European International Journal of Pedagogics (ISSN: 2751-OOOX)

Received: 31 December 2025 | Accepted: 22 January 2026 | Published: 28 February 2026

DOI: 10.55640/eijp-06-02-29

RESEARCH ARTICLE

Methodology for Determining and Assessing Students' Achievement Level in Biology Using Mobile Applications

Salimova Sarvinoz Farxodovna
d.p.p.s., associate professor at Department of Biology at Bukhara State University, Uzbekistan

Kenjaeva Nozima Jobirovna
1st year master's student at Bukhara State University, Uzbekistan

VOLUME: Vol.06 Issue02 2026

PAGE: 130-134

Abstract

Mobile applications can make assessment in school biology more frequent, timely, and instructionally useful, yet "digital quizzes" alone do not guarantee valid measurement or fair decisions about achievement. The approach combines evidence-centered design, formative feedback principles, and psychometric and learning-analytics procedures so that curriculum outcomes are explicitly linked to mobile tasks, scoring rules, and achievement-level decisions. Biology achievement is treated as a multi-component construct covering conceptual understanding, inquiry and experimentation, data interpretation, and scientific communication. Mobile applications deliver micro-assessments, scenario-based items, and virtual-lab tasks while capturing limited process indicators as secondary evidence. Achievement estimation is conducted through layered scoring: calibrated objective items (IRT where feasible), rubric-based scoring for explanations and performance tasks, and a composite achievement index used for mastery classification and growth monitoring. An illustrative synthetic dataset demonstrates reliability, alignment with an external criterion, and decision accuracy, and the paper discusses implementation constraints, equity, and data governance.

KEYWORDS

Biology education; mobile assessment; formative feedback; evidence-centered design; learning analytics; mastery learning; psychometrics.

INTRODUCTION

Biology learning in school is expected to include more than memorizing terms. Students must explain biological phenomena, interpret data, and demonstrate inquiry practices, yet classroom assessment often over-relies on short written tests because practical evaluation is time-consuming and inconsistent. Formative assessment research highlights that learning gains increase when students receive feedback that is timely, specific, and oriented toward improvement rather than judgment [1–2]. Mobile applications can support this loop by enabling short checks of understanding, immediate feedback, and structured opportunities for retrieval practice, which strengthens long-term retention when learners repeatedly recall information rather than only re-study it [3–5].

However, mobile assessment creates methodological risks. A quiz-heavy app can narrow learning to recognition tasks, while analytics dashboards can create false certainty if digital traces are treated as achievement without a validity argument.

Differences in access to devices and connectivity can distort achievement levels by confounding learning with opportunity. Therefore, a defensible methodology must specify what is being measured, what counts as evidence, how evidence is scored, how levels are set, and how decisions remain fair and transparent.

The methodology adopts evidence-centered design (ECD), which treats assessment as an evidentiary argument connecting claims about competence to observable performances and principled scoring rules [6–8]. ECD is especially suitable for mobile assessment because the technology can generate many observations, but only a subset are meaningful evidence. In the proposed approach, the assessment argument is documented in four linked models. The construct map defines what "biology achievement" means in the local curriculum and what progression looks like. The task model specifies what students do in the app to elicit evidence. The evidence model defines how observations are scored and interpreted. The assembly model defines how tasks are selected across time so that evidence covers the construct without over-testing. This logic is complemented by formative assessment principles that emphasize feedback usable for improving learning [1–2, 11].

Biology achievement is operationalized across four domains: conceptual understanding of core ideas, inquiry and experimentation, data interpretation, and scientific communication. Conceptual understanding emphasizes causal and systems reasoning rather than recall alone. Inquiry and experimentation includes posing investigable questions, identifying variables, predicting outcomes, and drawing conclusions from evidence. Data interpretation includes reading graphs and tables, recognizing variability and trends, and evaluating claims supported by data. Scientific communication includes constructing short explanations that connect claims with evidence and using biological terminology appropriately.

Achievement is reported through four ordered levels: emerging, basic, proficient, and advanced. Level descriptors are written so they are observable in mobile tasks. For instance, "emerging" may reflect correct recognition with frequent hints and weak explanation, while "proficient" reflects consistent accuracy, correct variable reasoning in scenarios, and explanations that include causal links. "Advanced" reflects transfer to unfamiliar contexts and critique of flawed conclusions.

Mobile applications deliver three complementary task families. The first family consists of calibrated knowledge items that include misconception-sensitive distractors and spaced retrieval schedules. The use of spaced retrieval is justified both pedagogically and measurement-wise: repeated low-stakes prompts improve retention and produce multiple observations for more stable achievement estimates [3–5]. The second family consists of scenario-based inquiry items and virtual-lab tasks, where students select investigative steps, manipulate variables, and interpret simulated results. The third family consists of short explanation prompts, diagram labeling, and data-story tasks that require evidence-based communication, helping prevent the assessment system from collapsing into recognition-only evidence.

The app records outcome data (accuracy, partial credit, rubric score) and limited process indicators (e.g., revisions, hint use, step sequences). Process indicators are treated as secondary evidence, primarily for formative diagnostics and validity checks. Because process data are easily confounded by reading speed, interface familiarity, or distraction, they are never used alone for high-stakes achievement decisions.

Scoring is layered to match evidence types. Objective items are scored dichotomously or polytomously and, when item banks and sample sizes are sufficient, calibrated using IRT to support comparable scores across different forms and across time. If a full IRT infrastructure is not feasible, the methodology recommends a calibrated classical approach with anchor items and periodic item analysis.

Performance and explanation tasks are scored with analytic rubrics aligned to the four achievement levels. Teacher scoring inside the app is supported with exemplar responses and brief scoring notes to reduce rater drift. Automated scoring, if used, is limited to low-stakes feedback support and triage, because opaque automated decisions can undermine fairness and interpretability.

Feedback design follows evidence-based principles: it should be timely, specific, and oriented toward next steps [2, 9, 11]. In a mobile context, this means feedback that references the objective and suggests an action, such as re-running a simulation with a different variable setting or rewriting an explanation by adding evidence.

For overall reporting, a composite achievement index (CAI) is computed as a weighted combination of domain scores: $CAI = wK \cdot K + wI \cdot I + wD \cdot D + wC \cdot C$, where K, I, D, and C denote

scaled domain scores and weights sum to 1. Weight setting is treated as a curricular decision and documented for transparency. Cut scores on CAI assign overall achievement levels, and minimum domain rules can be added to prevent strong recall scores from masking weak inquiry or communication. Growth monitoring is supported by repeated CAI measures and by objective-level mastery probabilities updated after each relevant task.

Content alignment is ensured by blueprinting tasks to curriculum outcomes and cognitive demand, avoiding over-representation of easy recognition items. Reliability is monitored for objective items via internal consistency and, when using IRT, test information, and for rubric tasks via inter-rater agreement. Criterion-related evidence is gathered by relating mobile indices to an external test or to structured teacher judgments. The methodology also incorporates learning analytics hygiene: missingness patterns are inspected so that non-participation is not automatically interpreted as low achievement [12–13]. Data governance follows minimization and transparency principles consistent with international mobile learning guidance: clear notice of what is collected, secure storage, limited retention, and role-based access [10].

To demonstrate the computations required by the methodology without claiming results from a specific school, an illustrative synthetic dataset was generated to mimic a biology unit supported by mobile micro-assessments and a virtual-lab task. The dataset included 180 students, pre- and post-unit scores (0–100), a 28-item mobile quiz form administered through an app, and a rubric-scored virtual-lab performance task. Although synthetic, the dataset supports transparent demonstration of reliability, criterion alignment, and decision rules.

Internal consistency of the 28-item mobile quiz component, computed as Cronbach's alpha, was 0.92. Mean pre-unit achievement was 50.9 points, while mean post-unit achievement was 57.2 points. The mean gain was 6.26 points with a paired-sample effect size of d = 0.85. The normalized gain was 0.13. In domain terms, mean scaled scores were approximately K = 63.4, I = 64.8, D = 61.2, C = 57.6, producing an average CAI = 62.1.

Criterion-related evidence can be explored by comparing a mobile-based mastery index with an external criterion. In the synthetic demonstration, a mastery index derived from mobile quiz performance correlated with post-unit achievement at r = 0.89. When a mastery threshold of 0.70 was used to classify students as proficient-or-above and the external criterion was post-unit score ≥ 70, the decision rule achieved sensitivity = 0.97, specificity = 0.65, and overall accuracy = 0.71. This illustrates how threshold choices influence the balance between minimizing missed proficient students and avoiding over-classification.

To demonstrate level assignment, CAI cut scores were set to create four ordered categories: CAI < 40 as emerging, 40–59 as basic, 60–79 as proficient, and ≥ 80 as advanced. In the illustrative dataset, the resulting distribution was 21.1% emerging, 23.9% basic, 25.0% proficient, and 30.0% advanced. This type of distribution check supports standard-setting conversations and helps detect mis-targeted task difficulty.

For the rubric-scored virtual-lab task, inter-rater reliability was illustrated using two independent ratings on a four-level rubric aligned to emerging, basic, proficient, and advanced. Quadratic-weighted Cohen's kappa in the synthetic demonstration was 0.82, reflecting substantial agreement when raters are trained and rubrics are anchored with exemplars.

The proposed methodology addresses a common gap in school practice: mobile assessment is often adopted for convenience, yet the underlying measurement argument remains implicit. By grounding mobile assessment in evidence-centered design, the approach clarifies how each digital observation is intended to function as evidence of biology achievement and how multiple evidence sources can be combined into transparent achievement levels. This clarity is important for fairness and trust, particularly when results influence grades.

A second strength of the methodology is its alignment with formative feedback and learning science. Mobile micro-assessments can support retrieval practice, and timely feedback can guide students' self-regulation when it is specific and actionable rather than merely evaluative [2–5]. The methodology emphasizes that feedback should be tied to learning objectives and should prompt improvement actions, such as revisiting a simulation with a different variable setting or rewriting an explanation with explicit causal links. This design reduces the risk that mobile apps become point collectors detached from biological reasoning.

Implementation constraints are nontrivial. Device availability

and internet connectivity vary widely, and equity concerns must be addressed so that assessment does not reflect differential access. Offline modes, school-provided devices, and scheduling policies can reduce these inequities. From a governance perspective, data minimization, informed consent, and secure storage are prerequisites for responsible mobile assessment and align with international guidance on mobile learning policy [10]. Schools should also consider the social consequences of constant measurement, avoiding surveillance dynamics by keeping analytics focused on learning support rather than punishment.

Teacher capacity building is part of the methodology, not an implementation afterthought. Rubric scoring and cut-score setting require a shared interpretation of "proficient" biology performance, so schools should schedule brief calibration activities and use standard-setting procedures that combine professional judgment with evidence, such as a modified Angoff discussion for objective items and a borderline review for rubric tasks. When teachers participate in these processes, the achievement levels become more transparent, more consistent across classes, and easier to explain to students and parents.

The methodology has limitations. The strongest psychometric approaches, including IRT calibration and bias analyses, require sufficient sample sizes and stable item banks. Process data such as time-on-task can be misleading and should remain secondary evidence. Finally, achievement levels depend on cut scores and descriptors that require local validation and professional consensus. Future empirical work should test the framework with real cohorts, examine how teachers use the reports, and evaluate effects on long-term reasoning and motivation.

Mobile applications can improve how biology teachers determine and assess students' achievement levels, but only when assessment is designed as a coherent evidentiary system rather than as a collection of quizzes. The methodology presented here integrates evidence-centered design, formative feedback, and layered scoring models to connect curriculum outcomes with mobile-delivered tasks and interpretable achievement levels. The illustrative computations show how reliability and decision-accuracy indicators can be computed and used to refine the system. Implemented responsibly, mobile assessment can strengthen feedback loops and support deeper biological understanding and inquiry competencies.

REFERENCES

1. Black P., Wiliam D. Assessment and classroom learning // Assessment in Education: Principles, Policy & Practice. 1998. Vol. 5, No. 1. P. 7–74. DOI: 10.1080/0969595980050102.
2. Shute V.J. Focus on formative feedback // Review of Educational Research. 2008. Vol. 78, No. 1. P. 153–189. DOI: 10.3102/0034654307313795.
3. Hattie J., Timperley H. The power of feedback // Review of Educational Research. 2007. Vol. 77, No. 1. P. 81–112. DOI: 10.3102/003465430298487.
4. Nicol D.J., Macfarlane-Dick D. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice // Studies in Higher Education. 2006. Vol. 31, No. 2. P. 199–218. DOI: 10.1080/03075070600572090.
5. Karpicke J.D., Roediger H.L. The critical importance of retrieval for learning // Science. 2008. Vol. 319, No. 5865. P. 966–968. DOI: 10.1126/science.1152408.
6. Roediger H.L., Karpicke J.D. Test-enhanced learning: taking memory tests improves long-term retention // Psychological Science. 2006. Vol. 17, No. 3. P. 249–255. DOI: 10.1111/j.1467-9280.2006.01693.x.
7. Dunlosky J., Rawson K.A., Marsh E.J., Nathan M.J., Willingham D.T. Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology // Psychological Science in the Public Interest. 2013. Vol. 14, No. 1. P. 4–58. DOI: 10.1177/1529100612453266.
8. Mislevy R.J., Steinberg L.S., Almond R.G. On the structure of educational assessments // Measurement: Interdisciplinary Research and Perspectives. 2003. Vol. 1, No. 1. P. 3–62.
9. Mislevy R.J. A brief introduction to evidence-centered design. Princeton, NJ: Educational Testing Service, 2004.
10. Mislevy R.J. Implications of evidence-centered design for educational testing // Educational Measurement: Issues and Practice. 2006.
11. UNESCO. Policy guidelines for mobile learning. Paris: UNESCO, 2013. 41 p. ISBN 978-92-3-001143-7.
12. Siemens G. Learning analytics: the emergence of a

discipline and where it might go // American Behavioral Scientist. 2013. DOI: 10.1177/0002764213498851.

13. Ferguson R. Learning analytics: drivers, developments and challenges // International Journal of Technology Enhanced Learning. 2012. Vol. 4, No. 5/6. P. 304–317. DOI: 10.1504/IJTEL.2012.051816.

14. Traxler J. Defining, discussing and evaluating mobile learning: the moving finger writes and having writ… // International Review of Research in Open and Distributed Learning. 2007. Vol. 8, No. 2. DOI: 10.19173/irrodl.v8i2.346.

15. Crompton H. A historical overview of mobile learning: toward learner-centered education // Handbook of Mobile Learning / ed. by Z.L. Berge, L.Y. Muilenburg. New York: Routledge, 2013. P. 3–14.