Program Evaluation Primer: A Review of Three Evaluations

Craig M. McGill, Ardith Clayton-Wright, and Mia Heikkila
Florida International University, USA

Abstract: Program evaluation is an essential process to program assessment and improvement. This paper overviews three published evaluations, such as reduction of HIV-contraction, perceptions of teachers of a newly adopted supplemental reading program, and seniors farmers' market nutrition education program, and considers important aspects of program evaluation more broadly.

Few human resource development (HRD) scholars, professionals, and practitioners would argue that the sub-field of program evaluation is not essential to the learning and performance goals of the HRD profession. Program evaluation, a "tool used to assess the implementation and outcomes of a program, to increase a program's efficiency and impact over time, and to demonstrate accountability" (MacDonald et al., 2001, p. 1), is an essential process to program assessment and improvement. Program evaluation (a) establishes program effectiveness, (b) builds accountability into program facilitators and other stakeholders, (c) improves the implementation and effectiveness of programs, (d) assists with the allotment and management of limited resources, (e) is important for marketing a program, (f) helps to justify existence of budget for program, and in its ultimate purpose, (g) is critical for the continuous development and improvement of the program. Program evaluations can be approached from a number of different paradigms, and this paper focuses on Kirkpatrick's (1975) four-level model of evaluation. The purpose of this paper is to review three program evaluations in differing fields to examine similarities within the three program evaluations based on the Kirkpatrick's model. In order to understand the basis of the reviews, Kirkpatrick's model is discussed in the following section.

Kirkpatrick's Four-Level Model of Evaluation

Although there are many different possible ways to approach program evaluation, one model has been in operation for six decades. First introduced in 1959, Kirkpatrick's four-level model is one of the most commonly used approaches of program evaluation. Bassi et al. (1996) reported that 96% of companies surveyed used some form of the Kirkpatrick framework to evaluate training and development programs. Twitchell, Holton, and Trott (2000) performed a meta-analysis of studies done in the last 40 years. Their research indicates the following ranges for the use of Kirkpatrick's four levels: level 1 (86-100% of surveyed programs), level 2 (71-90% of surveyed programs), level 3 (43-83% of surveyed programs), and level 4 (21-49% of surveyed programs). Although some companies do not use the model for all four levels, all four levels of the evaluation should be utilized to avoid biased conclusions (Kirkpatrick & Kirkpatrick, 2006).

The versatility of Kirkpatrick's (1975) model allows it to be used for the evaluation of the effectiveness of any program. Many things should be considered when conducting a program evaluation: (a) determining program needs, (b) setting objectives, (c) determining subject content, (d) evaluating the program, (e) selecting participants, (f) determining the best schedule, (g) selecting appropriate facilities, (h) selecting appropriate instructors, (i) selecting/preparing audiovisual aids, and (j) coordinating the program (Kirkpatrick & Kirkpatrick, 2006).

In designing his model, Kirkpatrick (1975) considered what impact the training would have on participants in terms of their reactions (level one), learning (level two), behavior (level three), and organizational results (level four).

Level one, reaction, simply evaluates the extent to which the trainees liked the program (Kirkpatrick, 1975). First, the evaluators must quantify the key determinants of the program expectations, design the program around the key expectations, and then, evaluate the trainees' reaction to the program designed around the expectations. Determining the information that is needed to refine the evaluation process and design an evaluation that will quantify the reactions of the participants is crucial during this period. Second, the evaluators must create a form to measure participant reaction and decide how to capture it. A set of standards is needed to measure the reaction of the evaluation process. It is important that the participants' perception of the evaluation process is positive, which should be reflected in the immediate written response with comments. In addition, they need to encourage written comments in addition to the multiple choices (in such design). For most accurate results, 100% of the answer sheets should be collected, which can be maximized by the agenda design. If the program participants are allowed to complete the evaluation before leaving the training, such as prior to a price drawing that engages the audience, a maximum results can be achieved (Kirkpatrick & Kirkpatrick, 2006).

Specific objectives of the program need to be developed for level two, learning (Kirkpatrick & Kirkpatrick, 2006). This is the phase in which the learning evaluations should be targeted to the specific objectives of the program and should be used to evaluate all projects. Learning can be measured immediately after the training or very shortly after the training has occurred. The evaluators should consider whether or not the participants understood the concepts, principles, and techniques presented by trainers and whether or not the trainees acquired new and improved skills or attitudes. Learning can be evaluated by measuring knowledge, skills, and attitudes by (a) measuring knowledge, skills and attitudes before and after the training, (b) using paper-pencil test for knowledge and attitudes, and (c) developing performance measures. A 100% response is desirable, and using a control group would enhance the design, although it is often not practical.

Evaluation of level three, behavior, attempts to answer the question of whether the training has been transferred to daily activities: "Are the newly acquired skills or attitudes being used in the environment of the learner?" (Kirkpatrick & Kirkpatrick, 2006). Behavior can change on if the condition is conducive (Mind Tools, n.d). Measuring behavior changes is one of the most important and often most neglected particulars of evaluation (Kirkpatrick & Kirkpatrick, 2006). The fact that a trainee succeeded in the learning objective does not translate to the trainee behavior changes at the work environment. The change is not necessarily in any way linked to the trainee. The changes may not have occurred due to various reasons, including supervisor resistance to apply changes, lack of trainees' positive attitude regarding the changes, lack of opportunities, changes in job description, policy changes and other reasons unrelated to achieving the learning objectives. The goal of the Level 3 evaluation measures not only if behavior changes occurred despite the multiple factors that may have prevented it, but also attempts to identify the reasons it may not have occurred. In measuring the participant's behavior, the following guidelines are recommended: (a) evaluate before and after training; (b) provide adequate time period for change (3-6 months); (c) collect information via survey or interview from all parties involved; (d) obtain 100% feedback from all parties involved; (e) when

possible, use a control group and a treatment group; and (f) consider the cost of the evaluation compared to the possible benefits (Kirkpatrick & Kirkpatrick, 2006).

The ultimate goal in Kirkpatrick's model is for the corporation to receive desired benefits or results (Level four: Results; Kirkpatrick & Kirkpatrick, 2006). This represents the phase for measuring the effectiveness of the program and its expected outcomes. Depending on the type corporation, the benefit may be monetary, humanitarian, service-oriented, and other. Although evaluating the results is desirable, it is often difficult to draw cause and effect relationship between training conducted and consequent results. The time gap between training and results may be lengthy, and multiple other factors may contribute to training program success besides the program in itself. However, the program developers and conductors must justify the positive impact the program has for its trainees. Otherwise, the program may be cut. The developers need to define the results in measureable terms, such as monetary benefits, increase in efficiency, improved morale, refined teamwork, and more satisfactory customer service, such as reduced number of complaints and more expressions of appreciation.

Application of Kirkpatrick's Framework to Published Evaluations

In this section, we discuss programs from three published papers: (a) reduction of HIV-contraction in the Latino community (two cases from Conner, 2004); (b) perceptions of teachers of a newly adopted supplemental reading program (Inman, Marlow, & Barron, 2004); and (c) evaluation of the South Carolina seniors farmers' market nutrition education program (Kunkel, Luccia, & Moore, 2003).

Reduction of HIV-Contraction in the Latino Community

Conner (2004) examines evaluations on two separate cases, each dealing with programs aiming to reduce the contraction of HIV in the Latino community. The purpose was to discuss the importance of culturally sensitive designs in evaluating programs. Conner frames the chapter with the intent of refining the concept of multicultural validity, which is "the accuracy, correctness, genuineness, or authenticity of understandings (and ultimately, evaluative judgments) across dimensions of cultural difference" (Conner & Kirkhart, 2003, p. 1). This is significant because

> Cultural issues and differences can be important factors in understanding which variables did and did not cause differences in programs (internal validity), which effects generalize over other settings and times (external validity), and what effects mean for higher-order constructs and implications (construct validity). (Conner, 2004, p. 52)

Although the purpose of the programs and the subjects of the larger Latino community were commonalties, the programs, and thus, the respective evaluations of each, are very different from each other.

The first case study looks at the *Tres Hombres sin Fronteras* (Three Men Without Borders) program, which was developed to educate Latino farmers about the dangers of unprotected sex with prostitutes and the risk of contracting HIV/AIDS. It consisted of 89 participants who were surveyed in both study groups. For the program, farmers worked with developers to create an eight-page *fotonovela*, telling the story of three famers who come into contact with prostitutes and the ramifications of having unprotected sexual intercourse. To augment this fotonovela, they developed a booklet with instructions on the proper way to use a condom. The goal of this program "was to test the effectiveness of the educational program in changing HIV-related knowledge, attitudes, and practices" (Conner, 2004, p. 54). The purpose of the evaluation was to see if the program was effective in conveying the dangers of unprotected

sex to Latino farmers in Mexico (measured by decreased rates of HIV contraction) and to determine whether to continue the program, and if so, how it could be improved.

The second case described Proyecto SOLAAR, a group aimed at educating urban-dwelling gay and bisexual Latino men. The group of men was especially vulnerable to HIV and other sexually transmitted infections because they are caught in the middle of two cultures with different norms and assumptions. The purpose of the Proyecto SOLAAR program was to educate these men not only about US cultural norms, but also about developing healthy behaviors and decision-making. The program was conducted as a weekend retreat, "during which a small group of men discuss issues and engage in some exercises and games that focus on topics that include relationships, dating, communication, self-concept, and HIV/AIDS" (Conner, 2004, p. 56). The program included facilitators helping participants to develop an individualized "dating plans" and "HIV risk reduction plans" (p. 56).

Supplemental Reading Program

In 2003 -2004, the state of Louisiana implemented EduSTRAND, a program designed to examine the perceptions of teachers of a newly adopted supplemental reading program in Louisiana (Eladrel Technologies, LLC, 2011). No Child Left Behind (NCLB) of 2002, a federal law that was supposed to reduce the reading achievement gap by 2014, was the impetus behind this program. The program incorporated 153 public schools between first and eighth grades and included 600 teachers whose goal was to analyze the effects of the reading program on student's performance. The sample size represented 10% of the Louisiana's schools. The mean average of teacher-student ratio was 14:1. At the time of the study, Louisiana had a total of 1,484 schools, with 124 school districts and 48,481 teachers (Eladrel Technologies, LLC, 2011). The ultimate goal of this program was to ensure that all students achieved the highest possible performance on the standard achievement measures. This study design utilized a mixture of Quasi-experimental and non-experimental methods. The method used to obtain data was past students' academic achievements, which provided a benchmark for comparison data and surveys which were mailed to teachers for their feedback responses.

Seniors Farmers' Market Nutrition Education Program

United States Department of Agriculture's (USDA) Commodity Credit Corporation administered the Seniors Farmers' Market Nutrition Program (SFMNP) in 2001, and the program evaluation was published two years later (Kunkel et al., 2003). An extension of a larger government program introduced by Massachusetts Department of Food and Agriculture in 1986, the initiative is a social and educational program targeting low-income seniors and local farmers. The SFMNP's purpose was to (a) provide locally grown fresh fruits, vegetables and herbs to impoverished seniors, (b) increase the consumption of domestic, agricultural products, and (c) assist in development of additional community-driven, agricultural enterprises such as Farmer's Markets, and roadside stands. A fourth purpose appears as to find evidence in support of or opposition to additional government funding for the program. At the time of registration, South Carolina Department of Social Services (SCDSS) distributed five 10-dollar vouchers for each of the 15,000 participants with a pamphlet containing nutrition information of available produce. The registration sites included churches, Farmers' Markets, Council on Aging, and Community and Senior Centers, among others. At the end of the program, a survey was mailed to a random sample comprised of 1,500 participating seniors with a 44% survey return rate, and 102 farmers with a 53% survey return rate (Kunkel et al., 2003), which were used for evaluation purposes.

Common Elements of Program Evaluation

Kirkpatrick's (1975) model evaluates reactions, learning, behavior, and results. The four programs evaluated by the three evaluators utilized Kirkpatrick's four-level systematic approach, however after the evaluation, we looked among the papers for other subcategories of commonalities. Twelve detailed commonalities important to program evaluation were found, most of them falling under Kirkpatrick's four levels. The following categories were observed: (a) define target group, (b) delineate expected outcomes, (c) operationalize success, (d) how the program was received, (e) unintended exclusion of target group members, (f) learning by target group, (g) valuable information not learned due to a design flaw, (h) behaviors changes of the target group, (i) gaps in program design, (j) intended results, (k) unintended results, and (l) suggestions for program improvement. In Table 1, we detail each of these throughout the three published program evaluations.

Program Evaluation Summary

Program evaluation design should be based on expected and desired results. For example, consumer-oriented approaches rely on understanding on consumers' perception on the product whereas judicial approaches investigate the pros and cons of the program. Moreover, accreditation approaches evaluate how the program would measure up to other similarly accredited programs, and utilization-focused methods concentrate on the way stakeholders will use the findings (Preskill & Russ-Eft, 2005). Although a variety of evaluation methods are available, the three authors utilized Kirkpatrick's (1975) four-level model to critique the programs for its widespread recognition as a comprehensive program evaluation model (Twitchell, Holton, & Trott, 2000).

The implications of the review of three program evaluations is that, in fact, common elements can be teased out of distinctively different program evaluations to understand the impact of the program for the participants and most importantly the effect on the organizational success to achieve its intended goals. Therefore, underlying similarities exists in program evaluations across the fields. Program success can be measured in infinite ways but the reality is that a program funded by a specific corporation is not successful unless it translates to measurable benefits for the corporation funding the program. Therefore, it is of utmost importance to the program evaluators to understand organizational goals and measure them effectively. Consequently, if the measures indicate that the program did not produce favorable results, which may be monetary, human service oriented, or other, the program evaluators might make suggestions, adjustments, and arguments for programmatic changes that would produce favorable future results.

Program evaluation relies on theory, but it is truly measured in practice. The evaluators must be committed to understanding the real-life, practical goals of the organizational and how to guide the organization to achieve its desired results. Theory in itself will not complete the job but the actual findings, recommendations, adjustments, and final results will define success of the program evaluation journey. In order for evaluators to arrive at valid conclusions, draw implications and make recommendations, it is critical that they pay attention to (and base their procedures upon) these types of validity. However, it is more than just knowing the correct methods of each approach of evaluation: "the evaluator must learn about and respond to the *context* [emphasis added] of the evaluation and its culturally related components, as well as to the *participants* in the evaluation and the cultural issues relevant to them" (Conner, 2004, p. 52). In particular, whenever a program evaluation is done, it is critical to pay attention to the characteristics of its participants. "To meaningfully assess and engage these culturally sensitive

programs, evaluators need to develop and implement evaluations sensitive to the cultural issues that characterize and are important to the populations, as well as to program participants and stakeholders" (Conner, 2004, p. 51).

As can be surmised from the cases discussed above, five factors must be part of a multiculturally-sensitive evaluation process: (a) involving participants in the evaluation study planning, (b) speaking the literal language of the participants, (c) speaking the figurative language of the participants, (d) working collaboratively with participants during implementation, and (e) sharing the benefits. The five factors are critical when evaluating a program through a multicultural lens. However, in the study of perceptions of teachers, the perspectives of gender, race, and socioeconomic variables were not discussed. Demographics of the teachers' years of experience and education level, school enrollment, class size, and the students' grade levels were included in the study. However, race, gender and socioeconomic status of teachers and students were omitted.

When writing up results on program evaluation, it is critical to provide the reader with enough details about the program, the considerations for evaluation, the methods used, and the results of the evaluation. Although the cases discussed in Conner (2004) provided a clear description of the purposes of the program, it was hard to gauge the adequacy of the evaluation plan because too little detail was provided. For example, there was very little justification and explanation for the one-month intervals for the illiterate Mexican farmers. The sessions were described in one paragraph, but given that so much of the procedure was relationally-driven, not much was said about the interactions between the facilitator and the various groups. Was the dynamic different among the groups? How did the impact learning? More detail would probably have given the reader a better idea of the methods used for the evaluation. The implications for stakeholders were given no attention. Indeed, it is not entirely clear who, beyond the farmers themselves, are the stakeholders. The study involving teacher's perceptions did a better job in this aspect—here, the evaluator was wise in involving his stakeholders, the teachers, in the direct development and implementation of the educational tool.

References

Bassi, L., & McMurrer, D. (2007). Maximizing your return on people. *Harvard Business Review, 85*(3), 115-123.

Conner, R. F. (2004). Developing and implementing culturally competent evaluation: A discussion of multicultural validity in two HIV prevention programs for Latinos. *New Directions for Evaluation, 2004*(102), 51-65.

Eladrel Technologies, LLC (2011). *Louisiana public school statistics: Demographics*. Retrieved from http://publicschoolsk12.com/all-schools/la/

Inman, D., Marlow, L., & Barron, B. (2004). Evaluation of a standards-based supplemental program in reading. *Reading Improvement, 41*(3), 179-187.

Kirkpatrick, D. L. (1975). *Evaluating training programs*. Boston, MA: McGraw-Hill Education.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels* (3rd ed.). San Francisco, CA: Berrett-Koehler.

Kunkel, M. E., Luccia, B., & Moore, A. C. (2003). Evaluation of the South Carolina seniors farmers' market nutrition education program. *Journal of the American Dietetic Association, 103*(7), 880-883.

MacDonald, G., Starr, G., Schooley, M., Yee, S. L., Klimowski, K., & Turner, K. (2001). *Introduction to program evaluation for comprehensive tobacco control programs.* Washington, DC: Department of Health and Human Services: Centers for Disease

Control and Prevention. Retrieved from http://www.cdc.gov/tobacco/tobacco_control_programs/surveillance_evaluation/evaluation_manual/pdfs/evaluation.pdf.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Preskill, H., & Russ-Eft, D. (2004). *Building evaluation capacity: 72 activities for teaching and training*. Los Angeles, CA: Sage.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Los Angeles, CA: Sage.

Twitchell, S., Holton, E., & Trott, J. (2000). Technical training evaluation practices in the United States. *Performance Improvement Quarterly, 13*(3), 84-109.

Table 1

Aspects of Program Evaluation of the Three Published Studies

	HIV Reduction		Teacher Perception	Farmers' Market
Target group	Illiterate Mexican Farmers	Latino Gay Men	1 - 8 Graders in Louisiana Public School	Low-Income Elderly and Farmers
Expected outcomes	Reduce incidence of HIV/AIDS		Increase in reading levels	Increase nutritional intake of elderly; revenue of farmers
Operationalize success	Decrease HIV/AIDS among Latinos		1 - 8 grade readers reach expected reading levels	Healthier nutritional habits of elderly; increase in farmers' income
How was the program was received (Level 1)	Too little detail included	Difficulties getting participants	Scores increased at all levels except in sixth grade.	Seniors and farmers expressed appreciation and satisfaction
Individuals excluded by program design	Farmers who could not make it to group meetings	No information available	10% of the targeted population in Louisiana public schools were selected	Seniors without transportation to the farmers' markets.

Learning occurred (Level 2)	Correct use of condoms	Views on cultural norms, dating, safe sex	Ways to increase reading scores among grades 1-8	Nutritional information, prices and quality of produce; seniors inclined to shop at farmers' markets
Valuable information not learned	Too little detail included	No information available	Demographic variables such as socioeconomic, gender and race were not part of the study.	Seniors did not learn to try *new* produce
Behavior change (Level 3)	Too little detail included	No information available	Responses only surveyed the teachers and not the students	Annually, 89% of seniors consumed more produce and intended to increase visits to farmers' markets
Gaps in program design	Too little detail included	Too little detail included	Need to design lessons to address special needs, especially with 6th graders	Exposure to new fruits and vegetables did not entice seniors to buy them; cooking lessons should be explored
Intended results (Level 4 Results)	Changes in HIV/AIDS-related knowledge	No information available	Increased reading levels were met with all grades from 1-8 except with 6th graders	98% of seniors used at least one voucher; 86% used all vouchers; 89% would continue shopping at farmers' markets; 100% of farmers were willing to participate again; farmers cashed 86% of the vouchers for a total of $643,300

Unintended results	No information available	Difficulty getting participants lead to a year-long recruitment campaign	No improvement in sixth grader's reading levels; possibly a Hawthorne Effect	A 10-dollar voucher had to be used in one stand whether or not the total amounted to 10 dollars
Suggestions for program improvement	Meet the farmers at their location for easy program access	Consider privacy aspects of the program	Inclusion of other demographic factors such as gender, race, socioeconomic status	Offer (a) transportation to seniors (b) smaller voucher denominations, and (c) cooking instructions