SYNTIA – Systems and Technologies in Artificial Intelligence Applications.
Volume 1 Issue 1 (2025). Page: 25-32.

Augmented Reality–Based English Learning Media Using a Target Tracking Model Approach

Gunawan Mulyadi[1*], Rizka Ardiansyah[2], Hajra Rasmita Ngemba[3]

[1] *Informatics Engineering Study Program, Faculty of Engineering, Universitas Tadulako, Palu, Indonesia.*
[2] *Department of Information Technology, Faculty of Engineering, Universitas Tadulako, Palu, Indonesia.*
[3] *Doctoral Programme, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Malaysia.*

**Corresponding author: gunawanmulyadi238@gmail.com*

Abstract: Young learners often find English vocabulary hard because spelling differs from pronunciation, and teaching by explanation alone is monotonous. This study built BINGDAR, an Android application in which pointing a phone at a real object (fruit, classroom object, or animal picture cube) shows a rotatable, zoomable 3D model with audio and text, using Vuforia model targets in Unity, and evaluated it with 30 third-grade students. All 15 interface functions passed the black-box test. In the target-detection test, the models were recognized between 30 and 60 cm for all three categories but never at 5 cm, and the viewing angle mattered in only 2 of 45 distance–angle conditions. Mean scores rose within each of five pretest–posttest sessions (45→64, 65→75, 81→89, 94→96, 97→98); the shrinking percentage gains (42% to 1%) mostly reflect the ceiling, since the normalized gain stayed between 0.29 and 0.42. Because there was no control group, these gains cannot be attributed to the application alone. The mean System Usability Scale score was 91.3 (95% CI 89.8–92.8), and all six User Experience Questionnaire scales (2.03–2.44) exceeded the "excellent" benchmark even at the lower confidence limit, although perspicuity had low reliability (α = 0.50). The application is well accepted; evidence for its learning effect needs a comparison group and individual scores.

Keywords : augmented reality; English vocabulary; model target; markerless tracking; System Usability Scale.

Received: November 03, 2025 | Revised: November 27, 2025 | Accepted: December 17, 2025 | Published: December 31, 2025

1. Introduction

English is taught from early primary school in many countries, and young learners struggle with vocabulary because English spelling often does not match pronunciation; lessons that consist of explanation followed by exercises can also bore them. Augmented reality (AR), which overlays virtual objects on the real world in real time, has been proposed to make such lessons more engaging, and reviews report positive effects on motivation and learning, together with practical difficulties such as usability and technical problems [1]–[3].

AR applications need a way to know what the camera sees. Marker-based tracking recognizes printed markers such as QR codes or image cards, which must be prepared in advance and limit what children can explore [4]. Markerless approaches recognize the scene or an object directly, and pose-estimation methods differ in what they require from the object and the camera [5]. Among them, a model target recognizes a real object from a registered 3D model of its shape, so a child can point the phone at a fruit or a toy and see the matching content. The trade-off is that every registered model competes with the others, and objects of similar shape may trigger the wrong content. Related work in Indonesia includes English-learning AR for elementary students on Android [6] and a model-generated-target application that shows a 3D coronavirus model for any spherical object [7].

This article describes BINGDAR, an AR English-vocabulary application based on model targets, and evaluates it in a primary school. We ask three questions. (RQ1) Does the application work as specified, and

SYNTIA – Systems and Technologies in Artificial Intelligence Applications is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

under which distances and viewing angles are the model targets recognized? (RQ2) How do students' test scores change over repeated pretest–posttest sessions, and how much of the change can be attributed to the application? (RQ3) How do students rate its usability and user experience, and how reliable are these ratings? The contributions are an account of the design and the 18 registered targets, a reanalysis of the detection tables, the learning results with normalized gains in place of percentages alone, and a reanalysis of the raw System Usability Scale and User Experience Questionnaire data with confidence intervals and reliability. Section 2 describes the method, Section 3 reports and discusses the results, and Section 4 concludes.

2. Method

2.1. Participants and setting

The participants were 30 third-grade students of Madrasah Ibtidaiyah Swasta Karya Thayyibah (MIS KT) LIMRAN. Each student used the application on a smartphone. The students were 8 to 9 years old. The school and the parents gave permission, and the whole third-grade class took part, so the sample is a convenience sample and not a random draw. The source does not record the smartphone models or the lighting of the tests.

2.2. Application development

The application was built with the Waterfall model [8]: requirements analysis, design with UML diagrams (use-case, activity, sequence, and class diagrams), implementation, and testing. Eighteen 3D models were registered as model targets with the Vuforia Model Target Generator (version 9.8.5) and exported to Unity 2019.4.12f1: six fruits (mango, watermelon, orange, apple, banana, avocado), six classroom objects (globe, table, book, pencil, clock, chair), and six animals (chicken, cat, cow, goat, bird, rabbit). Fig. 1 shows the development steps and the evaluations. To limit confusion between targets, the scanning screen is organized by category, so that only the targets of the chosen category are active.

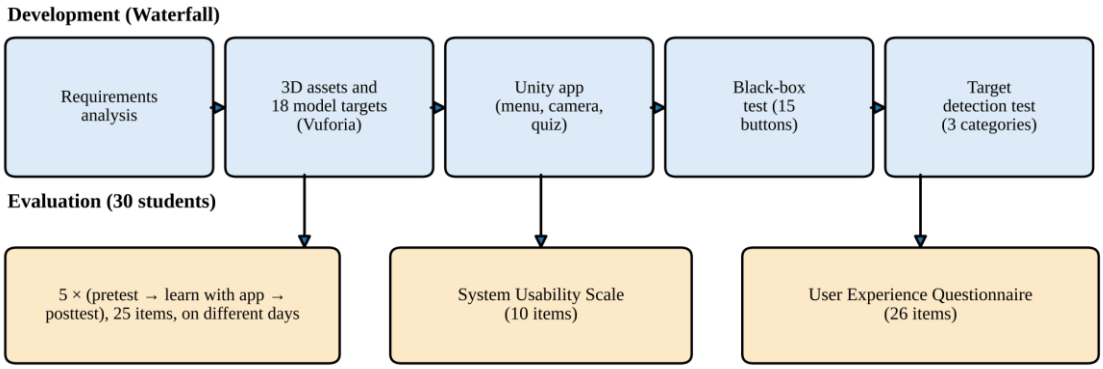


Figure 1. Development steps (top) and evaluations (bottom) of BINGDAR.

2.3. Evaluation design

Functional test. Fifteen buttons and functions were exercised once each and compared with the expected result [9]. *Target detection.* Each category was tested at five distances (5, 10, 30, 60, and 75 cm) and three viewing angles (30°, 55°, and 90°), with twelve scanned objects per category: the objects matching the registered targets and look-alike objects (for example a basketball for the globe, a pen for the pencil). A cell of the result table is marked as detected or not detected; the source does not define the mark further. *Learning outcome.* A one-group pretest–posttest design was repeated in five sessions on different days: a 25-item multiple-choice pretest, learning with the application, and a posttest. Only the class means of each test are reported. Besides the percentage change used in the source, we compute the normalized gain, the share of the possible improvement that was realized [10]:

$$\Delta_k = \frac{M_k^{\text{post}} - M_k^{\text{pre}}}{M_k^{\text{pre}}} \times 100\%, \qquad g_k = \frac{M_k^{\text{post}} - M_k^{\text{pre}}}{100 - M_k^{\text{pre}}} \tag{1}$$

where M^{pre}_k and M^{post}_k are the class means of session k on a scale of 0–100. *Usability.* The System Usability Scale (SUS) has ten items; the responses are converted to contributions c_{ij} from 0 to 4 and scaled to 0–100 [11], [12]:

$$\text{SUS}_i = 2.5 \sum_{j=1}^{10} c_{ij} \tag{2}$$

User experience. The User Experience Questionnaire (UEQ) has 26 semantic-differential items on a seven-point scale, transformed to −3…+3 and averaged into six scales; the scale means are compared with the published benchmark [13], [14]. For SUS and UEQ we recomputed the statistics from the per-student data of the source: the mean, the standard deviation of student-level scores, the 95% t-interval, and Cronbach's α [15] for the items of each scale. The data are ordinal ratings treated as interval, as is common in practice [16]. The UEQ matrix was read from the table printed in the source (30 students × 26 items), and our scale means reproduce all 26 item means and all six scale means of the source exactly.

3. Results and Discussion

3.1. The application

The main menu has Play, Quiz, and Exit buttons, with Info and Help pages (Fig. 2a). Play leads to the category screen (Fig. 2b) and then to the AR camera page (Fig. 2c), where the matching 3D model appears over the real object, with buttons for audio and text, rotation, and zoom in and out. The quiz contains 10 multiple-choice questions of 10 points each, on the vocabulary learned in the camera page (Fig. 2d).



Figure 2. BINGDAR screens: (a) main menu; (b) category selection; (c) AR camera page with the 3D mango; (d) quiz page.

3.2. Functional test and target detection (RQ1)

All 15 functions gave the expected result (Table 1), including navigation between pages, selection of the three categories, audio and text, enlarging, shrinking, and rotating the object, going back, and closing the application. Each was run once with valid input, so the test shows that the interface responds, not that it is robust to unusual input.

Table 1. Black-box test of the interface: 15 functions, one run each.

Group	Functions	Number	Result
Menu and navigation	Play, Quiz, Info, Help, Exit, Home	6	All as expected
Category selection	Objects, Fruits, Animals, category button	4	All as expected
AR camera controls	Audio and text, enlarge, shrink, rotate	4	All as expected
Back	Return to previous page	1	As expected

Table 2 and Fig. 3 summarize the detection test. No object was detected at 5 cm in any category. Detection worked in a band of 30–60 cm for all three categories: for fruits and animals all 36 cells at both distances are marked as detected, and for the objects 27 of 36 at 30 cm and 33 of 36 at 60 cm. At 10 cm the animals were detected at every angle and the fruits at 55° and 90° only; at 75 cm the fruits were not detected, the objects in 24 of 36 cells, and the animals only at 30°. The viewing angle mattered in just two of 15 distance conditions per category (fruits at 10 cm, animals at 75 cm), so distance, not angle, determines whether a model target is found, which agrees with what one expects when the recognition is based on the overall shape. Among the objects, small items such as the book and the pencil were lost at 75 cm and large ones such as the tables at 30 cm, which is consistent with the apparent size in the camera image. The pen, a look-alike of the pencil, was never marked as detected (0 of 15).

Table 2. Cells marked as detected (out of 36 per distance: 3 angles × 12 scanned objects) for each category and distance.

Distance (cm)	Objects	Fruits	Animals
5	0	0	0
10	0	24	36
30	27	36	36
60	33	36	36
75	24	0	12
All	84 (47%)	96 (53%)	120 (67%)

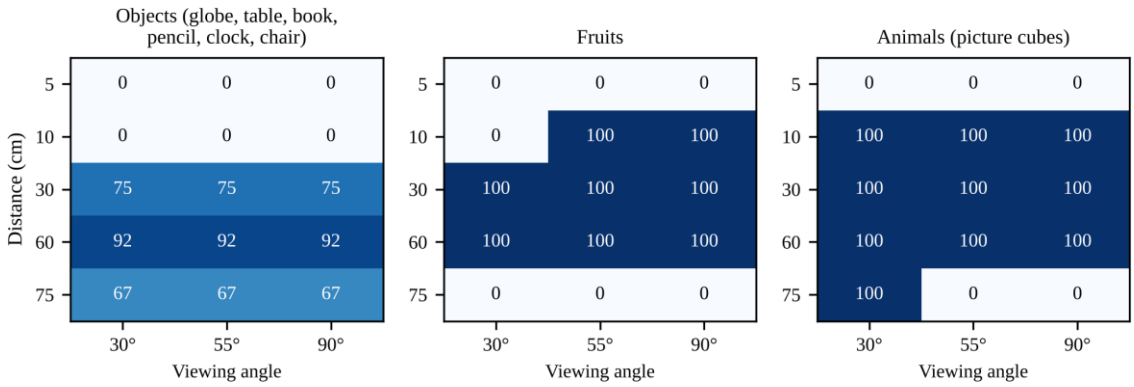


Figure 3. Percentage of the twelve scanned objects marked as detected, by distance and viewing angle, for each category.

Four points limit the interpretation. First, the mark is not defined: the table includes look-alike objects (basketball, tire, helmet, softball, boxes, Rubik's cube, and dice), and a mark for such an object means either a correct rejection or a false match. If a mark means that the 3D model appeared, the look-alikes were matched with the model in most cells at 30–60 cm, which is the failure that the category-wise scanning is meant to prevent; the source gives no count of false matches. Second, the animal targets are cubes with a picture of the animal on them, scanned against boxes, a Rubik's cube, and dice, so all targets have the same shape and shape-based recognition cannot separate them. Third, the titles of the two tables are swapped in the source (the table titled fruits lists the objects and the table titled objects lists the fruits), and in the fruit table the column headers pair model names with unrelated objects, so we group the results by the content of the table. Fourth, each cell is a single trial, with no information on lighting, device, or repetition. The source gives no counts of correct detections, misses, and false matches per object and no repeated trials under stated conditions.

3.3. Learning outcomes (RQ2)

Table 3 and Fig. 4 give the class means of the five sessions. In every session the posttest mean is higher than the pretest mean: by 19, 10, 8, 2, and 1 points. In percentage terms the gain falls from 42% to 1%, as

in the source. Our recomputation agrees except for session 3, where (89 − 81)/81 is 9.9%, i.e. 10%, not 9%; the first posttest mean is 64, and the 65 stated in the earlier version of this article is the pretest of session 2.

Table 3. Class means of the pretest and posttest (30 students, scale 0–100), gain in points, percentage gain Δ, and normalized gain g.

Session	Pretest	Posttest	Gain	Δ (%)	g
1	45	64	19	42.2	0.35
2	65	75	10	15.4	0.29
3	81	89	8	9.9	0.42
4	94	96	2	2.1	0.33
5	97	98	1	1.0	0.33

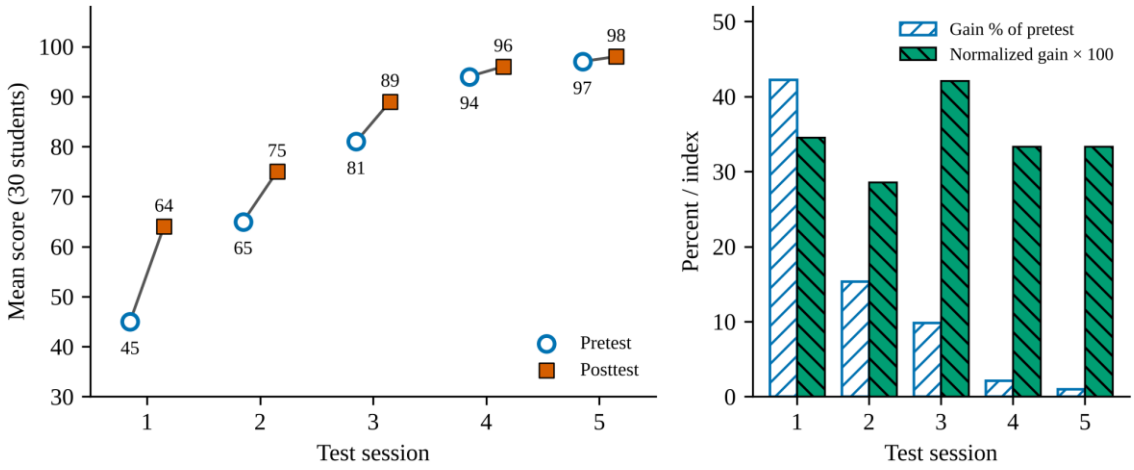


Figure 4. Pretest and posttest class means (left) and percentage gain and normalized gain (right) by session.

The shrinking percentage gain is largely a ceiling effect. The normalized gain stays between 0.29 and 0.42 (mean 0.34): each session closes about a third of the remaining distance to 100, whether the pretest is 45 or 97. The decline from 42% to 1% therefore does not mean that the application became less effective, and a pretest of 97 in session 5 leaves almost no room to show an effect. Scores also carry over between sessions: the pretest of each session is 1, 6, 5, and 1 points above the previous posttest, so the cumulative rise from the first pretest (45) to the last posttest (98) is 53 points, only 19 of which occurred inside the first session.

What this shows is limited by the design. There is one group and no control, and the same students were tested five times, so a rise can arise from familiarity with the test items, from ordinary lessons between sessions, and from maturation, as well as from the application. Only class means are reported, so no paired test, confidence interval, or effect size per student can be computed, and it is not stated whether the five tests used the same items or whether the test was reliable. The statement of the source that the application "improves motivation" is not supported either, as motivation was not measured. The data support the narrower statement that scores were higher after each learning session than before it. The source gives no individual scores and does not describe the teaching between sessions or any use of the application outside the sessions.

3.4. Usability and user experience (RQ3)

The mean SUS score of the 30 students is 91.3 (SD 4.0; 95% CI 89.8–92.8; range 82.5–97.5; median 92.5), against a commonly cited average of about 68 [12]. Every student scored above 80, so the application is rated in the highest usability categories, as the source states. Two properties of the data should be noted. The source prints only the converted contributions (all between 2 and 4), not the raw answers, and with such a narrow range the internal consistency of the ten items cannot be assessed (Cronbach's α is −0.15).

In addition, SUS and UEQ were designed for adult users; the students were third-graders who filled in the forms with the researchers' help, which tends to raise ratings.

Table 4 and Fig. 5 give the UEQ scales. All six scale means lie between 2.03 and 2.44, above the "excellent" threshold of the benchmark [14], and the lower ends of the 95% intervals (1.71–2.23) also exceed the thresholds (1.60–2.00), so the statement that all scales are excellent holds on the benchmark even allowing for sampling error. The overall mean of the 26 items, a simple UX index [17], is 2.27, and 52% of all 780 responses are the maximum, +3; only 8 responses (1.0%) are negative. Student 27 gave +3 to every item, and three students have an overall mean of 1.27 or less. The lowest item means are "conservative–innovative" (1.73), "unpredictable–predictable" (1.80), and "fast–slow" (1.93). The ranking of the scales, attractiveness highest and novelty lowest, is the same as in the source.

Table 4. UEQ scales for 30 students: mean, standard deviation of student-level scores, 95% confidence interval, Cronbach's α, and benchmark category of the point estimate [14].

Scale	Items	Mean	SD	95% CI	α	Benchmark
Attractiveness	6	2.44	0.55	2.23–2.64	0.81	Excellent (>1.84)
Perspicuity	4	2.29	0.53	2.09–2.49	0.50	Excellent (>2.00)
Efficiency	4	2.37	0.59	2.15–2.59	0.76	Excellent (>1.88)
Dependability	4	2.25	0.73	1.98–2.52	0.79	Excellent (>1.70)
Stimulation	4	2.14	0.77	1.86–2.43	0.77	Excellent (>1.70)
Novelty	4	2.03	0.87	1.71–2.36	0.63	Excellent (>1.60)

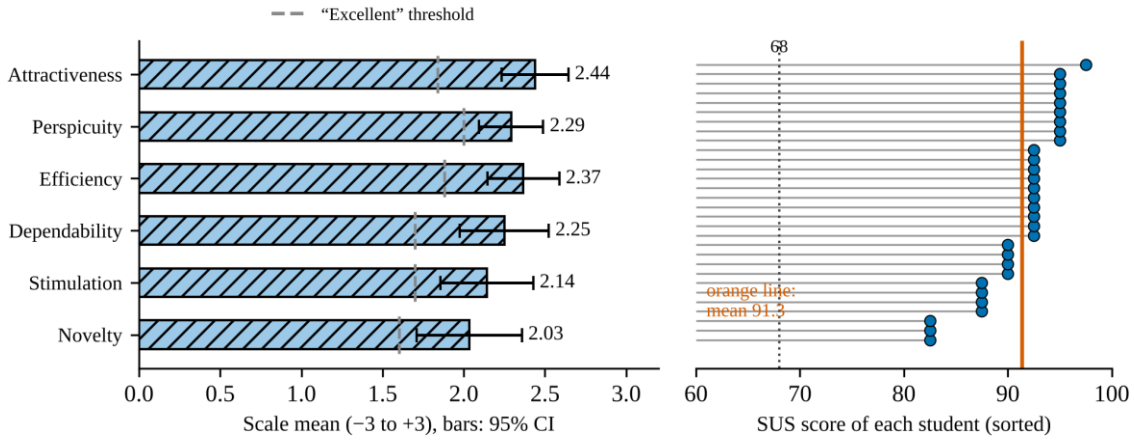


Figure 5. (a) UEQ scale means with 95% confidence intervals; dashed marks are the lower bounds of the "excellent" category [14]. (b) SUS score of each student, sorted; the dotted line marks 68.

Reliability is acceptable for attractiveness, efficiency, dependability, and stimulation (α = 0.76–0.81), borderline for novelty (0.63), and low for perspicuity (0.50); the whole instrument has α = 0.92. Because perspicuity has only four items and the sample is small, α is unstable [14], and the scale mean should be read together with its items. The three results (learning, SUS, UEQ) agree in direction, but they come from the same 30 students under guidance, so they are not independent confirmations.

3.5. Limitations

The study is one class in one school, with no control group and only class means for the learning outcome. The detection test has one trial per condition, undefined marks, and swapped labels. SUS and UEQ were completed by children with help, and the source keeps only transformed data. The comparison of AR with a non-AR version of the same lesson, the retention of vocabulary after weeks, and the pronunciation outcome, which the introduction names as the difficulty, were not measured. The number of targets (18) is small, and the performance with more registered models, where mismatches are expected to increase, is untested.

4. Conclusion

BINGDAR shows that model-target AR can present 3D vocabulary objects with audio and text on a phone: all 15 interface functions work, the targets are recognized from 30 to 60 cm in all three categories but not at 5 cm, and the viewing angle matters little. Thirty third-graders scored higher after each of five learning sessions, with a normalized gain of about one third per session, and rated usability (SUS 91.3) and user experience (all UEQ scales above the "excellent" benchmark, even at the lower confidence limit) very highly. The learning gain cannot be attributed to the application, because there is no control group, the same test was repeated, and the scores are near the ceiling; the detection test does not separate correct rejections from false matches.

The next study should compare the application with a lesson without AR using a control group and individual scores, test retention after a delay, report a confusion matrix of detections with look-alike objects under stated lighting, and collect usability data with an instrument validated for children.

AI Disclosure Statement

The authors used Generative AI and AI-Assisted Technologies throughout the preparation of this manuscript: to draft and revise portions of the text, to write and run the scripts behind the reanalysis of the SUS, UEQ, detection, and test data, to generate the figures, and to format the document from the authors' own research data. The authors verified the scripts and their outputs, reviewed the resulting text, cross-checked every reference against its original source, and accept full responsibility for the content of this publication.

Author Contributions Statement

Gunawan Mulyadi: Conceptualization, Methodology, Software, Investigation, Data Curation, Writing.
Rizka Ardiansyah: Software, Validation, Writing – Review and Editing.
Hajra Rasmita Ngemba: Methodology, Investigation, Formal Analysis, Supervision, Writing – Review and Editing.

References

[1] I. Radu, "Augmented reality in education: A meta-review and cross-media analysis," *Pers. Ubiquitous Comput.*, vol. 18, no. 6, pp. 1533–1543, 2014, doi: 10.1007/s00779-013-0747-y.

[2] M. Akçayır and G. Akçayır, "Advantages and challenges associated with augmented reality for education: A systematic review of the literature," *Educ. Res. Rev.*, vol. 20, pp. 1–11, 2017, doi: 10.1016/j.edurev.2016.11.002.

[3] J. Garzón, J. Pavón, and S. Baldiris, "Systematic review and meta-analysis of augmented reality in educational settings," *Virtual Real.*, vol. 23, no. 4, pp. 447–459, 2019, doi: 10.1007/s10055-019-00379-9.

[4] M. Billinghurst, A. Clark, and G. Lee, "A survey of augmented reality," *Found. Trends Hum.–Comput. Interact.*, vol. 8, no. 2–3, pp. 73–272, 2015, doi: 10.1561/1100000049.

[5] E. Marchand, H. Uchiyama, and F. Spindler, "Pose estimation for augmented reality: A hands-on survey," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 12, pp. 2633–2651, 2016, doi: 10.1109/TVCG.2015.2513408.

[6] F. S. Munir and P. Wanti, "Development of English learning media for elementary school students assisted augmented reality technology based Android," *Edumaspul J. Pendidik.*, vol. 6, no. 2, pp. 1839–1849, 2022, doi: 10.33487/edumaspul.v6i2.4392.

[7] R. Permana, E. P. W. Mandala, and D. E. Putri, "Augmented reality dengan model generate target dalam visualisasi objek digital pada media pembelajaran" (in Indonesian), *Maj. Ilm. UPI YPTK*, vol. 30, no. 1, pp. 7–13, 2023, doi: 10.35134/jmi.v30i1.143.

[8] W. W. Royce, "Managing the development of large software systems: Concepts and techniques," in *Proc. 9th Int. Conf. Softw. Eng. (ICSE)*, 1987, pp. 328–338.

[9] S. Nidhra, "Black box and white box testing techniques – A literature review," *Int. J. Embedded Syst. Appl.*, vol. 2, no. 2, pp. 29–50, 2012, doi: 10.5121/ijesa.2012.2204.

[10] R. R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *Am. J. Phys.*, vol. 66, no. 1, pp. 64–74, 1998, doi: 10.1119/1.18809.

[11] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the System Usability Scale," *Int. J. Hum.–Comput. Interact.*, vol. 24, no. 6, pp. 574–594, 2008, doi: 10.1080/10447310802205776.

[12] J. R. Lewis, "The System Usability Scale: Past, present, and future," *Int. J. Hum.–Comput. Interact.*, vol. 34, no. 7, pp. 577–590, 2018, doi: 10.1080/10447318.2018.1455307.

[13] B. Laugwitz, T. Held, and M. Schrepp, "Construction and evaluation of a user experience questionnaire," in *HCI and Usability for Education and Work*, Lecture Notes in Computer Science. Berlin, Germany: Springer, 2008, pp. 63–76, doi: 10.1007/978-3-540-89350-9_6.

[14] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Construction of a benchmark for the user experience questionnaire (UEQ)," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 4, no. 4, pp. 40–44, 2017, doi: 10.9781/ijimai.2017.445.

[15] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, 1951, doi: 10.1007/BF02310555.

[16] G. M. Sullivan and A. R. Artino, "Analyzing and interpreting data from Likert-type scales," *J. Grad. Med. Educ.*, vol. 5, no. 4, pp. 541–542, 2013, doi: 10.4300/JGME-5-4-18.

[17] A. Hinderks, M. Schrepp, F. J. Domínguez Mayo, M. J. Escalona, and J. Thomaschewski, "Developing a UX KPI based on the user experience questionnaire," *Comput. Stand. Interfaces*, vol. 65, pp. 38–44, 2019, doi: 10.1016/j.csi.2019.01.007.