Indonesian Journal of Educational Development (IJED)
Volume 6, Issue 4, 2026, pp. 1511-1525
ISSN: 2722-1059 (Online); ISSN: 2722-3671 (Print)
DOI: https://doi.org/10.59672/ijed.v6i4.5693

The Kirkpatrick model is utilized in Indonesian language-literature training programs for madrasa teachers in West Java

Faizal Arvianto*)[1], Rosita Rahma[2]
[1]Universitas Timor, Kefamenanu, Indonesia; faizal_arvianto@unimor.ac.id
[2]Universitas Pendidikan Indonesia, Bandung, Indonesia; rositarahma@upi.edu
*)Corresponding author: Faizal Arvianto; E-mail addresses: faizal_arvianto@unimor.ac.id

Article Info

Article history:
Received November 27, 2025
Revised December 12, 2026
Accepted January 09, 2026
Available online February 15, 2026

Keywords: Evaluation training program, Kirkpatrick model, Madrasa teachers

Abstract. An evaluation is required to determine whether a program has achieved its objectives. This process involves collecting data to inform decisions on whether the program should be continued, discontinued, or revised. This article evaluates the Indonesian Language and Literature Learning Evaluation Training program for madrasa teachers in West Java, employing the four levels of the Kirkpatrick evaluation model: reaction, learning, behavior, and results. Primary data were collected from training participants, while secondary data were obtained from evaluation tools created by the participants. Sampling was conducted using saturated and convenience sampling, with data collected via questionnaires, tests, and document analysis. The results indicate a positive reception from participants (reaction), a significant increase in cognitive scores from pre-test to post-test by 37.25% (learning), and a high rate of knowledge transfer, with 82% of participants producing highly appropriate evaluation instruments (results). These findings demonstrate that the training effectively bridged the gap between theoretical knowledge and classroom implementation. Based on these outcomes, the authors provide recommendations for program sustainability, including optimized scheduling, module development, intensive monitoring, and the establishment of periodic evaluation success criteria.

Introduction

Evaluation plays a central role in ensuring that any educational or training program operates effectively and fulfills its intended goals. In the broader context of educational development, evaluation is viewed not only as a technical activity but also as a strategic component that strengthens the alignment between program planning and real-world implementation. As emphasised in previous studies, evaluation enables institutions to monitor progress, identify strengths and weaknesses, and provide evidence-based insights to support decision-making (Altowaijri et al., 2019; Isma & Yusuf, 2025). Without systematic evaluation, programs risk becoming routine activities that continue to operate without clarity on their impact or relevance. Thus, evaluation functions as both a quality assurance mechanism and an instrument for continuous improvement (Khofifah et al., 2025).

In educational settings, the need for program evaluation has become increasingly urgent amid dynamic changes in curriculum development, pedagogical approaches, and teacher competency standards (Widana et al., 2023). Program evaluation is described as a form of providing information

that can be used as consideration in determining the objectives achieved, design, implementation, and impact to assist in decision making, accountability, and improving understanding of the phenomenon (Antonie, 2011; Hosseini et al., 2022; Powell & Bodur, 2019). This definition highlights that evaluation is not merely an afterthought but an integral part of the educational cycle. The growing emphasis on accountability and transparency across institutions further reinforces the need for evaluations that are systematic, reliable, and aligned with measurable indicators.

Furthermore, evaluation must be conducted through clear stages to ensure objectivity and consistency. The stages of program evaluation implementation include: 1) setting objectives, 2) establishing analysis and criteria, 3) determining the sample, 4) conducting the program evaluation, and 5) formulating decisions to be followed up with recommendations (Christie & Fierro, 2010; Frye & Hemmer, 2012; Kus, 2025; Owston, 2008; Pratomo & Shofwan, 2022). Each step provides a structured pathway to prevent oversight and ensure that conclusions drawn from the evaluation process are well-founded. For example, defining clear objectives is essential for guiding which aspects of a program to measure, whereas establishing proper criteria ensures that judgments are based on valid evidence rather than subjective perception.

External factors often influence program implementation, and evaluation helps uncover barriers and potential opportunities for improvement. The evaluation results will reveal several obstacles that need to be overcome, thereby improving existing opportunities (Pardo, 2011; Philibert et al., 2018). These findings can provide policymakers, educators, and program planners with critical insights, enabling them to refine strategies and improve the effectiveness of future activities. In essence, evaluation is a reflective process that encourages stakeholders to step back, assess performance, and make informed choices about next steps.

In addition, the object of evaluation varies widely; it can be a program, product, process, or policy. An essential aspect of this definition is that evaluation focuses on a specific object and aims to influence how people think about actions and or change their behaviour in relation to the object (Ilhami, 2024; Levy-Feldman, 2025; Netzer et al., 2018; Reed et al., 2021). This reinforces the notion that evaluation is inherently action-oriented and transformative. Rather than merely identifying deficiencies, evaluation aims to promote changes that enhance educational effectiveness and learner outcomes.

The purpose of program evaluation is to contribute to research and to provide feedback for improving teaching and learning practices, as interpreted by those involved in learning and decision-makers (Conole & Oliver, 2006; Hosseini et al., 2022; Rallis & Bolland, 2004; Rianyansa & Maisarah, 2024). Its role in supporting pedagogical innovation is particularly crucial in teacher training programs. As educational challenges evolve, teachers need continuous professional development to remain responsive to student needs and curriculum demands. Program evaluation helps ensure that the design and delivery of such training remain relevant, evidence-based, and aligned with global standards.

The urgency required in program evaluation includes: 1) the existence of fund expenditures with the achievement of program objectives and targets, 2) program decisions made, and 3) the collection of information in recommendations for the development of the following program action. Studies in program evaluation are conducted to collect and synthesise information about the status, value, achievements, significance, or quality of programs, products, people, policies, proposals, or plans, which are also included in program evaluation actions (Martens, 2023; Suklani, 2023; Zomorrodian & Matei, 2010). Through this multidimensional approach, program evaluation supports both short-term improvements and long-term strategic planning.

This evaluation activity also needs to be carried out for the Indonesian language and literature learning evaluation training program for madrasa teachers in West Java, organised by the Indonesian Language and Literature Education Study Program, Universitas Pendidikan Indonesia. As a long-running program, it has contributed substantially to the professional development of madrasa teachers, yet it has not been rigorously evaluated using a structured framework. The absence of a comprehensive evaluation poses risks, including the inability to measure learning gains, unclear behavioral change among participants, and the possibility of program stagnation.

So far, evaluation has been limited to reporting the number of participants who attended and the obstacles encountered during the activity. Such an approach does not provide sufficient insight into whether the program has achieved its educational objectives, nor does it offer a robust foundation for improvement. Therefore, it is necessary to conduct a more appropriate and systematic evaluation of the program. The author conducted a scientific study to evaluate the Indonesian language and literature learning evaluation training program for madrasa teachers in West Java using the Kirkpatrick evaluation model, which includes: reaction level evaluation, learning level evaluation, behaviour level evaluation, and result level evaluation (Farjad, 2012; Kirkpatrick, 2007), followed by recommendations for the program's sustainability.

A comprehensive evaluation using the Kirkpatrick model is expected to provide a deeper understanding of how the program contributes to teachers' professional growth, their instructional practices, and the broader institutional outcomes within madrasa education. Ultimately, the findings will serve not only as documentation of program performance but also as a strategic foundation for enhancing future training initiatives.

The novelty of this research lies in its comprehensive application of the four-level Kirkpatrick evaluation model (covering reaction, learning, behavior, and results) specifically tailored to an Indonesian language and literature training program for madrasa teachers. While previous evaluations of teacher training often focus solely on participant satisfaction (reaction) or immediate knowledge gains (learning), this study extends the analysis to track changes in instructional behavior and the tangible quality of evaluation instruments produced by teachers in their respective schools. By focusing on madrasa educators in West Java, this study addresses a specific institutional gap, as madrasa-based professional development often faces unique pedagogical and administrative challenges compared to general schools.

Method

This evaluation research employs a descriptive, quantitative approach, utilising survey techniques. According to Creswell & Creswell, surveys can provide quantitative descriptions of trends, attitudes, and opinions within a population. Survey techniques can be used to answer three types of research questions, namely: 1) descriptive questions; 2) questions to examine the relationship between two variables; and 3) questions to predict the relationship between two variables (Creswell & Creswell, 2018). In this study, the type of question referred to is the first type, namely descriptive questions. The evaluation analysis model used in this study is the Kirkpatrick model. This model was chosen because it offers several advantages, including a straightforward system and language for collecting data, as well as descriptive and evaluative information that is suitable for the required results (Khan & Ali, 2022; Yu, 2025). Additionally, it provides a more practical approach to the evaluation process, which is often complex (Bates, 2004).

According to Kirkpatrick, the evaluation process consists of four levels. These levels are reaction, learning, behaviour, and results (Kirkpatrick, 2007). The first level is reaction, which refers to how participants view and subjectively evaluate the relevance and quality of the training. According to

Kirkpatrick, every program should be assessed at this level to enable improvement. At this level, evaluation measures the participants' satisfaction. The second level is learning. Analysis at this level can be described as the extent to which participants' attitudes change, their knowledge increases, or their skills are expanded as a result of training. The third level of evaluation is behaviour. This level examines the changes in the behaviour or performance of training participants resulting from the training. The fourth level is results. Level four evaluation attempts to assess training in terms of the results or output produced after training is implemented.

The data in this study are primary, consisting of responses from training participants, their abilities, and evaluation tools developed by the training participants themselves. The data sources in this study include both primary and secondary data. The primary data sources were obtained from training participants, while the additional data sources were gathered from documents, including evaluation tools developed by the training participants. The sampling techniques employed in this study included saturated and convenience sampling (Etikan, 2016). Saturation sampling was used to collect data on training participants' responses at both the reaction and learning levels. Convenience sampling, which involves sampling based on the availability of respondents or social ties (friends, colleagues, and others) (Vehovar et al., 2016), was employed at both the behavioural and the results levels.

The data collection techniques employed in this study included questionnaires, tests, and document analysis. Questionnaires were used at the reaction and behaviour levels to collect data on participants' responses to the program. The questionnaires were developed using the Likert scale, which features five response options (Grassini & Laumann, 2020; Taherdoost, 2019). The questionnaire contained 13 items with questions related to the implementation of the training program, namely: (1) ease of delivery of material; (2) suitability of material to the competencies that teachers must have; (3) suitability of material to the current needs of teachers; (4) suitability of material to the curriculum; (5) the usefulness of the material in increasing general knowledge; (6) the accuracy of the material in relation to theoretical studies; (7) the usefulness of the material in school assessment applications; (8) the usefulness of the material in improving teacher competencies; (9) the usefulness of the material in improving teacher knowledge; (10) the clarity of illustrations and examples; (11) the clarity of the media used to deliver the material; (12) the clarity and accuracy of the language used in general; and (13) the clarity and accuracy of the language used in the material.

Testing techniques are used at the learning level to collect data on participants' abilities before and after the training program. Testing techniques are developed in accordance with the material provided during training. At the results level, the data collection technique used is document analysis. The documents assessed and analysed are those related to the evaluation tools developed and used by training participants in schools. Validity testing in a study assesses whether the instruments used are appropriate with respect to content parameters (Masuwai et al., 2024; Sugiyono, 2013). The results of validity testing also indicate that the measured variables are indeed the variables intended by the researcher (Cooper & Schindler, 2014). The validity test commonly used in research is Pearson's bivariate correlation, which is performed in SPSS 25.

The data analysis technique used in this study is the percentage technique. This method is used to determine the frequency with which respondents answer the research questions. Additionally, the percentage technique is used to determine the proportion of each answer, thereby facilitating analysis (Moleong, 2018). The percentage technique procedure includes 1) data examination; 2) data classification; 3) tabulation; 4) calculating the frequency of responses; 5) visualising data in the form of tables or graphs; 6) interpreting data in accordance with the research questions. The percentage technique formula used in this study is as follows.

$$P = \frac{F}{N} \; x \; 100\%$$

Explanation:
P = percentage value
F = frequency of responses
N = total number of respondents (Moleong, 2018)

Results and Discussion

The Kirkpatrick model of program evaluation was implemented in the Indonesian language and literature learning evaluation training for madrasa teachers in West Java, organised by the Indonesian Language and Literature Education Study Program at the Universitas Pendidikan Indonesia. The Kirkpatrick evaluation model comprises four levels: reaction, learning, behaviour, and results. Analysis was conducted at each level to obtain information on the implementation of the training program.

Reaction Level Evaluation

A training program can be considered successful if participants respond positively to the entire process carried out during the activity (Bhat & Rainayee, 2025; Sugandini et al., 2025). During the activity, 205 participants completed the questionnaire. Each participant filled out the questionnaire via a Google Form link provided by the committee. The questionnaire required participants to respond to 13 questions about the activity, such as (1) ease of delivery of the material; (2) suitability of the material to the competencies that teachers should have; (3) suitability of the material to the current needs of teachers; (4) suitability of the material to the curriculum; (5) the usefulness of the material in broadening general knowledge; (6) the accuracy of the materials substance in relation to theoretical studies; (7) the usefulness of the material in relation to assessment applications in schools; (8) the usefulness of the material in improving teacher competencies; (9) the usefulness of the material in improving teacher knowledge; (10) the clarity of illustrations and examples; (11) the clarity of the media used to deliver the material; (12) the clarity and accuracy of the language used in general; and (13) the clarity and accuracy of the language used in the material. Data related to participants' responses/reactions to the implementation of the training program can be shown in Image 1 below.

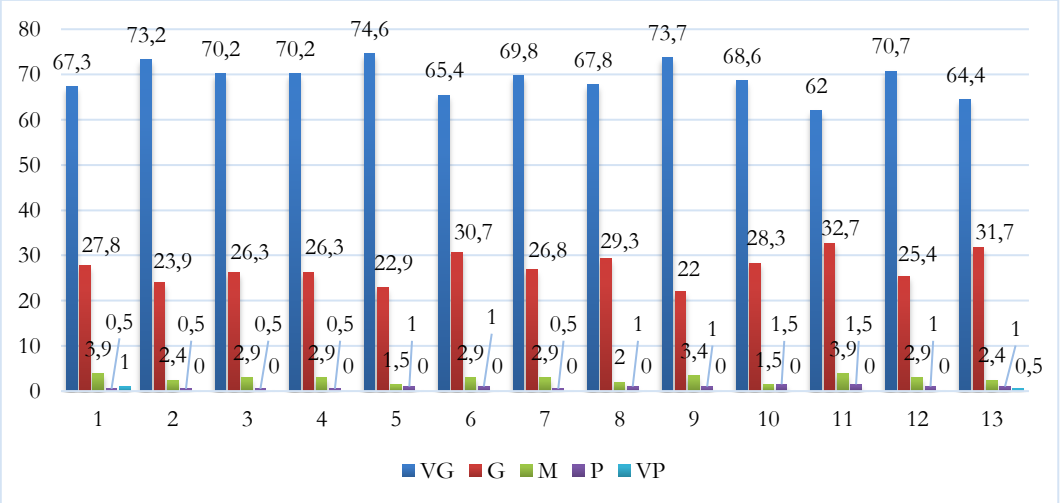

Image 1. Training Participants Responses

Description:
1) Ease of material delivery, 2) Relevance of material to the competencies that teachers must possess, 3) Relevance of material to current teacher needs, 4) Relevance of the material to the curriculum, 5) The usefulness of the material in broadening general knowledge, 6) Accuracy of material content with theoretical studies, 7) The usefulness of the material in relation to assessment applications in schools, 8) The usefulness of the material in improving teachers competence, 9) The usefulness of the material in enhancing teachers knowledge, 10) Clarity of illustrations and examples, 11) Clarity of the media used to deliver the material, 12) Clarity and accuracy of language use in general, 13) Clarity and accuracy of language use in teaching materials.

Based on the questionnaire above, the average percentage of participants who responded 'Very Good' to all questions was 69.07%, and the percentage who responded 'Good' was 27.24%. Thus, these positive responses, when accumulated, amounted to 96.31%. 'Moderate' responses amounted to 2.73%, 'Poor' responses amounted to 0.88%, and 'Very Poor' responses amounted to 0.12%. Thus, the adverse reactions amounted to 2.85%. The problems that can be identified in relation to these negative responses are: (1) Participants internet connections were unstable, causing many participants to leave and re-enter the training room, or encounter obstacles in accessing information during the activity; (2) Participants focus was divided because the activity coincided with their teaching schedule at school; (3) Some terms in the material were not well understood by participants, making it difficult for some participants to understand the material presented. These issues need to be evaluated for the implementation of similar training activities in the following year. Based on the data presented, it can be concluded that the Indonesian language and literature learning evaluation training program was received positively by training participants, who were madrasa teachers in West Java.

Learning Level Evaluation

Learning evaluation assesses the extent of improvement in the knowledge, attitudes, and skills of training participants (Azmy & Setiarini, 2023; Shek & Chak, 2012; Shewchuk et al., 2023). Learning evaluation focuses more on assessing learning outcomes or outputs (Caro et al., 2026; Coates, 2015; Garvey & Kiegaldie, 2023). Therefore, learning measurement utilises performance assessments to evaluate the knowledge acquired, changes in attitudes, and skills developed or improved. Knowledge evaluation is conducted using pretests and posttests at the beginning and end of the session (Jayaratne et al., 2025; Shivaraju et al., 2017). Meanwhile, attitudes and skills are assessed during the training session through performance evaluations (Jupri et al., 2025; Ozogul & Sullivan, 2009). A pretest was conducted before the training activity began by distributing the question link via G-Form. At the same time, posttests were administered after the training activity was completed, with participants receiving a G-Form link containing the same material but different questions. The results of the knowledge evaluation analysis, comparing the pretest and posttest, are presented in Images 2 and 3 below.

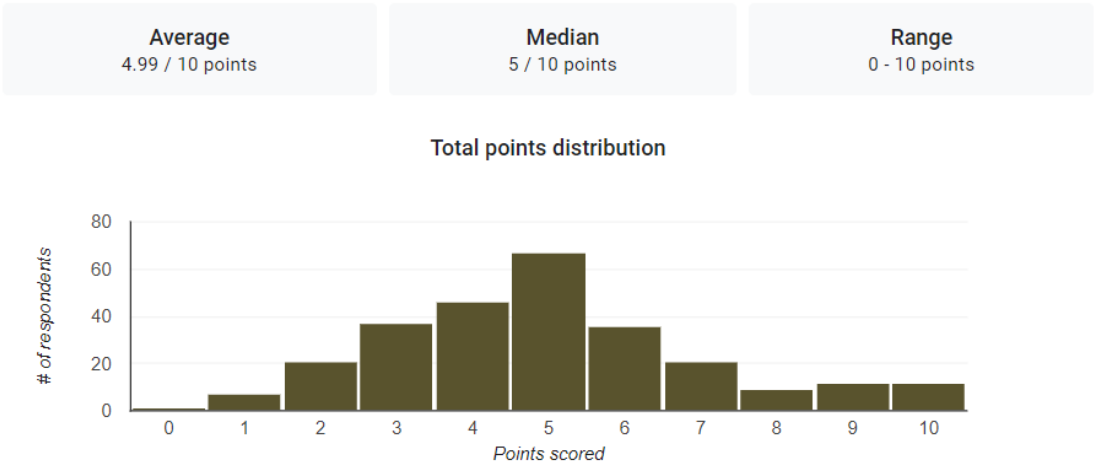

Image 2. Pretest Results of the Training Activity

Image 2 shows that the average pretest score of the participants was 4.99 out of a total score of 10, with a median of 5. This score indicates that the participants' understanding before attending the training was very low because the average score was below the median.

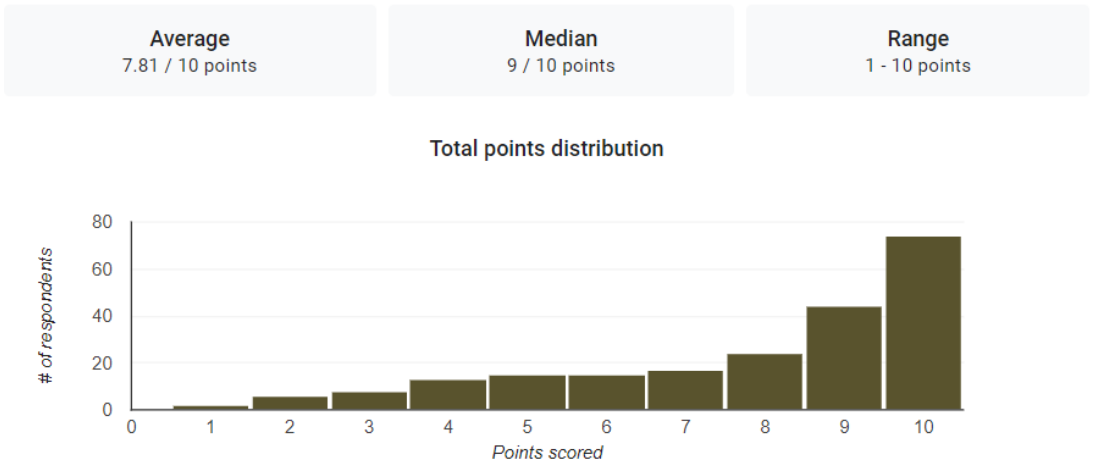

Image 3. Posttest Results of the Training Activity

Image 3 shows that participants' average posttest score was 7.81 out of 10, with a median of 9. Thus, the score increased by 37.25% from the pretest to the posttest. This suggests that participants' understanding increased after attending the training. Attitude and skill assessments were conducted through performance evaluations (Nollen & Gaertner, 1991; Pratiwi & Wahjoedi, 2024; Sudirman et al., 2023). Training participants were asked to develop a set of tests assessing Indonesian language and literature competencies. The participants' questions were then presented and evaluated during the second training session. The results of the performance assessment in the second session are presented in Image 4 below.

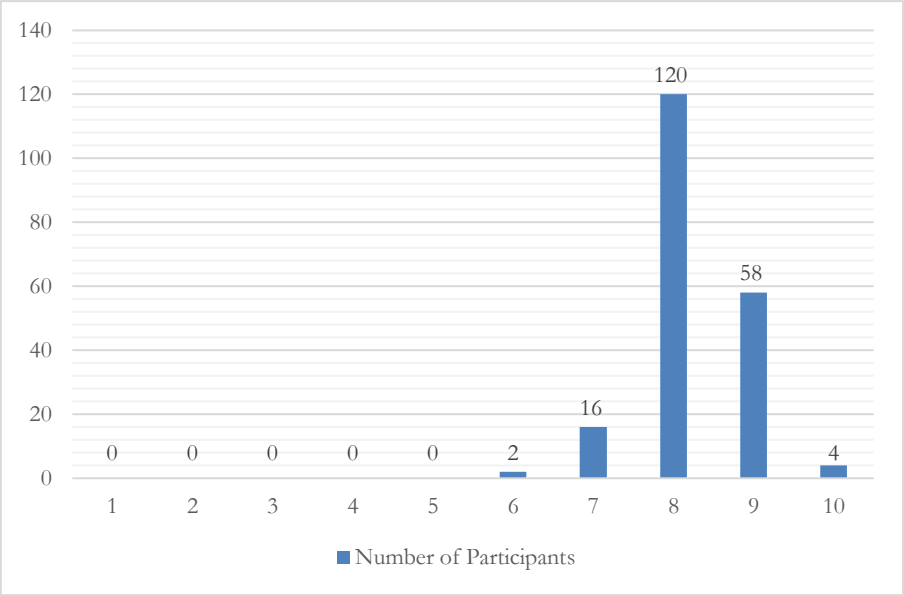

Image 4. Training Participants Performance Scores

Image 4 shows the performance scores of training participants during the second session. Of the total 200 participants involved in the second training session, there were 2 participants (1%) who scored 6; 16 participants (8%) who scored 7; 120 participants (60%) who scored 8; 58 participants (29%) who scored 9; and 2 participants (2%) who scored 10. From this data, the group average score was 8.23. Compared with the pretest average score of 4.99, the training participants' scores improved significantly.

Behaviour Level Evaluation

The evaluation at the behavioural level aims to assess whether the knowledge and skills acquired during training are applied in the work environment of the training participants (Ikramina & Gustomo, 2014; Rouse, 2011). To obtain data on the results of the behavioural-level evaluation, a survey was distributed to training participants via a Google Form link 1 month after the training. A total of 187 participants completed the link. The survey results are presented in Images 5 and 6 below.

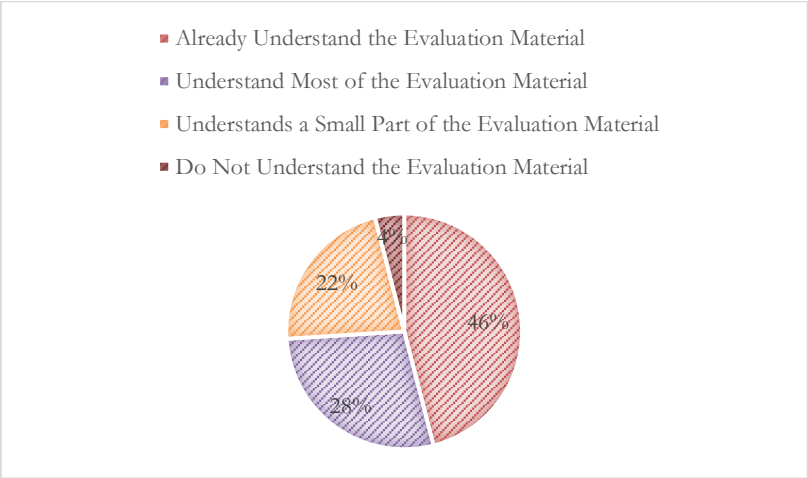

Image 5. Survey Results on Material Comprehension After Training

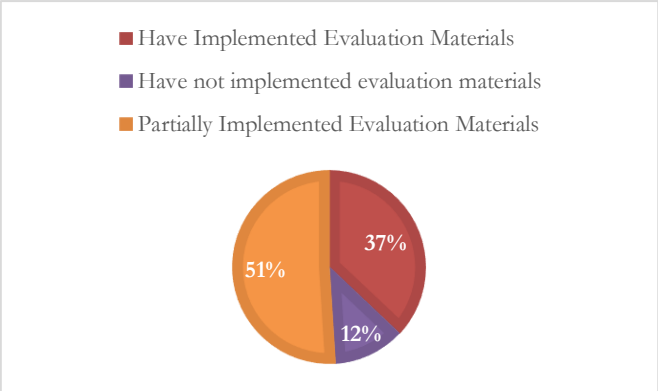

Image 6. Survey Results on Material Implementation After Training

Image 5 shows that 46% of participants claimed to have understood the training material even after the training activity was completed; 28% of participants understood most of the material obtained; 22% of participants understood a small part of the material obtained; and 4% of participants claimed not to understand or remember the evaluation material during training. Thus, most participants were assessed as still remembering and understanding the evaluation material after training.

Image 6 shows that most training participants (51%) have implemented the material learned in the training; 37% have only partially implemented the evaluation material; and 12% have not implemented the evaluation material in teaching at school. Thus, it can be concluded that most participants have implemented the evaluation material from the training. Based on the data results in Images 5 and 6, it can be concluded that the evaluation training program implemented has had an impact on the performance behaviour of madrasa teachers in developing learning evaluations in accordance with the conceptual understanding presented in the training material.

Result Level Evaluation

Evaluation at the outcome level aims to identify participants' performance outcomes after they have received training materials (Mehale et al., 2021; Rindarti, 2021; Rouse, 2011; Satyani, 2020; Steensma & Groeneveld, 2010; Zareisaroukolaei et al., 2024). To obtain data at the assessment level, a document analysis was conducted of the evaluation instruments developed by teachers at the school. A total of 30 documents were analysed, representing a sample of training participants. The results of the analysis of these documents are presented in Image 7 below.

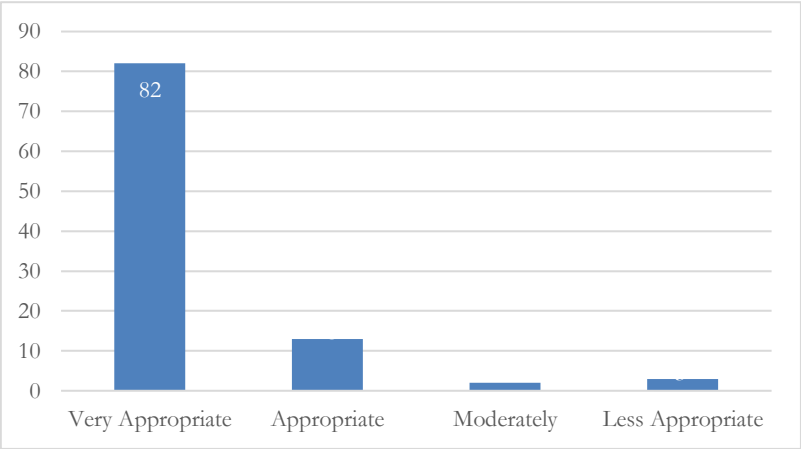

Image 7. Results of the Analysis of Teacher Evaluation Tools

Image 7 shows that 82% of the evaluation tools developed by teachers after receiving training were highly appropriate for the material; 13% of the tools were categorised as applicable; 2% of the evaluation tools were categorised as moderately appropriate; and 3% of the evaluation tools were classified as inappropriate for the evaluation material taught. Thus, based on the sample documents analysed, it can be concluded that the training program implemented had an impact on the competence and performance of madrasa teachers in developing learning evaluation tools.

The novelty of these research findings demonstrates a significant correlation between structured training interventions and improvements in teacher competencies. The data reveal that the training did not merely result in a 37.25% increase in cognitive scores, but more importantly, it facilitated a high rate of knowledge transfer into professional practice. The fact that 82% of the evaluation tools developed by teachers post-training were categorized as "highly appropriate" provides empirical evidence that the Kirkpatrick model can effectively bridge the gap between theoretical pedagogical training and classroom-level implementation in a madrasa context.

Despite the comprehensive insights provided by this evaluation, several limitations of the study warrant consideration. First, the use of convenience sampling for the behavior and results levels may constrain the generalizability of the findings to the broader population of madrasa teachers across Indonesia. Second, the assessment of instructional changes relied heavily on participants' self-reports through digital surveys, which introduces a potential risk of social desirability bias compared to direct classroom observations. Furthermore, the evaluation of the results level (Level 4) was conducted within a relatively short timeframe of one-month post-training; consequently, the long-term sustainability of the instructional improvements and their direct impact on student learning outcomes could not be fully captured in this study. Future research should consider longitudinal approaches and objective observational methods to further validate the training's long-term effectiveness.

The findings of this research provide significant insights for both the theoretical and practical aspects of educational program evaluation, especially regarding teacher professional development in madrasa education. From a theoretical standpoint, this study strengthens and underscores the Kirkpatrick four-level evaluation model as a comprehensive framework that encompasses not only participants' immediate reactions and learning outcomes but also the transformation of behaviors and measurable performance results. By empirically illustrating a distinct advancement from cognitive enhancement (37.25% increase in post-test scores) to modifications in instructional behavior and the creation of high-quality assessment products (82% of evaluation instruments deemed highly appropriate), this study reinforces the notion that the Kirkpatrick model adeptly addresses the enduring disconnect between training theory and actual classroom practice. Additionally, this study broadens the scope of existing evaluation literature by situating the model within the context of Indonesian language and literature education in madrasa environments, a field that has seen limited attention in previous empirical inquiries. This contextual enhancement adds depth to the conversation surrounding culturally and institutionally adaptive evaluation models in teacher education.

From a practical perspective, the findings provide policymakers, curriculum designers, and training coordinators with tangible evidence that structured, evaluation-focused professional development initiatives can result in lasting improvements in teachers' assessment skills. The clear transfer of training results into actual instructional materials underscores the need to align training content with teachers' immediate classroom needs and the specific contexts of their institutions. In practical terms, the results advocate the implementation of systematic evaluation processes as a fundamental aspect of teacher training programs, especially in madrasa settings, where instructional evaluation practices often need reinforcement. Furthermore, the research offers practical recommendations

for program enhancement, including improved scheduling, the development of comprehensive training modules, and systems to monitor post-training effectiveness. Ultimately, the practical significance of this study resides in its capacity to guide evidence-based decision-making, improve the quality of learning evaluation practices, and contribute to the overarching objective of elevating instructional quality and accountability within the Indonesian madrasa education system.

Conclusion

Based on the analysis presented in the previous sections, it can be broadly concluded that the Indonesian Language and Literature Learning Evaluation Training Program for Madrasa Teachers in West Java has demonstrated significant value and therefore merits continuation and further development. Several key considerations support this recommendation. *First*, the program scheduling plays a crucial role in ensuring the success and full participation of madrasa teachers. Conducting the training on weekends or national holidays is strongly recommended, as this arrangement allows participants to engage without the competing demands of teaching schedules and administrative obligations. When teachers can devote their attention exclusively to the training materials and activities, knowledge absorption, participation in discussions, and engagement in practical sessions become more effective. This scheduling strategy also supports a more relaxed and focused learning atmosphere, ultimately enhancing the program's impact. *Second*, developing a comprehensive training module is an essential improvement. The module should compile all materials delivered throughout the program, including theoretical foundations, practical guidelines, examples of evaluation instruments, case studies, and reflective exercises. Providing such a resource enables participants to review and deepen their understanding after the program ends. Long-term retention of knowledge is more likely when participants have structured materials they can revisit and apply in their classroom and institutional contexts. Additionally, the module can function as a reference for teachers who wish to share what they have learned with colleagues, thereby multiplying the program's reach and impact. *Third*, continuous monitoring by the Head of the Study Program is crucial for maintaining quality and ensuring alignment with institutional goals. Regular oversight—from early planning to final reporting—facilitates the identification of both strengths and weaknesses. This process helps the organizing team refine technical arrangements, adjust content as needed, and respond promptly to emerging challenges. Such monitoring also increases accountability and strengthens the academic integrity of the training. *Fourth*, establishing clear, measurable criteria for program success is necessary. These criteria may include participant satisfaction, improvements in teachers' knowledge and skills, the relevance of materials to classroom practice, and evidence of follow-up implementation in madrasa settings. Having well-defined indicators not only enables systematic evaluation but also allows organizers to track progress over time, compare outcomes across cohorts, and develop strategies for continuous improvement. Transparent criteria also help participants understand program expectations and outcomes. *Finally*, it is strongly recommended that a dedicated evaluation team be formed to conduct assessments before, during, and after program implementation. Pre-evaluation helps map participants' baseline competencies and needs, guiding the adaptation of the training materials. Monitoring during the program ensures that sessions run smoothly and provides real-time feedback for facilitators. Post-evaluation, meanwhile, captures the program's effectiveness and its longer-term impact on teachers' professional practices. A specialized evaluation team can thus contribute significantly to the program's sustainability, credibility, and long-term effectiveness. In conclusion, with structured scheduling, well-developed learning resources, strong managerial oversight, measurable success indicators, and an integrated evaluation mechanism, the training program will be better positioned to enhance the competence of madrasa teachers in West Java in conducting evaluations of Indonesian Language and Literature learning. Strengthening these aspects will not only improve the quality of future training cycles but also contribute to the broader goal of educational quality in madrasa environments.

Bibliography

Altowaijri, S., Rahman, A. U., Alfawareh, H. M., & Altowaijri, S. M. (2019). Program outcomes assessment and evaluation for continuous improvements. *IJCSNS International Journal of Computer Science and Network Security*, *19*(4), 40. https://www.researchgate.net/publication/338611508

Antonie, R. (2011). The role of program evaluation in the decision-making process. *Transylvanian Review of Administrative Sciences*, *33E*.

Azmy, A., & Setiarini, N. Y. (2023). Kirkpatrick model as evaluation training program for assessor: case study of government employee. *International Journal of Management, Accounting and Economics*, *10*(9), 2383–2126. https://doi.org/10.5281/zenodo.10115328

Bates, R. (2004). A critical analysis of evaluation practice: the kirkpatrick model and the principle of beneficence. *Evaluation and Program Planning*, *27*(3), 341–347. https://doi.org/10.1016/j.evalprogplan.2004.04.011

Bhat, Z. H., & Rainayee, R. A. (2025). Linking 'positive reactions' to utility reactions and trainee satisfaction: a structural equation modeling approach. *International Journal of Training Research*, *23*(1), 20–46. https://doi.org/10.1080/14480220.2024.2314014

Caro, M. F., Flórez, E. P., & Muñoz, I. C. (2026). A formal model for assessing the learning outcomes of academic programs. *Evaluation and Program Planning*, *114*, 102644. https://doi.org/10.1016/j.evalprogplan.2025.102644

Christie, C. A., & Fierro, L. A. (2010). Program evaluation. In *International Encyclopedia of Education*. Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.01618-3

Coates, H. (2015). *Assessment of learning outcomes BT - The european higher education area: between critical reflections and future policies* (A. Curaj, L. Matei, R. Pricopie, J. Salmi, & P. Scott (eds.); pp. 399–413). Springer International Publishing. https://doi.org/10.1007/978-3-319-20877-0_26

Conole, G., & Oliver, M. (2006). *Contemporary perspectives in e-learning research* (1st Editio). Routledge. https://doi.org/10.4324/9780203966266

Cooper, D., & Schindler, P. (2014). *Business research methods: 12th Edition* (12th Editi). MCGRAW-HILL US HIGHER ED. https://books.google.co.id/books?id=AZ0cAAAAQBAJ

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches (fifth edition)*. Sage Publications, Inc.

Etikan, I. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1. https://doi.org/10.11648/j.ajtas.20160501.11

Farjad, S. (2012). The evaluation effectiveness of training courses in university by kirkpatrick model (Case study: Islamshahr university). *Procedia - Social and Behavioral Sciences*, *46*, 2837–2841. https://doi.org/10.1016/j.sbspro.2012.05.573

Frye, A. W., & Hemmer, P. A. (2012). Program evaluation models and related theories: AMEE Guide No. 67. *Medical Teacher*, *34*(5), e288–e299. https://doi.org/10.3109/0142159X.2012.668637

Garvey, L., & Kiegaldie, D. (2023). *Assessment and evaluation in nursing education: A simulation perspective bt - comprehensive healthcare simulation: Nursing* (J. M. Kutzin, K. T. Waxman, C. M. Lopez, & D. Kiegaldie (eds.); pp. 143–153). Springer International Publishing. https://doi.org/10.1007/978-3-031-31090-4_14

Grassini, S., & Laumann, K. (2020). Questionnaire measures and physiological correlates of presence: A systematic review. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00349

Hosseini, S., Yilmaz, Y., Shah, K., Gottlieb, M., Stehman, C. R., Hall, A. K., & Chan, T. M. (2022). Program evaluation: An educator's portal into academic scholarship. *AEM Education and Training*, *6*(S1). https://doi.org/10.1002/aet2.10745

Ikramina, F., & Gustomo, A. (2014). Analysis of training evaluation process using kirkpatrick's

training evaluation model at pt. bank tabungan negara (Persero) Tbk. *Journal of Business and Management*, *3*(1), 102–111.

Ilhami, M. R. (2024). Evaluation of the goals and objects of assessment of aspects of student learning through bloom's taxonomy. *International Journal of Health, Economics, and Social Sciences (IJHESS)*, *6*(2 SE-Articles), 401–406. https://doi.org/10.56338/ijhess.v6i2.4968

Isma, N., & Yusuf, M. (2025). *The influence of the implementation of extracurricular activities of the islamic propagation agency on the practice of religious worship at mutia rahma bulu cina middle school , hamparan perak district*. *5*(1), 211–215. https://doi.org/10.30596/jcositte.v1i1.xxxx

Jayaratne, K. S. U., Kumar Chaudhary, A., & Diaz, J. M. (2025). Knowledge testing options in pre-test post-test evaluation design: implications for extension program evaluation. *Advancements in Agricultural Development*, *6*(4), 64–75. https://doi.org/10.37433/aad.v6i4.659

Jupri, T. J., Tinus, A., & Wurianto, A. B. (2025). Mengembangkan berpikir kritis melalui penilaian autentik berbasis hots di bidang ekonomi (Developing critical thinking through hots-based authentic assessment in economics). *Indonesian Journal of Innovation Studies*, *26*(3 SE-Innovation in Economics, Finance and Sustainable Development), 10.21070/ijins.v26i3.1453. https://ijins.umsida.ac.id/index.php/ijins/article/view/1453

Khan, D., & Ali, S. (2022). Training evaluation models: Comparative analysis. *Research Journal of Social Sciences & Economics Review*, *3*(4), 2707–9015. https://doi.org/10.36902/rjsser-vol3-iss4-2022

Khofifah, J. M., Nurdin, D., & Herawan, E. (2025). Enhancing teacher professionalism through academic supervision: A CIPP model evaluation. *Indonesian Journal of Educational Development (IJED)*, *6*(2), 380–392. https://doi.org/10.59672/ijed.v6i2.4727

Kirkpatrick, J. (2007). *The hidden power of kirkpatrick's four levels* (Vol. 61).

Kus, M. (2025). Evolution of program evaluation: A historical analysis of leading theorists' views and influences. *Education Quarterly Reviews*, *8*, 142-155. https://doi.org/10.31014/aior.1993.08.01.561

Levy-Feldman, I. (2025). The role of assessment in improving education and promoting educational equity. *Education Sciences*, *15*(2), 224. https://doi.org/10.3390/educsci15020224

Martens, D. M. (2023). *Review of research and evaluation in education and psychology: integrating diversity with quantitative, qualitative, and mixed methods* (6th Editio). Sage Publications, Inc.

Masuwai, A., Zulkifli, H., & Hamzah, M. I. (2024). Evaluation of content validity and face validity of secondary school Islamic education teacher self-assessment instrument. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2308410

Mehale, K. D., Govender, C. M., & Mabaso, C. M. (2021). Maximising training evaluation for employee performance improvement. *SA Journal of Human Resource Management*, *19*, 1–11. https://doi.org/10.4102/sajhrm.v19i0.1473

Moleong, L. J. (2018). *Metode penelitian kualitatif (Qualitative Research Methods)*. PT. Remaja Rosdakarya.

Netzer, L., Gutentag, T., Kim, M. Y., Solak, N., & Tamir, M. (2018). Evaluations of emotions: Distinguishing between affective, behavioral and cognitive components. *Personality and Individual Differences*, *135*, 13–24. https://doi.org/10.1016/j.paid.2018.06.038

Nollen, S. D., & Gaertner, K. N. (1991). Effects of skill and attitudes on employee performance and earnings. *Industrial Relations: A Journal of Economy and Society*, *30*(3), 435–455. https://doi.org/10.1111/j.1468-232X.1991.tb00797.x

Owston, R. (2008). *Handbook of research on educational communications and technology* (D. Jonassen, M. J. Spector, M. Driscoll, M. D. Merrill, J. van Merrienboer, & M. P. Driscoll (eds.); 3rd Editio). Routledge. https://doi.org/10.4324/9780203880869

Ozogul, G., & Sullivan, H. (2009). Student performance and attitudes under formative evaluation by teacher, self and peer evaluators. *Educational Technology Research and Development*, *57*(3), 393–410. http://www.jstor.org/stable/40388636

Pardo, R. (2011). *The evaluation and optimization of trading strategies* (2nd Editio). Wiley.

Philibert, I., Beernink, J. H., Bush, B. H., Caniano, D. A., Coyle, J. J., Gilhooly, J., Kraybill, D. E., Larson, D., Nace, M. C., Robertson, W. W., Rubin, J. D., Sanford, T., Chow, A., & Moran, S. (2018). Improving the improvement process: 5 dimensions of effective program evaluation and improvement. *Journal of Graduate Medical Education*, *10*(1), 114–117. https://doi.org/10.4300/JGME-D-18-00071.1

Powell, C. G., & Bodur, Y. (2019). Teachers' perceptions of an online professional development experience: implications for a design and implementation framework. *Teaching and Teacher Education*, *77*, 19–30. https://doi.org/10.1016/j.tate.2018.09.004

Pratiwi, Y. A. I., & Wahjoedi, T. (2024). The influence of attitude, skill, and knowledge control on employee perfomance at pt. astha kencana mulia. *INCOME: Innovation of Economics and Management*, *2*(3), 19–23. https://doi.org/10.32764/income.v2i3.3749

Pratomo, R. Y., & Shofwan, I. (2022). Implementation of education and training program evaluation. *Edukasi*, *16*(2), 63–77. https://doi.org/10.15294/edukasi.v16i2.39863

Rallis, S. F., & Bolland, K. A. (2004). What is program evaluation? Generating knowledge for improvement. *Archival Science*, *4*(1–2), 5–16. https://doi.org/10.1007/s10502-005-6988-4

Reed, M. S., Ferré, M., Martin-Ortega, J., Blanche, R., Lawford-Rolfe, R., Dallimer, M., & Holden, J. (2021). Evaluating impact from research: A methodological framework. *Research Policy*, *50*(4), 104147. https://doi.org/10.1016/j.respol.2020.104147

Rianyansa, A. A., & Maisarah, I. (2024). Students' perception of literature as a teaching strategies in enrich the EFL students' vocabulary. *Indonesian Journal of Educational Development (IJED)*, *5*(2), 282–291. https://doi.org/10.59672/ijed.v5i2.3465

Rindarti, E. (2021). Implementasi coaching untuk meningkatkan kemampuan kepala madrasah melaksanakan evaluasi pembelajaran jarak jauh (Implementation of coaching to improve the ability of madrasah principals to carry out distance learning evaluations). *Indonesian Journal of Educational Development*, *2 Nomor 3*(November), 401-415 https://doi.org/10.5281/zenodo.5680948 __. https://doi.org/10.5281/zenodo.5680948

Rouse, D. N. (2011). Employing kirkpatrick's evaluation framework to determine the effectiveness of health information management courses and programs. *Perspectives in Health Inforrmation Managemen*, *8*(Spring), 1c.

Satyani, I. A. P. (2020). Mewujudkan metamorfosis sd negeri 8 mas melalui manajemen keterlibatan masyarakat lokal dan global (Realizing the metamorphosis of 8 Mas State Elementary School through local and global community involvement management) . *Indonesian Journal of Educational Development*, *1*(3), 417–428. https://doi.org/10.5281/zenodo.4285148

Shek, D. T. L., & Chak, Y. L. Y. (2012). Evaluation of the training program of the project p.a.t.h.s.: findings based on the perspective of the participants from different cohorts. *The Scientific World Journal*, *2012*, 1–10. https://doi.org/10.1100/2012/687198

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. https://doi.org/10.1057/s41599-023-02403-5

Shivaraju, P., Manu, G., M, V., & Savkar, M. (2017). Evaluating the effectiveness of pre- and post-test model of learning in a medical school. *National Journal of Physiology, Pharmacy and Pharmacology*, *7*(9), 1. https://doi.org/10.5455/njppp.2017.7.0412802052017

Steensma, H., & Groeneveld, K. (2010). Evaluating a training using the "four levels model." *Journal of Workplace Learning*, *22*(5), 319–331. https://doi.org/10.1108/13665621011053226

Sudirman, S., Ramdani, A., Doyan, A., Anwar, Y. A. S., Rokhmat, J., & Sukarso, A. (2023). Development of performance assessment on science practicum integrated with automated feedback to measure scientific attitude in university: A case study in Indonesia. *Path of Science*, *9*(12), 3001–3010. https://doi.org/10.22178/pos.99-7

Sugandini, W., Suyasa, P. W. A., Divayana, D. G. H., & Ariawan, I. P. W. (2025). Rwa Bhineda concept and weighted product method in the Alkin-Provus evaluation model modified

design. *Indonesian Journal of Educational Development (IJED)*, *6*(2), 547–560. https://doi.org/10.59672/ijed.v6i2.4944

Sugiyono. (2013). *Metode penelitian pendekatan pendekatan kuantitatif, kualitatif, dan r&d (Research methods include quantitative, qualitative, and r&d approaches)*. Alfabeta.

Suklani, S. (2023). Evaluation model and its urgency on elementary education programs. *Jurnal Obsesi : Jurnal Pendidikan Anak Usia Dini*, *7*(2), 1639–1650. https://doi.org/10.31004/obsesi.v7i2.4201

Taherdoost, H. (2019). What is the best response scale for survey and questionnaire design; Review of different lengths of rating scale / attitude scale / likert scale. *International Journal of Academic Research in Management (IJARM)*, *8*(1), 1–10. https://ssrn.com/abstract=3588604

Vehovar, V., Toepoel, V., & Steinmetz, S. (2016). Non-probability sampling. In *The SAGE Handbook of Survey Methodology* (pp. 329–345). SAGE Publications Ltd. https://doi.org/10.4135/9781473957893.n22

Widana, I. W., Sumandya, I. W., & Asih, N. P. R. T. (2023). Evaluative study: Literacy outreach program based on local wisdom at SDN 1 Apuan Bangli. *JISAE: Journal of Indonesian Student Assessment and Evaluation*, *9*(1), 26 - 36. https://doi.org/10.21009/jisae.v9i1.32533

Yu, J.-E. (2025). Reliability and validity of applying kirkpatrick model for evaluating exercise rehabilitation program. *Journal of Exercise Rehabilitation*, *21*(4), 200–209. https://doi.org/10.12965/jer.2550428.214

Zareisaroukolaei, M., Shams, G., RezaeiZadeh, M., & Ghahramani, M. (2024). Effectiveness evaluation indicators of organizational e-learning courses. *Computers in Human Behavior Reports*, *15*, 100432. https://doi.org/10.1016/j.chbr.2024.100432

Zomorrodian, A., & Matei, L. (2010). Program evaluation: Its significance and priority for shaping and modification of public policies: A comparative analysis. *Proceedings of ASBBS*, 979–996.