Creative Education
2012. Vol.3, Special Issue, 951-958
Published Online October 2012 in SciRes (http://www.SciRP.org/journal/ce)

Scientific Research
DOI:10.4236/ce.2012.326145

The Program Assessment and Improvement Cycle Today: A New and Simple Taxonomy of General Types and Levels of Program Evaluation

James Carifio
University of Massachusetts-Lowell, Lowell, USA
Email: James_Carifio@uml.edu

Received August 20th, 2012; revised September 22nd, 2012; accepted October 5th, 2012

There has been strong pressure from just about every quarter in the last twenty years for higher education institutions to evaluate and improve their programs. This pressure is being exerted by several different stake holder groups simultaneously, and also represents the growing cumulative impact of four somewhat contradictory but powerful evaluation and improvement movements, models and advocacy groups. Consequently, the program assessment, evaluation and improvement cycle today is much different and far more complex than it was fifty years ago, or even two decades ago, and it is actually a highly diversified and confusing landscape from both the practitioner's and consumer's view of such evaluative and improvement information relative to seemingly different and competing advocacies, standards, foci, findings and asserted claims. Therefore, the purpose of this article is to present and begin to elucidate a relatively simple general taxonomy that helps practitioners, consumers, and professionals to make better sense of competing evaluation and improvement models, methodologies and results today, which should help to improve communication and understanding and to have a broad, simple and useful framework or schema to help guide their more detailed learning.

Keywords: Program Evaluation; General Types of Program Evaluation; Program Evaluation Foci; A Program Evaluation Taxonomy; Program Life Cycles; Higher Education

Introduction

In the past decade, there has been strong pressure from just about every quarter for higher education institutions to evaluate and improve their programs, with this pressure currently being at its high point for the last century (American Council on Education, 2012; State Higher Education Executive Officers, 2012). This pressure has simultaneously come from parents, students themselves (and particularly if they are working to pay for their own education or/and borrowing substantial amounts of money for the same), government agencies, business leaders, the general public, accrediting bodies and professional associations, all of whom are both clients of and stakeholders in the higher education system, which adds several layers of complexity and types of nuances to any kind of program evaluation or improvement efforts done by institutions in terms of their goals, design, data structures, analyses and reporting (Scriven, 2010a). The program development, evaluation and improvement cycle today, consequently, is much different and far more complex than it was fifty years ago, or even two decades ago, and it is actually a highly diversified and confusing landscape from both the practitioner's and consumer's view of such evaluative and improvement information (Mets, 2011).

Part of today's pressure for program evaluation and improvement information has come from a reaction to rapidly and continually increasing higher education costs relative to those who actually pay for the higher education students receive relative to the degree to which these students are getting quality (or desired adequacy) in terms of what is being paid for, with quality defined in many different ways, which range from student satisfaction to parental satisfaction to the achievement of highly desired outcomes, which include personal development objectives and desired types of employment, or desired further education (US News, 2012; London Times, 2012). Another source of pressure for program evaluation and improvement information today is a strong accountability and stewardship factor that has increasing come to the fore in all areas of public and private endeavors, given many of the excesses of the 1980's and 1990's, which is about far more than the intended use of resources and the avoidance of moral hazards, but also the achievement of core ideals and values and societal obligations in the conduct of higher education on a daily basis (Burke, 2005; Lederman, 2009; Shavelson, 2010). Where the first force above may be referred to as individuals getting "getting value for their money," the second force can be seen as society "getting value for its money," and it does not take a lot of reflection to see that these two forces may be working and pressuring institutions at cross purposes, and particularly when it comes to program evaluation and improvement efforts and information.

Another part of today's pressure for program evaluation and improvement efforts and information in higher education has come from the Continuous Improvement or Total Quality Management (TQM) movement and approach to carrying out one's mission or charge. Total Quality Management (English & Hill, 1994; Harman, 1994; Dlugacy, 2006; Mulligan, 2012) has become a strong factor and institutionalized in all areas of endeavors, but actually came to education and allied health latter than other areas, as a more real time, dynamic and different kind of statistical approach to evaluation and improvement in a

changing and competitive environment, rather than a fairly static and club-like environment, which it could be said characterized education and health both nationally and internationally until about the mid 1970's. The work of Deming (1986) in particular stands out in this area, even though many in both education and health do not realize that this creative and revolutionizing genius is at the foundation of much of their methodologies and what they do, after Deming's work was recognized and successfully implemented in Japan. What also tends not to be realized by many is that Deming's work and approach runs counter to and is not easily or seamlessly compatible with classical evaluation and improvement models and methodologies (in education in particular), which have been developed and used over the last 150 years and have a pantheon of creative and revolutionary geniuses of their own ranging from Taylor to Thorndike to Tyler to Cronbach to Campbell to Stake to Sufflebeam and many others.

The current force and demand for continuous improvement and the continuous improvement movement, it should be further noted, often works at cross purposes to the other forces mentioned above, and the views and approaches of the classical evaluation and improvement models and forces tend to look for more definitive answers that typically take more time to produce (Stufflebeam, 2001). In particular, there are the often contradictory pressures of the last factor of the five factors in TQM, which is marketing and the marketing movement, for whom evaluation and improvement information are its life's blood. Higher education and its programs have never been more strongly marketed than now and that marketing is happening both nationally and internationally and has created a version of "agenda-driven" as opposed to more neutral program evaluation and improvement information and efforts that further confuses understanding this currently complex landscape in a more organized and systematic way.

I have spent 4 decades doing program evaluation and improvement in K-12 and higher education, business, health and the military, of just about every conceivable kind and just about at every level, as well as trying to help practitioners and those charged with doing program evaluation and improvement carry out their charge, mandate, or mission in sensible and valid ways. I have increasing found that the program development, assessment, evaluation and improvement cycle today is much different and far more complex than it was even a decade ago. In fact, the field now is actually a highly diversified and confusing landscape from both the practitioner's and consumer's view of such evaluative and improvement functions, activities and information, as there are many competing and conflicting forces, approaches and agendas, which make the whole area difficult for the non-specialist and even those who call themselves experts. To illustrate this point more concretely, Stufflebeam (2001) identified 23 different evaluation, assessment, and accountability models that reflect the four major movements and forces outlined above either singly or in combinations. Stufflebeam expanded these 23 models to 31 in 2007 (Stufflebeam & Shrinkfield, 2007), and more models and model variants have been added since then. Mets (2011) tried to analyze and better systematize the models and movements in this field, but concluded that a set of more macro and simplifying categories were needed that were steps towards one or more taxonomies that would help to better organize and represent this seemingly ever burgeoning field. Creating taxonomies, however, is not an easy task, as is well known, and all taxonomies have various advantages and disadvantages (Mezzich, 1980; Godfray, 2002) and relative "goodness's of fit" and usefulness in different contexts and situations.

In an attempt to help the practitioners, managers, colleagues, doctoral students and others with whom I work, I have developed a very simple taxonomy that is quite helpful in organizing and understanding the highly diversified and confusing program evaluation and improvement landscape today. This simple taxonomy allows one to locate and classify various kinds of program evaluation and improvement efforts, activities, methodologies and reports in their own right and as compared to others, but also in terms of the general and developmental nature of program evaluation and improvement efforts today. Like all taxonomies, the primary purpose of the one presented below is to help facilitate communication and discussion between people as well as better situating and contextualizing particulars and particular instances, and what in general they are and are about. Such classifications help to represent an approach or model more appropriately and understand what they provide and do not provide, so that one has more reasonable expectations in a given context, as well as evaluates what has or has not been done more reasonably. Therefore, the taxonomy presented below, as well as this article, is not meant as a definitive or detailed answer to the things it conceptualizes, categorizes, and discusses, but as an advanced organizer for the field currently to help those doing or consuming program evaluation and improvement information to have a broad, simple and useful framework or schema to help guide their more detailed learning. At one level, this simple taxonomy can be seen as one way to group Stufflebeam's 31 plus models of accountability and program evaluation into more macro categories and progressive levels, questions, and functions that are easier and quicker to ascertain, understand, and evaluate in Scriven's (2012) sense of this term.

The Program Assessment and Evaluation Cycle

Many people do not grasp, or do not seem to gasp when talking about program evaluation and improvements, that all programs (like many other things including institutions) are not eternal, and do not spring fully formed and developed from the left ear of Zeus like Athena, but rather have a life cycle and go through a life cycle from birth to suspended animation or death or rebirth of some kind, and that during this developmental life cycle the program and its evaluation and improvement is qualitatively different in several key and important ways at each stage of the cycle. This basic fact obviously means among other things that one may be trying to use the wrong or rather least appropriate evaluation models, methodologies, activities, tools, and information for a particular program, or desired improvement, given where the program is in its development life cycle and the improvement sought. These are several excellent descriptions of program (or product) life cycles (O'Rand & Krecker, 1990), but the critical point of importance here is that all program evaluation and improvement efforts begin and really cannot wisely progress without answering the prime core question, which typically tends not to have been answered when I am asked for consultative help and ask it. This prime core question is:

Where are you? *And what in general are you looking to do*? *Briefly define and characterize* (*give me a model of*) *your venture/program* (*and its general goals*) *and what general kind*

(*Level*) *of* "*evaluation*" *you want to do and why, with the implicit question here being, "where exactly are you on the program's life cycle and do you know and understand this key point that drives almost everything else.*"

Please note that a program of inquiry (or research) to evaluate a particular "venture" (i.e., set of activities or program at any level) can be designed to reflect multiple levels of sophistication and goals. Once you define and characterize your "venture" (and its goals), it is usually quite helpful to "profile" the kinds (levels) of evaluation you wish to do, or ultimately wish to do, and what each level will require both incrementally and developmentally. Your evaluation efforts may progress across levels and across levels over time, but the key is knowing where you are on your venture's developmental cycle, and what "business and evaluation business your are really in right now", as Richard Morley, president of "The Breakfast Club" has help generations of entrepreneurs understand (Morley, 2012). What level(s) you choose also typically depends on where you are in the "program development, improvement and evaluation cycle".

The levels of the program development, assessment and evaluation in a program's or venture's life cycle, and my simple general taxonomy that attempts to capture and organize them are as follows:

Level 1: Venture/Program (Status) Reporting

The goal of this level is to produce on-demand and fairly quick narrative (and quasi-quantitative) reports and similar stories about the program/venture and its current status or/and promise and progress (or not). This type of evaluation is often called managerial or practitioner evaluation but it is also quite often called qualitative evaluation of several different kinds. This type of evaluation is typically characterized by "background homework" activities (briefings) on the program (and its competitors), "census" surveys of various kind, program review activities (and reports), program auditing activities (and reports), news stories (and feature articles), press releases, testimonials, official testimony (and briefings), symposia, conferences, various kinds of case studies and similar activities, all of which often represent different evaluation models from different traditions and somewhat non-commensurate disciplines.

Managerial, practitioner and much qualitative evaluation is typically done without a formal underlying data structure for the venture/program (or a formal venture/program evaluation plan), and with shifting goals and priorities that are most often externally determined. The lack of these two aforementioned features (which are part of the underlying core foundation of more classical evaluation models and higher levels in this taxonomy) are some of the features that characterize "managerial" or "naturalistic" evaluation and its activities from other types and levels of evaluation (Lincoln & Guba, 1985; Pawson & Tilley, 2008).

Managerial, practitioner, and qualitative evaluation is most typically done in a fast-paced, fast-moving setting where there are many competing ventures and priorities and comparatively little response time, and where the external and internal environment must be responded to quickly with mission critical information relative to several different stake holders. Such settings tend to be "institutional" and "action-oriented" in character, and informal R&D and "change oriented" settings where making a quick (initial) response that is then modified at will (usually with comparatively little "high powered and hard" data) is key. Given the context and its features, a wide variety of more qualitative research and evaluation techniques (e.g., interviews, focus groups, open-ended questions) tend to be used at this level, as they are far quicker and easier to both develop and do, and speed and response deadlines are of the essence at this level (Denizen & Lincoln, 2005).

As explained in detail elsewhere (Carifio & Perla, 2009), all research and evaluation methodologies and techniques are "qual-quantifications" or "quant-qualifications" and different sides of the same blanket, so the "qualitative-quantitative controversies" are essentially irrelevant from this perspective, and all methodology is essentially "mixed methods" to some degree. Methodology, therefore, is typically a matter of the mixture as well as the precision and warrants one wants for claims, as well as the complexity of the design and data structure one needs to employ to make different kinds of decisions on different kinds of claims (Mertens, 2010). Consequently, evaluation (and research) methodologies are not competitive but complimentary and well to poorly suited singly or in combinations for the problem and questions to answer (Green et al., 2006).

It also follows from the above points that almost all evaluation and research methodologies are both qualitative and quantitative to some degree or mixture of degrees at the same time with the degrees and mixtures varying according to several factors, including what level of the taxonomy that is being presented here one is currently at or working to be at, as one progresses through the program or venture's life cycle, relative to where one "stops," decides to stay, or "exits" the life cycle. In general, as one progress up the levels of this taxonomy, the evaluation or research one does typically becomes more quantitative and more powerfully quantitative and statistical in sophistication, design and complexity, but that is in great part due to having built the measures and data structures in the activities carried out at lower levels of this taxonomy and thus having the capacity as well as the time needed to implement and carry out this more extensive, sophisticated, and complex quantitative and statistical type and level of evaluation and research. This later type of evaluation and research also does not spring full blown in an instant from the left ear of Zeus like Athena, but must be developed typically as a capacity of the venture and program evaluation efforts over a fairly considerable amount of time.

Further, just because one is doing evaluation or research that is somewhat more qualitative than quantitative (as a mixture) does not mean that one cannot employ an experimental or quasi-experimental approach and actually use an experimental or quasi-experimental/evaluative design, even in case studies (Yin, 2008), as the extensive work of Kleining (1982) has clearly shown. Such qualitative designs and analyses indeed do not have the "power" of more quantitative and statistical designs and analyses, but they can still establish and answer causal and similar type questions. In a word, there is more to qualitative methodology than ethnography and various forms of text and literary analyses and methods, and if done appropriately, one can get valuable and valid findings for decision making at the lower levels or this taxonomy just as one can get fairly useless and invalid quantitative and statistical findings at the higher levels of this taxonomy when blind empiricism and shotgun designs are used. These issues are just not that simple or easy to generalize about definitely in a few words or paragraphs and each case and design must be judged on its own

details, quality, adequacy and merit as Phillips (2005) has pointed out and analyzed in detail. However, venture or program status reporting and evaluation designs typically tend not to be of the Kleining or Yin kind and tend to be more along the lines of managerial and practitioner evaluations as described above.

Lastly, it should also be noted that managerial, practitioner, and qualitative evaluation is often done in settings where there are not the resources to do much else (Bamberger et al., 2004), and this basic fact is an important and contextualizing factor. This type and managerial level of evaluation may suffice and be quite adequate for many ventures, programs and activities, and particularly in their initial phases, but once the "stakes" increase and particularly relative to the claims and assertions one wants to make about the program or venture, higher and different levels and types of evaluation are needed.

Level 2: Data for Decision-Making

The goal here is to select and measure a small set of variables (e.g., percent passing grades, number of users, increases in knowledge etc.) and to link them in a regression "equation" or logical decision-making algorithm of some kind such that by inputting actual "quantitative" data and formulating critical thresholds, decisions can be made to continue/discontinue/expand or modify use of the venture or particular parts of it. In other words, is there an adequate "return on investment" (or "benefits" as compared to lack thereof or/and losses) or promise to continue the venture and keep working on it?

I call this "keep or kill" evaluation, and this kind of evaluation may be formative or summative or retrospective or prospective. The "decision equations" may include stake holder interests and "good will factors" as well as policy and organizational interests and goals. This kind of evaluation is usually the cheapest, less labor intensive and easiest formal evaluation to do as very simple criteria, relatively "low-powered" data and "keep or kill value equations" may be used (or not). Keep or kill (level 2) evaluation is often done with "initial ventures" and prototypes or to make better timely practical and management decisions about a given venture or program, and there is a formal minimum data structure of some kind at this level. Usually, attempts to start building some kind of formal data structure for the venture or program begins at this level and usually to improve the quality and sophistication of the keep or kill equation and statistical analyses so that they are something more than just blind and/or shotgun "number crunching." However, it is also usually at this level that institution and corporate evaluators, evaluation teams and units discover that the institution has an (applications and "business procedures") Management Information System (MIS) which is highly problematic and often fairly useless for the keep or kill evaluations and decision-making that is the goal at this level rather than an Evaluation Information Management System and associated generally useful data structures that are needed for higher quality and more sophisticated keep or kill evaluations and a period in decision making. Being more than "twice-burned" on this critical short coming and flaw usually begins to encourage management and institutional and corporate evaluators to start designing and building more generally useful Evaluation Information Management Systems and data structures so they have better capacities to do this level and higher and more sophisticated levels of program and venture evaluations.

The problems at the data for decision-making level tend to be problems of "good" (reliable and valid) measures that maximize variance and minimize measurement error on each variable included in the functional inputs-throughput-outputs (multiple regression) equation of some kind that will be used, and selecting "power" and "explanatory,"… i.e. as opposed to convenient and locally "believable" variables and their accompanying convenient post-hoc armchair narratives, as often happens in this typically popular blind empiricism approach, which can actually be shotgun evaluation or research of the quantitative or/and qualitative kind, given the setting and the institutional or corporate data available (Schick, 2000). Many evaluation experts have written about the positives of using data for decision-making (Scriven, 2001; Pawson, 2006; Dlugacy, 2006), but many have also written about the flaws, difficulties, poor designs and even poor logic of this approach and over-interpreting and over-generalizing the results, which tends to be very context bound (Phillips, 2005; Coryn, 2007; Sloane, 2008). However, used judiciously, wisely, and for what it is, this level of evaluation is very useful, efficient, fairly timely, and cost effective for making "keep or kill" decisions in particular.

Usually, in my experience, decision-makers tend to wait way too long to make the "kill" decision when doing "keep or kill" evaluations. At one level, I believe that this problem is due to the decision-makers being too invested in too many ways in the venture (including their reputations for championing the venture), and a natural tendency not to want to be disappointed or to disappoint others. However, I also believe that this delay and foot dragging comes from decision makers not being honest about the fact that this is the level and kind of evaluation that they are actually in and doing, and that they need to kill off non-performing and non-promising ventures fairly ruthlessly, as Deming and other counsels, even if they are making a mistake, as the "power of the approach" will eventually assert itself and usually in a better form (Suppe, 1974). Continuous improvement, never mind more extensive change, can be a very slow process, if decision makers drag their feet on the kill decisions or let politics impede these decisions or are engaged in "keep or kill" evaluations only cosmetically which also often happens.

Level 3: Review and Learn

This level is a typical early stage in the evaluation/research/inquiry process, with an emphasis on gathering more information from reviews of available literature, examination of archived records, focus-group style interviews with current users (faculty, students and other stakeholders), and actual questionnaires and "harder measures". The goals are to identify potential key variables worth investigating, how they might be related, and how they can be measured, what actually is "The Theory of the Program (venture)", how sound is it, has it been implemented appropriately and what problems and impediments are being encountered and what might be done about either or both (Aneshensel, 2002). This level is sometimes called "program improvement evaluation" or "getting the program up to its specs", so there is a valid version of the program/venture to evaluate and an appropriate framework and model to interpret the evaluative results.

It cannot be over-emphasized how important this level of program evaluation is in terms of developing a research-knowledge-base and theory for the venture, even if it is only in

first draft form, as these two components are necessary if the venture is not to "fly blind in a changing storm," and not be yet another example of old fashioned shotgun research and evaluation and "black box" empiricism (of both the quantitative and qualitative kind) that is known as logical positivism, which is a much used "vampire" model that more or less died fifty years ago, but is quite difficult to keep in its coffin (Schick, 2000), and particularly so when the "improvement fever" is high. One also needs to have some type of research-knowledge-base and initial first draft or proto-theory of the program if one is going to be "primed for and recognize" and not simply ignore highly important unanticipated consequences or outcome of one's program or venture and evaluation efforts, which may be a ground breaking discovery even if it is one of the petite kind. There is a long and well documented history world wide of accidental and unanticipated discoveries that have occurred in all areas and with all kinds of ventures that made the venture and its efforts a thousand times more valuable than its initial goals or theory. In fact, one could argue that it actually would not be research, evaluation or a major improvement effort, if there were not unanticipated positive (or negative) consequences observed. Obviously, it is the "venture changing" positive unanticipated consequences that are important, but one has to be primed to observe/discover them, and that requires having a research-knowledge-base and theory for the venture, which is also needed to some degree for the next levels in this taxonomy. All of the points are more fully explained and elaborated in Perla and Carifio (2011). A few concrete examples of "review and learn" (level 3) evaluations are Glass (2000), Kenney (2008), Carifio and Perla (2009), and Mets (2011). Also one should not miss that level 3 review and learn evaluations today tend to be quite quantitative and statistical and statistically sophisticated in nature ranging from various form of meta-analysis (Glass, 2000) to quantitative model building (Aneshensel, 2002) and even secondary data analysis and formative causal and structural equation modeling.

Level 4: Defining "Does It Work?"

Does the program/venture actually work in terms of its "advertised capabilities" (and underlying theories)? Is it accessible, trouble-free, convenient, hitting or exceeding the bench marks set on the goals and criteria chosen? One should note that defining "works" and "does it work" is often not an easy thing to do and usually takes considerable effort. For example, are program effects immediate (and how immediate) or delayed (and how delayed) and are they lasting (and how lasting) or temporary (and how temporary). Does the program help some subgroup of students or clients or hurt some subgroup of students or clients or both simultaneously (all forms of different kinds of interactions effects or "workings"). Are the subgroups helped (or/and their advocates) so important mission-wise and politically that it trumps the subgroups hurt (or/and their advocates), and the reverse of this statement. Does the program "stop facial tics" (target goal) but "cause stuttering" in doing so; namely, are there unanticipated consequences, outcomes, or collateral damages (Elton, 1988; van Thiel & Leeuw, 2002; Figlio, 201l). Are there key qualitative differences between the outcomes of the new program (increases comprehension but decreases retention of facts) versus the program it is replacing (always called the "traditional approach," even though it was once the new approach), which produces lower comprehension but higher retention of key facts, and what is the calculus of choice in such a situation, or for saying the new program works or not?

"Does it work" is a very hard question to answer most often and requires a great deal of *a priori* focus and clarity about what "works" actual means, as well as a decent evaluative design and adequate evaluative data structure. It is at this level that the relevancy and adequacy of the general data structure of the institution in which the program or venture is embedded begins to express itself even more strongly than at level 2, and the many weaknesses of the institution's evaluative data structures begin to be discovered relative to being able to actually answer questions about does some program or venture "work." Further if one must move across institutions to answer questions of "does it work," one typically encounters multiple incompatible measures and data structures, which not only impede one efforts, but also helps one to understand the current movements to develop a common standard student or client "unit record" at the K-12 and higher education levels and in the field of medicine as well, to begin to alleviate this major evaluative data structures" problem and enable much better "does it work" evaluations in a much more feasible and cost effective way (Brass et al., 2006).

Also, when it comes to the question "Does it work", there is an unfortunate truism that one must always keep in mind which is that "any program any human can conceive will work for someone somewhere at some point in time, or might appear to do so, if one of Campbell and Stanley now 20 evaluative design flaws are operating in the situation". One must always be extremely cautious that one is not so over focused and over customized in terms of one's program, goals, clients, and situation or context that one essentially has an "sample of one" on everything (i.e., uniqueness or its fuzzy equivalent) or a flawed design or flawed data structure or all three when it comes to questions of "Does it Work". The basic problem here is that if one does in effect have "samples of one" across the board, it really does not matter, as the situation, problem, set of circumstances or client type will never occur again most likely (or extremely rarely), and one is doing a lot of work and making a lot of hoopla for very little return, unless one is in the "rare and orphan disease" business and that is the nature of one's venture. "Does it work" is one of the trickiest questions to ask and answer and particularly in terms of the manner in which this question tends to be asked and often answered by the various stakeholders in this process, which tends to be in a fairly vague, imprecise, somewhat naive, and implicitly personally defined way. These various flaws are some of the major roots and sources of difficulties in answering this question, the others roots being inadequate data structures and designs to actually do the job of saying whether the program or venture actually works or not.

The "Does it work" level in this simple taxonomy means establishing a design and data structure that causally connects the program or venture to its inputs and outputs in a reasonably valid way that allows causal statements and claims to be made about the effects of the program on whom relative to what outcomes and why as opposed to other uncontrolled, unperceived or unknown (exogenous) factors (variables), which is by no means and easy thing to do for many different reasons, which is why "Does it work" designs today tend to be multivariate in character. The "Does it work" level is also focused on understanding the general class of the program/venture and the general theory underlying it, so the evaluation effort is not "over

customized" and a "tunnel vision" effort, but contributes something to the general knowledge-base of the program type and theory that guides it. It looks to "expand the view and knowledge-base" a little bit and build organizational understanding and insight into what kinds of things the venture does and what it and the things that are its chief foci are about in a broader and more general way. And it should be clearly noted that organizational knowledge and wisdom is quite often very important (sometimes called "understanding the business you are actually in as opposed to the business you think you are in"), and sometimes even more important than the much more generalized knowledge and wisdom all of the experts and experts sources in this area tend to focus on and discuss. This organizational knowledge and wisdom, given that it is reasonably valid locally, is the very pay off of these efforts, provide that it is indeed espoused and touted as such (i.e., "this works for us in our context for our goals and clients") and not something more through the various rhetoric and pufferies that are endemic to reporting and dissemination activities now.

Level 5: Formative and Summative Evaluation Research

This level of evaluation represents the standard model of program evaluation research, where various "Stake and Stufflebeam" quasi-experimental and experimental designs and decision-making models are used to do (in the end) confirmatory evaluation of the program/venture, possibly with comparisons to naturally existing "control groups" and purposefully constructed "control" groups as well. Are there unanticipated outcomes and/or side effects of various kinds? "Policy Research" and making decisions to scale a program/venture up and/or disseminate it usually are at this level and this level typically require even better designs, data structures and multivariate analytic techniques than are needed at level 4. Sometimes, this type of evaluation is "high stakes" evaluation, and usually it is also done to provide the program financers, potential program users, and the general public with reasonable information about the veracity and validity of the program's claims (i.e., external social action consumer reports).

This level of this simple taxonomy is well developed and well worked by the pantheon of experts who have assiduously labored at this level for the last century, and the reader is referred to the most representative of these texts (e.g., Stake, 2003; Stufflebeam & Shinkfield, 2007; Pawson & Tilley, 2008; Mertens, 2010), and particularly to Stufflebeam's (2001) classic article summarizing the major models and approaches to formative and summative evaluation that have been developed and used extensively in this type of evaluation work. I have only a few comments of importance to make about this type and level of evaluation in this simple taxonomy.

The first of these comments is to strongly emphasize Stufflebeam's view, which he has expressed in several places, that there really are no direct or straight forward and simple algorithmic connections between formative and summative research and evaluation. Nor are there simple and straight forward transformation of "formative" research and research efforts into "summative" research and research efforts, and the two are essentially different in kind and basically incommensurate. This point, it should be clearly noted, in no way means that one is better than the other, as each has an appropriate setting and context. The point only means that although there are indeed fuzzy overlaps between the two, each has its own appropriate and valid questions, designs, data, analyses, standards and decision-making sets that need to be used, and the formative sets are not necessarily valid or converted or transformed into the summative set and vice-versa. The two, therefore, are qualitatively different and one is actually not necessary for the other. Stufflebeam's important point helps to explain the "disconnects and disappointments" that are often observed between formative and summative evaluations of the same program and the "effects discounting" (diminutions) that typically occurs when the program is disseminated to other settings. However, Stufflebeam's point has also given rise to some promising new approaches at this level which have been exploring the formative and summative evaluations of programs or ventures as (macro-level) case-studies along the lines of those done in business and medicine, as opposed to the classical scientific model and paradigm that is and has been the classical paradigm for formative and summative evaluation at this level for several decades (Stake, 2010). This new line of inquiry has a great deal of potential and particularly relative to building up institutional knowledge and wisdom about an institution's programs and ventures of various kinds.

My second comment of importance here is that the previous four levels are the developmental precursors of this level, more or less to some degree, and may be understood (and even characterized) as missing one or more of the elements in the models that are used to conduct an acceptable and valid evaluation at this level (which is a highly informative way to view and understand each of the previous levels). Each of the previous levels, therefore, is an "approximation" of some kind to this level and the next one. It should also be noted that not a great deal (comparatively) is written about the first four levels in this simple taxonomy, which is why I wrote more about each one of them than these last two levels, nor are these previous levels typically located, situated and contextualize in terms of this level and the next, which is one of the several useful and valuable attributes of this simple taxonomy.

Level 6: "Hard" Research/Evaluation

This level is the most advanced and ambitious level of program or venture evaluation. The goals here are to examine the relationships that exist between multiple antecedent, mediator, and outcome variables through mixtures of regression analyses, quasi-experimental studies, and true experimental manipulations and even national and now international trials. This type of evaluation typically is "high stakes" evaluation and is almost always prospective in character, although approximate retrospective designs/efforts are sometimes possible in certain situations. Often one also tries to estimate the range of outcomes for the program (lower limit results and upper limit results that will be observed and under what conditions) and other similar parameters as well as the decay of effects of the program over time (all effects are usually initially inflated). Often, one also tries to assess how well the program or venture works independent of its originators/founders (is it person or stakeholder dependent) and the degree to which it is "context/site/practitioner" proof (dissemination vulnerabilities). The standards for assessing ROI (Return on Investment) are also usually higher as are the policy questions and evaluations. It is relatively straight forward to see how much more generalized level 6 is than level

5 in terms of its focus and the types of claims it seeks to make, and it's stronger and much tighter focus on causation and establishing strong evidence and warrants for causal claims. There is much ongoing debate about this evaluation level and its requirements (Phillips, 2005; Brass et al., 2006; Coryn, 2007; Sloane, 2008; Scriven, 2010b), and the context and conditions under which it should be attempted and occur, but it is the kind of program or venture evaluation that needs to occur on key issues and goals if we are to build truly generalizable learning, instruction and educational theory. It is also at this level that the availability of stable and general data structures over significant periods of time becomes both critical and key. And once again one sees the importance and value of current movements to develop a common standard student or client "unit record" at the K-12 and higher education levels and in the field of medicine as well, to the longitudinal program and venture evaluations we do that examine and assess the more remote antecedents and the longer range outcomes of the programs and ventures we evaluate at this level in far more sophisticated and higher quality ways.

Summary

As previously stated, there has been strong pressure from just about every quarter in the last twenty years for higher education institutions to evaluate and improve their programs. This pressure is being exerted by several different stake holder groups simultaneously, and also represents the growing cumulative impact of four somewhat contradictory but powerful evaluation and improvement movements, models and advocacy groups. Consequently, the program assessment, evaluation and improvement cycle today is much different and far more complex than it was fifty years ago, or even two decades ago, and it is actually a highly diversified and confusing landscape from both the practitioner's and consumer's view of such evaluative and improvement information. Therefore, the purpose of this article was to present and begin to elucidate a relatively simple general taxonomy that can help practitioners, consumers, and professionals to make better sense of competing evaluation and improvement models, methodologies and results today, which should help to improve communication and understanding and to have a broad, simple and useful framework or schema to help guide their more detailed learning. It is hoped that the simple level 6 taxonomy presented achieves these goals and simplifies this complex area for those involved in evaluating programs and ventures today.

REFERENCES

American Council on Education (2012). National and international projects on accountability and higher education outcomes. http://www.acenet.edu/Content/NavigationMenu/OnlineResources/Accountability/index.htm

Aneshensel, C. S. (2002). *Theory-based data analysis for the social sciences*. Thousand Oaks, CA: Pine Forge Press.

Bamberger, M., Rugh, J., Church, M., & Fort, L. (2004). Shoestring evaluation: Designing impact evaluations under budget, time and data constraints. *American Journal of Evaluation, 25,* 5-7.

Brass, C. T., Nunez-Neto, B., & Williams, E. D. (2006). *Congress and program evaluation: An overview of randomized control trials (RCTs) and related issues*. URL (last checked 24 October 2008). http://digital.library.unt.edu/govdocs/crs/permalink/meta-crs-9145:1

Burke, J. (2005). *Achieving accountability in higher education*: *Balancing public, academic, and market demands.* San Francisco: Jossey-Bass.

Carifio, J., & Perla, R. (2009). A critique of the theoretical and empirical literature on the use of diagrams, graphs and other visual aids in the learning of scientific-technical content from expository texts and instruction. *Interchange, 41,* 403-436.

Coryn, C. L. S. (2007). The "holy trinity" of methodological rigor: A skeptical view. *Journal of Multidisciplinary Evaluation, 4,* 26-31.

Deming, W. E. (1986). *Out of the crisis.* Cambridge, MA: Center for Advanced Engineering Study, Massachusetts Institute of Technology.

Denzin, N., & Lincoln, Y. (2005). *The Sage handbook of qualitative research.* ThousandOaks, CA: Sage.

Dlugacy, Y. (2006). *Measuring healthcare: Using quality data for operational, financial and clinical improvement.* San Francisco, CA: Jossey-Bass.

Elton, L. (1988). Accountability in higher education: The danger of unintended consequences. *Higher Education, 17,* 377-390.

English, F. W., & Hill, J. C. (1994). *Total quality education: Transforming schools into learning places*. Thousand Oaks, CA: Corwin Press.

Figlio, D. (2011). *Intended and unattended consequences of school accountability.* http://www.youtube.com/watch?v=e3aKEuctqy8

Glass, G. (2000). *Meta-analysis at 25*. URL (last checked 15 January 2007). http://glass.ed.asu/gene/papers/meta25.html

Godfray, H. (2002). Challenges for taxonomy. *Nature, 417,* 17-19.

Green, J., Camilli, G., & Elmore, P. (2006). *Handbook of complementary methods in educational research*. Mahwah, NJ: Erlbaum.

Harman, G. (1994). Australian higher education administration and quality assurance movement. *Journal for Higher Education Management, 9,* 25-45.

Kenney, C. (2008). *The best practice: How the new quality movement is transforming medicine.* Philadelphia, CA: Perseus Book Group.

Kleining, G. (1982). *An outline for the methodology of qualitative social research*. URL (last checked 22 October 2008). http://www1.unihamburg.de/abu//Archiv/QualitativeMethoden/Kleining/KleiningEng1982.htm

Lederman, D. (2009). Defining accountability. Inside higher education. http://www.insidehighered.com/news/2009/11/18/aei

Lincoln, Y., & Guba, G. (1985). *Naturalistic inquiry*. Thousand Oak, CA: Sage.

London Times (2012). World university rankings. http://www.timeshighereducation.co.uk/world-university-rankings/2011-2012/top-400.html

Mets, T. (2011). Accountability in higher education: A comprehensive analytical framework. *Theory and Research in Education March, 9,* 41-58.

Morley, R. (2012). R morley incorporated. http://www.barn.org/index.htm

O'Rand, A., & Krecker, M. (1990). Concepts of the life cycle: Their history, meanings, and uses in the social sciences. *Annual Review of Sociology, 16,* 241-262.

Mertens, D. (2010). *Research and evaluation in educational and psychology: Integrating diversity with quantitative, qualitative, and mixmethods approaches.* Thousand Oaks, CA: Sage.

Mezzich, J. E. (1980). *Taxonomy and behavioral science: Comparative performance of grouping methods.* New York: Academic Press.

Mulligan, R. (2012). The Deming University. http://paws.wcu.edu/mulligan/www/demingu.html

Pawson, R. (2006). *Evidence-based policy: A realistic perspective.* Thousand Oaks, CA: Sage.

Pawson, R., & Tilley, N. (2008). *Realistic evaluation.* Thousand Oaks, CA: Sage.

Perla, R., & Carifio, J. (2009). Toward a general and unified view of educational research and educational evaluation: Bridging philosophy and methodology. *Journal of Multi-Disciplinary Evaluation, 5,* 38-55.

Perla, R., & Carifio, J. (2011). Theory creation, modification, and testing: An information-processing model and theory of the anticipated and unanticipated consequences of research and development. *Jour-*

nal of Multi-Disciplinary Evaluation, 7, 84-110.
Phillips, F. (2005). The contested nature of empirical research (and why philosophy of education offers little help). *Journal of Philosophy of Education, 39,* 577-597.
Schick, T. (2000). *Readings in the philosophy of science: From positivism to postmodernism*. Mountain View, CA: Mayfield.
Scriven, M. (2010a). Rethinking Evaluation methodology. *Journal of Multidisciplinary Evaluation, 6,* 1-2.
Scriven, M. (2010b). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation, 31,* 105-117.
Scriven, M. (2012). Evaluating evaluations: A meta-evaluation checklist. http://michaelscriven.info/images/EVALUATING_EVALUATIONS_8.16.11.pdf
Shavelson, R. (2010). Accountability in higher education: Déjà vu all over again. http://www.stanford.edu/dept/SUSE/SEAL/Presentation/Presentation%20PDF/Accountability%20in%20hi%20ed%20CRESST.pdf
Sloane, F. (2008). Through the looking glass: Experiments, quasi-experiments and the medical model. *Education Researcher, 37,* 41-46.
Stake, R. (2003). *Standards-based and responsive evaluation.* Thousand Oaks, CA: Sage
Stake, R. (2010). *Qualitative research: Studying how things work.* New York: Guilford Press.
State Higher Education Executive Officers (2012). National commission on accountability in higher education. http://www.sheeo.org/account/comm-home.htm
Stufflebeam, D. (2001). Evaluation models. *New Directions in Evaluation, 89,* 7-98.
Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models and applications*. San Francisco, CA: Jossey-Bass.
Suppe, F. (1974). *The structure of scientific theories*. Urbana: University of Illinois Press.
US News (2012). *Best colleges and universities.* http://www.usnews.com/rankings
Van Thiel, S. & Leeuw, F. (2002). The performance paradox in the public sector. *Public Performance & Management Review, 25,* 267-281.
Yin, R. (2008). *Case study research: Design and methods (applied social research methods).* Thousand Oaks, CA: Sage.