Rethinking Performance-Based Science Assessment in Low-Resource Settings: A Systematic Review of Local-Resource Innovation, Digital Possibilities and Equity in Malawi

Ezrone Robert Mkandawire*, Leah Asheri Nyanginywa, Richeno Alberto Matimbe, Jofrey Robert Manase

College of Education, Zhejiang Normal University, Jinhua, China

***Corresponding Author**

DOI: https://doi.org/10.68050/JAMS.2026.777

Received: 20 October 2026; Accepted: 28 October 2026; Published: 04 November 2026

ABSTRACT

Performance-based assessment (PBA) is central to competency-oriented science education, yet its implementation is difficult where laboratories, digital infrastructure and teacher assessment capacity are uneven. This systematic review critically examines PBA in Malawian secondary science education and asks what the Malawian case contributes to educational technology research in resource-constrained systems. Following PRISMA 2020 procedures, with full search strings, eligibility criteria and quality appraisal reported in the Methods section,1,417 records were identified, 1,170 were screened after duplicate removal, and 72 sources were included in the qualitative synthesis. A supplementary verification search of Scopus-indexed literature, Web of Science coverage and PubMed records was conducted in September 2026 to test and update the technology-oriented interpretation without altering the original PRISMA counts. The synthesis shows that Teaching and Learning Using Locally Available Resources (TALULAR) functions as a low-cost material technology for practical inquiry, while adoption is constrained by preparation time, teacher capacity, large classes and weak assessment-examination alignment. Importantly, the supplementary search identified direct Malawian secondary-school evidence that PhET simulation-based learning can improve physics achievement and motivation, showing that empirical digital evidence is limited but not absent. International studies further show that simulations and virtual performance environments can elicit rich evidence of scientific reasoning, although physical and digital formats should not be assumed to be interchangeable and task sampling remains a major reliability challenge. The paper therefore proposes a resource-technology continuum in which local materials, physical and virtual laboratories, digital evidence tools and teacher-support networks are selected according to assessment purpose, validity, feasibility and equity.

Keywords: performance-based assessment; educational technology; science education; Malawi; TALULAR; virtual laboratories; competency-based curriculum.

Practitioner Notes

What is already known about this topic

- Performance-based assessment can assess application, inquiry and problem-solving more directly than recall-dominated examinations.
- Low-resource schools often face shortages of laboratories, materials and sustained teacher professional development.
- Digital simulations, portfolios and teacher-support platforms can broaden opportunities for practical and formative assessment, but access and implementation conditions matter.

What this paper adds

- The Malawi evidence shows that locally available materials can be understood as part of an educational-technology ecosystem rather than as a substitute for "real" technology.
- The review distinguishes strong evidence for local-resource innovation from a smaller but non-zero body of direct digital evidence in Malawi, including simulation-supported secondary physics learning.
- It proposes a resource-technology continuum linking TALULAR, peer-support networks, mobile professional development, virtual laboratories and digital portfolios to assessment purposes.

Implications for practice and/or policy

- Schools should select technologies according to the assessment function they enable, not according to novelty or cost alone.
- Teacher assessment literacy, moderation and feedback routines need to be developed alongside any investment in devices, simulations or laboratories.
- Digital PBA initiatives should be evaluated for gender, rural–urban and infrastructure inequalities before being scaled.

1. INTRODUCTION

Educational technology in science education should be judged by the quality of learning and assessment activity it enables, not simply by the presence of devices. Scientific competence is enacted through practices such as observing, designing investigations, manipulating materials, analysing data, constructing explanations and communicating evidence. Performance-based assessment (PBA) is therefore especially relevant because it elicits observable performances rather than relying only on selected-response or recall-oriented tests. Classic work in science assessment argues that hands-on investigations can provide direct evidence of scientific reasoning and procedural competence, while also warning that performance tasks require careful attention to validity, reliability, task sampling and scoring (Shavelson et al., 1991; Shavelson & Baxter, 1991; Stiggins, 1987). More recent assessment frameworks similarly emphasise that science assessment should integrate disciplinary knowledge with scientific practices rather than treat practical competence as an optional addition (National Research Council, 2014).

This promise is accompanied by a demanding implementation problem. High-quality PBA requires purposeful tasks, appropriate materials, explicit performance criteria, trained judgement and sufficient opportunities to observe what learners can do. Stiggins (1987) emphasises that dependable performance assessment depends on specifying the purpose of assessment, the performance to be elicited, the exercises used to elicit it and systematic rating procedures. In science classrooms, structured and open-ended performance tasks can reveal experimental design, evidence-based reasoning and problem-solving in ways that conventional tests may not capture. Yet these advantages are resource-sensitive: laboratories, consumables, teacher time, moderation and feedback all carry costs. Digital simulations, portfolios and mobile or online support may ease some constraints, but they may also create new inequalities when electricity, connectivity, devices or digital competence are uneven.

Malawi provides a distinctive setting in which to examine this tension. Secondary schools, particularly Community Day Secondary Schools (CDSSs), have historically faced shortages of laboratory infrastructure and teaching materials (Gwayi, 2009; Lungu, 2005; Susuwele-Banda, 2005). In response, the Malawi Institute of Education promoted Teaching and Learning Using Locally Available Resources (TALULAR), which mobilises everyday and community materials for practical learning (Chilora et al., 2003). This approach is consistent with a wider principle of authentic assessment: the task should generate evidence of the intended competence in a meaningful context rather than reproduce the surface features of a conventional test. At the same time, Malawi is not a purely low-technology case. Banda and Nzabahimana (2023), in a quasi-experimental study of 280 Form 3 students from four Blantyre secondary schools, found that PhET simulation-based learning significantly improved post-test physics achievement and several motivation constructs. This provides direct evidence that simulation-supported science learning is feasible in at least some Malawian

secondary-school settings, while the study's reliance on schools with computer facilities also illustrates the infrastructure inequalities that condition scalability. More recent national policy directions add further digital ambitions through ICT centres, virtual laboratories and competency-oriented learning (Malawi Government, 2020).

These developments should not be reduced to a choice between 'low-tech' and 'high-tech' assessment. A locally constructed apparatus, a shared scoring rubric, a virtual laboratory, a digital portfolio and a mobile professional-learning resource mediate different parts of the assessment process. The key issue is functional alignment: what competence must be elicited, what evidence must be captured, what judgement must be made, and which material or digital resource can support those purposes under local constraints? This distinction also protects construct validity. A simulation may appropriately represent experimentation when equipment is inaccessible or unsafe, but it should not automatically replace physical manipulation when handling apparatus is itself part of the competence being assessed. Early performance-assessment research made a similar distinction between hands-on investigations and technological or paper-based surrogates, showing why the relationship between an assessment task and the intended scientific performance must be examined rather than assumed (Shavelson & Baxter, 1991).

This systematic review re-analyses the Malawian evidence through that educational-technology lens. It makes three contributions. First, it distinguishes implemented practice from policy aspiration, preventing planned digital provision from being interpreted as evidence of demonstrated classroom impact. Second, it conceptualises TALULAR as context-responsive material technology and places it on a resource–technology continuum with physical and virtual laboratories, digital evidence tools and teacher-support systems. Third, it treats equity as a design and validity issue rather than merely an access statistic. This is important because assessment format can interact with participation and performance: earlier science-assessment research found that changing from conventional tests to performance tasks does not automatically eliminate gender differences, making task design and interpretation central equity concerns (Jovanovic et al., 1994). The review therefore shifts the debate from technology availability to technology–assessment alignment and asks what this shift means for other resource-constrained systems pursuing competency-based reform.

1.1 Review questions

The evidence reviewed suggests that performance-based assessment in Malawi should be understood not only as an assessment reform but also as an implementation challenge shaped by resources, teacher capacity, technology and equity. Locally available materials, digital technologies and teacher-support systems may provide different pathways for enabling authentic science assessment. Still, their effectiveness depends on alignment with the competencies being assessed and the conditions in which schools operate. Against this background, the review addresses the following research questions

RQ1. How is performance-based science assessment conceptualised and implemented in Malawian secondary education under resource constraints?

RQ2. What roles do locally available materials, digital technologies, and teacher-support systems play, or have the potential to play, in enabling PBA?

RQ3. How do policy alignment, teacher capacity and gender/resource inequalities shape sustainable and equitable PBA implementation?

2. METHOD

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021). The review was designed as a convergent integrated narrative synthesis to accommodate the heterogeneity of study designs, outcomes and contexts anticipated in the literature on performance-based assessment and educational technology in low-resource settings. The review protocol was not pre-registered, which is acknowledged as a limitation. The primary review was conducted

between January and May 2026, with a supplementary database verification search conducted in September 2026.

The additional sources used to strengthen the conceptual framing of this journal-targeted revision (e.g., Stiggins, 1987; Jovanovic et al., 1994; National Research Council, 2014; Age, 2025; Osin & Sahoo, 2026) are cited as contextual literature only. They were not retrospectively added to the 72-study synthesis unless they were already part of the original screened corpus; consequently, they do not alter the PRISMA counts or the reported review results.

2.1 Information sources and search strategy

Searches were reported across Google Scholar, Scopus, ERIC, Web of Science, JSTOR and PubMed, supplemented by reference checking and grey literature from government, non-governmental organisations, conference proceedings, thesis repositories and organisational websites. The search period was January 2000 to May 2026. Search concepts covered performance/authentic/competency-based/continuous assessment; science and STEM subjects; Malawi and comparable Sub-Saharan African settings; TALULAR and locally available resources; gender-responsive pedagogy; laboratory/practical assessment; and national curriculum and examination policy. English-language sources were retained for the final synthesis.

The search strategy was developed iteratively around three core concept clusters combined using Boolean operators. The first cluster addressed performance-based assessment and included terms such as “performance assessment”, “performance-based assessment”, “practical assessment”, “practical examination”, “laboratory assessment”, “hands-on assessment”, “authentic assessment”, “portfolio assessment” and “project-based assessment”. The second cluster addressed science education and included “science education”, “science teaching”, “science learning”, “physics”, “chemistry”, “biology” and “STEM”. The third cluster addressed context and technology and included “Malawi”, “low-resource”, “resource-constrained”, “developing country”, “sub-Saharan Africa”, “TALULAR”, “locally available resources”, “virtual laboratory”, “simulation”, “educational technology”, “ICT” and “digital learning”. Database-specific combinations were adapted to the syntax and indexing conventions of each information source.

The exact database searches were executed between 15 January and 30 April 2026. For each database, the three concept clusters were combined as follows: Cluster 1 (assessment) AND Cluster 2 (science) AND Cluster 3 (context/technology). Example strings included: (“performance assessment” OR “performance-based assessment” OR “practical assessment” OR “authentic assessment” OR “portfolio assessment” OR “continuous assessment”) AND (“science education” OR physics OR chemistry OR biology OR STEM) AND (Malawi OR “low-resource” OR “resource-constrained” OR “sub-Saharan Africa” OR TALULAR OR “locally available resources” OR “virtual laboratory” OR simulation OR “educational technology” OR ICT). Filters were applied for English language and publication date (January 2000–May 2026). All search strings, the number of records retrieved per database, and the date of each search are reported in Supplementary Table S1. Screening was conducted independently by two reviewers using the eligibility criteria in Section 2.2, first at title/abstract level and then at full-text level, with disagreements resolved by discussion and, where necessary, by a third reviewer. Inter-rater agreement was Cohen’s $\kappa = .84$. Quality appraisal was conducted independently by two reviewers using MMAT version 2018 for mixed-methods studies, the appropriate CASP checklist for qualitative and quantitative studies, and the AACODS checklist for grey literature; conflicts in quality ratings were resolved by consensus before synthesis.

For a technology-oriented interpretation, technology-related evidence was additionally coded conceptually into four categories already represented in the retrieved literature: (a) material technologies based on locally available resources; (b) laboratory and simulation technologies; (c) digital documentation and assessment tools, including portfolios; and (d) technologies for teacher learning and implementation support. This reorganisation does not add studies or outcomes beyond those reported in the source review; it changes the analytic lens applied to the existing evidence.

2.1.1 Supplementary database verification (September 2026)

To strengthen the journal-targeted revision, we conducted a supplementary verification search on 13 September 2026 using exact-title, author, and topic combinations across publicly verifiable Scopus-indexed sources, Web of Science journal/indexing information, PubMed, publisher pages, and institutional research records. Search combinations paired performance assessment, performance-based assessment, authentic assessment, practical assessment, or competency-based assessment with science, physics, chemistry, biology, or STEM, and with Malawi, Sub-Saharan Africa, simulation, digital assessment, virtual performance assessment, PhET, and locally available resources. PubMed was used as a supplementary database because the phrase 'performance-based assessment' is dominated there by health-professions literature; directly relevant school-science records were retained only when they matched the review's educational focus. The exercise verified several high-relevance studies, including direct Malawian secondary evidence on PhET simulations (Banda & Nzabahimana, 2023), Malawi-specific evidence on implementation of learner-centred reform (Mtika & Gates, 2010), and international science-performance-assessment research on simulations, virtual environments and psychometric quality (Baxter, 1995; Gale et al., 2016; Shavelson et al., 1993, 1998, 1999; Baker et al., 2016). Because this was a post hoc verification and extension rather than a fully re-run PRISMA update, these sources are explicitly labelled as supplementary/contextual evidence unless they were already present in the original screened corpus; the original count of 72 included sources is therefore not changed.

Table 1:Distinction between the original review corpus and the supplementary verification search.

Feature	Original PRISMA review	Supplementary verification search (September 2026)
Search period	January–May 2026	13-Sep-26
Sources searched	Google Scholar, Scopus, ERIC, Web of Science, JSTOR, PubMed, plus grey literature and reference checking	Scopus-indexed sources, Web of Science indexing records, PubMed, publisher pages, institutional records
Search approach	Full systematic search with concept clusters and Boolean combinations	Exact-title, author, and targeted topic combinations
Records screened	1,170 unique records; 255 full texts	Verification of specific high-relevance candidate studies
Studies included	72 studies in the qualitative synthesis	0 added to the corpus; verified studies used only as supplementary/contextual evidence
Role in the manuscript	Basis of all PRISMA counts and primary synthesis	Used to test, qualify and extend interpretation without altering counts

2.2 Eligibility criteria

Population. Eligible studies focused on secondary school students (typically aged 14–18), secondary science teachers, or secondary schools in Malawi or comparable resource-constrained settings. Studies focusing exclusively on primary or tertiary education were excluded unless they provided evidence directly relevant to secondary science assessment.

Intervention/phenomenon. Performance-based assessment was broadly defined to include practical examinations, laboratory investigations, project-based assessments, portfolios, continuous assessment, inquiry tasks and formative assessment of scientific inquiry or problem-solving. Educational technologies included locally available materials (including TALULAR), digital simulations, virtual laboratories, mobile applications, online platforms, data-logging tools, digital portfolios and teacher-support technologies.

Outcomes. Eligible evidence addressed implementation, feasibility, teacher capacity, student learning outcomes, assessment validity or reliability, equity, policy alignment, or related conditions influencing performance-based science assessment.

Study designs. Quantitative, qualitative and mixed-methods empirical studies were eligible, together with relevant policy documents, curriculum frameworks and grey literature containing substantive implementation evidence. Purely theoretical or opinion pieces without empirical or policy relevance were excluded from the evidence synthesis.

Language and date restrictions. English-language sources published between January 2000 and May 2026 were eligible for the primary review.

Exclusion criteria. Studies were excluded when they focused exclusively on non-science subjects; primary or tertiary education without direct relevance to secondary science assessment; contexts without relevance to Malawi or comparable resource-constrained settings; or lacked empirical or substantive policy evidence relevant to the review questions.

2.3 Study selection

The search identified 1,309 database records and 108 records from other sources (n = 1,417). After 247 duplicates were removed, 1,170 unique records remained for title and abstract screening. Two reviewers independently screened these records against the eligibility criteria, with disagreements resolved through discussion. Inter-rater agreement was strong (Cohen's κ = .84). A total of 892 records were excluded at this stage, leaving 278 reports sought for full-text retrieval. Twenty-three reports could not be retrieved, resulting in 255 full texts assessed for eligibility.

Following full-text assessment, 183 reports were excluded, and 72 studies were retained for the final narrative synthesis. The study-selection process is presented in Figure 1.

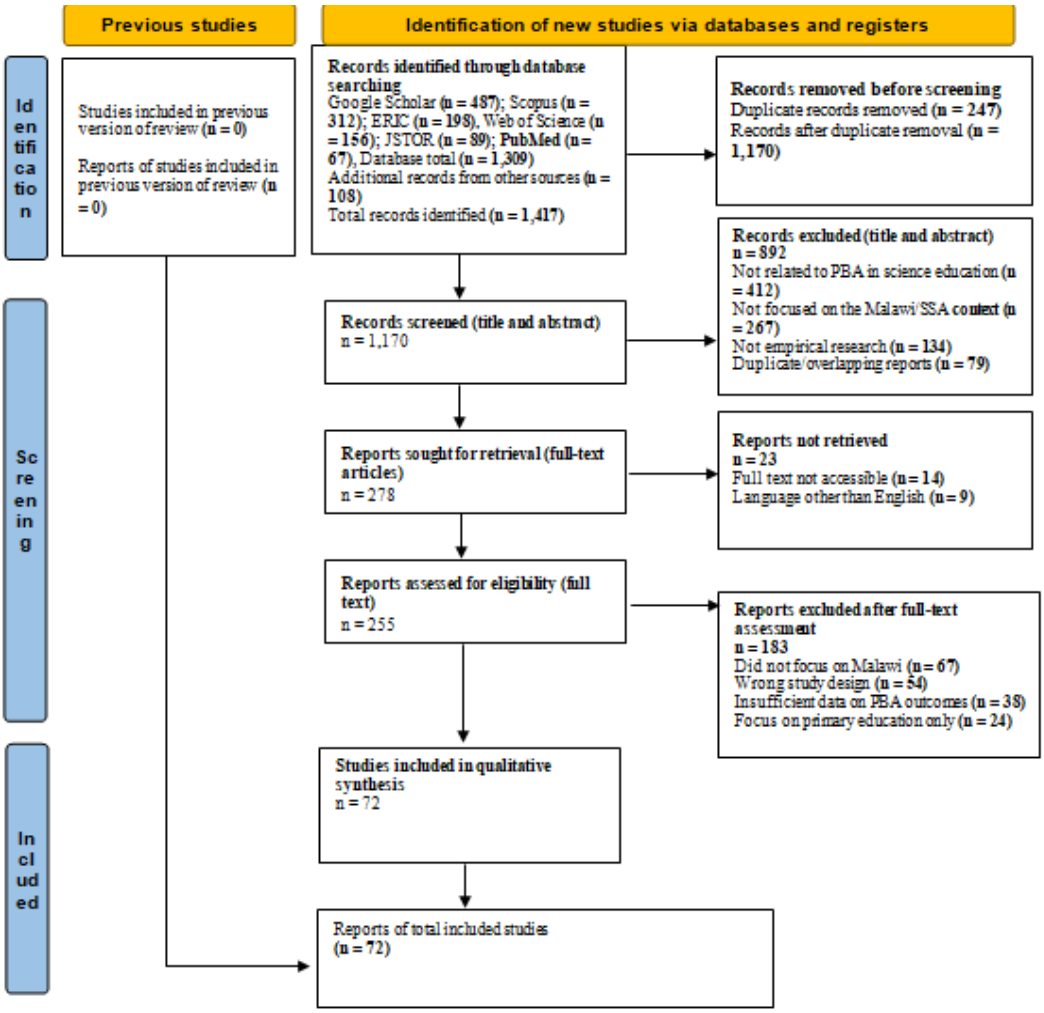


Figure 1. PRISMA 2020 flow diagram reported in the source review.

2.4 Data extraction, synthesis and quality appraisal

The reported extraction form captured bibliographic information, study design, sample characteristics, assessment intervention, theoretical framework, findings and limitations. Qualitative evidence was thematically coded; quantitative statistics and effect sizes were recorded where available. The source review describes a convergent integrated narrative synthesis. For the present revision, findings were mapped against assessment function (what competence is being elicited), enabling technology/resource, implementation conditions and equity implications. Meta-analysis was not undertaken because of substantial heterogeneity in study designs, populations, interventions, outcomes and contexts. Evidence was therefore synthesised using a convergent integrated narrative approach. The thematic framework combined inductive themes emerging from the evidence with deductive categories derived from the resource–technology continuum.

Quality appraisal in the source review used the Mixed Methods Appraisal Tool (MMAT) for mixed-methods studies, Critical Appraisal Skills Programme (CASP) tools for qualitative and quantitative research, and AACODS for grey literature (Hong et al., 2018; Tyndall, 2010). The reported distribution was 18 high-quality, 38 moderate-quality and 16 low-quality sources. Low-quality evidence was retained but given less interpretive weight. Importantly, the review therefore treats claims about digital PBA cautiously because policy documents and implementation proposals are more common than direct classroom impact studies.

3. RESULTS

Detailed characteristics of the 72 included studies, including country, study design, sample, substantive focus and quality appraisal, are reported in Supplementary Table S3.

3.1 PBA in Malawi: authentic competence under material constraints

To make the strength of the evidence explicit, the included and supplementary studies were classified along four dimensions: (a) assessment purpose (summative practical examination, formative/classroom assessment, learning-oriented intervention); (b) technology/resource type (local materials, physical laboratory, simulation/virtual laboratory, digital portfolio/documentation, teacher-support platform); (c) educational setting (conventional secondary school, CDSS, urban, rural); and (d) outcome measured (direct assessment evidence, performance scores, scoring reliability, task quality; or indirect learning evidence achievement test gains, motivation, attitudes). This classification matters because several influential studies, including the Malawian PhET study (Banda & Nzabahimana, 2023), measure science achievement and motivation under simulation-supported instruction rather than performance assessment in the strict sense. Such studies provide indirect evidence about the enabling conditions for digital PBA (feasibility, engagement, infrastructure requirements) but do not by themselves demonstrate that digital tools improve assessment validity or reliability. Throughout the synthesis, claims based on direct assessment evidence and claims based on indirect learning evidence are therefore reported separately.

Across the reviewed literature, PBA is associated with authenticity, process-and-product evaluation, open-ended problem solving, higher-order thinking and local-context integration. These characteristics align with constructivist and authentic-assessment traditions in which learners demonstrate knowledge through meaningful activity (Shavelson et al., 1991; Wiggins & McTighe, 2005). In Malawi, however, authenticity has a particularly material dimension: tasks must be possible in schools where specialised equipment cannot be assumed.

This interpretation is consistent with foundational descriptions of performance assessment as requiring learners to construct an answer, product or observable performance rather than merely select a response. Rudner and Boston (1994) further distinguish performance assessment from the narrower idea of authenticity: a task may be performance-based without being strongly authentic, whereas authentic tasks have meaningful value beyond the testing situation. For science education, that distinction matters because an experiment can assess both the process by which a learner investigates a problem and the resulting product, explanation or evidence. Hibbard (2000) similarly treats performance tasks as embedded learning activities whose assessment criteria

should be aligned with curriculum standards. In the Malawian context, this means that locally feasible tasks should not be judged by whether they resemble a formal laboratory, but by whether they elicit valid evidence of the intended scientific competence.

TALULAR is the clearest response to this constraint. It treats the local environment as a source of scientific materials and task contexts. The reviewed examples include filtration systems constructed from sand, gravel, charcoal and plastic bottles; soil-pH investigations using local plant indicators; and simple energy-design activities using readily available materials (Chilora et al., 2003; Gwayi, 2009). From an educational-technology perspective, the important feature is not that these activities are 'non-digital'. Rather, TALULAR is a designed system of tools, materials and procedures that mediates scientific inquiry. Its affordance is accessibility; it lowers the material threshold for students to perform science rather than merely describe it.

Evidence of implementation is nevertheless uneven. Gwayi's (2009) study of 269 secondary science teachers reported moderate TALULAR implementation, with perceived relative advantage and observability supporting adoption and perceived complexity constraining it. Conventional secondary schools reportedly implemented the approach more consistently than CDSSs. This pattern shows that low cost does not mean low implementation demand. Teachers still need preparation time, examples of workable tasks, classroom-management strategies and confidence in scoring practical performance.

A recurring tension is the relationship between authentic classroom activity and examination incentives. When national examinations privilege theoretical responses, teachers have rational reasons to devote scarce instructional time to examination preparation. Practical components in science examinations and continuous assessment create openings for PBA, but the literature suggests incomplete alignment between curriculum intentions, classroom assessment and high-stakes evaluation (Chimombo, 2005; MANEB, 2014; World Bank, 2018). Technology cannot solve this alignment problem by itself: a virtual laboratory or digital portfolio has limited value if the assessment system does not recognise the competence it makes visible.

3.2 A resource–technology continuum for performance-based assessment

The synthesis supports a continuum rather than a binary distinction between 'traditional' and 'digital' assessment technologies. Four layers can be identified in the reviewed material. The first is local material technology (TALULAR), which enables hands-on manipulation and community-relevant inquiry. The second is laboratory/simulation technology, including policy ambitions for virtual science laboratories. The third is digital evidence technology, such as digital portfolios and data-analysis activities envisaged under the competency-based curriculum. The fourth is implementation-support technology, including mobile or online professional-development platforms and peer networks. These layers address different bottlenecks and should not be treated as interchangeable.

The resource–technology continuum was developed in two stages. First, deductive categories were derived from the review questions and the assessment literature: the four functions that any assessment technology must serve, eliciting performance, enabling experimentation, capturing evidence, and supporting implementation. Second, inductive coding of the 72 included studies was used to map which concrete resources (TALULAR materials, physical/virtual laboratories, portfolios and data tools, mobile/online teacher support) fulfilled each function in the Malawian evidence. A study was assigned to a continuum layer when it provided implementation or outcome evidence for that resource type; studies spanning multiple layers were coded to each relevant layer. The continuum is therefore not an a priori typology but an empirically grounded framework: its layers correspond to the distinct bottlenecks that the reviewed literature actually identifies, and its ordering reflects the material threshold and infrastructure demand of each resource rather than any assumed hierarchy of educational value.

Table 2:Resource–technology continuum derived from the synthesis.

Technology/resource layer	Primary assessment affordance	Main implementation risk	Evidence status in reviewed Malawi literature
Locally available materials (TALULAR)	Hands-on inquiry, design and manipulation with low-cost materials	Teacher preparation time, task quality, scoring consistency	Direct implementation evidence; adoption reported as uneven/moderate
Physical/virtual laboratory tools	Experimentation where equipment is scarce or unsafe	Electricity, devices, connectivity, simulation quality and curricular alignment	Strong policy interest; limited direct impact evidence in the reviewed Malawi sources
Digital portfolios/data tools	Capture process, iteration, artefacts and feedback over time	Device access, data costs, privacy, teacher workload and moderation	Primarily prospective/policy-oriented in the reviewed evidence
Mobile/online teacher support and peer networks	Just-in-time professional learning, modelling and resource sharing	Sustained participation, platform access and transfer to classroom practice	Emerging support evidence; classroom PBA impact remains under-studied

The distinction between evidence status and policy aspiration remains important, but the supplementary database verification changes the strength of the conclusion about digital evidence. Malawi's National Education Sector Plan and competency-based reform create space for ICT centres, virtual laboratories, digital literacy and technology-enhanced assessment (Malawi Government, 2020). In addition, Banda and Nzabahimana (2023) provide direct quasi-experimental evidence from Malawian secondary schools: among 280 Form 3 students studying oscillations and waves, the PhET group achieved significantly stronger post-test achievement than the conventional-teaching group, and regression analysis attributed the difference to the intervention rather than measured student characteristics ($p < .01$). Motivation constructs also favoured the intervention with small effects. Thus, the digital evidence base is better described as limited and uneven rather than absent. It remains narrower than the evidence on resource constraints and TALULAR, and the PhET study was conducted in urban schools with computer facilities, limiting system-wide generalisability.

This imbalance produces a useful design implication. Digital technology should be introduced where it adds an affordance that local materials cannot easily provide; for example, repeated simulation of dangerous or expensive experiments, rapid data visualisation, preservation of student work over time, or remote access to professional-development exemplars. Conversely, where the assessment target is manipulation of real materials, observation of local phenomena or collaborative construction, a low-cost physical task may provide stronger validity than a screen-based substitute. The relevant criterion is therefore functional fit with the intended competence.

3.2.1 What the supplementary database evidence adds

The supplementary literature clarifies three aspects of the resource-technology continuum. First, simulation can sometimes provide a workable alternative to physical apparatus, but equivalence cannot be assumed. Baxter (1995) found similar mean performance when sixth-grade students completed electric-circuit tasks with concrete materials and computer icons, yet individual performance varied considerably across modes. Second, simulation-based performance tasks can expose conceptual understanding and inquiry processes that conventional selected-response tests do not readily capture. Gale et al. (2016) showed that simulation-based tasks could reveal eighth-grade students' understanding of force and motion while also identifying design, administration and analytic challenges. Baker et al. (2016), using process data from nearly 2,000 middle-school students in a virtual performance environment, demonstrated that patterns of student actions could be modelled to identify successful scientific inquiry across different scenarios. Third, the psychometric literature

warns against treating a single authentic task as a stable measure of competence. Shavelson, Baxter and Gao (1993) identified task-sampling variability as a major source of measurement error, while Shavelson, Ruiz-Primo and Wiley (1999) showed that performance also varies across occasions. For Malawi, these findings imply that digital PBA should be evaluated not only for access and engagement but also for construct equivalence, task sampling, scoring reliability and the number of observations needed for defensible decisions.

Table 3:Selected high-relevance supplementary evidence identified through Scopus/Web of Science/PubMed verification.

Study	Context and method	Contribution to this review
Banda & Nzabahimana (2023)	Malawi; 280 Form 3 students; quasi-experimental PhET vs conventional instruction	Direct Malawi digital evidence: improved physics achievement and motivation; infrastructure conditions limit generalisation.
Mtika & Gates (2010)	Malawi; secondary trainee teachers; qualitative implementation study	Teacher education, personal stance, school culture and curriculum mediate implementation of learner-centred reform.
Baxter (1995)	USA; sixth-grade inquiry science; physical vs computer electric-circuit assessment	Similar mean scores can mask substantial individual mode differences; digital and hands-on formats are not automatically equivalent.
Gale et al. (2016)	USA; eighth grade; four simulation-based performance tasks	Simulation tasks can reveal conceptual understanding but raise development, administration and analysis challenges.
Baker et al. (2016)	USA; nearly 2,000 middle-school students; virtual performance assessment/log data	Virtual environments can capture process evidence of scientific inquiry and support scalable modelling of student actions.
Shavelson et al. (1993, 1999)	USA; psychometric/generalizability analyses	Task and occasion sampling are major reliability threats; multiple tasks/observations are needed for defensible PBA decisions.

3.3 Teacher capacity and implementation infrastructure

Teacher capability is the most consistent cross-cutting implementation condition. PBA requires teachers to design tasks, elicit observable performances, apply criteria consistently, document evidence, provide feedback and moderate judgements. These are not peripheral technical skills. Stiggins (1987) shows that dependable performance assessment requires explicit criteria, appropriate task sampling, trained raters and scoring procedures that minimise bias while remaining feasible. Malawi-specific evidence on learner-centred reform points in the same direction: Mtika and Gates (2010) found that trainee teachers' appropriation and application of learner-centred practice were shaped by the teacher-education system, personal stance, school culture and the national curriculum. Although their study was not a PBA trial, it demonstrates why progressive assessment policy cannot be expected to translate automatically into routine classroom practice. Recent qualitative evidence from prospective science teachers in another low-resource context similarly identifies crowded classrooms, limited time, inadequate resources, weak rubric-design skills and insufficient training or monitoring as recurring barriers (Osin & Sahoo, 2026). Technology therefore adds not only an operational demand but an assessment-literacy demand: teachers must understand how a tool changes the evidence on which judgements are based.

The EQUALS project and cluster-based support described in the review are important because they shift professional development from one-off training toward networks in which teachers can share resources and implementation strategies. This logic is compatible with digital teacher-support platforms, but the key mechanism is social and pedagogical rather than technological. Video demonstrations, mobile resources or online communities are useful when they make practice observable, reduce perceived complexity and support adaptation to local classrooms, precisely the adoption conditions highlighted by Rogers' Diffusion of Innovations framework (Rogers, 2003; Gwayi, 2009).

Large classes and workload remain structural constraints. Individual observation, feedback and scoring are time-intensive, and digital systems may either reduce or increase workload depending on design. A portfolio platform, for example, can organise evidence but also generate an unmanageable volume of submissions. Scalable PBA therefore requires task sampling, group-performance structures where appropriate, concise rubrics, moderation procedures and selective use of technology to reduce rather than add friction.

3.4 Equity: gender, school type and the digital divide

The review's gender-responsive pedagogy strand broadens the usual access discussion in educational technology. Equity is not achieved merely by making a tool available. Participation can be shaped by classroom roles, task design, teacher expectations, scheduling and the relevance of science activities to learners' lives (FAWE, 2005, 2015; UNESCO, 2017). PBA can create multiple ways to demonstrate competence, but it can also reproduce bias if boys disproportionately handle equipment while girls record results, or if scoring criteria reward stereotyped expectations.

Digitalisation introduces an additional equity layer. The source literature supports policy ambitions for ICT and virtual laboratories, and the supplementary PhET evidence demonstrates that digital science learning can be effective in Malawi. However, the Banda and Nzabahimana (2023) study drew on four urban secondary schools selected in part because computer facilities were available. Its success therefore should not be interpreted as evidence of equitable national access. The documented differences between conventional schools and CDSSs in resource access suggest that digital PBA could widen gaps unless infrastructure, teacher support and maintenance are distributed deliberately. Technology-enhanced assessment should therefore report participation and outcomes disaggregated by gender, school type and location, alongside device/connectivity conditions. The PhET study included 44.6% female students but did not establish whether intervention effects were equivalent across gender groups, leaving an important equity question for future research.

Gender-responsive PBA also requires attention to validity. The assumption that performance assessment is inherently fair is not supported by the broader science-assessment literature. Jovanovic et al. (1994) examined whether performance-based science assessment eliminated gender differences and showed why the measurement method itself must be scrutinised rather than presumed neutral. For Malawi, this supports the use of transparent rubrics, structured group roles, varied opportunities to demonstrate competence and monitoring of participation as well as scores. Digital portfolios may eventually preserve multiple artefacts and make process evidence more visible, but this remains a research proposition rather than an established Malawian effect in the reviewed evidence.

3.5 Policy alignment and the competency-based reform window

Malawi's policy trajectory increasingly supports competence, practical skills and digital literacy. The National Education Policy, the National Education Sector Plan, continuous-assessment initiatives and the 2025 competency-based curriculum collectively create a more favourable environment for PBA than a purely examination-driven system (Ministry of Education, Science and Technology, 2013; Malawi Government, 2020). The review also notes policy interest in virtual laboratories and ICT infrastructure.

Yet policy alignment must extend to assessment standards and quality assurance. If schools use very different materials or digital platforms, comparability becomes a legitimate concern. Sustainable PBA therefore needs common performance constructs and scoring principles while allowing local variation in task materials. This is where educational technology can support, not replace, assessment governance: digital exemplars, shared rubrics, moderation repositories and teacher-learning resources can make standards more visible across dispersed schools.

The policy opportunity is thus to avoid equating competency-based reform with device procurement. Investments should be sequenced around assessment purposes: define the competence, specify acceptable evidence, identify the lowest-cost valid way to elicit that evidence, and add digital technology where it increases access, feedback, repeatability or documentation.

4. DISCUSSION

This review reframes Malawi's PBA literature as an educational-technology problem of alignment. The central finding is that authentic assessment in low-resource systems is not well served by a simple progression from 'no technology' to 'digital technology'. The evidence instead points to a layered ecosystem in which local materials, physical and virtual laboratories, documentation tools and professional-support networks perform different functions. TALULAR is especially important because it demonstrates that technology for learning can be materially simple while pedagogically sophisticated. It expands what students can perform with resources already embedded in their environment.

The supplementary database verification sharpens this conclusion. The direct Malawian PhET study means that the country's digital-assessment trajectory should not be characterised only as policy aspiration: there is already empirical evidence that simulation-supported physics learning can improve achievement and motivation under enabling infrastructure conditions (Banda & Nzabahimana, 2023). At the same time, the international evidence cautions against technological substitution without validation. Baxter (1995) found mode-related variation at the individual level even when physical and simulated electric-circuit assessments produced similar average performance, while Shavelson et al. (1998) argue that a science performance assessment should be treated as a designed system of task, response demand and scoring rather than as a device. Digitalisation therefore changes the medium of performance, but it does not remove the need to establish construct validity, reliability and feasibility.

The proposed resource–technology continuum offers a novel perspective for educational technology research. Much technology research begins with a named digital tool and asks whether it improves learning. In low-resource PBA, the more productive sequence may be reversed: begin with the performance that needs to be elicited, identify the constraints that prevent it, and then select the least burdensome technology capable of overcoming those constraints. This shifts evaluation from technological novelty to assessment validity, feasibility and equity.

This measurement perspective also addresses scalability. Generalizability studies show that task sampling is a major source of error in performance assessment and that student performance can vary across both tasks and occasions (Shavelson et al., 1993, 1999). Consequently, a single impressive project or simulation should not be treated as a complete measure of scientific competence. For system use, Malawi would need a deliberate sampling architecture: a manageable set of tasks spanning key competencies, common scoring criteria, moderation, and enough repeated evidence to support the intended decision. Digital platforms may reduce some administrative burden and can preserve process data, but they do not eliminate the fundamental sampling problem.

The continuum also connects the Malawian case to a longer history of assessment technology. Shavelson and Baxter's work treated hands-on investigations, notebooks, computer simulations and paper-and-pencil formats as alternative ways of representing or approximating scientific performance. The important question was not which medium appeared more advanced, but how faithfully each generated evidence about the target competence. The National Research Council (2014) extends this logic by arguing for assessment systems capable of eliciting integrated evidence of scientific practices and disciplinary understanding. Read together with Malawi's resource constraints, these perspectives suggest a principle of 'minimum sufficient mediation': select the least burdensome configuration of material and digital resources that still produces valid, interpretable and equitable evidence of the intended performance.

The review also exposes an evidence asymmetry. Claims about TALULAR adoption are grounded in direct teacher studies, whereas claims about virtual laboratories, digital portfolios and ICT-supported PBA are more often grounded in policy intentions or future-oriented recommendations. This distinction matters for concern with whether and how educational technologies improve education. Malawi currently offers a strong implementation research agenda but not yet a mature evidence base for the impact of digital PBA. Future studies should therefore test concrete configurations rather than assuming that digitalisation is inherently beneficial.

A useful next step would be comparative implementation trials across school types. For example, the same scientific competence could be assessed through (a) a TALULAR physical task, (b) a virtual simulation, and (c) a blended sequence in which students first simulate and then construct or investigate with local materials. Outcomes should include conceptual learning, quality of performance evidence, teacher workload, scoring reliability, student engagement, gender participation, cost and infrastructure demands. Such designs would move the field beyond access narratives toward evidence about which technology–assessment combinations work, for whom and under what conditions.

Teacher professional development should be studied with the same functional logic. The adoption findings associated with relative advantage, observability and complexity suggest that professional learning should show teachers complete examples of PBA, make scoring visible and reduce preparation burden. Mobile video exemplars, downloadable task banks and cluster moderation could be evaluated as a combined implementation package. The relevant outcome is not platform use but whether teachers design higher-quality tasks, provide better feedback and make more dependable assessment judgements.

Equity must also be treated as an outcome rather than a background variable. The gender-responsive literature in the review indicates that task participation and classroom roles can shape who benefits from practical science. Digital technology adds possible divides in device access, connectivity and digital competence. Evaluations should therefore examine whether technology changes participation patterns and whether apparently scalable solutions work similarly in CDSSs and better-resourced conventional schools.

A second theoretical implication concerns the meaning of infrastructure. In many educational-technology studies, infrastructure is treated as a background condition measured through connectivity, hardware or platform availability. The Malawi literature suggests a wider definition. For PBA, infrastructure includes physical materials, time, scoring routines, moderation arrangements, teacher knowledge and institutional incentives. A school can possess computers yet still lack assessment infrastructure if teachers do not have workable tasks or if examination rules discourage practical work. Conversely, a school with few digital devices may possess meaningful PBA infrastructure when teachers share locally feasible tasks, common rubrics and peer moderation. This broader definition helps explain why procurement-centred reforms can fail to change assessment practice.

The framework also clarifies the relationship between substitution and augmentation. A virtual laboratory can substitute for an experiment that cannot be conducted because equipment is unavailable, but it can also augment a physical activity by allowing students to rehearse procedures, vary parameters repeatedly or visualise processes that are difficult to observe directly. Similarly, a digital portfolio can substitute for paper storage, but its stronger educational value may lie in making revision histories, multimedia evidence and feedback cycles visible. Research should therefore identify the intended role of technology rather than treating use/non-use as the intervention. This distinction would support more precise explanations of why a technology improves or fails to improve assessment quality.

A third implication is that implementation should be analysed at multiple levels. Teacher perceptions matter, as the TALULAR diffusion evidence shows, but they interact with school-level resources and system-level incentives. A teacher may perceive strong relative advantage in a practical task and still avoid it because class size, timetable pressure or examination expectations make implementation costly. Likewise, a well-designed digital platform may show high usability but little classroom effect if data costs are transferred to teachers or learners. Multi-level implementation research is therefore necessary to avoid attributing system constraints to individual teacher resistance.

The Malawi case also contributes to debates about contextualisation in educational technology. Local relevance is often discussed as a content issue, yet in PBA it is also an assessment-design issue. Tasks based on local water, soil, energy or household materials can generate evidence of scientific reasoning while reducing dependence on imported kits. Contextualisation, however, should not become a justification for lower standards. The construct being assessed, such as experimental design, evidence interpretation or explanation, should remain explicit and comparable even when the materials differ. This balance between common

constructs and locally variable task forms is a promising area for technology-supported moderation and exemplar sharing.

Finally, the review suggests that sustainability should be evaluated as continued capability rather than continued use of a particular product. A sustainable PBA innovation is one that schools can maintain, adapt and quality-assure after external support declines. This favours modular approaches: locally available materials for core practical tasks, shared digital resources where connectivity permits, downloadable/offline exemplars where it does not, and cluster-based human support that does not depend on a single platform. Such an architecture is likely to be more resilient than a uniform technology solution imposed across schools with very different starting conditions.

4.1 Implications for educational technology design and policy

For designers, the priority is to preserve the authenticity of the target competence. A simulation is valuable when it enables experimentation that would otherwise be impossible, but it should not automatically replace manipulation of real materials when physical performance is itself part of the construct. Digital portfolios are valuable when process, iteration and feedback matter, but they need manageable evidence rules. Teacher-support platforms are valuable when they reduce complexity and make good practice observable, but they need to be integrated with local peer support.

For policymakers, technology procurement should be tied to assessment architecture. Common rubrics, moderation processes, exemplar tasks and teacher assessment literacy are infrastructure just as much as devices and laboratories. A sustainable strategy would support a baseline of locally feasible PBA in all schools and add digital tools progressively where they offer clear additional affordances. This approach reduces the risk that the competency-based curriculum produces a new divide between schools able to enact authentic assessment and schools able only to imitate it on paper.

Table 4:Practical decision criteria for selecting assessment technologies.

Decision question	If YES → preferred option	If NO → consider
Is manipulation of real materials part of the competence being assessed?	Local materials task (TALULAR) or physical laboratory	Simulation may be used for rehearsal only
Is the experiment dangerous, expensive, or impossible to perform with available equipment?	Simulation/virtual laboratory, validated for construct equivalence	Local materials adaptation of the task
Does the assessment require evidence of process, revision and feedback over time?	Digital portfolio/documentation tool, with manageable evidence rules	Sampled paper-based artefacts plus moderation
Is the binding constraint teacher preparation time, scoring confidence or isolation?	Mobile/online teacher support combined with cluster peer moderation	Centralised exemplar packs and shared rubrics
Are devices, electricity and connectivity reliably available to all target schools?	Digital tool can be scaled, with disaggregated equity monitoring	Offline/downloadable resources or low-tech alternatives
Does a single task need to support a high-stakes decision?	Multiple sampled tasks across occasions (per Shavelson et al., 1993, 1999)	No single-task decision is defensible

These implications also change how effectiveness should be measured. Student achievement remains important, but PBA technologies should be judged against a broader outcome set: the validity and richness of performance evidence, reliability of scoring, quality and timeliness of feedback, teacher preparation and marking time, accessibility, student participation, and total implementation cost. A technology that produces a small achievement gain while doubling teacher workload may be less scalable than a simpler approach that

improves feedback and participation at low recurring cost. Likewise, a tool that works in a well-resourced conventional school but cannot operate reliably in a CDSS has limited system-level value. Future evaluations should therefore combine learning outcomes with implementation, equity and economic indicators.

A related implication is that PBA should be treated as part of a multiple-measures assessment system rather than as a universal replacement for written testing. Performance tasks are especially valuable when the target construct requires investigation, design, manipulation, reasoning, explanation or creation, while more economical formats may remain appropriate for other purposes. The design problem is therefore one of evidence sufficiency: schools need enough observations and artefacts to support defensible judgements without creating an assessment workload that cannot be sustained. This principle is particularly important in large, resource-constrained classes, where the educational value of authenticity must be balanced against scoring reliability, teacher time and opportunities for feedback. A blended architecture of selected-response items, constructed responses, sampled performance tasks and portfolios may therefore be more scalable than attempting to assess every competency through extended performance tasks (National Research Council, 2014; Rudner & Boston, 1994; Stiggins, 1987).

The international relevance of the Malawi case lies in this emphasis on constrained choice. Many education systems are introducing competency-based curricula while simultaneously expanding digital infrastructure. The temptation is to treat these reforms as naturally complementary. The evidence reviewed here suggests that complementarity must be designed. Competency-based assessment specifies performances that need to be observed; technology offers possible ways to elicit, capture or support those performances; and institutional conditions determine whether those possibilities become routine practice. This three-part relationship assessment construct, technological affordance and implementation condition, provides a transferable analytic frame for research in other low- and middle-income settings.

4.2 Limitations and review transparency

The evidence base has important limitations. The source review reports geographical concentration in particular regions and districts, substantial reliance on self-report, variable study quality and exclusion of non-English sources. The policy landscape is also changing rapidly. Direct empirical evidence on digital science PBA in Malawi remains limited, but the supplementary database verification identified at least one strong secondary-school PhET study, so it is no longer accurate to describe the digital evidence as purely prospective or policy-based. Even this evidence is context-bound because participating schools had computer facilities and the study evaluated learning and motivation rather than a national performance-assessment system. The supplementary Scopus/Web of Science/PubMed verification was not a formal re-run of the original systematic search and screening workflow; therefore, it should not be used to revise the PRISMA counts without a documented updated search, deduplication and eligibility screen. In the present manuscript, these records are used transparently to test, qualify and extend the interpretation of the original 72-source synthesis.

To enhance review transparency, the complete included-study list and extraction matrix for the 72 included sources are provided as Supporting Information, enabling readers to examine study characteristics, quality judgements and the mapping of individual sources to synthesis claims. These materials complement the PRISMA flow diagram by providing study-level traceability for the review.

5. CONCLUSION

Malawi's experience shows that performance-based science assessment in low-resource settings is fundamentally an educational-technology design challenge. Locally available materials can make scientific performance possible where formal laboratories are scarce, while direct Malawian evidence now shows that PhET simulations can improve physics achievement and motivation in schools with enabling computer infrastructure. International research further demonstrates that simulations and virtual performance environments can capture conceptual and process evidence that conventional tests may miss. None of these technologies is sufficient on its own, and digital and physical formats should not be assumed equivalent without validation.

The contribution of this review is to replace a low-tech/high-tech binary with a resource-technology continuum organised around assessment function. The most appropriate technology is the one that elicits valid evidence of the intended competence with acceptable demands on teachers, infrastructure and learners. For Malawi and comparable systems, the immediate research priority is to test blended configurations of local materials, simulations, digital evidence tools and professional-support systems while measuring achievement, construct validity, scoring reliability, task/occasion sampling, teacher workload, cost and equity. Future updates of the review should formally incorporate newly identified evidence through a documented search, screening and extraction process, while the present review retains the original 72-study corpus and treats the September 2026 evidence as supplementary contextual evidence.

REFERENCES

1. Age, T. J. (2025). Performance-based assessment: A transformative approach to enhancing mathematics learning in Ubuntu classrooms across Sub-Saharan Africa. European Journal of STEM Education, 10(1), 04. https://doi.org/10.20897/ejsteme/16420
2. Baker, R. S., Clarke-Midura, J., & Ocumpaugh, J. (2016). Towards general models of effective science inquiry in virtual performance assessments. Journal of Computer Assisted Learning, 32(3), 267-280. https://doi.org/10.1111/jcal.12128
3. Banda, H. J., & Nzabahimana, J. (2023). The impact of Physics Education Technology (PhET) interactive simulation-based learning on motivation and academic achievement among Malawian physics students. Journal of Science Education and Technology, 32(1), 127-141. https://doi.org/10.1007/s10956-022-10010-3
4. Baxter, G. P. (1995). Using computer simulations to assess hands-on science learning. Journal of Science Education and Technology, 4(1), 21-27. https://doi.org/10.1007/BF02211578
5. Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7–74.
6. Chilora, H. G., Nampota, D., Chiwaula, L., Chisale, E., Chibwana, C., Kazembe, J., & Mchikoma, S. (2003). Teaching and learning using locally available resources (TALULAR): A manual for teachers. Malawi Institute of Education.
7. Chimombo, J. P. G. (2005). Quantity versus quality: The case of basic education in Malawi. International Journal of Educational Development, 25(3), 261–273.
8. FAWE. (2005). Gender-responsive pedagogy: A teacher's handbook. Forum for African Women Educationalists.
9. FAWE. (2015). Tuseme: Let's speak out! Gender-responsive pedagogy in practice. Forum for African Women Educationalists.
10. Gale, J., Wind, S., Koval, J., Dagosta, J., Ryan, M., & Usselman, M. (2016). Simulation-based performance assessment: An innovative approach to exploring understanding of physical science concepts. International Journal of Science Education, 38(14), 2284-2302. https://doi.org/10.1080/09500693.2016.1236298
11. Gwayi, S. M. (2009). Perceptions of innovations as predictors of TALULAR implementation levels among secondary school science teachers in Malawi: A diffusion of innovations [Doctoral dissertation, Virginia Polytechnic Institute and State University]. VTechWorks.
12. Hibbard, K. M. (2000). Performance-based learning and assessment in middle school science. Eye On Education.
13. Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., et al. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Education for Information, 34(4), 285–291.
14. Jovanovic, J., Solano-Flores, G., & Shavelson, R. J. (1994). Performance-based assessments: Will gender differences in science achievement be eliminated? Education and Urban Society, 26(4), 352–366.
15. Lungu, C. B. (2005). Status of science education in Malawi secondary schools. Malawi Institute of Education.
16. Malawi Government. (2020). Malawi 2063: An inclusively wealthy and self-reliant nation. Government of Malawi.

17. Malawi Institute of Education. (2018). Continuing professional development framework for teachers. Ministry of Education, Science and Technology.
18. MANEB. (2014). Annual examination statistics report. Malawi National Examinations Board.
19. Ministry of Education, Science and Technology. (2013). National Education Policy. Government of Malawi.
20. Mtika, P., & Gates, P. (2010). Developing learner-centred education among secondary trainee teachers in Malawi: The dilemma of appropriation and application. International Journal of Educational Development, 30(4), 396-404. https://doi.org/10.1016/j.ijedudev.2009.12.004
21. National Research Council. (2014). Developing assessments for the Next Generation Science Standards. The National Academies Press.
22. Osin, & Sahoo, M. (2026). Performance assessment in science: Issues and challenges in implementation. International Journal of Social Science Research, 3(2), 95–104. https://doi.org/10.70558/IJSSR.2026.v3.i2.30902
23. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71.
24. Rogers, E. M. (2003). Diffusion of innovations (5th ed.). Free Press.
25. Rudner, L. M., & Boston, C. (1994). Performance assessment. ERIC Review, 3(1), 2–12.
26. Shavelson, R. J., & Baxter, G. P. (1991). Performance assessment in science. Applied Measurement in Education, 4(4), 347–362.
27. Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. Journal of Educational Measurement, 30(3), 215-232. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x
28. Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessments: Political rhetoric and measurement reality. Educational Researcher, 20(4), 22–27.
29. Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. Journal of Educational Measurement, 36(1), 61-71. https://doi.org/10.1111/j.1745-3984.1999.tb00546.x
30. Shavelson, R. J., Solano-Flores, G., & Ruiz-Primo, M. A. (1998). Toward a science performance assessment technology. Evaluation and Program Planning, 21(2), 171-184. https://doi.org/10.1016/S0149-7189(98)00005-6
31. Stiggins, R. J. (1987). Design and development of performance assessments. Educational Measurement: Issues and Practice, 6(3), 33–42.
32. Susuwele-Banda, W. J. (2005). Classroom assessment in Malawi: Teachers' perceptions and practices in mathematics [Doctoral dissertation, University of Pretoria].
33. Tyndall, J. (2010). AACODS checklist. Flinders University.
34. UNESCO. (2017). Cracking the code: Girls' and women's education in science, technology, engineering and mathematics (STEM). UNESCO.
35. Wiggins, G., & McTighe, J. (2005). Understanding by design (2nd ed.). ASCD.
36. World Bank. (2018). Malawi education sector improvement project (EQUALS): Project appraisal document. World Bank Group.

AI Use Declaration

Generative AI was used for language polishing, structural revision and editorial assistance in preparing this manuscript. The author reviewed the resulting text and remains responsible for the accuracy of all claims, citations, interpretations and references.