International Journal of Progressive Education, Volume 10 Number 2, 2014

The Use of Outcome Mapping in the Educational Context

Anna Lewis*
University of South Florida, St. Petersburg

Abstract
Outcome Mapping is intended to measure the process by which change occurs, it shifts away from the products of the program to focus on changes in behaviors, relationships, actions, and/or activities of the people involved in the treatment program. This process-oriented methodology, most often used in designing and evaluating community development projects uses graduated progress markers to determine if the intervention is achieving the desired outcomes and forms the basis for additional monitoring and evaluation. This theoretical paper explores the use of Outcome Mapping as an alternative or supportive method of research design and evaluation in teaching and learning contexts. Outcome mapping can provide educational researchers with the tools to think holistically and strategically about the process and partners needed to achieve successful results. This paper discusses the relevance of this method and compares and contrasts it to the functionality, use, and outcome measures utilized in current educational assessments methods.

* Dr. Anna Lewis is a K-12 Science Educator. She has evaluated professional development programs throughout Florida, in addition to working with international grassroots organizations in developing program evaluations.

Correspondence: ARLewis@usf.edu

Introduction

Educational researchers, especially those funded by outside agencies, are under pressureto demonstrate that their programs (e.g. teacher education programs, new curriculum, or newteaching/learning strategies) result in significant and lasting change. The ultimategoal is to show that the educational program hasgoal is to show that the educational program has increased student academic achievement, often as indicated by standardized test scores. However, increases in student academic achievement is more the product of a confluence of events for which no single program, researcher, or agency can realistically claim full credit. As a result, assessing impacts and actual outcomes are problematic, "Yet many [researchers] continue to struggle to measure results far beyond the reach of their programs" (Earl and Carden 2001, p.2-3).

Table 1: Annual Reading & Math Growth (Hedges, 2010)

Grade Transition	Reading Growth Effect Size	Math Growth Effect Size
K-1	1.52	1.14
2 - 3	0.60	0.89
3 - 4	0.36	0.52
4 - 5	0.40	0.56
5 - 6	0.32	0.41
6 - 7	0.23	0.30
7 - 8	0.26	0.32
8 - 9	0.24	0.22
9 - 10	0.19	0.25
10 - 11	0.19	0.14
11 - 12	0.06	0.01

Based on work in progress using documentation on the national norming samples for the CAT5, SAT9, Terra Nova CTBS, Gates MacGinitie (for reading only), MAT8, Terra Nova CAT, and SAT10. 95% confidence intervals range in reading from +/- .03 to .15 and in math from +/- .03 to .22

Table 1(Hedges, 2010) illustrates one of the reasons why these impacts are so difficult tomeasure. This table shows the average effect size for annual student growth in reading and math learning for grades K-12. If one compares the optimistically expected effect size of Cohen (small=0.2, med= 0.5, large=0.8) or the empirically deduced effect sizes of Lipsey (small=0.15,med= 0.45, large=0.9) (Hedges, 2012) to Table 1, it is obvious that one year of schooling in the early grades can influence student learning to such an extent that it is extremely difficult to discern the impact of a teacher summer workshop, a 2-month program, or the implementation of new teaching/learning strategies with any accuracy. In the later grades the impacts are much more difficult to distinguish simply because achievement scores are more resistant to change, even with a full year of schooling. This task is a particularly onerous one when using standardized tests to demonstrate student achievement, as the discernible standard effect size of these tests is between 0.07 to 0.23 (Hedges & O'Muircheartaigh, 2010). Though not impossible, to do so requires projects that are expensive in both resources and time (i.e. random cluster trials and other such types of experimental designs).

An evaluation methodology is needed that can untangle program impacts on student achievement from normal student growth and other factors that influence students daily. Outcome Mapping offers one approach to this problem. It is an evaluation method that measures the *process* by which change occurs, instead of the end result of the change. This methodology assesses the contributions of educational projects/programs toward the achievement of increases in student

academic outcomes by documenting the changes in behavior, relationships, activities, and actions of a particular group of people or organizations (Rogers, 2012). This method is unique in that the documentation process is developed and maintained not only by the program management, but by all stakeholders, from those who fund the project, to project participants and its iterative, and open to change through periodic review.

This methodology was created for and has been used by international developmental agencies that face significant challenges very similar to those found in educational contexts (IDRC, 2005b). Challenges, such as outcomes or changes in participants can take place before orlong after the program ends. Outcomes may not take the form anticipated, or outcomes and participants may be influenced by those not directly involved in the program. "Outcomes interact with each other and the causes of change usually cannot be isolated", (Carden, Earl, & Smutylo, 2009, p.1), thus making it difficult to attribute change to a specific program or program components Outcome Mapping was adapted from 'Outcome Engineering' (Kibel, 1999) and is intended as a flexible and complementary approach to traditional measurement and evaluation methods. A key innovation of Outcome Mapping is to look at the *results* of a program as ***changes in behavior***.

This paper provides a discussion to briefly explore the differences between research and evaluation, to describe the basics of how Outcome Mapping is used in the developmental arena, and to discuss how systems thinking may be used as a framework in which to justify the use and utility of Outcome Mapping in education evaluations.

Research and Evaluation

There has been substantial debate as to what if any differences exist between evaluation activities and research. It is important to explore the distinction between the two so that one may clarify what knowledge and skills are needed to conduct evaluations and how these differ from the knowledge and skills of social or educational researchers. Early on, experts in the field were divided into one of two camps. Some, such as Michael Scriven, (1998) asserted that there were differences between research and evaluation, but that the two overlap. Others such as Trochim (1998,) argued that evaluation is no different from applied social science. In an effort to simplify the distinction between research and evaluation, these activities have often been caricaturized oroverly generalized which tend to mask the real differences and similarities between the two. Here are some common expressions used to distinguish the two:

- Research generalizes, Evaluation particularizes (Priest, S 2001)
- Research seeks to prove, Evaluation seeks to improve." (M. Q. Patton, 1997)
- Evaluation – so what? Research – what's so? (Mathison, 2008)
- Evaluation – how well it works - Research – how it works. (Mathison, 2008)

None of these expressions captures the complexity of either activity. For instance, where educational research often does make inferences regarding general populations through the use of population sampling, it also makes use of case studies to better understand individuals or instances of particular interest (e.g. narratives of war victims). In addition, though evaluations usually focus on a particular program or project, the outcomes may be generalized and implemented in a broader fashion, as with the case of Head Start programs. This program, first introduced in 1965, resulted in very positive program evaluations which found that six weeks of the program overcame five years of socioeconomic poverty (Currie & Thomas, 1995; Barnett, 1995). This type of evaluation results spurred the widespread proliferation of the Head Start program throughout the US and is one of the longest-running programs to address issues associated with children living in poverty.

The inception of *program evaluation* in the United States was prompted by the Elementary and Secondary Education Act (ESEA) which passed in 1965. The act required that distribution of public

funds be justified. Those that stepped up to do the job found that research methods focusing on hypothesis testing were not well suited to providing information regarding the complex social situations in which schools were embedded. Due to this failure of social science research methods alone to determine the value and efficacy of educational programs,evaluators borrowed from other disciplines and developed new models. As Mathison (2008) suggests, the question we are considering is much like asking "what is the difference between the discipline of mathematics and statistics?" How is the newer discipline, statistics, different from the older more established discipline of mathematics? Just as statisticians created theories and models to establish their unique work, there is also a particular logic followed in evaluation (e.g. Alkin et al., 1979, Patton et al., 1977 and Weiss and Bucuvalas, 1980; Fournier, 1995; Scriven, 1999; Mark and Henry, 2004) with many sub-theories (e.g. Practical Participatory Theory (Cousins & Whitmore, 1998), Values-engaged Theory (Greene, 2005a and Greene, 2005b), and Emergent Realist theory (Mark et al., 1998)).

Not only were new logic models and theories needed but new professional skills were needed to search for unintended outcomes and side effects, to discern significances within different points of view; to report controversial issues and beliefs, and to synthesize facts and principles (Coffman, 2003-2004). Mathison (2004a) suggests an anarchist epistemology had taken precedence in the practice of evaluation, where one method is rejected as supreme over any other and "evaluation as a practice shamelessly borrows from all disciplines and ways of thinking to get at both facts and values" (Mathison 2008). The priority for evaluation focused on what methods delivered the most meaningful information in a given context. Social science *research* methods such as Outcome Mapping, have value in education but when used in an *evaluation* context they are particularly salient as they focus not only on the outcomes but on the value and perspectives of the participants and stakeholders. Thus the essential difference between research and evaluation is the purpose for which they are conducted.

Both research and evaluation require accuracy, which is judged by the validity and reliability of the data collected. However, in addition to accuracy, evaluation is judged by itsutility, feasibility and propriety as described in the *Program Evaluation Standards* (Stufflebeam, 1999). Essential to all evaluation models is the attention to the participants' perspective. Educational research includes the participant but often only in reference *to* whom or *for* whom the data are collected rather than a consideration of the participants' vested interests. Evaluation is subjective in that it is *always* innately bound to the interests of all stakeholders including funders, program management, and participants. In broad strokes there are three distinct phases of program evaluation:

1. ***Needs evaluation*** is typically used in program planning. Just as one would develop a research plan one must determine an evaluation plan. This is done by identifying the stakeholder or client needs, program objectives, program priorities, and resources available and/or necessary in which to conduct the evaluation. Generally, needs evaluations are used to help develop new programs or justify existing program components.

2.*Process evaluation* is most often used to determine the fidelity in which the program is conducted. This phase of evaluation documents is how the program is being carried out by stakeholders compared to the proposed or intended program implementation. Often, process evaluations are used as "reality checks" to help guide implementation by program management and are used in the final program analysis to help better understand the data collected and the program's impact on participants.

3.*Outcome evaluation* characteristically determines the overall effects or impacts of the program in relation to the initial program objectives. Good outcome evaluation not only indicates whether the program objectives were met but also documents any unintended effects.

In the next section we will describe Outcome mapping as it is currently used to evaluate developmental programs while drawing comparisons to traditional education program evaluation practices and the three types of evaluation as described above.

Outcome Mapping (OM) as an Evaluation Tool

Many educational programs use a 'project centered' approach in evaluation, where management has the controlling role in the needs, process, and outcome evaluation design. Outcome Mapping (OM) as implemented by international development agencies brings a 'partnered center' approach into focus with a loosening of control at the management level. To accomplish this shift in perspective a very deliberate process is followed to enhance the development of all three phases of evaluation.

Experience has shown that development (like learning in science education) is a complex process that does not take place in isolation. "Linear cause and effect" thinking contradicts the understanding of development as a complex process that occurs in open systems" (Earl & Carden, 2002). Some simplification is necessary to create and implement programs, however, the contextual reality of any project or program must be acknowledged. In addition, development outcomes do not occur with a clear beginning, middle, and end delineation. Often programs make a difference that is incremental and cumulative rather than a single measurable event. It is also reasonable to expect that the intended outcome may actually be achieved after the program has ended or that outcomes may erode over time due to other influences entirely outside of the program activity. Persons involved in evaluation of curricula, professional development, learning resources, etc., can relate to each of the issues described.Developers of OM deal with these issues by focusing on the contributions rather than attribution of their programs. Perhaps the most notable characteristic of OM is that it focuses on ***outcomes as behavioral change***. OM attends to the problem of attribution of impact by increasing the value and attention placed on changes that "are clearly within the program's sphere of influence. . . this appears to suggest concentrating on easier, less important, short-term achievements, in fact it does the opposite. Rather the evaluation focuses attention on incremental, often subtle changes, without which the large-scale, more prominent achievements in human well-being cannot be attained or sustained" (Earl & Carden 2001, p.10). Attention is placed on targeted behaviors and relationships within the scope of the program, as well as increasing its effectiveness in relation to project goals. Reporting requires managers to demonstrate that they are progressing *toward* impact and improving effectiveness – but not accountable for the impact itself. In this way accountability becomes rational rather than empirical – the intended "impact" of the program is the ideal to aim for rather than the yardstick against which performance is measured. "Thus the threat of failing to discover 'hidden attribution' is eliminated when feedback on performance concentrates on improving, rather than on proving, on understanding rather than on reporting, and on creating knowledge rather than on taking credit" (Earl & Carden 2001, p.10). This is in contrast to the usual methods used in education evaluation which seek causal relations between the intervention and observable change.

Outcome Mapping is not intended to assess the relevance of a programming area or an evaluation of the cost-effectiveness of one approach compared to another. Nor is this method, as currently used in developmental work, useful for very small projects. Most importantly, if the project is not in a position to change the behavior of participants then this approach would not be appropriate (Earl & Carden, 2002). It is this last point that makes Outcome Mapping particularly suited to designing and evaluating education projects, as most educational endeavors have at least one component that is predicated upon teachers and/or students learning new skills and behaviors. In fact, the focus of Outcome Mapping is the change process which occurs in those who are ***directly*** interacting with the program.

Another important aspect of this method is to recognize that change is a reciprocal relationship (Rogers, 2012). By acknowledging that participants are not only influenced by interventions but also affect the intervention itself, evaluators can capture important information that goes beyond "pilot testing" or "program updates based upon user feedback" to understanding specific mediating factors in teaching and learning for the populations the program serves.

Outcome Mapping provides a continuous system for thinking holistically and strategically about achieving results. Outcome Mapping does this by monitoring three key areas: ***changes in the behavior***

of partners within a program, the program's strategies, and the way in which a program functions as an organizational unit.

The process of OM is not a discrete event but cyclical in nature where ***needs, process, and outcome evaluations*** are utilized and attended to throughout the life of the project. The needs assessment is conducted before the onset of the program, and is revisited regularly to attend to new needs or issues that may arise. The process assessment is also conducted regularly and involves identified stakeholders and boundary partners. Frequent feedback is elicited anddocumented to allow the program to be responsive to the needs of these partners. Since outcomes are based on behavior changes in the partners, tracking these changes is continually monitored and documented and not simply evaluated at the end of the project. OM developers recommend performing a three-stage process at the beginning of the project to develop the initial strategic plan, data collection tools and a tracking system which would be reviewed and updated as needed throughout the life of the project (IDRC, 2005b). OM helps developmental project managers establish who the stakeholders are, how they will be affected, by which activities in the programs, and how the outcomes will be documented and assessed. The following is a brief synopsis of the three stages of OM as used in developmental programs, and highlights how each stage may be used in educational contexts.

Stage 1 Intentional Design. This initial step is where the researchers, project management team, evaluators and project participants or boundary partners outline and clarify at the macro-level the outcome challenges they would like to support. These are the 'downstream' impacts the program is working to achieve. This outline provides reference points to guide strategy formulation and action plans (rather than acting as performance indicators). In addition, this outline is used to develop progress markers for each boundary partner, which in turn is used to track performance at each level. These progress markers identify incremental changes that the program may realistically influence which prompt behavioral change and build the foundations of sustained social change (Carden, Earl, & Smutylo, 2009).

It is envisioned that in implementing this step in educational projects one might employ concept maps to assist in identifying the behaviors and other affective components associated with academic achievement in addition to the standard tests and surveys now employed. What these affective components might look like would be impacted by the stakeholders present whowould be invited to openly share their experiences and perspectives. Especially important are the boundary partners who have unique perspectives that can assist in making the intangible process of learning visible.

Stage 2 Outcome and Performance Monitoring. At this stage a performance monitoring framework is designed based upon the ground work in Stage 1. Three common data collection tools are developed at this Stage: 1) an outcome journal that documents boundary partner actions and relationships, 2) a strategy journal that documents strategies and activities of all boundary partners, and 3) a performance journal that documents the organizational practices that keep the program relevant and viable (Earl & Carden 2002; Smutylo, 2005). With this framework a broad range of monitoring information may be identified and tracked. The challenge at this Stage is to identify what information is needed and at what level.

Applying Stage 2 in the educational context would entail selecting from the cadre of tests and surveys those that are appropriate to measure the project outcomes. More importantly though, this step would include choosing and/or developing the affective and behavioral components that have robust and reliable constructs and identifying or building instruments to measure these constructs.

Stage 3 is Evaluation Planning. Here priorities are set so that evaluation resources and activities may be targeted where they will be most useful. It is in this stage that the main elements of the evaluation design are pulled together and finalized. Here the details are decided upon, such as the priority evaluation topics, issues, and questions. Also, what data is to be collected, the person(s) responsible for collecting the data, the time frame and the cost of conducting the evaluation plan.As with

developmental projects, in educational contexts Stage 3 would entail the finalization of the evaluation plan. Here the three types of evaluations: Needs, Process, and Goals/Outcomes) are intertwined into one to ensure a cohesive program of evaluation that is conducted and maintained throughout the life of the project. It is important to note once again that this final plan is a holistic approach. It recognizes that the needs and processes to meet those needs may change to better meet the program goals. Therefore, the final plan always entails monitoring three key areas: **changes in the behavior of partners within a program, the program's strategies, and the way in which a program functions as an organizational unit.**

There are several assumptions and weakness inherent in this methodology that should be acknowledged. Some of these assumptions are:

- The belief that knowledge is socially constructed – this is especially true in educational contexts (what is taught and when for example).
- Evaluators are committed to the value of inclusion and the democratization of public conversations between all stakeholders.
- Evaluators are committed to act impartially, attending to the interests of all stakeholders and not privileging one group over another.
- Boundary partners have a level of self-awareness (or at least the ability to attain that level with training and practice) that will enable them to contribute to identifying behavioral markers.
- All stakeholders are encouraged to participate throughout evaluation but it is essential at the beginning and conclusion.

The project management team has the organizational will to integrate the evaluation which may entail modifying or adjusting strategies at each iteration of the evaluation review cycle.

A few of the inherent weaknesses are:

- Micro-politics that often appear at every level of the project may influence the success of the participatory approach.

- Entrenched values may prevent the use of findings in decision making (e.g. suggestions may be dismissed or marginalized).

- The practical logistics of gathering representatives from all stakeholder groups together for the initial meeting to develop behavioral markers and other evaluation markers.

- Participation of stakeholders may wane if not invested in the process either by financial obligation (paid to participate), emotionally committed (cares about the project), or professional support (administrative interest or obligation).

- Stakeholder groups may have competing interests which must be identified and resolved before this methodology can be implemented.

There is no single answer to address these weaknesses. However, by being aware of them, attention may be given to ameliorating their impact on the overall evaluation. As it has been described, Outcome Mapping offers a participatory methodology that assists evaluators in developing a system that can be used to meet both accountability and project assessment needs. In addition, this methodology has also shown promise for cross-program evaluations in that it can facilitate a standardization of indicators without losing the unique richness of each program,thus combining both quantitative and qualitative approaches. The next section explores the utility of OM and a rationale for its use in education evaluation.

General Systems Thinking

As discussed earlier, traditional social research methods have been found limiting for conducting comprehensive evaluations of complex social interactions such as those found in education. However, this is a topic for further discussion and validation in another paper. New approaches to education evaluation need to take into consideration the inherent unpredictability and the underlying values, norms and behaviors that shape responses to education interventions and programs. This requires a different mindset that creates the conditions for contextualized solutions. A "general systems thinking" or "systemic" approach offers compelling perspectives for all aspects of education especially in the areas of evaluation. Broadly speaking, systemics provides a methodological framework for understanding phenomena that emphasize the relationship *between* parts rather than simply focusing on the parts themselves. These relationships are driven by feedback loops which are often complex and invisible. Systemics has grown into widespread use in many areas of business, manufacturing, and economics because it offers an approach to complex and persistent issues, issues not unlike those found in education. The intent of this paper is not to fully explicate the use and function of systems thinking, but rather, to pull a filament from the systemic tool box to develop a context and rationale for the use of Outcome Mapping in evaluating education programs.

We live in a complex world of systems, made up of people, groups of people, things, rules, practices, and constraints. In each domain, systems create patterns of activities which help individuals accomplish their goals and most often help those individuals interact with one another (e.g. air traffic control systems, banking systems, wireless networks). Every system thatis created or occurs naturally embodies a tension: one of *responsiveness* at the local level or the parts of the system and at the same time the system as a whole provides *coherence*. From a systemic educational perspective we would like the *parts* of a system to be *responsive* to local circumstances and the system *as a whole* to be *coherent*. This may mean a single individual if one is looking at the "class" as the system, or it may mean how individuals and classes respond to themselves and each other if the system in question is the school. As Harrison & Henderson (2010) suggest "the more a system's parts are responsive to the diversity and dynamism of the world, giving people the ability to meet their needs, the less we can know about how the whole system will behave. The more the system drives towards coherence, the stronger the relationships between the parts and the less freedom each part has to adapt to its circumstances in unexpected ways" (Harrison & Henderson, 2010)). This interplay is often seen as a zero-sum choice in which a gain for one side or characteristic entails a corresponding loss for the other side or characteristic. In this example, to increase responsiveness one must lose control and continuity or lose responsiveness to maintain control and overall system coherence.

In the world of educational evaluation we might see Randomized Control Trials (RCTs) at one end of the spectrum (high cohesion) where there is very tight control over the evaluation. Here each variable is accounted for and assigned, allowing very little freedom or responsiveness in order to maintain reliability and validity over the variables being measured. On the other end one might place Phenomenological studies where the evaluator puts aside any structure, control, or preconceptions to document experiences and perspectives from the vantage point of the subject. Here there is great freedom or responsiveness on the part of the individuals and almost no structure or cohesion imposed by the researcher. Other dichotomies in education can be similarly viewed, such as teacher-centered or student-centered instruction and traditional vs3inquiry pedagogy. All of these can be illustrated as a binary choice along a zero-sum continuum. Figure 1 shows the dichotomy between educational evaluations.

Coherence (RCT) ---------- Responsiveness (Ethnographies)

Figure 1

However, Harrison & Henderson (2010) offer another proposition where these dichotomies are viewed systemically. They recommend we examine the tension itself – that we examine the tension in terms of the relationship between the two ends of the spectrum. Thus the zero-sum line transforms into

a trade-off curve, as shown in Figure 2. Trade-off curves are used in many types of design practices and describe the limits of performance that are possible within a given design approach. Typically they characterize the relationship between two or more key parameters.

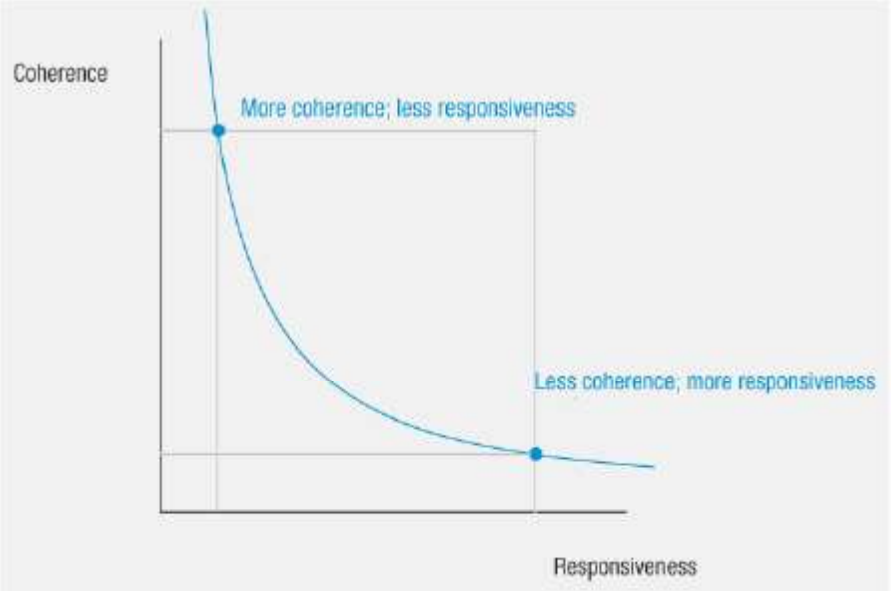

Figure 2

We can move from the zero-sum trade-off curve (as shown in Figure 2) by redesigning processes. This then makes it possible to do *worse* than zero-sum, a bad system can be both incoherent and unresponsive; or we can improve a system and do *better* than zero-sum by improving both coherence and responsiveness. In manufacturing, we can move to the higher trade-off curves by making processes more efficient or finding better materials. One can slide to the lower trade-off curve by poorly maintaining the factory or using inferior materials.

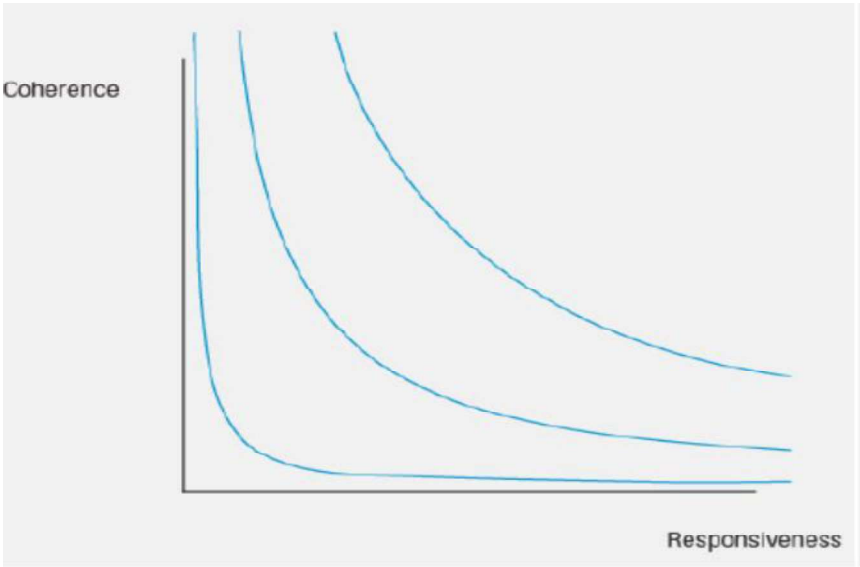

Figure 3

In this paper trade-off curves allow us to systemically conceptualize how outcome mapping can move us to a higher trade-off curve which increases both the coherence and responsiveness of education evaluation activities.

Moving Education Evaluations to Higher Trade-off Curves

Several caveats are in order before venturing further. First, the conceptual framework for the use of Outcome Mapping offered here is an initial, suggestive one. It has not been evaluated, although it is hoped that if it appears promising it will be subjected to extensive testing in practice. Second, the example given, though familiar, is for the most part created for purposes of this discussion. There is no literature beyond that found in international development where Outcome Mapping has been used. Third, it is hoped that this paucity of information should be seen as an opportunity for further investigation and exploration of this methodology in the context of education evaluation. We begin by examining an imaginary project and apply Outcome Mapping to introduce a higher trade-off curve into the project evaluation effort.

Take for example a project that is looking to test the impact and efficacy of a new curriculum for an earth science (any discipline would work here) class for all 6th grades students in a particular district. Most often, teachers are simply given new texts or resources and told to use them in their classrooms. Teacher training may or may not be provided in the particular content or pedagogy implemented in the curriculum being evaluated. As is often the case, textbook and curriculum developers tend to believe that their resources are teacher-proof and applicable across diverse classrooms and populations. Tips and pointers may be provided to teachers for students with special needs in the end notes of the teacher resource. The curriculum is developed with the intention of providing coherence to the educational event or system. Daily guides indicate what will be taught each day and administrators perform spot-checks in classrooms to ensure teachers are on task and the targeted content indicated by the curriculum map is being taught. This is often an effort to ensure the program is implemented with fidelity.

At the end of each unit and at the end of the semester standardized exams are distributed and scores documented.

The evaluation for this type of project would usually seek to show the curriculum positively impacted science achievement or knowledge of science content. It may also wish to show that teachers and students enjoyed the process and through the use of the curriculum students were motivated to continue to study science. To “prove” the impact of the curriculum it would be necessary that *coherence* to the curriculum (or system) would take priority over *responsiveness* to the teacher and/or student needs (parts of the system). This does not mean that the evaluation must be a Randomized Trial, but rather when any evaluation seeks to prove causal effect, coherence must take precedence over responsiveness otherwise the number of confounding variables would be so great no causal links could be made. When local needs don’t fit a system’s design, the drive for coherence can make local work inefficient. In response, people do whatever they can to adapt the system to their needs. For instance, in our example teachers may find their students need more time to learn the content than allotted on the curriculum map and therefore skip units to attend to student needs. It’s possible that the teachers may not know the science content being taught and therefore may pass along misconceptions or not address student misconceptions which result in poor test grades. Our trade-off curves show that when we examine the relationship between cohesion and responsiveness humans will tend to augment the system and reshape activities of the project to suit the reality of their experiences, thus making it difficult to adhere to cohesion and prove causality.

However, it has been shown that systems can move to higher trade-off curves through local and non-local adjustments (Harrison & Henderson, 2010). Outcome Mapping offers a means of providing local and non-local changes in interaction between cohesiveness andresponsiveness, which can be a profound means of moving to higher trade-off curves. By *shifting our attention from attribution to*

contribution and progressing toward impact and improving effectiveness, evaluators are capable of finding the right balance between generality and particularity. This shift in perspective allows for simple infrastructures to manage very complex interactions. The infrastructure recommended by OM is highly permeable and permits a high level of control by users, while still giving the collaboration or system as a whole the ability to maintain coherence.

Outcome Mapping provides a continuous system for thinking holistically and strategically about achieving results. Recall that OM does this by monitoring three key areas: **changes in the behavior of partners within a program, the program's strategies, and the way in which a program functions as an organizational unit**. In the above example, if we use OM to evaluate the new 6th grade curriculum, our focus shifts from "proving" the curriculum increases science scores to showing *changes in behavior that would lead to* increased science scores. In addition, since boundary partners such as administrators, teachers and students participate in defining and recording these behavior changes on a recurring basis and provide ongoing insight regarding obstacles to behavioral change, the system (current curriculum, curriculum developers and/or evaluators) can respond to these local needs by adapting the program's strategies as required. Such adaptations are *essential* for enabling systems to respond to a complex, diverse and changing world. Though the strategies may change the outcome from "this curriculum increases student achievement scores in science" to "this curriculum increase positive behaviors (specific behaviors would be identified and clearly articulated) that lead to higher achievement in science"; the newly developed curriculum remains steady, resulting inhigh system coherence. These local and non-local adjustments raise the level of the trade-off curve as the system becomes permeable to human concerns.

In addition, Miller and Campbell (2006) reviewed 46 studies where "stakeholder empowerment" though measured in diverse ways, found that when any stakeholder felt empowered, outcomes improved. This was especially true when ''group process in which the group collectively decided on the evaluation aims, evaluation design, and evaluation procedures and collaboratively collected, analyzed, and reported evaluation data'' (p. 305).

In the international development world, program stakeholders such as local community funders, service providers, and boundary partners have worked together to develop a common core set of indicators and measurement tools that can be used to regularly collect data and guide evaluation. Educational evaluators wishing to explore the use of OM could, over time, identify appropriate core outcome indicators and measures specific to their needs. As eluded to earlier in the Stage 2 description, these might entail such things as increases in class attendance and student participation, demonstrated curiosity, the willingness to learn something not previously known, tolerance to ambiguity, no expressed anxiety regarding test-taking, and an interest in sharing their knowledge. In addition, boundary partners could assist in capturing learning experiences that may illuminate how the learning process of participants is enhanced, and assist in making visible otherwise tacit activities, behaviors, and knowledge. This would entail identifying robust and reliable affective constructs and building instruments to measure these constructs. Many such instruments are available but lack the research base to validate them (Liu, 2010) due to the changing understanding of affective constructs These types of measures would be used in addition to the usual cadre of standard science content tests, surveys, and classroom observations.

Change in education is slow and laborious. The use of Outcome Mapping in conjunction with the usual evaluation methods would be a positive move in shifting our perspectives to a more systemic way of viewing education evaluation; an affirmation that the process of learning is truly a personal and individualized endeavor - that attention to the journey is as important as arriving at the destination.

Conclusion

Education is a complex endeavor; interventions, curriculums, and professional development are more like networked interactions between stakeholders than linear processes of problem articulation,

project design, and implementation. Complex problems require strategies that entail changes in established patterns of action. Utilizing systemic evaluation frameworks can play a role in helping to move educational developers, policy-makers, and researchers to embrace a more realistic approach to identifying patterns and resolving problems in education. Pressure to be accountable for impact leads to conceptualizing and evaluating programs as successful or unsuccessful. However, experience tells us that education, like all social development, is more complex and cannot be isolated from the actors with which it will interact, nor insulated from outside influences. We need to make sure that the tools we have at our disposal for evidence generation are sophisticated and nuanced to acknowledge this messy reality, and that we are sharing ideas on how to do this in a practical and affordable way. Outcome Mapping provides a continuous system for thinking holistically and strategically about achieving results. OM assumes that in reality it is the boundary partners who control real change and the programs themselves are simply facilitators of the process by providing opportunities, training, and resources.

References

Alkin, M., Daillak, R., & White, P. (1979). Using evaluations: Does evaluation make a difference? Beverly Hills, CA: Sage.

Ambrose, K., Kelpin, K., Smutylo, T. (2012) Oxfam: Engendering Change Program Mid-Term Learning Review FINAL REPORT. Southern Africa, Horn & East Africa, and Americas Workshops. Oxfam Canada EC Program Mid-term Learning Review: Final Report, May 8, 2012.

Barnett (Winter 1995). "Long Term Effects of Early Childhood Programs on Cognitive and School Outcomes". *The Future of Children* **5** (3): 25–50.

Carden, F., Earl, S., & Smutylo, T. (2009). *Outcome mapping: Building learning and reflection into development programs*. International Development Research Centre (IDRC)

Coffman, J. (2003-2004, Winter). Michael Scriven on the differences between evaluation and social science research. *The Evaluation Exchange, 9(4).*

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

Creswell, John W (2009). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage Publications, Inc; 3rd edition (July 15, 2008).

Currie J., Thomas D. (1995). "Does Head Start Make A Difference?" *American Economic Review* 85 (3): 341.

Fournier, D. (Ed.) (1995). Reasoning in evaluation: lnferential links and leaps (New Directions for Evaluation no. 48). San Francisco: Jossey-Bass.

Greene, J. C. (2005). A value-engaged approach for evaluating the Bunche-Da Vinci Learning Academy. New Directions for Evaluation, 106, 27–45.

Greene, J. C. (2005a). A value-engaged approach for evaluating the Bunche-Da Vinci Learning Academy. New Directions for Evaluation, 106, 27–45.Greene, J. C. (2005b). Evaluators as stewards of the public good. In S. Hood, R. K. Hopson, & H. T. Frierson (Eds.), The role of culture and cultural context: A mandate for inclusion, truth, and understanding in evaluation theory and practice (pp. 7–20). Greenwich, CT: Information Age Publishing.

Hansen, M., Alkin, M. C., & Wallace, T. L. (2013). Depicting the logic of three evaluation theories. Evaluation and Program Planning, 38, 34–43.

Harris, J, Henderson, A. (2012) Coherence and responsiveness. Interactions. Volume 19 Issue 5, September + October 2012 Pages 67-71 ACM New York, NY, USA.

Harris, J. , Henderson, A. (2012). Coherence and responsiveness. Interactions. 19, 5 (September 2012), pages 67-71.

Hedges, L.V. & O'Muircheartaigh, C.A. (2010). Improving generalization from designed experiments. Working Paper, Northwestern University.

Hedges, L.V. (2012). Sample Design for Group Randomized Trials, Presentation for the 2012 IES/NCER Summer Research Training Institute at Northwestern University.

IDRC, (2005b). Facilitation manual and facilitator summary sheets: http://www.idrc.ca/en/ev-62236-201-1-DO_TOPIC.html

Kibel. B. M. (1999). Success stories as hard data: an introduction to results mapping. Springer Publications.

Liu, X. (2010). Using and developing measurement instruments in science education: A Rasch modeling approach. Iap.

Luskin, R., Ho, T. (2013) Comparing the intended consequences of three theories of evaluation. Evaluation and Program Planning. 38 (2013) pages 61-66.

Mark, M. M., & Henry, G. T. (2004). The mechanisms and outcomes of evaluation influence. Evaluation, 10, 35–57.

Mark, M. M., Henry, G. T., & Julnes, G. (1999). Toward an integrative framework for evaluation practice. American Journal of Evaluation, 20, 177–198.

Mathison, S. (2004a). An anarchist epistemology in evaluation. Paper presented at the annual meeting of the American Evaluation Association, Atlanta.

Mathison, S. (2008). What is the difference between evaluation and research, and why do we care? In N. L. Smith & P. R. Brandon (Eds.), Fundamental issues in evaluation. New York, NY: The Guilford Press.

Melvin M. Mark, Gary T. Henry, George Julnes , Melvin Mark, Gary Henry (2011). Evaluation: An Integrated Framework for Understanding, Guiding, and Improving Policies and Programs

Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. American Journal of Evaluation, 27, 296–319.

Ottawa.Earl, S., & Carden, F. (2002). Learning from complexity: The International Development Research Centre's experience with Outcome Mapping. Development in Practice, 12(3-4), 518-524.

Patton MQ. (1997) Utilization-focused evaluation. Thousand Oaks, CA: Sage Publications.

Patton, M. Q. (1997). Utilization-focused evaluation: The new century text. Thousand Oaks, CA: Sage.

Patton, M. Q., Grimes, P. S., Guthrie, K. M., Brennan, N. J., French, B. D., & Blythe, D. A. (1977). In search of impact: An analysis of the utilization of federal health evaluation research. In C. H. Weiss (Ed.),

Priest, S. (2001). A program evaluation primer. Journal of Experiential Education, 24(1), 34-40.

Rogers, Patricia (2012). When to use outcome mapping. http://www.internationalpeaceandconflict.org/profiles/blog/show?id=780588%3ABlogPost%3A728981&xgs=1&xg_source=msg_share_post Retrieved May 1, 2013. Professor of Public Sector Evaluation at the Royal Melbourne Institute of Technology University.

Scriven, M. (1998, January 16). Research vs. evaluation. Message posted to bama.ua.edulcgi-binlwa?A2=ind9801C&L=evaltalk&P=R2131&1 =l&X= 2Fl1357E5870213C59&Y.

Scriven, M. (1999). The nature of evaluation: Part I. Relation to psychology. Practical Assessment, Research and Evaluation, 6(11).

Shaw, Ian Graham Ronald. (2006) The SAGE Handbook of Evaluation. London : SAGE Publications.

Smutylo, T. (2005). Outcome mapping: A method for tracking behavioral changes in development programs, ILAC Brief 7.

Smutylo,Terry (200) Outcome mapping: A method for tracking behavioral changes in development programs online under: http://www.outcomemapping.ca/resource/resource.php?id=182.

Stufflebeam, D. (1999). Metaevaluation checklist. Retrieved December 15, 2005, from http://www.wmich.edu/evalctr/archive_checklists/eval_model_metaeval.pdf

Trochim, W. (1998, February 2). Research vs. evaluation. Message posted to bama.ua.edulcgibinlwa?A2=ind9802A&L=evaltalk&P=R503&l=1&x=089CE94613356B8693 &YUS DOE (2002) “What Works Clearinghouse” http://ies.ed.gov/ncee/wwc/

Weiss, C. H., & Bucuvalas, M. J. (1980). Social science research and decision making. New York, NY: Columbia University Press.

Xiufeng Liu, Calvin S. Kalman (2010) Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach (HC) (Science and Engineering Education Sources) Information Age Publishing.