurring informal feedback mechanisms). For example, information gathered through evaluation can be used to make positive changes to a program's implementation and scope, pushing the program forward – and sometimes backward – through program phases. A fundamental point, and the focus of the next section, is that for any given program phase there is a corresponding and appropriate evaluation phase; when these are synchronized we refer to this as *alignment*. Alignment between program and evaluation phases is essential for ensuring that a program obtains the kind of information that is most needed at that point in the life of the program, and that program and evaluation resources are used efficiently.

The idea of matching programs to methods has existed for some time (Bannan-Ritlans, 2003; Cronbach et al., 1980; Ruegg & Jordan, 2007). Michael Scriven's (1967) distinction between formative and summative methods suggests the importance of appropriately yoking method to program phase. The types of questions asked at each program phase will differ as will the types of evaluation approaches employed to answer those questions. For example, Rossi, Lipsey, and Freeman (2003) argue that the evaluations of less mature programs should focus on a needs assessment and assessment of program theory, whereas more mature programs should utilize process evaluations, impact/outcome evaluations, and efficiency assessments. Similarly, Chen (2010) proposes a sequencing of evaluation efforts that begins by assessing the viability of an intervention (i.e., the degree to which it is practical, affordable, helpful, etc.) before trying to assess effectiveness or efficacy using methods such as the RCT. He argues that traditional top-down approaches to evaluation over-emphasize internal validity at the expense of external validity.

While these theorists help us to consider the relationship between programs and methods, we further advance this line of reasoning by grounding it in foundational theories from the life and developmental sciences. As discussed above, Evolutionary Evaluation considers both individual evaluation cycles as well as the accumulation of multiple evaluation cycles over a program's life, and also defines the relationship between methodology and validity. In the following section, we extend this thinking further by considering the implications of misalignment.

4.1. Alignment and misalignment

Fig. 3 depicts the relationship between program phases (on the *x* axis) and evaluation phases (on the *y* axis). (For readability the labels on each axis are in the form of phase numbers, as specified in Fig. 1.) For any given program, if the program and evaluation phases are perfectly aligned, the program would fall somewhere along the diagonal line. The circle labeled "A" is an example of a program in Phase IV-A, the "Dissemination" phase of its program evolution, with an evaluation design that is in the corresponding Phase IV-A – "Generalizability" evaluation phase.

In reality, program phases and evaluation phases are often not aligned; rather, they fall somewhere below or above the diagonal line. The consequences of misalignment vary depending upon where on the off-diagonal the program is situated. However being out of alignment, in either direction, amounts to a waste of resources and increases the chances of potentially costly bad decisions.

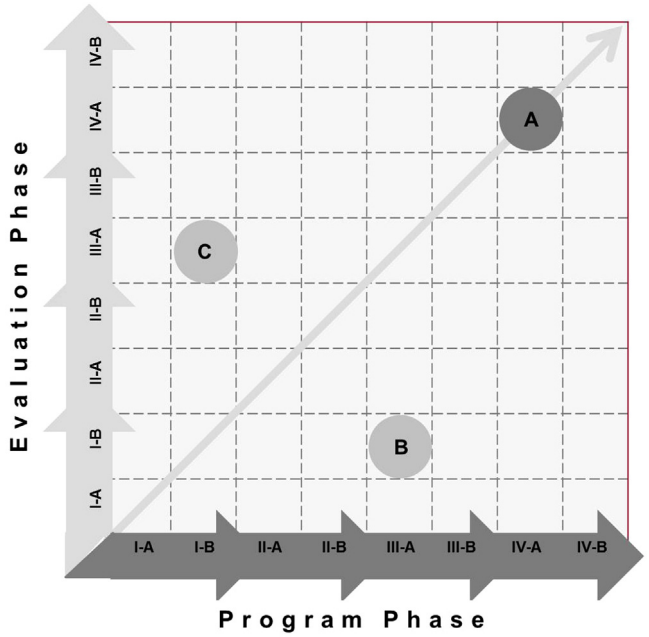

Fig. 3. Relationship between program and evaluation phases. The diagonal line indicates perfect alignment between program and evaluation phases.

A program that is below the diagonal line has a program phase that is more advanced than its evaluation phase. For example, program "B" is in Phase III-A – "Stability" of its program evolution, but it is in Phase I-B – "Process and Response" of its evaluation evolution. This is a program that has reached a stable state and should generally be conducting evaluations that focus on effectiveness using comparison and control designs. However, its current evaluation cycles are focused only on rapid feedback related to implementation or participant experience. This could result in an insufficiently effective program continuing without making important and necessary changes or improvements. If so, it may need to return to an earlier program phase and make needed changes or be retired. Alternatively, this could be a very successful program that others would benefit from receiving that will not get promoted or disseminated more widely because it cannot make strong enough claims regarding its effectiveness. In either case, the misalignment leads to suboptimal use of scarce program resources.

A program that is above the diagonal line has an evaluation phase that is more advanced than its program phase. Later-phase evaluations tend to be more expensive and often require more time to complete than earlier phase evaluations, so resource constraints alone suggest that they should be targeted for select situations. It is also important to take into account what kind of information the program, in its current state, really needs. For example, program "C" is in Phase I-B – "Initiation" of its program evolution, but in Phase III-A – "Comparison and Control" of its evaluation evolution. This is a program that is still changing rapidly and should be using an evaluation that provides rapid feedback on program implementation and works toward clarifying the key constructs. However, this program is engaged in what we might pejoratively label "premature experimentation". The design is appropriate for a more stabilized and standardized program and is potentially more expensive than necessary for Program C and could lead to bad programming decisions.

The potential for misguided programming decisions is particularly pronounced for a program that is still in an early program phase (as is the case for Program C). Due to the variability in

implementation typical of programs that are early in their program evolution, there is likely to be more "noise" in the data. These programs are still changing so rapidly that any results about outcome effectiveness are unlikely to be replicable in subsequent rounds of implementation (there are still too many moving parts and construct validity has not yet been established). If the evaluation happens to yield favorable results, one cannot be confident that these results will persist in subsequent rounds of implementation. If the results of the evaluation are unfavorable, it is not necessarily because the program does not work. Decisions based on findings from "premature experimentation" risk discontinuing an otherwise potentially effective program that has not yet reached a level of stability that would allow for the detection of positive effects, or the promotion of an otherwise poor program that happened to demonstrate positive results (but which may not be replicable over subsequent rounds of implementation).

It is not uncommon to have a program whose evaluation and program phases are not aligned. Budget constraints, program or evaluation capacity constraints, unavoidable external imperatives, and other understandable factors can lead to persistent misalignment. However inertia, lack of information, bias, and other less appropriate factors can play a role as well. It seems reasonable to presume that stakeholders would widely agree that evaluations need to strive both for societal well-being and the effective use of resources. Therefore, moving toward alignment of program and evaluation phases – and promoting the healthy evolution of the program – should be treated as a key goal of evaluation planning. For a program that is already being evaluated but whose program and evaluation phases are not currently aligned, the move toward alignment does not necessarily occur within one evaluation cycle. Rather, the focus should be on building evidence over successive evaluation cycles while simultaneously striving for phase alignment. The needs and resources of the individual program must be considered when developing a strategy for bringing program and evaluation phases into alignment.

This notion of alignment is also seen in evolutionary theory in the relationships between some species. "Symbiosis is a close ecological relationship between the individuals of two (or more) different species. Sometimes a symbiotic relationship benefits both species, sometimes one species benefits at the other's expense, and in other cases neither species benefits" (Meyer, 2012). One of the most familiar examples of this is the relationship of the flower and the bee. The flower provides nectar that is produced into honey, and the bee acts as the vehicle for plant sexual reproduction by moving pollen from one flower to another. Each provides something to the other and both benefit from the exchange. In the case of a program and its evaluation, the relationship is one in which the evaluation relies on, or exists because of, the program and although the program could exist without evaluation (e.g., driving blind), it is likely to evolve more successfully and survive longer if appropriate evaluation is conducted. An Evolutionary Evaluation framework encourages the symbiotic or co-evolutionary relationship between program and evaluation phases. This evolutionary understanding of programs and evaluation allows us to address the question of program–evaluation methodology fit in an entirely new way, yielding a new standard of rigor. The next section articulates this new standard which includes not only addressing program and evaluation phase alignment, but also the critical importance of addressing multiple types of validity.

5. Implications of an Evolutionary Evaluation perspective: the EBP case

In the past few decades, rigor has increasingly become associated with the use of randomized controlled trials (RCTs) as the basis for establishing that a program is "evidence-based" (Community Preventive Services Task Force, 2012; Institute of Education Sciences, 2012). RCTs are related most to the prioritization of internal validity over other validity types. Chen (2010) refers to this focus on internal validity as the top-down approach to program development. Alternatively, programs that initially focus on assessing viable validity are labeled bottom-up programs (Chen, 2010). Evolutionary Evaluation allows us to understand both of these approaches to program development, including their relative strengths, and how they relate to each other, to the complete validity typology, and to the definition of rigor.

Bottom-up programs tend to be based on informal theory or knowledge of local context and are responsive to local needs. These programs are typically practitioner driven and based in practitioners' ideas and expertise about what is likely to work (practice-derived theory), knowledge of the research literature, and/or adaptations of existing established programs. Programs developed from the bottom-up typically enter the evolutionary phase model at "Initiation" (Phase 1). This corresponds with the "Process and Response" (Phase 1) evaluation phase which emphasizes construct validity. Viable validity is also particularly salient as the "Initiation" phase typically involves a new program or a significant change to an existing program. Practitioners' relative strengths would most naturally tend to be associated with establishing the viability of a program (viable validity), bringing program ideas to life (construct validity) and understanding the community and cultural contexts within which programs might or might not work (external validity). External pressure to establish internal validity has increased in recent years, leading to pressure for practitioners to undertake evaluations that match neither their strengths nor the ontogenetic development of their programs. And, the imposition of late-stage methods to early-stage programs denies the validity-enhancing natural evolution of bottom-up programs.

Top-down programs are commonly based in more formal academic theory and explicitly linked with a research evidence-base. These programs are generally researcher driven and place an emphasis on establishing internal validity at the outset (Chen, 2010). Thus, top-down programs typically enter the evolutionary phase model at "Stability" (Phase 3). This corresponds with the "Comparison and Control" (Phase 3) evaluation phase which emphasizes internal validity. Researchers' relative strengths tend to be in establishing internal and conclusion validity congruent with their training and interest in research. The rush to experimentation before the nature of the program and its measurement have developed sufficiently runs the risk of making potentially promising programs look ineffective when instead they have not been provided sufficient time to establish the foundations of viability, construct and conclusion validity. It would be akin to saying that first graders did not perform effectively because they were not able to act in an intelligent, professional, adult manner. If evaluation is rightly seen as part of a societal selection mechanism, the danger of premature experimentation is that it will tend to incorrectly eliminate programs that could have been effective if they were allowed to develop appropriately.

5.1. Rethinking the EBP mandate

Regardless of whether a bottom-up or top-down approach is used, Evolutionary Evaluation suggests that all phases of program evolution, and their corresponding assessments of validity, should be addressed before a program is labeled "evidence-based". Establishing some types of validity does not guarantee that others will also necessarily be attainable for any given program. Bottom-up programs that initially establish viable, construct, and conclusion validity will not necessarily be able to establish internal and external validity. Similarly, top-down programs that initially establish internal and external validity will not necessarily

be able to establish viable, construct, and conclusion validity. Evolutionary Evaluation predicts this and even argues that it is beneficial for adaptive development. The fact that a program has persisted for a long time does not in and of itself justify its continued existence (although it does suggest that there were evolutionary and ecological factors that led it to evolve and survive to that point). Similarly, just because a program is effective in controlled circumstances does not mean that it will be effective in most real-world contexts (although it does not rule out the possibility that it could be effective in contexts other than the test context).

Typically, programs that are deemed "evidence-based" have taken a top-down approach to development, began with later-stage methods like RCTs, and have only completed part of the evolutionary phase circle. It would be premature to label such programs as "rigorous" or "evidence-based" when they have not addressed viability, whether the program reflects what was intended, whether the measures accurately reflect the outcomes, or whether they can work in any but the original testing contexts.

Given the current climate, there is incredible system pressure that a program be evaluated with an RCT design to be considered eligible for the "evidence-based program" label. This introduces a distortion that skews the global portfolio of programs away from those generated using a bottom-up approach and favors those generated using a top-down approach. It is important to underscore the consequences of this distortion when evidence-based programs are so narrowly defined. One important consequence apparent from Evolutionary Evaluation is that a failure to address viable and construct validity makes "evidence-based programs" vulnerable when disseminated in new ecological niches or contexts. The evidence-based program movement itself has identified the dissemination and diffusion of proven interventions as one of the most challenging problems (Grol & Grimshaw, 1999; Herbert, 2003; Kerner et al., 2005; Khoury et al., 2007; Nutley & Davies, 2000). Even though external validity may be assessed in the EBP perspective, it is still typically done in an artificial manner with highly trained program implementers, carefully selected participants, and ample resources. This does not adequately represent the real world in which programs will eventually be implemented and hope to survive (Chen, 2010; Wandersman & Lesesne, 2012).

Much is lost when programs derived from the bottom-up are undervalued (Kazdin, 2008). Evolutionary Evaluation would maintain that establishing viable, construct, conclusion and external validity are just as essential as establishing internal validity for a program's prolonged survival and success – and should be considered critically important components of rigorous evaluation. Bottom-up programs embed practitioner knowledge and expertise, and are particularly sensitive to local context and identifying a good program-environment fit. This knowledge and sensitivity to program-environment fit is essential regardless of whether a program is derived from the bottom-up or top-down (however, bottom-up programs have the advantage of considering fit early in program development). Attention to program-environment fit is important both when programs are first being launched in the real-world and also over time as the context changes and established programs need to adapt. Practitioners are particularly attuned to the local environment and changes in it and are therefore best positioned to have insights about what kinds of programmatic changes are needed (Durlak & DuPre, 2008). Theories (often implicit) that underlie bottom-up programs tend to be based more heavily on evolved experiential knowledge as opposed to more classically accepted academic theories. It is important to recognize that practitioners are more than just the implementers of empirically derived theories. Programs derived from the bottom-up can be an important source of programmatic innovation because of the unique knowledge and expertise of practitioners. Additionally, bottom-up programs provide the ecosystem with an important source of program variation that is necessary to avoid program monocultures. In short, by ignoring or undervaluing bottom-up programs we risk losing a valuable and much needed source of innovation, variation and adaptation. Evolutionary Evaluation emphasizes the importance of drawing on as many sources of variation as possible, including both programs derived from the bottom-up and the top-down. The current climate has tended to favor the latter at the expense of the former.

6. Applications and conclusion

Evolutionary Evaluation has implications that go beyond only the EBP debate. Here, we will briefly reflect on a few ways in which this perspective could inform decision making regarding the management of both programs and portfolios of programs.

6.1. Implications for management of individual programs

We have worked extensively with programs and program staff on implementing an Evolutionary Evaluation framework and this is described in the Guide to the Systems Evaluation Protocol (Trochim et al., 2012) which can be accessed at https://core.human.cornell.edu/research/systems/protocol/index.cfm. In our work with programs and program staff, we have found that they often face two conflicting pressures. First, they feel pressure from funders or other stakeholders to provide summative evaluations that assess program effectiveness even though the program is still in an early evolutionary phase. Second, programs and program staff often lack resources (e.g., time, money, and appropriate training) to properly support and conduct the kinds of evaluations that are being requested. When funders, program portfolio managers and practitioners conceptualize program evaluation from an evolutionary perspective, the consequences of misalignment between program and evaluation evolutionary phases become more apparent and better decisions can be made about whether to keep, change, or retire a program, and about what kinds of evaluations to conduct and fund.

For example, in response to the funder who is requesting a summative evaluation of an early evolutionary phase program, it would be important to describe the program in terms of its program and evaluation phases and provide a multi-year evaluation proposal that clearly explains that questions regarding program effectiveness will be addressed if and when the program reaches (survives to) the appropriate program development phase. From this basis, it will be possible to have a discussion with funders or program portfolio managers that can critically weigh the tradeoffs between the need for evidence of effectiveness and the potential risks and costs of premature experimentation.

Sometimes, however, time constraints make it difficult or impossible to complete a series of evaluation cycles that covers all program and evaluation evolution phases. Program managers may face constraints due to deadlines from other agencies, federal mandates, funders' reauthorization schedules, and so on (Brooks, 2012; Mark, 2012). The Evolutionary Evaluation perspective identifies clearly what type of knowledge has been established to date and what is still unknown or uncertain. The consequence of having an omitted phase of evaluation, or even of conducting evaluations from different evolutionary phases simultaneously, can be better understood and can inform program decisions that are being driven by external schedules. Moreover, by underscoring the connection between a program and its environment and the evolutionary importance of "fit" between the two, Evolutionary Evaluation promotes caution in prematurely disseminating programs and highlights the essential role of careful consideration of program adaptation to local contexts.

Evolutionary Evaluation also has implications for how we think about research-practice integration. Researchers' relative strengths are generally in theory and the research that serves that theory. They value the connections that practitioners have with the local community, in particular, access to and pre-existing relationships with the target population. But practitioners also have deep and evolved knowledge of the local context which can aid in developing better congruence between programs and the environment. This speaks to the value of engaging practitioners early in program design phases and not just seeking practitioners' feedback after a program has already been designed and tested. That is, efficiencies can be gained by addressing viable or construct validity prior to launching more costly assessments of internal and external validity.

In order to build the best environment for promoting societal and community well-being, partnerships and collaborations between researchers and practitioners would ideally be built not just around researcher-initiated programs, but also around practitioner-initiated programs (Kazdin, 2008). This bi-directional flow would capture an essential and often overlooked source of program innovation and variability, encouraging more rapid evolution of programs.

6.2. Implications for management of portfolios of programs

Evolutionary Evaluation emphasizes that any program is situated within a larger ecology of programs. It would maintain that the goal of evaluation is to ensure that knowledge about programs evolves more effectively using conscious artificial selection rather than allowing natural selection to play out as it will. Some (and even most) start-up programs will fail or need serious revisions and it should be recognized that this is a part of successful phylogenetic development. Moreover, the environment is constantly in flux and adaptations may be needed in response to such changes. Sometimes external change is even drastic enough to warrant a repetition of earlier phases (i.e., a reassessment of viable validity and construct validity).

In organizations that are simultaneously running or funding multiple programs, it is advantageous to think about the collection of programs as constituting a portfolio and encouraging variation of programs at different phases of development (a rich, diverse ecosystem). By examining the program phases of multiple programs in a portfolio, funders and program portfolio managers can make strategic decisions about where to invest evaluation resources in order to test whether longer-term outcomes are being achieved (Urban & Trochim, 2009).

Evolutionary Evaluation suggests that funders should be conscious of the evolutionary phases of the set of programs in their portfolio and make strategic decisions regarding the balance of programs desired at any given phase. There is also evolutionary value in including programs that are derived using both bottom-up and top-down approaches and encouraging innovation in both. In addition, they should be aware both of moving programs toward improved alignment of program and evaluation phases, and of encouraging the progression of programs through the program phases over successive evaluation cycles.

6.3. Conclusions

The theory of evolution is the foundation of the life sciences. Evolutionary epistemology argues that this theory also describes how knowledge evolves. Developmental systems theory, ecological theory and systems theory enrich our understanding of this evolutionary process. Programs can be viewed as a form of knowledge translated into practical application. Species of programs exist within a complex environment that naturally exerts selection pressure and that we attempt to influence through artificial selection. Program variations are essential to this ecosystem and provide the essential grist for selective retention of individual programs that have fitness to their environment. Evaluation is essential both for generating program variations and for selecting those that fit. This artificial selection of programs – essentially a form of program breeding – is at the center of the evaluation agenda. An individual program, essentially a program "organism", follows a developmental life course or "ontogeny" and exists within a "species" of knowledge about that family of programs that is continuously evolving in a "phylogenetic" manner. Evaluation is most effective when appropriately aligned or "symbiotic" with the program's stage of development, encouraging development of the individual program and enabling selective retention to occur.

Evolutionary Evaluation has the potential to enhance our understanding of a broad range of issues in contemporary evaluation. This was illustrated here in the context of evidence-based programs. In the current EBP climate, practitioners are increasingly expected to implement only or primarily programs that have been demonstrated to be effective through RCTs. While an emphasis on ensuring program effectiveness is clearly important, Evolutionary Evaluation suggests that there are significant risks in rigid interpretations of the evidence based idea. While in the short-term it may appear to be efficient, in the long-term, rigid adherence to EBP risks encouraging program monocultures and reducing important sources of variation for subsequent evolution and adaptation. Evolutionary Evaluation reminds us that, just as in nature, we need to be concerned with preserving sufficient diversity in the program ecosystem.

The potential for the application of evolutionary theory in evaluation is just beginning to be addressed and this paper can only be viewed as an early stage development in the ontogeny of Evolutionary Evaluation. As in all contexts where we might like to enhance the rate or shape the direction of evolution, we need to encourage a diversity of new thinking. We hope this paper provides a genesis for such evolution.

Acknowledgements

This research was supported by NSF grant number 0814364 awarded to the authors. The authors would also like to thank Thomas Archibald, Jane Buckley, Marissa Burgermaster, Claire Hebbard, Sarah Hertzog, and Margaret Johnson.

References

Bannan-Ritlans. (2003). The role of design in research: The integrative learning design framework. *Educational Researcher, 32*(1), 21–24.

Bertalanffy, L. V. (1972). The history and status of general systems theory. In G. J. Klir (Ed.), *Trends in general systems theory*. New York: Wiley-Interscience.

Bradie, M., & Harms, W. (2006). Evolutionary epistemology. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Stanford, CA: Stanford University.

Brooks, A. (2012). Commentary on "Expanding evaluative thinking: Evaluation through the program life cycle". *American Journal of Evaluation, 33*(2), 280–282.

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., & 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.

Caracelli, V. (1989). Structured conceptualization: A framework for interpreting evaluation results. *Evaluation and Program Planning, 12*(1), 45–52.

Chapel, T. J. (2012). Evaluation purpose and use: The core of the CDC program evaluation framework. *American Journal of Evaluation, 33*(2), 286–289.

Chen, H. (2005). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Thousand Oaks, CA: Sage.

Chen, H. (2010). The bottom-up approach to integrative validity: A new perspective for program evaluation. *Evaluation and Program Planning, 33*, 205–214.

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

Colby, A., Kohlberg, L., & Lieberman, M. (1983). A longitudinal study of moral judgement. *Monographs of the Society for Research in Child Development, 48*(1–2).

Colosi, L., & Brown, J. S. (2006). Towards a systems evaluation protocol. *Paper presented at the American Evaluation Association*.

Community Preventive Services Task Force. (2012). *The Community Guide* Retrieved from http://www.thecommunityguide.org/index.html.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Boston: Houghton Mifflin Company.

Cornell Office for Research on Evaluation (CORE). (2009). *The evaluation facilitator's guide to: Systems evaluation protocol*. Ithaca, NY: Cornell Digital Print Services.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R., Hornik, R. C., Phillips, D. C., et al. (1980). *Toward reform of program evaluation*. San Francisco: Jossey-Bass.

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.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3–4), 327–350 http://dx.doi.org/10.1007/s10464-008-9165-0

Green, L. W. (2008). Making research relevant: If it is an evidence-based practice, where's the practice-based evidence? *Family Practice, 25*(Suppl. 1), i20–i24.

Grob, G. F. (2012). A lifer's perspective on life cycle evaluations. *American Journal of Evaluation, 33*(2), 282–285.

Grol, R., & Grimshaw, J. (1999). Evidence-based implementation of evidence-based medicine. *The Joint Commission Journal on Quality Improvement, 25*(10), 503–513.

Hebbard, C., Trochim, W., Urban, J. B., Casillas, W., Cathles, A., Hargraves, M., et al. (2009). Alignment of program lifecycles and evaluation lifecycles. *Paper presented at the American Evaluation Association* https://core.human.cornell.edu/documents/lifecycleposterAEA2009.pdf.

Herbert, J. D. (2003). The science and practice of empirically supported treatments. *Behavior Modification, 27*(3), 412–430.

Institute of Education Sciences. (2012). *What Works Clearinghouse* Retrieved from http://ies.ed.gov/ncee/wwc/.

Kane, M., & Trochim, W. (2006). *Concept mapping for planning and evaluation*. Thousand Oaks, CA: Sage Publications.

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63*(3), 146–159.

Kerner, J. F., Guirguis-Blake, J., Hennessy, K. D., Brounstein, P. J., Vinson, C., Schwartz, R. H., et al. (2005). Translating research into improved outcomes in comprehensive cancer control. *Cancer Causes & Control, 16*, 27–40.

Khoury, M. J., Gwinn, M., Yoon, P. W., Dowling, N., Moore, C. A., & Bradley, L. (2007). The continuum of translation research in genomic medicine: How can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention. *Genetics in Medicine, 9*(10), 665–674.

Kohlberg, L. (1963). The development of children's orientations toward a moral order: I. Sequence in the development of moral thought. *Vita Humana, 6*, 11–33.

Kohlberg, L. (1984). Essays on moral development. In *The pyschology of moral development* (Vol. 2,). San Francisco: Harper & Row.

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. (2002). *Concepts and theories of human development*. Mahwah, NJ: Lawrence Earlbaum Associates.

Lerner, R. M. (2006). Developmental science, developmental systems, and contemporary theories of human development. In R. M. Lerner (Ed.), Hoboken, NJ: Wiley.

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

Mark, M. M. (2012). Program life cycle stage as a guide to evaluation decision making: Benefits, limits, alternatives, and future directions. *American Journal of Evaluation, 33*(2), 277–280.

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

Mayr, U. (2001). Age differences in the selection of mental sets: The role of inhibition, stimulus ambiguity, and response-set overlap. *Psychology and Aging, 16*(1), 96–109 http://dx.doi.org/10.1037//0882-7974.16.1.96

Meyer, J. R. (2012). *Symbiotic relationships* From http://www.cals.ncsu.edu/course/ent591k/symbiosis.html.

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

Molles, M. C. (2001). *Ecology: Concepts and applications* (2nd ed.). Boston: McGraw-Hill.

National Institutes of Health. (2012). *Commonfund* Retrieved from http://commonfund.nih.gov/.

National Library of Medicine. (2008). *What are clinical trial phases?* Retrieved from http://www.nlm.nih.gov/services/ctphases.html.

Nutley, S., & Davies, H. T. O. (2000). Making a reality of evidence-based practice: Some lessons from the diffusion of innovations. *Public Money & Management, 20*(4), 35–42.

Overton, W. F. (2006). Developmental psychology: Philosophy, concepts, methodology. In R. M. Lerner (Ed.), Hoboken, NJ: Wiley.

Overton, W. F. (2010). Life-span development: Concepts and issues. In R. M. Lerner (Ed.), *Handbook of life-span development* (Vol. 1,). Hoboken, NJ: Wiley.

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. New York: Guilford Press.

Pickett, S. T., Kolasa, J., & Jones, C. G. (1994). *Ecological understanding*. San Diego, CA: Academic Press.

Pollan, M. (2002). *The botany of desire: A plant's-eye view of the world*. New York: Random House.

Popper, K. (1973). Evolutionary epistemology. *Paper presented at the Sections I–VI of "The Rationality of Scientific Revolutions" at the Herbert Spencer Lecture*.

Popper, K. (1975). Evolutionary epistemology. In R. Harre (Ed.), *Problems of scientific revolution*. Oxford: Oxford University Press.

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

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

Richerson, P. J., Mulder, M. B., & Vila, B. J. (1996). *Principles of human ecology*. Needham Heights, MA: Simon & Schuster Custom Pub.

Rossi, P., Lipsey, M., & Freeman, H. (2003). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage Publications.

Ruegg, R., & Jordan, G. (2007). *Overview of evaluation methods for R&D programs: A directory of evaluation methods relevant to technology development programs*. US Department of Energy Office of Energy Efficiency and Renewable Energy.

Scheirer, M. A. (2012). Expanding evaluative thinking: Evaluation through the program life cycle. *American Journal of Evaluation, 33*(2), 264–277.

Scheirer, M. A., Scheirer, M. A., Mark, M. M., Brooks, A., Grob, G. F., Chapel, T. J., et al. (2012). Planning evaluation through the program life cycle. *American Journal of Evaluation, 33*(2), 263–294.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives on curriculum evaluation, AERA monograph series on curriculum evaluation* (Vol. 1, pp. 38–83). Skokie, IL: Rand McNally.

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Toulmin, S. (1967). The evolutionary development of natural science. *American Scientist, 55*(4), 456–471.

Trochim, W. (2005). *Research methods: The concise knowledge base*. Mason, OH: Cengage Learning.

Trochim, W. (1989). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning, 12*(1), 1–16.

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, 9*(4), 289–308.

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

Trochim, W. M. (2007). Evolutionary perspectives in evaluation: Theoretical and practical implications. *Paper presented at the Eastern Evaluation Research Society* http://www.socialresearchmethods.net/research/EERS2007/Evolutionary%20Perspectives%20in%20Evaluation%20Theoretical%20and%20Practical%20Implications.pdf.

Trochim, W. M., Hertzog, S., Kane, C., & Duttweiller, M. (2007). Building evaluation capacity in extension systems. *Paper presented at the American Evaluation Association*.

UNEP/CMS Secretariat. (2006). *Migratory species and climate change: Impacts of a changing environment on wild animals*. Bonn, Germany: UNEP/CMS Convention on Migratory Species and DEFRA.

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

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.

Walker, L., Gustafson, P., & Hennig, K. H. (2001). The consolidation/transition model in moral reasoning development. *Developmental Psychology, 37*, 187–197.

Wandersman, A. H., & Lesesne, C. A. (2012). If translational research is the answer, what's the question? Who gets to ask it?. In E. Wethington & R. E. Dunifon (Eds.), *Research for the public good: Applying the methods of translational research to improve human health and well-being* (pp. 33–51). Washington, DC: American Psychological Association.

Jennifer Brown Urban is Assistant Professor of Family and Child Studies at Montclair State University where she also directs the Developmental Systems Science and Evaluation Research Lab. Her scholarship has three interwoven strands including: (1) the development and testing of a systems science approach to program evaluation and planning to enhance internal evaluation capacity in STEM education contexts; (2) advancing the field of developmental science toward the application of systems science methodologies to developmental science questions; and (3) building an evidence-base within developmental science that addresses the role of multiple contextual factors (i.e., family, school, neighborhood) on adolescent development.

Monica Hargraves is Assistant Director of Evaluation for Extension and Outreach, at Cornell University. Her career began in economics with a PhD from the University of Rochester, positions on the faculty of Brown University and in the Research Department of the International Monetary Fund. She pursued a growing interest in

community-based work by joining the staff of Cornell Cooperative Extension in 1998. She moved to the Cornell Office for Research on Evaluation in 2008, where she is involved in researching, developing and testing the Systems Evaluation Protocol, with particular responsibility for evaluation capacity-building in the Cornell Cooperative Extension system.

William M. Trochim is Professor of Policy Analysis and Management at Cornell University and Professor of Public Health at the Weill Cornell Medical Center. He is the Director of Evaluation for the Weill Cornell Clinical and Translational Science Center, the Director of Evaluation for Extension and Outreach at Cornell, and the Director of the Cornell Office for Research on Evaluation. His research focuses on the development and assessment of evaluation and research methods. Dr. Trochim served for four years on the American Evaluation Association's Board of Directors and as President of AEA (2008).