International Journal of Medical Science and Health Research

Vol.10, No.05; 2026

ISSN: 2581-3366

Evaluating Medical Educators' Training Program in Myanmar Using the New World Kirkpatrick Model and Path Analysis

Aung Tin Kyaw[1], Ye Phyo Aung[2], Tin Tun I[3] & Tin Tun II[4]

[1]Department of Preventive and Social Medicine, University of Medicine, Magway, Myanmar
[2]Department of Medical Education, Defence Services Medical Academy, Yangon, Myanmar
[3]Department of Biochemistry, Manipal University College, Malacca, Melaka, Malaysia
[4]Department of Human Resources for Health, Ministry of Health, Nay Pyi Taw, Myanmar
Correspondence: Aung Tin Kyaw, Department of Preventive and Social Medicine, University of Medicine, Magway, Myanmar. Tel: +95-997-345-6853

doi: 10.51505/ijmshr.2026.10501 URL: http://dx.doi.org/10.51505/ijmshr.2026.10501

Received: Aug 21, 2026 Accepted: Aug 31, 2026 Online Published: Sep 15, 2026

Abstract

Structured faculty development is important for strengthening the pedagogical, assessment, leadership, and scholarly competencies of medical educators. In Myanmar, Diploma in Medical Education (Dip.Med.Ed.) programmes are delivered by the University of Public Health (UPH), Yangon and the Defence Services Medical Academy (DSMA), but comprehensive model-based evaluation across reaction, learning, behavioural transfer, and institutional results remains limited. This study evaluated medical educators' perceptions of training outcomes using the New World Kirkpatrick Model (NWKM), examined the four-level measurement structure, and estimated model-consistent associations among the four levels. A quantitative analytical cross-sectional study was conducted among 209 Dip.Med.Ed. graduates selected by simple random sampling from a population of 400. Data were collected at one point in time using a 39-item self-administered five-point Likert questionnaire. Mean scores were high for Reaction (4.12±0.52), Learning (4.18±0.53), Behaviour (4.08±0.54), and Results (4.16±0.57). Cronbach's alpha ranged from 0.934 to 0.954. The refined CFA showed $\chi^2/df=2.189$, RMSEA=0.075, CFI=0.881, TLI=0.872, and SRMR=0.050; however, the Fornell–Larcker criterion was not satisfied and the strong inter-construct correlations (r=0.742–0.877) indicated substantial construct overlap. In the path model, Reaction was associated with Learning ($\beta=0.868$, $p<0.001$), Learning with Behaviour ($\beta=0.809$, $p<0.001$), and Behaviour with Results ($\beta=0.617$, $p<0.001$). These coefficients should be interpreted cautiously because all variables were self-reported simultaneously and measurement overlap/common-method variance may have inflated the observed relationships. The findings are compatible with, but do not establish, the sequential NWKM pathway. Longitudinal, multi-source studies using objective behavioural and institutional indicators are needed.

Vol.10, No.05; 2026

ISSN: 2581-3366

Keywords: faculty development, medical educators, New World Kirkpatrick Model, confirmatory factor analysis, path analysis, Myanmar

1. Introduction

Medical educators require more than clinical or public health expertise. Effective teaching also depends on competence in curriculum design, teaching-learning methods, assessment, feedback, communication, leadership, and educational scholarship (Harden & Laidlaw, 2017; Cantillon et al., 2019). Faculty development programmes therefore serve as an important mechanism for strengthening the educational capacity of health-professions institutions. Systematic evaluation is equally important because it determines whether training produces meaningful changes beyond participant satisfaction and immediate knowledge gain (Kirkpatrick & Kirkpatrick, 2006; Steinert et al., 2016).

In Myanmar, formal medical educator development is provided through the Diploma in Medical Education (Dip.Med.Ed.) programmes conducted by the University of Public Health (UPH), Yangon and the Defence Services Medical Academy (DSMA). These programmes address educational theory, curriculum development, teaching-learning strategies, assessment, communication, leadership, and educational research. Earlier evaluations in Myanmar reported high satisfaction and perceived improvements in teaching competencies, but comprehensive evidence on how these outcomes progress from learning to workplace application and institutional impact remains limited (Tin Tun, 2015; Naing et al., 2022).

Kirkpatrick's four-level framework evaluates training through Reaction, Learning, Behaviour, and Results. Reaction concerns participants' satisfaction and perceived relevance; Learning reflects knowledge, skills, and attitudes acquired; Behaviour captures the application of learning in the workplace; and Results represent organisational or institutional outcomes (Kirkpatrick, 1998; Kirkpatrick & Kirkpatrick, 2006). The framework is widely used in health-professions education, yet evaluations often concentrate on the first two levels because behavioural and organisational outcomes are more difficult to measure (Yardley & Dornan, 2012). This limitation is important because favourable reactions and learning do not necessarily guarantee transfer to practice or institutional improvement.

The New World Kirkpatrick Model (NWKM) extends the four-level framework by emphasizing the pathways that connect the levels and the conditions required for transfer. In particular, Behaviour is conceptualised as the crucial transition between learning and results, while organisational support, reinforcement, opportunity, and accountability influence whether learning is applied in practice (Kirkpatrick & Kirkpatrick, 2016). Liao and Hsu (2019), applying a path-analytic approach, similarly reported that Behaviour had the strongest direct relationship with Results and that lower-level outcomes contributed to Results mainly through behavioural transfer.

Vol.10, No.05; 2026

ISSN: 2581-3366

Evidence from faculty development research generally supports positive effects on educators' knowledge, teaching skills, confidence, and professional practice (Masood & Usmani, 2015; Heydari et al., 2019; Ahmad et al., 2023; Kohan et al., 2023; Eyigör & Gürpınar, 2025). However, the organisational effects of faculty development and the structural relationships among Kirkpatrick levels remain comparatively under-investigated, particularly in Myanmar. A model-based evaluation can therefore provide both practical evidence for programme improvement and empirical evidence about whether the NWKM is a useful framework in this setting.

This study aimed to evaluate the effectiveness of medical educators' training in Myanmar by: (1) examining differences in Kirkpatrick evaluation scores across sociodemographic and professional groups; (2) evaluating the measurement properties of the Reaction, Learning, Behaviour, and Results constructs; (3) examining correlations among the four levels; and (4) estimating direct, indirect, and total associations among the four levels using path analysis.

2. Method

2.1. Study design and setting

This was a quantitative analytical cross-sectional study based on the New World Kirkpatrick Model. The study population comprised medical educators who had completed the Dip.Med.Ed. programme at UPH, Yangon or DSMA and were working in medical universities, teaching hospitals, or public health institutions in Myanmar.

2.2 Population, sampling, and data collection

The source population consisted of 400 Dip.Med.Ed. graduates. Using Yamane's formula with N=400 and a 5% precision level, the minimum sample size was 200; allowing 4.5% for non-response yielded a target of 209. Participants were selected using simple random sampling. Data were collected through a structured self-administered online questionnaire distributed through the KoboCollect platform (Yamane, 1967).

2.3. Measurement instrument

The questionnaire collected sociodemographic and professional information and assessed four Kirkpatrick domains. Reaction (11 items) measured satisfaction, relevance, trainer competence, programme organisation, and delivery. Learning (10 items) assessed perceived gains in curriculum design, teaching-learning strategies, educational objectives, assessment, communication, leadership, and workshop preparation. Behaviour (11 items) assessed participants' self-reported application of acquired competencies in teaching, assessment, communication, curriculum work, educational research, and workshop facilitation. Results (7 items) assessed participants' perceptions of changes in teaching practice, student engagement, faculty development participation, curriculum planning, educational research interest, and academic contribution. Thus, Levels 3 and 4 represented perceived behavioural application and

Vol.10, No.05; 2026

ISSN: 2581-3366

perceived results rather than independently observed behaviour or objective institutional outcomes. Items were rated from 1 (strongly disagree) to 5 (strongly agree), and composite mean scores were calculated for each level.

2.4. Statistical analysis

Frequencies and percentages summarised categorical characteristics; means and standard deviations summarised composite Kirkpatrick scores. Independent-samples t-tests and one-way ANOVA assessed differences in scores across demographic and professional categories. Internal consistency was evaluated using Cronbach's alpha. CFA was used to assess the four-factor measurement structure, including standardised factor loadings, composite reliability (CR), average variance extracted (AVE), and model-fit indices. Model fit was described using the chi-square/df ratio, root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker–Lewis index (TLI), and standardised root mean square residual (SRMR). Discriminant validity was assessed using the Fornell–Larcker criterion (Fornell & Larcker, 1981; Ab Hamid et al., 2017). Pearson correlation coefficients examined associations among the four composite scores. Path analysis estimated direct, indirect, and total associations specified by the NWKM. Statistical significance was interpreted using the p-values reported in the thesis analysis; exact p-values are reported where available.

3. Results

3.1. Participant characteristics

A total of 209 medical educators were analyzed. The mean age was 46.19±6.78 years. Participants were predominantly aged 41–50 years (47.4%), female (55.0%), and doctorate/PhD holders (56.5%). Nearly half had at least 16 years of teaching experience (45.5%), and 58.9% had 21 years or more of total service. UPH graduates represented 60.3% of the sample and DSMA graduates 39.7%.

Table 1: Sociodemographic and professional characteristics of participants (n=209)

Characteristics	n (%)
Age 30–40 years	42 (20.1)
Age 41–50 years	99 (47.4)
Age 51–67 years	68 (32.5)
Male	94 (45.0)
Female	115 (55.0)
Master's degree	91 (43.5)
PhD/Doctorate	118 (56.5)

Vol.10, No.05; 2026

ISSN: 2581-3366

Characteristics	n (%)
Basic science expertise	51 (24.4)
Clinical expertise	76 (36.4)
Other expertise	82 (39.2)
Lecturer and below	91 (43.5)
Associate Professor	65 (31.1)
Professor and above	53 (25.4)
Teaching experience ≤5 years	38 (18.2)
Teaching experience 6–10 years	39 (18.7)
Teaching experience 11–15 years	37 (17.7)
Teaching experience ≥16 years	95 (45.5)
Total service ≥21 years	123 (58.9)
UPH	126 (60.3)
DSMA	83 (39.7)

3.2 Kirkpatrick evaluation scores and reliability

Participants reported high scores across all four levels. Learning had the highest mean score (4.18±0.53), followed by Results (4.16±0.57), Reaction (4.12±0.52), and Behaviour (4.08±0.54). Cronbach's alpha values ranged from 0.934 to 0.954, and the overall 39-item instrument had α=0.981. Although these coefficients indicate very high internal consistency, values around or above 0.95 can also be compatible with item redundancy; therefore, the coefficients should not be interpreted as evidence of measurement quality in isolation. The lowest areas of agreement were related to balancing theory and practice, educational research planning, organising educational workshops, and contributing research papers to conferences.

Table 2: Descriptive scores and measurement reliability/validity

Construct	Items	Mean ± SD	α	CR	AVE	Loading range
Reaction (Level 1)	11	4.12 ± 0.52	0.934	0.925	0.560	0.703–0.811
Learning (Level 2)	10	4.18 ± 0.53	0.952	0.949	0.667	0.733–0.872
Behaviour (Level 3)	11	4.08 ± 0.54	0.954	0.951	0.659	0.716–0.913
Results (Level 4)	7	4.16 ± 0.57	0.936	0.929	0.670	0.693–0.884

3.3 Confirmatory factor analysis and discriminant validity

The initial CFA model had CFI=0.860, TLI=0.850, RMSEA=0.082, and SRMR=0.052. After theoretically justified correlations among selected error terms were introduced, the refined model showed χ^2/df=2.189, RMSEA=0.075, CFI=0.881, TLI=0.872, and SRMR=0.050. Although RMSEA and SRMR were acceptable, CFI and TLI remained below the commonly preferred 0.90 level and therefore indicate only marginal model fit. CR exceeded 0.70 and AVE exceeded 0.50 for all constructs. However, the Fornell–Larcker criterion was not satisfied: several inter-construct correlations exceeded the square roots of AVE. This substantial overlap raises uncertainty about whether the four self-reported domains were empirically distinct in this sample and should be considered when interpreting the subsequent path model.

Table 3: Fit indices for the refined four-factor measurement model

Fit index	Observed	Interpretive benchmark	Interpretation
χ^2/df	2.1887	<5	Acceptable
RMSEA	0.075	<0.08	Acceptable
CFI	0.881	≈0.90 or higher preferred	Marginal/acceptable
TLI	0.872	≈0.90 or higher preferred	Marginal
SRMR	0.050	<0.08	Good

3.4 Differences across participant characteristics

Most sociodemographic characteristics were not consistently associated with evaluation scores. Female participants reported higher Reaction scores than males (4.20±0.49 vs 4.03±0.53; p=0.03). Expertise area differed for Behaviour (p=0.04). Teaching experience was associated with Reaction (p=0.03), Learning (p=0.02), Behaviour (p<0.001), and Results (p=0.03), with the highest scores generally observed among educators with at least 16 years of teaching experience. UPH graduates reported higher Reaction (p<0.001), Learning (p=0.03), and Behaviour (p=0.01) scores than DSMA graduates. The Results difference by institution was not statistically significant in Table 4.10 (p=0.10).

3.5 Correlations and path analysis

All four Kirkpatrick levels were strongly and positively correlated. Reaction correlated with Learning (r=0.843), Behaviour (r=0.766), and Results (r=0.742); Learning correlated with Behaviour (r=0.877) and Results (r=0.836); and Behaviour correlated with Results (r=0.868). All correlations were significant at p<0.01.

Vol.10, No.05; 2026

ISSN: 2581-3366

Table 4: Direct, indirect, and total associations in the path model

Relationship	Direct β	p	Indirect β	p	Total β
Reaction → Learning	0.868	<0.001	—	—	0.868
Reaction → Behaviour	0.093	0.144	0.703	<0.001	0.796
Learning → Behaviour	0.809	<0.001	—	—	0.809
Reaction → Results	0.084	0.214	0.739	<0.001	0.823
Learning → Results	0.286	0.001	0.500	<0.001	0.785
Behaviour → Results	0.617	<0.001	—	—	0.617

Reaction was strongly associated with Learning (β=0.868, p<0.001), and Learning was strongly associated with Behaviour (β=0.809, p<0.001). Behaviour had the largest direct association with Results (β=0.617, p<0.001), while Learning also had a smaller direct association with Results (β=0.286, p=0.001). Direct paths from Reaction to Behaviour and Results were not significant; indirect associations were observed for Reaction with Behaviour (β=0.703) and Results (β=0.739), and for Learning with Results through Behaviour (β=0.500). The model explained 71.1% of Learning, 77.2% of Behaviour, and 78.0% of Results variance. These estimates describe associations within the specified model rather than temporal or causal effects. In view of the high inter-construct correlations, limited discriminant validity, and common self-report method, the magnitude of the path coefficients may partly reflect measurement overlap or common-method variance.

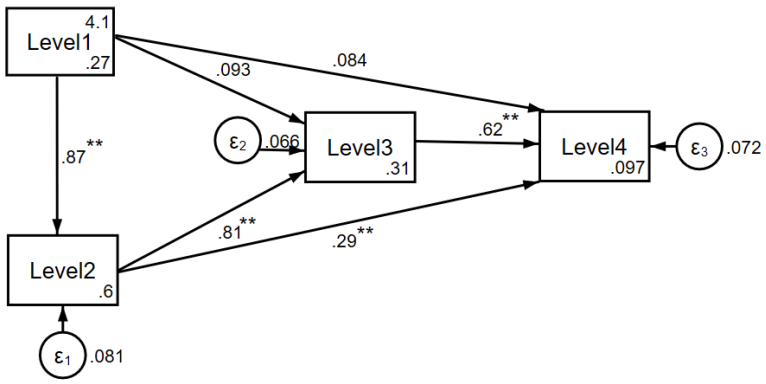

Figure 1: Path model of the four New World Kirkpatrick levels.

4. Discussion

4.1 Principal findings

This study provides a model-based, cross-sectional evaluation of medical educator training in Myanmar across all four Kirkpatrick levels. Participants reported favourable perceptions, with

Vol.10, No.05; 2026

ISSN: 2581-3366

mean scores above 4.0 for Reaction, Learning, Behaviour, and Results. Reliability and convergent-validity indices were high, but the measurement findings also revealed important limitations: CFI and TLI were marginal, discriminant validity was not established, and the constructs were strongly correlated. Accordingly, the observed Reaction–Learning, Learning–Behaviour, and Behaviour–Results pattern should be regarded as compatible with the NWKM rather than as confirmation of a temporal sequence or causal mechanism.

4.2 Measurement properties of the four-level evaluation

Cronbach's alpha, CR, and AVE indicated strong internal consistency and convergent validity. Nevertheless, the very high reliability coefficients require cautious interpretation because coefficients approaching or exceeding 0.95 may also suggest redundancy among items. The fit profile of the refined CFA was mixed: RMSEA and SRMR met commonly used thresholds, whereas CFI=0.881 and TLI=0.872 remained below the conventional 0.90 benchmark. More importantly, the Fornell–Larcker criterion was not met because correlations among the four constructs were very high. This substantial overlap raises the possibility that the questionnaire did not clearly distinguish four empirical dimensions and may partly capture a broader favourable perception of the programme. The overlap may also reflect the theoretical interdependence of the NWKM levels. Future validation should use additional discriminant-validity methods such as HTMT and, where feasible, multi-source or objective indicators.

4.3 Behavioural transfer as the key link to results

The path model showed an association pattern compatible with the NWKM: Reaction was closely associated with Learning, Learning with Behaviour, and Behaviour with Results. Reaction had no significant direct association with Behaviour or Results, whereas indirect associations were substantial. Behaviour had the largest direct coefficient with Results. However, this finding should not be interpreted as demonstrating that Behaviour is a causal determinant or proven mediator of institutional Results. Because the four constructs were measured simultaneously with the same self-report scale and showed limited discriminant validity, part of the structural pattern may reflect measurement redundancy or common-method variance. The pattern is broadly comparable with Liao and Hsu (2019), but longitudinal evidence is required to test the temporal sequence proposed by the NWKM.

The practical implication is that satisfaction and perceived learning, although important, are insufficient endpoints for judging faculty development. A programme may be well received and improve knowledge yet fail to produce sustained change if educators lack opportunities, reinforcement, resources, or institutional support. The NWKM emphasises these transfer conditions, and research in faculty development likewise highlights the importance of workplace context, follow-up, and institutional culture (Kirkpatrick & Kirkpatrick, 2016; Steinert et al., 2016).

Vol.10, No.05; 2026

ISSN: 2581-3366

4.4 Implications for medical education in Myanmar

The descriptive findings indicate that graduates perceived favourable effects of the Dip.Med.Ed. programmes on their teaching practice and institutional contribution; they should not be interpreted as independently verified programme effects. Strong agreement was reported for use of varied teaching strategies, assessment methods, student engagement, communication, and curriculum participation. Lower agreement was seen for planning educational research, organising educational workshops, publishing or presenting research, and some aspects of theory–practice balance. These areas provide practical targets for curriculum strengthening and post-training support.

Associations with teaching experience suggest that more experienced educators reported higher scores on several levels, but the cross-sectional data do not establish why these differences occurred. UPH graduates reported higher Reaction, Learning, and Behaviour scores than DSMA graduates, while the Results difference was not statistically significant. Because this study did not collect sufficiently detailed comparative information on programme delivery, cohort composition, workplace support, or other institutional factors, explanations for the UPH–DSMA differences would be speculative. Future comparative work should document programme structure and participant characteristics more fully before attributing differences to the training environment.

4.5 Limitations and directions for further research

Several limitations should shape interpretation. First, the study was cross-sectional and measured all four NWKM levels simultaneously; therefore, it cannot demonstrate the temporal progression implied by the model or establish causal relationships. Second, all levels were measured using the same self-reported questionnaire and response scale at one time point. Behaviour and Results therefore represent perceived application and perceived outcomes rather than observed behavioural change or objective institutional impact. This design is susceptible to common-method variance, which may partly explain the very high reliability coefficients, strong inter-construct correlations, and large path coefficients. Third, the failure of the Fornell–Larcker criterion and marginal CFI/TLI values indicate limited discriminant validity and raise the possibility of measurement redundancy. Fourth, graduates completed the programme in different years, creating potential recall bias. Time since graduation and cohort distribution were not available for the present analysis, so their effects could not be examined. Fifth, the study did not collect sufficiently detailed comparative data on UPH and DSMA programme structure or participant composition to explain institutional differences. Finally, organisational supports and barriers to transfer were not directly measured.

Future research should use longitudinal designs with measurements before training and at defined follow-up points, incorporate independent or objective Level-3 and Level-4 indicators, and obtain supervisor, student, or institutional data where feasible. Useful outcomes could include teaching-evaluation scores, educational leadership activities, curriculum-development

contributions, research productivity, promotion or role progression, and student-related outcomes. Future studies should also record time since graduation and cohort, assess common-method bias and HTMT, and examine organisational support, workload, resources, mentoring, reinforcement, and opportunity to apply learning.

4.6. Conclusion

Medical educators who completed the Dip.Med.Ed. programmes in Myanmar reported favourable perceptions across Reaction, Learning, Behaviour, and Results. The observed associations were compatible with a Reaction–Learning–Behaviour–Results sequence, and Behaviour showed the largest direct association with perceived Results. However, the cross-sectional, single-source design, marginal CFA indices, substantial construct overlap, and possible common-method variance preclude causal conclusions and may have inflated structural relationships. The findings therefore support the practical value of evaluating faculty development beyond satisfaction and perceived learning, while also demonstrating the need for stronger measurement of behavioural transfer and institutional outcomes. Longitudinal, multi-source evaluation using objective Level-3 and Level-4 indicators is required to confirm sustained programme impact.

References

Ab Hamid, M. R., Sami, W., & Sidek, M. M. (2017). Discriminant validity assessment: Use of Fornell–Larcker criterion versus HTMT criterion. *Journal of Physics: Conference Series*, 890(1), 012163.

Ahmad, S., Imtiaz, M., Aziz, R. S., Manzoor, N., Ahmed, S., & Habib, B. (2023). Kirkpatrick's approach to build faculty development programs in medical education. *Pakistan Journal of Medical & Health Sciences*, 17(3), 441–444.

Cantillon, P., Wood, D., & Yardley, S. (2019). ABC of learning and teaching in medicine (3rd ed.). Wiley-Blackwell.

Eyigör, H., & Gürpınar, E. (2025). Development of a blended clinical education faculty development program in postgraduate medical education and evaluation with the Kirkpatrick model. *Hamidiye Medical Journal*, 6(1), 50–58.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50.

Harden, R. M., & Laidlaw, J. M. (2017). Essential skills for a medical teacher: An introduction to teaching and learning in medicine (2nd ed.). *Elsevier.*

Heydari, M. R., Taghva, F., Amini, M., & Delavari, S. (2019). Using Kirkpatrick's model to measure the effect of a new teaching and learning methods workshop for health care staff. *BMC Research Notes*, 12(1), 388.

Kirkpatrick, D. L. (1998). Evaluating training programs: The four levels (2nd ed.). *Berrett-Koehler.*

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). Evaluating training programs: The four levels (3rd ed.). *Berrett-Koehler.*

Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). Kirkpatrick's four levels of training evaluation. ATD Press.

Kohan, M., Changiz, T., & Yamani, N. (2023). A systematic review of faculty development programs based on the Harden teacher's role framework model. *BMC Medical Education*, 23, 910.

Liao, S.-C., & Hsu, S.-Y. (2019). Evaluating a continuing medical education program: New World Kirkpatrick Model approach. *International Journal of Management, Economics and Social Sciences*, 8(4), 266–277.

Masood, R. Q., & Usmani, M. A. W. (2015). Evaluation of medical teacher's training program through Kirkpatrick's model. *Khyber Medical University Journal*, 7(2), 76–80.

Naing, T. T., Minamoto, Y., Aung, Y. P., & Than, M. (2022). Faculty development of medical educators: Training evaluation and key challenges. *The Asia Pacific Scholar*, 7(3), 23–32.

Steinert, Y., Mann, K., Anderson, B., Barnett, B. M., Centeno, A., Naismith, L., Prideaux, D., Spencer, J., Tullo, E., & Viggiano, T. (2016). A systematic review of faculty development initiatives designed to enhance teaching effectiveness in medical education. *Medical Teacher*, 38(8), 769–786.

Tin Tun. (2015). Effectiveness of faculty development programme: Medical educators' perspectives on Diploma in Medical Education course. Department of Medical Science, Ministry of Health, Myanmar.

Yamane, T. (1967). Statistics: An introductory analysis (2nd ed.). *Harper & Row.*

Yardley, S., & Dornan, T. (2012). Kirkpatrick's levels and education 'evidence'. *Medical Education*, 46(1), 97–106.