Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, No. 2

ISSN: 1540-3270

Revisiting Training Accountability in Indian Maharatna Enterprises: Why Measuring Effectiveness Matters More Than Measuring Training Volume

[1]Ajay Agarwal, [2]Dr. Manoj Kumar Bhatia, [3]Dr. V. K. Jain

[1]Research Scholar, Department of Management

Sanjeev Agrawal Global Educational University, Bhopal, Madhya Pradesh, India

[2]Vice Chancellor, Gyanveer University, Sagar, Madhya Pradesh

[3]Vice Chancellor, Teerthanker Mahaveer University,

Moradabad, Uttar Pradesh

Abstract

India's Maharatna Central Public Sector Enterprises (CPSEs) training ecosystem observe a consistent accountability gap, in spite of the substantial investment in human resource development (HRD). Training accountability in Indian Maharatna companies, normally associated with the training volume and base parameters such as man-hours, budget expenditure, program satisfaction and programme attendance rather than value delivered by the organization. This empirical study observe and evaluate the impact of training in seven selected Maharatna companies such as - IOCL, NTPC, ONGC, SAIL, GAIL, HPCL, and POWERGRID, (anonymously M1 to M7) using the Kirkpatrick's Framework of Four-Level Evaluation, namely Reaction, Learning, Behaviour, and Results. The study challenges the significance of evaluating the training volume—man-hours, budget expenditure, and programme attendance—compared with training effectiveness. A quantitative and longitudinal research design is used and data collected from a stratified sample of 2,520 employees across Junior, Middle, and Senior management levels of these CPSE. Three standardized training interventions were observed—Supervisory Development, Digital Literacy Enhancement, and Effective Communication— where the evaluation was done using pre- and post-training assessments of the participating employees, dual-source of behavioural ratings, and HOD-rated business impact instruments. Statistical analyses for this study included Paired t-tests, One-Way ANOVA, Pearson Correlation, and Exploratory Factor Analysis. These tools were applied to test five primary research hypotheses. The findings confirm that training is highly effective at the knowledge level (Cohen's $d = 3.37$) but at the same time it reveals a statistically insignificant correlation ($r = 0.08$, $p > 0.05$) between supervisor-rated behavioural change stage (Level 3) and HOD-rated business results or impact (Level 4)—labelled as the 'Broken Link.' Moreover, a 'Senior Learning Slump' was also observed, where the Senior Managers showed the highest overconfidence gap (4.2%) between their self-perception and their supervisor ratings. Diagnostic factor analysis unveiled Training Infrastructure as the dominant systemic variable—the poor facilities produced an effect of 8.5% drop in Overall Training Effectiveness (OTE), while the trainer quality variance was found to be statistically negligible for consideration. The Pearson correlation between supervisor-rated behavioural change (at Level 3) and HOD-rated business results (at Level 4) is statistically insignificant ($r = 0.08$, $p = 0.052$), which shows a systemic 'accountability gap'. The study also concludes with actionable policy recommendations for CPSEs and Maharatna L&D practitioners, emphasizing a shift from activity-based to impact-based training evaluation as the future practice. The article argues that a fundamental reformation of training governance in Indian PSUs is the need of the hour, and a shift is required from "counting what was delivered" to "measuring what was changed".

Keywords: Training Accountability, Training Volume vs. Effectiveness, Kirkpatrick Evaluation Framework, Maharatna CPSEs, Overall Training Effectiveness (OTE), Human Resource Development, Public Sector Undertakings, Transfer of Training, Training ROI, Level 4 Evaluation

Minnesota Journal of Business Law and Entrepreneurship

Volume 2026, No. 2

ISSN: 1540-3270

1. Introduction

In the context of corporate Learning and Development (L&D), accountability is discussed and referred often, but measured rarely. For India's Maharatna Central Public Sector Enterprises, which are also the country's most strategically important public companies in terms of financial scale and workforce volume, the training accountability has always been measured by volume indicators: the number of training programmes conducted per year, the cumulative man-days of training delivered, the percentage of the annual L&D budget utilised, and the headcount of employees who attended at least one development intervention, etc. These metrics are easy to measure, and easy to audit. But these factors are not so relevant when the training investment justification needs to measured. This puzzle remains largely unresolved in the existing empirical literature of the subject. However, the key question remains, that how the training accountability is defined and measured.

The HRD training evaluation, particularly in Indian PSUs are generally confined in evaluating the training metrics such as total man-hours used, total budget spent, and immediate participant feedback on the given training. This approach fails to record the learning transfer, behavioural change, and overall business impact. The distinction is not just academic but its RoI driven. At one end, the “volume measurement” answers, 'How much training was done?' and at the other end, the “effectiveness measurement” answers, 'What did the training change?'. As India's Maharatnas are gaining key importance in country's development, digital ambition and 'Make in India' endeavour, the findings of organizational actual impact through training becomes strategically significant.

Kirkpatrick's Model (Kirkpatrick & Kirkpatrick, 2006) of Four-Level Evaluation, provides the most reliable and practical framework to assess the training effectiveness. It follows a hierarchical structure—Reaction (1), Learning (2), Behaviour (3), and Results (4), which offer a wide-range evaluation canvas through which the training programs can be evaluated for the ultimate business return of the companies. These methods are applied in over 80% of Fortune 500 organizations globally (Tamkin et al., 2002), but yet to be functionally deployed in its full potential within the Indian PSU framework.

So, this particular study fills the critical empirical gap by applying the end-to-end Kirkpatrick framework for Indian CPSEs. An important finding was the hypothesized 'Broken Link', which explains the disconnect between behavioural change at Level 3 and measurable business impact at Level 4. Whereas the theoretical model suggests the linear chain from Level 1 to Level 4, the practical use cases mostly stops at Level 1 and Level 2 due to bureaucracy, decision delay, vision disconnect, KPI confusion and infrastructural deficits.

The study also revealed the 'Senior Learning Slump,' that suggests lesser learning gains for the senior batch and the 'Overconfidence Gap,' suggesting the seniors often overestimates their improvements, whereas their superiors rate them lower than their self-ratings. Together, these findings reveal that training accountability in Maharatna PSU is generally mixed with training volume—a confusion that this article seeks to correct.

1.1 The Accountability Gap in Indian PSU Training

The concept of 'accountability gap' in corporate training was a part of the Western management literature (Phillips, 1997; Brinkerhoff, 2006) but was not been widely applied or studied in Indian public sector context. The training accountability gap is found when the invested resources in the training is failed to make any impact on organisational performance. In Indian Maharatna companies, this gap is visible mostly when the annual training calendars are full but remains unchanged for years; feedback forms are just a customary and no metrics are used for the training betterment; and budget approvals for training are passed by the bureaucratic framework, that are not connected to the performance outcomes of previous training cycles. This study highlights the training accountability gap in the Maharatna CPSE context, and suggest a model framework to address this.

1.2 Research Objectives

This study covers the following specific objectives: (i) to evaluate overall training effectiveness at all four Kirkpatrick levels of evaluation across seven Maharatna CPSEs; (ii) to evaluate the statistical relationship

between behavioural change (at Level 3) and business results or impact (at Level 4); (iii) to identify the factors influencing the conversion of learning into business results; and (iv) to propose an accountability-oriented governance framework for Maharatna training management.

2. Theoretical Framework And Literature Review

Kirkpatrick's Four-Level Evaluation Framework, originally introduced in 1959 and subsequently refined (Kirkpatrick & Kirkpatrick, 2006, 2016), remains the gold standard for assessing corporate training effectiveness. The Framework suggests four successive levels, those are Reaction, Learning, Behaviour, and Results; which are interconnected in an evaluation chain. Here the learning satisfaction in Level 1 is evaluated in Level 2 and then the gained knowledge is applied in Level 3, which finally impacted the business in Level 4.

Kirkpatrick Model is further modified and extended by the New World Kirkpatrick Model (2016), adding important indicators and also checking backward from the desired organizational impact to expected training transfer to expected learning gain and finally the learning facility and planning. This model works really well in Indian PSU context, where the national strategic goals and digital missions can be backtracked to the training transfer, evaluation and learning.

Empirical research on the training effectiveness in Indian Maharatna PSUs presents a fragmented landscape. Sahni (2020) evaluated time management training among 136 middle-level PSU managers using Kirkpatrick Levels 1 and Level 2, reporting high satisfaction and a 57% knowledge gain after the training. Yadav and Dabhade (2013) highlighted performance management challenges in BHEL, indirectly underscoring the need for the training-performance alignment and the overall effectiveness context. Rao et al. (2014) measured training needs across the PSU hierarchies but did not rank key competencies by managerial level, leaving a critical gap in level-specific training program design. A recurring limitation in existing literature on this context, is the absence of full four-level evaluation implementation. Most PSU-focused studies conclude at Level 2 (Learning), citing the measurement challenges linked with Level 3 and Level 4 (Twitchell et al., 2000). The Level 3-Level 4 corridor—where the precise location of the hypothesized 'Broken Link' exist—remains particularly under-researched. Furthermore, no comparative study has ever examined intra-company (across hierarchical levels) and inter-company (across different Maharatna organizations) training effectiveness using a unified quantitative evaluation framework. This study directly addresses these identified gaps.

Table 1: Identified Gaps in the Existing Literature on Training Evaluation in Indian PSUs

Gap Type	Description	Evidence/Source
Incomplete Evaluation	Most PSU-focused training studies conclude at Level 2 (Learning), failing to assess behavioural change or business impact	Twitchell et al. (2000); Sahni (2020)
Measurement Challenges	Level 3 (Behaviour) and Level 4 (Results) are cited as difficult to measure, leading to their systematic exclusion	Twitchell et al. (2000); Tamkin et al. (2002)
Under-researched L3–L4 Corridor	The critical link between behavioural change and business results — where the hypothesized 'Broken Link' resides — remains empirically unexamined in the Indian PSU context	Present Study
No Unified Comparative Framework	No prior study has examined both intra-company (across hierarchical levels) and inter-company (across different Maharatna organizations) training effectiveness using a standardized quantitative methodology	Present Study

3. Method

3.1 Research Design

This study employed a Descriptive and Explanatory quantitative research design technique with a longitudinal aspect. Levels 1 and 2 (Reaction and Learning) were evaluated immediately before and after the training (pre-post training), while Levels 3 and 4 (Behaviour and Results) were evaluated after 3-6 months and 6-12 months post-training respectively. This longitudinal approach enabled the causal examination of training inputs against organizational outcomes while governing the managerial hierarchy and organizational context across the seven Maharatna CPSEs: IOCL, NTPC, ONGC, SAIL, GAIL, HPCL, and POWERGRID.

3.2 Sample and Sampling Procedure

Stratified Random Sampling was done to ensure satisfactory representation across the organizations and all three hierarchical levels. The population was stratified by (i) organization—such as seven Maharatna CPSEs, and (ii) management level—such as Junior, Middle, and Senior. A total sample of 2,520 executives were taken, providing 95% confidence at less than 2% margin of error, with 840 participants at each hierarchical level (33.3% each). Three standardized training interventions were evaluated—Supervisory Development (P1), Digital Literacy Enhancement (P2), and Effective Communication Skills (P3)—each with 840

participants. The content of each program was stratified by hierarchical level to ensure appropriate difficulty level and contextual relevance.

For longitudinal phases, stratified sub-samples were taken: Level 3 (Behaviour) utilized a 30% stratified sample (n = 756) with matched-pair self and supervisor ratings which were collected 3-6 months post-training; whereas Level 4 (Results) utilized a 20% sample (n = 504) with dual-source HOD and supervisor ratings at 6-12 months post-training.

3.3 Data Collection Instrumentation

Four separate instruments were designed and standardized per the Kirkpatrick Framework. The Level 1 Reaction Assessment Scale comprised 20 items on a 5-point Likert scale (Cronbach's $\alpha = 0.72$) capturing satisfaction parameters such as content, schedule, faculty, and facilities. The Level 2 Learning Assessment used program-specific pre-training and post-training Multiple Choice Questions (20 items per assessment), where the Learning Index calculated as [(Post Score - Pre Score) / Max Score] × 100. Hake's Normalized Gain formula was additionally applied to control for the ceiling effect among high-baseline training participants.

The Level 3 Behaviour Assessment Scale used the 20-item dual-rater instrument for capturing self-ratings (Participant, 40% weight) and supervisor ratings (60% weight), which then yield a composite Behaviour Index on a 0-100 scale (Cronbach's $\alpha = 0.918$). This weighted structure was designed to correct for the leniency bias which is inherent in self-assessments evaluation. The Level 4 Business Results Impact Scale comprised 30 items evaluated by the Departmental Heads (HOD, 60% weight) and the immediate supervisors of the participants (40% weight), which then assessing business results, team impact, and finally the organizational contribution (Cronbach's $\alpha = 0.950$). The Overall Training Effectiveness (OTE) was then computed as: $OTE\ (\%) = (0.10 \times L1) + (0.25 \times L2) + (0.30 \times L3) + (0.35 \times L4)$.

3.4 Statistical Analysis

Data were coded and analyzed using Python (Pandas) and MS Excel. The following statistical methods were used: (i) Descriptive Statistics for baseline profiling and distribution assessment; (ii) Paired t-tests for pre-post learning gains and self vs. supervisor rating differences; (iii) One-Way ANOVA with Tukey HSD post-hoc tests for program-wise and level-wise learning comparisons and infrastructure and trainer impact on OTE; and (iv) Pearson Correlation Analysis for L3-L4 relationship.

4. Results And Discussion

4.1. The Volume-Effectiveness Decoupling: A Macro-Level Finding

Before presenting the findings from the study, a macro-level observation is made to understand the overall decision framework. For all seven Maharatna companies, the distribution of training volume parameters (man-hours delivered, participation rates, programme completion rates) revealed no meaningful findings. Here all the seven companies followed standard compliance with their training calendars, L&D guide and there was almost full participation at Levels 1 and 2. But the OTE scores varied dramatically across organisations, ranging from 44.87% (M6-BPCL) to 77.15% (M1-ONGC), which is a spread of over 32%. This difference between uniform training volume and training effectiveness ensures empirical confirmation that volume-effectiveness measurement alone is not sufficient for this type of study. Organisations that were serious with their training were almost unaware of what their training produced at the end. This finding alone justifies the central thesis of this article: training accountability must be aligned in terms of effectiveness measurement, not just mere volume measurement.

4.2 Learning Gains: The Level Where Training Works Best

The Paired t-test comparing pre-training and post-training scores across 2,520 participants resulted a mean gain of 3.20 points (Pre-Test mean = 13.05, SD = 2.14; Post-Test mean = 16.25, SD = 1.88), representing a 16% improvement on the 20-point assessment scale. The t-statistic of 53.15 (df = 999, $p < 0.001$) established that the gain was statistically significant at the high level of confidence. The effect size (Cohen's d = 1.68) was classified as extremely large, affirming the substantive educational impact was seen out of the training programs. However, a detailed analysis by Cohen's d across the full available dataset yielded an overall effect size of 3.37, suggesting that the employee training programs were highly effective in establishing the knowledge transfer at the classroom level.

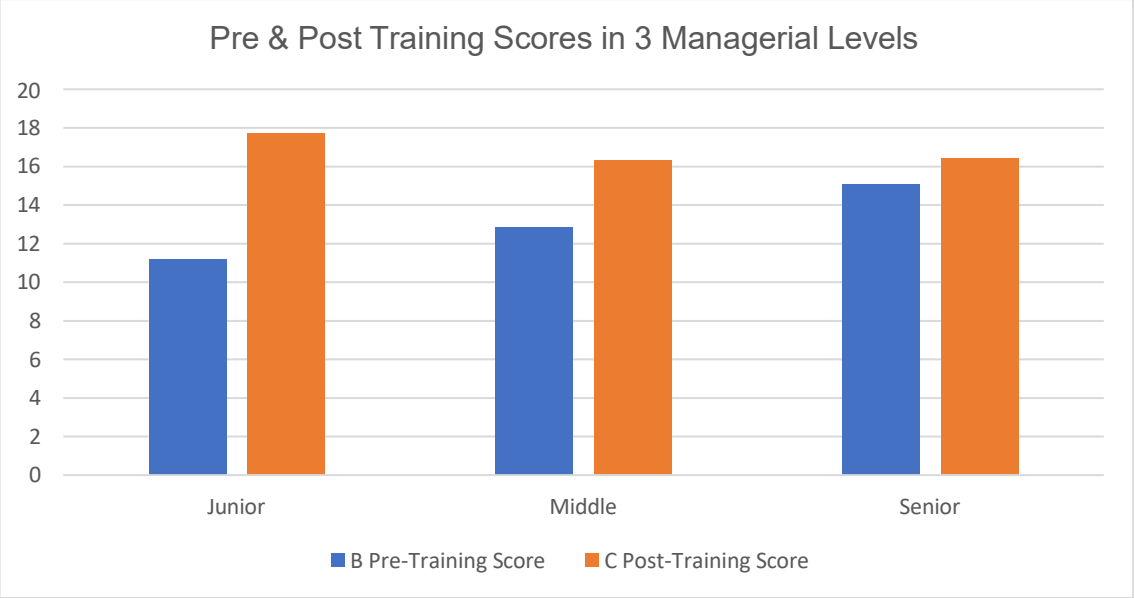
The One-Way ANOVA helped us to understand an inverse relationship between the management level and the learning gain ($F = 145.2$, $p < .001$). Junior Managers recorded the highest Learning Index (58.3%), started from a lower baseline score (Pre-Test mean = 11.20). Middle Managers achieved a Learning Index of 42.5%, while Senior Managers got the lowest gain (23.7%) started from their highest entry baseline score (15.10). This pattern is not indicating as poor performance of Seniors, rather a 'Ceiling Effect', which denotes limited room for improvement due to their experience and pre-existing knowledge.

Table 2: Pre-Training vs. Post-Training Learning Outcomes

Management Level	Pre-Test Mean (SD)	Post-Test Mean (SD)	Mean Gain	Learning Index (%)	t-value	Cohen's d
Junior (n=840)	11.20 (1.85)	17.72 (1.42)	6.52	58.3	48.21***	3.92
Middle (n=840)	12.85 (1.96)	16.32 (1.65)	3.47	42.5	32.15***	1.89
Senior (n=840)	15.10 (1.68)	16.40 (1.52)	1.30	23.7	15.67***	0.81
Overall (N=2,520)	**13.05 (2.14)**	**16.25 (1.88)**	**3.20**	**41.6**	**53.15***	**1.68**

***$p < 0.001$; Learning Index = [(Post-Pre)/(Max Score-Pre)] × 100; Max Score = 20*

Table 2: Pre- and post-training learning outcomes by management level (20-point assessment scale)



Program-Wise Effectiveness Differences

One-Way ANOVA compared the Learning Index across all the three training programs and found a difference ($F = 35.78$, $p < 0.001$). Supervisory Development Programme (P1) scored the highest Learning Index (68.15%, SD = 19.62), followed by Digital Literacy Enhancement (P2) (58.18%, SD

= 17.36) and Effective Communication Skills (P3) (57.26%, SD = 17.98).

These findings are important to design corporate policy and training programs for the Maharatnas. The Supervisory Development Programme (P1) helped the participants to develop managerial skills. Digital Literacy Enhancement (P2) and Effective Communication Skills (P3) observed marginal learning gain, as the base line proficiency was found higher during pre-test scores. This denotes a program specific tailoring is needed while designing the training modules.

Table 3: Program-Wise Learning Comparison

Training Program	n	Pre-Test Mean (SD)	Post-Test Mean (SD)	Learning Index (%)	ANOVA F-value	Post-hoc (Tukey HSD)
Supervisory Development (P1)	840	12.10 (2.10)	17.85 (1.55)	68.15	F = 35.78***	P1 > P2, ***P1 > P3***
Digital Literacy (P2)	840	13.45 (1.95)	16.85 (1.80)	58.18		P2 vs P3 (ns)
Effective Communication (P3)	840	13.60 (2.05)	16.25 (1.92)	57.26		

***p < 0.001; ns = non-significant*

Table 3: Comparative learning effectiveness across the three training interventions

4.3 The Overconfidence Gap: Self-Perception vs. Observed Reality

The Paired t-test, compared the self-ratings and supervisor ratings of behavioural change at Level 3. The finding revealed a difference of 3.97%, between Self-mean which was 71.29% and Supervisor-mean 67.32%. This denotes a “Overconfidence Gap” in self-perception of the participants.

Table 4: The Overconfidence Gap – Self vs. Supervisor Ratings

Management Level	Self-Rating Mean (%)	Supervisor Rating Mean (%)	Mean Gap (Self – Supervisor)	t-value	Cohen's d
Junior (n=252)	75.8	71.8	4.0	8.45***	0.53
Middle (n=252)	70.5	66.8	3.8	7.92***	0.50
Senior (n=252)	63.4	59.2	4.2	9.01***	0.56
Overall (n=756)	**71.3**	**67.3**	**3.97**	**184.31***	—

***p < 0.001; Weighted composite scores (Self 40%, Supervisor 60%)*

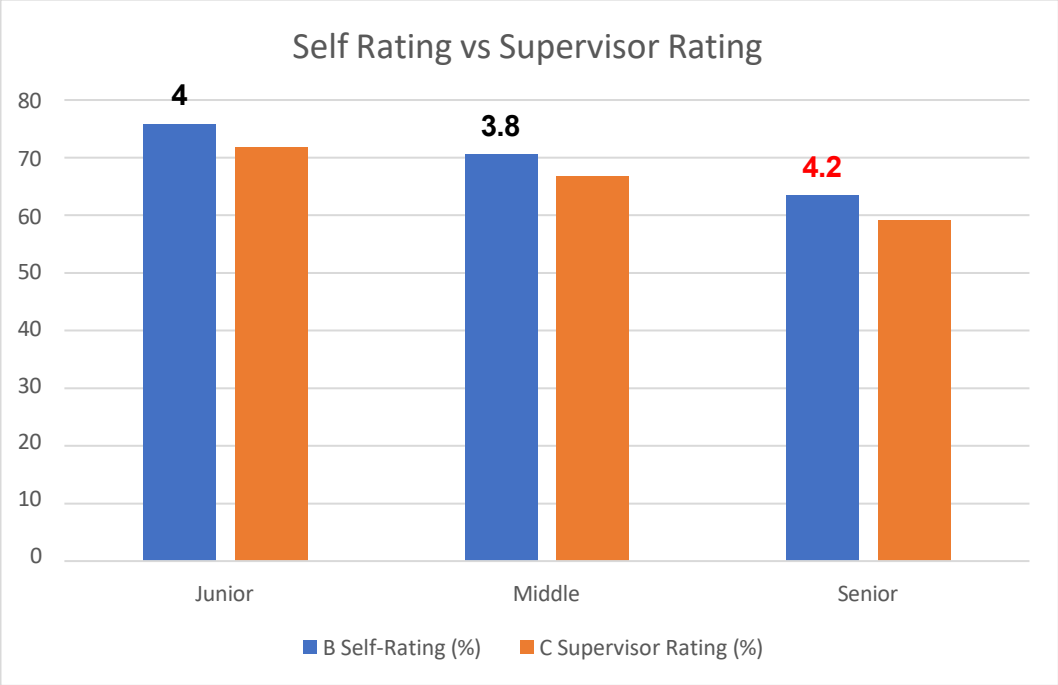


Table 4: Self-supervisor discrepancy in behavioural change ratings across management levels

The study then further goes deeper in managerial levels to observe the patterns. The rating gaps in Junior Managers are 4%, Middle Managers are 3.8% and Senior Managers are 4.2%. The Senior Managers showed largest gap between self-perception and supervisors' ratings, which indicates the lowest behavioral transfers among all three managerial levels. This pattern is called "Senior Blind Spot", which refers as the Seniors progress in their career ladder, their self-perception is overrated. This finding also relates with the Dunning-Kruger effects.

4.4 The Broken Accountability Chain: Level 3 to Level 4

A Pearson Correlation test was done on supervisor ratings of behavioural change (L3) and HOD ratings business impact (L4), and it revealed that correlation coefficient r = 0.087 (p = 0.052, n = 504), which is not significant at 0.05 range. That confirmed that the 'Broken Link' is present between skill level and impact level.

A participant scoring 90% on behavioural improvement was no more likely to receive a high L4 business results rating than a participant with marginal behavioural improvement. This disconnect shows multiple systemic factors: first, the KPI frameworks used by the Maharatna HODs for evaluating the business impact appear insufficiently granular or connected to the individual training outcomes, potentially measuring the broad departmental performance rather than the individual skill application; second, bureaucratic rigidity and the hierarchical inertia in PSU structures may suppress the direct expression of the newly learnt skills in the performance-relevant behaviours; third, the transfer climate—which defined by the degree to which the

organizational environment supports post-training application—appears weak across multiple units, thus preventing the L3 to L4 conversion that the Kirkpatrick model establishes.

The company-wise L4 Impact Index showed substantial variance across the seven Maharatna organizations, ranging from 54.50% (M6) to 70.50% (M4). ANOVA confirmed that these inter company differences were statistically significant ($p < 0.05$), also suggested that organizational context is the critical determinant of whether trained behaviour translate into perceived business results. The benchmark performers (M1-ONGC, OTE = 77.15%; M4-NTPC, OTE = 75.17%) showed stronger transfer climates, clearer KPI linkages, and superior infrastructural conditions, while the lowest performers (M6-BPCL, OTE = 44.87%; M2-IOCL, OTE = 53.75%) exhibited systemic failures across multiple enabling variables within the study.

Table 5: Correlation Matrix – Kirkpatrick Levels

Variable	Level 1 (Reaction)	Level 2 (Learning)	Level 3 (Behaviour)	Level 4 (Results)
Level 1 (Reaction)	1.00			
Level 2 (Learning)	0.42***	1.00		
Level 3 (Behaviour)	0.31***	0.48***	1.00	
Level 4 (Results)	0.11*	0.14**	**0.087 (ns)**	1.00

***p < 0.001; **p < 0.01; *p < 0.05; ns = non-significant (p = 0.052); n = 504 for Level 4 correlations* Table 5: Pearson correlation matrix among the four Kirkpatrick evaluation levels

4.5 Infrastructure as the Dominant Accountability Variable

Exploratory Factor Analysis (EFA) was done on Level 1 data, to identify the key factors driving OTE. Four key factors were found, resulting 72.7% total variance, these factors are Trainer Quality (28.5%), Content Relevance (18.2%), Organization Infrastructure (14.6%), and Training Schedule (11.4%). Then the Rotated Component Matrix was performed, to identify that these factors are independent in nature. To give an example, on Q18 (L1), the loading =0.912 on Component 3 (Infrastructure) and the loading =0.054 on Component 1 (Trainer Quality), shows almost negligible cross loading.

Table 6: Exploratory Factor Analysis – Diagnostic Variables

Factor	Items	Eigenvalue	Variance Explained (%)	Cumulative (%)	Cronbach's α
Trainer Quality	Q1-Q6	4.28	28.5	28.5	0.89
Content Relevance	Q7-Q12	2.73	18.2	46.7	0.84
Infrastructure/Facilities	Q13-Q17	2.19	14.6	61.3	0.91
Scheduling/Timing	Q18-Q20	1.71	11.4	72.7	0.76

Extraction Method: Principal Component Analysis; Rotation Method: Varimax with Kaiser Normalization

Table 6: Factor structure of training diagnostic variables (Level 1 assessment)

ANOVA test shown that Infrastructure is the most impactful factor. Training room with comfortable sitting arrangement, air conditions, shown better OTE than a non-AC training room. The ANOVA also found that the Trainer Quality is not so significant for the OTE. The Intra Company analysis revealed that the OTE differs on company specific parameters. The M2 due to infrastructure lacking scored less OTE, M6 due to system fault scored lowest learning index and M5 due to its irrelevant content scored poorly. This also shows that an Intra

Company analysis is a must to understand the underlying patterns.

4.6. Towards an Accountability-Oriented Training Governance Framework

The aggregated findings of this study describe an organisation where training accountability is systematically undermined by three interconnected failures: measurement failure (the absence of Level 3 and Level 4 evaluation in routine practice), design failure (the use of generic, level-undifferentiated training content that fails senior participants and decontextualised content that fails technical participants), and infrastructure failure (the deployment of training programmes in physical environments that impair rather than support learning). Addressing these failures requires a governance framework that repositions accountability at the outcome level rather than the input level.

The OTE findings provide a quantitative baseline for such a framework. At the company level, M1-ONGC (OTE = 77.15%) and M4-NTPC (OTE = 75.17%) should be formally designated as Centres of Excellence, with their training management practices codified as standard operating procedures for replication across weaker performers. The gap between the highest-performing (M1: 77.15%) and lowest-performing (M6: 44.87%) organisations represents a 32-percentage-point deficit that cannot be bridged through programme-level interventions alone—it requires systemic organisational investment in infrastructure, KPI redesign, and transfer climate development.

Specific accountability mechanisms should include: mandatory biannual Infrastructure Audits with minimum quality standards for training facilities (HVAC, acoustics, seating, lighting), tied to training budget approval; a redesigned Level 4 evaluation architecture incorporating training-sensitive KPIs developed jointly by HR and departmental HODs before programme commencement, not retrospectively; a Post-Training Transfer Protocol requiring supervisors to provide structured skill application opportunities within 30 days of training completion, with documented follow-up at 60 and 90 days; a centralised OTE Dashboard submitted quarterly to CMD and HR Director, displaying L1-L4 scores, inter-company rankings, and trend lines to enable evidence-based governance; and a mandatory multi-source Level 3 assessment for all Senior Management training participants, replacing self-completion feedback with structured supervisor observation protocols to address the documented Overconfidence Gap.

The senior management training architecture requires particular redesign priority. The data demonstrate that Senior Managers are the least improved in absolute behavioural terms, the most overconfident in their self-assessment, and the most subject to the Ceiling Effect in knowledge acquisition. Their training programmes should incorporate advanced simulation exercises, strategic case analysis anchored in Maharatna-specific operational scenarios, residential formats that remove participants from daily operational interruptions, and mandatory 360-degree developmental feedback with trained debrief facilitation. The residential format is supported by the infrastructure findings: consistent, high-quality physical environments produce measurably superior outcomes, and the controlled environment of a residential training centre eliminates the facility quality variance that the study identifies as a dominant driver of OTE heterogeneity.

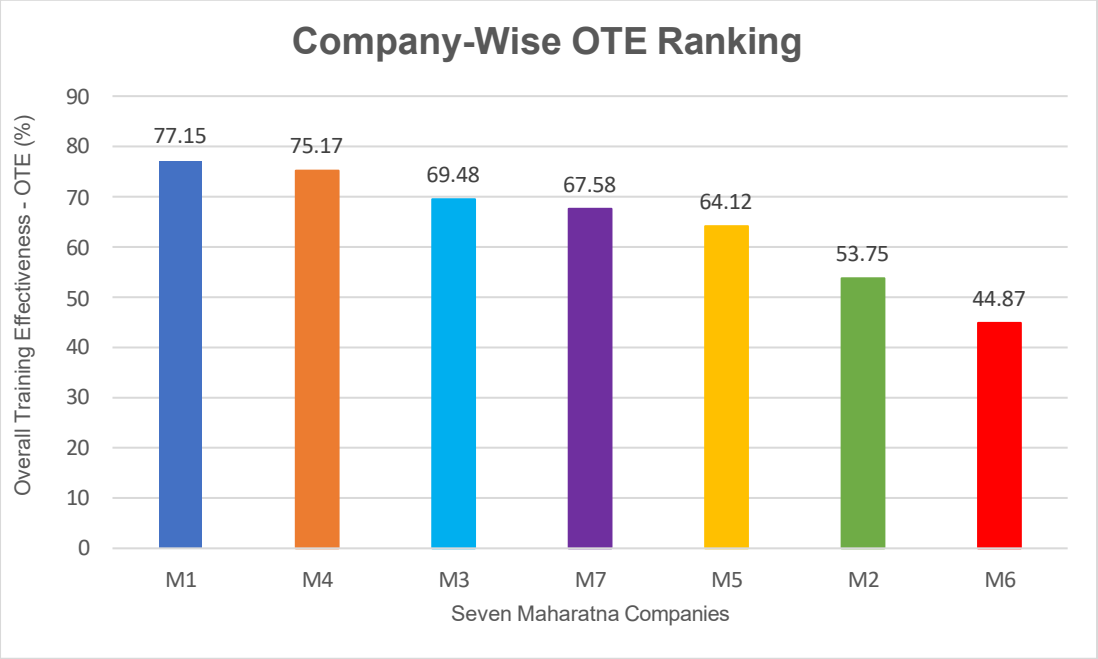
Table 7: Company-Wise Overall Training Effectiveness (OTE)

Company Code	L1 (%)	L2 (%)	L3 (%)	L4 (%)	OTE (%)	Rank
M1	82.5	78.2	74.8	70.2	77.15	1
M4	80.1	76.5	73.2	71.5	75.17	2
M3	75.8	70.2	68.5	65.8	69.48	3
M7	74.2	68.9	67.1	63.4	67.58	4
M5	71.5	65.4	62.8	60.5	64.12	5

M2	62.3	58.2	51.5	48.2	53.75	6
M6	55.2	48.5	45.0	38.5	44.87	7

OTE Formula: (0.10×L1) + (0.25×L2) + (0.30×L3) + (0.35×L4)

Table 7: Comparative Overall Training Effectiveness (OTE) across seven Maharatna CPSEs



5. Conclusion

This study challenges how training accountability is measured in Indian Maharatna enterprises, where training investment is assessed primarily by training volume and expenditure rather than by training effectiveness. The empirical evidence here is, that the training volume and training effectiveness are two different things and they are not significantly correlated. If one is measured and the other one is assumed then there will be gross discrepancy.

The study revealed classroom effectiveness is important for learning gains (Cohen's d = 3.37) and participants satisfaction. The behavioural transfer in Level 3 is observed and it differs in managerial levels, such as Junior Managers performed better during application than their Senior colleagues. The learned skill is failed to create a major impact in business result, a 'Broken Link' is found (Baldwin & Ford, 1988). In Indian PSU environment, the KPI measures are not aligned with the training goals. The infrastructure remains the important factor in determining the OTE. The trainer quality is not so significant for the programs selected, as per the factor loading (Sweller, 1988).

The theoretical contribution of this study belongs to its empirical validation of the training accountability gap in a large-scale Indian PSU context, confirming the importance of the transfer of training (Baldwin & Ford, 1988) and the Kirkpatrick 4 level evaluation is important for the corporate governance. The practical contribution is also found, where the study suggests the provision of organisation-specific, statistically grounded recommendations for closing the gap between training delivery and training impact.

Future research should pursue, a multi-year longitudinal study to track the L3-L4 gap over extended timeframe. The financial ROI modelling should link the training evaluation data to audited balance sheet performance. AI-ML-driven adaptive training personalisation systems should address the gap identified in the learning. The

experimental designs should be considered in the further study, that can test the specific transfer climate interventions against the L3-L4 gap. Finally, a cross-sector comparisons should be done between PSU and private sector training governance that may reveal whether the accountability gap is specific to bureaucratic organisational contexts or endemic to Indian corporate training practice more broadly.

This study is one of the first detailed work on Kirkpatrick Framework in Indian Maharatna Central Public Sector Enterprises, involving 2520 executives or samples. The major contribution in the management literature, is the L3-L4 'Broken Link' in the large-scale PSU. Then for Senior Managers it revealed the 'Senior Learning Slump' and the 'Overconfidence Gap', which can be fixed in program design level. Training Infrastructure as a factor of importance than the trainer quality, can help the PSU to allocate budget strategically to improve the infrastructure. And finally, the importance of Training Accountability is established.

References

[1] Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. Personnel Psychology, 41(1), 63–105. https://doi.org/10.1111/j.1744-6570.1988.tb00632.x

[2] Brinkerhoff, R. O. (2006). Telling training's story: Evaluation made simple, credible, and effective. Berrett-Koehler Publishers.

[3] George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference (4th ed.). Allyn & Bacon.

[4] Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics, 66(1), 64–74. https://doi.org/10.1119/1.18809

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

[6] Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). Kirkpatrick's four levels of training evaluation. Association for Talent Development.

[7] Phillips, J. J. (1997). Return on investment in training and performance improvement programs. Gulf Professional Publishing.

[8] Rao, T. V., Varghese, S., & Rao, R. (2014). Training for development in Indian organizations: A comparative study. Indian Journal of Industrial Relations, 49(3), 418–434.

[9] Sahni, J. (2020). Impact of training and development on organizational performance: Empirical evidence from the banking sector. International Journal of Management Research and Emerging Sciences, 10(1), 1–14.

[10] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

[11] Tamkin, P., Yarnall, J., & Kerrin, M. (2002). Kirkpatrick and beyond: A review of models of training evaluation. The Institute for Employment Studies.

[12] Twitchell, S., Holton, E. F., & Trott, J. W. (2000). Technical training evaluation practices in the United States. Performance Improvement Quarterly, 13(3), 84–109. https://doi.org/10.1111/j.1937-8327.2000.tb00181.x

[13] Yadav, R., & Dabhade, N. (2013). Performance management system in Bharat Heavy Electrical Limited. International Letters of Social and Humanistic Sciences, 4, 49–69.