360-Degree Multi-Source Assessment Framework for Evaluating Learning Outcomes in Experiential Science Education Among Middle-Stage Students: A Systematic Literature Review

Abstract

Experiential approaches to science instruction, such as project-based activities, lab experiments, field inquiries, and citizen science, are extensively advocated for middle-stage learners (around 10–15 years of age). Most assessment techniques used to evaluate the results of these activities are mainly single-source assessments, relying on teacher-assigned grades or standardized tests. They reduce the evidence about the cognitive, procedural, affective, and social-emotional outcomes that experiential pedagogy intends to develop. This literature review summarizes the literature: (a) the impacts of experiential learning methods on science education outcomes of middle-school-aged students and (b) the multi-source or 360-degree assessment (i.e. self-, peer-, teacher, and parent/guardian-based) in K-12 education and other educational settings. Following PRISMA 2020 guidance, records were identified in Scopus, Web of Science and ERIC/Education Source, supplemented by citation snowballing and Google Scholar. Due to heterogeneity in the research designs and outcome variables, the results were grouped thematically but not analyzed statistically. The key takeaways of the review: First, experiential science instruction is generally associated with gains in conceptual understanding, inquiry skills and motivation relative to didactic instruction, although effects vary with fidelity, duration and outcome measure. Second, while multi-rater designs enhance construct coverage and mitigate the bias associated with single raters, their application is infrequent in K-12 science classrooms, being predominantly utilized in higher education, medical training, and workplace evaluations. Third, no validated framework currently integrates multi-source assessment with experiential science pedagogy for middle-grade students. The review article explores the concept of tripling through the frameworks of experiential learning and multi-informant assessment research. It proposes a multi-source assessment approach, incorporating self-assessment, planned peer assessment, teacher records, parent or guardian feedback, and product-based evidence, to evaluate various aspects of a student's abilities. The article also addresses classroom practices, teacher training, and future research on validation, while acknowledging the review's limitations.

Keywords:

360-degree assessment; multi-source assessment; experiential learning; science education; middle school; learning outcomes

1. Introduction

The middle years of schooling, around the ages of 10 to 15, are a critical period for science learning when students' attitudes and confidence as science learners are developed through interaction and engagement. Their scientific inquiry patterns are also formed during these young years, and they are expected to carry through into later academic years. Therefore, the ways in which science is taught and evaluated at this level will impact on the educational career.

Experiential approaches to teaching are rooted in Dewey's (1938) view that education grows out of meaningful experience, and Kolb's (1984) framework of experiential learning emphasizes that learning involves a continuous cycle of experience, reflection, conceptualization, and experimentation, rather than a linear progression. According to Vygotsky (1978), understanding is mediated through interaction with others. In the classroom, these ideas translate into lab investigations, field inquiries, project work, and STEM units integrated as a package, and citizen science. These methods are alternatives to traditional teacher-centred, textbook-heavy instruction.

A review of several studies on experiential learning showed that students learned well in experiential approaches and had good grades (Burch et al., 2019). In addition, a fast evidence check of 4–14-year-old children's learning that academic gains are reported in most subjects and warnings are mentioned that the size of the effect may not be independent but a function of the implementation (fidelity), time (duration), and measuring (type of result). (Ranken, 2025). Science research has also provided similar results. For example, a middle school integrated STEM curriculum has been related to improved life science learning (Anwar et al., 2022).

The learning process and change in scientific identity can be explained through examples cited in the teaching of citizen science (Zhang et al., 2023), while ecological knowledge and engagement have a relationship through nature-based programmes (Carrier et al., 2013; Morag et al., 2013). Although education and pedagogy in science have evolved, assessment practice has not kept pace with these changes. Science assessment in many schools still relies on examinations, quizzes and end-of-unit tests scored by the teacher. This is a single-source model, and it captures only part of what experiential teaching sets out to develop. Collaboration, self-regulation, using learning in new contexts, and developing one's interest as well as sense of personal capabilities are difficult to observe from a paper test, which limits oneself to a particular point of view. Majumdar (2026) points out how the teacher-only, single-source methods of evaluating achievement in middle school experiential science have been predominant.

Multi-source, or 360-degree, could be an alternative solution to this problem. Initially developed for the assessment of work-related performance, this approach has also been adapted for higher education, professional development, and teacher training programmes. Tee and Ahmed (2014) explained that this

type of assessment integrates student self-assessment, peer evaluation by the student's classmates, and assessment and ratings made by supervisors or stakeholders to create a comprehensive picture of skills and capabilities.

The concept of multi-informant research that has grown in developmental and school psychology makes a significant point: that a teacher, parent, and student will each report different aspects of a child' s functioning; therefore, the agreement among them is limited, and the disagreement itself may provide information (Mudarra et al., 2022). The fact that students can be rated by a number of teachers has also been shown to lead to better quality measurement (Muschkin & Malone (2007). In other words, collecting ratings from different raters can yield the best assessment outcome.

Despite these advantages, multi-source assessment is seldom used in middle-grade experiential science education. Where multi-rater designs do appear in school-age populations, they cluster in social-emotional and behavioral measurement rather than science-specific learning outcomes. In this article, we first list a series of difficulties associated with the current assessment model for middle school science. For example, assessments rely almost completely on single raters and written exams, which are incompatible with the multifaceted nature of experiential learning. Second, multi-source assessments have mostly been the subject of studies beyond childhood. Third, no previous study has managed to bring together the two different fields of research into a developmental framework.

The manuscript outlines the state of the art on experiential learning science and multi-source assessments. Based on the evidence, a new framework is developed for middle-school science lessons through an interdisciplinary combination of both areas. To be more specific, objectives and research questions are stated first, followed by a literature review and an introduction of the review approach. Then, the integrated findings are presented and interpreted in a discussion section, while the article ends with closing remarks and recommendations. The presented framework is of a theoretical nature and is supposed to serve as an input for future empirical studies.

2. Review of Literature

2.1 Theoretical Foundations of Experiential Science Learning

Dewey (1938) stated that learning is a series of interactions and that one cannot fully grasp anything without connecting the current with the past as well as future. According to Kolb, this is a circular process made of four different states. Furthermore, Morris (2020) has updated the original idea and given recommendations for further development of the learning cycle, which is still an influential concept in education.

If the cycle is considered to be the way of learning and changing of one's knowledge, assessment should not occur only at a testing point but rather throughout the process in reflection, forming ideas, and testing

out new concepts. Vygotsky (1978) focused on teaching a child in a manner that the child can grasp the lesson and further suggested that peer interaction and the involvement of teachers are crucial parts of the learning process and can therefore be the sources of evaluation of learning.

2.2 Experiential Learning and Science Outcomes

The results of a meta-analysis on the experiential learning process within and across different areas show that learning is positively facilitated by the experiential learning method (Burch et al., 2019); however, there is no evidence about science specifically. Ranken (2025) conducted a literature review of the application of experiential learning in children between the ages of 4 and 14 years and stated better academic results, with the strength of the effect differing based on how faithful, how long, and how the intervention/measure was done. In the field of science, Anwar et al. (2022) tested the integration of STEM concepts with standard lessons to observe the effect on life science knowledge. The middle school students were found to learn better when an integrated STEM unit was given to them. Zhang et al. (2023) identified citizen science across different levels and linked active participation to learning scientific practice and the formation of identity.

Experiential educational approaches that combine outdoor/nature-based and hands-on activities have shown a close correlation with increased ecological literacy and long-lasting engagement (Carrier et al., 2013; Morag et al., 2013). However, the same authors highlight that the limited availability of time, underpaid staffing, and testing pressure are some of the biggest hindrances, which is why such approaches are rarely used in regular classrooms. Through an examination of the use of hands-on science instruction in middle schools, Majumdar (2026) pointed out additional advantages. Nonetheless, outcomes tend to be primarily gathered through single-source methods.

2.3 How Experiential Science Outcomes Are Currently Measured

The measures in the selected research primarily focus on student cognitive and achievement results. These results are most often obtained from teacher-graded or researcher-administered tests (Burch et al., 2019; Majumdar, 2026). Interest, self-efficacy, and attitude, along with outcomes involving collaboration and social-emotional dimensions, seem to be evaluated less frequently. Typically for these purposes, only student self-report is used as a source of data. Chan et al. (2021) report that outcomes of this type are of utmost importance, yet other individuals who are involved are hardly ever used as a source of evidence. Thus, at a basic level, it was found to be an inconsistency that experiential learning is considered the most effective pedagogy for the outcomes that require triangulation as the least of its methods.

2.4 Multi-Source and 360-Degree Assessment in Education

Multi-Source Framework gathers opinions or ratings from different stakeholders, including the assessee, their peers, immediate supervisors or teachers, and sometimes, others. This idea originated from appraisals in the workplace and later expanded to higher education, clinical training, and professional development for teachers (Tee & Ahmed, 2014). Three benefits have been repeatedly mentioned in the literature. Using several different raters together can counter the bias and peculiarities that any single rater might possess. This provides a complete picture of the thinking, feeling, and behavior of performance. It also allows for formative, developmental, and feedback, as well as the judgment of summative results. In higher education, it has served to promote self-regulated learning and leadership development; however, its use in K-12 classrooms seems to be very limited.

The supporting principles of the framework are based on formative assessment research. Black and Wiliam (1998) showed the learning-oriented value of feedback, Boud (1995) advocated for a model of self-assessment that is structured, and Topping (1998) conducted an overview of peer assessment and pinpointed the three elements that must be present: clear criteria, training, and calibration, in order for the peer assessments to be trusted.

2.5 Evidence from Multi-Informant Research

Muschkin and Malone (2007) examined multiple ratings of the same students by teachers, finding that multiple ratings resulted in a better measurement strategy and lower error; thus, the multi-rater approach is valid even if it does not involve parents or peers. Mudarra et al. (2022) compared ratings of teenagers' social-emotional skills by teachers, parents, and students themselves, finding rather low levels of agreement. Raters showed higher convergence when it came to behavior observable from the outside (attendance, work completed), and divergence for internal state measures, where students could provide additional information about their experience unavailable to others. Two conclusions can be drawn here. Raters are different, and averaging will lead to a loss of important information. Disagreement between raters should be interpreted as information about the context, not an error. Comparative analysis of self-, peer-, and instructor ratings of performance tasks leads to the same conclusion: self-and peer ratings differ from instructor ratings in a systematic manner.

2.6 Identified Gaps

Five gaps appear in the literature review. First, there is a gap between multidimensional experiential pedagogy and unidimensional assessment tools. Second, there is a target population gap because multi-source models have only been empirically proven for use in higher education and professional education programs. Third, there is an area or subject matter gap in that multi-rater studies with school children pertain only to social-emotional factors and disregard the more cognitive aspects, procedures, and applications associated with science. Fourth, there is a gap between research methods and practice in not having any

empirical model that includes the self and multi-rater informed structure of assessment. Finally, there is a practicality gap related to time, training, etc.

3. Objectives of the Study

The present review was undertaken to address the identified gaps and limitations in the existing literature concerning experiential science education and the assessment of learning outcomes among middle-grade students. The review aims to integrate research on experiential learning and multi-source assessment to establish a comprehensive conceptual foundation for evaluating science learning at the middle stage. The specific objectives are as follows:

1. To systematically examine and synthesize the available literature on the learning outcomes of experiential science education among middle-grade students and identify the assessment methods and evaluation techniques employed to measure these outcomes.
2. To critically analyse the characteristics, validity evidence, and implementation challenges associated with multi-source and 360-degree assessment approaches in educational settings and related disciplines, with particular emphasis on self-assessment, peer assessment, teacher assessment, and parent or caregiver assessment.
3. To formulate a comprehensive conceptual framework for a 360-degree multi-source assessment approach tailored to experiential science education at the middle stage, drawing upon the synthesized evidence and insights from the existing literature.
4. To propose evidence-informed recommendations for classroom assessment practices, teacher professional development, and educational policymaking, while identifying priorities and directions for future research on the assessment of experiential science learning.

The present study is literature-based and does not involve primary data collection. Instead, it brings together and critically examines distinct yet interrelated strands of research that have largely developed independently, including experiential science education, learning outcome assessment, and multi-source evaluation. By integrating these perspectives, the review seeks to bridge existing conceptual gaps and provide a coherent foundation for advancing assessment practices, informing educational decision-making, and guiding future empirical investigations in middle-stage science education.

4. Research Questions

RQ1: What learning outcomes have been linked to experiential learning approaches in middle-grade science teaching, and which evaluation methods have researchers employed to measure and judge those outcomes in classroom practice?

RQ2: Which multi-source or 360-degree assessment models, incorporating self-, peer-, teacher-, and parent-informed evaluations, have been implemented, across different education systems, and in what ways have the designs of these models been validated to confirm their quality and credibility?

RQ3: What design features are essential for developing a 360-degree, multi-source assessment model that is both authentic and practical for evaluating the learning outcomes of experiential science teaching at the middle-grade level, so that teachers, students, and families can use it with confidence?

5.1 Research Type

The investigation at hand is a systematic literature review which follows the PRISMA 2020 guidelines. As it integrates two contrasting areas of knowledge, experiential teaching and multi-source assessment, it has used a parallel-strand approach to both search and combine the relevant documents. The separate strands were searched first and then combined into an integrative synthesis, yielding a shared set of outcome domains. Since this is a conceptual and secondary study, no primary data were gathered from the participants.

5.2 Population

The population consisted of academic works in science education and assessment that focused on middle-stage learners aged 10–15, i.e., grade levels 6 to 8 or 9. On the educational side, the review concentrated its efforts on studies of experiential, inquiry-based, practical, and activity-oriented science teaching. As far as assessment was concerned, the review focused on papers using multi-source, multi-rater, or 360-degree methods that collected data from students, peers, teachers, or parents.

5.3 Sample

We used a systematic approach to select the sample and followed eligibility criteria for including the papers after searching four databases: Scopus, Web of Science, ERIC/Education Source, and Google Scholar. In addition, backward and forward citation tracking, grey literature, and recent conferences have been used to supplement the search. The search period began in 2013 and ended in 2026. In addition, a very small number of foundational works, Dewey (1938), Kolb (1984), and Vygotsky (1978) were excluded because they were not considered empirical evidence.

For example, records were regarded as meeting the inclusion criteria when they related to middle-stage students, addressed science education, involved experiential/inquiry-based instruction, or included multi-source assessment. Furthermore, they are peer-reviewed publications, conference papers, systematically reviewed publications, or meta-analyses that can be accessed as English full text. In contrast, records have been considered irrelevant when they dealt exclusively with early childhood, primary and adult audiences, or were unrelated to the topic framework. In addition, such records, which lacked components of science

education, or, were restricted to lecture-based teaching, or, were opinion articles, blogs, or theses that lacked empirical data.

5.4 Tools

To start with a set of search strings, a list of criteria for selection and a table of data extraction, were the three main tools of the review. Multi-source assessment means 360-degree, multi-source, multi-rater or multi-informant. So, this word was merged with the word assessment or evaluation and also with the words hands-on learning, science education, and middle school or teenagers. We did topic search in Web of Society using the same set of words while for ERIE/Education Source the words hands-on learning, science, peer, self and teacher assessment, and middle grades were chosen.

Two main criteria for excluding records were that they had to be peer-reviewed journal articles, conference proceedings, reviews, or meta-analyses, and their full texts had to be published in English. Screening consisted of two distinct phases. In the first phase, a check of the title and abstract was made, and records that were a good fit were moved to the next phase, where the full text was checked against the selection criteria. For each of the selected records, population and grade level, subject matter, pedagogical intervention, sources, types of assessment, outcome domains measured, and major findings were extracted.

5.5 Data Analysis

Because the studies involved differed in their design, participants, and outcome measures to some extent that prevented pooling of the statistical results and computing meta-analysis effect size, they decided not to do this. Instead, the data were processed using thematic analysis in three steps. First, the studies' results (outcomes) and the sources of information were identified (coded). Next, the codes produced were grouped into themes that were descriptive. Finally, analytical themes were built from these, which are the fundamentals of the framework presented. The use of two strands differs from the traditional one-strand PRISMA meta-analysis, and this limitation is mentioned in the Conclusion section.

6. Findings / Results

This review did not generate primary data; therefore, the findings of this study consist of themes generated during the thematic synthesis of the literature. Four sections follow below related to classroom practices: improving teaching methods, classroom management, student performance and training gaps. Each finding is connected to the research question.

6.1 Improvement in Teaching Methods

All the literature considered demonstrates a positive impact of the approaches based on experience and inquiry on the learning outcomes in comparison to traditional textbook-centred approach (Burch et al.,

2019; Ranken, 2025). Integrated STEM units (Anwar et al., 2022), citizen science education (Zhang et al., 2023), and field-based/nature-based programs (Carrier et al., 2013; Morag et al., 2013) demonstrate an improvement in conceptual knowledge, scientific practice, and sustained engagement, respectively. This is the answer to RQ1, as experiential approaches are associated with conceptual, procedural, and motivational outcomes.

However, there are two caveats to this conclusion. First, effectiveness depends on implementation fidelity, program length, and the type of assessment (Ranken, 2025). Second, the most effective programs, such as field-based programs, are hard to sustain within the middle school schedule because of time, staffing, and assessment limitations (Carrier et al., 2013). In addition, experiential approaches can be made even more effective through appropriate assessments. Assessment performed during the entire cycle of Kolb, reflection journals, peer observations, and project artifacts provide teachers with data needed to make changes during instruction, not only after.

6.2 Classroom Management

None of the included studies examined classroom management as the primary outcome of a multi-source assessment; therefore, this theme is inferential. These two points are supported by the literature. First, multi-source assessment is concentrated in higher education and clinical training, and documentation of its practical demands in ordinary middle-grade classrooms is very limited, especially regarding time, training, and rater burden. Second, the literature on formative, peer, and self-assessment indicates that these practices are effective when the criteria are explicit and the raters are trained and calibrated (Topping, 1998; Boud, 1995).

For classroom management, the synthesis points to several design needs. Instruments must be short and behaviorally anchored, with few scale points suited to young adolescents. The roles in group investigations must be clear enough for peers to observe and rate each other's contributions. Parent input must be brief and limited to what parents can observe at home. Digital data-collection tools can reduce the administrative load by collecting and displaying ratings from several sources. A phased introduction is advisable, beginning with self- and peer-assessment instruments that change teachers' routines least. These are design implications drawn from adjacent evidence and should be tested in pilot classrooms.

6.3 Student Achievement

Regarding achievement, the evidence is positive and rather limited. The effects reported are clustered in the realm of cognitive and achievement, using teacher-assessed or researcher-administered measures (Burch et al., 2019; Majumdar, 2026). Less consistent and generally using self-report data without confirmation from teachers, peers, and parents, are affective measures of interest, self-efficacy, and

attitudes, as well as socio-emotional and collaboration outcomes. It is precisely the latter, which are characteristic for experiential teaching, that receive the least triangulation. The multi-informant approach sheds light on this topic.

In comparative studies, there was moderate agreement between teachers, parents, and students in the evaluation of teenagers' social-emotional skills: raters agreed on externally observable behavior and disagreed on internal characteristics, which could be better assessed by students themselves (Mudarra et al., 2022). Averaging the scores of several teachers diminished measurement errors (Muschkin & Malone, 2007): In a comparison of self-, peer-, and instructor scoring of performance tasks, differences in scoring were systematically observed depending on the type of rater and student grade. These observations answer RQ2: the multi-rater approach is effective in terms of increasing the coverage of the construct and diminishing single-rater bias, but raters are not substitutable and need to be interpreted differently.

6.4 Training Gaps

Training comes up repeatedly as an issue in this review. One, the ability to interpret divergent ratings across various sources is something that is not common among teacher education programs cited in this literature review. Two, peer and self-rating involve training and calibration of students who rate their peers and themselves, which is rarely available in science class contexts. Three, tools for peer and parent reports that can assess science-based outcomes are much less available than those for teachers' and self-reports. Four, there is no information regarding teacher reactions to contradictory sources.

6.5 Synthesis: The Proposed Framework (RQ3)

In light of these topics, the review suggests an assessment framework in five outcome domains that uses a 360-degree approach: conceptual knowledge, science process and inquiry skills, affective and motivation outcomes, collaboration and social emotional skills, and real-world application (transfer) of learning. These five domains will be measured using a combination of self-, peer, and teacher assessments, parental feedback, and artifacts or performance (Table 1). Not all sources contribute to every domain. For example, parent feedback is not meant to measure content knowledge, whereas peer assessment does not apply to transfer.

Outcome domain	Self-assessment	Peer assessment	Teacher assessment	Parent/guardian input	Artifact / performance evidence

Conceptual / cognitive understanding	Reflective journal, concept map rating	Structured peer-review rubric on lab reasoning	Criterion-referenced content test, rubric-scored inquiry task	Not applicable / indirect	Lab report, project artifact scored against rubric
Science process / inquiry skills	Skills checklist self-rating	Peer observation during group investigation	Direct observation checklist, practical exam	Home-based experiment engagement log	Video-recorded practical task
Affective / motivational outcomes	Interest and self-efficacy scale	Peer perception of engagement	Teacher-rated engagement observation	Parent-rated interest and attitude at home	Portfolio reflection entries
Collaboration / social-emotional skills	Self-rated teamwork contribution	Peer rating of group contribution (multi-rater)	Teacher-rated collaboration rubric	Parent-reported social behaviour change	Group-process log / meeting notes
Transfer / real-world application	Self-reported application beyond class	Not typically applicable	Performance task in novel context	Parent-observed application at home/community	Community or citizen-science project output

Table 1. Proposed 360-degree multi-source assessment matrix for experiential middle-grade science education

Four implementation guidelines are associated with the matrix. The instruments must be distinguished as per rater, where the rating will be behaviorally anchored scales of 3 to 4 levels for students. The aggregation should be specific to the domain rather than an average. Teacher ratings and artifact ratings should carry more weight when seeking conceptual outcomes. For the internal state and out-of-class behavior, self, peer, and parent ratings should carry more weight. Raters' disagreements must be noted for teacher assessment and not averaged away. Peer and parent ratings can be affected by friendship, status, language, accessibility, and parental involvement. Therefore, they must be used only for diagnostic and judgmental purposes and not for grading purposes.

7. Discussion

The experiential science literature stresses outcome domains (conceptual, procedural, affective, collaborative, and transfer) that are difficult for a single teacher assessment to capture (Burch et al., 2019; Majumdar, 2026). It is mostly written exams and tests that an individual evaluator will work on. However, a student's cooperative efforts in an investigation, how a student's perception of science evolves, and even their utilization of the knowledge they have acquired will hardly be captured through this limited assessment method. In contrast, the multi-source literature developed in many different fields outside science K-12 already contains guidelines on rater triangulation and measurement design necessary for capturing all such outcomes (Tee & Ahmed, 2014). The framework presented here demonstrates how these concepts can be transferred to pre-teens and science education. Science educators of the experiential learning type realize what they seek to measure but have no way of effectively using these tools. The latter, multi-source assessment gurus, on the other hand, have tools but never implemented them in this context or age group.

Disagreements between the raters may come to light as a second issue. Typically, different ratings would be seen as a type of noise and should be limited as much as possible, with most of the history of measurement being focused on raters being in agreement. A contrario, multi-informant approach would rather argue against that view. Different people observing learners, teachers, parents, for example, would differ in their ratings to a certain extent because each rater is, simply put, witnessing the learner in a different setting (Mudarra et al., 2022). Imagine a student investigating an idea with tremendous energy in their house but at school, the same student stays silent in a structured group activity. The student will automatically get contradictory grades for the two activities and such grades will reflect where the student gets most involved and under which conditions. The discrepancy becomes a diagnostic element according to the frame it is a way of opening up the dialogue between teacher and pupil instead of being considered an imperfection which should be corrected. The position also resonates with the purpose of 360-degree feedback, which aims to help learners make their improvements over time rather than to merely arrive at one final result (Black & Wiliam, 1998; Tee & Ahmed, 2014).

The insights of this study reveal the relative value placed on different information sources. While self- and peer-assessments are helpful, they must be supported by well-defined rubrics and agreed-upon evaluation standards (Boud, 1995; Topping, 1998). Otherwise, these evaluations stand a good chance of being inconsistent, exaggerated, or simply reflecting the raters' personal inclinations rather than the merit of the work itself. Teacher evaluation continues to be the most trustworthy method for measuring students' knowledge of a subject. This is because only teachers are knowledgeable about the subject and professionally educated in its evaluation. The fact that involving more teachers reduces measurement errors (Muschkin & Malone, 2007) indicates that even within one teacher's evaluations, multiple observations lead to a stronger conclusion. Parental input becomes reasonable when it focuses only on a parent's observable contribution. That is, their child's behavior and attitude at home, e.g., curiosity, persistence, and

interest in science-related activities. To ask parents to evaluate scientific understanding would go beyond what they can realistically see and, hence, this would be adding extra load without increasing the value.

The most effective use of each input was determined by the type of information provided by the input. However, certain limitations must be considered when interpreting these findings. Much of the investigation into multisource assessment has taken place at different stages, especially in universities, healthcare setting exams, and socio-emotional assessments. Therefore, the use of the findings of such research in middle school science would be an extension, and the components of the proposed framework need to be treated as hypotheses to be tested rather than as findings that are already settled. The learning by doing literature itself is inconsistent in its implementation of strategies and measures of outcomes (Ranken, 2025), which makes it challenging to conduct a cross-study comparison and hampers the robustness of the overall evidence generated.

Feasibility seems to be the biggest challenge. Teachers currently face a lot of time pressure; thus, a multi-source assessment could significantly increase their workload. In addition, a large number of rating instruments that students and parents need to complete may lead to rating fatigue, which affects not only their responses but also the level of engagement and the quality of the feedback. There could also be a bias element, either from close friendships among students or from parent–child relationships. Any of these issues could lead to underperformance in terms of the implementation of the framework; therefore, they all require direct examination during pilot testing.

One protective measure could be to not make peer or parent ratings a part of the final grade, which would lower the intensity of the situation and reduce the tendency to overrate. Teachers are required to interpret conflicting findings as an extra step in the assessment process.

If a teacher reads through the situation carefully and takes it seriously, they would have to make a good guess of each student. That takes the time and the skill. Therefore, the suggestions given later put particular attention to the professional development of teachers so that the teacher can take the help of the differences among different assessments to make the right decisions and give the students some support to learn better.

8. Conclusion

This systematic review analyzed the existing literature on the results of experiential science education for middle-stage learners, as well as on multi-source and 360-degree assessment in education and related fields. The review identified a consistent pattern showing the advantages of experiential teaching practices in relation to conceptual knowledge, inquiry skills, and motivation. In addition, the persistent use of a single-source, achievement-oriented approach to assessment was identified despite the complex nature of the learning results created through experiential education. Multi-source assessment was shown to increase

construct coverage and minimize single-rater bias, although its implementation in K-12 science classes was scarce.

Based on the findings, the review provided a model of 360-degree multi-source framework including five dimensions of outcomes and four sources of raters' judgments along with artifact-based evidence with recommendations about instrumentation, dimension-specific aggregation, feasibility, and equity considerations. The key assumption of the proposed framework is that raters' disagreement carries information and needs to be interpreted rather than averaged.

A narrative synthesis of two studies was conducted, as there were no pooled effect sizes; thus, a strictly defined single-strand PRISMA meta-analysis was not performed. Retaining two strands means that some degree of heterogeneity can be admitted. It is unlikely that a search across three databases and snowballing would provide an exhaustive collection of grey literature, non-English sources, and very recent sources. Several elements of the framework were borrowed from other areas. The framework remains conceptual and not validated through empirical research. Empirical piloting of the matrix in middle school science classes, development and validation of psychometric properties of peer/parent rating measures of science outcomes, research on teachers' perception of discrepancies between multiple sources of information, and longitudinal studies of the prediction of future science engagement and achievement by multiple sources are important topics for further study.

9. Recommendations

The curriculum and assessment group must develop unit-level assessments for the hands-on science units based only on the framework given earlier. It may add multi-source assessments as a sort of badly designed addition to existing assessments. However, this would not make sense in many cases, especially where you may already have a variety of other means of assessment. The students may feel more overwhelmed by the variety of methods, and thus the learning outcomes may be affected. Professional development programs and teacher training courses should cover ways of dealing with multi-source assessment results, including rater training and the use of calibration tools. Implementing the framework should be done gradually, starting with self- and peer-assessments, and teachers should be given tools or means of collection and presentation of ratings from different sources as means of reducing the teachers' work-load, especially if they have to collect ratings on hundreds of students per grade level. The school policy on grading and reporting should be such that the use of multi-source information is separate from that used for formative or diagnostic purposes. This is because peer and parent-report sources could have equity and bias issues; therefore, those sources of information should only be used for summative purposes. The scientists involved in the development and validation of multi-rater measures should focus on the development of instruments that are specific not only in the age of the target subjects but also in the desired outcome, especially for the

middle-school science outcomes. In addition, they should ensure that pilot tests cover control and identification of rater patterns among various student subgroups, which is quite a common phenomenon.

10. References

Anwar, S., Menekse, M., Guzey, S., & Bryan, L. A. (2022). The effectiveness of an integrated STEM curriculum unit on middle school students' life science learning. *Journal of Research in Science Teaching, 59*(7), 1204–1234. https://doi.org/10.1002/tea.21756

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Boud, D. (1995). *Enhancing learning through self-assessment*. Routledge.

Burch, G. F., Giambatista, R., Batchelor, J. H., Burch, J. J., Hoover, J. D., & Heller, N. A. (2019). A meta-analysis of the relationship between experiential learning and learning outcomes. *Decision Sciences Journal of Innovative Education, 17*(3), 239–273.

Carrier, S., Tugurian, L., & Thomson, M. (2013). Elementary science indoors and out: Teachers, time, and testing. *Research in Science Education, 43*, 2059–2083.

Chan, H. H.-K., Kwong, H. Y. C., Shu, G. L. F., Ting, C. Y., & Lai, F. H.-Y. (2021). Effects of experiential learning programmes on adolescent prosocial behaviour, empathy, and subjective well-being: A systematic review and meta-analysis. *Frontiers in Psychology, 12*, Article 709699.

Dewey, J. (1938). *Experience and education*. Kappa Delta Pi.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice-Hall.

Majumdar, S. (2026). Examining the effects of experiential science education in middle school classrooms. *American Journal of STEM Education, 20*, 201–226. https://doi.org/10.32674/m38kqg15

Morag, O., Keren-Rotem, T., & Tal, T. (2013). Long-term educational programs in nature parks: Characteristics, outcomes, and challenges. *International Journal of Environmental and Science Education, 8*, 427–449.

Morris, T. H. (2020). Experiential learning – A systematic review and revision of Kolb's model. *Interactive Learning Environments, 28*(8), 1064–1077.

Mudarra, M. J., Álvarez-González, B., García-Salguero, B., & Elliott, S. N. (2022). Multi-informant assessment of adolescents' social–emotional skills: Patterns of agreement and discrepancy among teachers, parents, and students. *Behavioral Sciences, 12*(3), Article 62. https://doi.org/10.3390/bs12030062

Muschkin, C. G., & Malone, P. S. (2007). Multiple teacher ratings: An evaluation of measurement strategies. *Educational Research and Evaluation, 13*(1), 71–83. https://doi.org/10.1080/13803610601058215

PRISMA. (2020). Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020. *BMJ, 372*, Article n71.

Ranken, E. (2025). The effect of experiential learning on academic achievement of children aged 4–14: A rapid evidence assessment. *The Curriculum Journal*. Advance online publication. https://doi.org/10.1002/curj.304

Tee, D. D., & Ahmed, P. K. (2014). 360 degree feedback: An integrative framework for learning and assessment. *Teaching in Higher Education, 19*(6), 579–591.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Zhang, E. Y., Hundley, C., Watson, Z., Farah, F., Bunnell, S., & Kristensen, T. (2023). Learning by doing: A multi-level analysis of the impact of citizen science education. *Science Education, 107*(5), 1324–1351.