

Article

Evaluating Reflective Thinking-Based Teacher Training Using Kirkpatrick's Model: A Multi-Source and Cross-Level Analysis

Rahma Tri Wulandari * , Pedro Jurado de los Santos  and Antoni Navío Gámez 

Department of Applied Pedagogy, Autonomous University of Barcelona, Cerdanyola del Vallès, 08193 Bellaterra, Spain; pedro.jurado@uab.cat (P.J.d.l.S.); antoni.navio@uab.cat (A.N.G.)

* Correspondence: rahmatri.wulandari@autonoma.cat

Abstract

Evaluations of teacher professional development programmes are often constrained by reliance on single-source self-report data and partial evaluation designs, limiting the validity of conclusions on training effectiveness. This study evaluates a Reflective Thinking-Based Training programme in Indonesia using Kirkpatrick's model, strengthened through multi-source triangulation involving teachers, school principals, peer teachers, and students. A convergent mixed-methods design was employed, integrating quantitative and qualitative data across all four levels. The sample comprised 30 lower secondary school teachers. External evaluations were provided by three principals and four peer teachers at Level 3 (Behaviour), and by 266 students at Level 4 (Results). Findings indicate positive outcomes at early levels, with lower scores observed at higher levels. Mean scores ranged from 4.08 (student-reported outcomes) to 4.77 (teacher reactions). The gradual decline across levels suggests that transfer of training into classroom practice and student learning experiences is not automatic. Cross-level analysis indicates that relationships among evaluation levels are conditional rather than deterministic. These findings provide preliminary support for the feasibility of combining a comprehensive four-level evaluation with multi-source triangulation in resource-constrained contexts. For policymakers and practitioners, results underscore the importance of institutional support mechanisms for sustaining the transfer of professional learning into practice.

Keywords: teacher professional development; Kirkpatrick model; reflective thinking; training evaluation; multi-source triangulation; cross-level analysis

1. Introduction

Evaluating the effectiveness of teacher professional development (TPD) programmes is essential for ensuring that such initiatives not only enhance teachers' knowledge and skills, but also promote sustained changes in instructional practice and improve students' learning experiences (Darling-Hammond, 2017; OECD, 2019a). Accordingly, TPD evaluation cannot be limited to short-term outcomes; it must also capture the process of transfer from training into professional practice and its implications for classroom learning.

Empirical evidence, however, indicates that TPD evaluation continues to face substantial limitations. Among more than 1300 studies reviewed, only nine met rigorous standards of evidence (Yoon et al., 2007), while an analysis of 139 TPD programmes across 14 countries found that most had not undergone rigorous evaluation (Popova et al., 2022). This gap is evident across contexts. In low- and middle-income countries, the relationship between programme implementation and effectiveness remains weak (Mitchell et al., 2024;



Academic Editor: Federico Corni

Received: 15 April 2026

Revised: 18 May 2026

Accepted: 22 May 2026

Published: 27 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Popova et al., 2022), whereas in OECD countries, participation in TPD has not consistently translated into changes in teaching practice (OECD, 2019b). These findings suggest that the central challenge in TPD lies not only in programme design but also in the adequacy of evaluation approaches.

One of the most widely used evaluation frameworks is Kirkpatrick's model, which organises evaluation into four levels: Reaction, Learning, Behaviour, and Results (Alsalamah & Callinan, 2021; Nawaz et al., 2022). Despite offering a comprehensive structure, its application in TPD contexts presents three major limitations.

First, Kirkpatrick-based evaluations tend to be partial, with a predominance of measurement at Level 1 and Level 2, while Levels 3 and 4 reflecting training transfer and real-world impact are rarely assessed systematically (Bates, 2004; Kennedy et al., 2013; Shewchuk et al., 2023). As a result, evaluations often prioritise indicators that are easier to measure rather than those most relevant for explaining changes in instructional practice.

Second, most evaluations rely heavily on self-report data, which are vulnerable to perceptual bias and do not always represent independently verifiable behavioural change (Shewchuk et al., 2023; Smither et al., 2005). Although the literature has emphasised the importance of multi-source triangulation to enhance validity, its application in TPD evaluation remains limited (Ambu-Saidi et al., 2024; Oanh et al., 2024; Shewchuk et al., 2023).

Third, the model is often interpreted in a linear manner, implying that relationships across levels are automatic and deterministic. However, both classical and contemporary studies indicate that these relationships are context-dependent and not always empirically strong (Alliger & Janak, 1989; Holton, 1996; Nawaz et al., 2022). This limitation is rarely tested explicitly, leading many studies to implicitly reproduce assumptions of linearity without empirical verification.

Collectively, these limitations suggest that the key challenge in TPD evaluation lies not only in the use of Kirkpatrick's model itself, but in how it is operationalised methodologically. In particular, studies that systematically integrate all four evaluation levels with multi-source triangulation involving teachers, principals, peer teachers, and students remain scarce. Consequently, existing evaluations often fail to capture behavioural change and learning outcomes comprehensively.

Despite these limitations, Kirkpatrick's model remains highly relevant when supported by more comprehensive and evidence-based evaluation designs (Bates, 2004; Nawaz et al., 2022). In this regard, a multi-source triangulation approach is essential, not only to reduce self-report bias but also to enable cross-validation across perspectives and strengthen the credibility of findings (Stufflebeam & Coryn, 2014). Such an approach also enables empirical re-examination of assumptions about relationships across evaluation levels, rather than treating the model as a purely normative structure.

In response to these gaps, this study evaluates a Reflective Thinking-Based Training (RTBT) programme in Indonesia, a country with approximately 4.7 million teachers and a rapidly evolving TPD system that is increasingly prioritised in national education policy (OECD, 2019b). The study applies Kirkpatrick's model across all four levels with multi-source triangulation involving teachers, school principals, peer teachers, and students, and conducts cross-level analysis to examine relationships across evaluation levels.

This study addresses two main research questions: (1) to what extent is the RTBT programme effective when evaluated using Kirkpatrick's model with a multi-source triangulation approach, and (2) what patterns of relationships emerge across the evaluation levels? The study makes three contributions to the TPD literature. First, it provides a comprehensive evaluation encompassing all four levels of Kirkpatrick's model. Second, it demonstrates the methodological value of multi-actor triangulation by integrating perspectives from teachers, principals, peers, and students to capture training effects across

professional and classroom contexts. Third, it offers interpretive findings suggesting that relationships across levels are conditional rather than deterministic. These contributions are relevant not only to the Indonesian context but also to other education systems facing similar challenges in conducting accountable and evidence-based TPD evaluation.

2. Literature Review

2.1. *Reflective Thinking as a Foundation for Teacher Professional Development*

Within the TPD literature, reflection is widely regarded as a central mechanism for meaningful learning and sustained professional change (Kolb, 1984; Schön, 1983). Reflective thinking enables teachers to evaluate their experiences, understand the basis of their professional decisions, and design contextually relevant improvements (Schön, 1983). Through this process, teaching experience becomes a primary source of ongoing professional learning (Kolb, 1984).

Several models of reflective thinking have been developed to support teacher learning. Schön's (1983) concept of the reflective practitioner positions reflection as an integral component of professional practice, encompassing both reflection-in-action and reflection-on-action. This perspective is extended in Kolb's (1984) experiential learning framework, which conceptualises reflection as a key phase in a cyclical process beginning with concrete experience and leading to conceptualisation and active experimentation. Gibbs (1988) further develops a more structured and operational model of reflection, offering systematic stages that facilitate its application in professional training contexts. While these frameworks provide important insights into reflective learning, their integration into systematic, contextually grounded, and empirically evaluable training designs remains limited in practice.

Clarke and Hollingsworth's (2002) model of professional growth provides an important framework for understanding how reflection contributes to changes in practice through interactions among personal, practice, external, and outcome domains. Similarly, Opfer and Pedder (2011) emphasise that teacher professional change is systemic, shaped by the interaction of individual, contextual, and learning-related factors. Thus, the effectiveness of reflective training depends not only on programme design but also on supportive environments that facilitate transfer into practice. Darling-Hammond et al. (2017) further confirm that effective TPD is consistently associated with opportunities for professional collaboration and strong institutional support. Reflective training therefore cannot be separated from the broader systemic context in which it operates.

In this context, the RTBT programme was designed to integrate multiple reflection frameworks into a structured training model and to connect them explicitly with professional practice through follow-up mechanisms. This design is intended to support evaluation not only at the level of learning but also at the level of behavioural change and its impact on classroom practice.

The programme was developed within the Indonesian national framework and implemented through government-accredited teacher training institutions in East Java province. It consists of a 32-h training structure covering four main modules: (1) foundational concepts of reflective thinking; (2) international reflection models (Kolb, Schön, Gibbs); (3) application of reflection in instructional planning and teaching; and (4) collaborative reflection. Follow-up activities are embedded in participants' professional practice over a three-month period after the training. This design explicitly links training learning with real-world practice, enabling cross-level evaluation.

2.2. Training Evaluation: Conceptual Foundations and Purposes

A growing body of research indicates that TPD programmes do not consistently lead to sustained changes in instructional practice (Borko, 2004; Darling-Hammond et al., 2017; Timperley et al., 2007). The impact of training on classroom practice is often not immediately observable, as professional change occurs gradually and is shaped by contextual factors (Avalos, 2011). This underscores the need for evaluation approaches capable of capturing gradual, context-dependent professional change.

Desimone's (2009) framework identifies five core features of effective TPD: content focus, active learning, coherence, duration, and collective participation as mediators between training design and student learning outcomes. However, this relationship is indirect and mediated through changes in teachers' knowledge, beliefs, and instructional practices. This perspective is reinforced by Opfer and Pedder (2011), who argue that changes in practice are not linear but emerge from complex interactions among individual characteristics, school contexts, and programme features. Consequently, TPD evaluation requires approaches capable of capturing these dynamic and multidimensional processes.

2.3. Kirkpatrick's Model: Global Application and Persistent Limitations

Kirkpatrick's model is one of the most widely used frameworks for training evaluation across educational and professional contexts, including teacher and principal development (Alsalamah & Callinan, 2021; Oanh et al., 2024). In practice, however, evaluations often do not encompass all four levels comprehensively. Shewchuk et al. (2023) report that only a small proportion of studies (approximately 8.5%) assess outcomes at Level 4, indicating that evaluation practices remain concentrated at the lower levels and do not consistently capture broader impacts.

These limitations are partly driven by practical challenges associated with measuring Level 3 and Level 4 outcomes, including resource constraints, limited institutional support, and methodological complexity (Cahapay, 2021; Kennedy et al., 2013). Yet, changes in instructional practice and their impact on student learning are widely recognised as the most meaningful indicators of training effectiveness (Guskey, 2002b).

Theoretical critiques of Kirkpatrick's model further highlight two key limitations. Alliger and Janak (1989) demonstrate that the assumed hierarchical, causal, and correlational relationships among levels are not consistently supported by empirical evidence. Holton (1996) argues that the model is better understood as a taxonomy of training outcomes rather than a comprehensive evaluation model, as it does not account for critical mediating variables such as motivation, transfer conditions, and workplace environment. In addition, Bates (2004) notes that the model is frequently applied in a mechanistic manner, without sufficient consideration of contextual factors. Together, these critiques suggest that the use of Kirkpatrick's model requires a more analytical and context-sensitive approach.

2.4. Multi-Source Evaluation as a Methodological Response

Single-source evaluation, particularly self-report, is vulnerable to perceptual bias and social desirability effects (Fang, 1996; Smither et al., 2005). In contrast, a multi-source approach enables the integration of complementary perspectives, thereby enhancing the accuracy and credibility of evaluation (Goe et al., 2008).

This approach allows for cross-validation between participants' self-reports and external observations, which is critical given that teachers' beliefs and reported practices do not always reflect actual classroom behaviour (Fang, 1996). Conway and Huffcutt (1997) found that interrater correlations are often low, suggesting that different evaluators capture distinct dimensions of performance. Scullen et al. (2000) argue that a substantial portion

of rating variance originates from rater-specific perspectives, commonly referred to as idiosyncratic rater effects.

A multi-source approach is also aligned with Guskey’s (2002a) evaluation framework, which emphasises that professional development evaluation must encompass multiple levels and cannot be adequately represented by a single instrument or data source. Furthermore, Darling-Hammond et al. (2017) highlight that effective TPD programmes consistently incorporate feedback from multiple stakeholders as part of the professional learning process.

Thus, multi-source triangulation not only strengthens methodological validity but also responds to critiques of Kirkpatrick’s model regarding its limited consideration of contextual variables (Bates, 2004; Holton, 1996). By involving external actors such as principals, peers, and students, evaluation becomes more context-sensitive and better able to capture behavioural changes as they occur in authentic practice.

3. Methods

3.1. Evaluation Design

This study employed a convergent mixed-methods evaluation design (Creswell & Plano Clark, 2017), grounded in Kirkpatrick’s four-level model. Quantitative and qualitative data were collected and integrated to provide a comprehensive assessment of training effectiveness across all four evaluation levels.

Quantitative data were primarily used at Level 1 (Reaction) and Level 2 (Learning), while Levels 3 (Behaviour) and 4 (Results) combined quantitative and qualitative data from multiple sources. This design directly responds to methodological gaps in the TPD literature, particularly the dominance of single-source evaluation and the partial application of early-level measures (Oanh et al., 2024; Shewchuk et al., 2023). In addition, the design enables cross-level analysis to examine relationships across evaluation levels. Thus, the study not only evaluates programme effectiveness but also examines the relational logic underlying Kirkpatrick’s model.

3.2. Participants

The participants comprised three main groups, as summarised in Table 1. These groups were purposively selected to align with the evaluation requirements at each level, enabling triangulation across perspectives and strengthening the validity of findings.

Table 1. Participant groups and roles in the study.

Participant	N	Role in Evaluation			
		Level 1	Level 2	Level 3	Level 4
Teachers	30	√	√	√	√
School principals	3			√	
Peer teachers	4			√	
Students	266				√

Teachers functioned as the core respondents, contributing data across all evaluation levels to capture both immediate and sustained outcomes of the training. These teachers had between 3 and 20 years of teaching experience, reflecting the range of experience commonly found among in-service teachers in the Indonesian TPD context. School principals and peer teachers were included as external evaluators at the behavioural level, offering supervisory and collegial perspectives to assess changes in instructional practice. Meanwhile, students provided evidence at the results level through their perceptions of classroom learning

experiences, reflecting how the effects of the training were manifested in actual teaching and learning processes. These students were drawn from classes taught by the participating teachers and represented Grades 7 to 9 (aged 12–15 years).

3.3. Operationalisation of Kirkpatrick's Model

The operationalisation of Kirkpatrick's four-level model is presented in Table 2, which outlines the alignment between evaluation levels, evaluation focus, data sources, measurement instruments, and data collection timing.

Table 2. Operationalisation of Kirkpatrick's Model in Teacher Training Evaluation.

Evaluation Level	Evaluation Focus	Data Sources	Measurement Instruments	Data Collection Timing
Level 1–Reaction	Participants' responses to training content, methods, facilitators, and relevance	Teachers	Closed-ended questionnaire	Immediately after training
Level 2–Learning	Learning outcomes, particularly understanding and reflective thinking ability	Teachers	Closed-ended questionnaire	Immediately after training
Level 3–Behaviour	Application of training outcomes in teaching practice and professional behaviour	Teachers; principals; peer teachers	<ul style="list-style-type: none"> • Closed-ended questionnaire (teachers) • Closed- and open-ended questionnaires (external evaluators) 	<ul style="list-style-type: none"> • Teachers: immediately after training • External evaluators: ~3 months post-training
Level 4–Results	Impact on instructional practice and students' learning experiences	Teachers; students	<ul style="list-style-type: none"> • Closed-ended questionnaire (teachers) • Closed- and open-ended questionnaires (students) 	<ul style="list-style-type: none"> • Teachers: immediately after training • Students: ~3 months post-training

It should be noted that teacher-reported data at Levels 3 and 4, collected immediately after training, reflect self-perceived readiness and anticipated application rather than evidence of actual behavioural transfer in authentic classroom settings. Empirical validation at both levels was based primarily on data obtained from external evaluators (principals, peer teachers, and students), approximately three months after programme implementation, allowing sufficient time for observable instructional changes to emerge.

The quantitative instruments were adapted from the training evaluation framework developed by the Centre for Teachers and Education Personnel of East Java Province, the implementing authority of the RTBT programme, and were further informed by empirical studies on teacher professional development and reflective practice. Across all instruments, the Kirkpatrick model was operationalised as an evaluation framework rather than a fixed measurement instrument, with indicators adapted to stakeholder roles and evaluative purposes (Alsalamah & Callinan, 2021; Matolić et al., 2023; Yu, 2025).

Instrument development followed a unified measurement structure across the four evaluation levels. For teacher self-assessment, each level was operationalised as a single

latent construct measured through five indicators: reaction (Level 1), learning outcomes (Level 2), self-reported behavioural change (Level 3), and instructional impact (Level 4), yielding a 20-item integrated instrument. External evaluator instruments were developed for principals and peer teachers by adapting the teacher instrument into observational formats capturing the same underlying behavioural construct from supervisory and collegial perspectives. The principal instrument comprised 10 items, while the peer teacher instrument comprised 15 items. At Level 4, student perceptions of classroom instruction and perceived changes in teacher practice were measured using a 15-item instrument. All instruments employed a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

Content validity was established through expert judgement involving six specialists in teacher professional development and training evaluation. Experts independently evaluated item relevance, clarity, and representativeness, and feedback from expert panel review was used to refine item wording and structure prior to field implementation.

3.4. Data Analysis

Given the evaluative purpose of this study, quantitative data were analysed descriptively using means and standard deviations across all respondent groups. Scores from the five-point Likert scale were interpreted using effectiveness thresholds adapted from teacher training evaluation guidelines developed by a government-accredited teacher training institution in East Java province, Indonesia: ≤ 3.54 (ineffective), 3.55–4.04 (moderately effective), 4.05–4.54 (effective), and 4.55–5.00 (very effective).

Qualitative data from open-ended responses were analysed using thematic analysis following [Braun and Clarke \(2006\)](#), including data familiarisation, systematic coding, theme development, and narrative construction.

In addition, a cross-level pattern analysis was conducted to examine whether outcomes at earlier levels were consistently associated with results at subsequent levels. This analysis was interpretive and comparative, focusing on directional trends and convergence across quantitative and qualitative findings, rather than inferential tests of association. This approach is consistent with convergent mixed-methods evaluation designs in which data integration across levels serves to build a comprehensive interpretive account rather than establish statistical causation ([Creswell & Plano Clark, 2017](#)). The analysis directly addresses ongoing debates regarding assumptions of linearity within Kirkpatrick's model ([Holton, 1996](#); [Nawaz et al., 2022](#)), contributing a pattern-based examination of cross-level relationships within this specific programme context.

4. Results

4.1. Level 1—Reaction: Participants' Responses to the Training

The evaluation at Level 1 aimed to capture participants' responses to the training across five dimensions: content relevance, training methods, facilitator competence, training duration and activities, and overall satisfaction.

The results presented in [Figure 1](#) indicate that all dimensions were rated within the "very effective" category. The highest mean score was observed for the appropriateness of training duration and activities ($M = 4.80$), while the lowest score was recorded for overall satisfaction and expectations ($M = 4.73$).

Other dimensions, including content relevance, instructional methods, and facilitator competence, showed consistently high scores ($M = 4.77$), indicating a uniformly positive response across all aspects of the training.

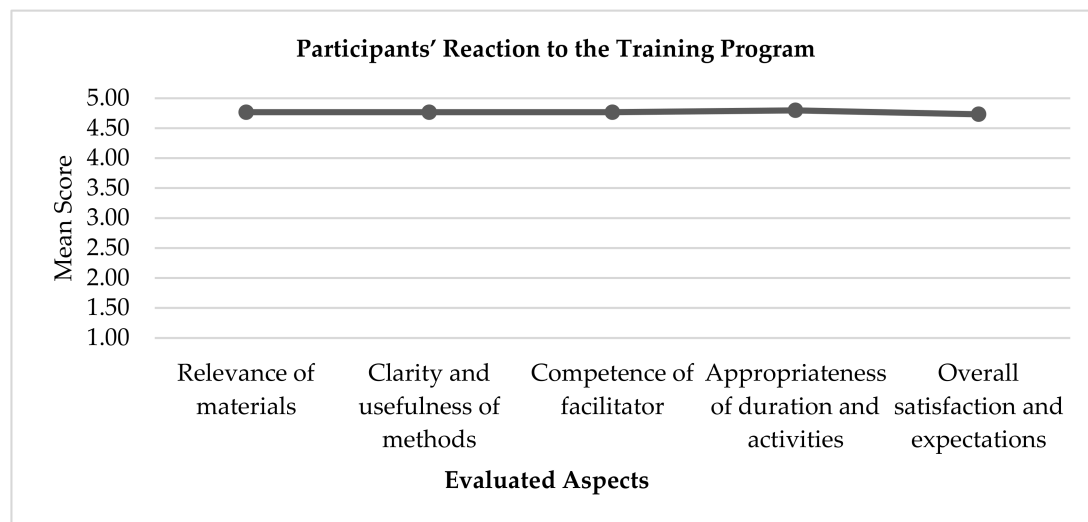


Figure 1. Level 1-Reaction.

4.2. Level 2—Learning: Participants' Learning Outcomes

The Level 2 evaluation assessed participants' learning outcomes, including conceptual understanding, reflective skills, ability to identify reflective methods, confidence in conducting reflective activities, and theory–practice integration.

As presented in Figure 2, the results show that the overall mean of learning outcomes fell within the “effective” category ($M = 4.51$). The highest score was found for reflective skills ($M = 4.67$), followed by conceptual understanding of reflective thinking ($M = 4.53$). Meanwhile, the ability to identify reflective methods and connect theory with practice yielded equal scores ($M = 4.47$).

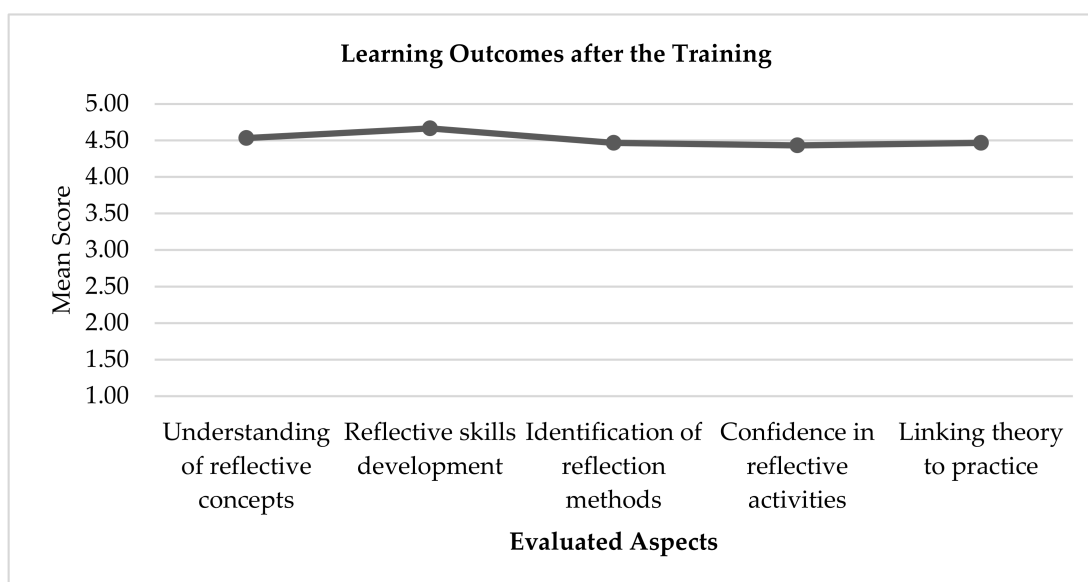


Figure 2. Level 2-Learning.

The lowest mean score was found for confidence in conducting reflective activities ($M = 4.43$). This result remains within the “effective” category, indicating that participants were developing confidence in reflective practice, though it had not yet reached the level of other learning dimensions.

4.3. Level 3—Behaviour: Changes in Teaching Practice

The evaluation at Level 3 presents changes in teaching practice based on three data sources: teachers' self-assessments, principal evaluations, and peer teacher assessments.

4.3.1. Teachers' Self-Assessment of Behavioural Change

The self-assessment results indicate that all behavioural indicators were rated within the "effective" category, with mean scores ranging from 4.30 to 4.53.

Figure 3 shows that the highest scores were observed for adapting instructional practices and sharing reflective practices ($M = 4.53$), followed by integrating reflection into lesson planning ($M = 4.50$). Lower scores were found for the frequency of reflection ($M = 4.30$) and the direct application of reflective methods ($M = 4.40$).

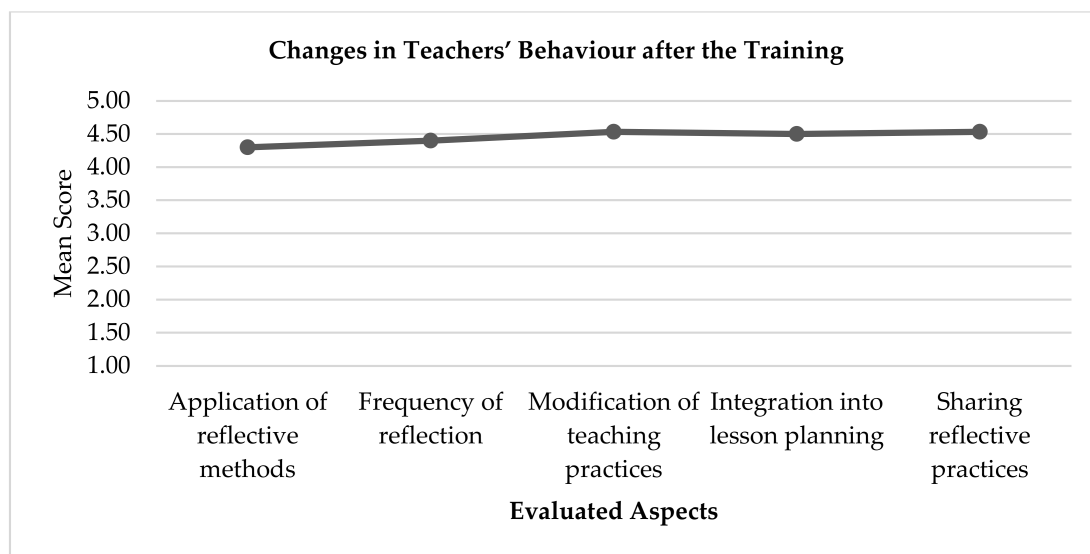


Figure 3. Level 3-Behaviour.

4.3.2. External Validation by School Principals and Peer Teachers

To strengthen validity, behavioural changes were also assessed through external evaluations by school principals and peer teachers, as summarised in Table 3.

Table 3. External Evaluation of Changes in Teachers' Behaviour (Level 3).

Source	N	Mean	SD	Category
Principals	3	4.67	0.306	Very Effective
Peer Teachers	4	4.20	0.493	Effective

Principal evaluations yielded the highest mean score ($M = 4.67$), while peer teachers reported a mean score of 4.20. Both sources indicate that behavioural changes fall within the "effective" to "very effective" range.

Qualitative findings from principals and peer teachers further support the quantitative results by providing contextual evidence of observed changes. Principals reported more varied and well-managed instructional practices, improved teacher–student interaction, and increased openness to professional collaboration. Peer teachers confirmed improvements in pedagogical and professional competence, as well as the use of post-lesson reflection to inform instructional improvement.

The combined findings suggest that reflection is beginning to move beyond an individual activity toward a more systematic approach to instructional improvement. The

alignment between teachers' self-reports and external evaluations suggests that some reported behavioural changes may also have been observable within the school context. Given the small number of external evaluators, however, this alignment should be interpreted as preliminary rather than as systematic confirmation of behavioural change across settings.

Overall, the Level 3 findings suggest that training outcomes have begun to transfer into teaching practice and professional behaviour, although the intensity and consistency of this transfer vary across individuals and school contexts.

4.4. Level 4—Results: Impact on Students' Learning Experiences

The Level 4 evaluation focused on both teachers' and students' perceptions of classroom learning experiences.

4.4.1. Teachers' Perceptions

Teachers' perceptions were used to examine how reflective practices developed during the training influenced classroom dynamics, particularly in terms of instructional quality, classroom interaction, and student engagement.

As depicted in Figure 4, teachers' perceptions of training impact were rated within the "very effective" category, with mean scores ranging from 4.47 to 4.63. The highest scores were observed for instructional quality and student engagement ($M = 4.63$), followed by learning outcomes ($M = 4.60$), management of instructional challenges ($M = 4.50$), and professional collaboration ($M = 4.47$).

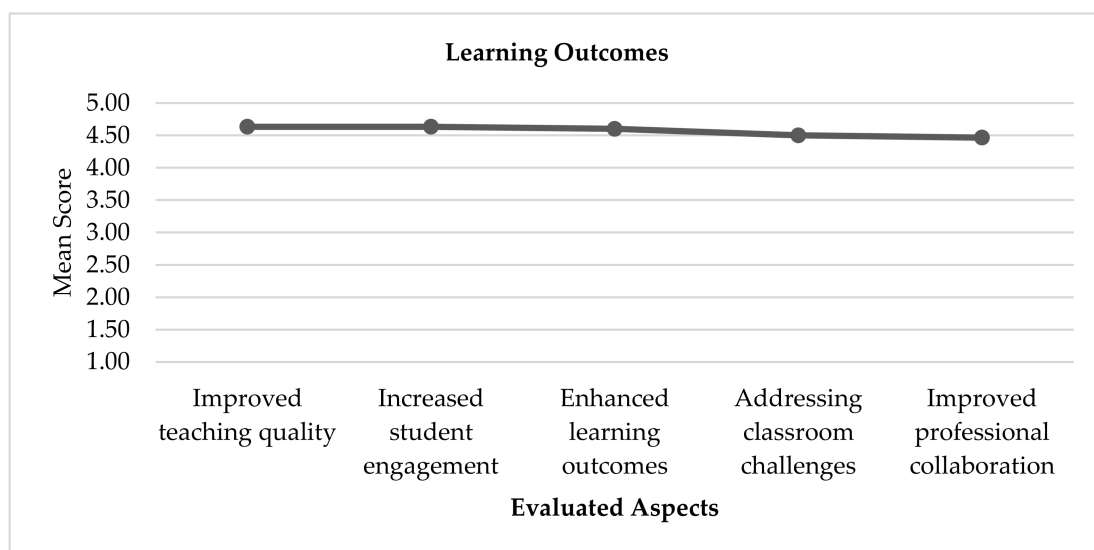


Figure 4. Level 4-Results.

4.4.2. Students' Perceptions of Classroom Instructional Practices

Students' perceptions were used as an external data source to assess the extent to which changes in teaching practice were reflected in classroom learning experiences. Quantitative data were collected through a closed-ended questionnaire consisting of 15 items covering instructional clarity, classroom interaction, student engagement, use of learning media, and variation in teaching strategies and assessment, as presented in Table 4.

Table 4. Students' Evaluation of Teachers' Classroom Instructional Practices (Level 4).

Source	N	Mean	SD	Category
Students	266	4.08	0.563	Effective

As displayed in Figure 5, the composite student mean score of 4.08 (SD = 0.563) falls within the “effective” category, indicating that students generally perceived positive changes in their teachers’ instructional practices. Item-level scores ranged from 3.81 to 4.50, with the highest ratings for opportunities to ask questions and support for learning difficulties, and the lowest for alternative explanations when needed and the application of new teaching methods. The moderate standard deviation reflects variation across classrooms, likely attributable to differences in classroom context, student characteristics, and the consistency of reflective practice implementation.

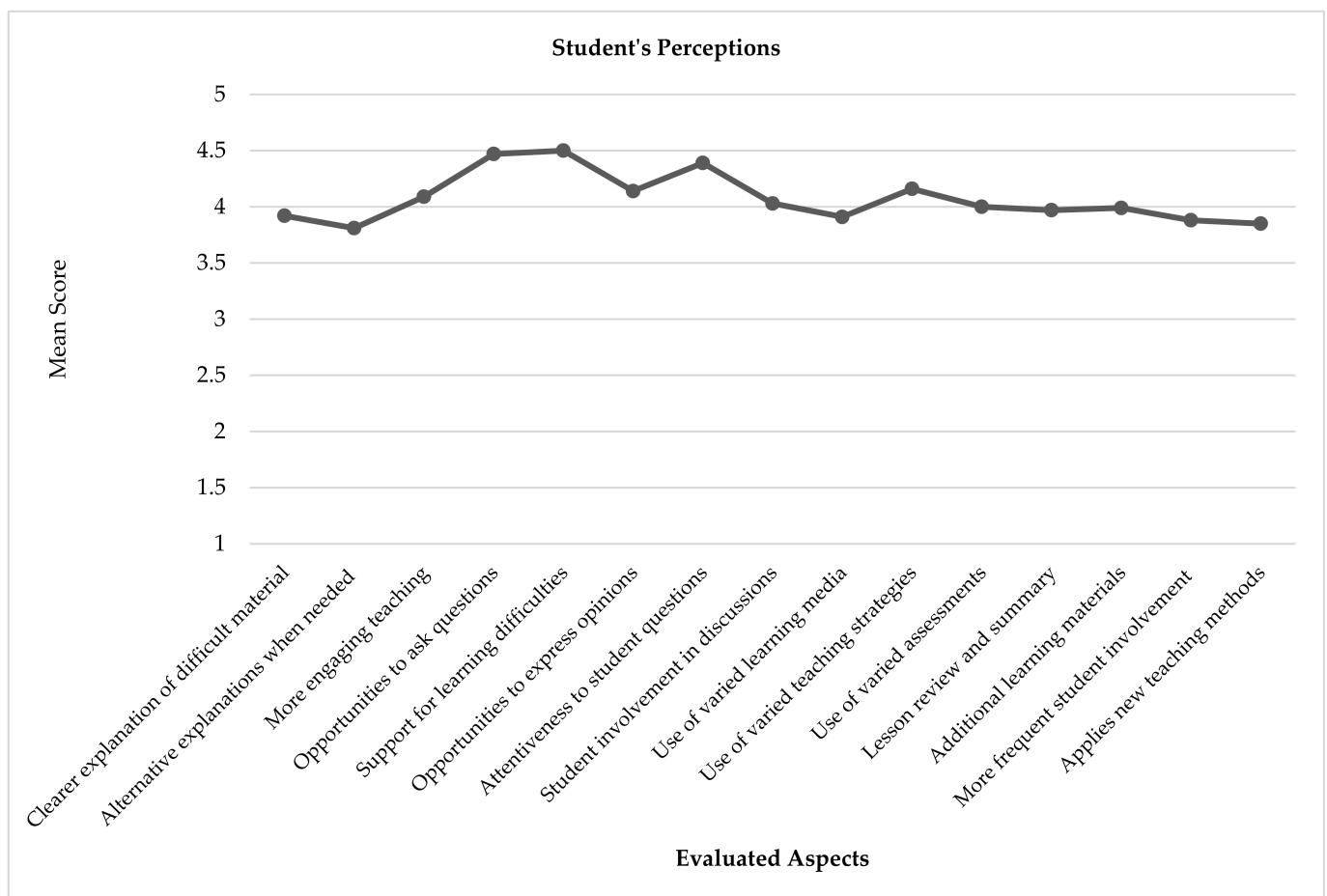


Figure 5. Mean Scores for Students’ Perceptions.

These quantitative findings are supported by qualitative data indicating clearer and more structured explanations, more varied instructional methods, and increased interactive activities (e.g., discussions, group work, learning games). Students also reported higher levels of comfort, motivation, and confidence in participating during lessons, collectively suggesting that learning was perceived as more comprehensible and engaging.

Overall, the alignment between teachers’ self-reports and students’ experiences suggests that the training has generated perceived outcomes at Level 4, particularly in instructional clarity, student participation, and the diversification of teaching practices.

4.5. Cross-Level Analysis: Relational Patterns Across Kirkpatrick Levels

Table 5 presents a summary of mean scores across all evaluation levels, along with the main patterns observed at each level.

Table 5. Cross-Level Summary and Inter-Level Patterns in RTBT Evaluation.

Level	Composite Mean	Category	Key Evaluator (s)	Observed Inter-Level Link
Level 1–Reaction	4.77	Very Effective	Teachers	Consistently high scores across all dimensions (≥ 4.73); uniformly positive responses
Level 2–Learning	4.51	Effective	Teachers	Strong learning outcomes; reflective skills higher than self-efficacy
Level 3–Behaviour	4.45	Effective	Teachers	Behavioural change confirmed across evaluators; consistent patterns despite variation
	4.67	Very Effective	Principals	
	4.20	Effective	Peer Teachers	
Level 4–Results	4.57	Very Effective	Teachers	Student ratings lower than teacher ratings; greater variability observed
	4.08	Effective	Students	

The cross-level summary shows a gradual decline in mean scores from Level 1 ($M = 4.77$) to Level 4 ($M = 4.08$ from the student perspective). Variability across evaluators also increases at higher levels, as reflected in the wider score range at Level 3 (4.20–4.67) and the divergence between teacher and student perspectives at Level 4 (4.08–4.57). Nevertheless, all results remain within the “effective” to “very effective” categories.

Three main relational patterns emerge from the cross-level analysis. First, high scores at Level 1 (≥ 4.73 across all dimensions) correspond with strong learning outcomes at Level 2 ($M = 4.51$), with reflective skills as the highest-performing dimension and self-efficacy as the lowest. Second, conceptual understanding and skills acquired at Level 2 precede and align with behavioural changes confirmed at Level 3 by multiple independent evaluators. Third, behavioural changes identified at Level 3 correspond with perceived classroom impact at Level 4, although with greater variability, particularly in student ratings.

Differences in evaluator scores at Level 3 principals (4.67), teachers (4.45), and peer teachers (4.20) indicate variation across perspectives. Similarly, at Level 4, teacher ratings (4.57) are higher than student ratings (4.08), with greater dispersion among student responses. This suggests variability in the consistency of training transfer across classrooms and contexts.

5. Discussion

5.1. Teachers’ Responses to the Training (Level Reaction)

The findings at the Reaction level indicate that the RTBT programme was received positively by participants across all assessed dimensions, including content relevance, training design, instructional methods, and facilitator competence. This finding aligns with literature emphasising programme clarity, content relevance (Brugha et al., 2024; Wilson et al., 2025), and facilitator support (Darling-Hammond et al., 2017) as key characteristics of effective professional development. Within training evaluation frameworks, positive reactions are generally considered an initial condition that supports engagement in the learning process (Kraiger et al., 1993; Salas et al., 2012).

The marginally lower overall satisfaction score ($M = 4.73$), relative to the ratings across individual dimensions, warrants brief interpretive consideration. Within the Kirkpatrick evaluation framework, participant reactions are understood as a multidimensional construct. Accordingly, an overall satisfaction rating does not necessarily represent a simple arithmetic combination of scores across specific dimensions (Kirkpatrick & Kirkpatrick, 2016). This interpretation is also consistent with research on subjective evaluation processes, which suggests that global judgments are often influenced by participants’ overall impres-

sions at the time of evaluation, including factors beyond the aspects explicitly measured (Schwarz & Strack, 1999). Nevertheless, the observed difference of 0.07 scale points is practically negligible.

However, positive reactions do not directly guarantee learning outcomes; rather, they are generally conceptualised as enabling conditions that facilitate learning at subsequent levels (Guskey, 2002a; Kirkpatrick & Kirkpatrick, 2016).

5.2. Teacher Learning and the Strengthening of Reflective Capacity (Level Learning)

At the Learning level, the findings show that participants achieved strong learning outcomes, particularly in the development of reflective skills, understanding of reflective thinking, and the ability to connect teaching experience with instructional analysis. This indicates that the training did not merely enhance conceptual understanding, but also strengthened participants' readiness to operationalise reflection in professional practice. These outcomes are consistent with professional development frameworks that emphasise the close relationship between teacher learning and classroom practice (Darling-Hammond et al., 2017; Desimone, 2009; Timperley et al., 2007). More specifically, the active learning and coherence features embedded in the RTBT design appear to have enabled reflection to develop as an ongoing professional practice.

Nevertheless, teachers' self-efficacy for independent reflection was lower than the other indicators, suggesting a gap between procedural understanding and the internalisation of reflection as a professional habitus (Loughran, 2002). From a self-efficacy perspective, this represents a predictable developmental stage in which confidence develops gradually through repeated practice (Bandura, 1997). Clarke and Hollingsworth (2002) similarly argue that knowledge gains do not automatically transform practice, but are mediated by repeated cycles of reflection and enactment. Lower confidence at this stage should therefore be interpreted as a normal transitional feature rather than as evidence of programme failure (Avalos, 2011).

To support the development of confidence in reflective practice, structured post-training mechanisms are recommended, such as coaching sessions, peer mentoring (Darling-Hammond et al., 2017) or professional learning communities (Wenger, 1998) that provide opportunities for repeated reflective cycles within authentic classroom contexts. Without such structures, the gap between procedural understanding and the internalisation of reflection as a professional habitus is likely to persist (Bourdieu, 1990). This is because professional confidence develops through sustained, practice-based reinforcement rather than a single training intervention (Bandura, 1997; Loughran, 2002).

5.3. Transfer of Learning and Changes in Teaching Behaviour (Level Behaviour)

The findings at the Behaviour level suggest early signs that reflective practice is beginning to shape teachers' professional behaviour, consistent with Schön's (1983) conception of the reflective practitioner. These patterns were evident in the use of more varied and student-centred teaching strategies, the emergence of post-lesson reflection as a developing habit, and the incorporation of instructional media and project-based approaches. Such developments are consistent with the possibility that learning achieved at the Learning level has begun to transfer into classroom practice, though this inference should be interpreted cautiously given the descriptive design and limited external evaluator sample.

However, external validation indicates that these changes remain largely confined to the individual level and have not yet developed into collective practice at the school level. This finding is consistent with Baldwin and Ford's (1988) transfer of training model, which emphasises that transfer is shaped by the interaction of trainee characteristics, training design, and work environment. In this context, the limited collective transfer suggests

that the work environment has not yet fully supported the systemic internalisation of reflective practice.

From a broader perspective, individual behavioural change is a necessary condition, but insufficient on its own, for the development of a shared professional culture (Borko, 2004; Wenger, 1998). Without supportive collaborative structures, reflective practice is likely to remain individual rather than becoming a collective norm at the school level. This interpretation is consistent with the synthesis by Darling-Hammond et al. (2017), who highlight professional feedback and school-based collaboration as key conditions supporting the sustainability of long-term professional development.

It should be emphasised that these interpretations are derived from a descriptive mixed-methods evaluation without a control group and with a limited number of external evaluators at Level 3. Accordingly, the observed behavioural patterns should be interpreted as programme-related associations within this evaluation context rather than as causal evidence of behavioural transfer.

5.4. Training Impact on Students' Learning Experiences (Level Results)

The findings at the Results level suggest that some changes in teachers' instructional behaviour may have been reflected in students' learning experiences, particularly in improved classroom interaction and teacher responsiveness. While there is general alignment between teacher and student perceptions, the lower and more variable student ratings suggest that the perceived effects of training on classroom practice are uneven and context-dependent.

The use of student perceptions as an indicator is supported by research showing that student ratings constitute a valid source of information for evaluating instructional quality (Goe et al., 2008; Marsh & Roche, 1997). In addition, this finding is consistent with the concept of visible learning, which emphasises the importance of feedback-informed reflection in improving the quality of teaching and learning (Hattie, 2008).

The difference between teacher and student perceptions, combined with the greater variation in student data, indicates that the perceived effects associated with the training were not distributed uniformly. This pattern is consistent with literature showing that perceptions of instructional quality are shaped by differing observational positions and the complexity of classroom contexts (Fraser, 2012; Tomlinson, 1999; Wisniewski et al., 2022). Furthermore, teachers inherently adapt instruction to the specific needs of students and the unique dynamics of each classroom, leading to varied learning experiences even when the foundational training programme is identical (Darling-Hammond et al., 2017). Accordingly, this variability is more appropriately interpreted as evidence of implementation heterogeneity rather than as evidence of programme ineffectiveness.

It should also be noted that the Results level in this study focused on students' learning experiences rather than academic achievement. This approach is consistent with the frameworks proposed by Desimone (2009) and Guskey (2002b), both of which argue that the effects of professional development on student outcomes are indirect and mediated through changes in instructional practice.

5.5. Practical Implications

The findings of this study have practical implications at three interrelated levels: programme design, school-level institutional practice, and national TPD policy. At the programme design level, a comprehensive evaluation covering all four Kirkpatrick levels enables identification of training effects that cannot be captured through partial evaluation. The finding that teacher self-efficacy at the Learning level was lower than the other indicators suggests the need for structured post-training follow-up, such as coaching sessions or professional learning communities, to support the internalisation of reflection as a pro-

fessional habitus (Loughran, 2002; Wenger, 1998). Programme designers should therefore recognise that strong outcomes at the Reaction and Learning levels do not automatically produce systemic behavioural change without a supportive work environment (Baldwin & Ford, 1988; Holton, 1996).

At the institutional level, the finding that behavioural change remained largely individual and had not yet developed into collective practice underlines the strategic role of principals as facilitators of reflective culture rather than merely evaluators. School-based mentoring programmes involving peer teachers as reflective partners may strengthen the transfer of training from the individual domain to the collective domain (Borko, 2004; Darling-Hammond et al., 2017; Wenger, 1998). In addition, the disparity between principal ratings and peer teacher ratings indicates the need for more consistent and standardised observation practices among internal evaluators so that changes in teacher behaviour can be assessed more accurately and comparably (Goe et al., 2008).

At the policy level, this study offers a multi-source evaluation model that could be adapted for large-scale TPD systems in Indonesia without disproportionate resource demands. The finding that student ratings were lower and more variable than teacher perceptions offers preliminary grounds for policymakers to incorporate student perspectives as an indicator in national TPD evaluation, given that student feedback is a valid source of evidence in assessing instructional quality (Goe et al., 2008; OECD, 2019b). This approach responds directly to the growing demand for evidence-based accountability in global education reform (Popova et al., 2022).

5.6. Limitations and Future Research

This study has five limitations that should be considered when interpreting the findings.

First, the number of external evaluators was limited to three principals and four peer teachers, which constrains the generalisability of the Level 3 findings. Future studies should expand the sample through multi-school designs covering more diverse geographical and educational contexts, to enable more robust verification of behavioural change patterns.

Second, the absence of a control group and longitudinal measurement means that the relationships identified in the cross-level analysis should be interpreted as associative rather than causal. Future research adopting quasi-experimental designs with pre-post measurement would provide stronger evidence for empirically testing the relational assumptions of Kirkpatrick's model.

Third, the Level 4 measure was limited to students' perceptions of learning experiences and did not include objective indicators of academic achievement. Future studies should integrate learning outcome data, such as formative assessment scores or examination results, as complementary indicators so that training impact can be evaluated not only in terms of process but also in terms of outcomes (Desimone, 2009; Guskey, 2002b).

Fourth, the three-month post-training observation period was insufficient to capture the long-term consolidation of reflective practice. Longitudinal studies spanning at least one year or longer would provide a more comprehensive understanding of the sustainability of RTBT programme effects, consistent with recommendations for assessing teacher professional development over adequate time horizons (Avalos, 2011; Timperley et al., 2007).

Fifth, teacher-reported data at Levels 3 and 4 were collected immediately after training, prior to any opportunity for classroom transfer. These ratings should therefore be interpreted as indicators of perceived readiness and anticipated application rather than as evidence of actual behavioural change. Future studies should adopt a longitudinal or time-lagged data collection design in which teacher self-assessments at higher levels are collected at the same point as external evaluator data, ensuring that comparisons across sources reflect equivalent observational timeframes.

6. Conclusions

This study comprehensively evaluated an RTBT programme using Kirkpatrick's model across all four levels, supported by a multi-source triangulation approach involving teachers, school principals, peer teachers, and students. Within the bounded scope of this descriptive evaluation, the findings indicate positive outcomes across all four levels, including favourable participant responses, gains in reflective understanding, and reported behavioural changes. The convergence between self-report and external evaluator data at Level 3 offers preliminary support for the feasibility of a multi-source evaluation design in this context, rather than constituting a definitive demonstration of externally confirmed behavioural change across settings.

Cross-level analysis identified generally consistent relational patterns across the four evaluation levels, while also confirming that these relationships are conditional rather than deterministic. High outcomes at the Reaction and Learning levels were associated with behavioural changes at the Behaviour level; however, the intensity and consistency of transfer varied across individuals and school contexts. At the Results level, the alignment between teacher and student perceptions indicates perceived classroom impact, although greater variability in student ratings suggests that the transfer of reflective practice has not yet been uniformly distributed across classrooms.

Theoretically, the findings contribute to ongoing discussions regarding Kirkpatrick's model in three ways. First, the integration of multi-source data across Levels 3 and 4 demonstrates the methodological value of triangulating self-report with external evaluator perspectives, while also underscoring the need for larger and more diverse evaluator samples in future research. Second, the gradual decline in mean scores across levels, alongside increased variability at higher levels, is consistent with the view that cross-level relationships are associative and context-dependent rather than the automatic progressions assumed by linear interpretations of Kirkpatrick's model. Third, taken together, these findings support an analytical rather than mechanistic operationalisation of Kirkpatrick's framework, particularly in resource-constrained contexts where comprehensive evaluation remains rare.

From a practical perspective, the study offers an evaluation framework whose design principles may inform future evaluation efforts in similar TPD contexts, with broader applicability contingent on validation with larger and more diverse samples. These findings suggest that comprehensive and multi-source cross-level evaluation serves not only as a methodological enhancement, but also as a practical means of tracing how teacher training is reflected in classroom practice and students' learning experiences.

Author Contributions: Conceptualization, R.T.W., P.J.d.I.S. and A.N.G.; methodology, R.T.W., P.J.d.I.S. and A.N.G.; validation, P.J.d.I.S. and A.N.G.; formal analysis, P.J.d.I.S. and A.N.G.; investigation, R.T.W. and R.T.W.; writing—original draft preparation, R.T.W.; writing—review and editing, R.T.W., P.J.d.I.S. and A.N.G.; visualization, R.T.W., P.J.d.I.S. and A.N.G.; supervision, P.J.d.I.S. and A.N.G.; project administration, R.T.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Indonesia Endowment Fund for Education (LPDP), Ministry of Finance, Republic of Indonesia, grant number LOG-5299/LPDP/LPDP.3/2023. The APC was also funded by LPDP.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki. The study involved voluntary participation of teachers, principals, and students. For student participants under the age of 18, institutional permission was obtained, and written consent was secured from legal guardians. All data were anonymised to ensure confidentiality.

Informed Consent Statement: Informed consent was obtained from all participants involved in the study. For students under the age of 18, written consent was obtained from legal guardians prior to participation.

Data Availability Statement: Data available on request due to ethical restrictions.

Acknowledgments: The authors would like to extend their sincere gratitude to the teachers, principals, and students who voluntarily participated in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TPD teacher professional development
RTBT Reflective Thinking-Based Training

References

- Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology*, 42(2), 331–342. [CrossRef]
- Alsalamah, A., & Callinan, C. (2021). Adaptation of Kirkpatrick's four-level model of training criteria to evaluate training programmes for head teachers. *Education Sciences*, 11(3), 116. [CrossRef]
- Ambu-Saidi, B., Fung, C. Y., Turner, K., & Lim, A. S. S. (2024). A critical review on training evaluation models: A search for future agenda. *Journal of Cognitive Sciences and Human Development*, 10(1), 142–170. [CrossRef]
- Avalos, B. (2011). Teacher professional development in teaching and teacher education over ten years. *Teaching and Teacher Education*, 27(1), 10–20. [CrossRef]
- Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, 41(1), 63–105. [CrossRef]
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. W.H. Freeman and Company.
- Bates, R. (2004). A critical analysis of evaluation practice: The Kirkpatrick model and the principle of beneficence. *Evaluation and Program Planning*, 27, 341–347. [CrossRef]
- Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher*, 33(8), 3–15. [CrossRef]
- Bourdieu, P. (1990). *The logic of practice*. Stanford University Press.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. [CrossRef]
- Brugha, M. E., Arif, I., Peters, S., Ahmed, F., Piccini, C., Bermudez, G. M. A., Goodland, J., Raghavendra, D., & Weeden, K. (2024). Educators' perceptions and experiences of online teacher professional development. *Journal of Interactive Media in Education*, 2024(1), 1–15. [CrossRef]
- Cahapay, M. B. (2021). Kirkpatrick model: Its limitations as used in higher education evaluation. *International Journal of Assessment Tools in Education*, 8(1), 135–144. [CrossRef]
- Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education*, 18(8), 947–967. [CrossRef]
- Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, 10(4), 331–360. [CrossRef]
- Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.
- Darling-Hammond, L. (2017). Teacher education around the world: What can we learn from international practice? *European Journal of Teacher Education*, 40(3), 291–309. [CrossRef]
- Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute. Available online: <https://learningpolicyinstitute.org/product/effective-teacher-professional-development-report> (accessed on 4 November 2025).
- Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38(3), 181–199. [CrossRef]
- Fang, Z. (1996). A review of research on teacher beliefs and practices. *Educational Research*, 38(1), 47–65. [CrossRef]
- Fraser, B. J. (2012). Classroom learning environments: Retrospect, context and prospect. In B. J. Fraser, K. G. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 1191–1239). Springer. [CrossRef]

- Gibbs, G. (1988). *Learning by doing: A guide to teaching and learning methods*. Oxford Polytechnic.
- Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Institute of Education Sciences. Available online: <https://eric.ed.gov/?id=ED521228> (accessed on 16 November 2025).
- Guskey, T. R. (2002a). Does it make a difference? Evaluating professional development. *Educational Leadership*, 59(6), 45–51.
- Guskey, T. R. (2002b). Professional development and teacher change. *Teachers and Teaching: Theory and Practice*, 8(3), 381–391. [CrossRef]
- Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Holton, E. F. (1996). The flawed four-level evaluation model. *Human Resource Development Quarterly*, 7(1), 5–21. [CrossRef]
- Kennedy, P. E., Chyung, S. Y., Winiecki, D. J., & Brinkerhoff, R. O. (2013). Training professionals' usage and understanding of Kirkpatrick's Level 3 and Level 4 evaluations. *International Journal of Training and Development*, 18(1), 1–22. [CrossRef]
- Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). *Kirkpatrick's four levels of training evaluation*. ATD Press.
- Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice-Hall.
- Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, 78(2), 311. [CrossRef]
- Loughran, J. J. (2002). Effective reflective practice: In search of meaning in learning about teaching. *Journal of Teacher Education*, 53(1), 33–43. [CrossRef]
- Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, reliability, and utility. *American Psychologist*, 52(11), 1187–1197. [CrossRef]
- Matolić, T., Jurakić, D., Jurakić, Z. G., Maršić, T., & Pedišić, Ž. (2023). Development and validation of the EDUcational course assessment TOOLkit (EDUCATOOL)—A 12-item questionnaire for evaluation of training and learning programmes. *Frontiers in Education*, 8, 1314584. [CrossRef]
- Mitchell, R., Ayinselya, R. A., Barrett, A. M., Cortez Ochoa, A. A., David, O., Imaniriho, D., Nwako, Z. A., Weldemariam Reda, N., & Singh, M. (2024). *Teacher professional development in Africa: A critical synthesis of research evidence*. Bristol Working Papers in Education Series School of Education.
- Nawaz, F., Ahmed, W., & Khushnood, M. (2022). Kirkpatrick model and training effectiveness: A meta-analysis 1982 to 2021. *Business & Economic Review*, 14(2), 35–56. [CrossRef]
- Oanh, P. T. K., Sau, N. T. U., Nhung, T. T., & Thuong, L. T. T. (2024). Overview research on teacher training evaluation based on scopus data from 2000 to 2024. *Journal of Education and Practice*, 15(3), 1–9. [CrossRef]
- OECD. (2019a). *Education policy outlook 2019: Working together to help students achieve their potential*. OECD Publishing. [CrossRef]
- OECD. (2019b). *TALIS 2018 results (volume I): Teachers and school leaders as lifelong learners*. OECD Publishing. [CrossRef]
- Opfer, V. D., & Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research*, 81(3), 376–407. [CrossRef]
- Popova, A., Evans, D. K., Breeding, M. E., & Arancibia, V. (2022). Teacher professional development around the world: The gap between evidence and practice. *World Bank Research Observer*, 37(1), 107–136. [CrossRef]
- Salas, E., Tannenbaum, S. I., Kraiger, K., & Smith-Jentsch, K. A. (2012). The science of training and development in organizations: What matters in practice. *Psychological Science in the Public Interest*, 13(2), 74–101. [CrossRef] [PubMed]
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.
- Schwarz, N., & Strack, F. (1999). Reports of subjective well-being: Judgmental processes and their methodological implications. In N. S. D. Kahneman, & E. Diener (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 61–84). Russell Sage Foundation.
- Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85(6), 956–970. [CrossRef]
- Shewchuk, S., Wallace, J., & Seibold, M. (2023). Evaluations of training programs to improve capacity in K*: A systematic scoping review of methods applied and outcomes assessed. *Humanities and Social Sciences Communications*, 10(1), 887. [CrossRef]
- Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology*, 58(1), 33–66. [CrossRef]
- Stufflebeam, D. L., & Coryn, C. L. S. (2014). *Evaluation theory, models, and applications* (2nd ed.). Jossey-Bass.
- Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development*. Ministry of Education, New Zealand. Available online: <https://www.educationcounts.govt.nz/publications/series/2515/15341> (accessed on 4 November 2025).
- Tomlinson, C. A. (1999). *The differentiated classroom: Responding to the needs of all learners* (2nd ed.). Association for Supervision and Curriculum Development.
- Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.
- Wilson, M. M., Zafar, F., & Nichol, C. (2025). Fostering inquiry: The impact of cross-curricular professional development on STEM teacher practices. *Education Sciences*, 15(4), 421. [CrossRef]
- Wisniewski, B., Röhl, S., & Fauth, B. (2022). The perception problem: A comparison of teachers' self-perceptions and students' perceptions of instructional quality. *Learning Environments Research*, 25(3), 775–802. [CrossRef]

- Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. Available online: <http://ies.ed.gov/ncee/edlabs> (accessed on 4 November 2025).
- Yu, J. (2025). Reliability and validity of applying Kirkpatrick model for evaluating exercise rehabilitation program. *Journal of Exercise Rehabilitation*, 21(4), 200–209. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.