Commonwealth Research Consulting, Inc.
Bulletin

CwRC–NIC Bulletin 7
June 2010

National Institute of Corrections Training Evaluation Project

Series Conclusion: Improved Evaluation Tools and Lessons from the Project

Stephen Parson, M.S.
James B. Wells, Ph.D.
Kevin I. Minor, Ph.D.

This is the seventh and final issue in the current series of research bulletins on NIC's Training Evaluation Project. This bulletin is based largely on a 2009 Master's Thesis completed by Mr. Parson (and supervised by Dr. Wells and Dr. Minor) at Eastern Kentucky University entitled *Development and Application of an Evaluation Matrix: A Structured Response to Deficiencies in the Kirkpatrick Model*. The project is being conducted by a team of researchers from Commonwealth Research Consulting, Inc. (CwRC), in collaboration with NIC's Division of Research and Evaluation. The purpose of the project, and the bulletin series, is to enhance understanding of training

Highlights

- ***Kirkpatrick's Approach to Training Evaluation includes four levels:***
  1. Reaction: Did participants like the training? Were they satisfied with the training?
  2. Learning: Did participants improve their job-related knowledge, skills, or attitudes?
  3. Behavior: Did participants apply the training and/or improve their behavior on the job?
  4. Results: Did organizational metrics improve as a result of training?
- ***Benefits of the Kirkpatrick Model:***
  - ⇒ It is simple, elegant, and commonsensical.
  - ⇒ It has increased evaluation interest and clarity.
  - ⇒ It has considerable practical utility.
  - ⇒ It provides a systematic, consistent structure for future development.
- ***Limitations of the Kirkpatrick Model:***
  - ⇒ It is a biased approach; neglects key stakeholders.
  - ⇒ It is simplistic, incomplete, and lacking in both theoretical and empirical support.
  - ⇒ It is not suited for full implementation.
- ***NIC's Training Evaluation Matrix improves on the Kirkpatrick approach by:***
  - ⇒ Expanding the scope of the model.
  - ⇒ Increasing specificity and precision in the model.
  - ⇒ Enhancing the potential for rigorous application of the model.

Major findings from the Project:

- ***Evaluability*** assessments contributed to a more effective allocation of limited resources by identifying trainings suitable for further evaluation.
- ***Target audience*** assessments suggest participant selection was generally appropriate for all trainings, and revealed no evidence of significant bias in participant selection, treatment, or benefit.
- ***Program implementation*** results were broadly favorable. All programs were in high demand, attendance and participation were good, and dropout rates were low. Results suggest all trainings were well organized, well-appointed, and implemented as designed.
- ***Reaction*** to training was broadly favorable, revealing generally high levels of satisfaction with nearly all dimensions of training and trainers. Results were mixed with regard to 6 of 46 trainers.
- ***Learning*** was good, though subjective reports typically suggested greater learning than objective measures.
- **Behavior/transfer** results were broadly favorable. Participants generally reported significant increases in training-related behaviors on the job and moderate progress implementing training objectives and action plans.
- ***Organizational change*** results were generally positive for trainings with a reasonable expectation of producing such change. For example, moderately rigorous evaluations of CLD and MDF showed some evidence of improvement on several organizational metrics.

Training Evaluation Project Primary Staff

For Commonwealth Research Consulting, Inc:

James B. Wells, Ph.D.
President and Chief Research Consultant
jbwells@cwrc.us

Kevin I. Minor, Ph.D.
Senior Research Consultant
kiminor@cwrc.us

J. Stephen Parson, M.S.
Research Consultant
jsparson@cwrc.us

For the National Institute of Corrections:

Christopher A. Innes, Ph.D.
Chief, Research and Evaluation
cinnes@bop.gov

Dee L. Halley
Correctional Program Specialist and Project Monitor
dhalley@bop.gov

Acknowledgements

The National Institute of Corrections Training Evaluation Project is made possible by the support of NIC via Cooperative Agreements 05A28GJF9, 06PEI01GJM1, 07PEI12GJQ7, and 08PEI21GJX1.

CwRC staff wish to acknowledge the support and cooperation of the many persons who helped make this project possible. Morris Thigpen, Larry Solomon, Tom Beauclair, Chris Innes, Dee Halley, Bob Brown, John Eggers, Leslie LeMaster, Launa Kowalcyk, Virginia Hutchinson, Belinda Watson, Fran Zandi, Cheryl Paul, Robbye Braxton-Mintz, Rob Jeffreys and others at NIC have provided essential support for this project. We also wish to acknowledge our support staff, whose daily efforts further the project in so many ways. Finally, we want to express our appreciation to the growing number of NIC trainers and training participants who have taken time out of their busy schedules to graciously share their insights with us.

Although many persons and organizations contributed to the project described in this bulletin, any errors or omissions are those of the authors alone.

The findings, interpretations, and views presented in this bulletin are those of the authors and do not necessarily reflect the positions or policies of the National Institute of Corrections, or any other organization or individual.

programs, and when appropriate, facilitate program improvements to better serve the field.

Previous bulletins in the series include:

1. ***Participant Demographics, Overall Evaluation of Training, and Applicability Ratings*** (February 2007)[1] provides a demographic sketch of 458 training participants, a discussion of early results from the evaluation project, and a preliminary profile of organizational resources and barriers to the implementation of training objectives in the workplace.
2. ***Participant Evaluation of Trainers*** (July 2007)[2] focuses on 34 trainers involved in 20 Academy Division trainings offered during the pilot phase of the project (2005-2006), and provides a discussion of both quantitative and qualitative findings.
3. ***Training Results, Activity Level Changes, and Implementation Results*** (February 2008)[3] discusses findings from a series of multivariate analyses of the relationships between demographic characteristics, training quality, post-training environments, and the successful implementation of training objectives in the organization.
4. ***2008 Evaluation Results: Satisfaction, Learning, and Action Plan Progress*** (November 2008)[4] provides preliminary evaluation results from four FY08 Jails Division and Prison Division trainings.
5. ***2008 Evaluation Supplement: Learning, Application and Action Plan Progress*** (March 2009)[5] updates and expands the evaluation described in Bulletin 4 to include recently collected follow-up data, and findings from a series of multivariate analyses.
6. ***Training, Leadership, and Organizational Change: Focus on CLD and MDF*** (July 2009)[6] examines the individual and organizational dimensions of leadership, and the relationships between training, leadership, and organizational change, based on several evaluations of Correctional Leadership Development (CLD) and Management Development for the Future (MDF).

These bulletins are available at: www.nicic.org/research.[7]

The current bulletin concludes this series by discussing two major accomplishments of the Training Evaluation Project:

1. **The development of improved evaluation tools.**
2. **Important insights gained from the large body of evaluation findings made possible by those tools.**

The bulletin begins with a review of the strengths and limitations of the most widely used training evaluation model. It continues with a discussion of the development and application of an improved approach to training evaluation. The bulletin concludes with a summary of findings from the Training Evaluation Project and future directions.

The Kirkpatrick Model

Training is important. Everyone is a stakeholder of training, both in terms of paying for it, and in terms of benefiting from (or suffering) its consequences. Everyone has a stake in their own training and the training of those with whom they interact in everyday life. The effects of adequate or inadequate training can be cumulative and compounding both between and within individuals as they move through life and interact in a multitude of ways.

Whether funded via tax dollars, consumer purchases, or other means, training costs permeate the economy. In a recent large scale review of the training and development literature, Aguinis and Kraiger[8] found that US organizations alone spend over $126 billion a year on employee training and development. They found "overwhelming evidence" that training is beneficial for individuals, organizations and society.[9] Yet only a fraction (10-30%) of training transfers to the workplace in terms of improved individual or organizational performance.[10-11] Thus, on a gross national scale, training appears to be effective, but not very efficient.

Given the importance of training, and growing evidence that most training doesn't fully transfer into desired outcomes, it is important to identify factors that aid or impede the transfer of training. As demonstrated by findings from the Training Evaluation Project (see especially Bulletins 3 and 5)[12-13] and a growing body of training evaluation literature, transfer is not solely a function of training quality; it is also influenced by individual characteristics of trainees, and the organizational context and conditions to which they return after training.

These issues fall under the purview of training evaluation, the most frequently ignored, and poorly executed element of the training enterprise.[14-15] Moreover, Kirkpatrick's four level model, by far the most popular approach to training evaluation,[16-20] is deeply flawed,[21-22] has undermined progress on evaluation theory, and has hindered efforts to conduct meaningful evaluations.[23-29] Given the importance of training and training evaluation, and the widespread use of a flawed evaluation model, it is critical to improve or move beyond the Kirkpatrick Model.

The significance of the bulletin is twofold. On the one hand, ***development of the NIC Evaluation Matrix*** and associated tools provide an evidence-based solution to a long-standing problem in the fields of training and evaluation. This solution, a direct and coordinated response to the limitations of the Kirkpatrick Model, carries a variety of implications for policy and practice. On the other hand, ***application of the NIC Evaluation Matrix*** in 25 training evaluations over the last five years has produced a large body of empirical findings.[30-35] These findings demonstrate the utility of the Evaluation Matrix, and provide a number of insights into improving effectiveness and efficiency in both training and evaluation.

Overview of the Kirkpatrick Training Evaluation Model

In 1959 and 1960 Donald Kirkpatrick published a series of four articles in what was then known as the *Journal of the American Society of Training Directors* (ASTD). He used those articles to lay out his four level approach to training evaluation which consists of:

1. Reaction: Did participants like the training? Were they satisfied with the training?
2. Learning: Did participants improve their job-related knowledge, skills, or attitudes?
3. Behavior: Did participants apply the training and/or change their behavior on the job?
4. Results: Did organizational metrics improve as a result of training?

Kirkpatrick's stated purpose in publishing these articles was "…to stimulate training directors to increase their efforts in evaluating training programs."[36] He wrote only a few statements that alluded to a theoretical basis for the model, and provided only general implementation guidelines, using the term "four steps." Kirkpatrick never intended the four steps to be a complete model, and didn't begin to use the terms "model" or "levels" until after they had been popularized and widely accepted by others.[37]

With regard to Level 1 (reaction) Kirkpatrick suggests that evaluators design a reaction sheet to quantify participants' thoughts and feelings about the training program, as well as various aspects of the program. He suggests making it anonymous to encourage honesty, inviting additional written comments to encourage thoroughness, and obtaining an immediate response rate of 100%. Other suggestions include developing minimum standards and measuring reactions against those standards.

At Level 2 (learning) Kirkpatrick suggests using validated, objective tests, on a pre-post basis, with a control group, and analyzing the data statistically. Again he stresses the importance of obtaining a 100% response rate. Level 2 includes three dimensions of learning: knowledge, skills, and attitudes. According to Kirkpatrick evaluation at Level 2 is more difficult, complicated, and expensive than Level 1, often requiring the assistance of a statistician.

To evaluate at Level 3 (behavior) Kirkpatrick suggests using a control group where feasible, collecting both pre and post data, and conducting statistical analyses. At this level he suggests collecting data not only from the trainees, but also from their subordinates, supervisors, and peers, if possible, and selecting 100 trainees (or an appropriate sampling). The distinction between knowing something and actually practicing it on the job is the difference between Level 2 and Level 3. However, it is also important to allow

time for behavior change to occur, so he suggests waiting at least three months before collecting follow-up data. Kirkpatrick observes that evaluating job behavior is more difficult than evaluating Levels 1 and 2; the assistance of statisticians, researchers, or consultants may be required. Finally, he suggests weighing the cost of the evaluation against the potential benefit, implying that a Level 3 evaluation is not always appropriate.

Kirkpatrick defines Level 4 (organizational results) as "a measure of the final results that occur due to training, including increased sales, higher productivity, bigger profits, reduced costs, less employee turnover, and improved quality."[40] Other examples of Level 4 results include facility safety, staff absenteeism, inmate grievance rates, job satisfaction, and organizational commitment. Again Kirkpatrick suggests using a control group, if feasible, measuring both before and after training, and allowing enough time for results to be achieved. According to Kirkpatrick, evaluation at Level 4 is the most difficult, and usually requires the assistance of researchers, consultants, or statisticians. He suggests weighing evaluation costs against potential benefits, implying a Level 4 evaluation is not always appropriate. Finally, Kirkpatrick recommends being satisfied with evidence if proof isn't possible.

Despite numerous alternative evaluation frameworks and models[41-45] this simple, unfinished model has become the most widely used and influential approach to training evaluation.[46-53] Although much has changed in the 50 years since Kirkpatrick's articles first appeared,[54] his four level model has "weathered well" over the years[55-57] and today remains almost unchanged from its original form.[58] The model has become a standard in training evaluation[59-60] and "… has now attained near universal acceptance with virtually all training evaluations in some way paying homage to Kirkpatrick's four levels."[61] The popularity and success of the Kirkpatrick Model is due in large part to its simplicity and utility, discussed next.

Benefits and Contributions of the Kirkpatrick Model

The Kirkpatrick Training Evaluation Model has served the training domain well.[62-65] As Bates puts it, "There is no doubt that Kirkpatrick's model has made valuable contributions to training evaluation thinking and practice."[66] A review of relevant literature suggests the most important benefits and contributions of the model include:

- It is simple, elegant, and commonsensical.
- It has increased evaluation interest, clarity, and focus.
- It has considerable practical utility.
- It provides a systematic, consistent structure for future development.

The Kirkpatrick Model is widely acknowledged to be simple, elegant, and appealing to common sense.[67-78] Those four simple words: ***reaction, learning, behavior,*** and ***results*** have an intuitive appeal, as if reflecting the natural order of things—a natural system, as opposed to a manmade contrivance. The temporal order and logical soundness of the four levels seem so obvious as to be undeniable. Arranged in this order, the four levels suggest a logical chain of causality beginning at Level 1. While Kirkpatrick has been roundly criticized for assuming or implying linear causality between the levels[79-96] several researchers have reported at least some evidence of relationships between the levels.[97-106]

Much of the increased interest in training evaluation is due to the simplicity and popularity of the Kirkpatrick Model.[107] It has helped draw attention to outcomes[108-110] and fostered the recognition that single outcome measures are inadequate for most evaluations.[111] The model has highlighted the distinction between learning and transfer, and the importance of transfer.[112-113] The increased clarity and interest in training evaluation has helped raise awareness of its importance and cost. Such heightened awareness is crucial at a time when management increasingly demands proof that training is effective and valuable, yet often fails to budget for adequate evaluation.[114]

The utility of the model has also drawn attention in the literature. For example, several writers note that Level 1 can aid in marketing training, and increasing training reputation, attendance, and funding.[115-116] Level 1 can also aid in program revision and decision making, i.e., formative and summative actions.[117-118] Bowers et al. note that the model can help determine not only what works and what does not, but "why."[119] Moreover, an immediate evaluation can serve not only as a check, but also as a reinforcement.[120] Likewise, Levels 2 and 3 can produce evidence of program effectiveness. With regard to Level 4, Bates writes "…this bottom-line focus is seen as a good fit with the competitive profit orientation of their sponsors."[121] Although Kirkpatrick indicates that the complexity and difficulty of evaluation increases at each level,[122] Nickols states that the necessary knowledge and tools exist to apply all four levels.[123]

Finally, a major benefit and contribution of the Kirkpatrick Model lies in its systematic approach to evaluation. Several writers note that the model employs a common language and structured format.[124-128] Dye goes further in pointing out that it provides a structure and consistency for criticism and further development.[129] In doctoral dissertations both Dye and Pulichino[130] assert that the Kirkpatrick Model, despite falling short of being a complete, researchable model, can nonetheless be used as a basis to build such a model. In fact, it has spawned many such modifications and extensions toward that end.[131-135]

These benefits and contributions closely mirror the goals expressed by Kirkpatrick.[136] In this sense he appears to have accomplished what he set out to do. Nonetheless, many writers believe some of these benefits have substantial downsides, rest on flimsy or faulty foundations, and don't go nearly as far as they should, and could. These and other criticisms and limitations are discussed next.

Limitations and Criticisms of the Kirkpatrick Model

Kirkpatrick's approach to training evaluation went largely unchallenged for several decades. Then in 1978 John Newstrom, a professor of management and industrial relations at the University of Minnesota, published a brief but convincing conceptual critique of several unsupported assumptions in the model.[137] Over the next several decades criticism grew and perhaps peaked with Edward F. Holton's 1996 article, *The Flawed Four-Level Evaluation Model*.[138] Where Newstrom's critique was largely conceptual, Holton's critique was theoretical. He found Kirkpatrick's model severely deficient, lacking all six elements of a theoretical model—the famous "taxonomy" critique, i.e., it is simply a classification scheme, not a fully researchable model. Still others have criticized the model on ethical, practical, and methodological grounds. Among the many criticisms identified in the literature, three common themes emerge:

- Kirkpatrick's work is biased in favor of training practitioners, ignoring other stakeholders.
- The model is simplistic, incomplete, and lacking in both theoretical and empirical support.
- The model is not suited to full implementation and routinely fails to produce the evidence deemed most valuable by the model itself.

These themes are discussed next.

A Biased Approach that Neglects Key Stakeholders

Over the course of 50 years Kirkpatrick has demonstrated a clear bias for the needs, status, and perspectives of trainers and training directors. Although an advocacy role can be appropriate in some situations, in others it is grossly inappropriate and potentially damaging. Kirkpatrick seems unconcerned that training evaluation has many constituencies and that the use of his model is important to a variety of stakeholders. His interests center on what is best for the training practitioner[139-140] and discount other stakeholders.[141-150] Moreover, his primary rationale for evaluation is justification of the training department.[151-152] Kirkpatrick himself states "When training directors can prove their programs have been effective in terms of learning as well as reaction, they will have objective data for selling future programs and increasing their status in the company." [153]

It is disturbing that the dominant evaluation model is based on such a narrow, self-justifying view of evaluation that it subordinates or dismisses the interests of other stakeholders and the organization itself. It is interesting to note that after decades of criticism on this point, the only change Kirkpatrick has made to the model is to list his rationales for evaluation in a slightly different order, with formative and summative rationales now appearing before self-justification.[154] The basis, substance, and thrust of his model, and his writings about the model, remain clearly trainer-centric. In essence he has popularized his model in large part by appealing to the self-interests of training practitioners.

Designers of evaluation methods, models, and tools have an ethical obligation to be diligent, thorough, and responsible in their work. They should know that their work has potentially far-reaching implications and should thus produce balanced and beneficial products, not systematically biased products that further the interests of one group at the risk or expense of others. Especially problematic, according to Bates, is "…the inability of the model to effectively address both the summative question (Was the training effective?) and the formative question (How can the training be modified in ways that increase its potential for effectiveness?)"[155] Bates points out that evaluators have an ethical obligation of beneficence to individuals and organizations, and that the Kirkpatrick Model may violate this principle because its limitations and risks outweigh its benefits.

Kirkpatrick purports to further evaluation science by discussing control groups, pre/post testing, response rates, intervening variables, and other methodological issues associated with rigorous science. At the same time, he makes no effort to incorporate these things in his model. Several writers have noted that his model is outdated,[156-159] deficient in theoretical foundation, and lacking empirical support.[160] In fact, Kirkpatrick shuns rigorous or scholarly approaches altogether.[161] He claims that he never intended his four levels to be a complete model—that he just wanted to provide some guidance to get people started on evaluation.[162-163] Yet at Levels 2, 3, and 4 he makes statements such as "this is more difficult than the previous level" and "you may need the help of statisticians, researchers, or consultants".[164] This begs the question,

> ***"To whom does he intend to provide guidance?" If he intends the model to be used by training practitioners, then why not provide a model they can use without professional help? On the other hand, if he intends for researchers, consultants, or professional evaluators to implement the model, then why not respond to their criticisms and improve the model?***

In fact, the closest Kirkpatrick has come to directly answering his critics is to respond to Holton's taxonomy claim, "Perhaps he is correct. I don't care whether it's a model or a taxonomy as long as training professionals find it useful in evaluating training programs."[165] However, more egregious than perpetuating numerous well-criticized and problematic assumptions (discussed in detail in the following sections), by refusing to address or correct them, Kirkpatrick persists in implying they are valid and true.[166-167] This cavalier, antiscientific attitude has given rise to a simplistic and incomplete evaluation model which is not suited for full implementation, and routinely fails to produce the very findings deemed most valuable and necessary by the model itself.

Model is Simplistic, Incomplete, and Lacking in both Theoretical and Empirical Support

Kirkpatrick's training evaluation model is simplistic, incomplete, and lacks detail.[168-182] What at first appears to be elegant simplicity is revealed upon closer examination to be vast oversimplification. Holton notes that Kirkpatrick's four level model is in fact a taxonomy (classification scheme), not a complete evaluation model, in that it ***lacks all six elements of a theoretical model*** *(constructs, relationships, boundaries, system states, deductions, and predictions).*[183]

The Kirkpatrick Model also fails to address such key issues as:

- Needs assessment[184-193]
- Individual or pre-training states, e.g., motivation, demographics, target audience, etc.[194-203]
- Post-training, organizational, situational, or contextual factors, transfer climate, politics, culture, resources, etc.[204-217]
- Efficiency, return on investment, added value, cost-benefit analysis, business requirements, etc.[218-228]

Kirkpatrick ignores individual differences, contextual factors, even the most rudimentary application of theory, and a host of other issues vital to the conduct of effective evaluations. Thus, it is no surprise that the model has been criticized for insufficient rigor.[229-233] However, Kirkpatrick further compounds the deficiencies of his model with a startling array of unsupported assumptions.

Among the most frequently criticized assumptions in the model is that of linear causality between the levels, i.e., positive reactions cause learning, learning causes behavior change, etc. Although he is careful to clarify that positive results at each level are not ***sufficient*** to ensure results at the next level, he nonetheless consistently states or implies that positive results at each level are ***necessary*** to ensure results at the next level. Over the years Kirkpatrick has made numerous statements such as: "people must like a training program to obtain the most benefit";[234] "if training is going to be effective, it is important that trainees react favorably";[235] "without learning, no change in behavior will occur"[236] despite lack of theoretical or empirical support for such relationships.[237-254]

In fact, according to some writers there can be no evidence of linear causality between the four levels as the model is currently constructed. Perhaps Holton said it best:[255]

> *The problem is not that it is a taxonomy but rather that it makes or implies causal statements leading to practical decisions that are outside the bounds of taxonomies. Causal conclusions, which are a necessary part of evaluation, require a more complete model. ... Attempts to test causal assumptions within a taxonomy are futile because, by definition, taxonomies classify rather than define causal constructs.*

In short, the problem is not that it is a taxonomy, but that it is a taxonomy masquerading as a model.

The causal linkage assumption is especially troubling because it has fostered the perception that reaction measures can be used as surrogates for learning, behavior, and results.[256] This has contributed to an over-reliance on Level 1 (reaction) and the under-utilization of Levels 2, 3, and 4. This narrow focus on reaction diverts attention from developing and delivering a truly effective training, and identifying ways to improve the training, and instead promotes entertainment over learning.[257] Kirkpatrick seems unaware that learning can be difficult and uncomfortable, and that reactions, especially gross reaction measures that don't distinguish affect (feelings/ enjoyment) from other dimensions of reaction (such as perceived quality, utility, etc.) may well correlate inversely or not at all with learning, behavior, or results. Moreover, even if Level 1 was shown to correlate directly with the other levels, this would not imply causality due to the numerous conceptual, theoretical, methodological, and other limitations of the Kirkpatrick Model.

A closely related and oft criticized assumption in Kirkpatrick's model is that the four levels are hierarchical in the sense that each level is more important and more informative than the previous level.[258-266] For example, Kirkpatrick has written "Evaluation becomes more difficult, complicated, and expensive as it progresses from Level 1 to Level 4—and more important and meaningful."[267] Yet he offers no theoretical basis or empirical evidence to support this assumption. At best he attempts to support the assumption by implying a parallel to yet another assumption, i.e., each level is more difficult to evaluate than the previous level, therefore each level must be more important and more informative than the previous level.

Further evidence of inadequate theoretical development is apparent in that the model ignores intervening, mediating, moderating, and confounding variables.[268-274] Clearly Kirkpatrick has been aware of the likely influence of such variables since early in his career. For example, in 1960 he wrote of "complicating factors" that can make evaluation of some programs difficult, and went on to quote E.C. Keachie regarding the difficulties associated with "'the separation of' variables.' That is, how much of the improvement is due to training as compared to other factors?"[275] Yet despite his awareness of the likely influence of such variables, and repeated criticisms over the years for not accounting for them, Kirkpatrick has made no effort to include such variables in his model.

The Kirkpatrick Model is also simplistic, incomplete, and unrefined in the sense that it assumes evaluation is definitive[276-277] and speaks to finite intervention, not continuous learning.[278] Similarly, Kirkpatrick's view of reaction as Level 1 implies that evaluation does not begin until after the training is finished.[279-280] This oversight is akin to the aforementioned lack of attention to needs assessment, individual differences, and other pre-training states. Likewise, the model is lacking with regard to the

follow-up evaluation, e.g., it fails to specify follow-up measures of utility reactions,[281-282] retained learning, satisfaction, etc. This oversight is akin to the aforementioned lack of attention to relevant post-training issues such as organizational, situational, and contextual factors, transfer climate, etc. It has also been noted that the model focuses solely on behavior or performance improvement, and fails to address establishing, maintaining, or extinguishing behavior.[283]

Some writers criticize the Kirkpatrick Model for not establishing desired outcomes up front.[284-287] However, others criticize it for being too focused on outcomes (the what) while ignoring the why/how, i.e., process and implementation issues that identify why the program was or was not effective, how it can be improved, etc.[288-290] Thus the model appears to be both deficient with respect to specific outcomes and excessive with respect to general outcomes. As Kearns points out, a well designed evaluation informs the learning process and brings together the various stakeholders of training.[291]

Other criticisms of the model have focused on specific levels. For example, despite the recognized importance of reaction measures, Kirkpatrick's formulation of Level 1 has been criticized for not being well researched or understood.[292-293] Brown notes specifically the lack of research on the construct validity of reactions.[294] In fact, several writers have suggested that Kirkpatrick's Level 1 is too simplistic and uni-dimensional to capture relevant dimensions of reaction identified in the literature, e.g., affective reactions, perceived utility, perceived difficulty, satisfaction with design, satisfaction with delivery, etc.[295-300] Level 2 has similarly been criticized as being too simplistic and uni-dimensional,[301-305] as has Level 3.[306-308]

When specific levels are criticized, Level 4 (organizational results) also draws considerable attention. At times Kirkpatrick confuses individual and organizational measures. For example, in his 1996 reprint of the original Level 4 article, Kirkpatrick gives an example of teaching typing and measuring words per minute as an example of Level 4 results.[309] This is incorrect. A trainee's tested typing speed before and after training is in fact a Level 2 measure of learning. Actual improvement on the job would constitute Level 3 behavior, or transfer of training. In both cases these are individual level measures, not organizational level results. In addition to his apparent confusion with regard to individual (micro level) and organizational (macro level) measures, Kirkpatrick assumes that outcomes can be aggregated across levels.[310] But organizational level outcomes are more than, and different from, the simple sum or aggregate of individual level outcomes. Elsewhere he directs the reader to focus on Levels 1-3, seeming at a loss to expound on Level 4. The original Level 4 article was dominated by phrases such as "…difficult if not impossible…", "…proceeding at a slow pace…", and "…eventually we may be able… ."[311] This seems understandable given the seminal nature of his work at the time. On the other hand, Kirkpatrick has made almost no changes to his model since publishing the original articles 50 years ago. Thus it is not surprising that contemporary writers criticize Level 4 as being too broad, too vague, and too uni-dimensional.[312-316]

Finally, Kirkpatrick assumes that training is designed (and able) to effect change at Level 4.[317-319] For example, as he wrote in his original Level 4 article and re-published in 1996 "The objectives of most training programs can be stated in terms of the desired results, such as reduced costs, higher quality, increased production, and lower rates of employee turnover and absenteeism. It's best to evaluate training programs directly in terms of desired results."[320] Yet most training programs are <u>not</u> designed to produce Level 4 change.[321] The audience of training programs is individuals, and most trainings are designed to produce change in the learning and behavior of those individuals. With rare exception, trainings are simply not delivered to entire organizations, nor intended to produce direct change in organizational level metrics. Kirkpatrick's confusion in this regard may help explain why he assumes that trainers are accountable for Level 4 change.[322]

Model is Not Suited for Full Implementation

Given the problems with Kirkpatrick's approach discussed thus far, it is not surprising that the model is difficult to implement and rarely utilized at Levels 3 and 4. Despite the enormous popularity of the model, it is seldom fully implemented.[323-338] Although estimates vary, most sources agree that 70-90% of training evaluations attempt to assess Level 1 (reaction), 30-40% attempt Level 2 (learning),10-20% attempt Level 3 (behavior/transfer), and something less than 10% attempt Level 4 (organizational results).[339-344] Kirkpatrick has stated that the less frequent use of the higher levels stems from the increased complexity and expense involved at higher levels.[345] Although he may be partially correct, a more compelling case can be made that frequent lack of full implementation stems largely from the substantial limitations of the model.

With higher levels of the model so seldom implemented, "We have then a situation whereby the most widely accepted framework for training evaluation consistently fails to produce the very evidence that it purports is necessary to demonstrate value".[346] Moreover, attempts to use the Kirkpatrick taxonomy as if it was a complete model typically leads to overgeneralizations and misunderstandings, and ultimately has limited our thinking and hindered our ability to conduct meaningful evaluations.[347-354] Recognizing this, researchers and practitioners have promoted novel modifications to the model in efforts to improve its effectiveness. While this has produced some potentially useful adaptations and modified models, it has generally produced scattered, inconsistent, and incompatible modifications.[355] Moreover, despite the many adaptations and modifications which have been put forth, relatively few have actually been applied or tested.

The NIC Evaluation Matrix

Improving the Kirkpatrick Model

Over the course of the Training Evaluation Project the research team employed a variety of strategies to address deficiencies in the Kirkpatrick Model. However, what began as work within the framework of the existing model evolved into a concerted effort to absorb the Kirkpatrick approach into a broader, deeper, and better specified model. Though certainly not complete, this process is well underway, and has been substantially aided by:

- A growing body of training evaluation literature reflecting wider recognition of Kirkpatrick Model deficiencies.
- An ongoing and productive collaboration between Commonwealth Research Consulting and NIC's Division of Research and Evaluation.
- A variety of empirical findings from data collected with the improved model.

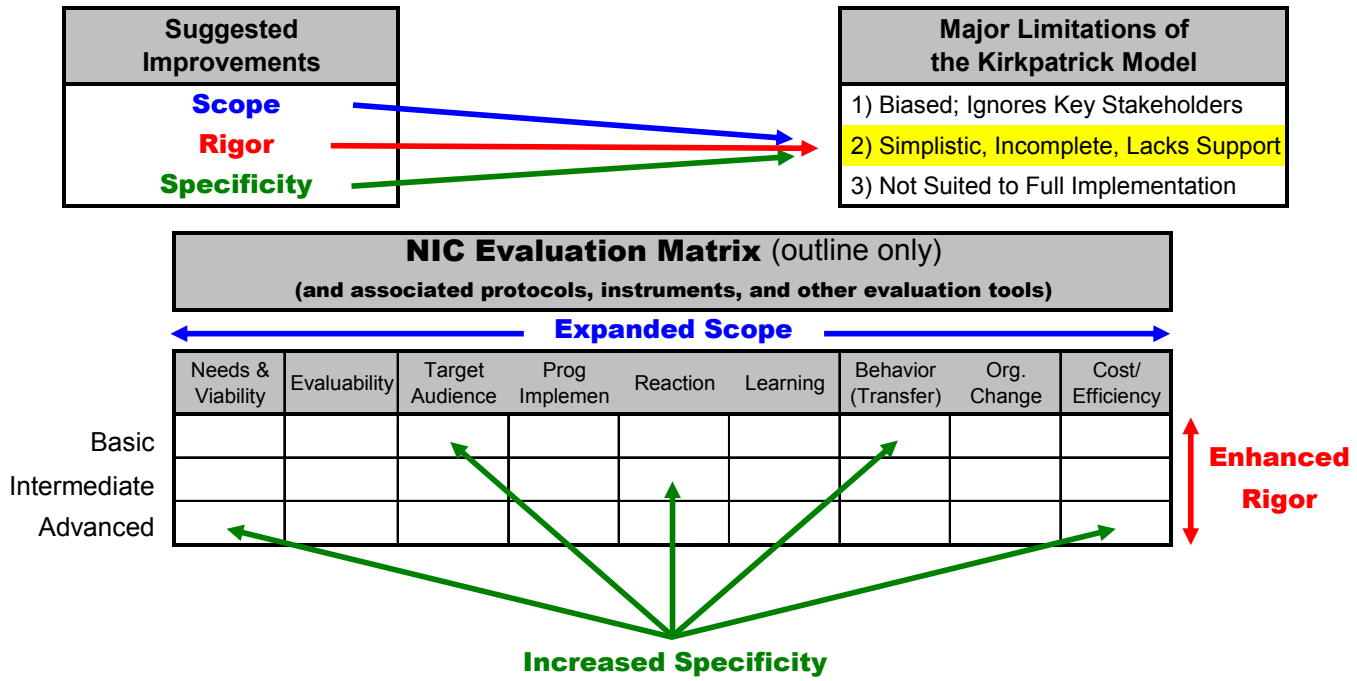
Critics of Kirkpatrick's approach have suggested numerous modifications and extensions to improve the model. A review of the many suggestions in the literature reveal three common themes. Most modifications and extensions aim to improve the model's:

- Scope (breadth)
- Rigor (depth) or
- Specificity (precision/detail)

Clearly these three areas don't align in a symmetric, one-to-one correspondence with the three major limitations of the model discussed previously (Figure 1, upper right). Improved scope, rigor, and specificity all primarily target the second limitation: the simplistic and incomplete nature of the model. Apparently most writers feel that simply identifying and publicly criticizing the bias inherent in the model (first limitation) is sufficient. Likewise, most writers apparently see little need to target recommendations directly at the implementation issue (third limitation), as this issue is symptomatic of the other limitations, and thus will improve if they improve.

Consistent with the principal themes identified in the literature, the research team employed a variety of techniques to improve the scope, specificity, and rigor of the Kirkpatrick Model. In addition to developing and validating a battery of evaluation instruments, the research team developed evaluation protocols and other tools to facilitate evaluation design and execution. One such tool is the NIC Evaluation Matrix (see Figure 1: outline version; Appendix A: full version; Appendix B: basic version). This tool provides evaluation guidance in the form of a matrix (grid) and companion narrative. The Matrix was designed to address limitations of the Kirkpatrick Model and thereby assist practitioners and researchers in designing and conducting sound evaluations. The horizontal dimension of the Matrix expands the breadth or scope of the Kirkpatrick model, while the vertical dimension expands the depth or rigor of the model. Likewise, narrative in the individual cells

Figure 1. The principal improvements required by the Kirkpatrick Model are embodied in the structure of the Evaluation Matrix. (See Appendix A for a draft version of the full matrix.)



of the Matrix and on its opening and closing pages provide guidance for conducting each type of evaluation at various levels of rigor, thus improving the specificity of the model.

Taken together, the Evaluation Matrix and associated instruments, protocols, and other tools (not appended) provide a more complete, balanced, and refined approach to evaluation than the Kirkpatrick Model. Moreover, utilization of the Matrix in approximately 25 training evaluations over the last five years has given rise to a substantial body of univariate, bivariate, multivariate, and qualitative evaluation findings (see Bulletins 1-6).[356-361] In addition to providing a rational, objective, and defensible basis for a variety of programmatic and policy initiatives, these findings suggest the beginnings of a theoretical foundation for the improved evaluation model. The following sections describe the development and application of the Evaluation Matrix and associated tools, and discuss how improvements in scope, rigor, and specificity address limitations of the Kirkpatrick Model.

Note that some degree of overlap exists between scope, specificity, and rigor, as these areas are naturally intertwined to a point. Thus the following sections each focus on a single area, while including elements of the others as appropriate.

Expanding the Scope of the Model

As discussed previously, the Kirkpatrick Model has been criticized for failure to address a variety of issues relevant to evaluation. One such area of omission stems from the implicit assumption that the evaluation begins only after the training has ended. Such an approach assumes away individual differences in training participants and ignores other important pre-training considerations such as needs and viability assessment, evaluability assessment, target audience, and program implementation. The Evaluation Matrix and related protocols, instruments, and other tools provide for the execution of these and other important pre-training evaluation steps not included in the Kirkpatrick Model. In fact, immediately apparent in the Evaluation Matrix (Appendix A) is that there are many types of evaluation, some of which should be conducted in conjunction with deciding if training is needed, while others should be conducted prior to or concurrently with training design. In no case should evaluation be an afterthought.

The most important pre-training evaluation type absent from the Kirkpatrick Model is the ***needs and viability assessment***. It is vital to know if the proposed training is in fact needed to address individual or organizational deficiencies.[362-372] Likewise, it is important to understand the evaluation needs of the various stakeholders.[373-374] As detailed in the "Needs and Viability Assessment" column of the Matrix (Appendix A), the purpose of this type of evaluation is to identify a target audience with important unmet needs that could likely be addressed effectively and efficiently with training. The Matrix provides a series of guiding questions to consider, and advice on data collection and analytic strategies for conducting needs assessments at various degrees of rigor. A needs assessment typically will be of benefit even in the case of mandated trainings, e.g., law or policy may dictate that a certain training must be done, yet offer little guidance on how it should be done. Thus, a needs assessment is useful not only in determining "if" a training should be done, but also why, how, when, and for whom it should be done.

Another important pre-training evaluation type absent from the Kirkpatrick Model is the ***evaluability assessment***. As explained in the "Evaluability Assessment" column of the Matrix, this type of evaluation attempts to answer the basic question, "Can this training reasonably be evaluated?" There are many reasons to conduct an evaluation; in fact, at least one reason for each of the nine types of evaluation listed in the Matrix, e.g., to determine if the training is needed, if it was implemented as designed, if the trainees mastered the learning objectives, etc. Nonetheless, even if there are many important reasons to evaluate a training, in some cases it is likely to be inadvisable to attempt an evaluation. There are many things to consider when determining if an evaluation is warranted and feasible. The evaluability assessment helps determine the extent to which any given training lends itself to evaluation by answering such questions as, "Are the training objectives clear and measurable?" "Will principal stakeholders support the evaluation?" "To what extent can the organization provide necessary data?" etc. In addition to guiding questions such as these, the Matrix provides advice on data collection and analytic strategies for conducting evaluability assessments at various degrees of rigor. Likewise, sample protocols, cover letters, informed consent forms, and other guidance and documents absent from the Kirkpatrick Model were designed to accompany the Matrix.

Target audience assessment is another important pre-training evaluation type not addressed by Kirkpatrick. The purpose of this assessment is to determine the extent to which the training is accessible and being delivered to the appropriate audience. (Note: the target audience should have been identified and specified as part of the needs assessment.) Although the target audience assessment can not be completed at least until final participant selections have been made, planning for the target audience assessment should begin well before the training. For example, a necessary part of any training design is a carefully specified description of the target audience. The target audience assessment could include a comparison of this element of the design against the target audience identified by the needs assessment. Once participant selections are announced, it is important to verify actual selections are consistent with the intended target audience. This is often difficult without collecting demographic data from participants. Thus the Matrix includes demographic instruments and protocols for administering them. The Training Evaluation Project includes numerous examples of qualitative and quantitative analyses conducted to ensure participant selection is consistent with the target audience, and participants are selected and treated without bias based on race, gender, age, etc. (see Bulletins 1 and 3-6)[375-379]

Before attempting to measure training outcomes, it is necessary to evaluate the extent to which the training actually happened, i.e., the extent to which it was implemented as designed. A ***program implementation assessment*** evaluates the degree to which staffing, budget, materials, and other training resources were available and actually utilized in all training locales consistent with the training design. Whether or not a subsequent outcome evaluation indicates that training objectives were met, it is beneficial to know if outcomes were due to target audience irregularities, training design/content, implementation, delivery, post-training conditions, or other factors.

Finally, as discussed previously, many writers have criticized the scope of the Kirkpatrick Model for not including ***cost or efficiency*** evaluations such as return on investment, added value, or cost-benefit assessment. Unlike outcome evaluations which seek to determine if a training achieved the desired outcomes, efficiency evaluations seek to determine if the outcomes were worth the costs. Such costs can be both direct, as in training design and delivery costs, and indirect, as in lost productivity while employees are being trained rather than working. As noted in the Evaluation Matrix, it only makes sense to conduct cost/efficiency evaluations when findings from outcome evaluations are favorable; one would not expect efficiency in the absence of effectiveness. Guiding questions offered by the Matrix include: "Do stakeholders consider the various benefits/results of the training worth the various costs?" "To what extent was the program efficacious from a financial standpoint?" Such questions go beyond traditional training measures to tap various business metrics. The Matrix provides guidance on conducting a cost/efficiency evaluation at various level of rigor.

The scope of the Matrix and associated protocols, instruments, and other evaluation tools is approximated by the nine evaluation types described in its vertical columns (Appendix A). This represents a substantial expansion over the Kirkpatrick Model which addresses only four of the nine, each in a more limited manner than the Matrix. Modifications and extensions with regard to the specificity or precision of the Kirkpatrick Model are discussed next.

Increasing Specificity in the Model

Note that many elements in this section also contribute to a more rigorous application of the Kirkpatrick Model.

As discussed previously, the Kirkpatrick Model is frequently criticized for being vague, simplistic, and failing to specify relevant dimensions of the four levels. Efforts to improve the specificity and precision of the model typically involve adding constructs or splitting the levels into more precise constructs and relevant dimensions of those constructs. Where efforts to increase the scope of the model typically involve adding new evaluation types or other broad elements **outside** the boundaries of the model, increased specificity more often involves adding smaller elements or dimensions **within** the boundaries of the model. The Evaluation Matrix and associate tools incorporate a number of such changes to the four levels.

Although Kirkpatrick provides for an evaluation of training satisfaction via gross ***reaction*** measures, he fails to distinguish affective reactions (enjoyment, irritation, etc.) from perceived utility and other relevant dimensions. Based on recommendations in the literature and evaluation project findings, the research team developed a multi-dimensional reaction instrument to collect separate reactions to training content, pace, relevance, physical context, and other dimensions. Moreover, rather than only seeking reactions to training delivery overall, the form includes numerous items about each individual trainer, such as demonstrated expertise, professionalism, preparation, organization, enthusiasm, interaction, and other areas. Recognizing that no survey can cover all possible relevant areas, the form includes several open-ended items to probe for additional commentary. Content analyses of participant responses have produced numerous important findings reported throughout the bulletin series.[380-385] Moreover, recognizing that initial reactions to training may not persist beyond the training environment, follow-up surveys revisit many of these reaction items 3-12 months after training.

Specificity at Level 1 was also increased by including several self-reported measures of Level 2 and Level 3 variables. For example, although self-reported learning can be utilized as a very basic (non-rigorous) Level 2 (learning) measure, it can also be utilized as a Level 1 (reaction) measure, i.e., a feeling or perception of learning, rather than an objective assessment of learning itself. Similarly, the form asked participants to rate the extent to which they anticipated applying what they learned in the training to their jobs on a scale of 0 to 10. Although this can be viewed as a very basic or non-rigorous Level 3 (behavior) measure, it can also be utilized as a reaction measure, i.e., a self-reported prediction or anticipation of behavior, rather than behavior itself.

In order to further probe the issue of application/transfer (behavior), and to act as a pre measure for the follow-up, participants were asked to complete a training action plan. This instrument asked participants to set several measureable goals they intended to pursue in the workplace based on their participation in the training, and to estimate their progress on each over the next three months. Again, although any actual future pursuit of those goals would constitute Level 3 behavior, the initial estimates can also be utilized as reactions to training.

With regard to ***learning*** (Level 2) Kirkpatrick appears to have dismissed any notion of a basic evaluation, recommending instead a rigorous full experimental design. He suggests using a control group, a pre-post design with objective tests based on training objectives, and statistical analyses to "prove" learning has occurred. However, a full experimental design is rarely feasible, and in any case, Kirkpatrick offers insufficient practical guidance for designing or implementing such a rigorous evaluation of learning. Likewise, he alludes to knowledge, skills, and

attitudes, implying various dimensions of learning, yet fails to specify how to measure these distinct dimensions. The Evaluation Matrix and related tools increase the specificity of the model by providing sufficient guidance for conducting such a rigorous evaluation. The Matrix also provides tools and guidance for conducting a more modest basic level evaluation. In fact, the Matrix provides for a range of learning evaluation strategies from basic self-assessments at a single point in time, to more rigorous valid, reliable, and objective measures on a pre, post, and post 2 basis. Moreover, various guidelines and instruments are provided for trainers, coworkers, and others to offer assessments of participant learning.

In describing Level 3, job ***behavior or transfer*** of training, Kirkpatrick again appears to dismiss any notion of a basic evaluation. Instead he makes a series of suggestions regarding control groups, large sample size, and other topics associated with rigorous evaluation, while offering little practical guidance in designing or implementing such a rigorous evaluation of behavior. Likewise, Kirkpatrick treats job performance in a very isolated manner. He ignores important pre-training measures such as trainee ability, motivation, and other background characteristics. He assumes relationships between reaction, learning, and behavior while providing no practical guidance for testing or measuring such relationships. He largely ignores organizational factors, providing no guidance in the identification of post-training resources and barriers to the transfer of training, despite repeated evidence that only a small fraction of training actually transfers to the job, due largely to post-training and organizational difficulties.[386-390]

The Evaluation Matrix and related tools provide increased specificity in the evaluation of post-training job performance by taking a broader and more detailed view of training transfer. Sample cover letters, consent forms, demographic forms, evaluation instruments, protocols, analytic strategies and other tools are provided to facilitate a range of evaluation strategies from basic self-assessments to more rigorous advanced evaluations. Evaluation instruments include a large number of training quality, organizational context, and participant background variables. Various evaluation protocols and analytic strategies examine potential relationships between these variables and post-training job performance (see Bulletins 3 and 5).[391-392]

Kirkpatrick's Level 4, ***organizational change***, also lacks adequate development. At times he confuses individual and organizational measures. Elsewhere he directs the reader to focus on Levels 1-3, without expounding on Level 4. He offers no practical guidance for measuring organizational change, or drawing causal connections from any particular training to organizational change. But most concerning, Kirkpatrick continues to assume that training is designed (and able) to effect organizational change. This is simply not true for many trainings. This assumption undermines both the formative and summative functions of evaluation by complicating efforts to improve the training, and inevitably finding that training failed to produce the desired (but unreasonable) organizational outcomes.

To address these limitations, the Evaluation Matrix begins by asking, "Was the training designed to achieve organizational level results?" Rather than assuming the answer is yes, this question can be examined as part of several types of evaluation. First, the needs and viability assessment can provide important information on the organization's needs and viable means to address those needs. For example, if the organization needs to reduce absenteeism (an organizational level metric), this assessment can help determine if training is likely to produce that outcome, i.e., training is not always the answer. Second, an evaluability assessment can help determine if training goals are clear, measurable, and feasible. For example, even in situations where training is determined to be a potentially viable solution to an organizational level problem, it is inadvisable to conduct a Level 4 evaluation if the training goals are not clear and measurable. Other examples can be drawn from each column (evaluation type) in the Matrix. For example, an organizational change evaluation is not likely to be productive unless the training was delivered to the intended target audience and implemented as designed. Thus, the Matrix improves on Level 4 by asking the question "Is this training designed to produce organizational level outcomes?" rather than assuming the answer is always yes. Moreover, the Matrix helps answer the question by drawing on findings produced by several types of pre-training evaluation not addressed by Kirkpatrick.

Consistent with efforts to improve specificity elsewhere in the model, the Evaluation Matrix provides practical guidance in the form of evaluation protocols, guiding questions, instruments, analytic strategies, sample cover letters, consent forms, and other tools to facilitate a range of organizational change evaluation strategies. For example, the research team developed a number of protocols and instruments for collecting self-ratings of organizational change from training participants. Such methods are appropriate for a basic level evaluation. However, the Matrix also provides guidance for conducting more rigorous evaluations of organizational change. For example, it provides protocol for administrating several valid and reliable instruments that tap organizational variables such as the Multifactor Leadership Questionnaire,[393] the Organizational Commitment Questionnaire,[394] and the Job Descriptive Index.[395] In an advanced evaluation, these instruments, and others developed by the research team, could be administered to training participants and others at various levels in the organizational hierarchy, both before and at some point 3-12 months after the training. Likewise, relevant data would be collected from organization records, such as turnover, absenteeism, disciplinary, etc. Bulletin 6 provides an analysis of organizational change data.

Enhancing Rigor in the Model

The Kirkpatrick Model is frequently criticized for being too simplistic and incomplete to implement with any appreciable degree of rigor. The research team employed a variety of strategies to improve upon the model. Previously discussed strategies to improve the scope and specificity of

the model in many cases also serve to enhance rigor. Nonetheless, additional modifications and extensions were targeted specifically at improving rigor.

Perhaps the most obvious element related to rigor is the vertical dimension of the Matrix itself (Appendix A). The bulk of the Matrix is devoted to descriptions of evaluation strategies at various degrees of rigor, from basic (near the top) to advanced (bottom row). Included are suggestions for evaluation design, measures, data collection, possible analytic strategies, etc. The Matrix and accompanying narrative explain how each of these change depending on desired degree of rigor, and how rigor impacts the range of possible conclusions and the confidence appropriate for conclusions at varying degrees of rigor. Evaluation protocols, instruments, and other documents used in conjunction with the Matrix provide additional instructions and tools for designing and implementing evaluations at various degrees of rigor.

An important goal of the research team was to absorb the productive elements of Kirkpatrick's four levels into the broader framework of the Evaluation Matrix while reducing or eliminating the bias, unsupported assumptions, and other limitations of the original model. Shifting the focus of evaluation from self-justification to effective evaluation reduces bias, highlights the need to include the needs and perspectives of other stakeholders, and facilitates more rigorous and meaningful evaluation. Moreover, a stakeholder approach naturally deals with the politics of evaluation, further reducing barriers to evaluation and facilitating more rigorous application of the model. The Matrix and related tools were designed with input from a variety of stakeholders. Likewise, evaluations based on the Matrix provide for stakeholder input as a matter of protocol.

Another important goal of the Matrix, as discussed previously under *Expanding the Scope of the Model*, was to include several pre-training planning and design issues not addressed by Kirkpatrick. When self-justification is deemphasized as a rationale for evaluation, model inadequacies with regard to planning and design become more apparent. For example, when stakeholder input is considered important, and the formative and summative functions are valued, the absence of needs and viability assessment, evaluability assessment, and target audience assessment become more obvious. The absence of these important pre-training evaluation types undermine not only training design and implementation, but also the quality and rigor of subsequent evaluation efforts. The Evaluation Matrix and related tools stress the importance of early evaluation planning. In fact, the very structure of the Matrix stresses this point simply by including these important pre-training evaluation types.

Rigorous application of the Kirkpatrick Model is further undermined by his isolated view of training that ignores trainee characteristics and organizational context, coupled with his unsupported assumptions of simple, direct relationships between the four levels. As reflected in the Evaluation Matrix and related tools, the research team collected reaction data, demographics, and other background data both from training participants and trainers in an effort to isolate reactions to training from reactions based on personal or individual characteristics. Collecting age, race, gender and other demographic and background data on participants and trainers allowed the research team to test for and control for the influence of non-training variables in both participant and trainer reactions. For example, it was possible to ascertain whether participants responded more favorably to trainers of a certain gender or race. Furthermore, such data facilitated the target audience and program implementation evaluations. Finally, these data provided the basis for several analyses to test for bias in participant selection and treatment. Credible statements about reaching the intended target audience and the absence of bias are not possible without collecting and analyzing demographic and background data.

The Evaluation Matrix and related tools aid in identifying and productively utilizing the various relationships between trainee, training, and organizational variables (see Bulletins 3-6.)[396-399] For example, participants were asked on a three, six, or twelve month follow-up not only to rate the extent to which they have applied their training on the job, but also to give progress ratings on each action plan goal they established previously during training. Organizational context is established for these measures by having participants rate a series of items on the extent to which they represented resources or barriers to participants' application of training on the job. These application, progress, and organizational context measures are then compared to estimates made at the time of training. The final analytic strategy incorporates all these measures along with participant background variables, all available measures of training quality and satisfaction, and other post-training contextual measures such as responses to a series of open-ended items on the follow-up.

Given the well-documented deficiencies of the Kirkpatrick Model, the research team is working to improve conceptual clarity within the four levels and to establish an adequate theoretical foundation for the four levels and the Matrix overall. For example, when incorporating reaction, learning, behavior/transfer, and organizational results into the Matrix, the research team has endeavored to define terms more precisely, clearly specify levels of analysis, and improve construct validity. Moreover, where appropriate, instruments include one or more qualitative items, and quantitative items are phrased in causal terms. Finally, research is underway to more clearly specify expected causal linkages, identify mediating, confounding, and intervening variables, and to further test and validate measures. This work, together with improved evaluation protocols and methodologies developed in conjunction with the Evaluation Matrix, enhances the potential for more rigorous evaluations of reaction, learning, behavior, and results.

Note that the Evaluation Matrix is a developmental stage product. In it's current form it represents a substantial improvement over the Kirkpatrick Model. However, it is not yet a complete and researchable model.

Lessons from the Project

This section provides a brief summary of findings from the Training Evaluation Project.[400] First, univariate/descriptive findings are organized around each of the nine types of evaluation addressed by the Evaluation Matrix and associated tools. Next, multivariate/inferential findings that span more than one type of evaluation are discussed. Implications for training and evaluation policy and practice are discussed where appropriate.

Training needs and viability assessments were not included in the scope of services for cooperative agreements associated with the Training Evaluation Project. However, based on informal discussions with various NIC personnel, it appears that NIC has drawn on a number of such assessments in developing and updating their portfolio of trainings.

Evaluability assessments also were not part of the formal scope of services for cooperative agreements on which the project was based. However, informal evaluability assessments were conducted in the course of developing each agreement, i.e., in deciding which trainings to include in the evaluation project, evaluability was a prime consideration. Discussions between project staff and key stakeholders addressed a number of evaluability issues such as those covered by the guiding questions in the Evaluation Matrix (Appendix A).

Target audience assessment is an important consideration in any evaluation. Participant demographic data were collected to facilitate target audience assessments in all trainings evaluated during the project (see Bulletins 1 and 4). Based on these and other data, and target audience profiles provided by NIC, several quantitative analyses and qualitative reviews were conducted to ensure that participant selection was consistent with the target audience, and that participants were selected and treated without bias based on age, race, gender, or other variables examined (see Bulletins 3-6 for participant demographic profiles and discussion of target audiences.) Analyses suggest participant selection was generally consistent with the target audience in all trainings. Analyses revealed no evidence that participant selection, treatment, satisfaction, or benefit from training was significantly influenced by bias or discrimination based on age, race, gender, or other variables examined.

Program implementation assessments were conducted both formally, as part of the overall training evaluation, and informally via discussions and interviews with various stakeholders. Results were broadly favorable. All programs evaluated were in high demand, attendance was excellent, participation was good, and dropout rates were low. Results suggest the programs were well organized, well staffed, and implemented as designed. There appeared to be no significant shortage of required resources. In most cases evaluation results were similar for various administrations of a training from place to place and year to year. Of note, however, trainer evaluations for five Meeting the Needs of Female Offenders trainings showed unusual variation from 2005 to 2006. For additional discussion of this and other potential concerns see Bulletin 2.

Participant reaction to training was evaluated via one or more multi-dimensional instruments developed to measure satisfaction with training content, pace, relevance, enjoyment, physical context, etc. Results were broadly favorable (see especially Bulletins 1 and 4). Similarly, the instruments addressed various aspects of each individual trainer. Again the results were quite positive, revealing high levels of satisfaction with 40 of the 46 trainers evaluated (see Bulletins 2 and 4). However, results were mixed for 6 of the 34 trainers evaluated in 2005-2006 (see Bulletin 2 for additional details.) All reaction instruments contained both quantitative and qualitative (open-ended) items. Qualitative results (including content analyses of narrative responses) both supported and clarified quantitative results by displaying a consistent pattern of findings, yet with greater variation (see for example figures 7a-7c in Bulletin 2.) Follow-up ratings of satisfaction collected 3-12 months after the trainings typically remained quite high, yet slightly lower than initial ratings (see Bulletins 1, 4, and 5).

Participant learning was assessed via a variety of methods, including self-reported learning, trainer ratings of participant learning, and objective pre-post1-post2 (follow-up) tests, depending on the evaluation design for each training. Results were generally positive, though subjective measures (both self-report and trainer estimates) typically indicated greater learning than objective measures (see Bulletins 3-5.)

Behavior/performance change (application or transfer of learning) was assessed on a pre/post basis via a variety of methods, depending on the evaluation design for each training. Methods included self-reports, team reports, and 360 degree reports involving feedback from the training participant and numerous coworkers, direct reports, supervisors, and others in the organization. Behavior was assessed both with regard to actions/efforts and the results/outcomes of those efforts, e.g., progress on action plan goals or success implementing training objectives in the workplace. Results were broadly favorable. Participants generally reported significant changes in behavior as a result of training, such as increased training-related behaviors or applying training on the job (see Bulletins 3-5). Similarly, both training participants and others typically reported that those actions bore fruit in the sense of successfully implementing training objectives on the job (Bulletin 3), making progress on action plan goals (Bulletins 4-5), or improving leadership practices (Bulletin 6). In cases where progress was less than anticipated, about 75% of participants reported that their agencies could have better supported their efforts, while only 14% indicated that

NIC could have done anything to improve their progress (see Bulletin 5.)

Organizational change was similarly evaluated by collecting data from a variety of sources, again depending on the evaluation design for the training in question. Sources of organizational change data included training participants, team members of participants who also attended the training in question, team members of participants who did not attend the training, and other members of participants' organizations who did not attend the training, e.g., their supervisors, subordinates, coworkers, and others. Results were generally quite favorable. For example, CLD participants on average reported significantly increased organizational involvement in key training-related activities after they attended training (see Bulletin 6). Likewise, MDF participants reported moderate to exceptional progress improving selected organization-wide key performance indicators after attending training (see Bulletin 6.) Perhaps most important, however, members throughout the organizational hierarchy at dozens of correctional institutions reported significant positive changes in leadership after a senior leader at their facility attended CLD (see Bulletin 6.)

Cost-Benefit or Efficiency evaluations were not conducted in the pilot phase (2005-2006) of the project, nor in the standardization/streamlining phase of the project (2007-2008). The focus during the first phase was exploration and capacity building, while the focus in the second phase was validation and simplification. Based on the improved evaluation designs, instruments, protocols, and other tools, some current (and perhaps most future) evaluations will employ cost-benefit, return-on-investment, added-value, or other efficiency evaluations. Both basic level/subjective and more advanced level/objective instruments are currently under development for use in a related project.

Several themes emerge from the project that reach across multiple types of evaluations. Consistent with findings and recommendations in the literature, ***pre-training evaluation activities,*** such as evaluability and target audience assessments, proved invaluable to the project. For example, in several cases informal evaluability assessments contributed to a more effective allocation of scarce resources by identifying trainings suitable (and not suitable) for further evaluation efforts. Similarly, target audience assessments conducted throughout the project failed to produce significant evidence of bias, discrimination, or differential benefit (e.g., learning, behavior/performance improvement, etc.) based on age, race, gender, or other demographic variables examined (See Bulletins 3-5). These and other pre-training evaluation activities are critical to the success of any evaluation. In no case should evaluation be an afterthought.

Qualitative data, such as those drawn from narrative responses to open-ended survey items, were essential in identifying important areas and dimensions untapped by quantitative (closed-ended) items. For example, see Bulletin 2 (p. 7) for a discussion of the role of qualitative data in the evaluation of trainer strengths and limitations. Qualitative data can also provide the foundation for more sophisticated analyses. For example, see Bulletin 3 for a discussion of the predictive value of organizational resources and barriers in multivariate analyses of training outcomes. Note that these resource/barrier items were initially derived and subsequently refined via content analysis of narrative responses to open-ended items, i.e., qualitative data. ***Organizational resources and barriers*** subsequently proved to be excellent predictors (in many cases the best predictors) of post-training outcomes in most evaluations. Finally, such data can serve to clarify and corroborate quantitative results. For example, see Bulletin 2 (p. 10) for a discussion of the role of qualitative data in magnifying and confirming more subtle patterns observed in quantitative trainer evaluation data.

Input from a variety of stakeholders is crucial to the success of any training or evaluation endeavor. Trainers and training directors are important stakeholders, but certainly not the only stakeholders. Likewise, trainees are important sources of evaluation data, but certainly not the only viable source of evaluation data. Although more advanced evaluation designs will necessarily involve more sophisticated data collection strategies than basic evaluations, all evaluations should consider collecting ***balanced data***. Even a basic evaluation, for example, can benefit from collecting trainer feedback in addition to participant feedback. Likewise, even in situations where collecting objective data is not a viable option, it can still be beneficial to, for example, circulate a brief subjective (opinion-based) cost-benefit survey to 8-10 stakeholders in various departments. This can be both a means to increase buy-in and a potentially valuable source of information. Line staff, middle management, and upper administration may have very different needs and views regarding the value of training. For example, higher ranking correctional personnel were found to be more critical of training (in the sense of suggesting more improvements) than lower ranking personnel. This is especially interesting given that the two groups did not differ to any practical extent on training satisfaction ratings, learning, action plan progress, or any other available training evaluation measure (see Bulletin 4, pgs. 7-8).

Finally, in one of the more ***rigorous evaluations*** in the project, Correctional Leadership Development was found to be an effective method of improving leadership among senior leaders and achieving positive organizational change. Moreover, the quality of the training, as rated by leader-participants, was predictive of the leadership ratings they would receive six months after the training from their direct reports. These and related findings are the subject of Bulletin 6. Evaluations such as this one require a high degree of cooperation and coordination among the various stakeholders, considerable pre-training evaluation planning, and the execution of an intermediate to advanced evaluation design as specified by the NIC Evaluation Matrix and related tools.

Future Directions

Our partnership with NIC over the last several years has yielded a lot of useful information about training, and training evaluation. Over the last three years we have produced a half dozen bulletins that have not only informed NIC training professionals about how to improve training programs, but as well, have allowed NIC to build additional capacity in conducting its own training evaluations.

The partnership has also benefited CwRC in many respects. As a result of being involved in ever more challenging training evaluations with NIC, we have become well-versed in the best practices training literature, and developed a unique set of evaluation skills and capabilities. A good illustration of this is evident from the contents of this bulletin that describe how we have developed improved evaluation tools as well as gained important insights from the large body of evaluation findings made possible by these tools.

The current training evaluation we are conducting on behalf of NIC in is the most challenging yet. The purpose of the study is to rigorously evaluate the relative merits of web-based instruction versus classroom-based instruction. The study will employ randomized experimental design in evaluating the effectiveness and efficiency of web-based and classroom-based versions of a specially designed new training for juvenile justice supervisors (Conducting Employee Evaluations). This training was designed from the ground up to address a serious problem supervisors face in properly documenting and evaluating the performance of subordinates. To ensure that both the online and face-to-face training media contain a balance of instructional impact, NIC Curriculum Development Specialists provided valuable technical assistance to the trainers involved in developing the training curriculum.

The amount of training conducted online is increasing rapidly, both within and outside the field of corrections. Online learning presents lots of advantages and opportunities for staff development. These include: increased access, greater flexibility, cost savings, and increased collaboration. Despite these advantages, there are several disadvantages of online staff development. They include: quality of content and process, hidden costs, and readiness of the online learner.[401] The current study will evaluate online training both in terms of its relative advantages and disadvantages, and in terms of its effectiveness and efficiency relative to classroom training.

This NIC-funded project will utilize a very sophisticated research methodology referred to as a "classic experimental design." This design is considered the most rigorous of all research designs because training participants will be randomly assigned to either an experimental group that receives the online training, or a control group that receives the traditional classroom training. Because both groups are probabilistically equivalent, and both trainings were stringently designed to differ only in delivery modality, any differences between the groups can be assumed to result from the experimental treatment, i.e., online delivery of the training. As a result, this type of design is considered the "gold standard" in terms of establishing cause-effect relationships. Currently, most research on online learning has primarily been descriptive or exploratory and considered low quality, particularly in terms of establishing cause-effect relationships.[402] Even more recent online research has been unable to provide true experimental data to identify the causal relationships behind this new type of instructional technology.[403]

In addition to utilizing evaluation measures that are now standard to most NIC trainings (e.g., Participant Evaluation of Training, Trainer Evaluation of Participants, Training Action Plan, etc.), we will also utilize other instruments to measure the following:

- the learning environment of the training
- the learning style of the participants
- the participant's motivation to learn and anxiety about the learning experience
- knowledge gained and retained from the training
- the amount and type of learning that transferred to the participant's work environment
- both participant and external ratings of the participant's performance
- cost versus benefit of the training

Approximately half of the above mentioned measures are existing instruments that have already been shown to be valid and reliable in previous studies. The remaining instruments are measures that we developed. In order to ensure that these newly developed instruments are valid and reliable, they will be pre-tested and pilot-tested after they are subjected to both internal and external review by subject matter experts.

As a result of this project, NIC will have excellent data and a variety of findings regarding the impact of online training relative to classroom training. In addition to training effectiveness and efficiency results with regard to participant knowledge, attitude, and behavioral change, NIC will also gain valuable insights regarding online learner outcomes, online learner characteristics, the online course environment, and online cost versus benefit factors.

Notes

[1] Wells, J. B., Minor, K. I., & Parson, J. S. (2007, February). *Participant demographics, overall evaluation of training, and applicability ratings* (Bulletin No. 1). Eastern Kentucky University, Center for Criminal Justice Education and Research.

[2] Wells, J. B., Minor, K. I., & Parson, J. S. (2007, July). *Participant evaluation of trainers* (Bulletin No. 2). Eastern Kentucky University, Center for Criminal Justice Education and Research.

[3] Wells, J. B., Minor, K. I., & Parson, J. S. (2008, February). *Training results, activity level changes, and implementation results* (Bulletin No. 3). Lexington, KY: Commonwealth Research Consulting, Inc.

[4] Wells, J. B., Minor, K. I., & Parson, J. S. (2008, November). *2008 Evaluation results: Satisfaction, learning, and action plan progress* (Bulletin No. 4). Lexington, KY: Commonwealth Research Consulting, Inc.

[5] Wells, J. B., Minor, K. I., & Parson, J. S. (2009, March). *2008 Evaluation supplement: Learning, application, and action plan progress* (Bulletin No. 5). Lexington, KY: Commonwealth Research Consulting, Inc.

[6] Wells, J. B., Minor, K. I., & Parson, J. S. (2009, July). *Training, leadership, and organizational change: Focus on CLD and MDF* (Bulletin No. 6). Lexington, KY: Commonwealth Research Consulting, Inc.

[7] The location of the bulletins is subject to change. If a search of the NIC website does not locate the bulletins, please contact Dr. James Wells at jbwells@cwrc.us or 859.806.5748 for copies.

[8] Aguinis H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology, 2009, 60,* 451-474.

[9] Aguinis & Kraiger (2009), p. 467.

[10] Eseryel, D. (2002). Approaches to evaluation of training: Theory & Practice. *Educational Technology & Society, 5(2),* 93-98.

[11] Liebermann, S. & Hoffmann, S. (2008). The impact of practical relevance on training transfer: Evidence from a service quality training program for German bank clerks. *International Journal of Training and Development, 12(2),* 74-86.

[12] Wells et al., 2008a, Bulletin 3.

[13] Wells et al., 2009a, Bulletin 5.

[14] Burrow, J., & Berardinelli, P. (2003). Systematic performance improvement – refining the space between learning and results. *Journal of Workplace Learning, 15 (1),* 6-13.

[15] Eseryel, 2002.

[16] Aguinis & Kraiger, 2009.

[17] Alvarez, K., Salas, E., & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human Resources Development Review, 3(4),* 385-416.

[18] Bates, R. (2004). A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence. *Evaluation and Program Planning, 27,* 341-347.

[19] Holton, E. F. (1996). The flawed four-level evaluation model. *Human Resources Development Quarterly, 7(1),* 5-21.

[20] Sutton, B., & Stephenson, J. (2005). A review of 'Return on Investment' in training in the corporate sector and possible implications for college-based programmes. *Journal of Vocational Education and Training, 57(3),* 355-373.

[21] Newstrom, J. W. (1978). Catch-22: The problems of incomplete evaluation of training. *Training and Development Journal, November 1978,* 22-24.

[22] Holton, 1996.

[23] Abernathy, D. J. (1999). Thinking outside the evaluation box. *Training & Development, February 1999,* 19-23.

[24] Bates, 2004.

[25] Bernthal, P. R. (1995). Evaluation that goes the distance. *Training & Development, September 1995,* 41-45.

[26] Dye, K. L. (2002). Effective HRD evaluation: An expanded view of Kirkpatrick's four levels. *Dissertation Abstracts International, 63* (01), 53A. (UMI No. 3038608)

[27] Eseryel, 2002.

[28] Islam, K. (2004). Alternatives for measuring learning success. *Chief Learning Officer, November 2004,* 32-37.

[29] Pulichino, J. (2007). Usage and value of Kirkpatrick's four levels of training evaluation. Dissertation Abstracts International, 68(04). (UMI 3264775)

[30] Wells et al., 2007a, Bulletin 1.

[31] Wells et al., 2007b, Bulletin 2.

[32] Wells et al., 2008a, Bulletin 3.

[33] Wells et al., 2008b, Bulletin 4.

[34] Wells et al., 2009a, Bulletin 5.

[35] Wells et al., 2009b, Bulletin 6.

[36] Kirkpatrick, D. L. (1996). Revisiting Kirkpatrick's four-level model. *Training & Development, January 1996,* 54-59.

[37] Kirkpatrick, 1996. Note also that many have argued, as we do in this bulletin, that the term "model" is a misnomer when applied to Kirkpatrick's evaluation taxonomy. Nonetheless, for the sake of convenience and consistency with common usage, we retain and use of the term here.

[38] Kirkpatrick, 1996, pgs. 57-58.

[39] Kirkpatrick, 1996, p. 58.

[40] Kirkpatrick, 1996, p. 56.

[41] See Alvarez et al., 2004.

[42] Eseryel, 2002.

[43] Holton, 1996.

[44] Simkins, T., Coldwell, M., Close, P., & Morgan, A. (2009). Outcomes of in-school leadership development work: A study of three NCSL programmes. *Educational Management Administration & Leadership, 37(1),* 29-50.

[45] Sutton & Stephenson, 2005.

[46] Aguinis & Kraiger, 2009.

[47] Alliger, G. M., Tannenbaum, S. I., Bennett, W., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology, 1997, 50,* 341-358.

[48] Alvarez et al., 2004.

[49] Bates, 2004.

[50] Eseryel, 2002.

[51] Holton, 1996.

[52] Naugle, K. A., Naugle, L. B., & Naugle, R. J. (2000). Kirkpatrick's evaluation model as a means of evaluation teacher performance. *Education, 121(1),* 135-144.

[53] Nickols, F. (2000). *Evaluating training: There is no "cookbook" approach.* Retrieved May 1, 2009, from http://www.nickols.us/.

[54] For a brief historical perspective, see Nickols, 2000.

[55] Abernathy, 1999.

[56] Bernthal, 1995.

[57] Nickols, 2000.

[58] Kirkpatrick, 1996.

[59] Holton, 1996.

[60] Pulichino, 2007.

[61] Sutton & Stephenson, 2005, p. 357.

[62] Bowers, C. A., Hitt, J. M., Hoeft, R. M., & Dunn, S. (2003). Applying training evaluation models to the clinical setting. *Military Psychology, 2003, 15(1),* 17-24.

[63] Holton, 1996.

[64] Pulichino, 2007.

[65] Sekowski, G. J. (2002). Evaluating training outcomes: Testing an expanded model of training outcome criteria. Dissertation Abstracts International, 63(12), 6130B. (UMI No. 3076227)

[66] Bates, 2004.

[67] Alliger et al., 1997.

[68] Alvarez et al., 2004.

[69] Bates, 2004.

[70] Bernthal, 1995.

[71] Holton, 1996.

[72] Ilian, H.(2004).Levels of Levels: Making Kirkpatrick Fit the Facts and the Facts Fit Kirkpatrick. In B. Johnson, V. Flores, & M. Henderson (Eds.), *Proceedings of the 6th Annual Human Services Training Evaluation Symposium* (pp. 89-104). Berkeley, CA: California Social Work Education Center.

[73] Kearns, P. (2005). From return on investment to added value evaluation: The foundation for organizational learning. *Advances in Developing Human Resources, 7(1),* 135-145.

[74] Pulichino, 2007.

[75] Sekowski, 2002.

[76] Sutton & Stephenson, 2005.

[77] Wells, J. B. (2008). How rigorous should your training evaluation be? *Corrections Today, October 2008,* 116-118.

[78] Wong, P., & Wong, C. (2003). The evaluation of a teacher training programme in school management. *Educational Management & Administration, 31(4),* 385-401.

[79] Alliger et al., 1997.

[80] Alvarez et al., 2004.

[81] Bates, 2004.

[82] Brown, K. G. (2005). An examination of the structure and nomological network of trainee reactions: A closer look at "smile sheets". *Journal of Applied Psychology, 90 (5),* 991-1001.

[83] Burrow & Berardinelli, 2003.

[84] Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85(5),* 678-707.

[85] Dye, 2002.

[86] Holton, 1996.

[87] Ilian, 2004.

[88] Kearns, 2005.

[89] Kirkpatrick, 1996.

[90] Martin W. M., & Lomperis, A. E. (2002). Determining the cost benefit, the return on investment, and the intangible impacts of language programs for development. *TESOL Quarterly, 36(3),* 399-429.

[91] Naugle et al., 2000.

[92] Pulichino, 2007.

[93] Richmond, H. (2008). Beyond Kirkpatrick: An evaluation dilemma. *Training Journal, March 2008,* 51-54.

[94] Sekowski, 2002.

[95] Sutton & Stephenson, 2005.

[96] Wong & Wong, 2003.

[97] Aguinis & Kraiger, 2009.

[98] Alliger et al., 1997.

[99] Alvarez et al., 2004.

[100] Brown, K. G., 2005.

[101] Colquitt et al., 2000.

[102] Faerman, S.R., & Ban, C. (1993). Trainee satisfaction and training impact: Issues in training evaluation. *Public Productivity & Management Review, 16(3),* 299-314.

[103] Liebermann & Hoffmann, 2008.

[104] Sekowski, 2002.

[105] Sitzmann, T., Brown, K. G., Casper, W. J., Ely, K., & Zimmerman, R. D. (2008). A review and meta-analysis of the nomological network of trainee reactions. *Journal of Applied Psychology, 93(2),* 280-295.

[106] Warr, P., Allan, C., & Birdi, K. (1999). Predicting three levels of training outcome. *Journal of Occupational and Organizational Psychology, 72,* 351-375.

[107] Bates, 2004.

[108] Bates, 2004.

[109] Eseryel, 2002.

[110] Richmond, 2008.

[111] Bates, 2004.

[112] Bates, 2004.

[113] Naugle et al., 2000.

[114] Mendosa, R. (1995). Is there a payoff? *Sales and Marketing Management, 147(6),* 64-70.

[115] Alliger et al., 1997.

[116] Brown, K. G., 2005.

[117] Brown, K. G., 2005.

[118] Sitzmann et al., 2008.

[119] Bowers et al., 2003.

[120] Kirkpatrick, D. L., & Kirkpatrick, J. D. (2008). Assessing training results. *Security Management, March 2008,* 103-104.

[121] Bates, 2004, p.342

[122] Kirkpatrick, 1996.

[123] Nickols, F. (2004). *A stakeholder approach to evaluating training.* Retrieved May 1, 2009, from http://www.nickols.us/.

[124] Alliger et al., 1997.

[125] Alvarez et al., 2004.

[126] Bates, 2004.

[127] Dye, 2002.

[128] Martin & Lomperis, 2002.

[129] Dye, 2002.

[130] Pulichino, 2007.

[131] Bates, 2004.

[132] Eseryel, 2002.

[133] Holton, 1996.

[134] Sutton & Stephenson, 2005.

[135] Wells, 2008.

[136] Kirkpatrick, 1996.

[137] Newstrom, 1978.

[138] Holton, 1996.

[139] Pulichino, 2007.

[140] Nickols, 2004.

[141] Bates, 2004.

[142] Burrow & Berardinelli, 2003.

[143] Eseryel, 2002.

[144] Islam, 2004.

[145] Kearns, 2005.

[146] Nickols, 2000.

[147] Nickols, 2004.

[148] Pulichino, 2007.

[149] Richmond, 2008.

[150] Sutton & Stephenson, 2005.

[151] Dye, 2002.

[152] Sutton & Stephenson, 2005.

[153] Kirkpatrick, 1996, p. 57.

[154] Kirkpatrick, 1996.

[155] Bates, 2004, p. 346.

[156] Abernathy, 1999.

[157] Dye, 2002.

[158] Kearns, P. (2004). From trainer to learning consultant: An evolution. *Training Journal, May 2004,* 40-44.

[159] Richmond, 2008

[160] Pulichino, 2007.

[161] Pulichino, 2007.

[162] Kirkpatrick, 1996.

[163] Stoel, D. (2004). The evaluation heavyweight match. *Training & Development, January 2004,* 46-48.

[164] Kirkpatrick, 1996, pgs. 57-58.

[165] Kirkpatrick, 1996, p. 55.

[166] Pulichino, 2007.

[167] Sekowski, 2002.

[168] Bates, 2004.

[169] Bernthal, 1995.

[170] Dye, 2002.

[171] Eseryel, 2002.

[172] Holton, 1996.

[173] Ilian, 2004.

[174] Islam, 2004.

[175] Kearns, 2004.

[176] Kearns, 2005.

[177] Newstrom, 1978.

[178] Pulichino, 2007.

[179] Richmond, 2008.

[180] Sekowski, 2002.

[181] Sutton & Stephenson, 2005.

[182] Wells, 2008.

[183] Holton, 1996.

[184] Aguinis & Kraiger, 2009.

[185] Alvarez et al., 2004.

[186] Bernthal, 1995.

[187] Burrow & Berardinelli, 2003.

[188] Colquitt et al., 2000.

[189] Eseryel, 2002.

[190] Kearns, 2004.

[191] Martin & Lomperis, 2002.

[192] Sutton & Stephenson, 2005.

[193] Wells, 2008.

[194] Aguinis & Kraiger, 2009.

[195] Holton, 1996 in Alvarez et al., 2004.

[196] Bates, 2004.

[197] Bernthal, 1995.

[198] Colquitt et al., 2000.

[199] Tannenbaum, 1993 in Alvarez et al., 2004.

[200] Holton, 1996.

[201] Ilian, 2004.

[202] Sitzmann et al., 2008.

[203] Warr et al., 1999.

[204] Alvarez et al., 2004.

[205] Bates, 2004.

[206] Tannenbaum, 1993 in Alvarez et al., 2004.

[207] Bernthal, 1995.

[208] Brown, R., M., & Eggers, J. T. (2005). Management development for impact. *Corrections Today, December 2005,* 100-103.

[209] Colquitt et al., 2000.

[210] Dye, 2002.

[211] Faerman & Ban, 1993.

[212] Holton, 1996.

[213] Ilian, 2004.

[214] Richmond, 2008.

[215] Sekowski, 2002.

[216] Simkins et al., 2009.

[217] Warr et al., 1999.

[218] Aguinis & Kraiger, 2009.

[219] Burrow & Berardinelli, 2003.

[220] Goldwasser, D. (2001). Beyond ROI. *Training, 38(1),* 82-88.

[221] Holton, 1996.

[222] Honeycutt, E. D., Karande, K., Attia, A., & Maurer, S. D. (2001). An utility based framework for evaluating the financial impact of sales force training programs. *Journal of Personal Selling & Sales Management, 21 (3),* 229-238.

[223] Phillips, 1996 in Ilian, 2004.

[224] Islam, 2004.

[225] Kearns, 2005.

[226] Martin & Lomperis, 2002.

[227] Wells, 2008.

[228] Wong & Wong, 2003.

[229] Aguinis & Kraiger, 2009.

[230] Dye, 2002.

[231] Holton, 1996.

[232] Newstrom, 1978.

[233] Wells, 2008.

[234] Kirkpatrick,1959a in 1996, p. 55.

[235] Kirkpatrick, 1994, p. 27, in Bates, 2004.

[236] Kirkpatrick, 1994, p. 51, in Bates, 2004.

[237] Alliger et al., 1997.

[238] Alvarez et al., 2004.

[239] Bates, 2004.

[240] Brown, K. G., 2005.

[241] Burrow & Berardinelli, 2003.

[242] Colquitt et al., 2000.

[243] Dye, 2002.

[244] Holton, 1996.

[245] Ilian, 2004.

[246] Kearns, 2005.

[247] Kirkpatrick, 1996.

[248] Martin & Lomperis, 2002.

[249] Newstrom, 1978.

[250] Pulichino, 2007.

[251] Richmond, 2008.

[252] Sekowski, 2002.

[253] Sutton & Stephenson, 2005.

[254] Wong & Wong, 2003.

[255] Holton, 1996, p.7.

[256] Bates, 2004.

[257] Bates, 2004.

[258] Bates, 2004.

[259] Bernthal, 1995.

[260] Burrow & Berardinelli, 2003.

[261] Kearns, 2004.

[262] Kearns, 2005.

[263] Newstrom, 1978.

[264] Pulichino, 2007.

[265] Sekowski, 2002.

[266] Sutton & Stephenson, 2005.

[267] Kirkpatrick,1996, p. 56.

[268] Aguinis & Kraiger, 2009.

[269] Holton, 1996.

[270] Kearns, 2005.

[271] Mendosa, 1995.

[272] Pulichino, 2007.

[273] Richmond, 2008.

[274] Simkins et al., 2009.

[275] Kirkpatrick, 1960 in 1996, p. 59.

[276] Alvarez et al., 2004.

[277] Dye, 2002.

[278] Abernathy, 1999.

[279] Kearns, 2005.

[280] Sutton & Stephenson, 2005.

[281] Alliger et al., 1997.

[282] Sekowski, 2002.

[283] Dye, 2002.

[284] Burrow & Berardinelli, 2003.

[285] Kearns, 2005.

[286] Nickols, 2000.

[287] Sutton & Stephenson, 2005.

[288] Bates, 2004.

[289] Eseryel, 2002.

[290] Kearns, 2005.

[291] Kearns, 2005.

[292] Brown, K. G., 2005.

[293] Sitzmann et al., 2008.

[294] Brown, K. G., 2005.

[295] Aguinis & Kraiger, 2009.

[296] Alliger et al., 1997.

[297] Brown, K. G., 2005.

[298] Honeycutt et al., 2001.

[299] Sekowski, 2002.

[300] Warr et al., 1999.

[301] Alliger et al., 1997.

[302] Kraiger et al., 1993 in Bowers et al., 2003.

[303] Colquitt et al., 2000.

[304] Sekowski, 2002.

[305] Warr et al., 1999.

[306] Alliger et al., 1997.

[307] Tannenbaum, 1993 in Alvarez et al., 2004.

[308] Warr et al., 1999.

[309] Kirkpatrick, 1996.

[310] Dye, 2002.

[311] Kirkpatrick, 1960b in Kirkpatrick, 1996.

[312] Alliger et al., 1997.

[313] Burrow & Berardinelli, 2003.

[314] Ilian, 2004.

[315] Islam, 2004.

[316] Sekowski, 2002.

[317] Bates, 2004.

[318] Burrow & Berardinelli, 2003.

[319] Sekowski, 2002.

[320] Kirkpatrick, 1996, p. 59.

[321] Burrow & Berardinelli, 2003.

[322] Bernthal, 1995.

[323] Aguinis & Kraiger, 2009.

[324] Brown, K. G., 2005.

[325] Burrow & Berardinelli, 2003.

[326] Eseryel, 2002.

[327] Gray, G. R., Hall, M. E., Miller, M., & Shasky, C. (1997). Training practices in state government agencies. *Public Personnel Management, 26(2),* 187-202.

[328] Holton, 1996.

[329] Kearns, 2004.

[330] Lovell, K. (2007). Getting the value out of evaluation. *Training Journal, January 2007,* 43-46.

[331] Martin & Lomperis, 2002.

[332] Nickols, 2004.

[333] Olsen, J. H. (1998). The evaluation and enhancement of training transfer. *International Journal of Training and Development, 2(1),* 61-75.

[334] Pulichino, 2007.

[335] Sitzmann et al., 2008.

[336] Sutton & Stephenson, 2005.

[337] Taylor, P. J., Lamers, A., Vincent, M. P., & Driscoll, M. P. (1998). The validity of immediate and delayed self-reports in training evaluation: An exploratory field study. *Applied Psychology: An International Review, 47(4)*, 459-479.

[338] Wong & Wong, 2003.

[339] See Alliger, 1997.

[340] Meyer M. K., & Elliott V. (2003). *Training Evaluation: A review of literature.* (NFSMI Item No. R-59-03). University of Mississippi, National Food Service Management Institute.

[341] Eseryel, 2002.

[342] Pulichino, 2007.

[343] Sitzmann, 2008.

[344] Sutton & Stephenson, 2005.

[345] Kirkpatrick, 1996.

[346] Sutton & Stephenson, 2005, p. 357.

[347] Abernathy, 1999.

[348] Bates, 2004.

[349] Bernthal, 1995.

[350] Dye, 2002.

[351] Eseryel, 2002.

[352] Islam, 2004.

[353] Pulichino, 2007.

[354] Richmond, 2008.

[355] Dye, 2002.

[356] See Wells et al., 2007a, Bulletin 1.

[357] See Wells et al., 2007b, Bulletin 2.

[358] See Wells et al., 2008a, Bulletin 3.

[359] See Wells et al., 2008b, Bulletin 4.

[360] See Wells et al., 2009a, Bulletin 5.

[361] See Wells et al., 2009b, Bulletin 6.

[362] Aguinis & Kraiger, 2009.

[363] Alvarez et al., 2004.

[364] Bernthal, 1995.

[365] Burrow & Berardinelli, 2003.

[366] Colquitt et al., 2000.

[367] Eseryel, 2002.

[368] Martin & Lomperis, 2002.

[369] Nickols, 2000.

[370] Sutton & Stephenson, 2005.

[371] Wells, 2008.

[372] Adam, 2001 in Wong & Wong, 2003.

[373] Bernthal, 1995.

[374] Pulichino, 2007.

[375] See Wells et al., 2007a, Bulletin 1.

[376] See Wells et al., 2008a, Bulletin 3.

[377] See Wells et al., 2008b, Bulletin 4.

[378] See Wells et al., 2009a, Bulletin 5.

[379] See Wells et al., 2009b, Bulletin 6.

[380] See Wells et al., 2007a, Bulletin 1.

[381] See Wells et al., 2007b, Bulletin 2.

[382] See Wells et al., 2008a, Bulletin 3.

[383] See Wells et al., 2008b, Bulletin 4.

[384] See Wells et al., 2009a, Bulletin 5.

[385] See Wells et al., 2009b, Bulletin 6.

[386] Baldwin & Ford, 1988 in Eseryel, 2002.

[387] Robinson & Robinson, 1995 in Liebermann & Hoffmann, 2008.

[388] Warr et al., 1999.

[389] Wells et al., 2008a, Bulletin 3.

[390] Wells et al., 2009a, Bulletin 5.

[391] See especially Wells et al., 2008a, Bulletin 3.

[392] See especially Wells et al., 2009a, Bulletin 5.

[393] Avolio, B.J., Bass, B.M., & Jung, D.I. (1999). Re-examining the components of transformational and transactional leadership using the Multifactor Leadership Questionnaire. J*ournal of Occupational and Organisational Psychology, 72*, 441-462.

[394] Mowday, R. T., Steers, R. M., & Porter, L. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior, 14*, 224-247.

[395] Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement.* Chicago: Rand McNally.

[396] See Wells et al., 2008a, Bulletin 3.

[397] See Wells et al., 2008b, Bulletin 4.

[398] See Wells et al., 2009a, Bulletin 5.

[399] See Wells et al., 2009b, Bulletin 6.

[400] Due to the large number of findings reported in the bulletin series, not all are summarized here. While an effort was made to include the most important findings, the selection process was subjective. See individual bulletins for additional findings and discussion. For bivariate correlations see especially Bulletin 4.

[401] Killion, J. (2000, October). Online staff development: promise or peril? *NASSP Bulletin, 84(618)*, 38-46.

[402] Bernard, R.M., Abrami, P.C., Lou, Y., Borokhovski, E., Wade, A., Wozney, L., Wallet, P.A., Fiset, M., & Huang, B. (2004, Fall). How does distance education compare with classroom instruction? A meta-analysis of the empirical literature. *Review of Educational Research, 74(3)*, 379-439.

[403] Tallent-Runnels, M.K., Thomas, J.A., Lan, W.Y., Cooper, S., Ahern, T.C., Shaw, S.M, & Liu, X. (2006, Spring) Teaching courses online: a review of the research. *Review of Educational Research, 76(1)*, 93-135.

Appendix A.

Commonwealth Research Consulting, Inc.
Evaluation Matrix

CwRC–NIC Evaluation Matrix
January 2010
(Draft 4)

National Institute of Corrections Training Evaluation Project

Selecting Evaluation Type and Rigor: The NIC Evaluation Matrix

James B. Wells, Ph.D.
Kevin I. Minor, Ph.D.
Stephen Parson, M.S.

This document is designed to assist NIC staff and trainers, as well as internal and external evaluators, in understanding how different types of evaluations can be conducted at varying degrees of rigor. The purposes of the document are to assist in identifying: (a) the type(s) of evaluation appropriate and feasible for a given training situation, and (b) the degree of rigor required to validly address the questions of interest, and yet feasible to implement. Types of evaluation and degrees of rigor are described in the upper panel of table below. The lower panel illustrates how the two can be integrated to form a conceptual tool.

All evaluation types can be beneficial; they are distinguished by the questions which guide them and the research methods appropriate for addressing those questions. Depending upon the particulars of a given training, it may be neither desirable nor necessary to perform all evaluation types.

Likewise, all degrees of rigor can provide information useful to various stakeholders. As with evaluation types, the degrees of rigor are distinguished by the questions of interest as well as the designs, measures, and other methods needed to address those questions. But degrees of rigor are also distinguished by the degree of confidence desired for the reliability and validity of research findings. To the extent that an evaluation involves more advanced rigor, there is more attention to reliable and valid measurement, greater effort to track and control influences that might bias inferences drawn, and there may be greater corroboration of information obtained.

By design, basic evaluations are less rigorous than those conducted at the intermediate or advanced level. Nonetheless, all evaluations should consider: 1) needs and viability; 2) evaluability; 3) target audience; 4) program implementation;

Continued on Page 4

Evaluation Types	Evaluation Rigor
• **Needs and Viability Assessment** • **Evaluability Assessment** • **Process Evaluations** (four): target audience; program implementation; participant reaction; and participant learning • **Outcome Evaluations** (two): learning transfer/ behavior change; and organizational change • **Cost/Efficiency Evaluation**	• Evaluations employing **basic rigor** are primarily descriptive of the group studied and do not support inferences about cause-effect relationships. • **Advanced evaluations** can support valid and reliable inferences about: (a) individuals and groups beyond those in the study, and (b) cause – effect relationships with regard to the training. • The **intermediate degree** represents a bridge between the basic and advanced degrees, with movement toward more advanced rigor.

Evaluation Matrix (Overview)

Evaluation Type: / **Evaluation Rigor:**	Needs & Viability Assessment	Evaluability Assessment	Process Evaluations	Outcome Evaluations	Cost / Efficiency
Basic -Univariate/Descriptive					
Intermediate -Univariate/Descriptive -Bivariate/Inferential					
Advanced -Univariate/Descriptive -Bivariate/Inferential -Multivariate/Inferential					

See full Evaluation Matrix (pages 2-3) for details on various types of evaluations conducted with various degrees of rigor.

NIC Evaluation Matrix (full version)

Evaluation Type:			Needs and Viability Assessments	Evaluability Assessment	Process Evaluations	
					Target Audience	Program Implementation
Guiding Questions:			· To what extent are the field's needs, and/or standards being met by existing training programs? · What must be done to meet those needs and standards? · Who is the target population? · What forms of training are likely to be necessary, viable, effective, and/or attractive to meet the needs and/or standards of this particular correctional target population?	· To what extent are program goals clear, measurable and feasible? · Is there agreement about the intended effects of the training program? · To what extent does the program lend itself to evaluation? · Is evaluation desired? · To what use will the findings be put? · Who are the principal stakeholders? · What is stakeholder interest in the evaluation? · How will stakeholders use the findings? · To what extent can organizations supply information? · What effect(s) is the program expected to have on the training participants?	· To what extent is the program reaching the intended target population? · How many persons are receiving the training? · Are persons receiving the training those intended to receive the training? · To what extent are members of the target population aware of the program?	·To what degree was the program implemented as designed? ·To what extent are the various program functions being performed as intended? · Is staffing sufficient? · Is the amount, type and quality of training consistent with what theory suggests? · Is the program well organized? · Are resources sufficient to support the program? · Are there variations in resources across training sites?
Eval. Rigor	**Possible Analytic Strategies**					
Basic	**Univariate Analyses** ·Frequencies ·Descriptive Statistics	**Design & Measures**	· Data collected from unsystematic sample (e.g., convenience, snowball, etc.) · Use of non-validated instrumentation or measurement · Existing data sources (e.g., social indicators that estimate size of problem, surveys, agency records) · Informants (potential clientele, leaders)	· Review program documents, prepare program description, interview program personnel & stakeholders, identify evaluation users, achieve agreement to proceed - stakeholders should agree on: program objectives, program components to be analyzed, design of evaluation, priorities for undertaking work, commitment of required resources, necessary cooperation and collaboration, and plan for utilizing evaluation results.	Evaluation strategy based on participant's self rating of the following: · Demographic Information & Background · Awareness of NIC training	Evaluation strategy based on using basic measures of program implementation and delivery (e.g., participation, attendance, etc.).
Intermediate	**Univariate Analyses** ·Frequencies ·Descriptive Statistics **Bivariate Analyses** ·Analysis of Change ·Analysis of Relationships/ Associations ·Basic Inferential Statistics	**Design & Measures**	· Attempt to use systematic sampling · Use of measurement instrument in the process of being validated or able to estimate validity of measure used · Existing data sources (e.g., social indicators that estimate size of problem, agency records) · Informants (potential clientele, leaders), · Qualitative survey (e.g., focus groups) to obtain rich information on the nature of the problem	· Review program documents, prepare program description, interview program personnel & stakeholders, identify evaluation users, achieve agreement to proceed - stakeholders should agree on: program objectives, program components to be analyzed, design of evaluation, priorities for undertaking work, commitment of required resources, necessary cooperation and collaboration, identification of potential resources, identification of potential barriers to both program implementation and evaluation, plan for utilizing evaluation results, plan for efforts required from program staff to strengthen the evaluability potential of program components not currently amenable to evaluation, & approach for subsequently building them into the evaluation effort. · Emphasis placed on working toward describing theory underlying the program.	Evaluation strategy based on: participant's self rating of the following: · Demographic Information & Background, awareness of NIC training · Comparison of target group characteristics with actual participant characteristics	Evaluation strategy based on using: · Basic measures of program implementation and delivery (e.g., participation, attendance, etc.). · Description of program's actual operation based on interviews and surveys. · Emphasis placed on working toward monitoring of frequency, duration, intensity of program delivery as well as receptivity of participants
Advanced	**Univariate Analyses** ·Frequencies ·Descriptive Statistics **Bivariate Analyses** ·Analysis of Change ·Analysis of Relationships/ Associations **Multivariate Analyses** ·Inferential Statistics ·Analysis of Relationships controlling for other factors	**Design & Measures**	· Use of probability sample · Use of validated measure · Use of experimental or quasi-experimental design · Existing data sources (e.g., social indicators that estimate size of problem, agency records) · Key person surveys · Informants (potential clientele, leaders), · Qualitative survey · Quantitative assessment to provide reliable estimate of the extent and distribution of the problem	· Review program documents, prepare program description, interview program personnel & survey stakeholders, scout program (conduct site visits, observe program functions), review social science literature, identify evaluation users, achieve agreement to proceed - stakeholders should agree on: program objectives, program components to be analyzed, design of evaluation, priorities for undertaking work, commitment of required resources, necessary cooperation and collaboration, identification of potential resources, identification of potential barriers to both program implementation and evaluation, plan for overcoming potential barriers to both program implementation and evaluation, plan for utilizing evaluation results, plan for continued efforts required from program staff to strengthen the evaluability potential of program components not currently amenable to evaluation, & approach for subsequently building them into the evaluation effort. · Emphasis placed on clearly describing theory underlying the program as well as comparing theoretical results with needs identified.	Evaluation strategy based on: · Participant's self rating of the following: demographic information & background, awareness of NIC training · Comparison of target group characteristics with actual participant characteristics · Assess program overcoverage and undercoverage by examining program records, surveys of program participants, and surveys of field. · Assess bias in program coverage by comparing program users, eligible nonparticipants, and dropouts.	Evaluation strategy based on: · Description of program's actual operation based on interviews and surveys. · Comprehensive monitoring of frequency, duration, intensity of program delivery as well as receptivity of participants. · Analysis of the degree the program is consistent to its design. · Analysis of the extent the program is being implemented as intended. · Comparison of implementation across multiple sites.

NIC Evaluation Matrix (full version)

Process Evaluations		Outcome Evaluations		Cost / Efficiency
(Level 1: Reaction)	(Level 2: Learning)	(Level 3: Behavior)	(Level 4: Organizational Change)	(Level 5: ROI / Added Value)
· Were participants satisfied with the training (i.e., course content & applicability, & instructor) immediately at the conclusion of the training? · Were participants satisfied with the training over a follow-up period?	· To what extent was there any knowledge/skill/attitude change? · To what degree did these changes persist over 3-12 months after the program?	· Is the training effective in attaining the desired behavioral goals or benefits? · To what extent was there transfer of knowledge/skills/attitude to the workplace? · To what extent do the training's performance measures approximate outcomes?	· Was the training designed to achieve organizational level results? · Is the training effective in attaining the desired organizational goals or benefits? · To what extent was there organizational change? · To what extent is there evidence that the training had the expected effects on the participant's organization?	· Do stakeholders consider the benefits/results of the training to be worth the costs? · To what extent was the training efficacious from a financial standpoint? · To what extent was the training a good investment in relation to its returns? · Did the training produce effects that extended beyond the organization? Note: Evaluation at this level should only be conducted on programs that demonstrate desired outcomes.
Evaluation strategy based on participant's rating of Overall Evaluation of Training immediately following the training; no attempt made to distinguish dimensions of training (using non-validated instruments)	Evaluation strategy based on participant's rating of the following using non-validated instrument: · Knowledge/skill/attitude change	Evaluation strategy based on: · Participant's self rating of behavioral change · Use of research design that focuses on outcomes. · Identification of both proximal and distal outcomes (no attempt to distinguish the two).	Evaluation strategy based on: · Participant's self rating of organizational change · Use of research design that focuses on outcomes. · Identification of both proximal and distal effects (no attempt to distinguish the two).	Evaluation strategy based on: · One or more stakeholders' self ratings of whether benefits exceeded costs · Attempt to conduct cost-benefit analysis by estimating costs associated with program and comparing that to existing data on program outcomes as regards cost savings.
Evaluation strategy based on participant's rating of the following immediately after the training and/or over a follow-up period: · Overall Evaluation of Training · Applicability of Training · Anticipated and Encountered Barriers to and Resources for application · Evaluation of Trainers (using instruments in the process of being validated)	Evaluation strategy based on ratings of some combination of the following (using instruments in the process of being validated) · Participant's self perceptions of knowledge/skill/attitude change (e.g. unstandardized pre/ post tests) · Collect perceptual data from third parties to substantiate participant's self perceptions of knowledge/skill/attitude change (e.g. survey of supervisors, coworkers, subordinates, and/or correctional program specialists) · Development and piloting of knowledge/skill/attitude change tests in an effort to move toward the use of reliable and valid measures.	Evaluation strategy based on some combination of the following (using instruments in the process of being validated): · Participant's self perceptions of behavior change (e.g. unstandardized pre/ post tests) · Collection of behavioral data from third parties to substantiate participant's self perceptions of behavior change (e.g. survey of supervisors, coworkers, subordinates, and/or correctional program specialists) · Development and piloting of knowledge/skill/attitude change tests in an effort to move toward the use of reliable and valid measures. · Use of research design that focuses on associations and relationships · Identification of both proximal and distal outcomes and moving toward distinguishment of proximal and distal outcomes.	Evaluation strategy based on at least one of the following (using instruments in the process of being validated): · Participant's self perceptions of organization change (e.g. unstandardized pre/ post tests) · Collection of organizational data from third parties to substantiate participant's self perceptions of organization change (e.g. survey of supervisors, coworkers, subordinates and/or correctional program specialists) · Development and piloting of organizational change instruments in an effort to move toward the use of reliable and valid measures (e.g. Action Leadership Plan). Such instruments to be designed by Correctional Program Specialists. · Use of research design that focuses on associations and relationships. · Identification of both proximal and distal outcomes and moving toward distinguishment of proximal and distal outcomes.	Evaluation strategy based on: · Conducting retrospective cost-benefit and/or cost-effectiveness analysis by estimating costs associated with program and comparing that to either data on cost savings pertaining to program benefits (cost benefit) or data on program outcomes besides cost savings.
Evaluation strategy based on participant's rating of the following immediately after the training and/or over a follow-up period: · Overall Evaluation of Training · Applicability of Training · Anticipated and Encountered Barriers to and Resources for application · Evaluation of Trainers (using validated instruments) · Completing tasks assigned during training (not same as behavioral change)	Evaluation strategy based on ratings of some combination of the following (using validated instruments) · Participant's self perceptions of knowledge/skill/attitude change (e.g. standardized, objective & reliable pre/ post tests) · Use of valid and reliable tests designed to detect actual rather than perceived change in knowledge/skill/attitude (pre/post/post2) · Collection of perceptual data from third parties to substantiate participant's self perceptions of knowledge/skill/attitude change (e.g. survey of supervisors, coworkers, subordinates, and/or correctional program specialists) · Correctional Program Specialist rating of participant knowledge/ skill/attitude change	Evaluation strategy based on ratings of some combination of the following (using validated instruments) · Participant's self perceptions of behavioral change (e.g. standardized pre/ post tests) · Collection of perceptual data using valid & reliable instruments designed to substantiate self reported behavior change (e.g. MLQ 360, MLQ 5X) as observed and reported by an independent third party (e.g. survey of supervisors, coworkers, subordinates, and/or correctional program specialists) · Use of research design to increase confidence causal assertions can be made (e.g., identification and elimination of threats to internal validity, control groups, etc.) · Identification and distinguishment of both proximal and distal outcomes. · Measurement of unintentional outcomes (e.g., focus groups, qualitative surveys) · Assessment of effect of attrition (i.e., data loss)	Evaluation strategy based on ratings of some combination of the following (using validated instruments): · Participant's self perceptions of organization change (e.g. standardized pre/ post tests) · Collection of organizational change data using valid and reliable instruments (e.g. JDI, OCQ, Prison Social Climate Survey) from third parties to substantiate participant's self perceptions of organization change (e.g. survey of supervisors, coworkers, other employees, and/or correctional program specialists) · Use of existing agency/organization records and statistics as indicators of organization change (e.g., turnover, absenteeism, etc.) · Use of research design to increase confidence causal assertions can be made (e.g.,identification and elimination of threats to internal validity, control groups, etc.) · Identification and distinguishment of both proximal and distal outcomes. · Measurement of unintentional outcomes via focus groups, qualitative surveys, etc. · Assessment of effect of attrition (i.e., data loss) Note: Assumption that the evaluator must have a close, colloborative relationship with the training participants' agency/organization.	Evaluation strategy based on: · Conducting pre and post cost-benefit and/or cost-effectiveness analysis by estimating costs associated with program and comparing that to either data on cost savings pertaining to program benefits (cost benefit) or data on program outcomes besides cost savings. · Comparing cost-benefit and cost-effectiveness of different program modalities (e.g., traditional versus web).

5) participant reaction; 6) participant learning; 7) learning transfer/behavior change; 8) organizational change; and 9) cost and efficiency, e.g., return on investment, added value, or cost/benefit. To better understand the goals of each type of evaluation, refer to the Guiding Questions column of the Evaluation Matrix on pages 2-3. These questions provide a means of conveying the extent and purpose of each type of evaluation. While each type of evaluation should be considered, some might not be appropriate for inclusion in the final evaluation plan.

Further, while evaluations at various degrees of rigor share some common characteristics, there are designs and strategies that make each unique. A ***basic evaluation,*** for example, may address all nine types of evaluation. Yet each would reflect a more relaxed standard of rigor than required for an intermediate or advanced evaluation. (See the "Basic" row of the Evaluation Matrix, pages 2-3.)

In such a study, a needs and viability assessment would be limited by the use of an unsystematic sample, non-validated instrumentation, and reliance on existing data sources. An evaluability assessment conducted at the basic level may be limited to a review of program documents/descriptions, or interviews of certain stakeholders to estimate the feasibly of further evaluation efforts. Process evaluations at this level would also be less rigorous. For instance, data regarding the extent to which the program is operating as planned would be based on attendance, participant satisfaction ratings, and the use of non-validated or self-reported learning measures. Outcome evaluations, if utilized, would likely consist of participants' self- ratings of individual or organizational change assessed using non-validated instruments. Cost/Efficiency evaluations would merely include self-ratings by stakeholders to estimate return on investment or cost/benefit, rather than relying on data collection to establish more objective measures.

On the other hand, more rigorous, ***advanced evaluations*** are warranted in some cases. However, as the degree of rigor increases, so does the complexity and cost of the overall evaluation. Depending on such factors as scope of the training, its intensity, length, and projected cost (in both monetary and human resources) as well as commitment of the participating agencies, a variety of advanced evaluations may be required. (See the "Advanced" row on pages 2-3.)

Initially, thorough needs, viability and evaluability assessments may be required, given the importance of identifying training needs as well as the extent to which the training program lends itself to evaluation. Following this, the evaluation may need to include a thorough process evaluation to collect and analyze a large amount of target audience and program implementation data, as well as data on participant reaction to training and any change in knowledge, skills, or attitudes. Additionally, outcome evaluations may be needed to gather objective data on the extent to which the training influenced any behavior change of the participants or the organizations to which they belong. Last, advanced evaluations will likely also conduct pre and post cost-benefit and/or cost-effectiveness analyses by measuring costs associated with the program and comparing that to data on cost savings.

Two key characteristics of advanced, rigorous evaluations pertain to the sources of the data and how it is analyzed. First, in addition to data collected from training participants, advanced evaluations may also collect multiple sources of behavioral and organizational change data from participants' co-workers, subordinates, superiors and other employees. In addition, advanced evaluations could also utilize existing agency/organization records and statistics as indicators of organization change (e.g., turnover, absenteeism, etc.). Second, rather than simply describing the perceived benefits of the training to those providing data, advanced evaluations often employ more complex research designs in an effort to demonstrate that: a) the training caused certain outcomes, and b) the outcomes generalize (or apply) beyond those who provided data or participated in the program.

Finally, a few caveats about the Evaluation Matrix are in order. First, typologies of this kind are most easily distinguished on paper. In the real world of evaluation, there will often be blending across types of evaluation and degrees of rigor, coupled with confusion and ambiguity over what fits where. The matrix is meant to guide discussion and thinking about evaluation design; it is not a rigid protocol for compartmentalizing activities. Second, the matrix is geared mainly toward design, measurement, and analysis with less attention to work like data entry, project management, and reporting. However, the latter is just as important and grows more time consuming and expensive as the scope and rigor of the evaluation increases. Third, rather than being a finished product, this matrix is a work in progress, subject to continual refinement based on feedback and use. An important goal in this respect is to make it understandable and useful to persons not familiar with evaluation research.

While the matrix reflects the body of literature on evaluation research in general and training evaluation in particular, it is principally grounded in the seminal book *Evaluation: A Systematic Approach* (7th ed.) by Peter H. Rossi, Mark W. Lipsey, and Howard E. Freeman (Thousand Oaks, CA: Sage Publications, 2004). Insights were also drawn from the work of Donald Kirkpatrick (training evaluations), Jack Phillips (return on investment), and Paul Kearns (added value).

Acknowledgements

The National Institute of Corrections Training Evaluation Project was made possible by the support of NIC via Cooperative Agreements 05A28GJF9, 06PEI01GJM1, 07PEI12GJQ7, and 08PEI21GJX1.

CwRC staff wish to acknowledge the support and cooperation of the many persons who helped make this project possible. Morris Thigpen, Larry Solomon, Tom Beauclair, Chris Innes, Dee Halley, Bob Brown, John Eggers, Leslie LeMaster, Launa Kowalcyk, Virginia Hutchinson, Belinda Watson, Fran Zandi, Cheryl Paul, Robbye Braxton-Mintz, Rob Jeffreys and others at NIC have provided essential support for this project. We also wish to acknowledge our support staff, whose daily efforts further the project in so many ways. Finally, we want to express our appreciation to the growing number of NIC trainers and training participants who have taken time out of their busy schedules to graciously share their insights with us.

Although many persons and organizations contributed to the development of this evaluation matrix, any errors or omissions are those of the authors alone.

The views expressed in this document are those of the authors and do not necessarily reflect the positions or policies of the National Institute of Corrections, or any other organization or individual.

Commonwealth Research Consulting, Inc.
Basic Evaluation Matrix

CwRC–NIC Basic Evaluation Matrix
January 2010
(Draft 2)

National Institute of Corrections Training Evaluation Project

NIC Evaluation Matrix (simplified version for basic evaluations)

James B. Wells, Ph.D.
Kevin I. Minor, Ph.D.
Stephen Parson, M.S.
Adam K. Matz, M.S.

What is the Basic Evaluation Matrix?

The Basic Evaluation Matrix is a tool designed to assist NIC staff and trainers, as well as other agencies or evaluators, in selecting the appropriate types of evaluation for a given program or training situation. The matrix provides a brief description of five evaluation types including: needs and viability assessment; evaluability assessment; process evaluation; outcome evaluation; and cost/efficiency evaluation. The purpose of each type of evaluation is listed in the top row of the matrix. Guiding questions are provided to further clarify which evaluation type may be appropriate. The bottom row of the matrix provides a brief description of activities associated with each type of evaluation.

While some situations call for one or more of these evaluations to be conducted in a scientifically rigorous manner, such advanced evaluations are frequently cost prohibitive or otherwise not feasible. However, in most cases evaluations can be quite beneficial, without being scientifically rigorous. The procedures described in the Basic Evaluation Matrix utilize a more relaxed or basic standard of rigor in the conduct of evaluations. The reduced cost and complexity of basic evaluations provide viable options for organizations that don't have access to research personnel or substantial evaluation budgets. Findings from basic evaluations can provide significant insights for program improvements, evidence of program effectiveness, and support for program continuation.

In situations where more rigorous or advanced evaluations may be necessary and feasible, refer to the full Evaluation Matrix developed by Wells, Minor, and Parson (2007). Advanced evaluations may be necessary to provide more detailed findings, track changes over time, or to increase confidence in the extent to which evaluation findings are reliable, valid, generalizable, and replicable.

This project is made possible by support from the National Institute of Corrections. However, the authors are solely responsible for the content of this document, including all errors and omissions.

How Does it Work?

To use the Basic Evaluation Matrix, first, determine the goal and purpose of your evaluation, i.e., what you want to learn from the evaluation. For example: Is there a need for the program? Did training participants like the training? Did the program achieve the intended outcomes? Did the benefits of a training outweigh the costs? With these questions in mind, use the purpose and guiding questions of the matrix to locate the appropriate type(s) of evaluation. Next, review the descriptions of basic evaluation procedures for that evaluation type (bottom row) to help you determine what data to collect, where to get the data, and how to collect it. For example, you may collect satisfaction data from training participants with a brief written survey, or you may review documents and conduct interviews with personnel from finance and human resources to establish the total cost of a program.

Other Considerations

The matrix is a tool meant to foster discussion and initiate evaluation development. It is not a step-by-step recipe for evaluation in a particular context. Other activities such as data entry, project management, and reporting must also be considered when developing an evaluation agenda. Further guidance in conducting evaluations may be available from the full version of the Evaluation Matrix, research or evaluation personnel at your organization, outside consultants, or one of the many evaluation texts available.

The matrix is a work in progress, subject to continual refinement based on feedback and use. An important goal in this respect is to make it understandable and useful to persons not familiar with evaluation research. The authors welcome feedback on this document. Comments can be sent to:

Dr. James B. Wells, 4160 Kentucky River Parkway, Lexington, KY 40515 or jbwells@cwrc.us

While the matrix reflects the body of literature on evaluation research in general and training evaluation in particular, it is principally grounded in the seminal book *Evaluation: A Systematic Approach* (7th ed.) by Peter H. Rossi, Mark W. Lipsey, and Howard E. Freeman (Thousand Oaks, CA: Sage Publications, 2004). Insights were also drawn from the work of Donald Kirkpatrick (training evaluations), Jack Phillips (return on investment), and Paul Kearns (added value).

Simplified Matrix for Basic Evaluations*

Evaluation Type:	Needs/Viability Assessment	Evaluability Assessment	Process Evaluation	Outcome Evaluation	Cost/Efficiency Evaluation
Purpose:	To identify the nature and extent of needs to be addressed by the program, and to assess the likelihood that the proposed program will meet those needs.	To assess the extent to which a program lends itself to evaluation, and the likelihood that the evaluation will produce useful information to help improve the program.	To assess the extent to which a program is implemented as designed, reaches the target audience, and proves satisfactory and beneficial from the perspective of the participant.	To determine the extent to which the program is producing the intended outcomes, or having the desired effects.	To assess and compare the total costs and benefits of a program, the return on investment, and/or added value.
Guiding Questions:	Who is the target population for the program, and what are their needs?	Are stakeholders willing and able to support evaluation efforts? Are program goals clear, measurable, and feasible?	To what extent is the target population being served by the program as intended?	To what extent did the program have the intended outcomes with respect to promoting individual behavioral change and/or organizational change?	Where desired outcomes have been demonstrated, to what extent do the returns from the program exceed its cost?
Basic Evaluation Procedures:	Review available data (e.g., from agency records, government reports, or social indicators that estimate the size of a problem, etc.) or collect new data to measure the perceived need.	Review program documents, interview program personnel and stakeholders, ensure that all agree on the program objectives, components, design of evaluation, source of resources, and ensure there is adequate cooperation and collaboration in developing and conducting the evaluation.	Compare the demographic profile of participants with that of the intended target audience. Measure the extent to which participants' feel training or program objectives were covered. Measure participants' satisfaction with the training or program as well as participants' self-rating of improvement (e.g., knowledge, skills, attitudes, etc.) as a result of the training or program.	Measure participants' self-rating of behavioral and organizational change after the program's completion, typically six months to one year after the program has concluded.	Measure or document the cost of conducting the program and compare it with program outcomes. Measure multiple stakeholders' ratings of benefits versus costs of the program.

* For more detailed evaluation guidance please consult the full version of the NIC Evaluation Matrix, or contact your organization's research division.