IDEC DP² Series Vol. 3 No. 4

Developmnet Discussion Policy Paper

Testing Kirkpatrick's Four-Level Hierarchy of Training Evaluation: Evidence from Thailand's Automotive Industry

Tassanee Homklin, Yoshi Takahashi,
and Kriengkrai Techakanont

May, 2013

Department of Development Policy
Division of Development Science
Graduate School for International
Development and Cooperation (IDEC)
Hiroshima University
1-5-1 Kagamiyama, Higashi-Hiroshima
739-8529 JAPAN

Testing Kirkpatrick's Four-Level Hierarchy of Training Evaluation: Evidence from Thailand's Automotive Industry

Tassanee HOMKLIN[1a], Yoshi TAKAHASHI[a], and Kriengkrai TECHAKANONT[b]

[a]*Hiroshima University, Japan*

[b]*Thammasat University, Thailand*

Abstract

Several studies of training evaluation have failed to confirm the hierarchy relationship of reaction, learning, and behavior to results because of the difficulty of evaluating training. Furthermore, research in this area has tended to downplay the importance of level one (reaction) evaluation. In this study, we proposed investigating Kirkpatrick's four-level hierarchy of training evaluation, focusing specifically on two types of reactions, affective and utility, to predict training outcomes. The results of this study expand our understanding of the progressive causal relationship of reaction, learning, and job behavior to results. In particular, this study highlighted the utility reactions in predicting training effectiveness. Implications and future research directions suggested by the results are also discussed.

Key words: Kirkpatrick's training evaluation model, training effectiveness, training evaluation.

1. Introduction

Training is the most important strategy as well as commonly used human resource development activity by organizations to help employees improve knowledge and skills to meet environmental challenges. Organizations have come to spend more time and money on training; therefore, it is important that they evaluate the effectiveness of their training efforts more than ever (Cascio, 1989).

Among training evaluation models, Kirkpatrick's four-level model is the most extensively accepted and used, as it is simple, clear, and easy to implement, as training evaluators expect. The model shows four levels of training outcomes: reaction, learning,

[1] Corresponding author, Hiroshima University, Graduate School for International Development and Cooperation (IDEC), 1-5-1 Kagamiyama Higashi-Hiroshima, Hiroshima 739-8529 Japan.
E-mail address: h-tassanee@hiroshima-u.ac.jp (Homklin Tassanee).

behavior (transfer), and results. Organizations often evaluate training effectiveness using one or more of Kirkpatrick's criteria (Kirkpatrick, 1994). However, there are three limitations of Kirkpatrick's model that have implications for the ability of training evaluators to deliver benefits and, further, to satisfy the interests of organizations. These include the incompleteness of the model, the assumption of causality, and the assumption of the increasing importance of information as the levels of outcomes rise (Bates, 2004).

This study highlights one important discussion point concerning Kirkpatrick's model, that is, its emphasis on the progressive causal relationship of reaction, learning, and job behavior to results. For instance, trainees' satisfaction is important in making learning effective. Without learning, behavioral change will not occur (Kirkpatrick, 1994). Several studies of training evaluation have failed to confirm the hierarchical relationship of reaction, learning, and behavior to results because of the difficulty of evaluating training. Two meta-analyses of training evaluation studies, Alliger & Janak's (1989) and Alliger, et al.'s (1997), investigated the relationship among training criteria by using Kirkpatrick's model. They found little evidence either of substantial correlations between measures at different outcome levels or evidence of the linear causality suggested by Kirkpatrick (1994). Thus, as the model is still widely but only partially used in academic circles and by businesses, training evaluation academics tend to emphasize the need to examine all four of Kirkpatrick's evaluation levels.

The measurement of the reaction which generally takes place at the end of a course is the most commonly evaluated by organizations (Swanson & Sleezer, 1987; Arthur, Bennett, Edens & Bell, 2003). However, the previous studies did not provide a clear picture of the relationship between reaction and learning. That is because past research may have been limited by the criteria of reactions as a single dimensional construct. This is a considerable gap in trainee reaction for assessing the effectiveness of training. However, whether or not trainees are satisfied with the training they received does not provide an in-depth understanding of the effectiveness or other results of the training (Kirkpatrick, 1967). Alliger, et al. (1997) suggest that many trainee reaction items can be collapsed into a single affective dimension. Thus, when designing training programs and evaluating the results, various critical aspects of trainee reactions should be considered rather than focusing only on affective reactions such as whether the trainee enjoyed the training. Furthermore, their reaction forms should include utility judgments (Alliger, et al,

1997). This leads to an increased understanding of the role specific reactions play in training effectiveness.

Discussion about the insufficiency of reaction measures and research in this area has tended to downplay the importance of level 1 evaluation (Giangreco, et al., 2009). In fact, for several decades, the distinction between learning and job behavior has drawn increased attention to the importance of the learning transfer process in making training truly effective (Bates & Coyne, 2005). However, evaluation of reactions should not be ignored. In this respect, the following four reasons for reaction evaluation should be emphasized. First, positive training experiences may well have a beneficial impact on employee attitudes and behaviors (Alliger & Janak, 1989; Arthur, et al., 2003; Clement, 1982). Second, reaction evaluations can help organizations identify particular problems or weaknesses in their current training and improve their future training (Brown & Gerhardt, 2002; Mann and Robertson, 1996; Tannenbaum & Woods, 1992; Brinkerhoff, 1986; Ford & Wroten, 1984). Third, it shows trainees that the trainers are there to help them do their job better and that they need feedback to determine how effective they are (Kirkpatrick, 1994). Finally, reaction is more practically acceptable for training evaluation as a potential predictor of more costly criteria for training effectiveness—measures of learning, measures of on-the-job behavior, and measures of organization results. Thus, it is still important to examine the level of reaction to training.

Most often in Thailand, training evaluation is based on the participants' satisfaction survey of the program, trainers' subjective evaluation, and whether the trainees can understand and absorb the knowledge and skills from the training. Although these indicate Kirkpatrick's level one (reaction) and level two (learning) approaches, few studies have used all four levels of Kirkpatrick's model to evaluate Thai industries, including the automotive industry, the subject of the present study. Because of the difficulty of evaluating training, much training in Thailand either ignores behavior (level three) and results (level four) or approaches it through reaction and learning only.

Based on the arguments above, the main purpose of this study is to investigate Kirkpatrick's four-level hierarchy of training evaluation, focusing specifically on the type of reaction criteria, including affective and utility reactions, in predicting training outcomes. To achieve the purpose of this research, the authors pose the following research questions: What is

the relationship of reaction, learning, and behavior to results? In particular, how do trainees' affective and utility reactions influence learning?

2. Training effectiveness: Kirkpatrick's model

Most of the research on training evaluation has depended on Kirkpatrick's (1967) four-level typology to explain the effectiveness of training. Level 1, reaction, is trainees' feelings about and like of a training program. Reaction does not measure what trainees have learned, but rather indicates the trainee's motivation to learn. Although a positive reaction may not ensure learning, a negative reaction probably reduces the possibility that learning occurs. Note that a reaction measure is conceived in attitudinal rather than behavioral terms. Level 2, learning, is defined as the "principles, facts, and techniques understood and absorbed by the trainees" (Alliger & Janak, 1989). No change in behavior can be expected unless one or more of these learning objectives have been accomplished (Kirkpatrick, 1994). Learning is most often assessed by giving the trainees tests that tap declarative knowledge (Kriger, et al., 1993). This level of evaluation allows trainees to demonstrate their understanding of specific knowledge and/or skills within the learning program. Level 3, behavior change or transfer, refers to the knowledge and skills transferred to the job by trainees. This level attempts to determine whether trainees (who can apply the acquired specific knowledge and/or skills) use their new knowledge and/or skills when returning to the work environment. Level 4, results, refers to the final results that occurred because the trainees attended the program (Kirkpatrick, 1994). These could include the attainment of organizational objectives such as a reduction in absenteeism and personnel turnover, productivity gains, and cost reduction.

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 (Bates, 2004). Although Kirkpatrick is not clear about the causal linkages between training outcomes, his model can imply that a simple causal relationship exists between the levels of evaluation (Holton, 1996). In one of Kirkpatrick's more recent publications he argued that "if training is going to be effective, it is important that trainees react favorably and without learning, no change in behavior will occur" (Kirkpatrick, 1994). Research on training evaluation has largely failed to confirm such causal linkages. Two meta-analyses of training evaluation studies using Kirkpatrick's model (Alliger & Janak, 1989; Alliger et al.,

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).

Many studies that have evaluated training on two or more of Kirkpatrick's levels have reported different effects from training for different levels. However, few studies on training evaluation have tried to investigate the hierarchy of training outcomes and even fewer studies indicate the application of the four categories other than at the reaction level (Clement, 1982; Brandenburg, 1982; Parker, 1986; Alliger & Janak, 1989; Brinkeroff, 1989; Alliger, et al., 1997). For example, Alliger and Janak (1989) noted that only three out of 203 empirical studies examined all four levels. They found that reaction had a very weak correlation with learning ($r = .07$) but found stronger relations between learning and behavior ($r = .13$), learning and results ($r = .40$), and behavior and results ($r = .19$). Furthermore, Clement (1978) found the strongest evidence in support of the hierarchy by using path analysis and the results show that trainee reactions had a causal impact on learning, and learning had a significant influence on behavior change. Clement (1982) also found that reactions were positively related between learning and improvement in communicating behavior. However, only a few training evaluation studies have provided indirect support for the hierarchical model and demonstrated that satisfaction with training, learning, and behavior change occurs jointly (Fromkin et al., 1975; Latham, Wexley, & Purcell, 1975). Thus, this study tests the hierarchy relationship of training evaluation. We hypothesize that:

Hypothesis 1: There will be a hierarchy relationship of reaction, learning, and job behavior to results.

Discussion about the role of reaction measures has been prevalent in the literature of training evaluation. It is recognized that trainees cannot reap the full benefits of training without considering the role of reaction. However, some researchers such as Holton (1996) raised the question of the appropriateness of trainee reaction as a criterion of training effectiveness because a number of studies presented only a minor systematic relationship between reaction and learning. Most research related to training evaluation has focused on measuring trainee reaction to the training program and the degree of learning from the program (Tracey, et al., 1995).

Many studies on training effectiveness have concluded that reaction is positively related to learning (Brown, 2005; Kirkpatrick, 1994; Mathieu, Tannenbaum, & Salas, 1992; Noe &

Schmitt, 1986; Tracey et al., 2001; Warr, et al., 1999; Lin, Chen, and Chuang, 2011). However, some studies found little correlation between reaction and learning (Colquitt, Lepine, & Noe, 2000; Alliger, et al., 1997; Alliger & Janak, 1989; Dixon, 1990; Noe & Schmitt, 1986; Warr & Bunce, 1995). In addition, some researchers have even argued that trainee reactions are unrelated to learning (Holton, 1996; Hook & Bunce, 2001; Noe & Schmitt, 1986).

Furthermore, past research on training reaction and effectiveness may have been limited by the treatment of reaction as a unidimensional construct (Morgan and Casper, 2000). A number of studies have examined the relationships between reaction and learning. Particular facets or dimensions of trainee reactions appear to hold more promise, such that Alliger, et al. (1997) distinguish between affective and utility judgments of reactions. They found that utility reactions have a modest but significant relationship to immediate learning ($r = .26$); affective reactions to training do not. This study reporting a combined scale of affective and utility reactions has a significant relationship to immediate learning ($r = .14$) and to behavior or skill demonstration learning, the Level II distinction made by those researchers-- ($r = .12$). More recently, Tan, Hall, and Boyce (2003) found that both affective and cognitive/intention reaction scales did significantly correlate to a modest degree with the learning criteria. On the contrary, Hook and Bunce (2001) found that affective and utility reactions were not related to immediate learning. Moreover, Cannon-Bowers, et al. (1995) proposed that trainees' reactions, including satisfaction and perceived utility, were not related to declarative knowledge acquisition. The empirical research on facets or dimensions of trainee reaction remained equivocal.

As discussed above, previous empirical results have been inconclusive for the purpose of investigating the relationship between reaction and learning. Therefore, this study proposes to investigate the two facets of reactions, that is, affective and utility reactions. We collected measures of reaction and learning in order to determine if the training program was effective and examine the pattern of relations among the different types of criteria. Thus, we develop the hypotheses below:

Hypothesis 1A: Combined trainee reactions will be positively related to learning.

Hypothesis 1B: Trainee affective reactions will be positively related to learning.

Hypothesis 1C: Trainee utility reactions will be positively related to learning.

In addition to the relationship between learning and behavior, trainees must have the ability to retain knowledge and skills instilled during the training program to facilitate the transfer process. Baldwin and Ford (1988) argue that learning retention outcomes are directly associated with the generalization and maintenance of training effects on the job. They argue that in order for trained skills to be transferred, they first must be learned and retained. Furthermore, Velada, et al. (2007) also found that when trainees retain training content, they are more likely to perceive that they have transferred the training to the work context. Based on the literature reviews above, we hypothesize that:

Hypothesis 1D Learning will be positively related to behavior.

Fewer previous studies have investigated the relationship between behavior and results compared with those studies on the relationship between reaction and learning and the relationship between learning and behavior. The first important reason is there can be a long delay from the improvement in job behavior to desired organizational results. The second reason is there are more variables, both inside and outside the organization, which can influence this relationship (Clement, 1982). The final reason is greater difficulty in evaluating training at the higher levels of Kirkpatrick's model. However, while considering Kirkpatrick's original idea that there are causal relationships through all four levels, including from behavior to results, in this study we hypothesize that:

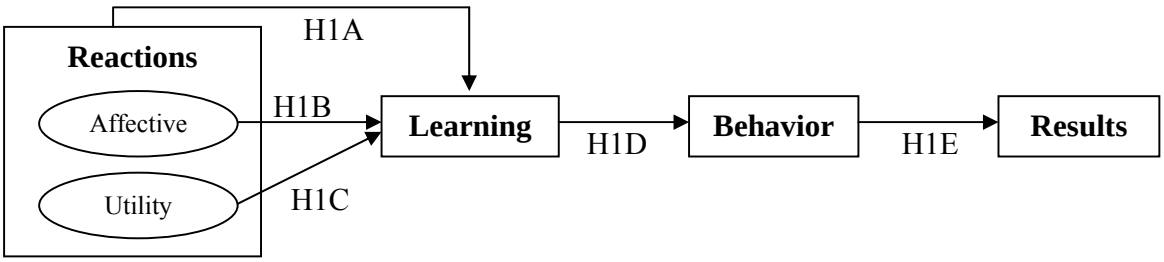
Hypothesis 1E: Behavior will be positively related to results.

3. Methodology

3.1 Conceptual framework

The conceptual framework for this study is shown in Figure 1. A focus of this study is testing Kirkpatrick's four-level hierarchy of training evaluation and investigating two facets of reactions, including affective and utility reactions, to predict training effectiveness. Specific hypotheses for each of the relationships are illustrated in Figure 1.

Fig 1. Conceptual Framework



3.2 Data collection context and sample

The case of the present study, the skill certification system for the automotive industry in Thailand, was one of the sub-programs under the Automotive Human Resource Development Program (AHRDP) and is expected to be very significant because of its potential impact on the whole industry. AHRDP was implemented from 2006 to 2011, as part of the Japanese Official Development Assistance (ODA) program, in cooperation with the Thai government and private sectors in both countries. Specifically for the skill certification system, at the start, Japanese experts from an automotive assembler, Nissan, supported knowledge transfer to local prospective examiners and trainers. They in turn transferred acquired skills and know-how to employees in local firms through training and examination. Through 2012, 363 people were certified in sixteen subjects: (1) die and mold finishing, (2) mechanical assembly finishing, (3) lathe with numerical control, (4) milling with numerical control, (5) handwritten mechanical drawing , (6) mechanical drawing by CAD, (7) electronic device assembly, (8) sequence control, (9) hydraulic system adjustment, (10) mechanical maintenance, (11) electrical maintenance, (12) metal press work/stamping, (13) plastic injection, (14) machining (lathe, milling), (15) ferrous casting, and (16) pneumatic circuits and apparatus device assembling. All of those subjects included theoretical and practical sessions. Questionnaires were distributed to all the participants in the sub-programs while 228 provided valid responses yielding a response rate of 62.8%.

Of the 228 study participants, 148 people participated in examiner training, 225 in trainer training, while the remaining 61 people attended courses for ordinary training. A participant could attend multiple levels and study various training subjects. The subjects studied by trainees included electrical maintenance (11.2%), mechanical maintenance (9.5%), pneumatic circuits and apparatus device assembling (8.8%), metal press work/stamping (8.4%), hydraulic system

adjustment (8.2%), three courses for die and mold finishing, electronic device assembly, plastic injection (6.5%), ferrous casting (6.0%), sequence control (6.0%), milling with numerical control (5.0%), machining (lathe, milling) (4.7%), lathe with numerical control (4.5%), mechanical assembly finishing (3.0%), handwritten mechanical drawing (3.0%), and mechanical drawing by CAD (2.2%). Among the sample, 98.7% of the participants were male. Regarding their age, 48.0% of the samples were between 31 and 40 years old, 40.1% were between 21 and 30 years old, whereas 11.9% were older than 40. 38.9% graduated from university, and 33.3% graduated from a vocational school. 55.5% of the respondents worked for automotive assembler and automotive parts manufacturers.

3.3 Measures

Variables in this study, as well as their corresponding sources of information, are described below.

Reactions. Twenty-seven items adopted from Morgan and Casper (2000) were used to assess trainees' feelings for and like of a training program. Affective reactions measure the extent to which a participant "liked" or was satisfied with different components of the training (e.g. course structure, testing process, instructors, materials, training management and administration process). Utility reactions consider the extent to which the participants can apply the content of training to their job. Sixteen items assessed the affective reactions of the trainee and five items were used to assess the participants' utility reactions to the training program. Reponses were made on a five-point Likert scale, with 1 = very dissatisfied and 5 = very satisfied.

Learning. Based on Kirkpatrick's model, learning refers to the knowledge, skills, and attitude acquired by trainees. Learning aims at understanding trainees' comprehension of instruction, principles, ideas, knowledge and skills from training. The learning measure consisted of sixteen items adopted from previous studies (e.g. Kirkpatrick, 2006; Leach and Liu, 2003). Reponses were made on a five-point Likert scale, with 1 = disagree strongly and 5 = agree strongly.

Behavior refers to the knowledge and skills transferred to the job by trainees (Kirkpatrick, 1994). Behavior consisted of thirteen items adopted from previous studies (e.g. Kirkpatrick,

2006; Leach & Liu, 2003, Velada, et al., 2007; Xiao, 1996). Reponses were made on a five-point Likert scale, with 1 = disagree strongly and 5 = agree strongly.

Results refer to the final results that occurred because the trainees attended the program (Kirkpatrick, 1994). These could include the attainment of organizational objectives and individual benefits. The results consisted of eighteen items adopted from previous studies (e.g. Kirkpatrick, 2006; Leach & Liu, 2003, Velada, et al., 2007; Xiao, 1996). Reponses were made on a five-point Likert scale, with 1 = disagree strongly and 5 = agree strongly.

In this research, the reliability of all remaining items was examined using one-dimension assessment. As a test of reliability, Cronbach's α was adopted to represent internal consistency. Cronbach's α for each scale of the questionnaire is acceptable (Reaction: .709, Learning: .665, Behavior: .647, and Results: .639), with all values greater than the threshold of .60. Therefore we conclude that the items are reliably measuring the defined constructs and variables.

4. Analysis of measurement model

In accordance with Gerbing and Hamilton's (1996) recommendation, we followed a three-stage approach. First, the measurement scales of latent variables were examined using exploratory factory analysis (EFA) in SPSS 19. Some items were eventually eliminated using this process. Then, all remaining items from the four measures were entered into a confirmatory factor analysis (CFA) in LISREL 9.10 using maximum likelihood (ML) estimation. Finally, to test the proposed hypotheses, the structural equation model was assessed. The criteria were used to evaluate the fit of the models in this study by taking suggestions from Bollen (1989), Joreskog and Sorbom (1993), and Hu and Bentler (1995) and all the criteria were satisfied. The scale internal structure fit measures abstract is shown in Table 1. The CFA results of reaction, learning, behavior, and results were appropriate (RMSEA = 0.020, 0.036, 0.068, and 0.046, respectively).

Means, standard deviations, and correlations among all measurements are reported in Table 2. Correlation analyses by Pearson product-moment indicated that the facet of reactions, including affective reactions, have a positive significant correlation with utility reactions (r = .151, $p < 0.05$), learning (r = .234, $p < 0.01$), behavior (r = .299, $p < 0.01$), and results (r = .276, $p < 0.01$). Another facet of reactions was that utility reactions have a positive significant correlation with learning (r = .324, $p < 0.01$), but were not significantly correlated with behavior (r = .127, $p > 0.05$). Furthermore, learning has a positive significant correlation with behavior (r

= .312, $p < 0.01$). However, both utility reactions and learning were not significantly correlated with results ($r = .080$ and $r = .029$ respectively, $p > .05$).

Table 1. Goodness of fit of scale internal structure.

	Criteria	Reactions	Learning	Behavior	Results
GFI	>0.90	0.935	0.951	0.965	0.950
SRMR	<0.06	0.047	0.052	0.069	0.064
RMSEA	<0.08	0.020	0.036	0.068	0.046
AGFI	>0.90	0.898	0.926	0.917	0.909
NNFI	>0.90	0.992	0.968	0.677	0.949
CFI	>0.90	0.994	0.975	0.829	0.966
PNFI	>0.50	0.600	0.688	0.394	0.591
PGFI	>0.40	0.593	0.627	0.402	0.518
χ^2/df	<2.00	1.095	1.397	2.049	1.304

Note. n = 228 for all models. GFI = goodness of fit index, SRMR = standardized root mean square residual, RMSEA = root mean square error of approximation, AGFI = adjusted goodness of fit index, NNFI = non-normed fit index, CFI = comparative fit index, PNFI = parsimony normed fit index, PGFI = parsimony goodness of fit index.

Table 2. Means, standard deviations, and correlations of variables

Variables	M	SD	1	2	3	4
1. Affective reactions	4.193	0.230				
2. Utility reactions	4.261	0.359	.151*			
3. Learning	4.094	0.353	.234**	.324**		
4. Behavior	4.051	0.342	.299**	.127	.312**	
5. Results	4.087	0.316	.276**	.080	.029	.283**

Note: Mean and standard deviation of all reaction are 4.209 and 0.206. Combined reactions demonstrated a statistically significant and positive correlation with learning, behavior, and results (r =.333, .307, and .268, respectively $p < 0.01$)

*$p < 0.05$, **$p < 0.01$*

5. Results and Discussions

5.1 Overall fit evaluation results

To test the fit of the hypothesized model, a structural equations analysis was conducted using LISREL 9.10 (Joreskog and Sorbom, 1993). The initial results of the hypothesis to test Kirkpatrick's four-level hierarchy of training evaluation by combining reactions in Model 1 showed that the overall chi-square was statistically significant ($\chi^2 = 281.11$ $df = 186$, $p < .001$);

the GFI was 0.891, the SRMR was 0.069, the RMSEA was 0.047, the AGFI was 0.865, the NNFI was 0.757, the CFI was 0.785, the PNFI was 0.518, the PGFI was 0.718, and the χ^2/df was 1.511.

Table 3. Goodness of fit of structural model.

	Criteria	Model 1	Model 2
GFI	>0.90	0.891	0.894
SRMR	<0.06	0.069	0.068
RMSEA	<0.08	0.047	0.046
AGFI	>0.90	0.865	0.867
NNFI	>0.90	0.757	0.772
CFI	>0.90	0.785	0.800
PNFI	>0.50	0.518	0.525
PGFI	>0.40	0.718	0.712
χ^2/df	<2.00	1.511	1.489

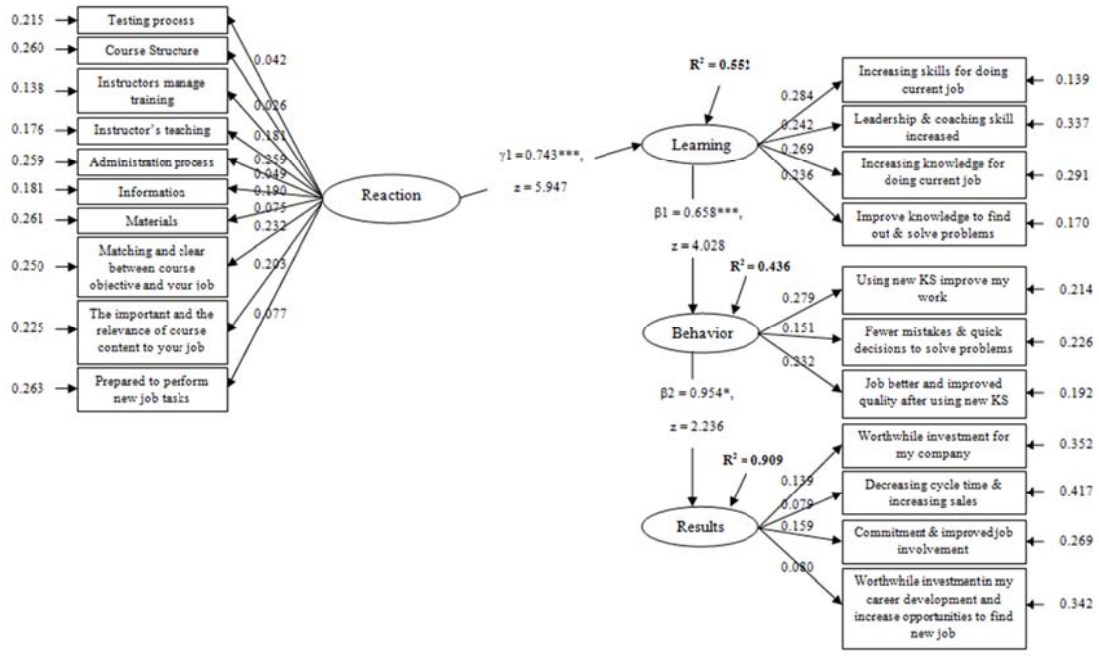
In addition, the further analyses tested two facets of reactions, including affective and utility reactions, to predict training outcomes in Model 2. The fit of the hypothesis showed that the overall chi-square was statistically significant ($\chi^2 = 274.04$; $df = 184$, $p < .001$); the GFI was 0.894, the SRMR was 0.068, the RMSEA was 0.046, the AGFI was 0.867, the NNFI was 0.772, the CFI was 0.800, the PNFI was 0.525, the PGFI was 0.712, and the χ^2/df was 1.489 (Table 3). From this perspective, it is therefore advisable to use the χ^2 value in conjunction with other fitness indices. In this study the fitness of the overall model is assumed to be appropriate according to good fitness indices including GFI.

5.2 Study hypothesis test results

With respect to our specific research hypotheses, there were hierarchy relationships of reaction, learning, and job behavior to results. Hypothesis 1 and the sub-hypotheses, including hypotheses 1A, 1D, and 1E, were supported. First, trainee reaction was positively related to learning ($\gamma 1 = 0.743$, $z = 5.947$, $p < .001$). Reaction explained 55.2% of variance of learning. Second, learning was positively related to behavior ($\beta 1 = 0.658$, $z = 4.028$, $p < .001$). Reaction and learning explained 43.6% of variance of behavior directly and/or indirectly. Third, behavior was positively related to results ($\beta 2 = 0.954$, $z = 2.236$, $p < .05$). From the residual, the results can be explained by reaction, learning, and behavior directly and/or indirectly at a 90.9% rate. The results for the hypothesized model are depicted in Figure 2.

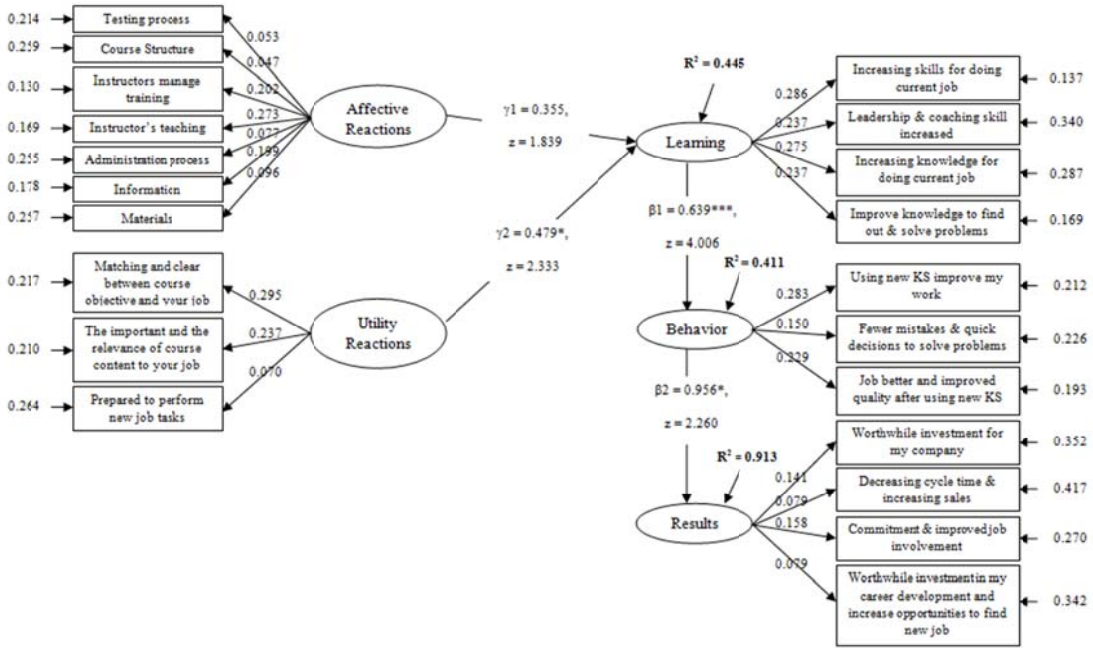
For the next model, we also examined two dimensions of reactions, that is affective and utility reactions, to predict training effectiveness. The results partially supported the two sub-hypothesis. The results provided support to hypotheses 1C while hypothesis 1B was not supported. As can be seen in Figure 3, first, affective reactions were not significantly related to learning ($\gamma 1 = 0.355$, $z = 1.839$). Only utility reactions were positively related to learning ($\gamma 2 = 0.479$, $z = 2.333$, $p < .05$). Reactions explained 44.5% of variance of learning. Second, learning was positively related to behavior ($\beta 1 = 0.639$, $z = 4.006$, $p < .001$). Utility reactions and learning explained 41.1% of variance of behavior directly and/or indirectly. Third, behavior was positively related to results ($\beta 2 = 0.956$, $z = 2.260$, $p < .05$). From the residual, the results can be explained by reaction, learning, and behavior directly and/or indirectly at a 91.3% rate.

Fig. 2. Estimated results of the model for testing Kirkpatrick's four-level hierarchy of training evaluation (Model 1)



*$p < 0.05$, ***$p < 0.001$

Fig. 3. Estimated results of the model for expanding the facets of reactions in predicting training effectiveness (Model 2).



*$p < 0.05$, ***$p < 0.001$*

5.3 Discussion

This study makes two specific contributions. First, it expands on the approach to the measurement of the impact of training using Kirkpatrick's model to expand the hierarchy relationship of reaction, learning, and job behavior to results. The results from CFA of the proposed model showed that combined reactions were significantly related to learning, learning was significantly related to behavior, and behavior was significantly related to results. The results of this study fully supported previous findings in the literature on training effectiveness (Alliger and Janak, 1989; Alliger et al., 1997; Leach and Liu, 2003; Kirkpatrick, 1996; Tan, et al., 2003; Warr, Allan, and Birdi, 1999). That is, reaction was significantly related to learning. This result is consistent with Alliger, et al.'s, (1997) meta-analysis and supports Kirkpatrick's (1967) original suppositions on the hierarchical nature of the relationship among the four primary training criteria. Reaction and learning from training play a critical role in the process of training evaluation. Positive reaction may influence an individual's willingness to use newly acquired knowledge and to attend future training programs (Tracey, et al., 2001) and learning is a fundamental requirement for transferring training to the workplace.

However, the relationship between behavior and results was even weaker than the relationship between learning and behavior. Trainees may not be able to immediately apply their acquired knowledge and skills to the job effectively due to the long delay between the change in job behavior and the desired organizational results. Moreover, there are many variables in the organization which can interfere with this relationship. Within the organization, we should consider the influence of the supervisor or manager, peers, and organizational support, as well.

Second, the other model tested the facets of reactions that were articulated in Kirkpatrick's model of training effectiveness. Two kinds of reactions, affective and utility reactions, were hypothesized to impact learning. The results of the present study underlined that trainee utility reactions had a significant relationship to learning. This result is consistent with Alliger, et al., (1997) who found that utility judgments hold potential as a predictor of learning and subsequent on-the-job use of the training content. In other words, our results showed that trainee utility reactions focus on the potential applicability of the material to the person's job. Utility reactions to training are more likely to be associated with changes in work behavior because trainees who see the program as relevant to their work are more likely to transfer their learning than those for whom it has low relevance (Warr & Bunce, 1995). This empirical support for a utility dimension is noteworthy.

In contrast, trainee affective reactions were not significantly related to learning. This means that trainees, even if they are satisfied with different components of training, such as course structure, testing process, instructors, materials, training management, and the administration process, didn't tend to achieve higher learning. It may be that trainees may enjoy a training activity which is not at all connected with his or her work activities, or may dislike learning something which is nevertheless of considerable importance to their job (Warr & Bunce, 1995). Furthermore, consistency with Kirkpatrick's (1967) original idea that trainees are satisfied with the training they received does not provide an in-depth understanding of the effectiveness of the training.

However, trainees' reactions are useful criteria to evaluate training programs. In particular, practitioners should examine participant reactions in terms of utility rather than affective reactions such as whether the participant enjoyed the training. Furthermore, practitioners should consider whether their reaction forms collect utility judgments or trainees'

reactions to whether their training can be used on the job and has merit, and these should be incorporated into comprehensive reaction forms (Alliger, et al., 1997; Mogan & Casper, 2000).

Implications. The results of this study have several implications for future practice in the field of human resource development. In this study, the success of Kirkpatrick's four-level model may provide some beneficial information that increases the clarity of which training criteria should be selected and how to adequately measure it. However, the implications of the expanded hierarchy model of training evaluation are quite important for training professionals. The most important thing to consider when we assess training evaluation is the background the trainees come from and the environment to which he or she returns. Practitioners using the four-level approach alone will be quite likely to remain terribly uninformed about critical aspects of training effectiveness and will consequently arrive at erroneous conclusions about their training programs (Holton, 1996). For training evaluation, if the extent of behavior does not improve as intended, we should examine the amount and types of learning that occurred. However, we should also think about the opportunities that trainees have had to use the training on the job. Furthermore, if organizational results such as improved productivity do not occur, we should examine the quality of job behavior improvement.

Limitations and future research. Although this study led to some important results, several limitations should be discussed. First, this study relied on self-assessment measures, which may have caused some common-method variance problems that may inflate observed relationships between variables. Future studies may consider using a research design in which multiple sources of data collection are used, such as from direct supervisors. Second, this study controlled for a variety of course features in the analysis. For reasons of confidentiality we were not able to control for demographic variables that may influence trainees' experiences and evaluation of the training they received, such as their age, gender, income, and hierarchical position. Third, although the use of data collected more than one year after the end of a training program is acceptable for examining reactions to the training, the cross-sectional nature of the data involved prevented rigorous testing of the causal relationship between the dependent and independent variables of interest in the study. Finally, although this study is based on a varied sample of companies, trainees, and types of training courses, the extent to which the results can be generalized to other cultural and institutional contexts remains open to question. Thus, future research should seek to examine the extent to which the present results can be reproduced in

different countries and should cover a full set of individual controls. Moreover, we also note that future research should incorporate questions that address trainee expectations about the program and how their expectations about the program were met. The study also suggests the need for better integration of work environment and individual characteristic variables in Kirkpatrick's model to better understand training effectiveness.

6. Conclusions

In conclusion, the result of this study expands our understanding of the progressive causal relationship of reaction, learning, and behavior to results. In particular, this study highlighted the utility reactions in predicting training effectiveness. Although additional research is required, this study takes a step toward a more comprehensive understanding of training effectiveness. Furthermore, future research on training evaluation should consider individual trainee characteristics and environmental variables beyond the training course that may have interfered with the results.

References

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of criteria: Thirty years later. *Personnel Psychology, 42,* 331-341.

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

Arthur, W., Bennett, W., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88,* 234-245.

Baldwin, T. T. & Ford, J. K. (1988). Transfer of Training: a review and directions for future research. *Personnel Psychology, 41,* 63-105.

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

Bates, R. and Coyne, T. H. (2005). Effective evaluation of training: Beyond the measurement of outcomes. Paper presented at the Academy of Human Resource Development International Conference (AHRD) (Estes Park, CO., Feb 24-27, 2005), 371-378 (Symp. 16-1).

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: John Wiley & Sons.

Bradenburg, D. C. (1982). Training evaluation: what's the current status?. *Training and Development Journal, 36*(8), 14-19.

Brinkerhoff, R. O. (1986). *Achieving results from training.* San Francisco: Jossey-Bass.

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.

Brown, K. G., & Gerhardt, M. W. (2002). Formative evaluation: An integrated practice model and case study. *Personnel Psychology, 55,* 951-983.

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.

Cascio, W. (1989). Using utility analysis to assess training outcomes. In I. L. Goldstein (ed.). *Training and development in organization,* 63-88. San Francisco: Jossey-Bass.

Clement, R. W. (1978). *An empirical test of the hierarchy theory of training evaluation.* Unpublished doctoral thesis, Michigan State Univesity.

Clement, R. W. (1982). Testing the hierarch theory of training evaluation: An expanded role for trainee reactions. *Public Personnel Management Journal, 11*(2), 176-184.

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,* 678-707.

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

Ford, J. K., & Wroten, S.P. (1984). Introducing new methods for conducting training evaluation and for linking training evaluation to program redesign. *Personnel Psychology, 37,* 651-666.

Fromkin, H. L., Brandt, J., King, D. C., Sherwood, J. J., & Fisher, J. (1975). An evaluation of human relations training for police. *Catalog of Selected Documents in Psychology, 5,* 206-207.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factory analysis. *Structural Equation Modeling, 3,* 62-72.

Giangreco, A., Sebastiano, A., & Peccei, R. (2009). Trainees' reactions to training: An analysis of the factors affecting overall satisfaction with training. *The International Journal of Human Resource Management, 20*(1), 96-111.

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

Hook, K., & Bunce, D. (2001). Immediate learning in organizational computer training as a function of training intervention affective reaction, and session impact measures. *Applied Psychology: An International Review, 50,* 436-454.

Hu, L. T., & Bentler, P. M. (1995). *Structural Equation Modeling: Concepts, Issues and Applications.* CA: Sage.

Joreskog, K. G., & Sorbom, D. (1993). *LISREL 9.10.* Morresville, IN: Scientific Software.

Kirkpatrick, D. L. (1967). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: A guide to human resources development.* New York: McGraw-Hill.

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

Kirkpatrick, D. L. (1996). Invited reaction: Reaction to Holton article. *Human Resource Development Quarterly, 7,* 23-25.

Kirkpatrick, D. L. and Kirkpatrick, J.D. (2006). *Evaluating Training Programs: The four levels.* San Francisco: Berrett-Koehler.

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,* 311-328.

Latham, G. P., Wexley, K. N., & Purcell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60*(5), 550-555.

Leach, M. P., & Liu, A. H. (2003). Investigating interrelationships among sales training evaluation methods. *Journal of Personal Selling & Sales Management, 13*(4), 327-339.

Lin, Y. T., Chen, S. C., & Chuang, H. T. (2011). The effect of organizational commitment on employee reactions to educational training: An evaluation using the Kirkpatrick four-level model. *International Journal of Management, 28*(3), 926-938.

Mann, S., & Robertson, I. T. (1996). What should training evaluations evaluate?. *Journal of European Industrial Training, 20*(9), 14-20.

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

Morgan, R. B., & Casper, W. J. (2000). Examining the factor structure of participant reactions to training: a multidimensional approach. *Human Resource Development Quarterly, 11*(3), 301-317.

Noe, R. A., & Schmitt, N. (1986). The influence of trainee attitudes on training effectiveness: Test of a model. *Personnel Psychology, 39,* 497-523.

Parker, B. L. (1986). Summative evaluation in training and development. *Journal of Industrial Teacher Education, 23*(2), 29-55.

Swanson, R. A., & Sleezer, C. M. (1987). Training effectiveness evaluation. *Journal of European Industrial Training, 11*(4), 7-16.

Tan J. A., Hall, R. J., & Boyce, C. (2003). The role of employee reactions in predicting training effectiveness. outcomes. *Human Resource Development Quarterly, 14*(4), 397-411.

Tannenbaum, S. I., & Woods, S. B. (1992). Determining a strategy for evaluating training: Operating within organizational constraints. *Human Resources Planning, 15*(2), 63-81.

Tracey, J. B., Hinkin, T. R., Tannenbaum, S., & Mathieu, J. E. (2001). The influence of individual characteristics and the work environment on varying levels of training outcomes. *Human Resource Development Quarterly, 12*(1), 5-23.

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

Velada, R., Caetano, A., Michel, J. W., Lyons, B. D., & Kavanagh, M. J. (2007). The effect of training design, individual characteristics and work environment on transfer of training. *International Journal of Training and Development, 11*(4), 282-294.

Warr, P., & Bunce, D. (1995). Trainee characteristics and the outcomes on open learning. *Personnel Psychology, 48,* 347-375.

Warr, P., Allan, C., & Birdi, K. (1999). Predicting three levels of training outcome. *Journal of Occupational and Organizational Psychology, 72*(3), 351-375.

Xiao, J. (1996). The relationship between organizational factors and the transfer of training in the electronics industry in Shenzhen, China. *Human Resource Development Quarterly, 7*(1), 55-73.