ORIGINAL ARTICLE

DEVELOPMENT AND EVALUATION OF A HIGHER-ORDER THINKING NAHWU ASSESSMENT INSTRUMENT

Dwi Elis Nawati[1*] Jumhur[2] Qoim Nurani[3]
[123] Universitas Islam Negeri Raden Fatah, Indonesia
*Corresponding Author: elisnawatidwi@gmail.com

ABSTRACT
Higher-order thinking assessment can move Arabic grammar evaluation beyond recall toward the analysis and application of grammatical rules. This study aimed to develop, validate, and conduct a classroom evaluation of a higher-order thinking nahwu assessment instrument delivered through Quizizz. A simplified research-and-development procedure informed by Borg and Gall was used, followed by a quasi-experimental pretest-posttest comparison involving 35 students in an experimental class and 35 in a control class. Material- and media-expert validation yielded feasibility scores of 94% and 84%, respectively. The experimental-class mean increased from 52.28 at pretest to 85.42 at posttest, while the control-class mean increased from 50.14 to 62.57. The posttest difference was statistically significant, $t(68) = -7.799$, $p < .001$, with an estimated absolute Cohen's d of 1.86. Mean normalized gain was 0.626 (SD = 0.208) in the experimental class and 0.259 (SD = 0.123) in the control class. The findings support the product's feasibility and indicate improved nahwu learning outcomes under the conditions of this single-site trial. Because item design, platform features, and classroom implementation were delivered as a package, the study does not isolate their separate effects. Direct claims about critical thinking, engagement, and transfer require validated measures and multisite evaluation.

Keywords: Arabic grammar, digital assessment, higher-order thinking, Quizizz

INTRODUCTION

Nahwu is the branch of Arabic grammar that explains sentence structure, syntactic function, and changes in final inflection (iʿrāb). Mastery of these relations is important for accurate reading and sentence construction, yet learners often experience nahwu as abstract and rule-intensive. At the study site, preliminary observation indicated that evaluation largely required the recall of rules and basic identification of iʿrāb. Students consequently had limited opportunities to diagnose grammatical errors, compare alternative analyses, or construct rule-governed sentences. This practical mismatch between the intended analytical use of grammar and recall-oriented classroom assessment motivated the present development study.

Assessment should elicit evidence of what learners can do with disciplinary knowledge rather than only whether they can reproduce it. The revised taxonomy distinguishes remembering and understanding from more complex processes of applying, analyzing, evaluating, and creating (Anderson & Krathwohl, 2001). Higher-order questions therefore need contextual prompts, plausible alternatives, and scoring rules that require reasoning. Brookhart (2010) emphasizes that cognitive demand lies in the thinking required by a task, not in surface difficulty. A digital multiple-choice item may be higher order when it requires analysis of an unfamiliar construction, whereas a difficult-looking item may remain lower order if it merely asks for memorized terminology.

The validity of a higher-order assessment cannot be inferred from expert approval alone. Validity concerns the interpretation and use of scores and requires a coherent argument supported by content, response-process, internal-structure, and consequential evidence (American Educational Research Association et al., 2014; Messick, 1995). Item development should therefore document the construct, blueprint, cognitive-level mapping, scoring, item performance, and reliability. Expert review is valuable for content relevance and representation, but transparent reviewer qualifications, rating procedures, and agreement indices are needed (Downing, 2003). These principles frame the 94% and 84% feasibility scores in this study as

evidence of judged appropriateness, not definitive proof of score validity or learning effectiveness.

Digital quiz platforms can facilitate rapid delivery, immediate scoring, repeated practice, and feedback. Formative assessment becomes productive when evidence is used to reduce the gap between current and intended performance (Black & Wiliam, 2009). Feedback effects vary with content and design, but synthesis research generally supports feedback that is timely, task-focused, and actionable (Hattie & Timperley, 2007; Wisniewski et al., 2020). Retrieval practice can also strengthen learning compared with repeated study, particularly when tests prompt effortful recall and are followed by feedback (Adesope et al., 2017; Roediger & Karpicke, 2006). Thus, a Quizizz-delivered instrument may support learning through several mechanisms, including retrieval, feedback, attention, and opportunities to correct errors.

Gamification introduces game-design elements such as points, competition, time constraints, and visual progress into non-game activities. Meta-analytic evidence indicates small positive average effects on cognitive, motivational, and behavioral outcomes, alongside substantial variation across designs and contexts (Sailer & Homner, 2020). Specific elements can influence psychological need satisfaction differently; combinations of goals, feedback, and social comparison should therefore be designed deliberately (Sailer et al., 2017). Reviews also warn that novelty, competitive pressure, accessibility, and implementation quality can moderate outcomes (Dichev & Dicheva, 2017; Zainuddin et al., 2020). The platform should consequently be treated as an instructional environment rather than an active ingredient whose effect is assumed.

Research on game-based response systems provides relevant but indirect evidence. A review of Kahoot! studies reported generally favorable effects on achievement, classroom dynamics, attitudes, and anxiety, while noting design and methodological limitations (Wang & Tahir, 2020). A controlled study of gamified quizzes likewise found that classroom gamification can influence learning, although effects depend on how quizzes are integrated (Sanchez et al., 2020). Quizizz-specific work suggests that self-paced multiplayer quizzes can support participation and immediate response, but much of this evidence is contextual and should not be generalized automatically (Zhao, 2019). The present study differs by combining higher-order item design with Arabic syntactic content and a classroom comparison.

Within the literature reviewed for this revision, evidence is abundant for gamification and formative assessment in general but limited for a validated higher-order nahwu instrument delivered through Quizizz. The novelty is therefore practical and contextual: the product translates analytical nahwu outcomes into digitally delivered items, subjects the product to expert judgment, and examines classroom learning outcomes. It does not propose a new theory of gamification or establish that Quizizz is superior across settings. Accordingly, this study aimed to develop the instrument, examine its material and media feasibility, and compare nahwu learning outcomes between a class using the developed product and a class receiving conventional evaluation.

METHOD

Research design and setting. The study used a research-and-development design informed by the Borg and Gall approach and was conducted at MA Patra Mandiri Plaju, Palembang, South Sumatra, Indonesia. The implemented sequence consisted of problem identification, information collection, product design, design validation, design revision, product testing, and final product revision. The sequence was simplified for a bounded school-level product trial; it should not be interpreted as completion of every stage of a large-scale educational product-development program.

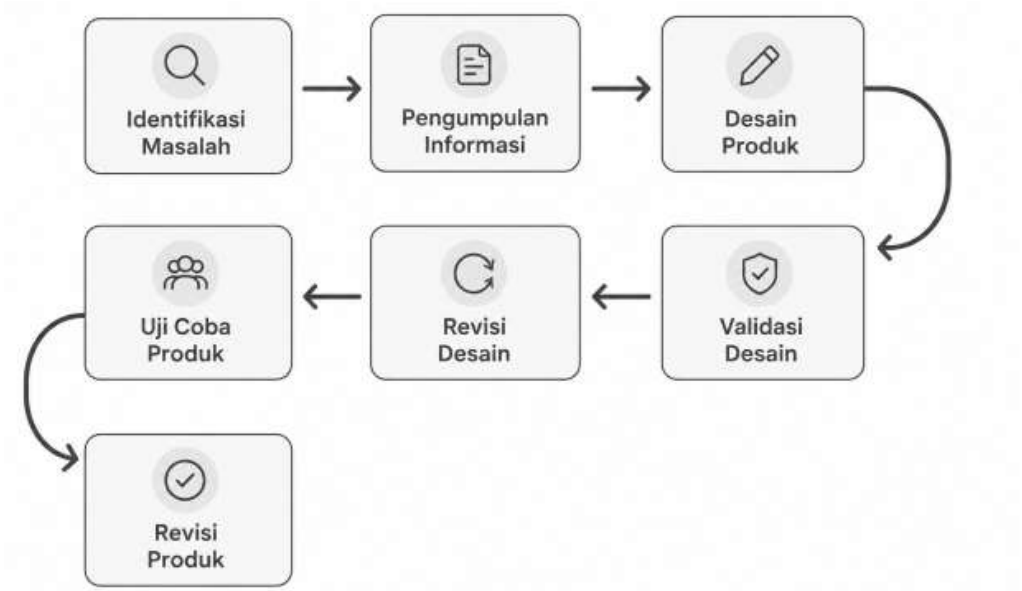

Figure 1. Simplified development stages used in the study.

Participants and sampling. Statistical output identified 35 students in the experimental class and 35 in the control class (N = 70). Students were classroom users of the assessment, while material and media experts judged product feasibility. The source record did not specify grade, age, gender, recruitment, class-allocation procedure, study period, or validator number and qualifications. These omissions limit evaluation of selection bias and transferability. They are reported transparently rather than filled with assumptions and must be completed from institutional records before journal submission.

Instrument development. Observation sheets, expert-validation questionnaires, a student-needs questionnaire, and nahwu learning-outcome tests were used. Observation and document review identified assessment needs. The development team prepared an item grid, wrote tasks intended to assess higher-order use of nahwu, implemented the items in Quizizz, sought expert review, and revised wording and display. Content alignment should be judged against the intended rules, syntactic functions, and cognitive operations. In line with contemporary standards, a complete submission should attach the blueprint, representative items, scoring rules, item difficulty and discrimination, reliability estimates, validator-rating method, and evidence that intended response processes occurred (American Educational Research Association et al., 2014; Downing, 2003).

Classroom procedure. The experimental class completed the Quizizz-delivered higher-order assessment, whereas the control class received conventional evaluation. Pretest and posttest scores were collected for both classes. The available source did not define intervention duration, session number, teacher equivalence, time on task, feedback settings, devices, connectivity, contamination controls, or fidelity checks. These factors are important because game-based learning depends on the interaction of cognitive, motivational, affective, and sociocultural features (Plass et al., 2015). Any causal interpretation must therefore concern the intervention package rather than the platform alone.

Data analysis. Shapiro-Wilk tests were used to examine score distributions. Posttest means were compared using an independent-samples t test. Because Levene's test was nonsignificant, the equal-variances row was interpreted. The standardized group difference was estimated from the reported t statistic and equal group sizes. Effect sizes complement p values by expressing magnitude, but they should be accompanied by uncertainty and interpreted in relation to design

quality (Lakens, 2013). Normalized gain was summarized as (posttest − pretest)/(maximum score − pretest). Gain scores are descriptive here; baseline-adjusted analysis is preferable when individual-level data are available because analysis of covariance can address residual baseline imbalance and often improves precision (Vickers & Altman, 2001).

Ethics and data governance. The supplied manuscript did not document ethics approval, institutional permission, student assent, guardian consent, confidentiality, or governance of data generated through Quizizz. Before submission, the authors should report the applicable review or exemption, consent process, account arrangements, data minimization, anonymization, access, retention, and data-availability conditions. No statement has been invented in this revision.

RESULTS AND DISCUSSION

Needs Analysis and Product Development

Observation, teacher interviews, student worksheets, and lesson plans indicated that existing nahwu evaluation relied largely on conventional lower-order questions and made limited use of digital media. The student-needs questionnaire produced an average score of 86.5%, interpreted in the source as strong demand for attractive, varied, digitally delivered evaluation. Because respondent characteristics, items, scoring, and category thresholds were not included in the source record, this percentage is preliminary evidence of perceived need rather than a population estimate.

The product was designed around a grid linking learning indicators, nahwu content, intended cognitive processes, and item presentation. Tasks were implemented in Quizizz to provide a self-paced interface and rapid scoring. This design is consistent with formative-assessment research when responses are used to identify misconceptions and guide next steps (Black & Wiliam, 2009; Gikandi et al., 2011). It also reflects evidence that repeated testing can improve retention, although benefit depends on meaningful retrieval and feedback rather than digitization itself (Adesope et al., 2017). Figure 2 shows the authoring and learner-facing interfaces.

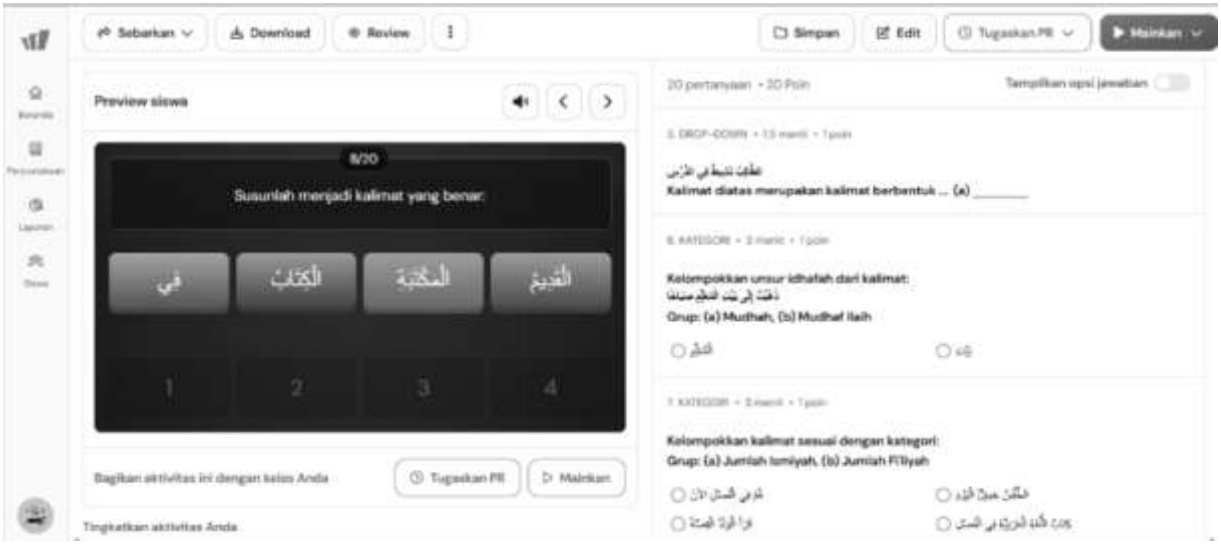

Figure 2. Quizizz interface used to design and deliver nahwu items.

Expert Validation and Revision

Material experts assigned a feasibility score of 94%, categorized as very good and valid in the source manuscript, while media experts assigned 84%, categorized as good. These ratings support the perceived relevance of the content and appropriateness of the digital presentation.

They do not, by themselves, establish reliability, construct validity, or effectiveness. A defensible validity argument must connect the intended interpretation of scores with multiple sources of evidence (Messick, 1995). Reporting the validators' expertise, scale, calculation, agreement, and item-level judgments would allow readers to evaluate the strength of this evidence.

Expert comments led to revision of item wording, alignment of cognitive demand, and improvement of the display. Figure 3 illustrates one before-and-after change. The revision process is pedagogically important because game elements cannot compensate for ambiguous stems, implausible options, or weak content representation. The final article should include a revision matrix and sample items to show exactly how feedback changed the product.

Figure 3. Example of the product before and after expert-guided revision.

Table 1. Pretest and Posttest Mean Scores by Class

Class	Pretest mean	Posttest mean
Experimental (n = 35)	52.28	85.42
Control (n = 35)	50.14	62.57

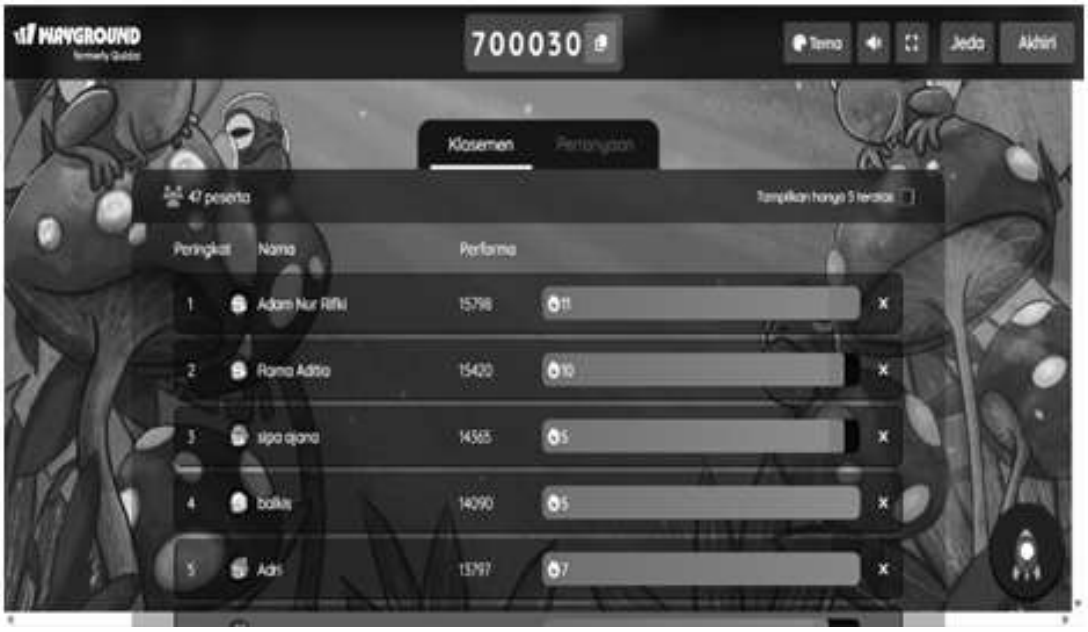

Figure 4. Quizizz results interface during classroom implementation.

The classes had similar, though not identical, pretest means. The experimental-class mean increased by 33.14 points and the control-class mean by 12.43 points. At posttest, the experimental mean exceeded the control mean by 22.85 points based on the descriptive table. The SPSS t-test output used a group ordering that produced a reported mean difference of −21.00; this inconsistency between the descriptive means and inferential output should be reconciled against the raw dataset before publication. Figure 4 displays the classroom results interface but does not provide evidence of implementation fidelity.

Table 2. Shapiro-Wilk Tests of Normality

Measure	Class	W	df	p
Pretest	Experimental	.940	35	.058
Posttest	Experimental	.966	35	.347
Pretest	Control	.941	35	.062
Posttest	Control	.976	35	.611

All reported Shapiro-Wilk p values exceeded .05. The distributions were therefore treated as approximately normal. Nonsignificant normality tests do not demonstrate perfect normality, so graphical checks and robust analysis would be useful if raw scores become available.

Table 3. Independent-Samples t Test for Posttest Scores

Statistic	Value
Levene's F	3.340
Levene's p	.072
t	−7.799
df	68
p (two-tailed)	< .001
Mean difference	−21.000
Standard error	2.693
95% CI	[−26.373, −15.627]
Estimated |Cohen's d|	1.86

Levene's test was nonsignificant, supporting interpretation of the equal-variances row. The posttest comparison was statistically significant, $t(68) = -7.799$, $p < .001$. Based on the reported t statistic and equal group sizes, the estimated absolute Cohen's d was 1.86, a large standardized difference. The negative sign reflects group coding; the descriptive means indicate higher performance in the experimental class. Because allocation was not reported and no baseline-adjusted model was available, this magnitude should not be interpreted as an unbiased causal effect. Pretest-posttest control designs require attention to baseline balance, regression to the mean, and the correlation between measurements (Morris, 2008; Vickers & Altman, 2001).

Table 4. Descriptive Normalized-Gain Statistics by Class

Statistic	Control	Experimental
n	35	35
Mean	.2594	.6258
Standard deviation	.12293	.20840
95% CI lower	.2172	.5542
95% CI upper	.3016	.6974
Minimum	.07	.11
Maximum	.57	1.00

Mean normalized gain was 0.6258 in the experimental class and 0.2594 in the control class. The source classified the experimental value as medium. Normalized gain was popularized as a descriptive index of improvement relative to the maximum possible gain (Hake, 1998), but it can be sensitive to starting score and ceiling effects. It should therefore supplement, not replace, raw-score descriptives and a baseline-adjusted group comparison.

Interpretation, Implications, and Limitations

Taken together, the findings show favorable expert judgments and better observed nahwu test performance in the class receiving the developed assessment package. The result is compatible with several mechanisms: cognitively demanding retrieval, immediate task feedback, visual progress, repetition, and increased attention. Meta-analysis suggests that gamification has small positive average effects but that outcomes vary across contexts and configurations (Sailer & Homner, 2020). Studies of specific design elements similarly show that badges, leaderboards, and performance graphs affect learners differently (Sailer et al., 2017). The present design cannot determine which mechanism produced the observed difference.

The study's strongest contribution is the alignment of a local curricular problem with a concrete assessment-development process. Higher-order nahwu assessment requires learners to interpret syntactic relations, evaluate erroneous analyses, and apply rules to unfamiliar sentences. This is more defensible than labeling every digital or difficult item as higher order. The product should ultimately be supported by a published blueprint and psychometric evidence. Without item-level evidence, the study supports feasibility and classroom promise, not a fully validated scale of higher-order thinking.

A stronger next-stage validation study should begin with an explicit construct map. Nahwu content can be crossed with cognitive processes so that each cell specifies the evidence expected from students. For example, an analysis item may ask learners to distinguish the grammatical function of a word across two sentence contexts, whereas an evaluation item may require selection and justification of the best correction for a malformed construction. Items should then be reviewed for linguistic accuracy, cultural accessibility, cognitive demand, and the plausibility of distractors. Pilot data can be used to examine difficulty, discrimination, distractor functioning, internal consistency, and differential performance across relevant groups. These steps would link expert judgment to observable score behavior and clarify whether total scores support the interpretations proposed for them (American Educational Research Association et al., 2014; Downing, 2003).

Digital delivery also creates design decisions that may alter the construct being assessed. Strict time limits can unintentionally measure reading speed or device fluency; public leaderboards can change effort and anxiety; random item order may improve security but complicate the interpretation of shared classroom experiences; and immediate correctness feedback may support learning while also exposing answers during loosely supervised administration. For high-stakes use, equivalent access, stable connectivity, identity controls, and documented accommodations are essential. For formative use, the platform's greatest value may lie in rapid diagnostic information rather than ranking. Teachers can group errors by nahwu concept, revisit common misconceptions, and ask students to explain why distractors are incorrect. Such follow-up turns a quiz event into an assessment-for-learning cycle consistent with formative-assessment principles (Black & Wiliam, 2009; Nicol & Macfarlane-Dick, 2006).

The findings are directionally consistent with research on gamified classroom quizzes. Sanchez et al. (2020) found that gamified quizzes can influence learning, while Wang and Tahir's (2020) review identified generally favorable patterns alongside methodological

limitations. Quizizz offers self-paced delivery and immediate response (Zhao, 2019), but technology is not inherently formative. Effective feedback must communicate information that students can use, and its effect depends on whether it directs attention to the task and learning process (Hattie & Timperley, 2007; Wisniewski et al., 2020). Teachers should therefore review error patterns, explain misconceptions, and connect platform results to follow-up instruction.

For practice, Arabic teachers may adapt the sequence by specifying learning outcomes, constructing a content-by-cognition blueprint, reviewing items, piloting the instrument, examining item statistics, and using results for remediation. Implementation should define time limits, feedback timing, item order, retry settings, and procedures to reduce guessing and unauthorized collaboration. Schools should also consider device availability, internet stability, accessibility, teacher competence, assessment security, and data privacy. Gamification may increase participation, but competitive features should be optional or carefully managed because learners differ in how they respond to social comparison (Dichev & Dicheva, 2017; Smiderle et al., 2020).

The practical interpretation of the large standardized difference requires particular caution. Large effects are possible in short, focused interventions, but they may also reflect teacher effects, unequal instructional exposure, unfamiliarity with the control assessment, or alignment between practice and the posttest. Reporting raw-score distributions, confidence intervals, attrition, and individual change would help readers evaluate the pattern. A future trial could hold item content constant while varying delivery mode, or use a factorial design that separately varies cognitive demand and gamified features. Delayed posttesting would indicate whether gains persist after platform novelty and immediate feedback have faded. Qualitative interviews or think-aloud protocols could further show whether students actually applied syntactic reasoning rather than learning platform-specific cues.

Several limitations constrain interpretation. The study involved one madrasah and two intact classes, with no reported allocation procedure or participant demographics. Intervention duration, teacher equivalence, fidelity, raw-score standard deviations, missing-data rules, and item reliability were not reported. The expert-validation process lacked validator characteristics and agreement statistics. Critical thinking and engagement were not directly measured, and no delayed outcome was collected. The discrepancy between the descriptive posttest difference and the t-test mean difference also requires verification. Finally, ethical review, consent, and Quizizz data governance were not documented. These limitations restrict causal inference, replication, and generalization. Future research should use multisite samples, complete psychometric reporting, baseline-adjusted analysis, direct measures of the intended cognitive processes, and longer follow-up.

CONCLUSION

A higher-order thinking nahwu assessment instrument was developed and delivered through Quizizz. Expert feasibility ratings were favorable, and the experimental class showed higher posttest performance and normalized gain than the control class in this single-site trial. The evidence supports the product's feasibility and an association with improved nahwu learning outcomes under the reported conditions. It does not isolate the effects of higher-order item design, gamification, feedback, or platform delivery, and it does not directly demonstrate improved critical thinking or engagement. Publication-quality reporting requires completion of participant and validator information, intervention details, psychometric evidence, ethics and data-governance statements, and reconciliation of the descriptive and inferential mean differences.

REFERENCES

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. Review of Educational Research, 87(3), 659–701. https://doi.org/10.3102/0034654316689306

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

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.

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.

Dichev, C., & Dicheva, D. (2017). Gamifying education: What is known, what is believed and what remains uncertain: A critical review. International Journal of Educational Technology in Higher Education, 14, Article 9. https://doi.org/10.1186/s41239-017-0042-5

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. Medical Education, 37(9), 830–837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. Computers & Education, 57(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics, 66(1), 64–74. https://doi.org/10.1119/1.18809

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, Article 863. https://doi.org/10.3389/fpsyg.2013.00863

Looyestyn, J., Kernot, J., Boshoff, K., Ryan, J., Edney, S., & Maher, C. (2017). Does gamification increase engagement with online programs? A systematic review. PLOS ONE, 12(3), e0173403. https://doi.org/10.1371/journal.pone.0173403

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. Organizational Research Methods, 11(2), 364–386. https://doi.org/10.1177/1094428106291059

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education, 31(2), 199–218. https://doi.org/10.1080/03075070600572090

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

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. Computers in Human Behavior, 69, 371–380. https://doi.org/10.1016/j.chb.2016.12.033

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

Sanchez, D. R., Langer, M., & Kaur, R. (2020). Gamification in the classroom: Examining the impact of gamified quizzes on student learning. Computers & Education, 144, 103666. https://doi.org/10.1016/j.compedu.2019.103666

Smiderle, R., Rigo, S. J., Marques, L. B., Coelho, J. A. P. M., & Jaques, P. A. (2020). The impact of gamification on students' learning, engagement and behavior based on their personality traits. Smart Learning Environments, 7, Article 3. https://doi.org/10.1186/s40561-019-0098-x

Vickers, A. J., & Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. BMJ, 323(7321), 1123–1124. https://doi.org/10.1136/bmj.323.7321.1123

Wang, A. I., & Tahir, R. (2020). The effect of using Kahoot! for learning: A literature review. Computers & Education, 149, 103818. https://doi.org/10.1016/j.compedu.2020.103818

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. Frontiers in Psychology, 10, Article 3087. https://doi.org/10.3389/fpsyg.2019.03087

Zainuddin, Z., Chu, S. K. W., Shujahat, M., & Perera, C. J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. Educational Research Review, 30, 100326. https://doi.org/10.1016/j.edurev.2020.100326

Zhao, F. (2019). Using Quizizz to integrate fun multiplayer activity in the accounting classroom. International Journal of Higher Education, 8(1), 37–43. https://doi.org/10.5430/ijhe.v8n1p37