International Journal of Management, Economics and Social Sciences
2019, Vol. 8(4), pp. 266 – 279.
ISSN 2304 – 1366
http://www.ijmess.com

DOI:10.32327/IJMESS.8.4.2019.17

Evaluating A Continuing Medical Education Program: New World Kirkpatrick Model Approach

***Shih-Chieh Liao[1]**
Shih-Yun Hsu[2]

[1]School of Medicine, China Medical University, Taichung City, Taiwan
[2]College of Intelligence, National Taichung University of Science and Technology, Taichung City, Taiwan

The New World Kirkpatrick (NWKM) four-level model is a new vision of the Kirkpatrick Model. NWKM adds new elements to recognize the complication of the educational program background and to evaluate the effectiveness of continuing education. This study used data collected from subjects, distributed to 393 participants enrolled in an acupuncture training program in Taiwan from 2010 to 2017, to explore the implication of NWKM for evaluating the effectiveness of continuing medical education and to discuss the connection and transition among the four levels of NWKM. Exploratory factor analysis was used to address that the items in the survey were grouped in different categories and mapped onto the four levels of the NWKM. Path analysis was used to describe the directed dependencies among the levels of NWKM. The results of path analysis showed that a positive relationship exists between any two levels, but direct effects can be observed only between two consecutive levels. It means that L4 outcomes can only be directly predicted by L3, but neither L1 nor L2. L4 is the ultimate outcome of evaluating the effectiveness of continuing education, but it is hard to achieve. This research concluded that L3 is the key to evaluate continuing medical education.

Keywords: Continuing medical education, new world Kirkpatrick model, curriculum evaluation, acupuncture, exploratory factor analysis

JEL: I19, I21

The purpose of continuing education is to promote the employee's professional abilities to enhance the efficiency of the employing organization (Noe, 2016). To understand and increase the effectiveness of continuing education, different methods are used to examine the design, development, implementation, and outcome of the training programs (Wang and Wilcox, 2006). Researchers believe that a comprehensive evaluation of a training program should include appraisal before training, curriculum design and development, and after training (Goldstein and Ford, 2002). Different types of process data or outcome data are gathered depending on the chosen evaluation approaches (Blanchard and Thacker, 2007).

Kirkpatrick's four-level model (hereafter, KM), one of the most recognized project evaluation frameworks, emphasizes that clinical outcomes are the highest level of impact that educational interventions can achieve. Without evaluating the effectiveness of an educational program, Wong and

Manuscript received June 1, 2019; revised October 18, 2019; accepted November 27, 2019. © The Author(s); CC BY-NC; Licensee IJMESS
*Corresponding author: liaosjay@yahoo.com

Holmboe (2016) argue that few medical educational programs can successfully improve the treatment outcomes of patients. Therefore, there has been a recent call for application of KM model in evaluating the effectiveness of continuing medical education (Moreau, 2017; Sultan *et al.*, 2019; Yardley and Dornan, 2012).

Although KM remains the most commonly used model for evaluating continuing education and training (Rafiq, 2015), there are some criticisms that the original KM has faced. First, the links between the levels are not strong and a causal relationship cannot be assumed (Alliger and Janak, 1989; Bates, 2004; Dixon, 1990). Second, as the level increases, their actual usage in continuing education evaluation decreases (Chevalier, 2004; Van Buren and Erskine, 2002). Third, the original KM does not provide an evaluator with insights into the underlying mechanisms that inhibit or facilitate the achievement of the educational program's results (Parker *et al.*, 2011).

In response to the criticism, Kirkpatrick and Kirkpatrick (2016) modified the original KM to develop the New World Kirkpatrick Model (hereafter, NWKM). NWKM adds new concepts to recognize the complication of the educational program background and to improve the authority for completely evaluating the effectiveness of continuing education.

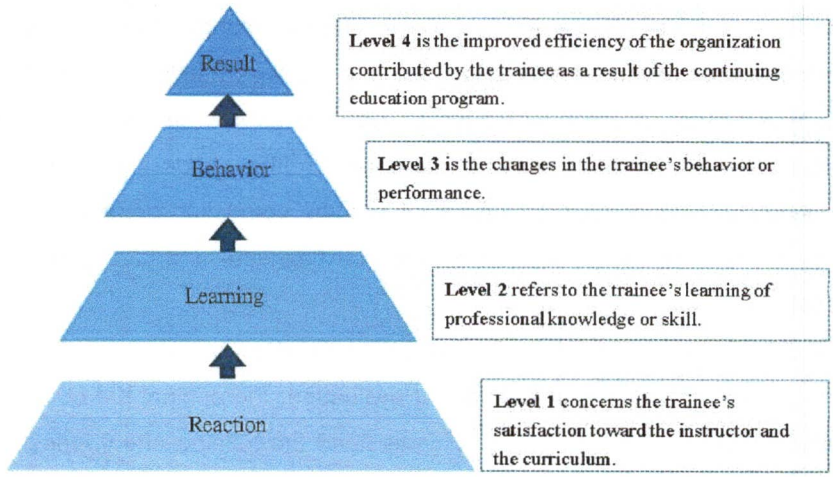
Therefore, the aim of this study is to apply NWKM to evaluate the effectiveness of continuing medical education as well as to understand the relations among the four level outcomes of NWKM. In other words, the major research question is to examine how lower level outcomes of NWKM might directly or indirectly predict higher level outcomes of NWKM.

LITERATURE REVIEW

The Original Kirkpatrick Model

KM was developed by Kirkpatrick (1959) to evaluate the effectiveness of continuing education (Praslova, 2010). As a type of outcome data evaluation, KM emphasizes the understanding of the training outcomes such as satisfaction toward the instructor, knowledge or skill gained, attitude or performance changed, and improved gains of the organization (Werner and DeSimone, 2011). These outcomes are graded into 4 levels depending on the amount of time required to achieve. Level 1 (L1), the reaction level, concerns the trainee's satisfaction toward the instructor and the curriculum. Level 2 (L2), the learning level, refers to the trainee's learning of professional knowledge or skill. Level 3 (L3), the behavior level, is the changes in the trainee's behavior or performance. Level 4 (L4), the result level, is the improved efficiency of the organization contributed by the trainee as a result of the continuing education program (Alliger and Janak, 1989; Kirkpatrick, 1998) (See Figure 1).

Original Kirkpatrick Model: Problems and Criticism



Source: Kirkpatrick (1998)

Figure 1. Levels of Kirkpatrick's Original Evaluation Model

It would be best if information from all four outcome levels of the original KM could be gathered and analyzed. However, due to man power and financial concerns, behavior and result levels of outcomes are not as often obtained as those of reaction and learning levels (Geber, 1995). The 2002 State-of-the-Industry Report (SIR) conducted by the American Society for Training and Development (ASTD) shows that 78 percent of the institutes surveyed evaluated the reaction level outcome (L1), 32 percent evaluated learning level outcome (L2), 19 percent evaluated behavior level outcome (L3), and only 7 percent evaluated result level outcome (L4) (Arthur *et al.*, 2003; Van Buren and Erskine, 2002). In medical education, most of KM implications of evaluating educational program reveal at L2 (learning) (Sultan *et al.*, 2019; Yardley and Dornan, 2012). It is not clear whether a high satisfaction rate in L1 (reaction) and L2 (learning) outcome would bring about an equally high satisfaction in L3 (behavior) and L4 (result) outcome.

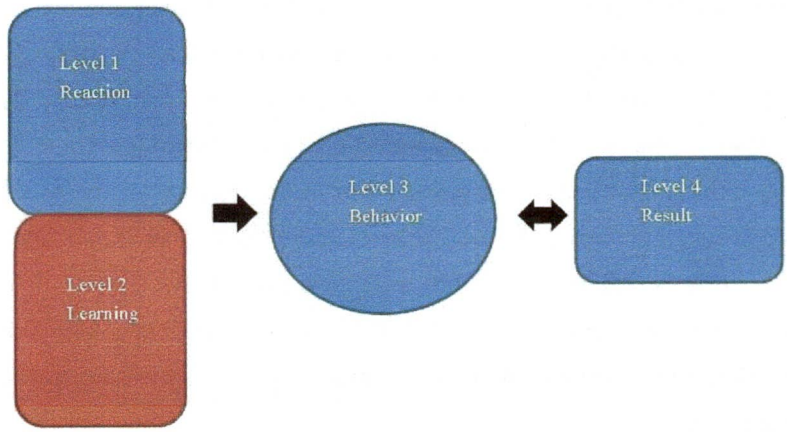
Noe (2016) argued that L1 (reaction) and L2 (learning) level outcome cannot be considered an index of training translation. In other words, results of L1 (reaction) and L2 (learning) evaluation cannot necessarily predict trainees' performance and attitude changes, how trainees apply the learning to problem solving at work, or what influences the training program might bring to the institute. Emphasizing L1 (reaction) and L2 (learning) and neglecting L3 (behavior) and L4 (result) might create happy participants but would not really bring about institute development.

The New World Kirkpatrick Model

Based on the original KM, the New World Kirkpatrick Model (NWKM) redefines the 4 levels of outcomes

and provides new explanations (Kirkpatrick and Kirkpatrick, 2016). In the NWKM (Figure 2), two new views are proposed. First, the original KM claims that the evaluations of continuing education can be separated into four levels, but the NWKM questions that the links between the levels are unclear. The NWKM, hence, argues that L1 (reaction) and L2 (learning) should be considered to be one larger category while L3 (behavior) and L4 (result) is the other. In other words, the NWKM contends that the correlation between L1 (reaction) and L2 (learning) and the correlation between L3 (behavior) and L4 (result) would be higher than the correlation between L2 (learning) and L3 (behavior) (Kirkpatrick and Kirkpatrick, 2016).

Second, the definitions of different level outcomes need to be determined backwards from L4 (result) to L1 (reaction). As an outcome model of evaluation, the original KM observes and measures the outcome of continuing education so as to understand the training effectiveness and ultimately to improve the overall performance of an institute. L4 (result) should indicate most important and desired outcome of training. In the NWKM, L4 (result) outcome is decided first and then the rest of the levels are defined following it. In this way, results of evaluation can provide better information and washback for the design and conduction of continuing education.



Source: Kirkpatrick & Kirkpatrick (2016)

Figure 2. The New World Kirkpatrick Model

Conceptual Framework

Kirkpatrick's original model (KM) is widely used for evaluating continuing education. New World Kirkpatrick Model (NWKM) expands the scope of the original KM by adding concepts and process measures to enable educators to interpret the results of evaluation, but with the aim of proving educational programs (Gandomkar, 2018). There are two major differences between the original KM

and NWKM.

1. In the NWKM, the outcomes of L4 is decided first and then the rest of the levels are defined following it.
2. The original KM claims that the four evaluation levels are separate, but NWKM argues that L1 and L2 should be considered to be one larger category while L3 and L4 is the other.

Based on the conceptual framework and the purpose of this study, this study proposes two hypotheses:

H_1: L1 (reaction) and L2 (learning) might be better considered as one category and L3 (behavior) and L4 (result) as the other.

H_2: L1 (reaction), L2 (learning), and L3 (behavior) outcomes might directly predict L4 (result).

METHODOLOGY

-Research Design

In order to apply NWKM to evaluate the effectiveness of continuing medical education as well as to understand the relationships among the outcomes of NWKM's different levels, the participants of an acupuncture training program offered by China Medical University were surveyed with endorsement from the Health Department of Taiwan.

The acupuncture training program is specially designed for physicians and dentists who are trained in Western medicine. The program contains a total of 192 hours of training on theoretical and philosophical perspectives of Chinese medicine and acupuncture, acupuncture skills, and bedside teaching. Table 1 (see Appendix-I) shows the course contents and number of hours for each topic. Upon finishing the program, trainees are allowed to take the acupuncture specialization certification test. After successfully passing the test, they can apply acupuncture in combination with their Western medicine medical practice.

-Measurement

Based on the research objective, a questionnaire was developed and face and content validities were ensured by three experts, including one medical educator (who is the main program designer and also a Chinese medicine physician with more than 30 years of clinical and medical continuing education experience) and two human resource experts (one of those is the first author and both experts have more than 15 years of human resource management experiences). The questionnaire contained 25 5-point Likert scale items (5 = Strongly agree, 1 = Strongly disagree). The items were designed following

the guidelines of NWKM and related literature (Alliger and Janak, 1989; Kirkpatrick, 1998; Kirkpatrick and Kirkpatrick, 2016; Werner and DeSimone, 2011). To further improve content and face validities, we asked the trainees, who were enrolled in the acupuncture training program a year before, to highlight any issue in questionnaire items. After collecting responses from study subjects, we revised the content and wording of the survey based on their feedback.

-Research Ethics

The study received an exemption recommendation (CMUH REC No. CRREC-107-078) from the Research Ethics Committee, China Medical University and Hospital, Taichung, Taiwan. A signed informed consent to participate was obtained from each participant and the rights about confidentiality, anonymity, voluntary withdrawal from study, and disposal of material containing personal information after the completion of the study were explained and assured to study subjects.

-Participants

A total of 393 (295 male) trainees enrolled in 14 different sessions of the acupuncture training program offered by China Medical University during the years between 2010 and 2017 were selected as study subjects. All of the 393 trainees were invited to participate in this research. The questionnaire was distributed to all the participants at the end of each session of the program. All of the study objectives, methods, and procedures of data collection were explained to the participants at the time of the survey.

Data Analysis

Among the 393 participants, 159 (124 male) valid surveys were collected with an effective response rate of 40.46 percent. Survey data were analyzed statistically by using exploratory factor analysis (EFA), chi-square test, one-way ANOVA, Pearson correlation, and path analysis. Due to the reason that NWKM has never been applied in evaluating an acupuncture training program, EFA was used to categorize participants' responses into a small number of main factors which could later be mapped to the four levels of NWKM. Chi-square was used to test whether the gender difference exists between male and female participants. One-way ANOVA was used to test if there is significant difference in the participants' satisfaction rate. Pearson correlation was used to understand the strength of every two consecutive NWKM levels (Lenhard and Lenhard, 2014). Finally, path analysis was conducted to check both direct and indirect effects among the NWKM levels.

According to Kirkpatrick and Kirkpatrick (2016), the role of human resource experts is critical in successfully implementing the model due to the reason that it requires expertise to assess the learning effectiveness and analyze final results of the training. Therefore, in the development of the survey

and the process of data analysis, we consulted the main program designer (also a Chinese medicine physician and educator), two human resource experts, and two participants to describe the learning transfer processes.

Factor Analysis

Principle Component Analysis with Varimax rotation method was conducted on the 25 items. The Kaiser-Meyer-Olkin (KMO) statistic was 0.899 and Bartlett's chi-square value was 3566.853 (df = 300; $p < 0.01$, a meritorious interpretation). Based on the Scree plot and eigenvalue criteria, three main factors were found explaining 65.84 percent of the variance. After consulting with two human resource experts, three items were deleted. The remaining 22 items had a Cronbach's alpha value of 0.940. Based on the same factor analysis procedures described above, the new KMO statistic was increased to 0.903 (Bartlett's χ^2 =3244.538; df = 231; $p < 0.01$, a marvelous interpretation) and the 3 factors explained 72.05 percent of the variance (see Table 2-Appendix-II).

RESULTS

When mapped onto NWKM, Factor 1 was considered to be L1 (reaction) outcome level with 12 items (mean = 4.204, SD = 0.517). Factor 2 was L2 outcome level with 4 items (mean = 4.326, SD = 0.563). Factor 3 could be mapped onto L3 (behavior) and L4 (result). Since NWKM has 4 levels, we divided Factor 3 into L3 (behavior) with 3 items (mean = 4.072, SD = 0.742) and L4 (result) also with 3 items (mean = 3.675, SD = 0.864). As the outcome level goes up from L1 to L4 (result), the participants' satisfaction rate goes down. One-way ANOVA revealed that other than L1 (reaction) and L2 (learning), there were significant differences in the participants' satisfaction rate in between any two consecutive NWKM levels ($p < 0.05$).

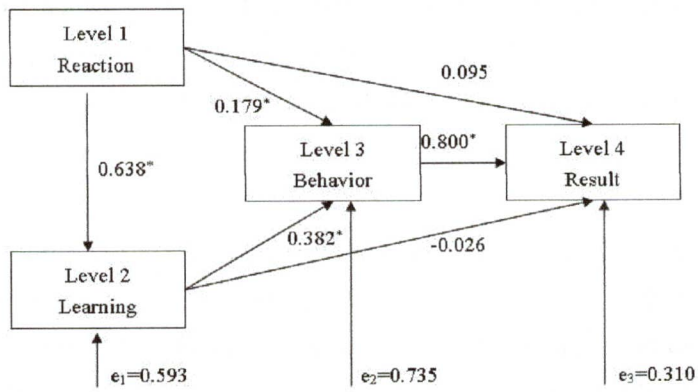
The relationship among the four levels in NWKM

All of the Pearson correlation coefficients were significant at $p < 0.01$. L1 (reaction) to L2 (learning) (r_{12} = 0.638), L1 (reaction) to L3 (behavior) (r_{13} = 0.423), L1 (reaction) to L4 (result) (r_{14}= 0.416), L2 (learning) to L3 (behavior) (r_{23} = 0.496), L2 (learning) to L4 (result) (r_{24}= 0.431),and L3 (behavior) to L4 (result) (r_{34} = 0.827). Comparisons of correlations revealed that the correlation between L1 (reaction) and L2 (learning) was higher than that between L2 (learning) and L3 (behavior), and the correlation between L3 (behavior) and L4 (result) was higher than that between L2 (learning) and L3 (behavior) ($p < 0.05$).

Path Analysis

The results of path analysis showed that L1 (reaction) had direct effect on L2 (learning) (β = 0.638, p

$p < 0.001$) and L3 (behavior) (β = 0.179, p = 0.047). L2 (learning) had direct effect on L3 (behavior) (β = 0.382, $p < 0.001$), and L3 (behavior) had direct effect on L4 (result) (β = 0.800, $p < 0.001$). No direct effect was found from either L1 (reaction) (β = 0.095, p = 0.109) or L2 (learning) (β = −0.026, p = 0.677) to L4 (result). L1 (reaction) also had an indirect effect of 0.638*0.382=0.244 on L3 (behavior) through L2 (learning), making a total of effect of 0.244+0.179=0.423. The indirect effect from L1 (reaction) to L4 (result) through L2 (learning) and L3 (behavior) was 0.638*0.382*0.800=0.195; the indirect effect from L2 (learning) to L4 (result) through L3 (behavior) was 0.382*0.800=0.306. Only L3 (behavior) had direct effect on L4 (result). Figure 3 shows the results of path analysis.



* $p<0.05$

Source: Study Analysis

Figure 3. Path Analysis

DISCUSSION

Based on the results, hypothesis one is confirmed. L1 (reaction) and L2 (learning) might be better considered as one category while L3 (behavior) and L4 (result) are considered as the other. Hypothesis two, however, is partially confirmed. It was found that only L3 (behavior) can directly predict L4 (result), and L1 (reaction) and L2 (learning) can only indirectly predict L4 (result). The results are discussed below.

L1 and L2 is considered as one category and L3 and L4 as the other

Among the 3 factors identified by factor analysis, Factor 1 could be mapped to L1 (reaction) of KM, which described the satisfaction toward the instructor and the curriculum, and Factor 2 mapped to L2 (learning), referring to knowledge and skill growth. Factor 3 was mapped to both L3 (behavior) and L4

(result), referring to behavior change and contribution to medical service. Then we wondered whether it was necessary to make L3 (behavior) and L4 (result) two separate outcome levels since they represented the same factor.

Careful scrutiny told us that distinction between L3 (behavior) and L4 (result) was necessary because they each had clearly defined outcomes to be measured. L4 (result), the ultimate desired outcome, concerns the growth and enhancement that the training brings to the institute, while L3 (behavior) concerns personal behavior change (Kirkpatrick and Kirkpatrick, 2016).

The fact that Factor 3 includes both L3 (behavior) and L4 (result) brings us back to the first research hypothesis, which is to understand whether L1 (reaction) and L2 (learning) are better considered one category and L3 (behavior) and L4 (result) the other. The answer to this question can be sought again from correlation analysis results.

The correlation coefficient between L2 (learning) and L3 (behavior) (r_{23} = 0.496) is significantly smaller than that between L1 (reaction) and L2 (learning) (r_{12}= 0.638) and that between L3 (behavior) and L4 (result) (r_{34} = 0.827). These statistical findings strongly suggest that L1 (reaction) and L2 (learning) can be considered one category and L3 (behavior) and L4 (result) another category.

The experts' opinion lent another support to the claim of two larger categories. The main program designer and the 2 participants pointed out that when trainees were satisfied with the instructor and the training curriculum (L1), the learning experience was positive and trainees often progressed in professional knowledge and skills (L2). However, as often constrained by the complex reality of the medical practice environment, the physician's professional knowledge and skills (L2) did not always or automatically translate into changes in medical practice (L3). In other words, there was a gap between L2 (learning) and L3 (behavior). This observation is in line with the claim by research that identify L3 (behavior) as the criterion for continuing education to really transfer to better workplace performance (Alliger *et al*, 1997; Cheng and Hampson, 2008). Based on both statistical analysis results, we argue that L1 (reaction) and L2 (learning) belong to one bigger category and L3 (behavior) and L4 (result) belong to the other.

The results of ANOVA show that while trainees were generally satisfied with the training program, degrees of satisfaction decreased as the KM levels increased. Arthur at al. (2003) also reported the phenomenon of the training effect size decreasing as the outcome levels increasing. An important assumption to draw from such results is that if continuing education is only evaluated up to L2 (learning), it is really hard to know L4 (result) outcome based on L1 (reaction) and L2 (learning) information.

Only L3 (behavior) can directly predict L4's outcomes (result). L1 (reaction) and L2 (learning) can only

indirectly predict L4 (result).

Path analysis results show that direct effects can only be observed from each level to the immediate next, not the level two or more spots higher (see Figure 2). Path analysis results also addresses the second research purpose that only L3 (behavior) has direct effect on L4 (result); L1 (reaction) and L2 (learning) have only indirect effect on L4 (result) through L3 (behavior). In response to the second research hypothesis, we claim that L1 (reaction) and L2 (learning) outcome cannot directly predict L4 (result) results. The two participants interviewed said that lack of clinical application in the real-life medical practice caused acupuncture skills to dwindle, which might help to explain why L1 (reaction) and L2 (learning) cannot directly predict L4 (result).

L3 (behavior) concerns the application of knowledge and skills learned from continuing education to real world practice in the workplace. A certain length of time is needed to really see positive L3 (behavior) outcome. Kirkpatrick and Kirkpatrick (2016) have suggested that to increase L3 (behavior) outcome, trainees need to have multiple occasions to apply their learning, and time is also needed for improved performance to happen. If coupled by a reward mechanism, sustained improved performances can be more reasonably expected. If L3 (behavior) outcome is still not satisfactory after all the above-mentioned measures, then the course contents and the curriculum design need to be examined.

The purpose of continuing medical education is to promote health providers' work performance and make the institute more efficient and successful. Based on the data in this study, L1 (reaction) and L2 (learning) in Kirkpatrick model should be considered one category of continuing medical education result, and L3 (behavior) and L4 (result) the other one. Besides, this study also shows that only L3 (behavior) can directly predict L4 (result), suggesting that the evaluation of continuing medical education should reach to at least L3 (behavior) level outcome to better understand how the program might help the trainees and the institute to achieve L4 (result), which is the ultimate goal of continuing medical education. When the evaluation of continuing medical education focuses on L1 (reaction) and L2 (learning) outcome and leaves out L3 (behavior) and L4 (result), it overemphasizes training effectiveness and underemphasizes effective training (Kirkpatrick and Kirkpatrick, 2016).

CONCLUSION

The purpose of continuing medical education is to provide on the job training where health providers can learn more about treatment methods, advance knowledge, and improve their skills to benefit patients in clinical practice. However, enhancing continuing medical education not only requires good courses but also good environment and motivation to enable health providers to apply what they have learned in clinical practices.

This research concludes that L3 is the key of evaluating the effectiveness of medical continuing education. There are two reasons to support our conclusion. First, if L4 (result) is difficult to measure, then L3 (behavior) should be attempted because in this study it was found that L3 (behavior) could directly predict L4 (result). The outcomes of L3 can help educators understand the process of enabling or hindering the application of knowledge or skills.

Second, when NWKM is to be used to evaluate continuing medical education, information about satisfaction toward the instructor and the curriculum (L1) and growth in professional knowledge and skill (L2) helps to improve the curriculum design and the learning environment, but it is not sufficient for understanding the trainees' performance in the workplace and how the institute is going to benefit from the training. On the other hand, the L3 outcome can feed back to improve the evaluation methods of L1 and L2. The feedback of L3 outcome offers educators to understand trainee's confidence and commitment, learner engagement and subject relevance are increased to L 2 and L1, respectively, to expand the scope of assessment of these two levels (Kirkpatrick and Kirkpatrick, 2016).

IMPLICATIONS

Researchers argue that most of implications of Kirkpatrick's model to evaluate the effectiveness of medical education program reveal at L2 (learning) (Yardley and Dornan, 2012; Sultan *et al*, 2019). This study claims that only L3 (behavior) can directly predict L4's outcomes (result), and L1 (reaction) and L2 (learning) can only indirectly predict L4 (result). L3 (behavior) can be considered a transfer criterion of continuing education. Therefore, we suggest that L3 (behavior) should be emphasized in future continuing medical education evaluation.

LIMITATIONS AND FUTURE DIRECTIONS

This research relied solely on the acupuncture program survey data to perform the examination of NWKM as it was applied in continuing medical education. Therefore, the findings and conclusions are limited in generalization. This research is also limited due to the reasons that it did not investigate why trainees' satisfaction decreased as the outcome level increased, and how the outcomes of L3 can be used to improve the evaluation methods of L1 and L2. More research is called for to understand the complex factors that might influence the transition of training outcome between different outcome levels.

REFERENCES

Alliger, G. M. & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology*, 42(2): 331-343.

Alliger, G. M., Tannenbaum, S. I., Bennett, J. W., Traver, H. & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology*, 50(2): 341-358.

Arthur, J. W., Bennett, Jr, 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(2): 234-245.

Bates, R. (2004). A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence. *Evaluation and Program Planning*, 27(3): 341-347.

Blanchard, P. N. & Thacker, J. W. (2007). *Effective Training: Systems, Strategies, and Practices* (4th ed.). New Jersey: Pearson/Prentice Hall.

Cheng, E. W. L. & Hampson, I. (2008). Transfer of training: A review and new insights. *International Journal of Management Reviews*, 10(4): 327-341.

Chevalier, R. (2004). Evaluation: The link between learning and performance. *Performance Improvement*, 43(4): 40-44.

Dixon, N. M. (1990). The relationship between trainee responses on participant reaction forms and posttest scores. *Human Resource Development Quarterly*, 1(2): 129-137.

Gandomkar, R. (2018). Comparing Kirkpatrick's original and new model with CIPP evaluation model. *Journal of Advances in Medical Education & Professionalism*, 6(2): 94- 95.

Geber, B. (1995). Does your training make a difference? Prove it! *Training*, 32(3): 27-34.

Goldstein, I. L. & Ford, J. K. (2002). *Training in Organization: Need assessment development and evaluation* (4th ed.). Belmont, CA: Wadsworth.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of the American Society of Training Directors*, 13, 3-9.

Kirkpatrick, D. L. (1998). *Evaluating Training Programs: The Four Levels* (2nd ed.). Philadelphia, PA: Berrett-Koehler.

Kirkpatrick, J. D. & Kirkpatrick, W. K. (2016). *Kirkpatrick's Four Levels of Training Evaluation*. Alexandria, VA: ATD Press.

Lenhard, W. & Lenhard, A. (2014). Hypothesis tests for comparing correlations. Retrieved February 16, 2019, from https://www.psychometrica.de/correlation.html

Moreau, K. A. (2017). Has the new Kirkpatrick generation built a better hammer for our evaluation toolbox? *Medical Teacher*, 39(9), 999-1001.

Noe, R. A. (2016). *Employee Training and Development* (7th ed.). New York: McGraw-Hill.

Parker, K., Burrows, G., Nash, H., & Rosenblum, N, D. (2011). Going beyond Kirkpatrick in evaluating a clinician scientist program: it's not "if it works" but "how it works". *Academic Medicine*, 86, 1389-1396.

Praslova, L. (2010). Adaptation of Kirkpatrick's four level model of training criteria to assessment of learning outcomes and program evaluation in higher education. *Educational Assessment, Evaluation and Accountability*, 22, 215-225.

Rafiq, M. (2015). Training evaluation in an organization using Kirkpatrick Model: A case study of PIA. *Journal of Entrepreneurship & Organization Management*, 4(3): 1-8.

Sultan, N., Torti, J., Haddara, W., Inayat, A., Inayat, H. & Lingard, L. (2019). Leadership development in postgraduate medical education. *Academic Medicine*, 94 (3): 440-449.

Van Buren, M. E. & Erskine, W. (2002). *The 2002 ASTD State of the Industry Report*. Alexandria, VA: American Society of Training and Development.

Wang, G. G. & Wilcox, D. (2006). Training evaluation: Knowing more than is practiced. *Advances in Developing Human Resources*, 8(4): 528-539.

Werner, J. M. & DeSimone, R. L. (2011). *Human Resource Development* (6th ed.). Mason, OH: South-Western Cengage Learning.

Wong, B. M. & Holmboe, E. S. (2016). Transforming the academic faculty perspective in graduate medical education to better align educational and clinical outcomes. *Academic Medicine*, 91 (4): 473-479.

Yardley, S. & Dornan T. (2012). Kirkpatrick's levels and education "evidence". *Medical Education*, 46, 97-106.

ACKNOWLEDGEMENT

This study was supported by the research grant No. MOST 108-2511-H-039-003-MY2 awarded by the Ministry of Science and Technology, Taiwan. The authors would like to thank the Center of Continuing Education at China Medical University for assistance with data collection. The authors are also grateful to the editor and anonymous reviewers for their useful comments to improve the quality of this paper.

Appendix–I

No.	Course Contents	Hours
1.	History of Chinese medicine	4
2.	Introduction of acupuncture	4
3.	Meridians	12
4.	Acupuncture points	24
5.	Acupuncture techniques	8
6.	Moxibustion	4
7.	Cupping	4
8.	Ear acupuncture	8
9.	Hand acupuncture	4
10.	Scalp acupuncture	4
11.	Electro acupuncture	4
12.	Laboratory experiment	4
13.	Physiology of acupuncture analgesia	4
14.	Modern research of acupuncture	4
15.	Therapeutics	20
16.	Special lectures on Chinese medicine and acupuncture	20
17.	Practice of acupuncture and Moxibustion	60
	Total	**192**

Table 1. Acupuncture Course Contents and Hours

Appendix-II

Denomination of factors	Components	Factor			Mean	SD
		1	2	3		
Reaction (Mean=4.204, SD=0.517)	Immediately reply to the intern complaint	0.894			4.126	0.691
	Inform trainee the precautions and important announcements	0.853			4.220	0.717
	Staff take the initiative to express care and assistance	0.849			4.170	0.695
	Deal with the trainee's problem immediately	0.822			4.164	0.674
	Reduce the incidence of the same problem	0.816			4.113	0.729
	The staff can give the trainees a correct reply	0.799			4.233	0.628
	The trainees feel respected	0.788			4.220	0.643
	Course management	0.531			4.145	0.664
	The schedule of bedside teaching	0.346			3.975	0.864
	Teaching methods	0.327			4.296	0.612
	Teachers have enough knowledge to response your questions	0.266			4.396	0.574
	Teachers can provide students with information for acupuncture learning	0.236			4.396	0.585
Learning (Mean=4.326, SD=0.563)	Upgrading my Chinese Medicine Knowledge		0.823		4.428	0.568
	Improving my acupuncture Knowledge		0.803		4.421	0.544
	Learning acupuncture treatment skills		0.765		4.214	0.732
	Learning acupuncture diagnosis and treatment skills		0.760		4.239	0.660
Behavior (Mean=4.072, SD=0.742)	Learning the clinical application of acupuncture			0.866	4.013	0.864
	Acupuncture knowledge can be combined with my expertise knowledge			0.821	4.063	0.862
	Useful for acupuncturist certificate examination			0.678	4.139	0.833
Result (Mean=3.675, SD=0.864)	Improve medical effectiveness			0.893	3.943	0.880
	Increase the number of my outpatient visits			0.757	3.333	1.077
	Improve patient satisfaction			0.867	3.748	0.907
Self-values (*eigenvalues*)		5.781	5.406	4.663		
Variance explained (%)		26.278	24.571	21.197		

Source: Calculated for this study

Table 2. Factors after Varimax Rotation