SWST – International Society of Wood Science and Technology

Special Issue: Scaling up Innovation for Sustainable Value Creation

BIOPRODUCTS BUSINESS

Multi-rater assessment in systematic reviews: A methodological innovation in forest higher education literature

Pipiet Larasatie[1]*, Bambang Sumintono[2], Sandra Rodríguez-Piñeros[3], Rene Zamora-Cristales[4], Annie Biju[5], Kamana Chamlagain[6]

Abstract: Systematic literature reviews (SLRs) are essential for synthesizing evidence in forest higher education, yet the reliability of article selection often hinges on subjective expert judgment. As forest education evolves to meet global challenges, such as climate change, digitalization, and market dynamics, educators must navigate an overwhelming volume of literature to identify high-quality science that fosters critical thinking and holistic understanding. This study introduces the Many Facet Rasch Model (MFRM) as a methodological innovation for evaluating multi-rater assessments of the forest higher education literature, offering a transparent and replicable framework for evidence synthesis. Five experts, who served as raters, assessed ten articles using six calibrated criteria (originality, comprehensiveness of literature review, methodology, scientific value of findings, related issues with forest higher education, and quality of analysis). The results demonstrate that MFRM can identify psychometrically sound evaluations, rank article quality, and diagnose criteria, particularly in comprehensiveness of the literature review and difficulty, most notably in literature review comprehensiveness and methodological rigor. This research provides practical guidance for forest higher education practitioners seeking to select pedagogically valuable resources. By enhancing transparency and reproducibility in literature evaluation, MFRM strengthens forest higher education's capacity to train future foresters with precision, integrity, and relevance.

Keywords: forest education, systematic literature reviews, Rasch model, evidence synthesis, psychometric analysis.

1. Introduction

Systematic literature reviews (SLRs) have become a cornerstone of evidence-based research, offering a reproducible and transparent methodology for synthesizing large bodies of scholarly work (Tranfield et al. 2003). In contrast to traditional narrative reviews, which often rely on implicit selection criteria and subjective synthesis, SLRs aim to minimize bias by applying structured protocols for article identification, screening, and evaluation. A well-conducted systematic review performs thorough literature searches of published studies to summarize extensive bodies of evidence and detect gaps that may exist among different studies (Cook et al. 1997b). This approach has gained prominence across disciplines, particularly in fields where fragmented evidence and methodological heterogeneity complicate decision-making.

The evidence-based movement originated in medical science, where researchers faced increasing difficulty in interpreting frequently contradictory findings (Ohlsson 1994). In the late 1980s, scholars began

1. Department of Sustainable Biomaterials, College of Natural Resources and Environment, Virginia Tech, 1650 Research Center Dr., Blacksburg, VA, USA. Email: pipiet@vt.edu. ORCID: https://orcid.org/0000-0001-5861-7618
2. Faculty of Education, Universitas Islam Internasional Indonesia, Jalan Raya Bogor Km 33,5, Cisalak, Sukmajaya, Depok, West Java, Indonesia. Email: bambang.sumintono@uiii.ac.id. ORCID: https://orcid.org/0000-0002-5490-3665
3. Facultad de Zootecnia y Ecología, Universidad Autónoma de Chihuahua Periferico Francisco R. Almada K. 1 Chihuahua, Chih. México C.P. Email: spineros@uach.mx ORCID: https://orcid.org/0000-0001-6698-6894
4. Department of Forest Engineering, Resources, and Management, College of Forestry Oregon State University, Corvallis, OR, USA. Email: rene.zamora@oregonstate.edu. ORCID: https://orcid.org/0000-0002-6090-6678
5. SankalpTaru Foundation, Bima Vihar, Dehradun, Uttarakhand, India. Email: anniebiju124@gmail.com. ORCID: https://orcid.org/0009-0009-4595-4100
6. School of Agriculture & Forestry, College of Applied & Natural Sciences, Louisiana Tech University, 1501 Reese Dr., Ruston, LA, USA. Email: kch062@email.latech.edu

* Corresponding author

BioProducts Business 11(3), 2026, pp. 35–46. ISSN 2378-1394
https://doi.org/10.22382/bpb-2026-003

to highlight the lack of rigor in secondary research, noting that evaluations of existing literature often relied on implicit and idiosyncratic approaches to data gathering and analysis (Mulrow 1987). These limitations occasionally led to misguided recommendations based on inadequate assessments. As a response, systematic reviews emerged as a methodological innovation to improve transparency, reproducibility, and decision-making quality. However, the reliability of SLRs depends not only on the comprehensiveness of the search strategy but also on the rigor of the article selection process (Cook et al. 1997a,b).

One effort to systematically summarize forest-related higher education was conducted by Larasatie et al. (2024), who identified 7,772 potential articles from the Web of Science and 16,270 from Scopus, resulting in a combined pool of 23,480 unique records, after removing duplicates. This massive number has potential risks in terms of how the inclusion and exclusion of articles were carried out. In such high-volume review scenarios, the risk of inconsistent article selection is amplified, particularly when expert judgment is used to determine inclusion.

To address the challenge, the present study applies the Many Facet Rasch Model (MFRM) as a methodological innovation to address these limitations, enhancing analytic rigor in selecting articles for the systematic review database. Specifically, this study addresses two primary research questions: (1) Can the MFRM improve the reliability and consistency of multi-rater evaluations during the article-selection stage of a SLR? (2) How do MFRM outputs, such as rater fit statistics, criteria difficulty, and rater's judgement, inform more transparent and defensible inclusion decisions in forest education research?

The MFRM has gained popularity for its ability to enhance control over the evaluation and test procedures (McNamara and Knoch 2012). In multi-rater review scenarios, where experts serve as raters to evaluate the relevance or quality of articles, subjective judgment introduces variability that can undermine the validity and reproducibility of the review (Linacre 2013). This approach is found to provide much better results compared to classical test theory tools for the same task, finding interrater reliability, such as Cohen Kappa, Fleiss Kappa and G-theory (Mohd Zabidi et al. 2022). Rater severity, leniency, inconsistency, and scale interpretation are common sources of error that remain largely unaddressed in conventional review protocols. Beyond its statistical advantages, MFRM offers a practical solution for educational settings where groups of faculty or curriculum committees must collectively evaluate scholarly evidence. By providing a common measurement framework across reviewers with differing expertise and perspectives, MFRM supports more consistent decision-making regarding which research should inform courses, programs, and educational policies.

2. Research context: Forest education

Traditional forest education was established in the late 19th century in Europe to train skilled foresters who could effectively manage state forests and address growing concerns about a potential scarcity of timber resources (Collett 2010; Hånell et al. 2005). Consequently, early forest education mainly trained individuals to be state forest managers, with curricula focused on timber harvesting and marked by a lack of concern for regeneration. As the global demand for timber increased, forest education programs were established across Africa, Asia, America, and Oceania. However, although the number and structure of institutions vary by region and forest type, timber extraction orientations continue to dominate curricula worldwide (Rekola and Sharik 2022).

Recent research emphasizes that forces such as climate change, digitalization, and market dynamics, are reshaping forest sector value chains, requiring radical sustainability transitions and new business models (Matthies et al. 2020). Nordic experiences highlight the need for proactive management and innovation to maintain competitiveness in circular and bioeconomy paradigms, which aligns with the growing emphasis on interdisciplinary competencies in forest higher education. These developments underscore the importance of integrating industrial competitiveness and sustainability transitions into forest education curricula, preparing graduates for roles that combine technical expertise with strategic sustainability management. In parallel, forest education has begun to respond to broader societal demands, incorporating competencies that address globalization, digitalization, and climate change (Bullard 2015; Chamlagain et al. 2025, 2026). Universities are gradually embedding economic and social dimensions into curricula, reflecting a shift

toward holistic and interdisciplinary approaches (Kanowski 2001). Today, forest education encompasses formal and informal programs that promote environmental literacy, sustainable forest management practices, and stewardship. Learners engage with topics such as biodiversity conservation, climate change mitigation, and ecosystem services, cultivating deeper connections with nature and contributing to the sustainable use of forest resources (Liu and Zhuang 2022).

The future of forest education has become an increasingly focal point for global organizations. The Food and Agriculture Organization (FAO) of the United Nations, the International Tropical Timber Organization (ITTO), and the International Union of Forest Research Organizations (IUFRO) have collaborated to examine deficiencies in forest education and propose long-term solutions (Rekola and Sharik 2022). As part of a Joint Initiative under the Collaborative Partnership on Forests, IUFRO and the International Forestry Students' Association (IFSA) established a Task Force on Forest Education to strengthen educational practices and make the sector more attractive to younger generations (IUFRO 2020). In partnership with the European Forest Institute (EFI), they conducted research on the future of education and employment in the forest sector. Regionally, the Task Force published inspirational materials to motivate youth in Africa to pursue forest-related studies and careers.

3. Theoretical background

3.1 Systematic literature reviews in forest literature

Systematic literature reviews (SLRs) are a method of synthesizing scientific evidence to address a specific research question in a transparent and reproducible manner, aiming to include all available published evidence on the topic and assess the credibility and quality of that evidence (Rother 2007). SLRs have emerged as a critical methodological tool for evidence synthesis, policy, and practice across diverse fields such as medicine, crime and justice, education, and conservation. However, their application in forestry and related fields has been comparatively slow and faces challenges, despite the field showing greater enthusiasm for the use of systematic maps (Petrokofsky and Savilaakso 2021). One contributing factor is the broad interdisciplinary scope of forestry related disciplines. The field encompasses traditional disciplines, such as silviculture and forest management, as well as contemporary fields including forest policy, environmental sciences, and sustainability studies (Chamlagain et al. 2025; Sheppard et al. 2020).

In the United States, forest education literature has evolved from the mid-1900s, focused primarily on timber production, to today's holistic approach that incorporates ecosystem services, soil science, wildlife management, hydrology, biodiversity conservation, and social sciences. (Barker 2025; Guldin and Brown 2005). In broader forest literature, SLRs have been used across a wide range of topics. For example, SLRs have synthesized methodological development and performance in forest biomass estimation, including the use of field-based allometric models, remote sensing technologies such as unmanned aerial vehicle (UAV platforms), and machine learning (Dashtpeyma and Ghodsi 2021; Latifah et al. 2025). SLRs have been also used in examining trends in carbon (C) storage research by employing bibliometric indicators to evaluate the development of forest C-storage estimation (Feng et al. 2025). Additional SLRs have explored forest management institutions by reviewing the regional variations in conceptualization, analyzing determinants of compliance, and assessing the methodological gaps in studies (Kimengsi et al. 2023).

Furthermore, SLRs methodology has been employed to enhance understanding of other forest-related topics, including urban forestry (Mundher et al. 2022), management practices (Vigna et al. 2021), and innovations (Maier et al. 2021). In forest-related higher education, SLRs have been used to identify inequality forms, issues in recruiting and retaining a more diverse workforce, and promising actions to address these issues (Bullard et al. 2024; Larasatie et al. 2024). Contemporary forest education increasingly emphasizes interdisciplinary competencies, integrating pressing issues such as globalization, digitalization, and climate change (Owuor et al. 2023). This evolution necessitates systematic reviews that can effectively synthesize evidence across multiple domains while maintaining methodological rigor. However, adopting systematic reviews in forestry education and related fields remains limited, with only a very small, recent, and thematically narrow

body of work available (Sjølie et al. 2025). This scarcity, combined with the field's methodological diversity, creates significant challenges for their effective application (Petrokofsky and Savilaakso 2021; Spake and Doncaster 2017).

Forestry studies range from quantitative forest measurement research to qualitative social science investigations of human-forest interactions (Rekola and Sharik 2022), each requiring different evaluation criteria and synthesis approaches. Additionally, the global expansion of forest education has led to research being published in multiple languages and cultural contexts, which may introduce selection bias in reviews (Larasatie et al. 2024; Rekola et al. 2025). Although researchers are interested in keeping SLRs up to date, evidence is dispersed across multiple publications and published in different places, making it difficult for researchers to locate all related work and then synthesize the entire body of evidence (Nepomuceno and Soares 2018). This challenge is further intensified by the substantial effort required to maintain and update SLRs over time, the limited availability of supporting tools and common repositories, and the dependence of any updated synthesis on the heterogeneous quality (Khan et al. 2019).

3.2 Rasch model in systematic literature reviews

Rasch analysis is a psychometric method designed to enhance the precision with which researchers construct instruments, assess instrument quality, and evaluate respondents' performances (Boone 2016). The Rasch model, originally developed for educational and psychological assessment, is now used by researchers and practitioners in education, health care, medical rehabilitation, business, government, and other fields that measure attitude, ability, or performance (Bond and Fox 2015; Boone 2016; Dabaghi et al. 2020; Gordon et al. 2021; Leung et al. 2014). In addition, some social-science researchers are also using Rasch techniques on test validity and the validation process (Reeves and Marbach-Ad 2016), although many continue to rely on classical test theory using instrument development and validation approaches.

The application of Rasch models in SLRs within the forest sector represents an emerging and predominantly unexplored methodological frontier. Current research reveals very few direct applications of Rasch models for conducting systematic reviews in forestry and related fields. One of the few existing applications involves integrating multiple soil variables into a single measure of site quality environmental data using Rasch-based approaches (Moral et al. 2011). Forestry and its related research have primarily focused on traditional meta-analytical approaches for systematic reviews (Spake and Doncaster 2017).

The MFRM extends the basic Rasch framework to accommodate multiple sources of variability in assessment contexts, making it more suited for SLRs. MFRM is an extension of the Rasch Measurement Model (RMM). While RMM enables calibration by computing item and person fit and aligns with a one-parameter model (Engelhard and Wind 2018; Wu 2017), MFRM is specifically focused on multi-rater analysis (Linacre 2013).

The MFRM model simultaneously estimates the difficulty of items, the rater severity/leniency, rater consistency, and the ability of people being assessed, all on the same logit scale (McNamara and Knoch 2012). In forestry education and related interdisciplinary fields, literature spans quantitative forest measurements, qualitative social science investigations, and diverse pedagogical approaches. MFRM in systematic reviews offers a methodological framework for improving transparency, rigor and reliability of article selection processes. In practice, for example, this means that groups of evaluators, such as accreditation teams, curriculum committees, peer reviewers, or faculty workgroups, may inconsistently assess the same set of articles when identifying evidence for curriculum redesign, selecting readings for core courses, evaluating teaching portfolios, or preparing self-study documentation. Such inconsistencies can influence which research is elevated, which findings inform educational decisions, and which pedagogical models ultimately reach students.

4. Methodology

This study aimed to evaluate forestry scientific articles using a multi-rater assessment approach grounded in the MFRM approach. The methodology was designed to achieve two objectives: (1) identify article quality through calibrated multi-rater judgments, and (2) analyze rater behavior, including severity and consistency, during the evaluation process.

4.1 Forestry articles, raters and assessment criteria

Ten forestry-related scientific articles were selected as the subjects of evaluation. These articles originated from a SLR project on forest higher education, which utilized the Web of Science and Scopus databases. The selection focused on articles where disagreements among the SLR team occurred regarding inclusion or exclusion, despite an established protocol. The list of articles is presented in Table 1.

Five experts served as raters in this study. They represented diverse backgrounds—three women and two men from Africa, America, and Europe—with academic and professional experience ranging from PhD-level students to retired professors and research associates in international forestry organizations. Their ages ranged from the 20s to the 70s, and all had prior experience in writing and reviewing scientific papers.

Six evaluation criteria were applied to assess each article: (1) originality; (2) comprehensiveness of literature review; (3) methodology; (4) scientific value of findings; (5) relevance to forest higher education; and (6) quality of analysis. The criteria were developed based on RASCH measurement (Khine 2020). The Rasch software will help to produce information about the quality of each article on a logit scale; the higher the number (positive logits), the better the paper based on the raters' judgment. Additionally, it provides the paper's fit statistics, which indicate its psychometric attributes. The same principle applies to other facets, including criteria and raters.

Each criterion was rated on a three-point Likert scale: sufficient (1), good (2), and very good (3). A three-point Likert scale was selected to balance simplicity for raters with the measurement requirements of Rasch/MFRM (see Bartholomeu et al. 2016). This scale reduces cognitive load and minimizes rater confusion, which is especially important when multiple experts from diverse backgrounds are involved. Although coarser than four- or five-point alternatives, and with limits in discrimination power and measurement sensitivity, three ordered categories are sufficient for MFRM analysis because the model focuses on the probabilistic ordering of categories rather than the number of response options.

Table 1. List of forestry scientific articles being assessed.

No.	Article title	Journal name	Year of publication
1	Natural Resource Service Learning to Link Students, Communities, and the Land	*Journal of Extension*	2012
2	Forest education and research in the United Kingdom	*Forest Science and Technology*	2005
3	Employment and education in forestry	*Journal of Forestry*	1999
4	Wood science education in a changing world: A case study of the UMASS-Amherst building materials & wood technology program, 1965-2005	*Forest Products Journal*	2007
5	Empowering Forestry Extension with geospatial technology	*Journal of Forestry*	2009
6	Impact of technician training programs on professional forestry education in the U.S.	*Journal of Forestry*	1968
7	North Dakota State University Horticulture and Forestry Program Assessment	*HortTechnology*	2010
8	Role of the faculty mentor in an undergraduate research experience	*Journal of Geoscience Education*	2013
9	Developing hybrid SWOT methodologies for choosing joint bioeconomy co-operation priorities by three Finnish universities	*Biofuels*	2016
10	Analysis and planning systems for multiresource, sustainable forestry: the Heureka research programme at SLU	*Canadian Journal of Forest Research*	2003

4.2 Procedures

Raters were invited via email and provided with detailed instructions about the study's purpose and procedures. Upon consent, each rater received a set of randomly assigned articles (in PDF format) along with an assessment form. Each rater evaluated six articles, and each article was assessed by three different raters, resulting in a planned dataset of 10 articles × 3 raters × 6 criteria = 180 observations (Table 2). Due to one missing rating, the final dataset comprised 179 observations.

All ratings were entered into Microsoft Excel and later analyzed using FACETS software, which applies the Many Facet Rasch Model to multi-rater data. This approach enabled simultaneous calibration of article quality, rater severity, and criterion difficulty on a common logit scale.

5. Results

5.1 Rater fit statistics

The first aspect examined is the overall quality of the raters' scoring behavior. Table 3 reports each rater's severity/leniency estimates, along with key psychometric indicators. Three severity groups emerged: R1 was the most severe rater (0.89 logits), R3 and R5 were moderately lenient, and R2 and R4 were the most lenient. All raters demonstrated acceptable fit indices (Outfit MnSq values between 0.5 and 1.5; ZStd values between –2.0 and +2.0; and positive point–measure correlations), indicating that their scoring patterns were consistent and psychometrically sound.

The separation (2.51; should be more than 2.0) and reliability (0.86; should be more than 0.67) confirm that the model successfully distinguished meaningful differences among raters, and the significant chi-square value (p = 0.00) supports the use of the MFRM model for analyzing these data. The close alignment between exact and expected inter-rater agreement (37.6%) further shows that raters worked independently, rather than converging artificially.

Rater severity estimates indicate systematic differences among raters, with some scoring more severely and others more leniently. Identifying these patterns allows the MFRM model to adjust article measures for rater effects, ensuring that logit values reflect article characteristics, rather than individual scoring tenden-

Table 2. Raters and articles.

Raters	Articles number										
	1	2	3	4	5	6	7	8	9	10	Total
1	V	V	V	V	V	V					6
2					V	V	V	V	V	V	6
3	V	V	V		V		V		V		6
4		V		V		V		V	V	V	6
5	V		V	V			V	V		V	6
Total	3	3	3	3	3	3	3	3	3	3	

Table 3. Rater measurement report.

No.	Rater	Model		Outfit		correlation
		Logit	Std.Error	MnSq	ZStd	Pt. Meas.
1	R1	0.89	0.28	1.00	0.00	0.50
2	R2	1.04	0.27	1.47	1.30	0.48
3	R3	−0.54	0.26	1.48	2.00	0.52
4	R4	−1.20	0.29	0.77	−0.50	0.58
5	R5	−0.58	0.27	0.74	−0.80	0.54

Separation 2.51; Strata 3.68; Reliability (not inter-rater) 0.86.
chi-square: 35.4; d.f.: 4; significance (probability): 0.00.
Exact Inter-Rater agreements: 67 = 37.6%; Expected: 75.9 = 42.7%.

cies. The severity information also indicates where variation in ratings originates, and which criteria may be applied differently across raters. These outputs form part of the diagnostic structure of MFRM and contribute to producing calibrated, rater-adjusted measures for all facets in the analysis.

5.2 Criteria difficulty level

Figure 1 presents the criteria difficulty estimates generated by the MFRM. These estimates show the relative strictness with which each criterion was applied across raters. Higher logit values indicate criteria that were rated more stringently, whereas lower values indicate criteria rated more easily. To assess the scientific articles, raters applied six criteria, and the model outputs show that *comprehensiveness of literature review* had the highest difficulty estimate (0.73 logits). The next highest criteria were *scientific value of findings* (0.40 logits) and *methodology* (0.32 logits). These values reflect only the statistical ordering of the criteria within the model and indicate the relative thresholds each criterion generated in the rating process.

The remaining criteria show progressively lower difficulty estimates. Quality of analysis was estimated at –0.11 logits, and originality at –0.28 logits. The lowest difficulty value appeared for relevance to forest higher education (–1.06 logits). These difficulty estimates represent the calibrated hierarchy of the six criteria and form part of the model's adjustment of article scores for criterion effects.

5.3 Raters' judgment to the article

Figure 2 presents the article measures generated by the MFRM after adjustment for rater severity and criterion difficulty. The logit values indicate the relative positions of the 10 articles on the model's latent scale after adjustment for rater severity and criterion difficulty. The model uses the calibrated rater severity estimates and the criterion difficulty hierarchy described in Sections 4.1 and 4.2 to produce these adjusted article measures. The rater severity estimates reported in Section 4.1 show variation in how raters used the scale, and these estimates were incorporated by the model when calculating the article measures. The inter-rater agreement values and the grouping of raters by severity represent the patterns of agreement and divergence detect-

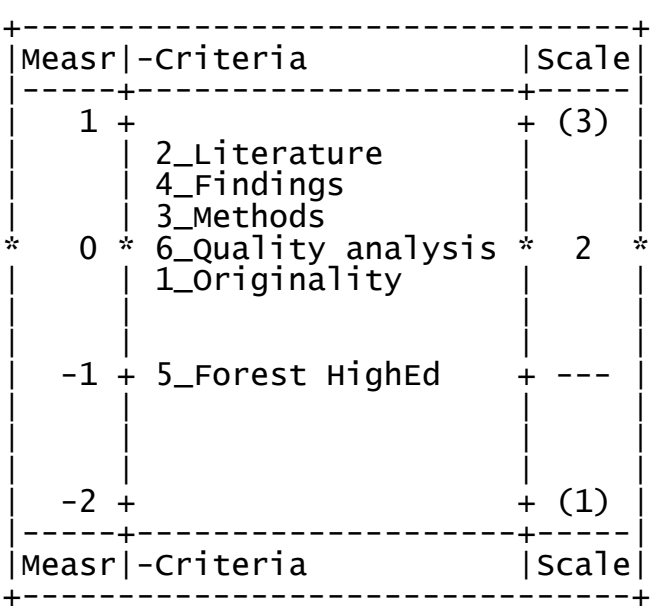

Figure 1. Criteria difficulty level.

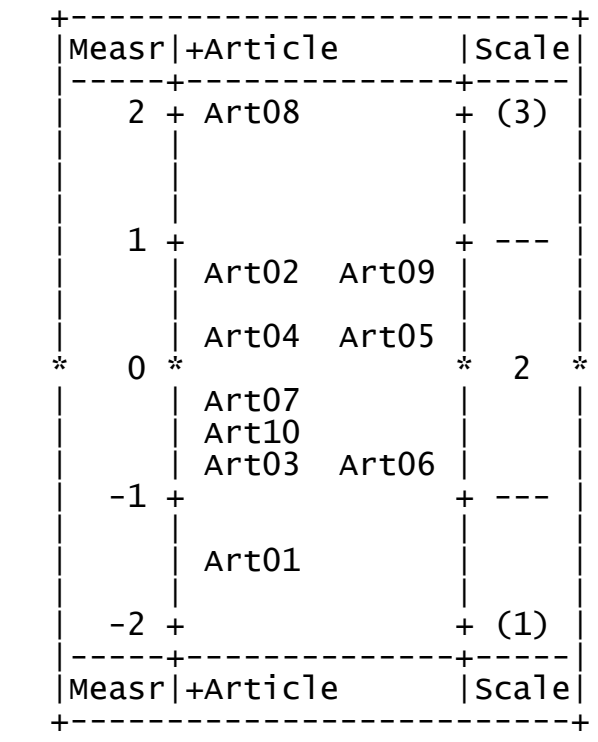

Figure 2. Rank of forestry scientific articles.

ed in the rater facet. These outputs indicate where variation in the ratings originated and form part of the information used by the model to generate the adjusted article logits.

The adjusted measures in Figure 2 show the relative ordering of the 10 articles on the latent scale. Articles with higher logit estimates are positioned toward the upper end of this scale, while lower values indicate placement toward the lower end. In this dataset, Article 8 received the highest logit estimate (1.96 logit), followed by Articles 2, 9, 4, and 5. The remaining articles were located progressively lower on the scale, with Article 1 receiving the lowest estimate (-1.51 logit). These values represent only the calibrated ordering of the articles within the model and do not imply external quality judgments beyond their relative positions in this analysis.

6. Discussions and conclusions

This study demonstrates the practical value of applying the Many Facet Rasch Model to improve the rigor and transparency of SLRs in forest higher education. By calibrating rater severity, assessing rating scale use, and identifying patterns of agreement and divergence, MFRM provides evidence that expert judgments can be analyzed in a reliable, psychometrically defensible way. These findings directly address the objectives of the study by showing that MFRM effectively reduces subjective variation across reviewers and supports more consistent article-selection processes.

The rater fit statistics indicate that all experts applied the rating criteria within acceptable ranges, and the separation and reliability values show that the model successfully distinguished meaningful differences in rater severity. This demonstrates that MFRM can calibrate rater behavior across individuals with diverse backgrounds, helping mitigate inconsistencies that often arise in multi-rater SLR workflows. Identifying which raters tended to be more severe or lenient provides review teams with actionable diagnostic information that can support rater training, calibration sessions, or improved protocol design.

The analysis also identified differences in the difficulty of specific evaluation criteria. While these patterns help illuminate how experts interpret methodological rigor, analytical quality, or relevance, they are best understood as secondary contextual insights rather than the primary aim of the study. Their value lies in illustrating how MFRM can detect uneven application of criteria, enabling teams to refine coding manuals, clarify inclusion standards, or adjust review procedures to enhance clarity and consistency.

Although article-level logit scores offer a ranked view of quality within the sample, the purpose of presenting them is to demonstrate how calibrated measures can support evidence-mapping and priority-setting in systematic reviews, not to make claims about the intrinsic quality of the included article. By converting subjective judgments into a common linear scale, MFRM provides review teams with a transparent method for comparing studies, identifying borderline cases, and supporting defensible inclusion decisions.

As forest education continues to evolve in response to global challenges such as climate change, digitalization, and equity, the need for robust and transparent review methodologies becomes increasingly urgent. This study contributes a methodological innovation that not only enhances the quality of systematic reviews but also supports the broader goal of advancing forest education through rigorous, evidence-informed scholarship. Taken together, the rater fit statistics, criteria difficulty patterns, and calibrated article rankings demonstrate how MFRM outputs can directly strengthen large-scale SLR and evidence-mapping projects beyond forest-related higher education, including forest-based bioeconomy research, and education in general. For education stakeholders including journal editors, curriculum designers, faculty mentors, and policy advisors, this approach provides a defensible method for selecting high-quality literature, refining peer review processes, and strengthening evidence-based decision-making. The calibrated difficulty levels of evaluation criteria highlight areas where forest education research may benefit from clearer standards, particularly in literature synthesis and methodological transparency.

6.1 Implications for practice

Forest higher education faces a dual challenge: the rapid expansion of available information and the persistent difficulty of identifying what is most relevant for instructional and curricular decisions. Instructors, mentors, and curriculum designers are increasingly tasked with curating content that not only reflects

scientific rigor but also cultivates critical thinking and interdisciplinary awareness. This study responds to that challenge by offering a calibrated framework, using the Many Facet Rasch Model, to evaluate the quality of scholarly literature through expert judgment. Rather than relying on intuition or ad hoc selection, educators can use MFRM-based insights to identify articles that meet high standards across key dimensions such as methodological transparency, analytical depth, and relevance to forest education.

Although daily instructional work often involves individual decision-making, several common processes in forestry education do rely on multi-rater evaluation. These include accreditation reviews (e.g., Society of Wood Science and Technology (SWST) and Society of American Foresters (SAF) accreditation), curriculum revision committees, team-taught or integrated courses, and graduate program oversight. In these contexts, groups of faculty collectively review literature, course materials, or evidence portfolios. MFRM can be applied to these group-based evaluations by calibrating differences in rater severity and producing adjusted measures that reflect the relative placement of materials on a common scale. This supports consistency and transparency in collaborative decision-making processes that have direct implications for curriculum quality and program development.

The calibrated difficulty levels of evaluation criteria also provide a diagnostic lens for improving student writing, guiding thesis supervision, and refining course materials. In peer review settings, this approach can help journals and academic programs ensure fairness, consistency, and transparency in evaluating submissions and teaching portfolios. The model supports a shift toward evidence-informed pedagogy, where the selection of teaching materials is grounded in reproducible analysis, and the training of future foresters is shaped by the best available science.

Beyond forest educational contexts, the findings also demonstrate clear value for large-scale evidence synthesis in the forest sector. For example, for forest-sector businesses and bio-products innovators, more reliable and transparent literature screening helps accelerate technology evaluation, identify credible research for investment decisions, and benchmark emerging trends in competitiveness or sustainability performance. Similarly, policymakers and governance bodies can use MFRM-enhanced SLRs to ensure that decisions about forest resources, circular bioeconomy strategies, or climate-related interventions are grounded in consistently evaluated evidence.

6.2 Limitations

While this study offers a novel application of the Many Facet Rasch Model in forestry higher education, several limitations should be acknowledged to contextualize its findings. First, the scope of analysis was intentionally narrow: 10 articles and five expert raters were selected to demonstrate methodological feasibility, not to represent the full diversity of forest education literature. A broader sample would allow for more nuanced insights across subfields, geographic regions, and institutional contexts.

Second, the evaluation criteria, though grounded in scholarly standards, were developed specifically for this study. Their relevance to other domains such as Forestry Extension, indigenous knowledge systems, or vocational training may require adaptation and validation. Without such refinement, the transferability of the model to other educational settings remains limited.

Third, while MFRM excels at quantifying rater behavior and item difficulty, it does not capture the qualitative reasoning behind expert judgments. The absence of narrative feedback or contextual interpretation means that some pedagogically valuable insights may remain hidden. Future studies could integrate mixed methods to bridge this gap between statistical calibration and educational meaning.

6.3 Future pathways

This study opens several promising avenues for advancing forest education research, evaluation, and practice. One direction is the development of standardized, Rasch-calibrated rubrics for evaluating forest education literature across institutions, languages, and cultural contexts. Such tools could support journal editors, grant reviewers, and curriculum committees in making transparent, reproducible decisions about scholarly quality.

Another opportunity lies in extending MFRM beyond literature reviews to classroom assessment, peer evaluation, and faculty development. By apply-

ing the model to student work, teaching portfolios, or capstone projects, institutions could bring psychometric rigor to educational quality assurance, supporting fairer and more consistent evaluation practices.

Future research might also explore how rater characteristics such as disciplinary background, cultural context, or training, affect severity and leniency patterns. Understanding these dynamics could inform more equitable review systems and improve cross-cultural collaboration in forest education.

Finally, researchers could investigate how calibrated article rankings influence student learning outcomes, curriculum coherence, or policy uptake. By linking methodological innovation to educational impact, forest education can evolve not only in what it teaches, but in how it builds knowledge for a sustainable future.

7. Acknowledgement

The authors thank the anonymous raters and reviewers who generously contributed their time and expertise to strengthen this manuscript. We also acknowledge that, with the exception of the second and last authors, all contributors were core members of the International Union of Forest Research Organizations (IUFRO) Task Force on Forest Education (2019–2024). This manuscript was developed as part of the Task Force's collective scientific efforts to advance forest science education.

8. References

Abas, A. (2023). A systematic literature review on the forest health biomonitoring technique: A decade of practice, progress, and challenge. *Frontiers in Environmental Science* 11, 970730. https://doi.org/10.3389/fenvs.2023.970730

Andrich, D., Marais, I. (2019). *A Course in Rasch Measurement Theory, Measuring in the Educational, Social and Health Sciences.* Dordrecth: Springer.

Aryadoust, V., Tan, H.A.H., Ng, L.Y. (2019). A Scientometric review of Rasch Measurement: The rise and progress of a specialty. *Frontier Psychology* 10, 2197. https://doi.org/10.3389/fpsyg.2019.02197

Barker, M. (2025). Forest education: past, present, and future. *Forests* 16(12), 1801. https://doi.org/10.3390/f16121801

Bartholomeu, D., Silva, M., Montiel, J. (2016) Improving the Likert Scale of the Children's Social Skills Test by means of Rasch Model. *Psychology* 7, 820–828. https://doi.org/10.4236/psych.2016.76085

Bentley, W. R. (1999). Professional forestry education in New York: An old lesson, a new model. *Journal of Forestry* 97(9), 29–32. https://doi.org/10.1093/jof/97.9.29

Bezruczko, N. (2005). *Rasch measurement in health sciences*. Maple Grove, MN: Jam Press.

Bond, T. G., Fox, C. M. (2015). *Applying The Rasch Model, Fundamentals Measurement in the Human Sciences*. 3rd edition. New York: Routledge. https://doi.org/10.4324/9781410614575

Boone, W. J. (2016). Rasch Analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education* 15(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecth: Springer.

Bullard, S. H. (2015). Forestry curricula for the 21st Century-maintaining rigor, communicating relevance, building relationships. *Journal of Forestry* 113(6), 552. https://doi.org/10.5849/jof.15-021

Bullard, S. H., Walker, T. J., Burger, L. (2024). Enhancing diversity in undergraduate degree programs in forestry and related natural resources: A brief review of critical issues and promising actions. *Journal of Forestry* 122(2), 107–122. https://doi.org/10.1093/jofore/fvad043

Chamlagain, K., Larasatie, P., Rubino, E., Knowles, S. (2025). From asking "would I be ready?" to "would I belong?": Preparedness perceptions of forest and natural resources university students in the United States to enter the workforce. *Forest Policy and Economics* 178, 103580. https://doi.org/10.1016/j.forpol.2025.103580

Chamlagain, K., Larasatie, P., Pelkki, M., Knowles, S., Chhetri, S., Rubino, E. (2026). What drives individuals to pursue careers in the forestry and related natural resources sectors? *Journal of Forestry*. In press.

Christensen, K.B., Kreiner, S., Mesbah, M. (2012). *Rasch in Health*. John Wiley & Sons.

Collett, N. G. (2010). A history of forestry education in Victoria, 1910–1980. *Australian Forestry* 73(1), 34–40. https://doi.org/10.1080/00049158.2010.10676307

Cook, D. J., Greengold, N. L., Ellrodt, A. G., Weingarten, S. R. (1997a). The Relation between Systematic Reviews and Practice Guidelines. *Annals of Internal Medicine* 127(3), 210. https://doi.org/10.7326/0003-4819-127-3-199708010-00006

Cook, D. J., Mulrow, C. D., Haynes, R. B. (1997b). Systematic reviews: Synthesis of best evidence for clinical decisions. *Annals of Internal Medicine* 126(5), 376–380.

Dabaghi, S., Esmaielzadeh, F., Rohani, C. (2020). Application of Rasch Analysis for development and psychometric properties of adolescents' quality of life instruments: A systematic review. *Adolescent Health, Medicine and Therapeutics* 11, 173–197. https://doi.org/10.2147/AHMT.S265413

Dashtpeyma, M., Ghodsi, R. (2021). Forest Biomass and Bioenergy Supply Chain Resilience: A Systematic Literature Review on the Barriers and Enablers. *Sustainability* 13(12), 6964. https://doi.org/10.3390/su13126964

Engelhard Jr, G., Wind, S. (2018). *Invariant Measurement with Raters and Rating Scale*s. New York: Routledge.

Engelhard Jr, G., Wang, J. (2021). *Rasch models for solving measurement problems: Invariant measurement in the social sciences*. Sage Publications. https://doi.org/10.4135/9781071878675

Feng, Y., Liu, J., Hu, H., Cui, P., Zhou, H., Ma, B., Liu, Z., Chen, D. (2025). Global patterns in forest carbon storage estimation: Bibliometric analysis of technological evolution, accuracy gains and scaling challenges. *Frontiers in Forests and Global Change* 8, 1649356. https://doi.org/10.3389/ffgc.2025.1649356

Gordon, R. A., Peng, F., Curby, T. W., Zinsser, K. M. (2021). An introduction to the many-facet Rasch model as a method to improve observational quality measures with an application to measuring the teaching of emotion skills. *Early Childhood Research Quarterly* 55, 149–164. https://doi.org/10.1016/j.ecresq.2020.11.005

Guldin, R. W., Brown, P. (2005). Forest education and research in the United States of America. *Forest Science and Technology* 1(2), 120–126. https://doi.org/10.1080/21580103.2005.9656278

Hou, S. K., Liu, Y. R., Li, C. Y., Qin, P. X. (2020, October). Dynamic prediction of rock mass classification in the tunnel construction process based on random forest algorithm and TBM in situ operation parameters. In *IOP Conference Series: Earth and Environmental Science* 570(5), p. 052056. IOP Publishing. https://doi.org/10.1088/1755-1315/570/5/052056

Hånell, B., Magnusson, T., Hallgren, J.-E., Karlsson, A. (2005). Swedish forest research and higher education-challenging issues and future strategies of forest research and education in Sweden. *Forest Science and Technology* 1(2), 98–103.

IUFRO. (2020). *IUFRO: Joint IUFRO-IFSA Task Force on Forest Education / Task Forces / Science in IUFRO*. International Union of Forest Research Organizations. https://www.iufro.org/science/task-forces/forest-education/

Kanowski, P. (2001). Forestry education in a changing landscape. *The International Forestry Review* 3(3), 175–183. https://www.jstor.org/stable/42609382

Khan, S., Memon, B., Memon, M. A. (2019). Meta-analysis: A critical appraisal of the methodology, benefits and drawbacks. *British Journal of Hospital Medicine* 80(11), 636–641. https://doi.org/10.12968/hmed.2019.80.11.636

Khine, M. S. (2020). *Rasch Measurement Applications in Quantitative Educational Research*. Springer Nature Singapore Pte Ltd.

Kimengsi, J. N., Owusu, R., Charmakar, S., Manu, G., Giessen, L. (2023). A global systematic review of forest management institutions: Towards a new research agenda. *Landscape Ecology* 38(2), 307–326. https://doi.org/10.1007/s10980-022-01577-8

Larasatie, P., Jones, E., Hansen, E., Lewark, S. (2024). A wake-up call? A review of inequality based on the forest-related higher education literature. *Environmental Science & Policy* 162, 103942. https://doi.org/10.1016/j.envsci.2024.103942

Latifah, S., Gandaseca, S., Afifi, M., Prasetyo, A. R., Purnama, M. I., Kertalam, L. R. A., Pratama, R. P. (2025). Three decades of forest biomass estimation in Southeast Asia: A systematic review of field, remote sensing, and machine learning approaches (1995–2025). *Jurnal Sylva Lestari* 13(3), 728–746. https://doi.org/10.23960/jsl.v13i3.1162

Leung, Y.-Y., Png, M.-E., Conaghan, P., Tennant, A. (2014). A systematic literature review on the application of Rasch Analysis in musculoskeletal disease—A Special Interest Group Report of OMERACT 11. *The Journal of Rheumatology* 41(1), 159–164. https://doi.org/10.3899/jrheum.130814

Li, J. (2019). A random dynamic search algorithm research. *Journal of Computational Methods in Sciences and Engineering* 19(3), 659–672. https://doi.org/10.3233/JCM-193522

Linacre, J. M. (2013). Facets computer program for many-facet Rasch measurement, version 3.71. 4. *Beaverton, Oregon: Winsteps. Com*.

Liu, G., Zhuang, H. (2022). Evaluation model of multimedia-aided teaching effect of physical education course based on random forest algorithm. *Journal of Intelligent Systems* 31(1), 555–567. https://doi.org/10.1515/jisys-2022-0041

Maier, C., Hebermehl, W., Grossmann, C. M., Loft, L., Mann, C., Hernández-Morcillo, M. (2021). Innovations for securing forest ecosystem service provision in Europe – A systematic literature review. *Ecosystem Services* 52, 101374. https://doi.org/10.1016/j.ecoser.2021.101374

Matthies, B., Korhonen, J., Toppinen, A. (2020). Current and future research themes in forest sector competitiveness: Case study of research orientations at the University of Helsinki. *BioProducts Business* 5(8), 87–106. https://doi.org/10.22382/bpb-2020-008

McNamara, T., Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing* 29(4), 555–576. https://doi.org/10.1177/0265532211430367

Mohd Zabidi, Z., Sumintono, B., Abdullah, Z. (2022). Enhancing analytic rigor in qualitative analysis: Developing and testing code scheme using many facet Rasch model. *Quality & Quantity* 56(2), 713–727. https://doi.org/10.1007/s11135-021-01152-4

Moral, F. J., Terrón, J. M., Rebollo, F. J. (2011). Site-specific management zones based on the Rasch model and geostatistical techniques. *Computers and Electronics in Agriculture* 75(2), 223–230. https://doi.org/10.1016/j.compag.2010.10.014

Mulrow, C. D. (1987). The Medical Review Article: State of the Science. *Annals of Internal Medicine* 106(3), 485. https://doi.org/10.7326/0003-4819-106-3-485

Mundher, R., Abu Bakar, S., Al-Helli, M., Gao, H., Al-Sharaa, A., Mohd Yusof, M. J., Maulan, S., Aziz, A. (2022). Visual Aesthetic Quality Assessment of Urban Forests: A Conceptual Framework. *Urban Science* 6(4), 79. https://doi.org/10.3390/urbansci6040079

Nepomuceno, V., Soares, S. (2018). Maintaining systematic literature reviews: Benefits and drawbacks. *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 1–4. https://doi.org/10.1145/3239235.3267432

Ohlsson, A. (1994). Systematic reviews—Theory and practice. *Scandinavian Journal of Clinical and Laboratory Investigation* 54(sup219), 25–32. https://doi.org/10.3109/00365519409088573

Owuor, J. A., Winkel, G., Giessen, L., Prior, L., Burns, J., Tegegne, Y. T., Poschen, P. (2023). Passion for nature: Global student motivations for forest related education and career aspirations. *International Forestry Review* 25(3), 358–371. https://doi.org/10.1505/146554823837586212

Petrokofsky, G., Savilaakso, S. (2021). The value of systematic evidence synthesis in forestry, land use and development to improve research, decision-making and practice. *Forests* 12(10), 1355. https://doi.org/10.3390/f12101355

Reeves, T. D., Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: a primer for discipline-based education researchers. *CBE—Life Sciences Education* 15(1), rm1. https://doi.org/10.1187/cbe.15-08-0183

Rekola, M., Sharik, T. L. (2022). *Global Assessment of Forest Education: Creation of a Global Forest Education Platform and Launch of a Joint Initiative under the Aegis of the Collaborative Partnership on Forests (FAO-ITTO-IUFRO project GCP/GLO/044/GER)*. Forestry Working Paper No. 32. Rome, FAO. https://doi.org/10.4060/cc2196en

Rekola, M., Taber, A. B., Sharik, T. L., Parrotta, J. A., Dockry, M. J., Babalola, F. D., Bal, T. L., Ganz, D., Gruca, M., Guariguata, M. R., Kungu, J., Larasatie, P., Nevgi, A., Rodriguez-Piñeros, S., Saengcharnchai, S., Sandström, N., Walji, K. (2025). Social and knowledge diversity in forest education: Vital for the world's forests. *Ambio* 54(4), 660–669. https://doi.org/10.1007/s13280-024-02104-6

Rother, E. T. (2007). Systematic literature review X narrative review. *Acta paulista de enfermagem* 20, v-vi. https://doi.org/10.1590/S0103-21002007000200001

Sharik, T. L., Lilieholm, R. J., Lindquist, W., Richardson, W. W. (2015). Undergraduate enrollment in natural resource programs in the United States: Trends, drivers, and implications for the future of natural resource professions. *Journal of Forestry* 113(6), 538–551. https://doi.org/10.5849/jof.14-146

Sheppard, J. P., Chamberlain, J., Agúndez, D., Bhattacharya, P., Chirwa, P. W., Gontcharov, A., Sagona, W. C. J., Shen, H., Tadesse, W., Mutke, S. (2020). Sustainable forest management beyond the timber-oriented status quo: Transitioning to co-production of timber and non-wood forest products—a global perspective. *Current Forestry Reports* 6(1), 26–40. https://doi.org/10.1007/s40725-019-00107-1

Sjølie, H. K., Lauritzen, T., Akin, D. (2025). Unveiling multiple pathways into tertiary forestry education: A mixed-method study. *Scandinavian Journal of Forest Research* 41(1), 68–83. https://doi.org/10.1080/02827581.2025.2566182

Spake, R., Doncaster, C. P. (2017). Use of meta-analysis in forest biodiversity research: Key challenges and considerations. *Forest Ecology and Management* 400, 429–437. https://doi.org/10.1016/j.foreco.2017.05.059

Tranfield, D., Denyer, D., Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management* 14(3): 207–222. https://doi.org/10.1111/1467-8551.00375

Vigna, I., Besana, A., Comino, E., Pezzoli, A. (2021). Application of the Socio-Ecological System Framework to Forest Fire Risk Management: A Systematic Literature Review. *Sustainability* 13(4), 2121. https://doi.org/10.3390/su13042121

Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling* 59(4), 453–470. https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2017_20171218/04_Wu.pdf

Wu, X., Zhou, Y., Xing, H. (2021). Studies on the evaluation of college classroom teaching quality based on SVM multiclass classification algorithm. In *Journal of Physics: Conference Series* 1735 (1), p. 012011). IOP Publishing. https://doi.org/10.1088/1742-6596/1735/1/012011