Int. j. adv. multidisc. res. stud. 2026; 6(6):342-349

Received: 16-09-2026
Accepted: 26-10-2026

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

Enhancing Economics Teachers' Competence in Identifying Prior Learning Readiness and Developing HOTS-Based Assessment for Differentiated Instruction

[1] Suratno Martodiryo, [2] Baseran Nor, [3] Rizal Iskandar, [4] Alfi Fadhilah Hakim Hrp, [5] Seprius Gantino

[1, 2, 3, 4, 5] Department of Economic Education, Faculty of Teacher Training and Education, Universitas Lambung Mangkurat, Banjarmasin, Indonesia

DOI: https://doi.org/10.62225/2583049X.2026.6.6.7133
Corresponding Author: **Suratno Martodiryo**

Abstract

Assessment literacy constitutes a fundamental professional competency for teachers, essential for designing differentiated instruction and developing evaluation instruments based on Higher-Order Thinking Skills (HOTS). This study rigorously evaluates the impact of a Subject Teacher Forum (MGMP)-based training intervention on enhancing the assessment literacy of economics teachers, employing the Rasch Measurement Model framework. Utilizing a quasi-experimental one-group pretest-posttest design, the research involved 20 senior high school economics teachers in Barito Kuala Regency, Indonesia. Shifts in teacher competency were evaluated using a 20-item instrument, analyzed via stacking calibration to map changes in individual ability, and racking calibration to measure item difficulty invariance. Model fit analysis confirmed that the instrument possessed excellent psychometric properties (Cronbach's Alpha = 0.82). The stacking analysis revealed a substantial increase in mean teacher ability of 1.74 logits, shifting from -0.19 logits at the pretest to 1.55 logits at the posttest. These findings were convergently supported by the racking analysis, which documented a precipitous decline in mean item difficulty of -1.48 logits, with the most pronounced intervention effects concentrated in the domains of cognitive taxonomy and HOTS item design. The results empirically demonstrate that structured technical training is highly effective in bridging the gap between teacher capacity and the cognitive demands inherent in assessment content.

Keywords: Assessment Literacy, Differentiated Instruction, Rasch Measurement Model

Introduction

Assessment literacy constitutes a foundational professional competency that determines a teacher's capacity to design educationally valid assessments, accurately interpret performance data, and translate evaluative evidence into consequential instructional decisions (Popham, 2009; Stiggins, 1991) [22, 24]. As educational reforms globally have elevated the centrality of higher-order thinking skills (HOTS), competency-based learning frameworks, and differentiated instructional approaches, the demands placed upon teachers to demonstrate sophisticated assessment competencies have grown commensurately (DeLuca *et al*., 2016; Willis *et al*., 2013) [9, 27]. Teachers who possess robust assessment literacy are demonstrably better positioned to align assessment tasks with curricular objectives, identify individual learning needs with precision, provide formatively meaningful feedback, and contribute to the overall validity and equity of educational evaluation systems (Xu & Brown, 2016) [30]. Policymakers, curriculum designers, and educational researchers across diverse national contexts have accordingly elevated teacher assessment literacy to a central professional development priority, recognising it as a powerful systemic lever for improving instructional quality and student learning outcomes at scale (Black & Wiliam, 1998; Gotch & French, 2014) [4, 12].

As a professional construct, assessment literacy encompasses multiple interlocking dimensions that collectively define a teacher's capacity to engage meaningfully and critically with the full cycle of educational assessment. Stiggins (1991) [24] provided a seminal foundational conceptualisation, identifying assessment literacy as comprising proficiency in selecting appropriate assessment methods, developing quality instruments, interpreting results defensibly, and communicating assessment information to relevant stakeholders in contextually appropriate ways. Popham (2009) [22] subsequently positioned assessment literacy not as a peripheral pedagogical skill but as a fundamental professional imperative, proposing classroom

measurement knowledge, formative assessment understanding, and score interpretation competency as its three constitutive pillars. These frameworks were progressively extended by subsequent scholars: Willis *et al*. (2013) [27] reconceptualised teacher assessment literacies as pluralistic and socially embedded practices shaped by curriculum mandates, institutional cultures, and evolving professional identities, while Xu and Brown (2016) [30] argued that assessment literacy develops primarily through sustained professional engagement with assessment evidence and collaborative discourse rather than through discrete declarative knowledge acquisition alone. Across these converging theoretical frameworks, competencies in cognitive taxonomy application, diagnostic assessment design, item construction, item quality analysis, and psychometric data interpretation consistently emerge as the core dimensions that professional development programmes must explicitly and systematically target (Anderson & Krathwohl, 2001; Popham, 2011) [1, 23].

An extensive body of international empirical research has generated converging and sobering evidence of persistent deficiencies in teacher assessment literacy across diverse educational systems, career stages, and institutional contexts. Gotch and French's (2014) [12] systematic review of assessment literacy measurement instruments documented substantive gaps between expected and demonstrated assessment knowledge, noting that existing evaluation instruments frequently revealed overconfidence in self-reported competency relative to objectively measured performance benchmarks. Zhang and Burry-Stock (2003) [31] corroborated these findings through a large-scale survey demonstrating that classroom assessment practices were frequently inconsistent with sound psychometric principles, particularly in item quality evaluation, rubric construction, and score interpretation. Volante and Fazio (2007) [26] identified comparable weaknesses among teacher candidates entering the profession, with pronounced deficiencies in cognitive taxonomy application and standardised assessment data interpretation. In their comprehensive international review, DeLuca *et al*. (2016) [9] concluded that assessment literacy gaps are pervasive across professional experience levels, persisting even among experienced practitioners who had completed prior training, underscoring the inadequacy of episodic professional development as a mechanism for building sustainable competency. Looney *et al*. (2018) [19] further demonstrated that teachers' assessment identity — encompassing their professional values, beliefs, and orientations toward assessment purposes — acts as a powerful mediating variable determining whether training translates into genuinely changed classroom practices. Collectively, these studies establish the systemic and multidimensional nature of the assessment literacy challenge: the demand for assessment-literate teachers is intensifying while the professional development mechanisms required to produce and sustain high-level competency remain underdeveloped (Black & Wiliam, 1998; Pastore & Andrade, 2019) [4, 21].

Professional development interventions designed to advance teacher assessment literacy have been implemented and evaluated across a range of national and institutional contexts, with accumulated evidence indicating that structured, content-focused training programmes can produce measurable improvements in assessment knowledge when they embody the key design features associated with effective professional learning (Desimone, 2009; Guskey, 2000) [10, 13]. Darling-Hammond *et al*. (2017) [8] synthesised findings from more than 35 rigorous studies of teacher professional development, identifying sustained duration, active engagement with disciplinary content, coherence with curriculum standards, and collaborative professional structures as the distinguishing features of high-impact programmes. Kennedy (2016) [15], through a comprehensive meta-analytic review, demonstrated that professional development programmes exerting the most enduring influence on teaching quality were those engaging teachers directly with the specific subject-matter knowledge and pedagogical skills targeted for improvement — a principle that directly validates training programmes embedding item construction, cognitive taxonomy application, and psychometric data analysis as core active learning activities. Despite this accumulating evidence, a persistent and critical methodological limitation pervades the training effectiveness evaluation literature: the vast majority of studies assessing training impact rely on gain scores derived from raw percentage comparisons, paired t-tests applied to ordinal-scale data, or self-report satisfaction surveys — approaches that impose fundamental psychometric constraints on the precision, interpretability, and generalisability of evaluative conclusions (Boone *et al*., 2014; Wright & Masters, 1982) [7, 29]. This measurement rigour deficit constitutes a structural impediment to the accumulation of robust, generalisable knowledge about which training designs are most effective, for which teachers, and under what institutional conditions.

In the Indonesian educational context, the urgency of advancing teacher assessment literacy has been substantially intensified by the implementation of the 2013 National Curriculum (Kurikulum 2013) and its successive policy revisions, which mandate the integration of HOTS, authentic assessment, and diagnostic learning approaches across all educational levels and subject disciplines (Kemendikbud, 2013) [14]. Economics education at the senior secondary level represents a particularly demanding case, requiring teachers to design assessments capable of evaluating complex analytical, evaluative, and applied economic reasoning processes — an undertaking presupposing mastery of cognitive taxonomy frameworks, item construction principles, and item quality analysis techniques that constitute the upper tier of assessment literacy competency. The Subject Teacher Forum — known formally in Indonesia as the Musyawarah Guru Mata Pelajaran (MGMP) — constitutes the primary institutionalised mechanism through which subject-specific professional development is delivered at the regional level, convening teachers from a common discipline to engage in collaborative learning, instructional planning, and competency development activities sanctioned by national educational policy (Bjork, 2005; World Bank, 2010) [3, 28]. MGMP-based training programmes are therefore a widely implemented and nationally legitimised professional development modality for advancing teacher competencies, including assessment literacy; yet the evidence base evaluating these programmes through objective, psychometrically sound measurement approaches remains critically underdeveloped, particularly for non-metropolitan regional contexts and subject-specific teacher communities.

A systematic examination of the extant literature reveals three substantive and interconnected research gaps that this

study is specifically positioned to address. First, while international scholarship on teacher assessment literacy is proliferating, investigations employing objective, scale-invariant measurement frameworks — rather than raw-score gain comparisons, ordinal-scale inferential tests, or self-report questionnaires — to evaluate training-induced changes in assessment competency remain markedly scarce, limiting the field's capacity to generate defensible and generalisable conclusions about training effectiveness (DeLuca *et al*., 2016; Gotch & French, 2014) [9, 12]. Second, and representing the study's primary methodological novelty, the application of Rasch measurement techniques — specifically the stacking technique for individual person ability calibration and the racking technique for comparative item difficulty analysis across assessment administrations — to the evaluation of assessment literacy training programmes has not been documented in the published literature, constituting a significant and actionable methodological frontier with substantial potential to advance evaluation rigour in this domain. Third, no published empirical studies have employed objective measurement frameworks to evaluate the effectiveness of MGMP-based assessment literacy training specifically for economics teachers in regional Indonesian educational contexts, leaving a critical evidence gap for policymakers responsible for professional development design, resource allocation, and quality assurance at the sub-national level.

To bridge these gaps with requisite methodological rigour, this study employs the Rasch Measurement Model as its primary analytical framework, incorporating two complementary calibration procedures — the stacking and racking techniques — that enable a comprehensive, diagnostically rich, and individually specific evaluation of training intervention effects. The Rasch model offers fundamental psychometric advantages over conventional evaluation approaches: it transforms ordinal raw scores into interval-level logit measures satisfying the mathematical requirements of fundamental measurement; it produces person ability and item difficulty estimates that are mutually independent of each other's sample-specific distributions, enabling sample-invariant comparisons; and it establishes a unified psychometric scale within which pre- and posttest performances are situated on an identical logit continuum, permitting direct, linear, and interpretively unambiguous comparisons of individual ability trajectories (Bond & Fox, 2015; Boone *et al*., 2014; Wright & Masters, 1982) [5, 7, 29]. The stacking technique concatenates pre- and posttest person ability estimates within a single calibration run using anchored item parameters, generating individual-level logit-scaled ability shift measures with direct substantive interpretability; the complementary racking technique isolates changes in item difficulty across administrations by anchoring person abilities, revealing precisely which content domains exhibited the greatest reduction in cognitive demand following the intervention — diagnostic intelligence that conventional evaluation methodologies are structurally incapable of providing (Linacre, 2021) [18]. Together, these two procedures yield convergent, dual-perspective evidence of intervention impact simultaneously from the learner and from the content, producing a quality of evaluative evidence substantially more precise, diagnostically actionable, and individually informative than any single-perspective approach within the conventional psychometric repertoire.

Motivated by these theoretical imperatives, empirical gaps, and methodological considerations, the present study pursues two interconnected objectives: to develop and psychometrically validate a Rasch-based assessment literacy instrument designed to measure economics teachers' competencies across four domains — diagnostic assessment in differentiated learning, cognitive taxonomy and HOTS item characteristics, item writing principles, and psychometric data interpretation; and to apply Rasch stacking and racking calibration techniques to evaluate rigorously the effectiveness of an MGMP-based assessment literacy training programme delivered to economics teachers in Barito Kuala Regency, South Kalimantan, Indonesia. Through these objectives, the study aims to establish an empirically defensible and scale-invariant evidence base for MGMP training effectiveness, to introduce a replicable Rasch-based methodological template for teacher professional development evaluation in analogous contexts, and to generate diagnostically actionable information to guide the refinement, targeting, and differentiation of future assessment literacy training designs in regional Indonesian educational settings.

Methodology

Research design and setting: This study employed a quantitative approach utilizing a quasi-experimental one-group pretest-posttest design. This design was specifically framed to evaluate the effectiveness of an assessment literacy training intervention on teachers' pedagogical abilities.

Participants and sampling: The participants consisted of 20 educators affiliated with the Economics Subject Teacher Forum (MGMP) in Barito Kuala Regency, South Kalimantan Province, Indonesia. A random sampling technique was utilized to minimize selection bias and ensure an adequate demographic representation within the target population.

Instrument(s) and data collection: Empirical data were collected through the administration of a multiple-choice test instrument specifically constructed to measure the participants' understanding of assessment literacy. The instrument comprised 20 items evaluating key topics taught during the training intervention, namely diagnostic assessment in differentiated learning, cognitive taxonomy and characteristics of Higher-Order Thinking Skills (HOTS) questions, and item analysis techniques using Winsteps. The construction of the instrument was mapped into four main domains representing comprehensive pedagogical competencies, as detailed in Table 1. The test was distributed to all participants prior to the intervention (pretest) and immediately following its completion (posttest) to precisely track the transition in their cognitive understanding.

Table 1: Distribution of Assessment Literacy Test Domains

Competency Domain	Item Numbers	Total Items
D1. Diagnostic assessment and differentiated learning	1–5	5
D2. Cognitive taxonomy and HOTS item characteristics	6–10	5
D3. Item writing principles and item quality	11–15	5
D4. Data interpretation (Rasch/Classical) and follow-up	16–20	5

Data analysis: All quantitative responses were evaluated using the Rasch Measurement Model via Winsteps software version 3.7.3. The initial stage of the analysis focused on item validation by evaluating fit parameters against rigorous acceptance criteria, which included Outfit Mean-Square (MNSQ) values between 0.5 and 1.5, Outfit Z-Standardized (ZSTD) values ranging from -2.0 to +2.0, and Point-Measure Correlation (Pt-Meas Corr) values within the 0.4 to 0.85 threshold. Measurement integrity was confirmed by testing the unidimensionality assumption, where the minimum threshold for raw variance explained by measures was set at 20%. Further psychometric quality was estimated by reporting person reliability, item reliability, and Cronbach's Alpha indices. To isolate and evaluate the intervention's impact within an equivalent metric framework, a comparative analysis between the pretest and posttest data was executed utilizing the stacking technique for individual ability (person) calibration and the racking technique for item difficulty calibration.

Results and Discussions

Assumption testing and model fit: Evaluation of the instrument's psychometric properties was conducted as a prerequisite prior to testing the intervention's effectiveness. The model fit analysis indicated that the test instrument exhibited excellent performance, characterized by the mean Infit and Outfit Mean-Square (MNSQ) values for both item and person falling within the ideal range (0.99 to 1.04), well within the strict acceptance threshold of 0.5 to 1.5. The unidimensionality assumption test revealed that the raw variance explained by the measure reached 27.6%. The instrument's reliability was confirmed through a Cronbach's Alpha index of 0.82, supported by a person reliability of 0.77 and an item reliability of 0.74, as well as adequate separation indices to distinguish ability strata (person separation = 1.81) and difficulty strata (item separation = 1.65). The summary of these instrument feasibility statistics is presented in Table 2.

Table 2: Summary of Assessment Literacy Instrument Psychometric Properties

Psychometric Parameter	Measurement Value
Item Infit MNSQ (Mean)	0.99
Item Outfit MNSQ (Mean)	1.04
Person Infit MNSQ (Mean)	0.99
Person Outfit MNSQ (Mean)	1.04
Item Separation	1.65
Person Separation	1.81
Raw variance explained by measure	27.6%
Item Reliability	0.74
Person Reliability	0.77
Cronbach's Alpha	0.82

Note. MNSQ = mean-square.

The results demonstrated that the assessment literacy test instrument developed in this study met the validity and reliability criteria based on Rasch parameters, confirming its suitability as an objective measurement tool for evaluating teachers' assessment literacy competencies. The mean Infit and Outfit Mean-Square (MNSQ) values for both items and persons ranged from 0.99 to 1.04 — values proximally close to the ideal criterion of 1.00 — indicating that each item consistently measured the same underlying construct without significant misfit (Bond & Fox, 2015; Boone *et al*., 2014) [5, 7]. These fit statistics reflect item-person interaction that is well-calibrated across the full ability continuum, a finding consistent with Rasch-based evaluations of teacher competency instruments in comparable professional development contexts (Boone, 2016; Fisher, 2007) [6, 11]. Regarding reliability, the Cronbach's Alpha of 0.82 surpassed the minimum threshold of 0.70 widely accepted in the measurement literature (Taber, 2018) [25], while person reliability (0.77) and item reliability (0.74) both exceeded the 0.67 threshold recommended by Fisher (2007) [11], collectively confirming that the instrument reliably differentiates participants across multiple ability strata. It is noteworthy that the person separation index (1.81) and item separation index (1.65) fell marginally below the conventionally preferred threshold of 2.0; however, this is contextually attributable to the study's restricted sample size (n = 20), as the statistical power to generate wider separation indices is inherently constrained by limited respondent variability in small-sample settings — a limitation consistently documented in training evaluation studies operating under similarly restricted sample sizes (Boone *et al*., 2014; Wright & Masters, 1982) [7, 29].

Stacking analysis: Individual ability calibration: The impact of the training intervention on participants' assessment literacy was evaluated utilizing the stacking technique to linearly compare teachers' abilities between the pretest and posttest phases. The person logit analysis demonstrated a widely distributed increase in ability among participants, with the mean ability substantially improving from -0.19 logits at the pretest to 1.55 logits at the posttest, resulting in a mean difference of 1.74 logits. Specifically, Teacher 13 recorded the highest ability surge with an increase of 5.43 logits. Minor anomalies were detected in two participants (Teacher 2 and Teacher 6) who experienced a slight decrease in ability by -0.23 and -0.38 logits, respectively, and Teacher 4, whose ability remained constant with no logit change. This shift in the ability profile is spatially visualized in Fig 1 and detailed in Table 3.

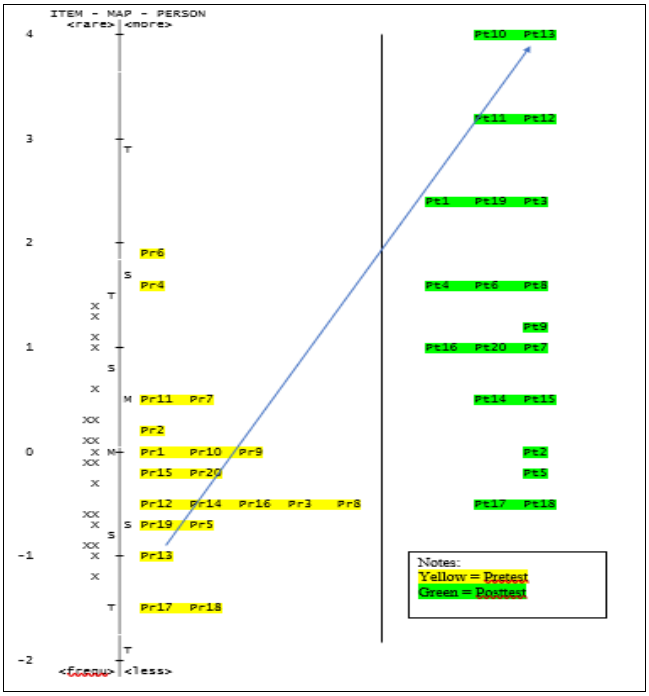


Fig 1: Wright Map of Individual Ability Shifts (Pretest vs. Posttest)

Table 3: Calibration of Individual Ability (Person Measure) in Logits

Participant	Pretest	Posttest	Difference
1	-0.01	2.43	2.44
2	0.22	-0.01	-0.23
3	-0.47	2.43	2.90
4	1.55	1.55	0.00
5	-0.71	-0.24	0.47
6	1.93	1.55	-0.38
7	0.45	0.95	0.50
8	-0.47	1.55	2.02
9	-0.01	1.23	1.24
10	-0.01	4.47	4.48
11	0.45	3.21	2.76
12	-0.47	3.21	3.68
13	-0.96	4.47	5.43
14	-0.47	0.45	0.92
15	-0.24	0.45	0.69
16	-0.47	0.95	1.42
17	-1.55	-0.47	1.08
18	-1.55	-0.47	1.08
19	-0.71	2.43	3.14
20	-0.24	0.95	1.19
Mean	**-0.19**	**1.55**	**1.74**

Note. Values are person measures in logits. Difference = posttest − pretest.

The stacking analysis yielded compelling evidence of the training intervention's effectiveness: the mean person logit increased from -0.19 logits at the pretest to 1.55 logits at the posttest, representing a mean gain of 1.74 logits — a trajectory that repositioned the average participant from a marginally below-average ability level to substantially above the mean item difficulty threshold. In the logic of the Rasch model, where 0.0 logits represents the average item difficulty, this upward shift signifies that the average teacher's ability following the intervention considerably exceeded the cognitive demands embedded in the assessment literacy content — a transformation that is both statistically meaningful and substantively significant (Bond & Fox, 2015; Linacre, 2021) [5, 18]. This finding aligns with the broader professional development literature, which consistently documents measurable gains in teacher knowledge and skills following structured, content-focused training interventions (Desimone, 2009; Guskey, 2000) [10, 13]. The stacking technique offers a distinct methodological advantage over conventional gain-score approaches or paired comparison tests, as it calibrates pre- and posttest person abilities onto a shared linear logit scale, rendering the comparison invariant of item order and sample composition and thereby producing a more rigorous and defensible estimate of intervention impact (Linacre, 2021; Wright & Masters, 1982) [18, 29]. Furthermore, the magnitude of the observed ability gain reinforces the substantive value of MGMP-based training as a professional development modality, contributing empirically to a growing body of literature recognising subject-specific teacher forums as effective platforms for advancing pedagogical competencies, including assessment literacy (DeLuca *et al*., 2016; Xu & Brown, 2016) [9, 30].

While the overall trajectory was markedly positive, the individual ability profiles revealed substantial heterogeneity in training responsiveness — a pattern that merits careful interpretation. Participants who entered the training with lower initial ability levels demonstrated the most pronounced gains: Teacher 13, who recorded the lowest pretest logit (-0.96), achieved the largest individual ability surge (+5.43 logits), a pattern consistent with the andragogical principle that adult learners entering an instructional experience with a wider knowledge deficit possess proportionally greater capacity for measurable gain from structured, needs-aligned instruction (Knowles *et al*., 2015; Merriam, 2001) [17, 20]. Conversely, the marginal decrements observed in Teacher 2 (-0.23 logits) and Teacher 6 (-0.38 logits) are interpretable through the lens of statistical regression to the mean — a well-documented phenomenon in pretest-posttest designs whereby participants positioned at the higher end of the initial ability distribution tend to produce slightly attenuated posttest scores due to natural measurement variability rather than genuine knowledge loss (Barnett *et al*., 2005) [2]. Teacher 6, who entered with the highest pretest logit (1.93) and was therefore already positioned well above the mean item difficulty at baseline, had a markedly narrowed bandwidth for detectable upward movement; this instrument-level ceiling effect, rather than stagnation in professional learning, plausibly accounts for the observed decrement. Teacher 4's unchanged logit value (1.55 at both time points) further supports this interpretation, suggesting that the instrument's difficulty range may have been insufficient to capture potential growth at the upper end of the ability continuum.

Racking analysis: Item difficulty shifts: Complementing the individual calibration, an evaluation of item difficulty invariance (racking) was conducted to confirm the extent to which the intervention reduced the material's difficulty level for the teachers. Comprehensive item mapping proved a drastic reduction in the difficulty level across almost all instrument items following the training provision. The mean item difficulty plummeted from 0.74 logits during the pretest to -0.74 logits during the posttest, marking an aggregate facility shift of -1.48 logits. Test item number 2 was the only question unaffected by the intervention, maintaining its absolute difficulty value at -1.02 logits in both the pretest and posttest phases. This downward shift in the item difficulty matrix is visually confirmed in Fig 2 and fully documented in Table 4.

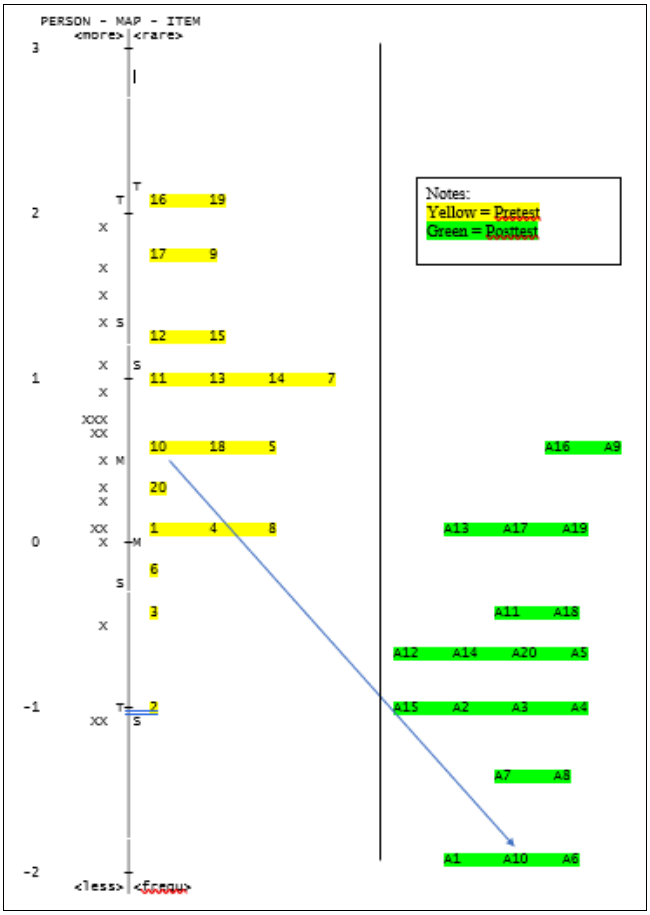


Fig 2: Wright Map of Item Difficulty Mapping (Pretest vs. Posttest)

Table 4: Calibration of Instrument Item Difficulty (Item Measure) in Logits

Instrument Item	Pretest	Posttest	Difference
1	0.09	-1.90	-1.99
2	-1.02	-1.02	0.00
3	-0.41	-1.02	-0.61
4	0.09	-1.02	-1.11
5	0.55	-0.70	-1.25
6	-0.16	-1.90	-1.74
7	1.00	-1.40	-2.40
8	0.09	-1.40	-1.49
9	1.77	0.55	-1.22
10	0.55	-1.90	-2.45
11	1.00	-0.41	-1.41
12	1.24	-0.70	-1.94
13	1.00	0.09	-0.91
14	1.00	-0.70	-1.70
15	1.24	-1.02	-2.26
16	2.08	0.55	-1.53
17	1.77	0.09	-1.68
18	0.55	-0.41	-0.96
19	2.08	0.09	-1.99
20	0.32	-0.70	-1.02
Mean	**0.74**	**-0.74**	**-1.48**

Note. Values are item measures in logits. Difference = posttest − pretest.

The racking analysis corroborated the stacking findings from a complementary analytical perspective: the mean item difficulty declined from 0.74 logits at the pretest to -0.74 logits at the posttest, marking an aggregate downward shift of -1.48 logits and confirming that the training systematically and substantially reduced the cognitive barriers associated with the assessment literacy content. In the Rasch framework, a simultaneous reduction in perceived item difficulty and an increase in person ability constitutes convergent evidence of learning: the intervention successfully narrowed the gap between participant capability and material demand, producing a mutually reinforcing narrative of knowledge acquisition from two independent analytical vantage points (Bond & Fox, 2015; Wright & Masters, 1982) [5, 29]. Items 10 (-2.45 logits), 7 (-2.40 logits), and 15 (-2.26 logits) exhibited the largest difficulty reductions; cross-referencing these with the instrument's domain mapping reveals that these items predominantly represent Domains D2 and D3 — cognitive taxonomy, HOTS characteristics, and item writing principles — which received the most intensive instructional treatment during the intervention, suggesting a close and meaningful alignment between pedagogical emphasis and measurable learning outcomes (Guskey, 2000) [13]. The sole exception was Item 2, which maintained an invariant difficulty of -1.02 logits across both administrations, indicating that the knowledge it assessed was likely already possessed by participants prior to training or fell outside the direct instructional scope of the intervention — a diagnostic signal of considerable value for refining the training curriculum in subsequent iterations. The convergence of stacking and racking evidence collectively positions the Rasch-based pre-post evaluation framework as a methodologically superior alternative to conventional training assessment approaches, simultaneously capturing growth from the learner perspective and from the content perspective to yield a richer and more diagnostically informative portrait of intervention impact (Linacre, 2021) [18].

Contributions and implications: This study contributes to the literature on teacher professional development evaluation by demonstrating the utility of Rasch-based stacking and racking techniques as precise, scale-invariant tools for quantifying the differential impact of training interventions on teacher ability and content difficulty. From a methodological standpoint, the application of these two complementary calibration strategies in the context of teacher assessment literacy training represents a relatively novel approach within the Indonesian educational research landscape, where training effectiveness has historically been evaluated through conventional pre-post score comparisons or qualitative feedback instruments that lack the psychometric rigour necessary to produce defensible, generalisable conclusions (DeLuca *et al.*, 2016; Gotch & French, 2014) [9, 12]. Substantively, this study provides empirically grounded, Rasch-scaled evidence of the MGMP forum's effectiveness as a structured vehicle for developing assessment literacy competencies among economics teachers — addressing a gap that existing literature on Indonesian teacher professional development has not adequately addressed through objective measurement frameworks. Beyond its immediate contextual contribution, the methodological template established here — comprising instrument development, stacking-racking calibration, and individual ability profiling — is directly replicable by researchers and programme evaluators assessing the impact of teacher training initiatives across diverse subject disciplines and educational levels.

The empirical evidence generated by this study carries direct and actionable implications for multiple educational

stakeholders responsible for designing, implementing, and evaluating teacher professional development programmes. For MGMP coordinators and training designers, the domain-level mapping of item difficulty shifts reveals that materials related to cognitive taxonomy, HOTS item design, and item analysis techniques were particularly responsive to the intervention, affirming that these topics should be retained as instructional priorities in future training cycles, while the pedagogical design of diagnostic assessment content may benefit from further refinement to maximise instructional impact. Educational supervisors and district authorities should consider institutionalising Rasch-based pre-post evaluation frameworks — incorporating stacking and racking analyses — as a standard assessment practice for teacher development programmes, given that this approach provides objective, data-driven evidence that substantially transcends the limitations of self-report satisfaction surveys (Kirkpatrick & Kirkpatrick, 2006) [16]. The observed heterogeneity in individual learning gains further underscores the importance of differentiated professional development pathways: teachers with higher entry-level competencies would likely derive greater benefit from advanced training modules presenting more cognitively challenging content, while the current training format demonstrates clear effectiveness for participants entering with lower baseline assessment literacy (Desimone, 2009; Knowles *et al*., 2015) [10, 17]. These implications collectively reinforce the Indonesian Ministry of Education's mandate for Continuing Professional Development, affirming MGMP-based training as an evidence-based, data-validated mechanism for fulfilling national teacher competency standards in the domain of assessment literacy.

Conclusion

This study establishes that the Subject Teacher Forum (MGMP)-based training program significantly enhanced the assessment literacy competencies of economics teachers, particularly in identifying students' prior learning readiness and designing Higher Order Thinking Skills (HOTS)-based evaluation instruments to support differentiated instruction. A comprehensive evaluation employing the Rasch Measurement Model via the stacking technique confirmed a substantial improvement in mean teacher ability by 1.74 logits, which is convergently supported by the racking technique demonstrating a mean reduction in test material difficulty by -1.48 logits. These findings empirically affirm that structured training interventions are highly effective in bridging the gap between teacher capacity and the cognitive demands of assessment literacy content, with the highest degree of efficacy observed among educators possessing lower baseline abilities.

Methodologically, this research offers a significant contribution by demonstrating the superiority of applying stacking and racking calibrations as invariant, linear-scaled measurement tools, thereby overcoming the limitations of conventional training evaluations that frequently rely on raw score comparisons or self-reported satisfaction surveys. The practical implications of this study recommend that educational authorities and MGMP administrators institutionalize this Rasch-based evaluation framework as an operational standard for assessing the impact of continuing professional development. Furthermore, the observed heterogeneity in individual ability improvements underscores the urgency of implementing differentiated training pathways in future iterations; teachers with high baseline assessment literacy require more complex instructional materials, whereas the current training format should be maintained for educators at foundational cognitive stages.

The present study is subject to certain limitations, primarily concerning the relatively small sample size (n = 20), which inherently constrains the magnitude of separation indices within the statistical modeling. Consequently, future research agendas should involve larger participant cohorts across diverse geographical regions and academic disciplines.

References

1. Anderson LW, Krathwohl DR. (Eds.). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman, 2001.
2. Barnett AG, Van Der Pols JC, Dobson AJ. Regression to the mean: What it is and how to deal with it. International Journal of Epidemiology. 2005; 34(1):215-220. Doi: https://doi.org/10.1093/ije/dyh299
3. Bjork C. Indonesian education: Teachers, schools, and central bureaucracy. Routledge, 2005. Doi: https://doi.org/10.4324/9780203957967
4. Black P, Wiliam D. Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice. 1998; 5(1):7-74. Doi: https://doi.org/10.1080/0969595980050102
5. Bond TG, Fox CM. Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.). Routledge, 2015. Doi: https://doi.org/10.4324/9781315814698
6. Boone WJ. Rasch analysis for instrument development: Why, when, and how? CBE-Life Sciences Education. 2016; 15(4), Article rm4. Doi: https://doi.org/10.1187/cbe.16-04-0148
7. Boone WJ, Staver JR, Yale MS. Rasch analysis in the human sciences. Springer, 2014. Doi: https://doi.org/10.1007/978-94-007-6857-4
8. Darling-Hammond L, Hyler ME, Gardner M. Effective teacher professional development. Learning Policy Institute, 2017. https://learningpolicyinstitute.org/product/effective-teacher-professional-development-report
9. DeLuca C, LaPointe-McEwan D, Luhanga U. Teacher assessment literacy: A review of international standards and measures. Educational Assessment, Evaluation and Accountability. 2016; 28(3):251-272. Doi: https://doi.org/10.1007/s11092-015-9233-6
10. Desimone LM. Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. Educational Researcher. 2009; 38(3):181-199. Doi: https://doi.org/10.3102/0013189X08331140
11. Fisher WP. Rating scale instrument quality criteria. Rasch Measurement Transactions. 2007; 21(1):1095. http://www.rasch.org/rmt/rmt211m.htm
12. Gotch CM, French BF. A systematic review of assessment literacy measures. Educational Measurement: Issues and Practice. 2014; 33(2):14-18. Doi: https://doi.org/10.1111/emip.12030
13. Guskey TR. Evaluating professional development. Corwin Press, 2000.

14. Kemendikbud. Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 81A Tahun 2013 tentang Implementasi Kurikulum [Regulation of the Minister of Education and Culture Number 81A of 2013 on Curriculum Implementation]. Kementerian Pendidikan dan Kebudayaan, 2013. https://jdih.kemdikbud.go.id/
15. Kennedy MM. How does professional development improve teaching? Review of Educational Research. 2016; 86(4):945-980. Doi: https://doi.org/10.3102/0034654315626800
16. Kirkpatrick DL, Kirkpatrick JD. Evaluating training programs: The four levels (3rd ed.). Berrett-Koehler Publishers, 2006.
17. Knowles MS, Holton EF, Swanson RA. The adult learner: The definitive classic in adult education and human resource development (8th ed.). Routledge, 2015. Doi: https://doi.org/10.4324/9781315816951
18. Linacre JM. Winsteps® Rasch measurement computer program user's guide, 2021. Winsteps.com. https://www.winsteps.com/winman/
19. Looney A, Cumming J, Van Der Kleij F, Harris K. Reconceptualising the role of teachers as assessors: Teacher assessment identity. Assessment in Education: Principles, Policy & Practice. 2018; 25(5):442-467. Doi: https://doi.org/10.1080/0969594X.2016.1268090
20. Merriam SB. Andragogy and self-directed learning: Pillars of adult learning theory. New Directions for Adult and Continuing Education. 2001; 89:3-14. Doi: https://doi.org/10.1002/ace.5
21. Pastore S, Andrade HL. Teacher assessment literacy: A three-dimensional model. Teaching and Teacher Education. 2019; 84:128-138. Doi: https://doi.org/10.1016/j.tate.2019.05.003
22. Popham WJ. Assessment literacy for teachers: Faddish or fundamental? Theory Into Practice. 2009; 48(1):4-11. Doi: https://doi.org/10.1080/00405840802577536
23. Popham WJ. Classroom assessment: What teachers need to know (6th ed.). Pearson Education, 2011.
24. Stiggins RJ. Assessment literacy. Phi Delta Kappan. 1991; 72(7):534-539. https://www.jstor.org/stable/20404630
25. Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in Science Education. 2018; 48(6):1273-1296. Doi: https://doi.org/10.1007/s11165-016-9602-2
26. Volante L, Fazio X. Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional practice. Canadian Journal of Education. 2007; 30(3):749-770. Doi: https://doi.org/10.2307/20466661
27. Willis J, Adie L, Klenowski V. Conceptualising teachers' assessment literacies in an era of curriculum and assessment reform. The Australian Educational Researcher. 2013; 40(2):241-256. Doi: https://doi.org/10.1007/s13384-013-0089-9
28. World Bank. Transforming Indonesia's teaching force: From pre-service training to classroom practice, 2010. https://openknowledge.worldbank.org/handle/10986/2871
29. Wright BD, Masters GN. Rating scale analysis. MESA Press, 1982.
30. Xu Y, Brown GTL. Teacher assessment literacy in practice: A reconceptualization. Teaching and Teacher Education. 2016; 58:149-162. Doi: https://doi.org/10.1016/j.tate.2016.05.020
31. Zhang Z, Burry-Stock JA. Classroom assessment practices and teachers' self-perceived assessment skills. Applied Measurement in Education. 2003; 16(4):323-342. Doi: https://doi.org/10.1207/S15324818AME1604_4