FEATURE

The Flawed Four-Level Evaluation Model

Elwood F. Holton III

The lack of research to develop further a theory of evaluation is a glaring shortcoming for human resource development (HRD). In this paper, I argue that the four-level system of training evaluation is really a taxonomy of outcomes and is flawed as an evaluation model. Research is needed to develop a fully specified and researchable evaluation model. Such a model needs to specify outcomes correctly, account for the effects of intervening variables that affect outcomes, and indicate causal relationships. I propose a new model based on existing research and accounts for the impact of the primary intervening variables such as motivation to learn, trainability, job attitudes, personal characteristics, and transfer of training conditions. A new role for participant reactions is specified. Key studies supporting the model are reviewed and a research agenda proposed.

Evaluation of interventions is among the most critical issues faced by the field of human resource development (HRD) today. Increasing global competition has led to intense pressure on HRD to demonstrate that programs contribute directly to the organization's "bottom line." Yet the dominant evaluation model, the four-level Kirkpatrick model, has received alarmingly little research and is seldom fully implemented in organizations (Kimmerling, 1993), leaving them ill-equipped to respond to this pressure. There is a critical need for new evaluation theory and research to give organizations a more sound methodology for allocating HRD resources.

The Kirkpatrick model for training evaluation (Kirkpatrick, 1976), also known as the four-level evaluation model, is acknowledged by many practitioners as the standard in the field. A number of modifications to the model have been suggested, including adding a fifth level to reflect training's ultimate value in terms of organization success criteria, such as economic benefits or human good (Hamblin, 1974) and societal value (Kaufman and Keller, 1994), or to focus more specifically on return on investment (ROI) (Phillips, 1995). Brinkerhoff (1987) proposed a six-level model that, in essence, added two

formative evaluation stages as precursors to Kirkpatrick's four levels. Although this work has contributed greatly to our conceptual thinking about evaluation, the models have received incomplete implementation and little empirical testing.

All of them are best labeled as *taxonomies*, which are simply classification schemes (Bobko and Russell, 1991). Bobko and Russell, citing Wallace (1983), noted that exploratory designs and case studies are the first steps in theory development, whereas the final steps are correlational and experimental studies. According to them, taxonomies are the link between the initial stages and the final confirmatory stages of developing theory. Although the Kirkpatrick model is elegant in its simplicity and has contributed greatly to HRD, the lack of research to develop further a theory of evaluation is a glaring shortcoming for the field. If HRD is to continue to grow as a profession, an evaluation model grounded in research is necessary.

One shortcoming of taxonomies is that they do not fully identify all constructs underlying the phenomena of interest, thus making validation impossible. Not surprisingly, Alliger and Janak (1989), in their comprehensive review of research on the four-level model, note that the implied causal relationships between each level of this taxonomy have not been demonstrated by research. Their search of the relevant academic literature located only 12 articles since 1959 reporting twenty-six correlations between levels in training programs out of 203 articles that reported *any type* of evaluation results. Furthermore, only three studies (Clement, 1982; Noe and Schmitt, 1986; Wexley and Baldwin, 1986) reported full four-level evaluations with correlations. The reported correlations varied widely, casting doubt on assumptions of linear causal relationships.

It can be argued that the correlations reported in these studies were not really a test of the model but rather an alternate approach to analyzing outcomes. For example, if only the four levels of outcomes are measured and a weak correlation is reported between levels two and three, all we really know is that learning from training was not associated with behavior change. In the absence of a fully specified model, we don't know if the correlation is weak because some aspect of the training effort was not effective or because the underlying evaluation model is not valid. Weak correlations might represent a well-functioning model reporting a poorly functioning training effort.

It is not surprising that the reported correlations were weak because the model is really only a taxonomy of training (and HRD) outcomes. Attempts to test causal assumptions within a taxonomy are futile because, by definition, taxonomies classify rather than define causal constructs. Kirkpatrick (1994) is unclear about causal linkages in his model. On the one hand, he discusses the influence of other factors such as organizational climate and motivation to learn on training outcomes, suggesting that the relationships between levels are not simple, linear ones. On the other hand, he makes statements that clearly imply a simple causal relationship between levels. For example, he says

that "if training is going to be effective, it is important that trainees react favorably" (p. 27) and that "without learning, no change in behavior will occur" (p. 51). 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.

Klimoski (1991, pp. 254–256), building upon Dubin (1976), noted that theories or models should have six components:

1. *Elements* or *units*—represented as constructs—are the subject matter.
2. There are *relationships* between the constructs.
3. There are *boundaries* or *limits* of generalization.
4. *System states* and *changes* are described.
5. *Deductions* about the theory in operation are expressed as propositions or hypotheses.
6. *Predictions* are made about units.

The four-level model does not meet any of these criteria. First, essential elements are not present. Noticeably absent are the major intervening variables that affect learning and transfer processes such as trainee readiness and motivation, training design, and reinforcement of training on the job (Clement, 1982). Others have proposed models of how individual differences affect training outcomes (Noe, 1986; Noe and Schmitt, 1986) and how factors affect the transfer of training (Baldwin and Ford, 1988; Broad and Newstrom, 1992). Previous evaluation studies identified by Alliger and Janak (1989) did not attempt to measure any intervening variables, which is one likely reason for the wide variation in the correlations reported. No evaluation model can be validated without measuring and accounting for the effects of intervening variables.

Because all of the elements are not present, the relationships between constructs are not fully specified. Considering the third criteria, the four-level model seems to have no limits of generalization within HRD specified. Without full specification of the elements and the relationships, it is questionable whether the model can be applied universally. Furthermore, the missing elements and relationships prohibit making accurate statements about system states, developing propositions and hypotheses, and making predictions.

There is a critical need for intensive research to move from a taxonomic evaluation approach to a fully specified model for HRD evaluation that meets these six criteria of good theories and models. The purpose of this paper is to take an initial step toward establishing such a model. An integrative evaluation model that accounts for the impact of the primary and secondary intervening variables is proposed. The model was developed by examining relationships and constructs from previous empirical research in a grounded theory building approach (Glaser and Strauss, 1967) and integrating the findings with an existing theoretical framework.

Research Supporting Integrative Evaluation Models

Two studies have demonstrated the importance and feasibility of the proposed approach by examining influences on training outcomes. Noe and Schmitt (1986) used path analysis to examine the relationship between trainee reactions to skill assessment, job involvement, career planning, exploratory behavior, locus of control, posttraining motivation, and pretraining motivation with the outcomes of reactions, learning, behavior change, and performance improvement. Mathieu, Tannenbaum, and Salas (1992) used a more powerful technique—structural equation modeling—to examine the relationships between situational constraints, assignment to training, motivation, previous education, trainee reactions, learning, and behavior change.

Specific results of these studies will be discussed in the following paragraphs, but several general conclusions can be noted that justify this line of research. First, both studies supported Alliger and Janak's (1989) suggestion that reactions are not linearly related to learning but instead may act as a moderator or mediator of learning. Neither found support for reactions as a primary outcome of training. Second, both studies found that complex relationships exist between various intervening variables and the primary outcomes of learning and performance change. Third, both suggest that further research holds promise for developing a fully integrated model.

These studies also suggest that the shortcomings of the four-level evaluation model are most apparent when it is used as a diagnostic tool. Consider the case where performance change or positive results are not found to occur. The only conclusion possible using data within the four-level model is that something is wrong with the training program. However, if the many intervening variables that remain unmeasured are considered, it is quite possible that the training program is well designed and that the problem lies outside the classroom with some element of the organization, job, or individual. Thus, to the extent that evaluations should provide information to make correct decisions about HRD interventions, the four-level taxonomy fails to do so when the outcomes are not those desired. Only with a fully specified model can the true problems with the intervention be isolated. When training outcomes achieve desired levels, this limitation is not as serious because the only risk is incorrectly attributing improvements to the intervention.

Through statistical analysis that controls for the effects of intervening variables, it might be possible to show an effective training program design even when overall group scores indicate poor outcomes. This is not to suggest that trainers are responsible only for what occurs in the training room, because if the outcomes are not positive then trainers have not accomplished their goals. Rather, it suggests that there is a complex system of influences on training outcomes that must be measured if training is to be accurately evaluated. This model would enable practitioners to diagnose correctly barriers to training effectiveness.

Proposed Model

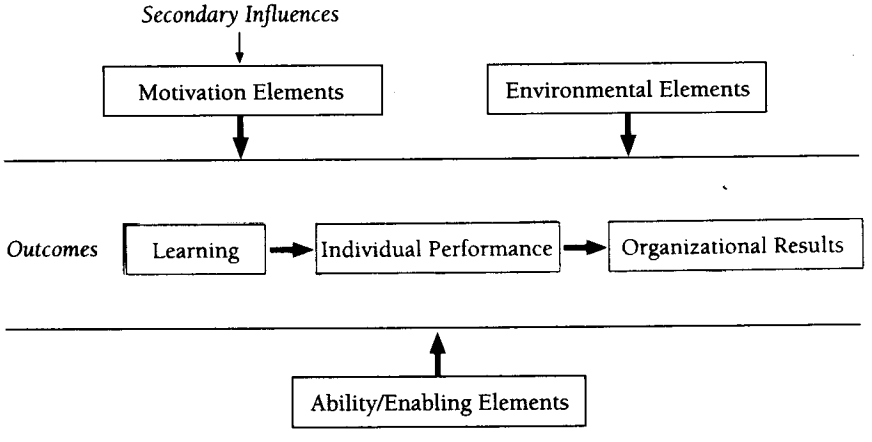
The proposed evaluation model is shown in conceptual format in Figure 1. Three primary outcome measures are proposed: *learning, individual performance,* and *organizational results.* These are defined, respectively, as achievement of the learning outcomes desired in an HRD intervention; change in individual performance as a result of the learning being applied on the job; and results at the organizational level as a consequence of the change in individual performance.

The first important difference between this model and the four-level taxonomy is the absence of reactions as a primary outcome (this will be discussed later in this paper). The second difference is that *individual performance* is used instead of *behavior* because it is a broader construct and a more appropriate descriptor of HRD objectives.

The third important difference between the two models is the inclusion of primary and secondary influences on outcomes. Noe (1986) proposed that a participant's behavior in training is a function of three factors: ability, motivation, and environment. His framework is used here to identify the primary intervening variables in the model.

Two of the outcomes—learning and individual performance—represent individual behaviors that an HRD intervention hopes to create. Learning can be classified as a behavior in the sense that individuals choose to learn or not to learn. However, learning is primarily an internal behavior whereas performance is usually a more external one. As will be discussed in more detail later, the same framework can be used at the organizational results level even though the unit of analysis shifts from the individual to the organization. Thus, HRD outcomes are hypothesized to be a function of ability, motivation, and

Figure 1. Conceptual Evaluation Model



environmental influences. Secondary influences are also included, particularly those that affect motivation.

The model describes a sequence of influences on outcomes occurring in a single learning experience. Over time, it would be expected that employees' successes at achieving results from learning experiences would enhance future motivation to learn. In other words, over time there are many cumulative feedback loops that are not shown in this model. The model is limited to evaluating outcomes from a single learning intervention although it also provides a conceptual view of the organizational HRD system at the macro level.

Influences on Learning Outcomes

The model assumes that there are three primary influences on learning: trainee reactions, motivation to learn, and ability.

Trainee Reactions. A frequently discussed aspect of the four-level model is the inclusion of reactions at the first level. Dixon (1990) demonstrated that there is little correlation between reactions and learning. Warr and Bunce (1995) recently divided reactions into three components (enjoyment, usefulness, and perceived difficulty) and also found no significant correlation between any of them and learning outcomes. Indeed, most learners would acknowledge that good learning can often be confusing and frustrating. Although these studies suggest that trainee reactions are unrelated to learning, as a practical matter few practitioners can afford to ignore totally the reaction of their trainees. Cognitive scientists point out that trainee reactions can play a role in building interest and attention and enhancing motivation (Patrick, 1992).

The question is whether trainee reaction should be a primary goal and outcome of training that is linearly related to learning or simply another intervening variable that has an impact on learning. Mathieu, Tannenbaum, and Salas (1992) found that reactions functioned as a moderator of the relationship between training motivation and learning as well as a mediator of other relationships. Their results offered the first explanation for the generally low correlations between reactions and learning observed by Alliger and Janak (1989). The authors noted that, had they examined the linear relationship alone, they too would have concluded that reactions had no significant relationship with learning. Instead, they concluded that reactions are important, but not in and of themselves. Similarly, Noe and Schmitt (1986) found no support for a direct link between reactions and learning.

Although these findings contradict conventional training practice, they suggest that trainee reactions should be removed from evaluation models as a primary outcome of training. The inclusion of trainee reactions as a primary outcome, particularly when defined as happiness (Kirkpatrick, 1994), is one of the greatest flaws of the four-level model. The effect has been to divert the field's attention away from the truly important HRD outcomes (for example,

performance) and focus many practitioners on activities that generate high ratings. Reactions and satisfaction with the learning experience are some of many measures of the performance improvement process but are not the primary outcomes of HRD interventions. Leaders of interventions, participants, and organizations need to focus more on the three primary outcomes and less on reactions to the process.

Reactions are included in this new model as a measure of the learning environment that affects learning behavior through a complex role moderating the relationship between motivation to learn and learning. More positive reactions to training may aid learning, and trainees who are more successful during learning are expected to have more positive reactions to the learning experience.

Motivation to Learn. Motivation to learn has a direct relationship with learning. Pretraining motivation and trainee attitudes have not received enough attention in the literature (Cohen, 1990). Four categories of variables are hypothesized to be primary influences on a participant's motivation to learn: readiness for the intervention, job attitudes, personality characteristics, and motivation to transfer learning. The last category will be discussed in the organizational results section of this paper.

Intervention Readiness. Several studies have examined influences on readiness to enter and participate in training programs. Hicks and Klimoski (1987) found that giving trainees the choice to attend training or not increased their motivation to learn and improved learning outcomes. Baldwin, Magjuka, and Loher (1991) found that trainees who had a choice of training content had greater motivation to learn. However, those who were allowed to choose but then not given their choice of training became less motivated than those who were not allowed to choose at all. Thus, the degree to which a trainee is involved in the needs assessment process and given choices about training would be expected to influence motivation to learn. The notion of meeting trainees' expectations and desires for training has also received support in the literature (Hicks and Klimoski, 1987; Tannenbaum, Mathieu, Salas, and Cannon-Bowers, 1991). Those trainees who do not feel the training will meet their needs will be less motivated and less likely to learn.

It is likely that motivation to learn will vary by trainees' readiness for the intervention. Readiness includes such variables as the degree to which trainees are involved in assessing needs, involvement in planning the training, degree to which expectations are clarified, degree of choice, and other unexplored influences.

Job Attitudes. Another unexplored influence on motivation to learn is the trainee's attitude toward the organization and the job. Given the rather large body of research on the relationship between job attitudes and overall motivation (Steers and Porter, 1991), it seems logical that job attitudes should affect motivation during learning interventions. However, only two studies could be located that tested this notion. Noe and Schmitt (1986) found a significant

relationship between job involvement and learning while Tannenbaum, Mathieu, Salas, and Cannon-Bowers (1991) found that more committed employees performed better in training. Although Mathieu, Tannenbaum, and Salas (1992) did not find a significant relationship between job involvement and motivation, they attributed it to the type of training in the study. It is likely that employees who exhibit more positive job attitudes would be more motivated to learn and, in turn, have more positive training outcomes.

Personality Characteristics. Selection researchers have long been interested in the validity of personality measures in predicting performance. The "Big Five" personality dimensions (extroversion, openness to experience, neuroticism, agreeableness, and conscientiousness) have been shown in a recent meta-analysis to have validity in explaining some of the variance in performance (Tett, Jackson, and Rothstein, 1991). Although HRD researchers have not explored these directly, other characteristics such as self-efficacy (Gist, Stevens, and Bavetta, 1991), locus of control (Noe and Schmitt, 1986), and need achievement (Baumgartel, Reynolds, and Pathan, 1984) have been shown to be related to training outcomes. Thus, certain personality characteristics would be expected to influence motivation to learn and, in turn, learning itself.

Ability. Overlooked in the evaluation literature is the role that general cognitive ability plays in influencing training outcomes. Psychologists have demonstrated that general cognitive ability has a significant impact on trainee success (Ree and Earles, 1991) and interacts with motivation (Kanfer and Ackerman, 1989) to enhance outcomes. Fleishman and Mumford (1989) provide an extensive review of the role general and specific cognitive abilities play in learning outcomes. It seems obvious that the ability of participants will affect the outcomes of an intervention. When evaluating a training program with a group of trainees who are relatively homogeneous in terms of job level and educational background, it is likely that there will be little variance in general ability and, therefore, little detectable influence on training outcomes. In situations where trainee groups are heterogeneous in cognitive ability or where programs are offered across an organization to groups of varying ability, it is likely that general cognitive ability will influence training outcomes. Because it is almost impossible to control for ability through random samples in most evaluation studies, it is essential to measure and control for it statistically.

Influences on Performance Outcomes

Learning is expected to lead to individual performance change only when three primary influences on transfer behavior are at appropriate levels. Following Baldwin and Ford (1988) and consistent with the Noe (1986) framework, the three primary influences proposed in this model are motivation to transfer, transfer conditions (environment), and transfer design (ability).

Motivation to Transfer. Trainees leave training programs with a certain level of motivation to utilize their learning on the job. Traditional evaluation

models acknowledge that transfer to the job is not certain but they fail to measure motivational influences on transfer. A variety of influences on transfer motivation have been suggested (Baldwin and Ford, 1988; Broad and Newstrom, 1992) and fall into four categories: intervention fulfillment, learning outcomes, job attitudes, and expected utility—or ROI—of results. The last category will be discussed in the organizational results section of this paper.

Intervention Fulfillment. Earlier research (Hicks and Klimoski, 1987; Hoiberg and Berry, 1978) has suggested that the degree to which trainees' expectations about training are met has a significant impact on posttraining attitudes. Goldstein (1985) stressed matching training to the needs and characteristics of learners as one of five factors underlying transfer. More recently, Tannenbaum, Mathieu, Salas, and Cannon-Bowers (1991) conducted a rigorous study of the effects of training fulfillment on a variety of training outcomes, including motivation. They operationalized training fulfillment as a combination of expectations with desires and perceptions of training related primarily to the relevance of training to the job. Their analyses controlled for the effects of pretraining attitudes, affective reactions to training (reactions), and performance in training itself. They found that training fulfillment played a significant role in understanding posttraining academic self-efficacy, commitment to the organization, and training motivation. Training motivation was similar to motivation to transfer because it was a measure of the trainees' perceived relationship between training success and future job performance.

It is expected that trainees who perceive that an intervention has met their expectations and fulfilled their need for performance-related learning will be more motivated to transfer learning into on-the-job performance.

Learning Outcomes. The outcomes of the learning intervention are also expected to have a secondary influence on motivation to transfer in addition to their primary influence on individual performance. Tannenbaum, Mathieu, Salas, and Cannon-Bowers (1991) also found that performance during training had an independent relationship with posttraining motivation. Expectancy theory (Vroom, 1964) suggests that individuals will be more motivated if they believe that their effort will lead to enhanced performance. More successful learners would be expected to feel better able to perform and, therefore, more motivated to transfer. In contrast, less successful learners would be expected to be less motivated to transfer learning. Their frustration might be particularly acute when they perceive that their performance-related learning needs were met but they were still not successful at completing the learning.

Job Attitudes. Just as job attitudes are expected to influence motivation to learn, they should also influence motivation to transfer learning to performance. Because of the paucity of research, the exact relationship is uncertain. However, expectancy theory would lead us to speculate that people with high commitment and job satisfaction would be more likely to exert effort to transfer and to perceive the rewards from transfer as having higher valence. Tannenbaum, Mathieu, Salas, and Cannon-Bowers (1991) offer some evidence of

this in their finding that organizational commitment and training motivation were correlated .53 before training and .55 after training. In general, participants with more positive job attitudes would be expected to be more motivated to transfer learning to performance.

Transfer Conditions. The notion that situational constraints affect workplace performance has been established in the literature (Peters, O'Conner, and Eulberg, 1985). Less certain is which constraints directly affect transfer of training. Recently, Rouillier and Goldstein (1993) reported on the development of an instrument to measure transfer climate. They proposed that transfer climate consists of seven dimensions: goal cues, social cues, task cues, positive reinforcement, negative reinforcement, punishment, and extinction. Although they were not successful in validating separate subscales in their instrument, their promising work demonstrated that transfer climate added significantly to the variance in performance explained after controlling for learning and unit performance. Tracy (1992) attempted to replicate their work, using their instrument along with his own organizational learning measures. He also found that transfer climate explained a significant portion of the variance in transfer.

Development of transfer instruments is still in its infancy but even when using developmental instruments research has shown what has been suggested in the practitioner literature for some time: nontraining factors such as supervisor support for training and rewards for using training affect a trainee's motivation to transfer learning into individual performance change on the job. Without controlling for the influence of transfer conditions, evaluation results are likely to vary considerably and yield erroneous conclusions about causes of intervention outcomes.

In this model, transfer conditions are posited to have a primary effect on performance and a secondary effect on motivation to transfer. Trainees who work in conditions supportive of learning transfer are more likely to transfer their learning to the job. In addition, people who work in positive transfer conditions are more likely to have high motivation to transfer.

Transfer Design. Another likely cause of failure to transfer is that the design of the training does not provide the ability to transfer the learning. That is, cognitive learning may well occur but the program participants may not have an opportunity to practice the training in a job context or may not be taught the manner in which to apply their new knowledge on the job. For example, several studies (Werner, O'Leary-Kelly, Baldwin, and Wexley, 1994; Wexley and Baldwin, 1986) have shown that goal setting during and after training improves transfer. In a more complex task situation such as negotiation training, improved results were obtained by augmenting goal setting with self-management training (Gist, Bavetta, and Stevens, 1990). Tziner, Haccoun, and Kadish (1991) showed that adding a relapse prevention module to training resulted in higher learning and greater transfer. Others cite numerous studies exploring dimensions of instructional design that enhance transfer of learning, including identical elements, conditions of practice, and overlearning (Baldwin and Ford, 1988; Patrick, 1992).

The degree to which transfer mechanisms are included in the design of the training itself is hypothesized to have a direct influence on the transfer of training. Trainees who are taught how to apply new knowledge and skills in a job context should have the ability to transfer learning which, when combined with motivation to transfer and positive transfer conditions, is likely to result in greater transfer.

Transfer design is difficult to measure because there are few definitive guidelines for what constitutes appropriate transfer designs. Transfer designs are expected to vary considerably depending on content, cultures, and other situational factors. Nonetheless, methods must be developed to assess the extent to which trainees acquire the ability to transfer learning. Clearly, even the most motivated trainee will be unable to transfer the learning if he or she does not know how to do so.

Influences on Organizational Results

Organizational results outcomes require a slightly different conceptualization because learning and performance are individual behavioral outcomes. However, organizational outcomes can still be roughly conceptualized as a function of ability, motivation, and environmental influences if viewed from the organizational perspective. That is, for an intervention to yield organizational results, it must have the ability to achieve results and motivate the organization and individuals to participate in it. It will also be affected by environmental factors. In other words, for results to occur, the intervention must be linked with organizational goals (ability), have utility or payoff to the organization and individual (motivation), and be subject to influences of factors outside HRD (environment). These factors are expected to be primary influences on organizational results, independent of learning and individual performance outcomes, and to influence other variables in the model.

Link to Organizational Goals. Effective interventions are based on extensive front-end analysis that begins with organizational analysis to identify the highest priority opportunities for performance improvement (Swanson, 1994). HRD interventions that are not linked to organizational mission, strategy, and goals are unlikely to produce results and particularly results that are valued by the organization (Rummler and Brache, 1990; Swanson, 1994). The analysis process generally proceeds in reverse order from evaluation outcomes and should be a continuous process of analysis leading to evaluation (Holton, 1995). If front-end analysis begins with organizational analysis and is done correctly, the resulting HRD intervention will be closely linked to organizational goals. Interventions closely linked to organizational goals are more likely to yield results while interventions not closely linked to organizational goals may not yield results, even with positive learning and individual performance change. Greater linkage to organizational goals would also tend to result in transfer designs that enhance transfer.

Expected Utility or Payoff. Organizations should not engage in HRD interventions unless the expected utility or payoff warrants investment of the resources. A variety of techniques can be utilized to calculate the financial benefits of HRD (Mosier, 1990) but generally the utility or payoff is not evaluated until the intervention is complete. However, Swanson and Gradous (1988) and Phillips (1991) correctly argue that the financial benefits should be forecast before the intervention begins. Interventions with low expected utility are less likely to demonstrate organizational results. From the organizational perspective, one reason is that low-utility interventions are simply less likely to receive resource allocations necessary to achieve profitable results.

This model also hypothesizes that the effects extend beyond resource allocation decisions to include a motivational component. It seems reasonable to expect that results are more likely to be achieved if the benefits are calculated and known to persons involved in the intervention, including both organizational sponsors and participants. Reber and Wallin (1984) showed that providing participants knowledge of the results of a safety training program increased their performance beyond the effects of training, goal setting, or training and goal setting. Generally, desired results were not achieved unless knowledge about results was included. Clark, Dobbins, and Ladd (1993) found that trainees who perceived training to have more job and career utility were more motivated. The correlation between training motivation and job and career utility were found to be .61 and .44, respectively, and to have significant paths in their structural model of training motivation.

Although much work remains to be done in this area, these findings are consistent with expectancy theory, which states that individuals will be more motivated if they perceive that their effort will lead to rewards they value. Interventions with high utility to the organization are also more likely to have high utility to the individual if there is a link between rewards and contribution to the organization. As shown in this model, high expected utility of organizational results from performance change should result in greater motivation to transfer learning into individual performance, and, in turn, in greater motivation to learn. Thus, organizational results are more likely to occur when an HRD intervention has a high expected utility or payoff to both the organization and the individuals.

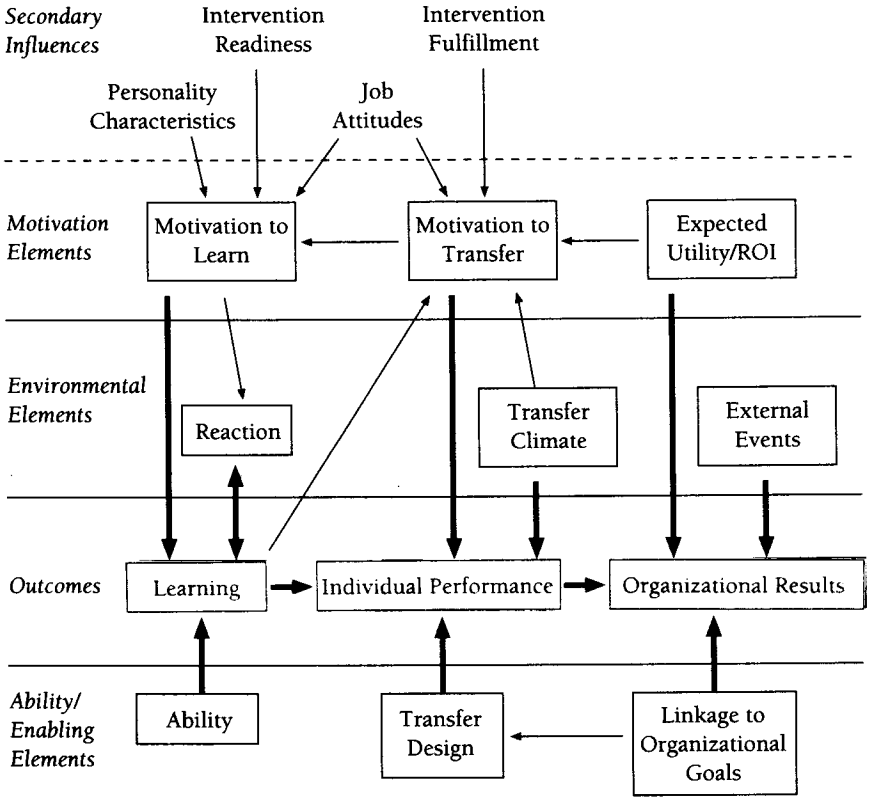
External Factors. One of the biggest challenges in evaluating organizational results is isolating the effects of training. A wide variety of factors that are completely outside the realm of training can affect the organizational results. For example, in a manufacturing environment productivity might be influenced by equipment failures, raw material shortages, price changes, absenteeism, and so on. The key external factors need to be identified and controlled for. These factors might be positive and amplify the intervention effects or negative and suppress results. A variety of measurement approaches can be used, ranging from simple control groups to more complex statistical methods. Because the strategies are unique to each organization and training program,

they cannot be specified in this model beyond a single global category of external factors.

Validating the Model

Figure 2 shows the complete model with all hypothesized relationships indicated by thick arrows (primary relationships) or lighter arrows (secondary relationships). Primary intervening variables (ability, motivation to learn, reaction to learning, transfer design, motivation to transfer, transfer conditions, expected utility, linkage to organizational objectives, and external events) are shown in boxes with arrows pointing directly to one of the outcomes. Secondary intervening variables (intervention readiness, job attitudes, personality characteristics, and intervention fulfillment) do not have boxes around them and have arrows pointing directly to one of the primary intervening variables. This is the measurement model that should be used for research purposes. Future research will need to operationalize the variables shown and test

Figure 2. HRD Evaluation Research and Measurement Model



the hypothesized relationships. Fortunately, models, scales, and methods already exist to make initial attempts at measuring most of the variables.

The validation of this model will clearly be an ambitious undertaking. One difficulty is in finding sites that are willing to donate time to collect this amount of data. Testing will require the use of sophisticated statistical techniques, the most promising of which is Structural Equation Modeling (LISREL), which can evaluate all causal relationships in the model simultaneously. This technique requires relatively large samples that are challenging to obtain. However, neither of these challenges are insurmountable.

From this discussion it should be evident that this effort needs to be undertaken. From a research perspective, the simple four-level taxonomy is inadequate. As suggested earlier, the role of reactions is very different from that specified in the four-level model: reactions are not a primary outcome. In addition, the four-level taxonomy can never be validated because there are too many unmeasured intervening variables. Furthermore, one can never really be sure what the answers provided by the taxonomy mean. Finally, the simple linear relationships between the levels that were implied in earlier writings and made more explicit in recent writing (Kirkpatrick, 1994) simply do not exist. It should be recognized that the four-level model is only a taxonomy and researchers should cease trying to validate it in its present form.

It should be equally evident, however, that the development and validation of a more complete model is feasible. The model proposed here is an initial step in that direction. Recent research points to the vital components of such a model and to the methodologies for measuring and validating it. The research agenda to move toward an integrative model should include the following:

- Validating the basic components of an integrative model
- Identifying the relationships within the model, including classifying direct, moderating, and mediating effects
- Identifying specific variables that should be measured within each of the major components of the model
- Developing and testing methodologies to analyze such a complex model
- Demonstrating that a fully integrated model can explain a major portion of the variance in training outcomes

It should be noted that this is not just a research tool; there are practical implications as well. Practitioners who use the four-level approach alone are quite likely to arrive at erroneous conclusions about their training programs. They need a model that can be used as a diagnostic tool, directing them to critical influences that need to be measured along with the outcomes and lead to accurate explanations of the outcomes obtained. When desired outcomes are not achieved, determining cause and effect is not merely an interesting research question but rather an essential step for making appropriate business decisions.

It is likely that after further research a simpler model may be developed for routine application where much of the variance will be explained by a smaller set of variables. For instance, it may be that measuring the primary intervening variables alone will be sufficient. Or it may be sufficient to measure a few key variables within each category. There is no theoretical reason why a practitioner model more sophisticated than the four-level approach but simpler to implement than the one described here could not be developed. Research is needed to compare results from this fully specified model with results from various combinations of simpler configurations.

Alternatively, coarse measures of the variables in the model might be found to be adequate surrogates for more precise measures. An interesting research question is whether end-of-course measures that collect self-reports of motivation to transfer, transfer design, motivation to learn, personal characteristics, transfer climate, and so on might be reasonable surrogate measures of these constructs. The method (self-report) and timing of data collection (end of course) might still be valuable as a coarse and inexpensive "first look" at evaluation. The key would be to stop asking "happiness" questions and focus on self-report estimates of the variables in a fully specified evaluation model. This line of research could lead to an evaluation system that is firmly grounded in validated theory but practical and efficient for routine use.

Conclusion

In this paper, I have argued for moving away from a taxonomic system and toward the development and testing of a true model for HRD evaluation. The model described here is an initial step in that direction and will surely be refined as research is conducted to test it. However, this discussion should make it clear that development and validation of a fully specified HRD evaluation model is within our reach. The methodologies to analyze such a model are now well established and sufficient research exists to develop reasonable research hypotheses. Reliance on the simple four-level taxonomy, and particularly on reactions as an outcome, only serves to minimize the value, impact, and sophistication of the intervention tools HRD employs and the results that can be achieved. If HRD is to grow as a discipline and a profession, it is imperative that researchers work deliberately to develop a more integrative and testable model.

References

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personal Psychology, 42,* 331–340.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41,* 63–100.

Baldwin, T. T., Magjuka, R. J., & Loher, B. T. (1991). The perils of participation: Effects of choice of training on trainee motivation and learning. *Personnel Psychology, 44,* 51–65.

Baumgartel, H., Reynolds, M., & Pathan, R. (1984). How personality and organizational/climate variables moderate the effectiveness of management development programmes: A review and some recent research findings. *Management and Labour Studies, 9,* 1–16.

Bobko, P., & Russell, C. (1991). A review of the role of taxonomies in human resources management. *Human Resource Management Review, 4,* 293–316.

Brinkerhoff, R. O. (1987). *Achieving results from training.* San Francisco: Jossey-Bass.

Broad, M. L., & Newstrom, J. W. (1992). *Transfer of training.* Reading, MA: Addison-Wesley.

Clark, C. S., Dobbins, G. H., & Ladd, R. T. (1993). Exploratory field study of training motivation. *Group and Organization Management, 18,* 292–307.

Clement, R. W. (1982). Testing the hierarchy theory of training evaluation: An expanded role for trainee reactions. *Public Personnel Management Journal, 11,* 176–184.

Cohen, D. J. (1990). The pretraining environment: A conceptualization of how contextual factors influence participation motivation. *Human Resource Development Quarterly, 1,* 387–398.

Dixon, N. M. (1990). The relationship between trainee responses on participation reaction forms and posttest scores. *Human Resource Development Quarterly, 1,* 129–137.

Dubin, R. (1976). Theory building in applied areas. In M. D. Dunnette (Ed.), *Handbook of Industrial/Organizational Psychology.* New York: Rand McNally.

Fleishman, E. A., & Mumford, M. D. (1989). Individual attributes and training performance. In I. L. Goldstein & Associates (Eds.), *Training and development in organizations* (pp. 183–255). San Francisco: Jossey-Bass.

Gist, M. E., Bavetta, A. G., & Stevens, C. K. (1990). Transfer training method: Its influence on skill generalization, skill repetition, and performance level. *Personnel Psychology, 43,* 501–523.

Gist, M. E., Stevens, C. K., & Bavetta, A. G. (1991). Effects of self-efficacy and posttraining intervention on the acquisition and maintenance of complex interpersonal skills. *Personnel Psychology, 44,* 837–861.

Glaser, B. G., and Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Hawthorne, NY: Aldine.

Goldstein, I. L. (1985). The applicability of a training transfer model to issues concerning rater training. In E. Locke (Ed.), *Generalizing from laboratory to field settings* (pp. 83–98). San Francisco: New Lexington Press.

Hamblin, A. C. (1974). *Evaluation and control of training.* New York: McGraw-Hill.

Hicks, W. D., & Klimoski, R. J. (1987). Entry into training programs and its effects on training outcomes: A field experiment. *Academy of Management Journal, 30,* 542–552.

Hoiberg, A., & Berry, N. H. (1978). Expectations and perceptions of Navy life. *Organizational Behavior and Human Performance, 21,* 130–145.

Holton, E. F., III (1995). A snapshot of needs assessment. In J. J. Phillips and E. F. Holton III (Eds.), *In action: Conducting needs assessment* (pp. 1–12). Alexandria, VA: American Society for Training and Development.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74,* 657–690.

Kaufman, R., & Keller, J. M. (1994). Levels of evaluation: Beyond Kirkpatrick. *Human Resource Development Quarterly, 5,* 371–380.

Kimmerling, G. (1993, Sept.). Gathering the best practices. *Training and Development,* 29–36.

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

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

Klimoski, R. (1991). Theory presentation in human resource management. *Human Resource Management Review, 4,* 253–271.

Mathieu, J. E., Tannenbaum, S. I., & Salas, E. (1992). Influences on individual and situational characteristics on measures of training effectiveness. *Academy of Management Journal, 35,* 828–847.

Mosier, N. R. (1990). Financial analysis: The methods and their application to employer training. *Human Resource Development Quarterly, 1,* 45–63.
Noe, R. A. (1986). Trainee attributes and attitudes: Neglected influences on training effectiveness. *Academy of Management Review, 11,* 736–749.
Noe, R. A., & Schmitt, N. (1986). The influence of trainee attitudes on training effectiveness: Test of a model. *Personnel Psychology, 39,* 497–523.
Patrick, J. (1992). *Training research and practice.* San Diego, CA: Academic Press.
Peters, L. H., O'Conner, E. J., & Eulberg, J. R. (1985). Situational constraints: Sources, consequences, and future considerations. In B. Rowland & G. Ferris (Eds.), *Research in personnel and human resources management,* Vol. 3 (pp. 79–113). Greenwich, CT: JAI Press.
Phillips, J. J. (1991). *Handbook of training evaluation and measurement methods* (2nd ed.). Houston, TX: Gulf Publishing.
Phillips, J. J. (1995). Return on investment—Beyond the four levels. In E. F. Holton III (Ed.), *Academy of HRD 1995 Conference Proceedings.*
Reber, R. A., & Wallin, J. A. (1984). The effects of training, goal setting, and knowledge of results on safe behavior: A component analysis. *Academy of Management Journal, 27,* 544–560.
Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g. Personnel Psychology, 44,* 321–332.
Rouillier, J. Z., & Goldstein, I. L. (1993). The relationship between organizational transfer climate and positive transfer of training. *Human Resource Development Quarterly, 4,* 377–390.
Rummler, G. A., & Brache, A. P. (1990). *Improving performance: How to manage the white space on the organization chart.* San Francisco: Jossey-Bass.
Steers, R. M., & Porter, L. W. (1991). *Motivation and work.* New York: McGraw-Hill.
Swanson, R. A. (1994). *Analysis for improving performance.* San Francisco: Berrett-Koehler.
Swanson, R. A., & Gradous, D. B. (1988). *Forecasting financial benefits of human resource development.* San Francisco: Jossey-Bass.
Tannenbaum, S. I., Mathieu, J. E., Salas, E., & Cannon-Bowers, J. A. (1991). Meeting trainees' expectations: The influence on the development of commitment, self-efficacy, and motivation. *Journal of Applied Psychology, 76,* 759–769.
Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44,* 703–742.
Tracy, J. B. (1992). *The effects of organizational climate and culture on the transfer of training.* Unpublished doctoral dissertation, State University of New York, Albany.
Tziner, A., Haccoun, R. R., & Kadish, A. (1991). Personal and situational characteristics influencing the effectiveness of transfer of training improvement strategies. *Journal of Occupational Psychology, 64,* 167–177.
Vroom, V. H. (1964). *Work and motivation.* New York: Wiley.
Wallace, M. (1983). Methodology, research practice, and progress in personnel and industrial relations. *Academy of Management Review, 8,* 6–13.
Warr, P., & Bunce, D. (1995). Trainee characteristics and the outcomes of open learning. *Personnel Psychology, 48,* 347–375.
Werner, J. M., O'Leary-Kelly, A. M., Baldwin, T. T., & Wexley, K. N. (1994). Augmenting behavior modeling training: Testing the effects of pre- and posttraining interventions. *Human Resource Development Quarterly, 5,* 169–183.
Wexley, K. N., & Baldwin, T. T. (1986). Posttraining strategies for facilitating positive transfer: An empirical exploration. *Academy of Management Journal, 29,* 503–520.

Elwood F. Holton III is assistant professor of human resource development, Louisiana State University, Baton Rouge.