INTERACTION: Jurnal Pendidikan Bahasa Vol. 13, No 2; September 2026
P-ISSN: 2406-9558; E-ISSN: 2406-9566

Development of Educandy-Based Evaluation Media for Grade XI Short Story Learning

Haisyah
haisyahaisyah@gmail.com

Fitriani
fitrianiyahya73@gmail.com

Herlina
alif.alifah7810@gmail.com

Universitas PGRI Pontianak

ABSTRACT

This study developed and examined the feasibility of an Educandy-hosted assessment package for Grade XI short story learning at SMA Negeri 1 Sungai Kakap. The product was not a new website; it consisted of short-story assessment items, instructions, and scoring settings configured on the existing Educandy platform. The study followed an adapted 4D sequence comprising Define, Design, Develop, and a limited classroom trial. Full dissemination beyond the development site was not completed. Participants reported in the study were 26 Grade XI students, one Indonesian language teacher, two media validators, one material validator, and two language validators. Data were obtained from a preliminary observation and interview, expert validation sheets, and teacher and student response questionnaires, then summarized descriptively as percentages. Media validation averaged 89%, material validation 92%, and language validation 92%, all within the study's very feasible category. The student response averaged 83.79%, while the response of the single teacher was 92%. These results support expert-rated feasibility and case-specific perceptions of usability, attractiveness, and practicality. They do not establish learning effectiveness, increased motivation, or improved comprehension. Important limitations include the single-school and single-class setting, reliance on one material validator and one teacher informant, unavailable item-level raw data and reliability coefficients, the absence of a comparative learning-outcome design, and incomplete dissemination. Further research should conduct item analysis, accessibility testing, broader validation, and controlled effectiveness testing.
Keywords: Educandy; Digital Assessment; Short Story; Evaluation Media; 4D Model

INTRODUCTION

Assessment is integral to instruction because it provides evidence about what students understand and where additional support is needed. When assessment is aligned with learning outcomes, its results can inform teaching decisions and help

learners recognize gaps in their understanding. Digital delivery can make assessment faster and more interactive, but the value of a digital tool depends on; the quality of the tasks, feedback, and interpretation of scores rather than on the platform alone (Black & Wiliam, 2009; Shute, 2008).

Game-based assessment adds features such as points, time limits, repeated attempts, and immediate feedback to an assessment activity. These features can influence attention and persistence, yet they do not automatically improve learning. Game-based learning research stresses the importance of aligning game mechanics with cognitive, motivational, and instructional goals (Plass et al., 2015). Likewise, the meta-analysis by Sailer and Homner (2020) found positive average effects of gamification while also showing that outcomes vary according to design and context. Consequently, a positive user response should be interpreted as evidence of perceived attractiveness or acceptability, not as direct proof of higher achievement.

Short story learning in Grade XI requires students to identify and interpret intrinsic elements such as theme, characterization, plot, setting, point of view, and moral message. These outcomes involve different levels of cognition. Remembering a named character is different from inferring a character trait from textual evidence or evaluating the plausibility of a moral interpretation. The revised taxonomy distinguishes remembering and understanding from analyzing and evaluating (Anderson & Krathwohl, 2001). Multiple-choice questions may sample some analytical processes when they use appropriate stimuli and plausible alternatives, but a small set of timed questions cannot fully represent open literary interpretation (Brookhart, 2010).

A preliminary interview with an Indonesian language teacher at SMA Negeri 1 Sungai Kakap on February 25, 2026 indicated that evaluation of short story learning relied mainly on written multiple-choice tests, essays, worksheets, and analysis assignments. The teacher reported that the existing evaluation routines were repetitive and that students often focused on recalling intrinsic elements instead of interpreting the relationships among those elements. This context created a practical need for a more varied digital assessment format while retaining alignment with the Grade XI short story learning outcomes.

Previous studies have developed Educandy-based evaluation activities in several contexts. Nurhabibah et al. (2021) reported the development of Educandy as an evaluation tool for Grade V Indonesian language learning. Purba and Dwi (2023) developed a digital evaluation activity for elementary science content, while Mufid and Utsman (2022) used Educandy for online Arabic language evaluation. Azizah et al. (2025) also reported Educandy development for Indonesian cultural-heritage content at the elementary level. These studies support the feasibility of configuring existing platform features for classroom evaluation, but their content, grade levels, and local contexts differ from Grade XI short story learning.

The contribution of the present study is therefore specific rather than universal. It documents the development and local feasibility of an Educandy-hosted package containing ten multiple-choice questions for Grade XI short story learning. It does not claim to have created the Educandy website, and it does not test whether the package improves achievement or motivation. The study addressed three questions: (1) How was the package developed through the Define, Design, Develop, and limited-trial sequence? (2) How did media, material, and language validators rate its feasibility? and (3) How did one teacher and 26 students evaluate its usability and acceptability after a limited trial?

LITERATURE REVIEW

Digital and Game-Based Assessment

Digital assessment should be understood as an assessment process delivered through technology, not merely as a change from paper to screen. Sound assessment requires a clear construct, appropriate items, understandable instructions, equitable access, and cautious interpretation. Feedback can support learning when it is timely and specific, but point systems and instant correctness indicators are not substitutes for explanatory feedback (Shute, 2008). In a platform such as Educandy, the developer must decide which learning outcomes the questions represent, how time and attempts will work, and what the resulting score can legitimately mean.

The distinction between game-based learning and gamification is also relevant. Game-based learning integrates learning activity with a game environment, whereas gamification usually adds selected game elements to a non-game task. Educandy-hosted quizzes use features such as points, time, and repeated attempts, so the assessment experience is gamified even when the underlying task remains a conventional multiple-choice question. Plass et al. (2015) explain that effective game-based environments coordinate cognitive, motivational, affective, and sociocultural dimensions. This coordination cannot be inferred solely from favorable percentages in a response questionnaire.

Short Story Assessment and Cognitive Demand

Short story comprehension ranges from locating explicit information to interpreting how narrative elements create meaning. According to Anderson and Krathwohl (2001), analysis requires students to differentiate relevant details, organize relationships, and attribute purpose, while evaluation requires judgments based on criteria. A question that asks students to select a stated setting mainly measures recall or understanding. A question that presents an extract and asks which evidence best supports an inferred theme can place greater demand on analysis.

The assessment package in this study included ten multiple-choice questions with a 30-second timer per question and up to three attempts. These settings may make the activity lively, but they also constrain construct coverage. A short timer can

disadvantage slower readers or students with unstable internet access, and repeated attempts without deduction may combine comprehension with persistence or guessing. Therefore, the package is treated as a supplementary formative activity. It is not presented as a stand-alone measure of analytical literary competence. Constructed responses, discussion, or written analysis remain necessary when teachers need evidence of students' reasoning and interpretation (Brookhart, 2010).

Previous Educandy Development and Study Position

The existing Educandy literature commonly reports development steps, expert validation, and learner responses. Nurhabibah et al. (2021) and Purba and Dwi (2023) show how platform features can be configured into classroom evaluation activities, while Mufid and Utsman (2022) report favorable user responses in an online language-learning context. Such findings establish practical precedent, but they do not remove the need to examine item quality, access, or learning outcomes in each new context. This study extends that practical line of inquiry to senior high school short story learning at one school. Its originality lies in the content selection, learning-outcome alignment, item writing, and local configuration of the package, not in the invention of the platform. The evidence is limited to development documentation, validator judgments, and user responses. This narrower position provides a more defensible basis for interpreting the reported percentages.

METHOD

Design and Samples

This Research and Development study used the 4D framework developed by Thiagarajan et al. (1974) as an organizing model. The study completed the Define, Design, and Develop stages, followed by a limited trial in one Grade XI class. It did not complete full dissemination through wider distribution, adoption, or evaluation beyond the development school. For accuracy, the final stage is therefore reported as a limited trial rather than as completed Disseminate activity.

The reported field-trial sample consisted of 26 Grade XI students and one Indonesian language teacher at SMA Negeri 1 Sungai Kakap. The validation group consisted of two media validators, one material validator, and two language validators. The roles indicate the intended areas of judgment: educational media, Indonesian language and short story content, and linguistic clarity. However, the retained research documentation does not provide the validators' academic qualifications, institutional positions, years of experience, selection criteria, or independence from the development team. The findings are consequently interpreted as ratings from the reported validation roles rather than as broad expert consensus.

The available aggregate record also does not document the students' prior Educandy experience, prior achievement, individual device ownership, connectivity quality,

absences, exclusions, or missing responses. The class is treated as a case-specific limited-trial group, and the single teacher response is treated as one informant's evaluation. No claim is made that these participants represent all Grade XI students or teachers.

Table 1. Development Sequence, Evidence, and Output

Stage	Main Activity and Data Source	Output and Decision
Define	Teacher observation/interview; student needs questionnaire; curriculum and learning-outcome review	Needs profile for a digital short story assessment package
Design	Selection of story elements; item and interface planning; choice of Educandy Quiz Questions mode	Prototype containing ten multiple-choice items and configured game settings
Develop	Media, material, and language validation; review of lower-rated dimensions	Revised package judged feasible by the reported validators
Limited trial	One Grade XI class and one teacher used the package and completed response forms	Case-specific usability and acceptability evidence; issues for further revision
Full Disseminate	Wider distribution, adoption, and multi-site evaluation	Not completed; recommended for subsequent research

Instruments and Procedures

The developed product was an assessment package hosted on the existing Educandy platform. The authors' work comprised selecting Grade XI short story outcomes, writing and organizing ten Indonesian-language multiple-choice questions, selecting the Quiz Questions mode, and configuring timing, attempts, and scoring. Each question was reported to have a 30-second timer; a correct answer could receive 27 points; and students could attempt an incorrect item up to three times without a stated point deduction. The package was shared through a web link and required a smartphone or computer with internet access. The submitted development record does not preserve the platform version, final public link, screenshots, or an archived copy of the item set. These missing elements limit independent inspection and long-term reproducibility.

The needs instruments covered four teacher-related dimensions and five student-related dimensions. Media validation covered interface and display, navigation, interactivity, technical quality, and suitability. Material validation covered content

alignment, content accuracy, linguistic quality, presentation, and evaluation quality. Language validation covered standardness, communicativeness, and student appropriateness. Student responses covered display, operation, clarity, attractiveness, perceived motivation and engagement, and practicality. The teacher form covered content, language, interface, ease of use, and instructional benefit.

The report states that the validation and response instruments used a five-point Likert scale. It does not retain the number of items within each dimension, the exact response anchors, negatively keyed items, source items, item-development process, sample statements, content-validity statistics, or reliability coefficients. Because an item-level response matrix is unavailable, Cronbach's alpha and score dispersion cannot be reconstructed. The reported 98% classroom-observation result is retained only as a descriptive note because the observer, number of observations, checklist items, and scoring rubric are not documented.

After the initial package was prepared, the media, material, and language validators reviewed the prototype. Lower ratings were used to identify priorities involving navigation, interface clarity, material presentation, and wording. The package was then used in one classroom trial. Students accessed the activity on available web-enabled devices, completed the ten questions under the configured timing and attempt rules, and provided responses after use. The teacher observed the activity and completed a separate response form. The retained record does not specify the trial date, session duration, division of teacher and researcher roles, technical incidents, reading accommodations, or the timing of data collection in sufficient detail for exact replication.

Table 2. Documented Product Components and Access Requirements

Component	Documented Configuration	Interpretive Note
Platform	Existing Educandy web platform, Quiz Questions mode	The study developed hosted content and settings, not a new website
Target users	Grade XI students studying short stories	Evidence is limited to one class at one school
Content	Ten Indonesian-language multiple-choice questions	Theme, characterization, plot, and moral message were identified as target elements
Timing	30 seconds per question	Rationale and accommodation for slower reading were not documented
Attempts	Up to three attempts after an incorrect response	Final score may reflect persistence or guessing as well as comprehension

Component	Documented Configuration	Interpretive Note
Scoring	27 points reported for a correct response; reported score range 123-270	The minimum score cannot be reconstructed from the stated scoring rule
Access	Web link on smartphone or computer; internet required	No offline alternative, archived link, platform version, or privacy configuration was retained

Data Analysis

The study used descriptive quantitative analysis. Each reported percentage followed P = (obtained score / maximum possible score) x 100. Feasibility ratings were interpreted using the study's thresholds: 81%-100% = very feasible, 61%-80% = feasible, 41%-60% = less feasible, 21%-40% = infeasible, and 0%-20% = very infeasible (Purba & Dwi, 2023; Sugiyono, 2020). Needs percentages are reported descriptively and are not treated as the same construct as expert feasibility or user preference.

The dimension percentages and averages were checked arithmetically. Table 3 averages 75.25%, the media validation average rounds from 88.5% to 89%, the student response dimensions average 83.79%, and the teacher dimensions average 92%. When source tables used whole numbers, results are retained at that precision; otherwise they are shown to two decimal places. The raw obtained scores, maximum scores, item counts, and respondent-level data were not available in the development record. Therefore, denominators, missing data, standard deviations, inter-rater agreement, and reliability estimates cannot be reported or independently verified. These constraints are carried into the interpretation rather than concealed by the aggregate percentages.

The manuscript reports only aggregate results and does not identify individual students or validators. The retained documentation, however, does not include an ethics approval or exemption number, written parental or guardian consent, student assent, voluntary-participation procedures, data-retention arrangements, or a formal privacy review of Educandy. It also does not clarify whether students used personal accounts or entered identifiable information. These omissions prevent a complete ethical audit of the trial and are acknowledged as a procedural limitation.

Internet dependence and timing also create accessibility concerns. The student needs result for access convenience was 51.74%, the lowest needs indicator, yet no offline alternative or formal accommodation procedure was documented. Subsequent trials should use school-approved consent and data-protection procedures, minimize personal data, provide an untimed or extended-time option

when appropriate, offer shared school devices or an offline alternative, and document the accessibility and privacy settings of the external platform.

RESULTS AND DISCUSSION

Define Stage: Needs Analysis

The Define stage indicated a strong case-specific need for a more varied evaluation format. The teacher-related dimensions produced a mean of 94% after rounding from 93.75%. This result came from the study's single teacher-related needs source and should not be generalized as a population estimate. Because the item count and raw score were not retained, the percentages describe the archived summary rather than a fully auditable scale.

Table 3. Teacher-Related Needs Profile

Observed aspect	Percentage	Descriptive category
Current evaluation conditions	92%	Very high need
Need for evaluation-media development	100%	Very high need
Technology and infrastructure readiness	83%	Very high need
Expected media characteristics	100%	Very high need
Mean	94%	Very high need

The 26-student needs summary averaged 75.25%. Interest in interactive media and expectations for the evaluation medium were the highest dimensions at 85.58%, while access convenience was substantially lower at 51.74%. The lower access result qualifies the recommendation for a web-based package. It suggests that interest in a digital format did not necessarily mean that every student had equally convenient access. The development decision therefore needs to be paired with device, connectivity, timing, and alternative-access planning.

Table 4. Student Needs Questionnaire Summary (N = 26 Reported)

Assessed aspect	Percentage	Descriptive level
Digital device utilization	74.36%	Needed
Interest in interactive media	85.58%	Very high need
Need for digital evaluation	79.00%	Needed
Media access convenience	51.74%	Moderate need
Expectations for evaluation media	85.58%	Very high need
Mean	75.25%	Needed

Design Stage: Assessment Content and Configuration

The Design stage translated the identified need into a ten-item package. The documented mapping focused on theme, characterization, plot types, and moral messages. The intended progression was from recognizing explicit story information toward interpreting relationships and implied meaning. Nevertheless, the retained records do not contain the item-level blueprint, answer key, source short story, cognitive-level code for each item, or the distribution of questions across indicators. It is therefore not possible to confirm how many items measured remembering, understanding, or analysis.

This gap changes the appropriate claim. The package can be described as a short story comprehension activity aligned in general terms with selected intrinsic elements. It cannot be described as a validated measure of analytical competence. The 30-second limit is also difficult to justify for text-based interpretation unless the story or relevant extract was read before the timed question. Future versions should retain the stimulus, blueprint, answer key, expert comments on every item, and evidence from pilot item analysis.

Table 5. Documented Content Mapping and Evidentiary Limits

Short story element	Intended assessment focus	Evidence retained
Theme	Identify or infer the central idea	Included in the general mapping; item count and cognitive level unavailable
Characterization	Identify traits or infer them from actions and dialogue	Included in the general mapping; representative item unavailable
Plot	Recognize sequence and plot type	Included in the general mapping; distribution across items unavailable
Moral message	Identify or interpret an implied lesson	Included in the general mapping; scoring rationale unavailable

Develop Stage: Expert Validation and Revision

The two media validators produced an overall average of 89%, within the very feasible category. The aggregate masks a substantial difference between Validator 1 (78%) and Validator 2 (99%). The largest dimension-level difference occurred in navigation, with 73% and 100%. This disagreement may reflect different expectations for instruction clarity, navigation sequence, device behavior, or

technical performance. Averaging the scores is useful as a descriptive summary, but it should not be interpreted as inter-rater agreement.

Table 6. Media Validation Results

Validation Aspect	Validator 1	Validator 2	Mean	Category
Interface and display	76%	96%	86%	Very feasible
Navigation	73%	100%	87%	Very feasible
Interactivity	80%	100%	90%	Very feasible
Technical quality	80%	100%	90%	Very feasible
Media suitability	80%	100%	90%	Very feasible

The single material validator rated the package at 92%. Evaluation quality received 100%, while content alignment and material presentation were the lowest dimensions at 87%. These ratings support the validator's perception that the content was feasible, but one rater cannot establish stable content validity. A second independent material validator and item-level content-validity analysis are needed before stronger claims are made.

Table 7. Material Validation Results (One Validator)

Validation Aspect	Percentage	Category
Content alignment	87%	Very feasible
Content accuracy	93%	Very feasible
Linguistic quality	93%	Very feasible
Material presentation	87%	Very feasible
Evaluation quality	100%	Very feasible
Mean	92%	Very feasible

The language validators both produced 92% overall. Their dimension ratings were close for standardness and communicativeness, but differed by ten percentage points for student appropriateness. This difference suggests that wording considered accurate by one validator may still require adjustment for Grade XI readability or familiarity. The overall result supports feasibility while leaving room for readability testing with a broader student sample.

Table 8. Language Validation Results

Linguistic aspect	Validator 1	Validator 2	Mean	Category
Standardness	90%	95%	93%	Very feasible

Linguistic aspect	Validator 1	Validator 2	Mean	Category
Communicativeness	90%	95%	93%	Very feasible
Student appropriateness	95%	85%	90%	Very feasible
Overall	92%	92%	92%	Very feasible

The development summary did not preserve verbatim validator comments or a dated before-and-after product log. Table 9 therefore records the revision priorities supported by the available scores and the actions reported at the level of the final package. It does not claim item-level changes that cannot be verified.

Table 9. Revision Priorities and Reported Development Actions

Evidence or issue	Revision priority	Reported action and remaining limitation
Media navigation: 73% vs. 100%	Make access and movement through the activity clearer	Navigation directions were reviewed before the limited trial; dated screenshots were not retained
Interface/display: 76% vs. 96%	Improve readability and visual consistency	Display settings were reviewed; device-specific evidence was not retained
Content alignment and presentation: 87%	Recheck alignment and question sequence	Items were reviewed against selected short story elements; item-level blueprint was not retained
Student appropriateness: 95% vs. 85%	Simplify or clarify wording	Instructions and terminology were edited; readability statistics were not calculated
Access convenience: 51.74%	Reduce access barriers	Smartphone and computer access were allowed; no offline version or formal accommodation was documented

Limited Field Trial: Student and Teacher Responses

The student response questionnaire averaged 83.79%, within the study's very feasible category. Practicality was the highest dimension at 88.00%, while ease of operation was the lowest at 81.09%. The pattern suggests generally favorable perceptions after use, but the narrow range among dimensions and the absence of respondent-level variability prevent conclusions about whether all students had similar experiences. In particular, the favorable practicality response should be read

alongside the lower 51.74% access-convenience result obtained during needs analysis.

Table 10. Student Response Results (N = 26 Reported)

Evaluated aspect	Percentage	Category
Interface and display	83.97%	Very feasible
Ease of operation	81.09%	Very feasible
Material clarity	82.69%	Very feasible
Media attractiveness	84.94%	Very feasible
Perceived motivation and engagement	82.05%	Very feasible
Practicality	88.00%	Very feasible
Mean	83.79%	Very feasible

The teacher response averaged 92%. Because it represents one teacher, it is evidence of one informant's favorable evaluation, not general teacher acceptance. The 95% scores for interface and instructional benefit indicate which dimensions this teacher valued most; they do not establish instructional effectiveness.

Table 11. Teacher Response Results (One Informant)

Evaluated Aspect	Percentage	Category
Content alignment	90%	Very feasible
Linguistic quality	90%	Very feasible
Media interface	95%	Very feasible
Ease of use	90%	Very feasible
Instructional benefit	95%	Very feasible
Mean	92%	Very feasible

The classroom observation was reported as 98%, and student game scores were reported to range from 123 to 270. Neither result is used as evidence of engagement or comprehension. The observation protocol lacks information about its observer, dimensions, item count, number of sessions, and scoring rule. The performance range also cannot be reconstructed from a simple rule of ten items multiplied by 27 points because 123 is not a multiple of 27. Time-based scoring or platform behavior may have contributed, but the retained record does not permit confirmation. Repeated attempts further complicate interpretation. When an incorrect response can be retried without a stated deduction, a final score may represent a combination of first-attempt knowledge, elimination, persistence, response speed, and guessing.

The score may still support classroom feedback, but it should not be treated as a validated achievement measure. First-attempt accuracy, response time, item difficulty, discrimination, distractor functioning, and a separate learning-outcome measure should be retained in future testing.

The expert and user-response results place this study within the pattern reported by earlier Educandy development research. Nurhabibah et al. (2021), Mufid and Utsman (2022), and Purba and Dwi (2023) similarly found that teachers and learners could use Educandy-hosted activities as digital evaluation alternatives. The present study adds a Grade XI short story context, but its evidence remains focused on feasibility and perception. The high percentages should not be interpreted as showing that the package improves motivation or comprehension. This distinction matters because engagement, usability, and achievement are related but separate constructs. Game features can attract attention and encourage persistence, yet learning depends on item quality, feedback, prior knowledge, and instructional integration (Plass et al., 2015; Sailer & Homner, 2020). The student response item labeled motivation and engagement represents perceived experience after one trial. It is not a validated motivation scale, a behavioral engagement measure, or a comparison with a non-game condition.

The lower access-convenience percentage is one of the most informative findings. A web-based activity may be attractive while still creating unequal participation for students with limited devices, unstable connectivity, slower reading speed, or accessibility needs. This result tempers the otherwise favorable response profile. Practical adoption should include device-sharing arrangements, flexible timing, alternative formats, and explicit privacy procedures rather than assuming that a web link is equally accessible to all learners. The media-validator disagreement also deserves more attention than the overall average. A difference between 73% and 100% for navigation may indicate that the validation rubric allowed different interpretations or that validators tested the package under different conditions. Without item-level ratings and qualitative comments, the cause cannot be determined. A subsequent validation round should use clearly defined criteria, record device and browser conditions, preserve comments, and calculate an appropriate agreement or content-validity index.

From an assessment perspective, the package's ten timed multiple-choice questions are best used for supplementary formative checking. Immediate correctness information and repeated practice can help teachers notice wher e students struggle, but explanatory feedback would provide stronger support for learning (Black & Wiliam, 2009; Shute, 2008). Higher-order short story outcomes also require tasks that elicit reasoning, textual evidence, and alternative interpretations. A balanced assessment design should combine the Educandy activity with discussion, constructed responses, or written analysis. The precise contribution of the study is the documented local development of an Educandy-hosted short story assessment package and the reporting of feasibility judgments from several stakeholder roles. The study did not invent the platform, complete full dissemination, or establish

effectiveness. This bounded contribution is still useful because it identifies practical design priorities, validator disagreement, access concerns, and scoring limitations that should guide a stronger next version.

The study is limited by one school, one class of 26 students, one teacher informant, and one material validator. It uses self-report and expert-rating aggregates without item-level raw scores, missing-data information, score distributions, reliability estimates, or documented inter-rater agreement. The item blueprint, short story source, final access link, platform version, and screenshots were not retained in the manuscript record. Trial dates, duration, technical support, and accommodations were also insufficiently documented. No pretest-posttest or comparison design was used, so learning effectiveness cannot be inferred. The timed, repeated-attempt scoring system was not psychometrically evaluated, and the reported performance range cannot be reconstructed from the stated point rule. Formal ethics, consent, privacy, accessibility, and data-storage documentation was unavailable. Full dissemination beyond the development site was not conducted.

Future research should preserve the complete item bank and blueprint, involve at least two qualified content validators, report validator credentials and comments, calculate content-validity evidence, and pilot the response instrument for reliability. Field testing should report first-attempt accuracy, item difficulty, discrimination, distractor performance, response times, missing data, and variability. A broader multi-school study should compare learning outcomes with an appropriate alternative, document ethics and privacy safeguards, test accessibility, and evaluate whether the package remains usable under different device and connectivity conditions.

CONCLUSION

This study produced an Educandy-hosted assessment package containing ten multiple-choice questions for Grade XI short story learning. The completed development sequence comprised Define, Design, Develop, and a limited trial; full dissemination was not completed. The reported media, material, and language validation averages were 89%, 92%, and 92%, respectively. Student responses averaged 83.79%, and the single teacher response averaged 92%. These findings support the package's expert-rated feasibility and case-specific perceived usability, attractiveness, and practicality. They do not demonstrate increased motivation, improved participation, or better short story comprehension. Before wider adoption, the package requires stronger item-level documentation, broader and better-qualified validation, reliability and item-quality testing, clear privacy and accessibility safeguards, and an effectiveness study using appropriate comparison and learning-outcome measures.

REFERENCES

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Azizah, S. N., Karmila, N., & Muhajang, T. (2025). Pengembangan evaluasi pembelajaran menggunakan Educandy pada materi kekayaan budaya Indonesia. *Didaktik: Jurnal Ilmiah PGSD STKIP Subang, 11*(4), 261–279. https://doi.org/10.36989/didaktik.v11i04.8714

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

Mufid, M., & Utsman, A. F. (2022). Pengembangan media evaluasi pembelajaran berbasis game internet Educandy sebagai sarana pembelajaran daring. *Muróbbî: Jurnal Ilmu Pendidikan, 6*(2), 251–265. https://doi.org/10.52431/murobbi.v6i2.952

Nurhabibah, P., Fikriyah, F., & Dewi, K. (2021). Pengembangan website Educandy sebagai alat evaluasi pada pembelajaran Bahasa Indonesia untuk siswa kelas V. *Fon: Jurnal Pendidikan Bahasa dan Sastra Indonesia, 17*(2), 255–264. https://doi.org/10.25134/fon.v17i2.4652

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist, 50*(4), 258–283. https://doi.org/10.1080/00461520.2015.1122533

Purba, F. D. S., & Dwi, D. F. (2023). Pengembangan evaluasi hasil belajar berbasis digital dengan aplikasi Educandy pada tema panas dan perpindahannya siswa kelas V di SD Negeri 107423 Tanjung Garbus. *ARMADA: Jurnal Penelitian Multidisiplin, 1*(7), 731–739. https://doi.org/10.55681/armada.v1i7.706

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review, 32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Sugiyono. (2020). *Metode penelitian dan pengembangan (Research and development)*. Alfabeta.

Thiagarajan, S., Semmel, D. S., & Semmel, M. I. (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. Leadership Training Institute/Special Education, University of Minnesota.