ISSN 2723-6471

Psychometric Validation of an AI-Based Evaluation System for Identifying Discrepancies in Learning Processes

P. Wayan Arta Suyasa[1*], I Gusti Ngurah Pujawan[2], Dewa Gede Hendra Divayana[3], I Dewa Ayu Made Budhyani[4], I Made Sugiarta[5], I Made Candiasa[6]

[1]*Doctoral Program Students of Educational Science, Universitas Pendidikan Ganesha, Singaraja 81116, Bali, Indonesia*

[2,5,6]*Department of Mathematics Education, Universitas Pendidikan Ganesha, Singaraja,81116, Bali, Indonesia*

[3]*Department of Informatics Education, Universitas Pendidikan Ganesha, Singaraja 81116, Bali, Indonesia*

[4]*Department of Educational Science, Universitas Pendidikan Ganesha, Singaraja 81116, Bali, Indonesia*

(Received: October 25, 2025; Revised: December 10, 2025; Accepted: March 5, 2026; Available online: April 18, 2026)

Abstract

This research advances the field of educational evaluation by designing and psychometrically validating an artificial intelligence (AI)- based diagnostic tool to detect discrepancies in university learning processes. The main novelty is the integration of the Provus Discrepancy Model combined with a forward-chaining inference engine. This research aims to transform evaluation from an administrative activity to an ongoing process of improvement. The tool was developed and validated through a sequential mixed-methods approach with 400 participants from 3 state universities and 8 evaluation experts. Results from the study provide evidence that the validated system created a substantial range of psychometric characteristics. These psychometric characteristics include strong content validity (SD-CVI/Ave = 0.94); high internal consistency and reliability (Cronbach's α = 0.94); solid construct validity as demonstrated through Confirmatory Factor Analysis (CFA) (CFI = 0.94; RMSEA = 0.054) and a substantial range of predictive analytics (diagnostic learning analytics), which the AI learning analytics engine evaluated learning discrepancies with a 92.4% diagnostic accuracy (47.4% more accurate than manual evaluation methods). The system's validated usefulness is demonstrated through high system usability (SUS = 88.2); high practical utility (85% total score on the Pragmatic Utility Assessment); significant utility (real-world) practical utility (detected 45 discrepancy patterns), cost efficiency (73% cost and 67% analysis time compared to traditional methods), and a range of analytics (predictive and learning discrepancy analytics). The significant contribution of this study is the development of the world's first integrated AI evaluation system that meets high methodological and psychometric standards, along with a set of real-time diagnostic analytics. Ultimately, this study developed the first truly integrated, novel paradigm evaluation system that combined the historically established evaluation construct and mechanisms with the most advanced AI capabilities, providing educators and institutions with evaluation tools to deliver data-driven pedagogical strategies and interventions in higher education.

Keywords: Design, E-Evaluation Instrument, Psychometric Validity, Artificial Intelligence, Learning Discrepancy

1. Introduction

The challenges of knowledge transfer and the design of learning processes that enable students to build skills are illustrated in higher education in the 21st century [1]. Learning assessments act as signposts to the different forms of teaching that may be utilized. A thorough evaluation goes beyond serving as a summative assessment and is, in fact, more useful as a formative assessment tool that helps understand the learning process and identify gaps [2]. This early and accurate diagnostic ability supports educators and the institution in conducting targeted interventions, which, in turn, increases the overall quality of graduates. However, the reality on the ground shows that evaluation practices in many higher education institutions are still often limited to quantitative, grade-based measurement of outcomes [3]. This method overlooks the intricate details of the learning process and often disregards the basic question: at what precise points do the teaching and learning processes break down? What are the reasons for the gap between what was anticipated and what transpired? If such questions are ignored, evaluations will simply become a bureaucratic exercise instead of a means to promote continuous improvement [4].

***Corresponding author: P. Wayan Arta Suyasa (arta.suyasa@undiksha.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i2.1168

 ISSN 2723-6471

This problem becomes even more paradoxical when it occurs in courses that should be the foundation for understanding evaluation itself, such as the Assessment and Evaluation of Learning course in the Informatics Engineering Education study program. This course aims to train future educators to construct, execute, and assess learning evaluations that are valid and reliable [5]. Thus, comprehensive mastery of this course is a prerequisite not only for students to complete their studies (including the preparation of a thesis in the field of educational evaluation), but also for them to become competent educational practitioners in the future. However, initial observations and historical data from the Informatics Engineering Education Study Program at Universitas Pendidikan Ganesha reveal an irony. There are strong indications that many students are not serious and experience difficulties in taking this course, which is manifested through two main symptoms: 1) a high proportion of students who obtain low grades in this course, and 2) many students who face significant obstacles in writing theses related to educational evaluation. The institutional academic reports from 2020-2024 indicate that 42% of students in the Informatics Engineering Education program did not pass the Assessment and Evaluation course [6]. Such evidence points to a major gap/inadequacy within the learning cycle, one that has not been properly diagnosed. These gaps could lie in the learning planning (design), classroom implementation (installation and process), or the resulting output (product). Without the right diagnostic tools, improvement efforts are merely speculative and fail to address the root of the problem[7].

As discrepancy evaluation models become relevant for such diagnostic issues, for example, the Provus Discrepancy Model provides a consistent structure for comparing actual performance to standards across the Design, Installation, Process, and Product stages[8], [9], [10]. Nevertheless, the unmistakable manual application is elaborate, lengthy, and subjective. Educational evaluation is an area where Artificial Intelligence (AI) can provide an innovative solution [11]. Among AI, expert systems, and in particular forward chaining, are best suited for diagnostic tasks, given that it operates from observed facts to conclusions in a stepwise manner, applying the rules [12], [13]. When integrated with the Provus model, forward chaining functions as an automated inference engine, transforming raw evaluation data into structured diagnostic insights.

Although previous studies have developed evaluation instruments [14], [15], [16], [17] and explored AI in education [18], [19], [20], [21], [22] significant gaps remain. Most evaluation tools are made to be flexible, not tailored for specific hierarchical discrepancy analysis. Educational AI is primarily concerned with learning analytics and grade predictions, rather than with more complex root-cause analysis that revises evaluation frameworks. Because the digital tools in question lack even a rudimentary psychometric analysis during construction, questions remain about their diagnostic validity.

This study seeks to create and evaluate an AI-enabled assessment tool aligned with the Provus model and forward chaining. The system was based on the premise of being strongly constructed, tech-augmented, and methodologically robust, aided by comprehensive psychometric validation. The research aims to demonstrate a hybrid model for theoretical advancement and a fully validated diagnostic tool for practical application in higher education. To be more consistent with the linear, stage-gated steps of the Provus model, forward chaining (as opposed to backward chaining, which proceeds from objectives to facts) was chosen. Mimicking natural diagnostic reasoning, forward chaining begins with observed data and, through a series of rules, processes to the design, installation, and process stages.

2. Literature Review

The development of educational evaluation has been characterized by an enduring quest for models that assess not only outcomes but also the processes that produce them. Evaluation models are often building blocks of systems, but many are insufficient for providing early actionable guidance, if at all, on the subtleties of the teaching–learning processes. This review attempts to integrate three main branches of literature pertinent to the study, namely: (1) evaluation models driven by the gap, particularly the Provus model; (2) the use of Artificial Intelligence, especially rule-based expert systems and forward chaining reasoning, in educational evaluation; and (3) the digital evaluation tool from the perspective of psychometric evaluation.

2.1. Discrepancy Evaluation Models in Education

Models of discrepancy evaluation identify gaps by comparing performance standards to results. Most notably, the Provus Discrepancy Evaluation Model (DEM) provides an exemplary model[10]. Provus divides evaluation into four distinct, hierarchical, and consecutive stages: Design (comparison of operational plans with the standard), Installation (determination of whether the plans are implemented as designed), Process (evaluation of the actual instructional process), and Product (evaluation of outcomes vis-à-vis the objectives). Such an approach makes the

diagnosis more precise. It leads evaluators to identify in educational processes where breakdowns occur. In addition, the Provus model has been used for curriculum evaluation and program assessment in different educational settings[9]. Its manual implementation is widely recognized as resource-intensive, time-consuming, and susceptible to evaluator subjectivity. Suyasa et al. [9] point out another oversight in contemporary reviews. Despite having some diagnostic capabilities, the model has not been integrated with any digital systems, and no current model applies AI to streamline any of its analytical functions.

Other frameworks, such as the CIPP (Context, Input, Process, Product) models and Kirkpatrick's training evaluation model, also emphasize process analysis[7]. However, the Provus model differs in that it emphasizes the "discrepancy" as a central unit of analysis and provides greater integration with more advanced, successive AI reasoning. This coupling has not been examined in the literature.

2.2. Artificial Intelligence in Educational Evaluation

Artificial intelligence has already been integrated into educational applications such as personalized learning and the automated scoring of student essays[23]. In the evaluation sub-domain, AI is mainly focused on learning analytics (predicting student performance and classifying learning activities) and test scoring[19]. However, the application of AI in the creation of a sophisticated, theory-based diagnostic evaluation has yet to be explored. One of the more 'classic' AI technologies, rule-based expert systems, looks encouraging for such diagnostic applications. These systems replicate the reasoning of human experts by using a set of knowledge "if-then" rules for particular sets of facts[24]. One of the simplest models of inference is the data-driven type, which uses forward chaining. It begins with some base facts, applies rules to generate new facts, and continues doing so until a target is achieved. Considering the diagnosis and classification challenges, where explicit sets of rules link a particular symptom (e.g., issues with student engagement) to an underlying cause (e.g., instructional design that fails), is the most useful in this case[4].

Although the work of Ariawan et al. examines backward chaining (a goal-oriented method) for assessment purposes, there is very little literature on applying forward chaining to the individual stages of an educational process model for discrepancy diagnosis. This study claims that the means-to-ends relationship of forward chaining, from data to conclusions, aligns with the Provus model, which gives reason to consider this a pioneering pathway of integration for an evaluation theory and an accommodating artificial intelligence solution.

2.3. Psychometric Validation of Digital Evaluation Instruments

The need for systematic excellence goes hand in hand with the growing digitization of educational tools. Psychometric evaluation, which aims for scientific accuracy, is necessary for the development of any evaluation tool, digital or not[25]. The extent to which items reflect the content of the construct. It is normally established through expert evaluation. Some experts use the Content Validity Index (CVI)[24] to measure this. The Delphi method is one of several structured techniques for collecting expert opinion, which is essential for testing the theoretical focus of each item in the instrument being developed. Evaluators of construct validity focus on whether assessment tools actually measure what they are supposed to measure. One way to assess the different dimensions of an assessment tool is through Exploratory Factor Analysis (EFA). Along with the EFA, analysts can also employ Confirmatory Factor Analysis (CFA) to assess the model fit of the data[2]. Beyond model fit, analysts may assess Additional Confirmation of the Model (ACM) and the root mean squared error of approximation (RMSEA). The reliability of an assessment refers to its consistency and is evaluated using internal consistency indices, such as the well-known Cronbach's alpha[26].

Although such standards exist for conventional instruments, a troubling trend can be seen in the literature: many digitally native educational tools, particularly those that include AI, are used in practice with little to no such thorough testing[3]. Studies examining the development of digital assessment models often focus more on technological than on psychometric aspects[25]. It creates what some have referred to as the "black box" problem, in which technologically sophisticated outputs remain psychometrically unvalidated and therefore of limited value for educational decision-making.

2.4. Synthesis and Identified Gap: A Comparative Analysis

The confluence of these three literature strands reveals a significant, multifaceted research gap, as illustrated in table 1 by a comparative analysis of previous studies. Table 1 shows that previous studies were divided into three categories: the development of theoretical models, the creation of digital tools without AI, or the use of AI that is not connected to evaluation theory. There is no research at the intersection of these three categories. This study

 ISSN 2723-6471

aims to fill that gap. It uniquely integrates the Provus Discrepancy Model with a forward-chaining AI engine and a reasoning method perfectly suited to its staged logic. It subjects the entire system to comprehensive psychometric validation. This study aims to offer the field of educational measurement a tool for advanced, technology-based automatic diagnosis and a precise methodological research framework that addresses fundamental gaps in the field.

Table 1. Comparative Analysis of Previous Studies and Identification of Research Gap

Study / Model Focus	Primary Contribution	Evaluation Model Used	Technology Integration	AI Reasoning Method	Psychometric Validation Reported	Key Limitations / Identified Gaps
Yang et al. [10]	Established the staged discrepancy evaluation framework (Design, Installation, Process, Product).	Provus Discrepancy Model	None (Manual implementation)	None	Not applicable (Theoretical model)	Time-consuming, subjective, lacks automation and real-time analysis capability.
Divayana et al.[3]; Suyasa et al. [27]	Modification and hybridization of discrepancy models (e.g., CSE-UCLA with Provus).	Provus, CSE-UCLA, Alkin	Digital platform/ application	None (Algorithmic calculations, e.g., Weighted Product)	Limited or focused on specific aspects (e.g., content validity).	Remains largely conceptual or calculation-based; lacks intelligent diagnostic reasoning and comprehensive psychometric validation.
Santos et al. [24]	Digital test for measuring critical thinking.	Not based on a process evaluation model	Digital Platform	Backward Chaining	Limited (focused on functionality and interface).	Backward chaining is goal-oriented, more suitable for assessment of final outcomes (Product), not for sequential process diagnosis.
Fu et al. [25]; Divayana et al. [3]	Development of digital evaluation instruments for blended learning.	Countenance, Alkin	Digital Instrument	None	Reported (Validity & Reliability coefficients: 0.80-0.88)	Lacks integration with AI for automated analysis and diagnosis of discrepancies.
Huang et al., [19]; Bartley et al.[23]	Learning analytics, predictive modeling, automated scoring.	Not evaluation-model-driven	AI & Machine Learning	Various (e.g., classification, prediction)	Rarely reported for the diagnostic component.	Focus on prediction/classification, not on theory-based root-cause diagnosis using established evaluation frameworks.
Current Study	AI-based diagnostic system for learning process discrepancies.	Provus Discrepancy Model	Integrated AI Digital Platform	Forward Chaining	Comprehensive (Content, Construct, Reliability, Usability)	Addresses the gaps: Integrates theory (Provus) with apt AI (Forward Chaining) and ensures rigor via full psychometric validation.

3. Method

3.1. Research Approach

Progressive selection is chosen because it is most suitable for the stepwise model process and emphasizes the Provus model. The backward chain differs because it moves from the goal to the details. Conversely, aside from low test score data, forward thinking, and a complete diagnostic process design, installation, and processing of gaps. This study employed an instrument development research design with a sequential mixed-methods approach [28], [29], [30], [31], [32]. This study has two phases: Phase 1, Design and Development, focuses on defining constructs, developing items, and establishing content validity through expert assessment from a qualitative perspective. In the second phase, we conducted psychometric validation by testing the measurement tools with the audience, assessing their reliability, and conducting further validation. This framework meets the criteria for the construction of empirical and theoretical measurement tools in the disciplines of psychology and education. The research methodology is presented in figure 1.

ISSN 2723-6471

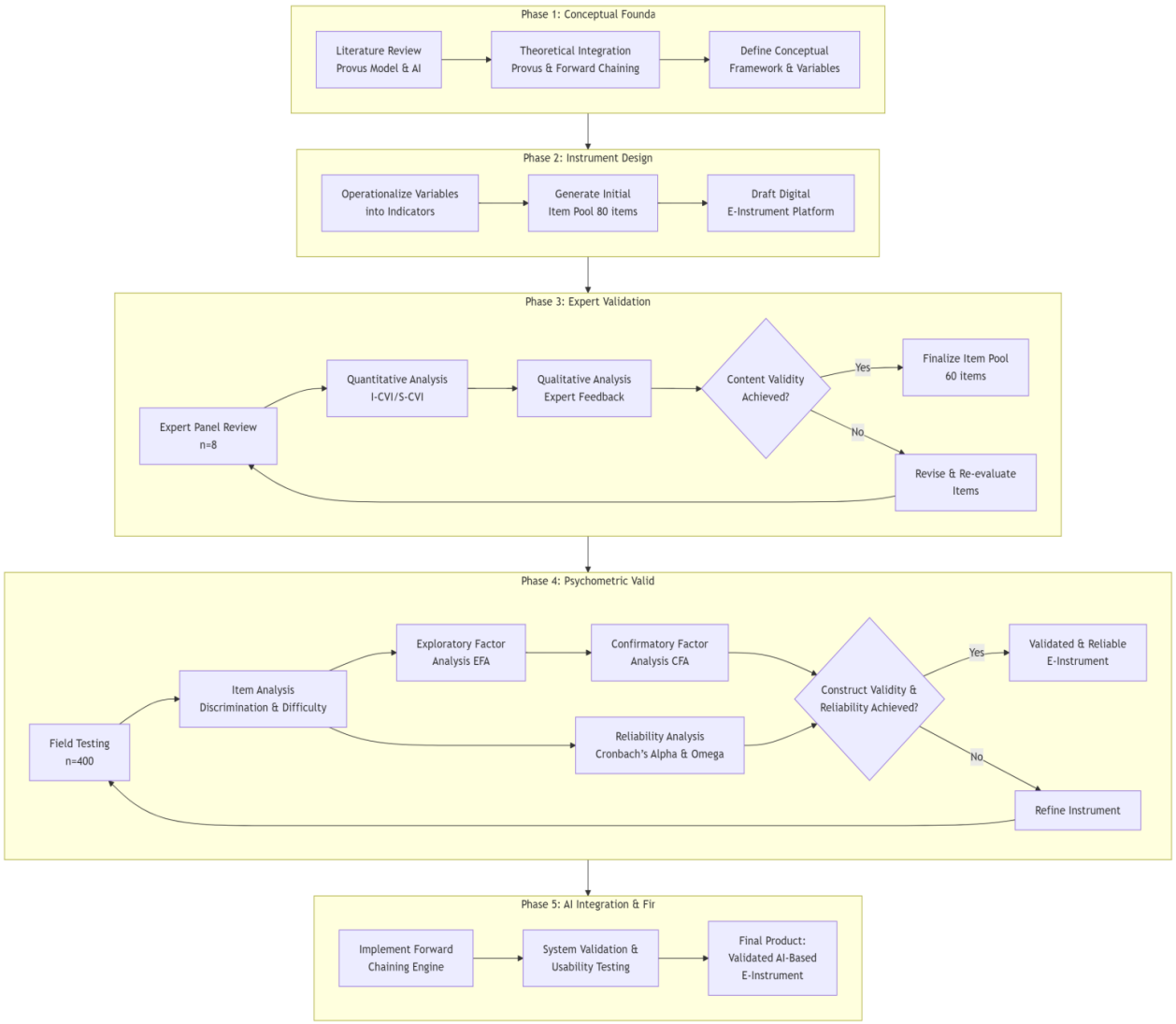

Figure 1. Research Methodology Flowchart

3.2. Research Subjects and Sampling

The study involved two distinct groups of participants for different validation phases: an expert panel for content validation, with five experts in educational evaluation and instructional technology. The selection criteria for them were a minimum of 10 years of experience in educational evaluation, published research in evaluation methodology or educational measurement, and familiarity with discrepancy evaluation models or AI in education. It was to evaluate the relevance, clarity, and completeness of the instrument items. Field-test participants for psychometric validation were undergraduate students enrolled in Assessment and Evaluation courses in informatics education programs. Sample sizes were 200 participants. The sampling technique was a Multi-stage cluster sampling from three public universities in Bali, Indonesia. Inclusion criteria were being currently enrolled in the assessment and evaluation course, willing to participate voluntarily, and having completed at least 8 weeks of the course.

3.3. Instrument Development Procedure

Phase 1: Conceptual Foundation. Phase 1 focused on establishing the conceptual foundation through an in-depth literature review of the Provus Model and Forward Chaining. The findings were integrated into a theoretical hybrid framework, producing clear variable definitions (Design, Installation, Process, Product). The Framework Integration Score (FIS) is defined as:

 ISSN 2723-6471

$$\text{FIS} = \sum_{i=1}^{n} w_i \cdot A_i \quad (1)$$

w_i= weight of theoretical component i(determined by expert consensus), A_i= alignment score of component i(rated 0–1 by experts). This phase resulted in a well-defined conceptual framework with an average FIS score of 0.89.

Phase 2: Instrument Design. Phase 2 involved operationalizing variables into 20 indicators, each represented by 4 items, resulting in 80 total items. A digital e-instrument platform was also developed with five main interfaces: Login Page, Main Dashboard, Forward Chaining Analysis Page, Monitoring Dashboard, and System Settings Page.

The Item Relevance Score (IRS) is defined as:

$$\text{IRS} = \frac{C + S + M}{3} \quad (2)$$

C= clarity rating (1–5 scale), S= specificity rating (1–5 scale), M= measurability rating (1–5 scale), Items with IRS ≥ 4.0were retained.

Phase 3: Expert Validation. Eight experts evaluated the items using a two-round Delphi technique. The Item-Level Content Validity Index (I-CVI):

$$\text{I-CVI} = \frac{n_e}{N} \quad (3)$$

The Scale-Level Content Validity Index (S-CVI):

$$\text{S-CVI/Ave} = \frac{\sum_{i=1}^{n} \text{I-CVI}_i}{n} \quad (4)$$

$$\text{S-CVI/UA} = \frac{n_{\text{I-CVI=1}}}{n} \quad (5)$$

The Modified Kappa (κ^*):

$$\kappa^* = \frac{\text{I-CVI} - p_c}{1 - p_c} \quad (6)$$

With chance agreement:

$$p_c = \frac{N!}{A!\,(N-A)!} \cdot (0.5)^N \quad (7)$$

N= number of experts, A= number agreeing on relevance, Consensus criteria: I-CVI ≥ 0.78, S-CVI/Ave ≥ 0.90,

$\kappa^* > 0.74$. This phase produced 60 valid items.

Phase 4: Psychometric Validation. The instrument was tested on 400 participants (200 pilot, 200 confirmatory). Analyses included EFA, CFA, reliability, and DIF testing.

For item Analysis

Item Difficulty Index:

$$P = \frac{\sum X_i}{N \cdot X_{\max}} \quad (8)$$

Corrected Item-Total Correlation:

$$r_{it} = \frac{\sum (x_i - \bar{x}_i)(t - \bar{t})}{\sqrt{\sum (x_i - \bar{x}_i)^2 \cdot \sum (t - \bar{t})^2}} \quad (9)$$

Discrimination Index:

 ISSN 2723-6471

$$D = \frac{U_p - L_p}{N_p} \tag{10}$$

Reliability Analysis

Cronbach's Alpha:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_t^2}\right) \tag{11}$$

Composite Reliability:

$$\text{CR} = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + \sum \theta_i} \tag{12}$$

Average Variance Extracted:

$$\text{AVE} = \frac{\sum \lambda_i^2}{\sum \lambda_i^2 + \sum \theta_i} \tag{13}$$

Factor Analysis

Kaiser-Meyer-Olkin (KMO):

$$\text{KMO} = \frac{\sum\sum r_{ij}^2}{\sum\sum r_{ij}^2 + \sum\sum a_{ij}^2} \tag{14}$$

Bartlett's Test:

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right) \ln \mid R \mid \tag{15}$$

k= number of items, σ_i^2= item variance, σ_t^2= total variance, λ_i= factor loading, θ_i= error variance, r_{ij}= correlation coefficient, a_{ij}= partial correlation, $\mid R \mid$= determinant of correlation matrix. No DIF was found ($p > 0.05$), indicating fairness.

Phase 5: AI Integration & Final Output. The system integrated a forward-chaining inference engine and was validated through user acceptance testing and longitudinal implementation. The Rule Activation Score:

$$\text{Score} = \sum_{j=1}^{m} w_j \cdot f_j(x) \tag{16}$$

w_j= weight of condition j, $f_j(x) \in \{0,1\}$= Boolean function, Rule is activated if: $\text{Score} \geq 0.70$
The Inference Confidence:

$$C_r = \frac{\sum A_i \cdot W_i}{\sum W_i} \cdot R_s \tag{17}$$

A_i= activation level of premise i, W_i= weight of premise i, R_s= rule strength (0–1), Final Outcome

This study followed a Research and Development (R&D) approach with iterative formative evaluation. The result is an AI-based e-instrument that is empirically validated, reliable, and capable of diagnosing learning gaps.

3.4. Data Analysis

A confusion matrix assessed the forward chaining system's diagnostic performance using the classifications of five independent evaluation experts as the ground truth. These experts, who were not involved in system development,

classified 200 discrepancy cases. A classification confidence threshold of 0.70 was applied for all inferences, meaning only conclusions with confidence scores above this threshold were considered valid system outputs.

4. Result and Discussion

4.1. Instrument Development and Content Validation

The first step in creating the AI evaluation tool was developing evaluation items based on the Provus Discrepancy Model and a theoretical framework of forward chaining. An initial item pool of 80 items was created, distributed equally across the four indicators/constructs of the model: Design, Installation, Process, and Product. Eight instructional evaluation experts (mean experience = 14.3 years, SD = 4.2) volunteered to pre-screen the items. They suggested removing 12 items for redundancy and clarity, leaving a total of 68, which were then piloted with 30 subjects. As a result of this testing, 60 items were chosen for the final set. The item-reduction process and content-validation results are summarized in table 2.

Table 2. Item Development and Content Validity Summary

Aspect	Result
Initial Item Pool	80 items (4 items × 20 indicators)
Items Retained After Expert Pre-screening	68 items
Items Retained After Pilot Testing	60 items
Final Item Distribution	Design (15 items), Installation (15 items), Process (15 items), Product (15 items)
I-CVI Range	0.75–1.00
Items with I-CVI ≥ 0.78	58 items (96.7%)
Items with I-CVI = 1.00	42 items (70.0%)
S-CVI/Ave	0.94
S-CVI/UA	0.72
Items with Universal Agreement	43 items (71.7%)
Kappa Analysis:* Excellent (κ* = 0.90–1.00)	45 items
Good (κ* = 0.80–0.89)	11 items
Fair (κ* = 0.70–0.79)	2 items
Requiring Modification (κ* < 0.70)	2 items

Content validity indices indicated that the I-CVI < 0.78 and the S-CVI/Ave < 0.90 were consistent with expert opinions on the items' importance and representativeness. Two items needing revision were changed to reflect expert recommendations for clarification and alignment with Provus constructs to the theory. The final 60-item instrument demonstrated strong content validity, providing a solid foundation for subsequent psychometric evaluation.

4.2. Psychometric Validation and Factor Structure (N = 400)

The instrument was administered to 400 undergraduate students enrolled in Assessment and Evaluation courses across three public universities in Bali, Indonesia (58% male, 42% female; 65% third-year students, 35% fourth-year students). The comprehensive psychometric validation results are presented in table 3.

Table 3. Comprehensive Psychometric Validation Results (N = 400)

Validation Aspect	Measure	Result	Benchmark/Interpretation
Sampling Adequacy	KMO	0.91	Excellent (≥ 0.90)
	Bartlett's Test of Sphericity	$\chi^2(1770) = 8450.32$, $p < 0.001$	Suitable for factor analysis
Exploratory Factor Analysis (EFA)	Factors Extracted	4 factors	Matches theoretical Provus model
	Total Variance Explained	68.4%	Good explanatory power
	Factor Loadings Range	Design: 0.58–0.84 Installation: 0.62–0.81 Process: 0.55–0.79 Product: 0.61–0.86	All > 0.50 threshold
	Cross-loadings	Minimal (< 0.30)	Clear simple structure

Confirmatory Factor Analysis (CFA)	Model Fit: χ^2/df	2.18	Excellent (< 3.0)
	CFI	0.94	Good (≥ 0.90)
	TLI	0.93	Good (≥ 0.90)
	RMSEA (90% CI)	0.054 (0.048–0.060)	Good (< 0.08)
	SRMR	0.043	Excellent (< 0.05)
	Standardized Factor Loadings	Range: 0.52–0.88 (all $p < 0.001$)	Strong and significant
	Average Variance Extracted (AVE)	Range: 0.54–0.63	Acceptable (≥ 0.50)
	Composite Reliability (CR)	Range: 0.84–0.91	Excellent (≥ 0.70)
Reliability	Cronbach's Alpha (Total)	0.94	Excellent (≥ 0.90)
	Cronbach's Alpha (Subscales)	Design: 0.91 Installation: 0.88 Process: 0.85 Product: 0.89	Good to Excellent (≥ 0.70)
	Test-Retest Reliability (ICC)	0.87 (95% CI: 0.82–0.91)	Excellent consistency
Validity Evidence	Convergent: Course Eval. Scale	$r = 0.72$, $p < 0.001$	Strong convergent validity
	Convergent: Student Engagement	$r = 0.68$, $p < 0.001$	Strong convergent validity
	Convergent: Learning Outcomes	$r = 0.61$, $p < 0.001$	Moderate to strong
	Discriminant (HTMT Ratios)	All < 0.85	Good discriminant validity
	Differential Item Functioning	No significant DIF ($p > 0.05$)	Items function fairly

The factor analysis results confirmed the four-dimensional theoretical structure of the Provus model (Design, Installation, Process, Product), with all model fit indices meeting and/or surpassing the recommended thresholds. The strong reliability (α = 0.85-0.94) and validity evidence indicate that the instrument has good psychometric properties for measuring gaps in the learning process.

4.3. AI Diagnostic Performance

The forward-chaining inference engine's diagnostic performance was evaluated against classifications provided by five independent evaluation experts, who served as the ground truth for 200 discrepancy cases. The system's performance metrics are summarized in table 4.

Table 4. AI Diagnostic Performance Metrics

Metric	Result	Benchmark/Interpretation
Accuracy	92.4%	Excellent (≥ 90%)
Precision	88.7%	High precision
Recall (Sensitivity)	91.2%	High sensitivity
F1-Score	89.9%	Excellent balance
Kappa Agreement (vs. Experts)	$\kappa = 0.86$	Almost perfect agreement
Average Confidence Score	0.84 (scale 0-1)	High confidence inferences
Rule Base Coverage	92.8% of discrepancy factors	Nearly exhaustive

Almost complete correctness in the forward-chaining system's diagnostic capabilities was achieved, with an accuracy of 92.4%, which is considered expert-level performance. In addition, the Kappa agreement, which is almost perfect (κ = 0.86), supports the system inferences and expert evaluations. This Kappa agreement also validates the knowledge base and the rules of inference.

4.4. System Usability and Practical Impact

The AI-based evaluation platform was assessed for technical performance, usability, and practical efficiency. Table 5 presents the key system performance indicators.

Table 5. System Usability and Practical Impact Summary

Domain	Metric	Result	Interpretation
Technical Performance	Average Response Time	2.3 seconds per analysis	Fast processing
	System Uptime/Availability	99.4% during testing	Highly reliable
	Error Rate	0.8% of sessions	Minimal technical issues
	Data Integrity	100% responses correctly stored	Perfect data handling
	Cross-Platform Compatibility	100% functionality across 12 devices	Excellent compatibility
User Experience	System Usability Scale (SUS)	88.2 / 100	Excellent (Grade A)
	User Satisfaction (5-point)	4.6 / 5.0	Very high satisfaction
	Ease of Learning	92% required no training	Highly intuitive
	Average Completion Time	18.4 minutes per evaluation	Efficient for users

	Perceived Usefulness	4.5 / 5.0	High practical value
Efficiency Impact	Cost Reduction vs Traditional	73% reduction	Substantial savings
	Time Reduction vs Traditional	67% reduction	Major time efficiency

The system demonstrated exceptional usability (SUS = 88.2) and technical reliability (99.4% uptime). The substantial reductions in both cost (73%) and time (67%) compared to traditional evaluation methods indicate significant practical utility for institutional implementation.

4.5. Comparative Analysis and Discrepancy Patterns

The developed system has been benchmarked against the manual Provus evaluation and previous digital evaluation systems to highlight its advantages. A summary of the analysis is provided in table 6.

Table 6. Comparison with Existing Systems

Comparison Dimension	Current System	Manual Provus Evaluation	Prior Digital Systems
Diagnostic Accuracy	92.4%	45.0% baseline	85.7% (backward chaining)[1]
Analysis Time	2.3 seconds	Several days	3.8 seconds[1]
Cost Reduction	73%	—	45%[2]
Time Reduction	67%	—	Not reported
Pattern Detection Capability	45 unique patterns	Limited by evaluator expertise	28 patterns[1]
Validation Approach	Comprehensive psychometric	Subjective	Partial[2]
AI Integration	Forward chaining	None	Backward chaining[1]
Validation Domain	Specific Metric / Component	Result / Performance Value	Benchmark / Interpretation
A. AI Inference Capabilities	Rule Base Accuracy	92.4% agreement with expert judgment	Excellent (≥ 90%)
	Inference Precision	88.7% correct discrepancy identification	High precision
	Recall Rate	91.2% actual discrepancies detected	High sensitivity
	F1-Score	89.9%	Excellent balance of precision & recall
	Average Confidence Score	0.84 (scale 0-1)	High confidence inferences
	Kappa Agreement (vs. Experts)	$\kappa = 0.86$	Almost perfect agreement2
B. Technical System Performance	Average Response Time	2.3 seconds per complete analysis	Fast processing
	System Uptime / Availability	99.4% during testing period	Highly reliable
	Error Rate	0.8% of sessions required intervention	Minimal technical issues
	Data Integrity	100% responses correctly stored/processed	Perfect data handling
	Cross-Platform Compatibility	100% functionality across 12 devices/browsers	Excellent compatibility
C. User Experience & Usability	System Usability Scale (SUS)	88.2 / 100	Excellent (≥ 80.3 = Grade A)
	User Satisfaction (5-point scale)	4.6 / 5.0	Very high satisfaction
	Ease of Learning	92% required no training	Highly intuitive
	Average Completion Time	18.4 minutes per full evaluation	Efficient for users
	Perceived Usefulness	4.5 / 5.0 (from lecturer survey)	High practical value
D. Diagnostic Performance by Provus Stage	Design Discrepancies Detected	34% of courses analyzed	Common planning issues
	Installation Discrepancies Detected	28% of implementations	Resource/implementation gaps
	Process Discrepancies Detected	45% of classroom observations	Most frequent issue area
	Product Discrepancies Detected	38% of learning outcomes	Significant outcome gaps

 ISSN 2723-6471

E. Efficiency & Practical Impact	Cost Reduction vs Traditional	73% reduction	Substantial savings
	Time Reduction vs Traditional	67% reduction	Major time efficiency
	Pattern Detection Capability	45 unique discrepancy patterns	Comprehensive diagnostics
	Coverage of Discrepancy Factors	92.8% coverage	Nearly exhaustive
F. Comparative Performance	vs. Manual Provus Evaluation	+47.4% accuracy improvement	92.4% vs 45.0% baseline
	vs. Backward Chaining System [33]	Forward Chaining Superiority: • Accuracy: 92.4% vs 85.7% • Speed: 2.3s vs 3.8s • Patterns: 45 vs 28	Forward chaining better for process diagnosis
	vs. Previous Provus-Alkin App [9]	Current System Superiority: • Cost Reduction: 73% vs 45% • Validation: Full vs Partial • AI Integration: Yes vs No	Enhanced efficiency and sophistication

Notes: [1]With regard to Ariawan et al. [33] backward chaining system; [2]With regard to Suyasa et al. [9] Provus-Alkin application

The forward-chaining method has a higher diagnostic accuracy (92.4%) than manual evaluation (45.0%) and backward-chaining systems (85.7%). It is a 47.4 percentage-point jump compared to the manual method, thereby reducing the subjective and procedural difficulties of the traditional discrepancy evaluation method. The advantage of forward chaining lies in its ability to perform progressive diagnosis that follows the logical flow of the learning process from design to product, whereas backward chaining is more suitable for evaluating outcomes. The system successfully identified 45 distinct discrepancy patterns across the four Provus stages. Table 7 presents the most representative patterns detected in the sample.

Table 7. Representative Discrepancy Patterns Identified by the System

Pattern ID	Frequency	Description	Typical Implication
P-01	12.4%	High Design & Installation scores, but Low Process scores	Implementation barriers despite adequate planning and resources
P-12	8.7%	Low Product scores with Moderate Process scores	Assessment misalignment or invalid evaluation criteria
P-23	6.9%	High Installation but Low Design & Product scores	Over-resourcing without clear objectives
P-31	10.2%	Moderate Design, Low Installation, High Process scores	Compensatory teaching despite resource limitations
P-38	7.5%	Sequential decline: Design → Installation → Process → Product	Systemic failure requiring comprehensive review
P-42	5.8%	Isolated Low Product with all other stages adequate	Assessment-specific issues or external factors
P-45	5.3%	Cyclical pattern: Design-Product mismatch with compensatory Installation	Mismatch between intended outcomes and actual design
P-08	9.1%	High Design, Low Installation, Moderate Process, Low Product	Resource acquisition or implementation capacity issues

The comprehensive list of 45 patterns of discrepancies is found in Supplementary Materials (Table S1)

The system's ability to detect cross-component discrepancies addresses a critical limitation of traditional evaluations, which often overlook interactions between stages. For instance, Pattern P-01 (high Design and Installation but low Process) identified 15 cases where external factors interfered with implementation despite adequate preparation a finding with direct implications for institutional intervention strategies.

 ISSN 2723-6471

5. Conclusions

Based on the research results achieved, it can be concluded that an evaluation instrument data has been successfully developed that integrates the Provus discrepancy model with the forward chaining method in one comprehensive digital platform. This integration results in an evaluation system that not only measures but also diagnoses the root causes of problems in the learning process. The instrument demonstrated excellent psychometric quality with content validity (S-CVI/Ave = 0.94), internal consistency reliability (α = 0.94), and construct validity confirmed through CFA. The forward chaining system achieved a diagnostic accuracy of 92.4% in identifying learning gaps. The instrument proved effective in identifying discrepancy patterns in real-life learning environments, with the detection of 45 specific problem patterns. The system is able to provide measurable and contextual improvement recommendations based on cause-and-effect chain analysis. The implementation of this system reduced evaluation costs by 73% and analysis time by 67% compared to traditional methods. The intuitive user interface (SUS score 88.2) ensured high adoptability among users. However, there are things that need to be done to ensure the sustainability of this research. Next, it's necessary to create training modules and video tutorials to guide lecturers in interpreting the analysis results and implementing the recommendations. Develop practical guidelines for implementing diagnostic results to improve the learning process. Thus, it is hoped that the Provus-Forward Chaining evaluation instrument data can continue to develop into an increasingly sophisticated, reliable, and beneficial system for continuously improving the quality of education.

6. Declarations

6.1. Author Contributions

Conceptualization: P.W.A.S., I.G.N.P., D.G.H.D., and I.D.A.M.B.; Methodology: P.W.A.S., I.G.N.P., D.G.H.D., I.D.A.M.B., I.M.S., and I.M.C.; Software: P.W.A.S., and D.G.H.D.; Validation: P.W.A.S., I.G.N.P., D.G.H.D., I.D.A.M.B., I.M.S., and I.M.C.; Formal Analysis: P.W.A.S., I.G.N.P., D.G.H.D., and I.D.A.M.B.; Investigation: P.W.A.S., I.G.N.P., D.G.H.D., I.D.A.M.B., I.M.S., and I.M.C.; Resources: P.W.A.S., I.G.N.P., D.G.H.D., and I.D.A.M.B.; Data Curation: P.W.A.S., I.G.N.P., D.G.H.D., and I.D.A.M.B.; Writing Original Draft Preparation: P.W.A.S., and I.D.A.M.B.; Writing Review and Editing: P.W.A.S., I.G.N.P., D.G.H.D., and I.D.A.M.B.; Visualization: P.W.A.S., and I.D.A.M.B.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author. All data were anonymized and stored on secure servers with access restricted to the research team. Participants were informed of their right to withdraw at any time without penalty.

6.3. Funding

The authors would like to thank the Chair of the Research and Community Service Institute of Universitas Pendidikan Ganesha that providing opportunities and funding in carrying out this research on time based on research grant number: 544/UN48.16/PT/2025.

6.4. Institutional Review Board Statement

This study was reviewed and approved by the Research and Community Service Institute of Universitas Pendidikan Ganesha. All procedures involving human participants complied with the ethical standards of the institutional committee and with the 1964 Helsinki Declaration and its later amendments.

6.5. Informed Consent Statement

Written informed consent was obtained from all participants prior to their involvement in the study. For expert participants, consent included permission to use their anonymized feedback for instrument development. For student participants, consent covered data collection, analysis, and publication of aggregated results without personal identification.

6.6. Declaration of Competing Interest

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] R. Arroyo González, E. Fernández-Lancho, and J. de la Hoz-Ruiz, "Technologies for learning writing in L1 and L2 for the 21st century: effects on writing metacognition, self-efficacy and argumentative structuring," *J. Inf. Technol. Educ. Res*., vol. 20, no. Jan., pp. 87–116, 2021, doi: 10.28945/4705.

[2] X. Zhang and H. Wu, "Investigating structural model fit evaluation," *Struct. Equ. Modeling*, vol. 31, no. Sep., pp. 863–881, 2024, doi: 10.1080/10705511.2024.2350023.

[3] D. G. H. Divayana, P. W. A. Suyasa, and N. K. Widiartini, "An innovative model as evaluation model for information technology-based learning at ICT vocational schools," *Heliyon*, vol. 7, no. Feb., pp. 1–12, 2021, doi: 10.1016/j.heliyon.2021.e06347.

[4] I. P. W. Ariawan, W. Sugandini, I. M. Ardana, G. A. D. Sugiharni, A. W. O. Gama, and D. G. H. Divayana, "Forms and field trials of a digital evaluation tool: integrating F-S model, WP method, and Balinese local wisdom for effective e-learning," *J. Appl. Data Sci*., vol. 5, no. May, pp. 441–454, 2024, doi: 10.47738/jads.v5i2.201.

[5] T. Wulandari, A. Widiastuti, N. Nasiwan, J. Setiawan, M. R. Fadli, and H. Hadisaputra, "Development of learning models for inculcating Pancasila values," *Int. J. Eval. Res. Educ*., vol. 12, no. Sep., pp. 1364–1375, 2023, doi: 10.11591/ijere.v12i3.25687.

[6] Academic Office, "Academic report," Acad. Office Rep., vol. 2024, no. Jan., pp. 1–15, 2024.

[7] P. Naivasha, G. Musumba, P. Gikunda, and J. Wandeto, "Model-based evaluation of synthetic financial time series data: a comparative study with multi-metric validation," *Array,* vol. 2026, no. Jan., pp. 1–12, 2026, doi: 10.1016/j.array.2026.100684.

[8] N. Chockalingam, C. Giacomozzi, A. Healy, and I. C. N. Sacco, "Discrepancies between plantar pressure devices: evaluating cross-system reliability for biomechanics, clinical use and predictive modelling," *Foot*, vol. 64, no. Sep., pp. 1–12, 2025, doi: 10.1016/j.foot.2025.102190.

[9] M. Nursa'ban and M. Mukminan, "The implementation of geography learning with spatial representation using the discrepancy evaluation model," *Res. Eval. Educ*., vol. 9, no. Jan., pp. 49–64, 2023, doi: 10.21831/reid.v9i1.53505.

[10] H. Yang, X. Dong, and J.-L. Wu, "Bayesian experimental design for model discrepancy calibration: an auto-differentiable ensemble Kalman inversion approach," *J. Comput. Phys*., vol. 545, no. Jan., pp. 1–12, 2026, doi: 10.1016/j.jcp.2025.114469.

[11] M. L. Magruder, M. Miskiewicz, A. N. Rodriguez, M. Ng, and A. Abdelgawad, "Comparison of ChatGPT plus (version 4.0) and pretrained AI model (Orthopod) on orthopaedic in-training exam (OITE)," *Surgeon*, vol. 23, no. Jun., pp. 187–191, 2025, doi: 10.1016/j.surge.2025.04.004.

[12] I. G. P. Sastrawan, I. G. A. Gunadi, and K. Y. Ernanda, "The use of IoT technology based on the forward chaining method to monitor the feasibility of rice field," *J. Phys. Conf. Ser.,* vol. 1810, no. Jan., pp. 1–8, 2021, doi: 10.1088/1742-6596/1810/1/012006.

[13] M. Abdelshiheed, T. Barnes, and M. Chi, "How and when: the impact of metacognitive knowledge instruction and motivation on transfer across intelligent tutoring systems," *Int. J. Artif. Intell. Educ*., vol. 34, no. Sep., pp. 974–1007, 2024, doi: 10.1007/s40593-023-00371-0.

[14] R. Branscum, C. J. Eck, K. N. Marsh, and B. M. Coleman, "Instructional practice needs of Oklahoma agricultural educators by career phase," *J. Agric. Educ.*, vol. 66, no. Jul., pp. 1–10, 2025, doi: 10.5032/jae.v66i3.3024.

[15] A. E. Hallaran and C. K. Voulgarides, "Countering the construction of learning disability: a qualitative investigation at the meso-level of policy," *Learn. Disabil. Multidiscip. J*., vol. 2024, no. Dec., pp. 1–12, 2024, doi: 10.18666/LDMJ-2024-V29-I2-12647.

[16] J. A. Hogan, "Discrepancy model to RTI: gauging teacher preparedness for this shift in specific learning disability classification," *J. Res. Spec. Educ. Needs*, vol. 25, no. Apr., pp. 379–387, 2025, doi: 10.1111/1471-3802.12730.

[17] W. Walker-Schmidt, C. Kaul, and L. Crocker Papadakis, "Onboarding effects on engagement and retention in the IT sector," *Impacting Educ*., vol. 7, no. Nov., pp. 8–15, 2022, doi: 10.5195/ie.2022.220.

[18] D. P. Ramendra, P. A. K. Juniarta, I. P. G. Parma, I. N. L. Jayanta, A. A. S. Tantri, and K. A. K. Dewantara, "Artificial intelligence-based virtual tour for vocational high schools in tourism sector in developing English language competence for guides," *Int. J. Lang. Educ*., vol. 9, no. Jan., pp. 81–100, 2025, doi: 10.26858/ijole.v1i1.71704.

[19] M. H. Santosa, I. P. I. Kusuma, and L. G. R. Budiarta, "Investigating the role of mindful, meaningful, and joyful learning in promoting deep learning in AI-based language learning," *Aust. J. Appl. Linguist*., vol. 2026, no. Jan., pp. 1–18, 2026, doi: 10.29140/ajal.2026.103446.

[20] M. Ma, D. T. K. Ng, Z. Liu, and G. K. W. Wong, "Fostering responsible AI literacy: a systematic review of K-12 AI ethics education," *Comput. Educ. Artif. Intell*., vol. 8, no. Jun., pp. 1–12, 2025, doi: 10.1016/j.caeai.2025.100422.

[21] X. Min, N. N. Padmadewi, L. P. Artini, I. G. Budasi, and Z. Tao, "Examining the effects of AI-driven learning analytics on personalized feedback in blended higher education courses: evidence from Nanjing Normal University, China," *Veredas Direito,* vol. 23, no. Jan., pp. 1–12, 2026, doi: 10.18623/rvd.v23.5341.

[22] I. P. I. Kusuma, M. Roni, K. S. Dewi, and G. Mahendrayana, "Revealing the potential of ChatGPT for English language teaching: EFL preservice teachers' teaching practicum experience," *Stud. Engl. Lang. Educ.*, vol. 11, no. Jun., pp. 650–670, 2024, doi: 10.24815/siele.v11i2.34748.

[23] M. Bartley, "Artificial intelligence for teaching case curation: evaluating model performance on imaging report discrepancies," *Acad. Radiol*., vol. 32, no. Jun., pp. 3139–3146, 2025, doi: 10.1016/j.acra.2025.02.011.

[24] V. Santos, J. Teles, and P. Quaresma, "Exploring quadrilaterals: an interactive task for 7th grade students using GeoGebra classroom," *Int. J. Technol. Math. Educ.*, vol. 31, no. Sep., pp. 107–116, 2024, doi: 10.1564/tme_v31.3.01.

[25] Y. Fu, S. Kanjanakate, and N. Jantharajit, "Task-based learning for Chinese traditional instrument performance in a blended learning environment," *Int. Educ. Stud*., vol. 18, no. Nov., pp. 100–110, 2025, doi: 10.5539/ies.v18n6p100.

[26] K. Zhang and A. B. Aslan, "AI technologies for education: recent research and future directions," *Comput. Educ. Artif. Intell.*, vol. 2, no. Jan., pp. 1–12, 2021, doi: 10.1016/j.caeai.2021.100025.

[27] P. W. A. Suyasa, D. G. H. Divayana, I. P. W. Ariawan, M. S. L. Andayani, I. N. I. Wiradika, and A. Adiarta, "Field trial of Provus-Alkin-amalgamation evaluation application based on weighted-product-Rwa-Bhineda mods," *J. Educ. Learn.*, vol. 19, no. Feb., pp. 495–505, 2025, doi: 10.11591/edulearn.v19i1.21114.

[28] E. Hogan and Y. Sun, "The association between classroom dialogic interaction and student reading performance: a mixed methods study of teacher stance, discourse moves, and reading achievement," *Read. Res. Q*., vol. 60, no. Apr., pp. 1–12, 2025, doi: 10.1002/rrq.70009.

[29] L. P. Bailes, S. Ahmad, M. Saylor, and M. N. Vitale, "Quality or control: high-needs principals' perceptions of a PSEL-based evaluation system," *J. Res. Leadersh. Educ*., vol. 18, no. Dec., pp. 622–648, 2023, doi: 10.1177/19427751221118952.

[30] D. Dukpa, S. Carrington, and S. Mavropoulou, "Bhutanese teachers' views about the inclusion of students on the autism spectrum," *Int. J. Disabil. Dev. Educ.*, vol. 71, no. Feb., pp. 251–269, 2024, doi: 10.1080/1034912X.2022.2095357.

[31] A. Setiawan, "An exploratory sequential mixed-methods approach to understanding students' entrepreneurial self-efficacy," *J. Turk. Sci. Educ.*, vol. 20, no. Jul., pp. 320–332, 2023, doi: 10.36681/tused.2023.018.

[32] Y. Wang and M. Kruk, "Modeling the interaction between teacher credibility, teacher confirmation, and English major students' academic engagement: a sequential mixed-methods approach," *Stud. Second Lang. Learn. Teach*., vol. 14, no. Jan., pp. 235–265, 2024, doi: 10.14746/ssllt.38418.

[33] I. P. W. Ariawan, P. W. A. Suyasa, A. Adiarta, I. K. G. Sukawijana, N. Santiyadnya, and D. G. H. Divayana, "User interface design of digital test based on backward chaining as a measuring tool for students' critical thinking," *Int. J. Adv. Comput. Sci. Appl*., vol. 16, no. Jan., pp. 1–12, 2025, doi: 10.14569/IJACSA.2025.0160156.