

THEY'RE HAPPY, BUT DID THEY MAKE A DIFFERENCE? APPLYING KIRKPATRICK'S FRAMEWORK TO THE EVALUATION OF A NATIONAL LEADERSHIP PROGRAM

Scott McLean
Gwenna Moss
University of Saskatchewan Extension Division
Saskatoon, Saskatchewan

Abstract: This article examines the Kirkpatrick evaluation framework through a case study of a national leadership development program. The authors introduce the program and the Kirkpatrick framework, and then describe the research processes and instruments through which the framework was applied to evaluate the pilot cohort of the program. The article concludes with several frank and practical insights about using the Kirkpatrick framework to evaluate non-credit educational programs. In areas such as leadership development education, Kirkpatrick offers an appealing framework for organizing an evaluation process. The framework enabled a productive formative evaluation process, and the demonstration of participant satisfaction and learning with the program was sufficient to facilitate the approval of funding for a second cohort. However, despite the investment of considerable resources, the evaluation of this program was not able to conclusively demonstrate that behaviour changes and resulting impacts on organizations and communities took place as a result of the program.

Résumé: Cet article examine le cadre d'évaluation de Kirkpatrick, à travers l'étude d'un programme national de développement de leadership. Les auteurs introduisent le programme et le cadre de Kirkpatrick, et ensuite décrit les processus de recherche et les instruments avec lesquels le cohorte pilote du programme a été évalué. L'article conclut avec des aperçus franc et pratique sur l'application du cadre de Kirkpatrick dans l'évaluation des programmes éducatifs sans crédits. Dans les domaines comme le développement de leadership, Kirkpatrick offre un cadre intéressant pour organiser une évaluation formative productive. Le

Corresponding author: Dr. Scott McLean, Extension Division, University of Saskatchewan, 117 Science Place, Saskatoon, SK S7N 5C8; <mcleans@duke.usask.ca>

cadre a donné la possibilité de démontrer suffisamment de satisfaction et d'apprentissage dans le programme pour mobiliser les fonds pour un second cohorte. Cependant, malgré l'investissement de ressources importantes, l'évaluation du programme n'a pas pu démontrer de façon concluante que ni les changements de comportement ni les impacts sur les organisations et les communautés étaient le résultat du programme.

The purpose of this article is to examine the Kirkpatrick framework as an evaluation tool through its application to a concrete case study — the evaluation of a national leadership development program. Non-credit educational programs, particularly programs that seek to develop soft skills such as leadership, present challenges to evaluators. Although it is common, and relatively simple, to determine participant satisfaction and perceived learning, it is more difficult, in a non-credit setting, to measure learning, behaviour change, and the eventual impact of such behaviour change on organizations and communities. The decision to employ the Kirkpatrick framework in this major non-credit educational initiative presented us with an opportunity to examine the utility of using this framework in programs with complex objectives such as the development of effective leaders.

THE CANADIAN AGRICULTURE LIFETIME LEADERSHIP PROGRAM

The Canadian Agriculture Lifetime Leadership (CALL) program was established in 1997 by the Canadian Farm Business Management Council (CFBMC) and the University of Saskatchewan Extension Division. Patterned after similar programs in the U.S.A., the mission of the 18-month program was to develop effective leaders for the Canadian agri-food industry. Within this mission, four program objectives were defined to express the means through which CALL would have an impact on the agri-food industry in Canada: (a) develop a pool of excellent leaders to serve national and regional organizations of importance to Canadian agriculture; (b) create a network of leaders able to link diverse sectors and regions within the Canadian agri-food industry; (c) create a network of leaders who will provide vision and leadership for the Canadian agri-food industry, and advocate for the Canadian agri-food industry both within Canada and internationally; and (d) create a network of leaders who will, through their impact as educators, mentors, and role models, improve production and farm business management practices across Canada. These program objectives presupposed a set of participant

objectives relating more concretely to the impact CALL would have on its learners. The CALL program set out to identify outstanding men and women with proven leadership potential, and to provide a forum for them to broaden their horizons of knowledge, practice the arts of leadership, and build effective networks. The participant objectives of learning, skill-building, and networking were used to structure dozens of specific learning objectives and activities.

Thirty participants took part in the first program cycle (September 1997 through April 1999). The CALL learning objectives were pursued through two main delivery strategies: a computer-mediated conference and six in-person seminars. The computer conference created a structured setting for dialogue on readings and issues related to leadership and agriculture. It enabled small groups of participants to work on assigned projects; facilitated private communication between CALL participants through private e-mail messages and chats; provided a virtual café for informal interactions; and created a forum for orientation and debriefing activities related to the six in-person seminars. The seminars, which were attended in person by all participants, were held in locations in Canada (Saskatoon, Vancouver, Guelph, Ottawa, Montreal, Fredericton, and Kananaskis), the U.S.A. (Albany, New York; Washington, D.C.; and McAllen, Texas), and Mexico (Saltillo, Monterrey, and Guadalajara). They ranged in length from four days to two weeks, and each addressed an important theme in agriculture (ranging from leadership to politics to international trade).

The 30 CALL participants, 16 men and 14 women, represented all ten Canadian provinces and ranged in age from just under 30 to just over 50. About two-thirds of the participants were involved in agricultural production, with the remaining participants employed in various non-farm agri-businesses and non-governmental organizations. The participants had a wealth of experience with agricultural leadership in farms, agri-businesses, non-governmental organizations, and rural communities. The size of the cohort was intentionally kept low (over 140 candidates applied to take part) because of our commitment to intensive teaching and learning techniques, the importance of peer networking to the program, and the logistical and cost constraints of organizing residential and study travel seminars for large groups. The cost of the program was about \$700,000, or just over \$23,000 per participant. Tuition accounted for \$5,000 per participant, and the CFBMC funded most other costs. Of the 30 participants, 28 graduated from the program. A second cohort of 27 participants began the program in September 2000.

THE KIRKPATRICK FRAMEWORK: A “STANDARD” MODEL IN TRAINING

Kirkpatrick first published his four-level evaluation framework in four issues of the *Journal of the American Society of Training Directors* (now called *Training and Development*) (Kirkpatrick, 1959a, 1959b, 1960a, 1960b). Thirty-five years later, the framework was published in a book, which didn't substantively change the 1959 framework (Kirkpatrick, 1994; a second edition was issued in 1998). Kirkpatrick's model encompasses four levels of evaluation: reaction, learning, behaviour, and results.

Level 1: Reaction. Frequently referred to as “happy face evaluation,” this level measures participant reaction to and satisfaction with the program and the learning environment.

Level 2: Learning. Changes in knowledge, skills, and/or attitudes constitute learning in the Kirkpatrick model. Excluded from this level of evaluation is the application of the learning on-the-job.

Level 3: Behaviour. This level determines whether changes in behaviour have occurred as a result of the program. Kirkpatrick stresses the importance of having information on levels 1 and 2 in order to interpret the results of level 3 evaluation. Specifically, if no behaviour change occurs, it is useful to determine whether this is due to participant dissatisfaction with the program (level 1) or a failure to accomplish the learning objectives (level 2), or whether the lack of change in behaviour is due to factors beyond the scope of the program (e.g., a lack of desire, opportunity, support, or rewards for changing behaviour).

Level 4: Results. Level 4 looks at the final results that occurred because the participants attended the program. Results can be thought of as “the bottom line”: the impact of the program.

The Kirkpatrick model quickly became a standard (perhaps *the* standard) model for evaluating training, and, to a great extent, remains so today. Newstrom (1978) described it as “a classic model” (p. 22), as did Bernthal (1995). Kaufman, Keller, and Watkins (1996) stated: “For more than 30 years the most popular, and probably the primary, model for evaluation among human-resource development professionals has been the four-level training evaluation approach proposed by Kirkpatrick” (p. 9). The widespread use of the

Kirkpatrick model in the training field means that trainers share a “common language” and a common tool that facilitates the comparison of results (Carliner, 1997, p. 14).

Kirkpatrick brings a practical, rather than a theoretical, perspective. He says: “People have asked me why the model is widely used. My answer: It’s simple and practical. Many trainers aren’t much interested in a scholarly, complex approach. They want something they can understand and use” (Kirkpatrick, 1996b, p. 55).

Numerous articles, mostly in the training literature, have used, built on, incorporated, or referred to the Kirkpatrick model. A series of articles in *The ASTD Training and Development Handbook: A Guide to Human Resource Development* (Craig, 1996) illustrate the continuing use of the model in fields that include sales and marketing training (Hahne & Schultze, 1996); computer skills training (Warfel, 1996); technical skills training (Robbins, Doyle, Orandi, & Prokop, 1996); job training (Nolan, 1996); and human performance technology (Rosenberg, 1996). Other examples include Moseley and Larson’s (1994) more general application of the model for evaluating workshops and conferences, and Pine and Tingley’s (1993) study of the application of all four levels of the Kirkpatrick model to soft-skills training, in the context of a two-day course in team building with intact work groups. Johncox (2000) reports an evaluability assessment, using Kirkpatrick’s model, of staff training in special care units for persons with dementia.

While the training literature consistently refers to the Kirkpatrick framework as a model, it appears less frequently in what might be called the more generic program evaluation literature. As Michalski and Cousins (2001) note, “the training evaluation and program evaluation literatures have developed largely in parallel, with few points of intersection” (p. 38), and “Well established program evaluation theory and practice remains largely outside of the mainstream in training evaluation practice” (p. 50).

Stufflebeam (2001) does not reference Kirkpatrick in his description and assessment of 22 approaches to, or models of, program evaluation. This is not inappropriate. Kirkpatrick’s framework could be used as part of almost all the models described by Stufflebeam: that is, the evaluator(s) and/or stakeholders could choose to evaluate reaction, learning, behaviour change, and results in any of a wide range of evaluation approaches. Lee and Pershing (2000) reviewed and

compared six evaluation perspectives for corporate training programs, including Kirkpatrick's four-level approach. None were referenced by Stufflebeam (2001). Although it is interesting to speculate on the reasons for the separation of the two bodies of literature, such a discussion is beyond the scope of this article.

Critiques of the Model

But is the Kirkpatrick framework a model?

One of the critiques of Kirkpatrick's framework is that it is frequently referred to as a model, and yet does not meet the usual criteria for a model. Holton (1996) argued that the model is really a taxonomy, and proposed a new model that met more of the criteria for a model. In response to the Holton critique, Kirkpatrick (1996c) said: "Personally, I don't care whether my work is called a model or a taxonomy as long as it helps to clarify the meaning of evaluation in simple terms and offers guidelines and suggestions on how to accomplish an evaluation" (pp. 23–24). Kirkpatrick (1996b) also noted: "In the November 1959 article, I used the term 'four steps.' But someone, I don't know who, referred to the steps as 'levels.' The next thing I knew, articles and books were referring to the four levels as the Kirkpatrick model" (p. 55). Alliger and Janak (1989) are likely close to the mark in saying: "Kirkpatrick's model may never have been meant to be more than a first, global heuristic for training evaluation. As such it has done well" (p. 339). They add that "the power of Kirkpatrick's model is its simplicity and its ability to help people think about training evaluation criteria. In other words, it provides a vocabulary and rough taxonomy for criteria" (1989, p. 331).

Difficulties with evaluating levels 3 and 4.

Evaluators, including Kirkpatrick, generally acknowledge that evaluation at levels 3 and 4 is more difficult, and tends to be avoided (Kaufman, Keller & Watkins, 1996; Kirkpatrick, 1994, 1996a; Shelton & Alliger, 1993; Twitchell, Holton, & Trott, 2000). Moss and Kehoe (2000), in surveys of university continuing education programmers in Western Canada, found that evaluation beyond level 1 in non-credit programming is relatively uncommon. Geber (1995) refers to the pressure to evaluate at levels 3 and 4 as

the trainerly equivalent of flossing your teeth. You *know* you're supposed to do it, you *know* it's good for you, you

know there might be dire consequences eventually if you don't, but let's be honest: How many people floss each day unless their dentists have warned them that their gums are flabby and their teeth are starting to wobble? (pp. 27–28)

Much of the challenge of evaluating at levels 3 and 4 is associated with the difficulty of attributing any measurable changes to the program or the intervention. This is a challenge that is only too well known by program evaluators. The distinction between correlation and causation is germane here: it may be sufficient to show that a relationship exists. Kirkpatrick himself is clear on the distinction between evidence of change versus proof of what caused the change, stating that often “we have to be satisfied with evidence instead of proof” (1994, p. 68).

Revisions of the Model

The model has been extended and revised by a number of individuals. Bernthal (1995) recommended extending evaluation beyond outcomes to collect information on process and other factors that can impinge on outcomes. Phillips (1996) added a fifth level, return on investment (ROI), which asks: Did the monetary value of the results exceed the cost for the program? He also extended level 1 to include not just participants' reactions to the program, but also their plans for the material. Partly to address the alleged difficulties of evaluating levels 3 and 4, Kaufman and Keller (1994) and Kaufman et al. (1996) proposed three modifications to Kirkpatrick's framework, as follows: (a) include a fifth level aimed at determining societal impact; (b) expand level 1 to include the value and worth of resources and methods; and (c) incorporate organizational interventions other than the training per se into the evaluation design. Watkins, Leigh, Foshay, and Kaufman (1998) extended these proposals to link Kirkpatrick with Kaufman's Organizational Elements Model, and hypothesized that the “Kirkpatrick Plus” framework will provide the “missing linkages” to help address infrequent use of levels 3 and 4 evaluation (p. 92). Stokking (1996) built on Kaufman and Keller's (1994) extension, in part by using Stufflebeam's CIPP model (and others) to clarify some components.

EVALUATING CALL

Agriculture and Agri-Food Canada, through the CFBMC, provided over three-quarters of the operating budget for CALL. As a pilot

project, CALL was provided with separate funding for an evaluation to be conducted by professional evaluators external to the CFBMC and to the University of Saskatchewan. The CFBMC provided a performance framework with which the evaluation design was to be aligned. It consisted of four dimensions: output, reach, direct impacts, and long-term impacts. Although the CALL program was not a training program, or even a human resource development program in the traditional sense, the Kirkpatrick approach provided a framework that met the program sponsor's evaluation requirements and promised to be useful in conceptualizing and organizing the evaluation.

A request for proposals was issued in April 1997, with the following description of the guiding questions that would structure the evaluation:

- What is the level of satisfaction of CALL participants with regard to the seminars and the computer conference?
- Have participants' leadership skills, knowledge, and personal networks improved through the program?
- Have participants' leadership commitments and behaviours changed through the program?
- Have participants' leadership activities had a beneficial impact on farm businesses, agri-food businesses, agricultural organizations, and rural communities?
- Have participants' activities led to a transfer of skill, knowledge, and networks to people in the participants' businesses, organizations, and communities?

By design, each of these questions corresponded to one of Kirkpatrick's four levels of evaluation. The fifth question was another means of assessing the indirect impact of participants' activities.

Nine proposals were received, and the successful consulting firm, in consultation with the program director, developed the evaluation design summarized in Table 1. The subsequent sections of this article describe the implementation and outcomes of this design.

Seminar Evaluation Instruments

The seminar evaluation instruments were primarily aimed at gauging the satisfaction of participants with the leadership development seminars in which they were taking part. However, given the very

explicit presentation of learning objectives in the categories of knowledge, skill development, and network building, participant satisfaction was intended to be related to their self-assessment of the extent to which they had learned from the seminar and its constituent sessions. Participants were not examined regarding the content of the seminars. Data from this phase of the evaluation served both summative and formative purposes.

Each of the six face-to-face seminars was evaluated in a consistent manner. Each day, participants evaluated each of the sessions on a scale from one (“poor”) to ten (“outstanding”), and had the opportunity to make brief comments about each of the sessions. Upon completion of the seminars, participants evaluated the seminar as a whole. For the two seminars that lasted over a two-week period, and that had sessions hosted in four or five locations, additional evaluation instruments were completed every four or five days (basically, whenever the seminar changed cities). Four questions were consistently asked in these end-of-seminar evaluations. First, participants were asked to identify the greatest strength(s) of the seminar, taking into consideration its overall structure and content. Second, participants were asked to rate the accomplishment of each of the learning objectives of the seminar on a scale from one (“poor”) to ten (“outstanding”). Learning objectives were always structured according to the three categories of knowledge, skills, and networks, and reflected the content of each seminar’s substantive focus. Third, participants were asked to rate, on a scale from one (“poor”) to ten (“outstanding”), four aspects of the seminar: preparation via computer conferencing; overall content; organization and logistics; and accommodation, meals, and meeting spaces. Fourth, participants were asked to provide any comments they may have had regarding the seminar, with a view to improving future seminars.

Table 1
Evaluation Framework for the CALL Program

Evaluation Component	Level 1	Level 2	Level 3	Level 4
Seminar instruments	✓	✓		
Questionnaire - mid-point	✓	✓	✓	✓
Questionnaire - end of program	✓	✓	✓	✓
Questionnaire - 2 years post program		✓	✓	✓
Leadership Practices Inventory			✓	
Observers' Questionnaire			✓	✓

Extensive Questionnaires Completed by Participants at the Mid-point and End of the Program, and Two Years Following Completion of the Program

Three extensive surveys were designed to move beyond the level of satisfaction to more fully explore learning, behaviour change, and impact. The three surveys shared a common structure. Each was composed of discrete sections, and each section began with a series of closed-ended questions that asked participants to rate, on a scale from one (“strongly disagree”) to ten (“strongly agree”), their level of agreement with statements about their experience of the program. Each section concluded with between one and four open-ended questions. The sections on the mid-term and final evaluation surveys were very similar. Each included sections on participant satisfaction, knowledge, skills, and networks; changing leadership practices; participant impact; diffusion of CALL benefits; and computer conferencing. In addition, the final evaluation survey asked participants to comment on their anticipated leadership practices in future years.

The types of behaviour change and impact promoted by the CALL program were complex and required time to be implemented and meaningfully assessed. Therefore a follow-up survey, completed two years after the program concluded, was designed to assess the longer-term outcomes of the program. It asked participants questions about their application of what they had learned through the program; their ongoing contact with other CALL graduates; the leadership roles in which they were currently engaged; changes to their leadership practices attributed to their participation in CALL; the direct and indirect impact of their leadership activities on others; and their self-assessment of what impact their participation in the program has had on them and their leadership activities.

Although these surveys relied on participant self-reports, we incorporated open-ended questions asking participants to provide concrete examples of learning, behaviour change, and impact. It was hoped that this approach, which required the participants to think more explicitly about and to articulate the changes that had occurred, would enhance the credibility and validity of the data.

Tables 2 and 3 present selected questions from the three major surveys to illustrate the use of each of the four levels of Kirkpatrick’s framework.

Table 2
The Major CALL Evaluation Surveys (I):
Examples of Open-Ended Questions and the Four Levels of Evaluation

Level	Mid-Term and/or Final	Follow-up
One	<p>Please identify two aspects of the CALL program about which you are most satisfied, and tell us why you are satisfied.</p> <p>Please identify two aspects of the CALL program which you would most like to see improved and tell us how.</p>	<p>Do you feel that the benefits of participating in the CALL program exceeded the cost (participant cost, sponsor cost, time)?</p>
Two	<p>Please tell us about two things you've learned through the CALL program that have improved your knowledge as a leader in agriculture. (Similar questions asked regarding skills and relationships.)</p> <p>The CALL program's mission is to develop effective leaders for the Canadian agriculture industry. Please use the space below to comment on whether (or not) and how the CALL program has helped you to become a more effective leader. Please comment on your learning and development of new knowledge, skills, networks, or any other aspect important to leadership.</p>	<p>Please provide two or three specific examples of how you have applied what you learned through CALL to your work as a leader.^a</p>
Three	<p>Please describe two leadership practices which you have improved through your involvement with the CALL program. In each case, give an example in which you have actually applied this practice.</p> <p>Please use the space provided below to explain whether (or not) your leadership practices and behaviours have changed as a result of taking part in CALL. Have you taken on new leadership responsibilities? Have you changed the way you lead in your existing (pre-CALL) roles?</p>	<p>Please provide two or three specific examples of how your leadership practices have changed since your participation in CALL. If your leadership practices have not changed, please say so.</p> <p>Please use the space below to explain, in your own terms, what the CALL program has meant for you as a leader, and as a person. Has your participation in the program had a lasting impact on you? Have you made changes in your leadership activities, or in your life more generally, that relate to your participation in CALL?</p>
Four	<p>Please give two specific examples of how your leadership practices have had an impact on an agriculture organization, agri-business, or local community since the start of the CALL program. Be sure to give enough detail so we can understand the positive impact you have had.</p>	<p>Please use the space below to describe (1) what impact (at any level: farms, other businesses, organizations, communities, families, etc.) you think your leadership activities have had since graduating from CALL and (2) how your activities have (or have not) led to a transfer of skill, knowledge, or networks to other people in your businesses, organizations, and communities.</p>

^a This question measures both levels 2 and 3: it asks respondents to identify and articulate what was learned (level 2), and to describe how that learning was applied (level 3).

Table 3
The Major CALL Evaluation Surveys (II):
Examples of Closed-Ended Questions and the Four Levels of Evaluation

Level	Mid-Term and/or Final	Follow-up
One	<ul style="list-style-type: none"> • I am generally satisfied with the CALL program. • I am generally satisfied with the CALL computer conference. • The CALL program is worth the \$2,500 tuition fee. 	N/A
Two	<p>Through the CALL program ...</p> <ul style="list-style-type: none"> • I have learned about different approaches to leadership. • I have a better understanding of the diversity of Canadian agriculture. • I have become a more effective public speaker. • I have become more skilled at analyzing issues. • I have become more skilled at working with groups. • I have broadened my access to a network of leaders and resource people from across Canada's agri-food industry. 	<p>As a result of my participation in CALL ...</p> <ul style="list-style-type: none"> • I improved my understanding of leadership. • I enhanced my knowledge of issues facing Canadian agriculture. • I developed useful leadership skills. <p>Over the past year ...</p> <ul style="list-style-type: none"> • I have been in touch with approximately _____ CALL graduates for personal or social purposes. • I have been in touch with approximately _____ CALL graduates for business or professional purposes.
Three	<p>As a result of participating in the CALL program ...</p> <ul style="list-style-type: none"> • I have become a more effective leader. • I am more conscious of my leadership practices. • I have become involved in a larger number of leadership roles. • I am better able to "inspire a shared vision" in my leadership roles. • I am a more effective ambassador for Canadian agriculture. • I am more confident in my ability to serve in leadership roles with provincial agriculture organizations. 	<p>Since the end of my CALL program ...</p> <ul style="list-style-type: none"> • I have been actively involved in leadership activities in my local community. • I have been actively involved in leadership activities with regional or provincial organizations of importance to agriculture. • I have been actively involved in leadership activities with national or international organizations or importance to agriculture. <p>I believe that my participation in CALL ...</p> <ul style="list-style-type: none"> • Helped me to become a more effective leader for the agriculture organizations that I serve.
Four	<p>Since the beginning of CALL ...</p> <ul style="list-style-type: none"> • My leadership practices have had a positive impact on other farm or agri-business operations. • I regularly talk to people about agriculture leadership issues. • I actively assist the people I work with to become more effective leaders. • I have used the networks built through the CALL program to help someone else find a service or resource they need. 	<ul style="list-style-type: none"> • My leadership activities have benefited others in the Canadian agriculture industry. • I am a mentor for one or more other leaders in the agriculture industry or my local community. • I actively seek out opportunities to share what I have learned through CALL with other leaders in the agriculture industry or my local community.

The three surveys conducted as part of the CALL evaluation process were clearly designed to elicit reflection from participants about their learning, behaviour change, and impact as a result of participating in the program. However, they were limited to self-assessment, and did not include any objective assessment or third-party observation.

Leadership Practices Inventory

As one way to move beyond purely self-reporting assessments of learning, behaviour change, and impact, we used a standardized leadership development instrument, The *Leadership Practices Inventory* (LPI), developed by Kouzes and Posner (1997) as a companion to their 1995 textbook, *The Leadership Challenge*. Because this text was required reading for CALL participants, the LPI was selected as a means of encouraging participants to reflect about their leadership practices. It also provided an indicator of behaviour change among participants over the course of the program.

The LPI-Self asks respondents to rate (on a scale from one to ten) the extent to which they typically engage in thirty different behaviours. The LPI-Observer asks those who have an opportunity to observe the individual being rated to rate that person on the same scale for the same thirty behaviours. The thirty behaviours are categorized into the following five “leadership challenges”: challenging the process, inspiring a shared vision, enabling others to act, modeling the way, and encouraging the heart. The six questions asked within each of these five categories are then averaged to come up with an individualized LPI profile, which can be compared against standardized results from thousands of previous leaders who have completed the LPI process.

We asked each participant to complete the LPI-Self and to have ten people, who were familiar with the participant’s leadership practices, complete the LPI-Observer on their behalf. In October 1997, all 30 participants and a total of 235 observers completed the LPI instruments. In February 1999, 17 of 28 participants and a total of 184 observers completed the instruments.

A Survey Completed by Selected Observers at the End of the Program

Self-reported data were also supplemented by a Peer Evaluation Survey designed to assess behaviour change and impact among par-

ticipants. A total of 40 respondents, who were in positions to observe the participants' activities, completed the survey. The closed-ended questions asked respondents to rate (on a ten-point scale) their level of agreement with 15 statements about the participant's leadership behaviours and impacts over the period of the program. The three open-ended questions were focused on determining the extent to which the respondents had observed evidence that CALL participants (a) had become more effective leaders and (b) had a beneficial impact on other people, organizations, or communities. Table 4 identifies the closed-ended questions.

The respondents to the Peer Evaluation Survey were not necessarily the same individuals who completed the LPI-Observer instrument. Data collected from peers and observers were used to corroborate the participants' claims that they had changed their leadership behaviours as a result of participating in the program, and that such behaviours had an impact on organizations and communities. Apart from the LPI instruments, such triangulation was ac-

Table 4
Peer Evaluation Survey: Closed-Ended Questions

Over the past two years . . .

- The participant has developed his or her leadership skills.
 - The participant has regularly talked to people about agriculture leadership issues.
 - The participant has actively assisted the people he or she works with to become more effective leaders.
 - The participant has served as a mentor for one or more other leaders in the agriculture industry
 - The participant has served as a mentor for one or more other leaders in their local community.

 - The participant has used the networks built through the CALL program to help someone else find a service or resource they need.
 - The participant has broadened their awareness of national issues of importance to agriculture in Canada.
 - The participant has broadened their awareness of international issues of importance to agriculture in Canada.
 - The participant's leadership practices have had a positive impact on his or her own farm or agri-business operation.
 - The participant's leadership practices have had a positive impact on other farm or agri-business operations.

 - The participant's leadership practices have had a positive impact on the agriculture organizations which he or she serves.
 - The participant has made positive contributions to their local community.
 - The participant has made positive contributions to the agriculture industry in Canada.
 - The participant has become a more effective public speaker.
 - The participant has become involved in a larger number of leadership roles
-

completed only at a composite level. Individual participants' claims were not compared to the claims made about them on the Peer Evaluation Survey. Rather, the evidence gathered by the peers was used to provide some modest external validation of participants' self-reported behaviour change and impact. The decision to not link, at an individual level, the analysis of data collected from participants and their peers was made due to the limited data analysis resources at our disposal, and the consistency of the overall evidence gathered from the two groups. Given the overwhelmingly positive response to the program from both participants and their peers, we limited our comparison of the two sources of data to the composite level.

REFLECTIONS ON THE KIRKPATRICK MODEL IN ACTION

The Kirkpatrick model provided a very useful organizing framework for the evaluation of the CALL program. Because the framework makes intuitive sense to non-expert audiences, the participants and the funding agency representatives understood the rationale for exploring these four levels. This facilitated communication with stakeholders (funders and participants) before, during, and after the evaluation activities. For the evaluators and the program director, the framework proved useful in conceptualizing and focusing the data collection, and in organizing and presenting the results.

As a formative tool, the evaluation process led to a number of significant changes to the program during its 18 months of activities. In comparison to the earlier seminars in the program, the later seminars involved fewer activities, a structured, small-group debriefing process, and relatively more participatory and group-building sessions. In the second year of the program, an opportunity was presented for each participant to undertake a structured and supported self-evaluation of their leadership skills, and develop an individualized skill development plan based on the outcome of that evaluation. The computer conference was restructured at the mid-point of the program. As a summative tool, the evaluation process assembled adequate evidence of the quality and value of the program for the funding agency to provide money for a second cohort. To this extent, the Kirkpatrick model worked as a practical means of organizing the evaluation of CALL.

What can we conclude about the CALL program based on the evaluation activities that were organized according to Kirkpatrick's framework? With regard to satisfaction, we assembled a tremendous

amount of data to document that the great majority of participants were satisfied with their experience of the program. With very few exceptions, participants expressed high levels of satisfaction with the overall program. Specific activities within the program were not always highly rated, and this led to ongoing adaptation of the program to better suit its participants. Given the time lag between our educational interventions and meaningful behaviour change or impact, Kirkpatrick's first two levels drove the formative evaluation agenda. In a like manner, the decision to fund and deliver a second cohort of the program was also driven largely by evidence about participant satisfaction. It was only with the follow-up survey two years following the end of the program that we began to gather good quality data about behaviour change and impact, while decisions to run a second cohort needed to be made soon after the completion of the program.

With regard to learning, we can say that a large majority of participants believe they developed their knowledge, skills, and networks through participating in the program, and that most were able to provide specific examples of their learning and its application to their leadership practices. Our evaluation of the learning accomplished in the program was largely restricted to inviting participants to self-report their perceived learning. Because CALL was a non-credit certificate program and its participants varied in levels of formal educational attainment from those who had not completed high school to those with post-graduate degrees, no examinations were required. Two activities not initially conceptualized as evaluation practices served to corroborate participants' self-reported learning over the course of the program. First, each participant worked in a group to complete an Issues Analysis Project that resulted in seven reports being written about key issues facing the agriculture industry in Canada. Although these reports were not formally assessed or graded, they do document a substantial amount of knowledge generation through the program. Second, a doctoral thesis is currently being written on the presence of critical and creative thinking exhibited in the CALL computer conference. The preparation of this thesis has involved the documentation of learning processes among a sub-group of CALL participants. In summary, the CALL participants claim to have learned through the program, and there is some external evidence to support that claim.

With regard to behaviour change, we are likewise largely dependent upon the subjective claims of the participants themselves. Many

CALL participants claim to have changed their leadership practices as a result of taking part in CALL, and most are able to provide examples of specific changes they have made and of concrete ways in which they have applied their learning from CALL to agricultural leadership work. These claims are supported by the modest survey of peers that was conducted. In contrast, the Leadership Practices Inventory turned out to be a pedagogically effective tool for assessing the leadership practices of the CALL participants and for structuring personal development activities by those participants, but it was not effective as a tool for measuring behaviour change. End-of-program measures did not meaningfully differ from the pre-tests. The fact that the LPI did not capture significant change in leadership behaviours of program participants may be interpreted in two ways. Either the behaviours of the participants did not change, or the standardized LPI instrument was not sensitive enough to capture behavioural changes. The very high scores achieved on the LPI-Observer instrument both at the beginning and end of the program suggest that the CALL participants were so highly rated by their observers at the start of the program that there was little room for improvement that could be measured by the instrument in a post-test. Participant scores on the LPI-Self increased from pre- to post-test, but not enough to be statistically significant. Given the complexity of the behaviour change promoted by this leadership development program, attributing such change to the experiences of the program would be very complex methodologically. The isolation of what took place within the program from the maturation effect and everything else that took place in the lives of its participants over an 18-month period would be very difficult to achieve.

The methodological challenges of measuring behaviour change in this complex leadership development program are reproduced when trying to substantiate the impact of such changed behaviours on organizations and communities. Participants were selected to the program because they demonstrated excellent potential as leaders for agricultural businesses and organizations. It should not be surprising that, after participating in the program, they have functioned as capable leaders for such businesses and organizations. What cannot be said with certainty, however, is the extent to which they would have functioned differently without participating in the program. Many participants claim that the CALL program had an impact on their leadership practices, and as a result on businesses and organizations of importance to Canadian agriculture, and are able to articulate specific examples, but this claim cannot be “proven” in an

absolute sense. Given the tremendous difficulty of creating control-group conditions, the attribution of behaviour change and impact to the specific interventions of the program is very difficult. As noted earlier, Kirkpatrick discusses the difficulty, even impossibility, of obtaining “proof” that a program caused specific changes and results. His approach is to assemble as much “evidence” of change as one can, including evidence from lower levels of the model. Part of the evidence that behavioural change and impact occurred as a result of a program is based on data on lower levels of satisfaction and learning. For a similar argument on the need to collect data on learning and not just behavioural outcomes, this one from the health services evaluation field, see Vingilis and Pederson (2001), p. 8.

It would have been technically feasible to assemble more robust evidence for Kirkpatrick’s third and fourth levels. We could have created a control group by ranking the top 60 applicants to the program, and randomly assigning those participants to program and control groups. Each group could have been subjected to comparable research protocols. However, three key issues led us to not take this approach to evaluating CALL. First, the cost of this more elaborate evaluation design would have meant a very substantial addition to the program’s budget. Second, it would have been difficult to motivate a control group to remain engaged in the research protocol over a four-year period. Third, the mission of the program is to have a real impact on the quality of leadership in the agriculture industry in Canada. Therefore, to potentially exclude some of the best applicants from the program, in order to strengthen the evaluation design, would be very difficult. From the outset of the program, we were aware of the difficulties of evaluating at levels three and four. We endeavoured to gather evidence, from the participants and from those in a position to observe the participants, about participant behaviour change and impact. We then inferred, given substantial self-reported learning, that at least some of that behaviour change and impact could be attributed to the program. This inference was supported by the ability of many participants to identify concrete examples of the connections between the program and their subsequent leadership practices.

Using the Kirkpatrick model to guide the evaluation of CALL involved substantial costs. In addition to the contract with the evaluation consultants, the program director needed to spend substantial time to design questions that would address the complex objectives of the program with regard to learning, behaviour change, and im-

pact. The effort to evaluate CALL at all four levels led to a significant burden being placed on participants. The three major survey instruments were very long (15, 9, and 9 pages, respectively), and the response rates were 22 of 30 for the mid-term survey, 23 of 28 for the final evaluation survey, and 15 of 28 for the follow-up survey. The Leadership Practices Inventory demanded some effort in order to identify and recruit observers, and while all participants completed the initial administration of the instrument, only 17 of 28 completed the second administration. With a less committed and motivated group of participants, the extent of these demands would be unrealistic.

The result of this substantial investment in evaluating CALL was a fairly convincing case to operate the program again. However, as has been noted above, this case was strongest at the level of participant satisfaction, and became weaker at each subsequent level of Kirkpatrick's model. The degree of difficulty in evaluating the program had an inverse relationship to Kirkpatrick's levels. That is, it was easiest to ascertain the level of participants' satisfaction with the program, and most challenging to determine the level of impact that their participation in the program had on the organizations and communities in which they held leadership roles. (For a similar viewpoint see Newstrom, 1978.) Our experience demonstrated that even with adequate funding, resources, and expertise it is challenging in practice to apply all four levels of Kirkpatrick's evaluation model. The inverse relationship between effort at each level and strength of evidence gathered at each level poses questions about the circumstances under which the need to document behaviour change and impact warrants the investment of time, money, and participant energy necessary to do so. For a large and relatively expensive program such as CALL, the investment seems appropriate. For smaller or less expensive programs, it may be hard to justify.

CONCLUSION

The purpose of this article has been to examine the Kirkpatrick framework as an evaluation tool through its application to a concrete case study. As such, it makes two fundamental contributions to the program evaluation literature. First, the article provides a detailed case study of evaluation practice in non-credit educational programming. The article describes evaluation practices as they actually happened, with frank recognition of the shortcomings of such practices. Case studies such as this one enrich evaluation theory

through describing its concrete application to the practice of evaluating programs. Second, the article provides a practical exploration of the Kirkpatrick framework. Did Kirkpatrick provide a useful framework for the evaluation of CALL? Yes. Our evaluation of CALL was more rigorous and convincing than the typical evaluation approach of similar leadership development programs in agriculture. Many related programs are evaluated through participant satisfaction surveys and the monitoring of graduates' leadership roles. Did Kirkpatrick provide a model that we could simply follow in order to achieve a satisfactory evaluation of the program? No. As this article demonstrates, our evaluation of CALL involved our best efforts, given resource and programmatic constraints, to gather evidence about the effectiveness of the program. Kirkpatrick provided a conceptual framework for our evaluation practices, but those practices were designed for the specific parameters of the CALL program. The Kirkpatrick framework is not a comprehensive program evaluation model, nor does it claim to be. It did, however, provide a useful organizing tool, given the focus of our evaluation of the CALL program. Extensions to the Kirkpatrick framework are needed (and have been proposed by others as cited earlier in this article) if it is to be included in a list of general program evaluation models. In spite of its limitations, we believe that the application of the Kirkpatrick framework was of practical benefit to our efforts to evaluate the CALL program.

REFERENCES

- Alliger, G.M., & Janak, E.A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology*, 42(2), 331–342.
- Bernthal, P.R. (1995, September). Evaluation that goes the distance. *Training and Development*, 49(9), 41–45.
- Carliner, S. (1997). Adapting the Kirkpatrick model to technical communication products and services. *Performance Improvement*, 36(4), 14–23.
- Craig, R.L. (Ed.). (1996). *The ASTD training and development handbook: A guide to human resource development* (4th ed.). New York: McGraw-Hill.
- Geber, B. (1995, March). Does your training make a difference? Prove it! *Training*, 32(3), 27–34.

- Hahne, C.E., & Schultze, D.E. (1996). Sales and marketing training. In R.L. Craig (Ed.), *The ASTD training and development handbook: A guide to human resource development* (4th ed.) (pp. 864–884). New York: McGraw-Hill.
- Holton, E.F., III. (1996). The flawed four-level evaluation model. *Human Resource Development Quarterly*, 7(1), 5–21.
- Johncox, V. (2000). Evaluability assessment of staff training in special care units for persons with dementia: Strategic issues. *Canadian Journal of Program Evaluation*, 15(Special Issue), 53–66.
- Kaufman, R., & Keller, J.M. (1994). Levels of evaluation: Beyond Kirkpatrick. *Human Resource Development Quarterly*, 5(4), 371–380.
- Kaufman, R., Keller, J., & Watkins, R. (1996). What works and what doesn't: Evaluation beyond Kirkpatrick. *Performance and Instruction*, 35(2), 8–12.
- Kirkpatrick, D.L. (1959a). Techniques for evaluating programs. *Journal of the American Society of Training Directors (ASTD)*, 13(11), 3–9.
- Kirkpatrick, D.L. (1959b). Techniques for evaluating programs: Part 2 - Learning. *Journal of ASTD*, 13(12), 21–26.
- Kirkpatrick, D.L. (1960a). Techniques for evaluating programs: Part 3 - Behavior. *Journal of ASTD*, 14(1), 13–18.
- Kirkpatrick, D.L. (1960b). Techniques for evaluating programs: Part 4 - Results. *Journal of ASTD*, 14(2), 28–32.
- Kirkpatrick, D.L. (1994). *Evaluating training programs: The four levels*. San Francisco: Berrett-Koehler.
- Kirkpatrick, D.L. (1996a). Evaluation. In R.L. Craig (Ed.), *The ASTD training and development handbook: A guide to human resource development* (4th ed.) (pp. 294–312). New York: McGraw-Hill.
- Kirkpatrick, D.L. (1996b). Great ideas revisited. *Training and Development*, 50(1), 54–59.
- Kirkpatrick, D.L. (1996c). Invited reaction: Reaction to Holton article. *Human Resource Development Quarterly*, 7(1), 23–25.

- Lee, S.H., & Pershing, J.A. (2000). Evaluation of corporate training programs: Perspectives and issues for further research. *Performance Improvement Quarterly*, 13(3), 244–260.
- Kouzes, J.M., & Posner, B.Z. (1997). *Leadership practices inventory*. San Francisco: Jossey-Bass Pfeiffer.
- Michalski, G.V., & Cousins, J.B. (2001). Multiple perspectives on training evaluation: Probing stakeholder perceptions in a global network development firm. *American Journal of Evaluation*, 22(1), 37–53.
- Moseley, J.L., & Larson, S. (1994). A qualitative application of Kirkpatrick's model for evaluating workshops and conferences. *Performance and Instruction*, 33(8), 3–5.
- Moss, G., & Kehoe, S. (2000). Beyond happy faces: Using evaluation to demonstrate program results. *Proceedings of the annual conference of the Canadian Association for University Continuing Education, 2000*, 124–132.
- Newstrom, J.W. (1978). Catch-22: The problems of incomplete evaluation of training. *Training and Development Journal*, 32(11), 22–24.
- Nolan, M. (1996). Job training. In R.L. Craig (Ed.), *The ASTD training and development handbook: A guide to human resource development* (4th ed.) (pp. 747–775). New York: McGraw-Hill.
- Phillips, J.J. (1996). Measuring the results of training. In R.L. Craig (Ed.), *The ASTD training and development handbook: A guide to human resource development* (4th ed.) (pp. 313–341). New York: McGraw-Hill.
- Pine, J., & Tingley, J.C. (1993). ROI [return on investment] on soft-skills training. *Training*, 30(2), 55–58.
- Robbins, D.W., Doyle, T.R., Orandi, S., & Prokop, P.T. (1996). Technical skills training. In R.L. Craig (Ed.), *The ASTD training and development handbook: A guide to human resource development* (4th ed.) (pp. 776–802). New York: McGraw-Hill.
- Rosenberg, M.J. (1996). Human performance technology. In R.L. Craig (Ed.), *The ASTD training and development handbook: A guide to human resource development* (4th ed.) (pp. 370–393). New York: McGraw-Hill.

- Shelton, S., & Alliger, G. (1993). Who's afraid of Level 4 evaluation? A practical approach. *Training and Development*, 47(6), 43–46.
- Stokking, K. (1996). Levels of evaluation: Kirkpatrick, Kaufman and Keller, and beyond. *Human Resource Development Quarterly*, 7(2), 179–183.
- Stufflebeam, D.L. (2001). *Evaluation models*. New Directions for Evaluation, 89. San Francisco: Jossey-Bass.
- Twitchell, S., Holton, E.F., 3rd, & Trott, J.W., Jr. (2000). Technical training evaluation practices in the United States. *Performance Improvement Quarterly*, 13(3), 84–110.
- Vingilis, E., & Pederson, L. (2001). Using the right tools to answer the right questions: The importance of evaluative research techniques for health services evaluation research in the 21st century. *Canadian Journal of Program Evaluation*, 16(2), 1–26.
- Warfel, S.L. (1996). Computer skills training. In R.L. Craig (Ed.), *The ASTD training and development handbook: A guide to human resource development* (4th ed.) (pp. 844–863). New York: McGraw-Hill.
- Watkins, R., Leigh, D., Foshay, R., & Kaufman, R. (1998). Kirkpatrick plus: Evaluation and continuous improvement with a community focus. *Educational Technology Research and Development*, 46(4), 90–96.

