ELSEVIER

Evaluation and Program Planning 27 (2004) 341–347

EVALUATION and PROGRAM PLANNING

www.elsevier.com/locate/evalprogplan

A critical analysis of evaluation practice: the Kirkpatrick model and the principle of beneficence

Reid Bates

Louisiana State University, Baton Rouge, LA, USA

Abstract

This chapter describes Kirkpatrick's four-level training evaluation model and the reasons for its popularity in organizations. Several fundamental limitations of the model are outlined and the potential risks these limitations raise for evaluation clients and stakeholders are discussed. It is argued that these risks, plus the inability of the model to effectively address both the summative question (Was training effective?) and the formative question (How can training be modified in ways that increase its potential for effectiveness?), limits the capacity of training and HRD professionals to fulfill the core ethical duty of beneficence.

Keywords: Kirkpatrick model; Training; Casual linkage; Training evaluationz

The models used to guide evaluations bear a close relationship to the effectiveness and utility of those evaluations. Inevitably there are also ethical dimensions to the models and the ways in which they are used. It is therefore important to subject our models to ongoing reflection and analysis from different perspectives. Such reflection should help answer a fundamental ethical question about evaluations: "Are we doing the right thing, and are we doing it well?" (Schwandt, 1998, p. 11). This chapter highlights some of the limitations of Kirkpatrick's training evaluation model and points to several risks for clients and stakeholders associated with the model and its assumptions. The analysis raises questions about the extent to which the model is consistent with the principle of beneficence or providing benefits to clients and stakeholders when the opportunity to do so is present.

1. The four level approach to training evaluation

By far the most popular approach to the evaluation of training in organizations today is Kirkpatrick's (1976) framework of four 'levels' of criteria. Kirkpatrick's (1976, 1994) training evaluation model delineates four levels of training outcomes: reaction, learning, behavior, and results. Level one includes assessment of training participants' reaction to the training program. Kirkpatrick (1959) originally discussed reactions in terms of how well participants liked a particular program. In practice, measures at this level have evolved and are most commonly directed at assessing trainees' affective responses to the quality (e.g. satisfaction with the instructor) or the relevance (e.g. work-related utility) of training. Learning measures, level two, are quantifiable indicators of the learning that has taken place during the course of the training. Level three behavior outcomes address either the extent to which knowledge and skills gained in training are applied on the job or result in exceptional job-related performance. Finally, level four outcomes are intended to provide some measure of the impact that training has had on broader organizational goals and objectives. In recent practice, the typical focus of these measures has been on organizational level financial measures.

1.1. The popularity of the four level model

The Kirkpatrick model has served as the primary organizing design for training evaluations in for-profit organizations for over thirty years. The overwhelming popularity of the model can be traced to several factors. First, the model addressed the need of training professionals to understand training evaluation in a systematic way (Shelton & Alliger, 1993). It has provided straightforward system or language for talking about training outcomes and the kinds of information that can be provided to assess the extent to which training programs have achieved certain objectives.

Second, Kirkpatrick insisted that information about level four outcomes is perhaps the most valuable or descriptive

0149-7189/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.evalprogplan.2004.04.011

information about training that can be obtained. For training professionals in organizations this bottom-line focus is seen as a good fit with the competitive profit orientation of their sponsors. The four-level model has therefore provided a means for trainers in organizations to couch the results of what they do in business terms. Many see this as critical if the training function is to become a true business partner and be seen as an active contributor to organizational success.

Finally, the popularity of the four-level model is also a function of its potential for simplifying the complex process of training evaluation. The model does this in several ways. First, the model represents a straightforward guide about the kinds of questions that should be asked and the criteria that may be appropriate. Second, the model reduces the measurement demands for training evaluation. Since the model focuses the evaluation process on four classes of outcome data that are generally collected *after* the training has been completed it eliminates the need for—or at least implies—that pre-course measures of learning or job performance measures are not essential for determining program effectiveness. In addition, because conclusions about training effectiveness are based solely on outcome measures, the model greatly reduces the number of variables with which training evaluators need to be concerned. In effect, the model eliminates the need to measure or account for the complex network of factors that surround and interact with the training process.

There is no doubt that Kirkpatrick's model has made valuable contributions to training evaluation thinking and practice. It has helped focus training evaluation practice on outcomes (Newstrom, 1995), fostered the recognition that single outcome measures cannot adequately reflect the complexity of organizational training programs, and underscored the importance of examining multiple measures of training effectiveness. The model promoted awareness of the importance of thinking about and assessing training in business terms (Wang, 2003). The distinction between learning (level two) and behavior (level three) has drawn increased attention to the importance of the learning transfer process in making training truly effective. The model has also served as a useful—if preliminary—heuristic for training evaluators (Alliger & Janak, 1989) and has been the seed from which a number of other evaluation models have germinated (e.g. Holton, 1996; Jackson & Kulp, 1978; Kaufman & Keller, 1994).

1.2. Limitations of the four-level model

There are at least three limitations of Kirkpatrick's model that have implications for the ability of training evaluators to deliver benefits and further the interests of organizational clients. These include the incompleteness of the model, the assumption of causality, and the assumption of increasing importance of information as the levels of outcomes are ascended.

1.2.1. The model is incomplete

The four-level model presents an oversimplified view of training effectiveness that does not consider individual or contextual influences in the evaluation of training. A broad stream of research over past two decades (Cannon-Bowers, Salas, Tanenbaum, & Mathieu, 1995; Ford & Kraiger, 1995; Salas & Cannon-Bowers, 2001; Tannenbaum & Yukl, 1992) has documented the presence of a wide range of organizational, individual, and training design and delivery factors that can influence training effectiveness before, during, or after training. This research has led to a new understanding of training effectiveness that considers 'characteristics of the organization and work environment and characteristics of the individual trainee as crucial input factors' (Cannon-Bowers, Salas, & Tannenbaum, 1995, p. 143). For example, contextual factors such as the learning culture of the organization (Tracy, Tannenbaum, & Kavanaugh, 1995), organizational or work unit goals and values (Ford, Quinones, Sego, & Sorra, 1992), the nature of interpersonal support in the workplace for skill acquisition and behavior change (Bates, Holton, Seyler, & Carvalho, 2000) the climate for learning transfer (Rouiller & Goldstein, 1993), and the adequacy of material resources such as tools, equipment, and supplies have been shown to influence the effectiveness of both process and outcomes of training. Kirkpatrick's model implicitly assumes that examination of these factors is not essential for effective evaluation.

1.2.2. The assumption of causal linkages

Kirkpatrick's model assumes that the levels of criteria represent a causal chain such that positive reactions lead to greater learning, which produces greater transfer and subsequently more positive organizational results. Although Kirkpatrick is vague about the precise nature of the causal linkages between training outcomes, his writings do imply that a simple causal relationship exists between the levels of evaluation (Holton, 1996). In one of Kirkpatrick's more recent publications he states that "if training is going to be effective, it is important that trainees react favorably" (Kirkpatrick, 1994, p. 27), and that "without learning, no change in behavior will occur" (p. 51). Research, however, has largely failed to confirm such causal linkages. Two meta-analyses of training evaluation studies using Kirkpatrick's framework (Alliger & Janak, 1989; Alliger, Tannenbaum, Benett, Traver, & Shotland, 1997) have found little evidence either of substantial correlations between measures at different outcome levels or evidence of the linear causality suggested by Kirkpatrick (1994).

1.2.3. Incremental importance of information

Kirkpatrick's model assumes that each level of evaluation provides data that is more informative than the last (Alliger & Janak, 1989). This assumption has generated the perception among training evaluators that establishing level four results will provide the *most useful* information about training program effectiveness. In practice, however,

the weak conceptual linkages inherent in the model and resulting data it generates do not provide an adequate basis for this assumption.

1.3. Ethical implications

Training practitioners in organizations as members of the human resource development (HRD) profession have taken on the obligation to advance the welfare of individuals and organizations for whom they work. As such, we have created the expectation among our clients and stakeholders that they will benefit from our efforts. It follows from this that the ethical principle of beneficence—doing good for others—is or should be a central ethical concern. In other words, we have an obligation to "think well and wisely about what it means to benefit others…" (Kitchener, 1984, p. 43). This obligation is ethically binding and is consistent with the core ethical principle of beneficence. The following sections describe the principle of beneficence and how it may be applied to training evaluation. An argument is made that the limitations of the four-level model may prevent evaluators from adequately addressing this principle.

2. The concept of beneficence

Along with autonomy, nonmaleficence, justice and fidelity, beneficence has been advanced as one of the five prima facie ethical principles of helping professions (Beauchamp & Childress, 1983). These core principles have been applied to various helping professions including health care delivery (Beauchamp & Childress, 1983), counselling psychologists (Kitchener, 1984), student affairs administrators in higher education (Upcraft & Poole, 1991), student services advisors (Brown & Krager, 1985), and more recently program evaluation (Newman & Brown, 1996). Beneficence can be defined as the quality of doing good, taking positive steps to help others, or the notion that one ought to do or promote action that benefits others. In its most general form, this principle asserts that it is an ethical duty to "help others further their important and legitimate interests" (Beauchamp & Childress, 1983, p. 148) and to confer benefits on clients, sponsors, and stakeholders when possible. The specific actions required to conform to the principle of beneficence derive from the relationships and commitments that are associated with the institutional roles that people play. For example, many now believe that those assuming the role of training evaluators should help organizations further their important and legitimate interests by striving to fulfill two fundamental evaluation goals: determining (a) if the program was effective and (b) what can be done to improve the training process.

Beauchamp and Childress (1983) suggest that the concept of beneficence goes beyond the principle of nonmaleficence (do no harm) and argue that the failure to provide a benefit when in a position to do so represents a violation of professional ethics when certain conditions are met. Newman and Brown (1996) have applied these conditions to the relationship between an evaluator and his or her client and stakeholders to suggest that evaluators have an obligation of beneficence when:

1. Clients, stakeholders, or other participants in an evaluation are at risk of significant loss or damage.
2. Action on the part of the evaluators is required to prevent the loss or damage.
3. The action of the evaluator would probably prevent the loss or damage.
4. The evaluator's action would not present a significant risk of harm to the client or stakeholder.
5. The benefits that clients, stakeholders, or other participants will accrue outweigh the harm that the evaluator is likely to suffer.

2.1. Kirkpatrick's model and the principle of beneficence

The principle of beneficence can be applied to ethical problems in training evaluation when questions arise about the risks and benefits of certain actions. Risks refer to potential future harm or loss and can address both the probability that a harm or loss will occur as well as its magnitude. In the context of training evaluation, risks for clients and stakeholders can result from any number of factors ranging from the failure to protect participant rights to the inability of the evaluation to convey accurate information about the object of study. Examining the potential risks associated with the evaluation models such as Kirkpatrick's can provide some insight into the capacity of those models to foster practice consistent with the principle of beneficence.

Benefits typically refer to some process, activity, or information that promotes the welfare and interests of clients and stakeholders (Beauchamp & Childress, 1983). Clearly the range of potential benefits from training evaluation is broad and varied. In general, however, the benefits that can be derived from evaluation are directly related to the capacity of the evaluation to develop information that increases the clarity of judgment and reduces the uncertainty of action for specific clients and stakeholders (Patton, 1997). For training evaluations in organizations, this is best done when descriptive and judgmental information is systematically collected that (a) assesses program effectiveness, and (b) helps improve the program relative to its goals (Goldstein & Ford, 2002; Holton, 1996; Swanson & Holton, 2000). Thus one general way to examine the potential benefits of Kirkpatrick's model as a guide to training evaluation is to analyze the extent to which it furthers or limits the ability to answer both the summative question (Was training effective?) and the formative question (How can the training process be modified in ways that increase its potential for effectiveness?).

The limitations of Kirkpatrick's model noted earlier carry with them some meaningful implications for risks and benefits that may accrue to clients and stakeholders in the evaluation process. For example, the failure of the model to promote consideration of key contextual input variables in training evaluation masks the real complexities of the training and evaluation process. With its exclusive focus on training outcomes, the model promotes the view that if training evaluators measure one or more of the four levels of outcomes then this will provide adequate evaluative information. In other words, the training program itself is presumed to be solely responsible for any outcomes that may or may not ensue. The problem with this approach is that, although it may provide some beneficial information about program outcomes (given the effective selection and measurement of criteria), when measurement is restricted to one or more of the four criterion levels no formative data about why training was or was not effective is generated (Goldstein & Ford, 2002). For example, the implicit assumption of the four-level model is that all of the work contexts to which trainees return will have equal effects on learning transfer. Such an assumption makes it very difficult to draw conclusions about training effectiveness if trainees all had similar reactions to training and displayed similar levels of learning yet exhibited wide variations in learning transfer. It is unclear, in the absence of more contextual information, whether the training program was not designed in ways that fostered effective transfer or whether other input factors blocked skill application.

When key input factors are not taken into account, the potential for misleading or inaccurate judgments about the merit and effectiveness of training increases. Such judgments represent significant risks to clients and stakeholders when, for instance, they lead to the cancellation of useful training programs or the promulgation of ineffective programs. Health, safety or other organizational performance problems may persist if effective training programs are mistakenly discontinued. Training resources may be wasted because ineffective programs are continued or because program improvement efforts are based on incomplete or misleading data. When misleading or inaccurate evaluations lead to poor decisions about training effectiveness, the credibility and effectiveness of future training efforts is undermined. It is also of general benefit to clients and stakeholders to know whether a training program's success or failure is a function of contextual factors such as proper equipment, adequate resources, organizational culture, performance consequences, managerial expectations and support, or other key input factors. In the absence of such information, organizations have little foundation upon which to make decisions about the extent to which the results of the same or a similar training program can be achieved with other individuals in other units or departments.

There is evidence that the causal linkage assumption, although not supported by research, is still popularly embraced and has spawned several evaluation practices that present potential risks to evaluation clients and stakeholders. First, this assumption has bred the perception that reaction measures can be used as legitimate surrogates or proxy measures for training outcomes in other domains (e.g. learning or behavior change). This perception has contributed to the overuse of reaction measures as the sole evaluative measures of training effectiveness. For example, over 94% of business organizations evaluate training using reaction measures (Bassi, Benson, & Cheney, 1996) while far fewer efforts are made to collect information about learning or behavior change on the job (Goldstein & Ford, 2002). In fact, most training evaluations in organizations has historically focused on collecting *only* reaction measures (Grider, Capps, & Toombs, 1988). Unfortunately, inferring some connection between reaction measures or learning measures and unmeasured outcomes at other levels will most likely furnish an incomplete, misleading view of training effectiveness and a distorted foundation upon which to make training decisions. Simply put, the causal linkage assumption has fostered a narrow focus on reaction measures that are insufficiently comprehensive to support credible judgements about the merit of training programs or what might be done to improve them.

The causal linkage assumption and the over-reliance on reaction measures also diverts trainers' attention away from efforts to make training truly effective to a focus on developing entertaining, amusing, and easy-going training that participants find enjoyable. It is often easier to develop a training program that will elicit positive reactions from participants than one that will lead to true learning and behavior change on the job. This presents a significant hazard for organizations to the extent it fosters a view that "having a good time becomes the mark of excellence [in training], a valuing of entertainment over education" (Michalski & Cousins, 2000, p. 249) or behavior change on the job. It also ignores the fact that learning is often difficult, and that effective learning often challenges participants to the point that they may experience training as uncomfortable (Knowles, Holton, & Swanson, 1998; Rodin & Rodin, 1972). In addition, the pressure on trainers to rely solely on reaction measures may increase as the cost or importance of a training program increases. Because trainers view the evaluation of training as parallel to the evaluation of their own performance (Michalski & Cousins, 2000), the level of personal risk rises substantially when a costly or high profile program is evaluated at higher level outcomes that are often more difficult to achieve than 'good' participant reactions. Evaluation clients and stakeholders are also in greater jeopardy in these situations because of the increased potential for loss or damage that may ensue from the misleading or inaccurate information that comes when reaction measures are used instead of or as surrogates for higher level outcomes.

The linkage between individual-level training outcomes and organizational outcomes is at best complex and difficult

to map and measure even when training is purposely designed to address organizational objectives. Nevertheless establishing this linkage through the collection of level four data—particularly organizational level financial data—is often portrayed as crucially important. The assumption that level four outcome measures in the form of organizational level financial information is the most useful for evaluating training has generated at least three elements of substantial risk for clients and stakeholders in the training process. First, it has fostered the belief among trainers and training evaluators that clients and stakeholders want to judge the utility of training using the same financial metrics they would use to judge other organizational initiatives. This conviction has generated a variety of efforts to apply financial formulas to calculate the business impact of training in dollar terms such as return-on-investment (ROI), reduced organizational costs, or other financial measures (e.g. Geber, 1995; Tesoro, 1998; Wang, 2003). Training professionals are drawn to this kind of evaluation because it is seen as a way to gain more power and influence in organizations (Hall and Goodale, 1986) and to shape management's view that the training function is "happily revenue producing instead of resource draining" (Alliger & Janak, 1989, p. 333). Aside from the complex causal connections linking individual learning to organizational outcomes, the tendency has been to disregard the fact that most training is not designed to directly address organizational level goals. This results in evaluations that provide relatively little valid data about training effects or which may grossly miscalculate and inflate the influence and value of training. In addition, level four data does not provide any kind of prescriptive information about training. Evaluations at this level virtually ignore formative data that could be used to improve training structures and process. In effect, a preoccupation with level four financial measures such as ROI is often self-serving and can divert attention from a focus on improving training. It also impairs the ability of organizations to develop training systems that are capable of continuous and ongoing improvement.

Second, the evaluation data generated by the Kirkpatrick model does not provide a reasonable warrant for validly making the conclusion that training causes changes in outcomes at the organizational level. In fact, the assumption that level four data is the most useful increases the risk of misleading or inaccurate results because (a) training initiatives rarely directly target costs or other financial metrics at the organizational level; (b) any number of factors can confound changes in financial and other performance measures at the organizational level; and (c) most training efforts have little capacity to directly affect level four criteria. For example, most training programs in organizations are of short or modest duration (e.g. 1 or 2 days) and are meant to have only a limited impact on the participants involved. Given the circumscribed impact of training and what we now know about the complex of factors that can influence training effectiveness, the evidentiary linkage from training to organizational results is likely to be too weak to be detected using the inferential chain provided by Kirkpatrick's model. The training process and the organizational context are just too complex and the causal linkages too poorly specified in Kirkpatrick's model to provide reasonable evidence or proof of conclusions of this nature. Results might indicate program success where there is none or point to program failure when the desired outcomes will take time to develop. In short, the question of whether training is solely responsible for skill acquisition and, by extension, organizational productivity or financial performance is "just too broad and imprecise; it neglects a myriad of potentially important intervening variables and contingencies" (Baldwin & Danielson, 2002, p. 27).

Finally, recent research by Michalski and Cousins (2000) suggests that different clients and stakeholder groups in organizations have appreciably divergent views about training outcomes and what is important to measure in assessing training effectiveness. Kirkpatrick's assumption that level four data are the most useful ignores the potential perceptual and expectation differences about training and training outcomes that may exist among key stakeholders groups (e.g. trainees, managers, trainers) in organizations. The risk for organizations in this approach is the potential it has to undermine what is one of the main consequences of training evaluation: the utilization of evaluative findings. Training evaluations that do not return the knowledge and information deemed most useful by key clients and stakeholders are unlikely to be used (Patton, 1997).

3. Discussion and conclusion

HRD as a profession is deeply concerned about issues of individual and organizational learning, change and success, and how, through its professional roles and activities, it can benefit people and organizations as they pursue various goals and interests. This intrinsic commitment means we have a driving obligation to use HRD tools and processes—including the selection and use of evaluation models and frameworks—in a manner that is ethically thoughtful and sound. Although there have been a number of conceptual and methodological criticisms of Kirkpatrick's model, few have evaluated the model from an ethical perspective. The goal of this chapter was to reflect on the assumptions and use of Kirkpatrick's model in an effort to respond to a fundamental ethical question about training evaluations: "Are we doing the right thing, and are we doing it well?" The principle of beneficence was used as a point of departure for addressing these questions. The principle of beneficence is an affirmative ethic in the sense that it demands that actions must be taken *when the opportunity arises* to actively contribute to the health and welfare of clients and stakeholders.

With regard to the first question, are we doing the right thing, there is little question that much of the heightened

interest in training evaluation and the increased incidence in the evaluation of training programs can be attributed to the simplicity and popularity of Kirkpatrick's model. In addressing the second question, this chapter has highlighted some of the limitations of Kirkpatrick's model. It has pointed to several risks for clients and stakeholders that can be generated from acceptance of the model and its assumptions. In addition, the inability of the model to effectively address both the summative question (Was training effective?) and the formative question (How can training be modified in ways that increase its potential for effectiveness?) suggests that potential benefits of the model may not outweigh the risks.

Along these same lines, it is also important to note that the principle of beneficence suggests that training evaluators may be at ethical risk if they fail to take advantage of advances in training evaluation methods, models, and tools when the opportunity arises. This proactive dimension of beneficence is well established in other helping professions. For instance, in medicine when methods of preventing certain diseases were discovered, there was universal agreement that not taking positive steps to provide this benefit through preventive programs would be immoral. Similarly gene therapy research is often justified on the basis of its potential to provide a more beneficial therapy than other alternatives (Beauchamp & Childress, 1983). This perspective suggests that training evaluators have an ethical obligation to improve their models and practice in ways that will enhance the capacity to more meaningfully benefit clients and stakeholders. Fortunately, research into training effectiveness over the last decade has generated new insights and has led to the development of more complete models of evaluation that have the potential to more effectively assess training outcomes and provide information needed to improve the training process (e.g. Holton, 1996). In addition, a number of valid, reliable, and easy-to-use assessment scales and instruments are available that complement such models and can help training evaluators examine a range of key input variables into the training process. For instance, recent research has led to the development of instruments measuring key pre-training factors (Weinstein et al., 1994), factors affecting learning transfer (Holton, Bates, & Ruona, 2000), and other contextual factors influencing training effectiveness (e.g. Tracy et al., 1995). Kraiger, Ford and Salas (1993) have forwarded a multidimensional model of learning outcomes from training and have described a process training evaluators could use to developing learning evaluation measures (Kraiger & Jung, 1997). Kraiger, Salas, and Cannon-Bowers (1995) developed and used a method for the assessment of an individual trainee's domain-specific knowledge and skills. Other researchers have provided tools for more accurately assessing the multidimensionality of participant reaction measures (Morgan & Casper, 2000) and models for thinking more clearly about the multiple dimensions of job performance (Campbell, McHenry, & Wise, 1990; Cannon-Bowers & Salas, 1997).

Training evaluators have an ethical obligation to "think wisely about what it means to benefit others" and to reflect on and evaluate the extent to which the models used in practice can help provide that benefit. The point of this chapter has not been to make a case that training evaluations using Kirkpatrick's model are unethical. However, it does raise questions about the extent to which the model benefits clients and stakeholders and suggests that consideration of these issues from an ethical perspective may be overdue.

References

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: thirty years later. *Personnel Psychology*, *42*, 331–342.

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

Baldwin, T. T., & Canielson, C. C. (2002). Invited reaction: Linking learning with financial performance. *Human Resource Development Quarterly*, *31*(1), 23–29.

Bassi, L. J., Benson, G., & Cheney, S. (1996). *Trends: Position yourself for the future*. Alexandria, VA: ASTD.

Bates, R. A., Holton, E. F., III, Seyler, D. A., & Carvalho, M. A. (2000). The role of interpersonal factors in the application of computer-based training in an industrial setting. *Human Resource Development International*, *3*(1), 19–43.

Beauchamp, T. L., & Childress, J. F. (1983). *Principles of biomedical ethics* (2nd ed). New York: Oxford University Press.

Brown, R. D., & Krager, L. (1985). Ethical issues in graduate education. *Journal of Higher Education*, *56*(4), 403–418.

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling the performance in a population of jobs. *Personnel Psychology*, *43*, 313–333.

Cannon-Bowers, J. A., & Salas, E. (1997). A framework for developing team performance measures in training. In M. T. Brannick, E. Salas, & C. Prince (Eds.), *Assessment and measurement of team performance* (pp. 45–62). Mahwah, NJ: Lawrence Erlbaum.

Cannon-Bowers, J. A., Salas, E., Tannenbaum, S. I., & Mathieu, J. E. (1995). Toward theoretically based principles of training effectiveness: a model and initial empirical investigation. *Military Psychology*, *7*, 141–164.

Ford, J. K., & Kraiger, K. (1995). The application of cognitive constructs and principles to the instructional systems design model of training: Implications for needs assessment, design, and transfer. *International Review of Industrial and Organizational Psychology*, *10*, 1–48.

Ford, J. K., Quinones, M., Sego, D., & Sorra, J. (1992). Factors affecting the opportunity to use trained skills on the job. *Personnel Psychology*, *45*, 511–527.

Geber, B. (1995). Does your training make a difference? Prove it!. *Training*, 27–34.

Goldstein, I. L., & Ford, J. K. (2002). *Training in organizations*. Belmont, CA: Wadsworth.

Grider, D. T., Capps, C. J., & Toombs, L. A. (1988). Evaluating valuations. *Training and Development Journal*, *42*, 11–12.

Holton, E. F., III (1996). The flawed four level evaluation model. *Human Resource Development Quarterly*, *7*(1), 5–21.

Holton, E. F., III, Bates, R. A., & Ruona, W. E. A. (2000). Development and validation of a generalized learning transfer system inventory. *Human Resource Development Quarterly*, *11*(4), 333–360.

Jackson, S., & Kulp, M. J. (1978). Designing guidelines for evaluating the outcomes of management training. In R. O. Peterson (Ed.), *Determining the payoffs of management training* (pp. 1–42). Madison, WI: ASTD.

Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: A guide to human resource development*. New York: McGraw Hill.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of ASTD*, *11*, 1–13.

Kirkpatrick, D. L. (1994). *Evaluating training programs: the four levels*. San Francisco: Berrett-Koehler.

Kitchner, K. S. (1984). Intuition, critical evaluation and ethical principles: the foundation for ethical decisions in counseling psychology. *The Counseling Psychologist*, *12*(3), 43–56.

Knowles, M. S., Holton, E. F., III, & Swanson, R. A. (1998). *The adult learner* (5th ed). Houston: Gulf.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, *78*(2), 311–328.

Kraiger, K., & Jung, K. M. (1997). Linking training objectives to evaluation criteria. In M. A. Quinones, & A. Ehrenstein (Eds.), *Training for a rapidly changing workplace* (pp. 151–175). Washington, DC: American Psychological Association.

Kraiger, K., Salas, E., & Cannon-Bowers, J. A. (1995). Measuring knowledge organization as a method for assessing learning during training. *Human Performance*, *37*, 804–816.

Jackson, S., & Kulp, M. J. (1978). Designing guidelines for evaluating the outcomes of maangemetn training. In R. O. Petersen (Ed.), *Determining the payoffs of management training* (pp. 1–42). Madison, WI: ASTD.

Michalski, G. V., & Cousins, J. B. (2000). Differences in stakeholder perceptions about training evaluation: a concept mapping/patter matching investigation. *Evaluation and Program Planning*, *23*, 211–230.

Morgan, R. B., & Casper, W. (2000). Examining the factor structure of k participant reaction to training: a multidimensional approach. *Human Resources Development Quarterly*, *11*, 301–317.

Newman, D. L., & Brown, R. D. (1996). *Applied ethics in program evaluation*. Thousand Oaks, CA: Sage.

Newstrom, J. W. (1995). Review of Evaluating training programs: the four levels by D.L. Kirkpatrick. *Human Resource Development Quarterly*, *6*, 317–319.

Patton, M. Q. (1997). *Utilization-focused evaluation* (3rd ed). Thousand Oaks, CA: Sage.

Rodin, M., & Rodin, B. (1972). Student evaluations of teachers. *Science*, *177*, 1164–1166.

Rouiller, J. Z., & Goldstein, I. L. (1993). The relationship between organizational transfer climate and positive transfer of training. *Human Resource Development Quarterly*, *4*(4), 377–390.

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology*, 471–497.

Schwandt, T. A., (1998). *How we think about evaluation practice*. Paper presented at the American Evaluation Association Conference, Chicago, IL.

Shelton, S., & Alliger, G. M. (1993). Who's afraid of level 4 evaluation? A practical approach. *Training and Development Journal*, *47*, 43–46.

Swanson, R. A., & Holton, E. F., III (2001). *Foundations of human resource development*. San Francisco: Berrett-Koehler.

Tannenbaum, S. I., & Yukl, G. (1992). Training and development in work organizations. *Annual Review of Psychology*, *43*, 399–441.

Tesoro, F. (1998). Implementing an ROI measurement process at Dell computer. *Performance Improvement Quarterly*, *11*(4), 103–114.

Tracy, J. B., Tannenbaum, S. I., & Kavanaugh, M. J. (1995). Applying trained skills on the job: The importance of work environment. *Journal of Applied Psychology*, *80*, 239–252.

Upcraft, M. P., & Poole, T. G. (1991). Ethical issues and administration politics. In P. L. Moore (Ed.), *New directions for student services, managing the political dimension of student affairs* (pp. 75–92). San Francisco: Jossey-Bass, No. 55.

Wang, G. (2003). Valuing learning: the measurement journey. *Educational Technology*, *43*(1), 32–37.

Weinstein, C. E., Palmer, D., Hanson, G., Dierking, D., McCann, E., Soper, M., & Nath, I., (1994). *Design and development of an assessment of readiness for training*. Paper presented at the annual conference of the Academy of Human Resource Development, San Antonio, TX.