ISEJ Indonesian Science Education Journal

ISSN Online: 2716-3350
https://siducat.org/index.php/isej/
Published by Yayasan Darussalam Bengkulu

OPEN ACCESS ARTICLE

Development and Preliminary Evaluation of Gamified HOTS Items for the Digital-Society Element of Phase D Informatics

Uswatun Hasanah[1], Syahrul[2], Anas Arfandi[3]
Vocational and Technology Education Study Program, Universitas Negeri Makassar, Makassar, Indonesia[1,2,3]
E-mail Corresponding: uswatunhasanah511@guru.smp.belajar.id

Received: March 9, 2026	Revised: May 10, 2026	Accepted: May 21, 2026

Abstract
This study developed and preliminarily evaluated a gamified Higher-Order Thinking Skills (HOTS) item bank for the Digital-Society (Dampak Sosial Informatika) element of Phase D Informatics. The initial pool comprised 20 four-option contextual multiple-choice items covering digital ethics, digital footprints and privacy, cyberbullying, and data security; 10 items targeted analysis (C4) and 10 targeted evaluation (C5). ADDIE served as the overarching development framework, while 4-D informed product specification and Tessmer informed formative evaluation. Expert appraisal involved one specialist each in Informatics content, educational assessment, and language. Field testing involved 60 Grade VIII students from three public junior secondary schools in Pinrang Regency, Indonesia. Domain-normalized expert-appraisal scores ranged from 90.77% to 98.19%. Exploratory classical-test screening identified 17 provisionally functioning items, whereas Items 6, 11, and 12 showed weak item-total associations and poor discrimination and remain pending revision and retesting. Cronbach's alpha for the unrevised 20-item pool was .735. Teacher and student practicality scores were reported separately (91.96% and 81.65%, respectively), because the questionnaires represented different respondent groups. The findings provide context-bound, preliminary evidence of item-bank feasibility and functioning; they do not demonstrate that gamification or pedagogical deep learning improved HOTS, computational reasoning, or learning outcomes.

Keywords: HOTS assessment; digital society; Phase D Informatics; gamification; preliminary item analysis

INTRODUCTION

Higher-Order Thinking Skills (HOTS) refer to cognitive processes that require learners to reorganize and apply knowledge rather than reproduce memorized information. In assessment, these processes are represented when students must interpret evidence, analyze relationships, compare alternatives, and evaluate the defensibility of a response (Lewis & Smith, 1993). The present study uses HOTS as the intended cognitive demand of selected-response items; it does not treat HOTS as synonymous with computational thinking, pedagogical deep learning, or gamification.

The relevance of these demands is visible in international assessment results. PISA 2022 reported continuing difficulties among Indonesian students in reasoning, interpreting information, and solving non-routine problems (OECD, 2023). This broad evidence does not by itself establish a deficit in Phase D Informatics, but it highlights assessment processes that are also required when students interpret digital situations, judge the credibility and consequences of online actions, and select defensible responses to privacy or security problems. The link to

Informatics therefore concerns shared reasoning demands rather than direct equivalence between PISA domains and the school subject.

Assessment can influence what teachers emphasize and what students practice. When classroom tests are dominated by factual recall, instruction may similarly privilege recognition and memorization. Conversely, tasks that require interpretation, comparison, justification, and knowledge transfer can provide more appropriate evidence for curriculum outcomes involving reasoning (Pellegrino, 2014). Validity, however, concerns the interpretation and use of scores and must be supported by multiple sources of evidence rather than by item appearance alone (Messick, 1995; American Educational Research Association et al., 2014).

Informatics includes several related but distinct constructs. Computational thinking concerns processes such as decomposition, abstraction, algorithmic reasoning, debugging, and solution evaluation (Wing, 2006; Grover & Pea, 2013; Weintrop et al., 2016; Shute et al., 2017). HOTS describes the cognitive level demanded by a task. Pedagogical deep learning describes an instructional orientation, while gamification describes the use of game-design features in delivery. Conceptual proximity among these constructs does not mean that evidence for one validates the others.

The official Phase D Informatics curriculum includes eight elements. The present item bank was not designed to represent the entire Phase D domain. It focused on the Digital-Society element (Dampak Sosial Informatika), whose outcomes include understanding the availability and openness of information, distinguishing public from private information, applying ethical conduct, and protecting oneself in digital society (Badan Standar, Kurikulum, dan Asesmen Pendidikan, 2025). The title and claims were therefore narrowed to this element.

Within that element, the items addressed four content clusters: digital ethics and online communication, digital footprints and privacy, cyberbullying, and data security. These clusters are relevant to the official Digital-Society outcomes, but they do not provide representative coverage of the broader Phase D elements of computational thinking, information and communication technology, computer systems, networks, data analysis, algorithms and programming, or cross-disciplinary computational projects. Score interpretations must consequently remain restricted to the sampled digital-society content.

Multiple-choice items can elicit analysis and evaluation when the stimulus contains essential information, the alternatives require comparison of plausible reasoning paths, and the distractors reflect misconceptions or incomplete reasoning (Haladyna et al., 2002; Scully, 2017). The current item bank therefore targets C4 analysis and C5 evaluation only. It does not claim to assess C6 creation, because students select among provided options rather than generate and defend an original product or solution with an appropriate scoring rubric.

High-quality assessment development requires explicit alignment among the intended content, cognitive process, item task, scoring rule, and proposed score interpretation. It also requires empirical examination of item functioning, internal structure, relations to external variables, and possible consequences of score use (American Educational Research Association et al., 2014). In this study, only early evidence from expert appraisal, exploratory classical-test statistics, and perceived practicality was available; the study does not present a complete validity argument.

Digital platforms can support efficient administration, automatic scoring, rapid reporting, and feedback (Redecker & Johannessen, 2013). Immediate feedback may assist formative use when it is timely and linked to the task (Hattie & Timperley, 2007; Shute, 2008). Gamification adds design features such as points, progress indicators, challenges, or leaderboards (Deterding et al., 2011). These features may influence participation, but their effects depend on implementation and can be weakened by competition, time pressure, or construct-irrelevant guessing (Dichev & Dicheva, 2017; Sailer & Homner, 2020).

The Indonesian pedagogical-deep-learning framework emphasizes learning that is mindful, meaningful, and joyful, together with experiences of understanding, applying, and reflecting (Anggraena et al., 2025). In the present study, these principles were used as design intentions:

attention to relevant information represented mindfulness, authentic digital scenarios supported meaningfulness, and interactive delivery and feedback were intended to support a positive experience. The study did not independently measure these dimensions and therefore does not claim that the item bank empirically established pedagogical deep learning.

The local needs analysis drew on teacher interviews, classroom observations, and review of existing assessment materials across the three participating schools. It indicated continued reliance on paper-based, recall-oriented questions and limited experience with contextual digital assessment. The available study record did not preserve event-level counts for observations or reviewed tests; these inputs are therefore used only to define local product requirements, not to estimate the prevalence of assessment practices in Pinrang Regency.

The closest instrument-development literature has primarily focused on computational-thinking tests rather than gamified Digital-Society HOTS items. Recent studies have developed and psychometrically examined computational-thinking assessments using item-response models, dimensionality analysis, invariance testing, and differential item functioning (Kong & Lai, 2022; El-Hamamsy et al., 2022, 2022, 2025; Zhang & Wong, 2023). Reviews likewise emphasize the need to distinguish the construct definition, task representation, and empirical evidence supporting score interpretation (Tikva & Tambouris, 2021; Ezeamuzie & Leung, 2022).

Gamified assessment studies often emphasize engagement, motivation, or platform comparison (Göksün & Gürsoy, 2019; Wang & Tahir, 2020; Bai et al., 2020; Zainuddin et al., 2020). The present contribution is narrower: it documents the development and exploratory screening of contextual C4-C5 items for the Digital-Society element delivered through a gamified platform. The novelty lies in combining curriculum-bounded item specifications, expert appraisal, item-level classical statistics, and separate user feedback, not in demonstrating superiority of gamified administration.

Accordingly, the study pursued four development-evaluation objectives: (1) to specify and develop contextual C4-C5 items aligned with the Digital-Society element of Phase D Informatics; (2) to obtain domain-specific expert appraisals of content, item construction, and language; (3) to examine preliminary item functioning, difficulty, discrimination, and internal consistency in a limited field sample; and (4) to describe teacher and student perceptions of practicality separately.

METHODS

Research Design

This study used Research and Development to construct and preliminarily evaluate a digital assessment product. ADDIE served as the overarching framework because it provided a coherent sequence from analysis to evaluation (Branch, 2009). The 4-D model informed product-specification decisions during define, design, and develop activities (Thiagarajan et al., 1974), while Tessmer's formative-evaluation procedures informed self-review, expert appraisal, individual review, an informal usability check, and field testing (Tessmer, 1993). The three sources were therefore nested rather than treated as competing full models.

Five operational stages were used: needs analysis, product design, development and expert appraisal, formative try-out and field testing, and finalization. Dissemination was removed from the framework because no formal dissemination procedure or outcome was documented. Table 1 maps each stage to its source, participants or inputs, procedures, output, and decision rule; Figure 1 summarizes the same framework visually.

Table 1. Stage-by-stage development process matrix

Stage	Model source	Inputs/ participants	Core procedure	Output/decision rule
Needs analysis	ADDIE Analysis; 4-D Define	Official curriculum; school-level needs inputs	Review curriculum, teacher input, local assessments, and access constraints	Restrict scope to Digital Society and define product requirements

Product design	ADDIE Design; 4-D Design	Blueprint and item specifications	Map content, C4-C5 processes, stimuli, options, key, scoring, and platform settings	Produce 20-item prototype: 10 C4 and 10 C5
Development and expert appraisal	ADDIE Development; 4-D Develop; Tessmer self/expert review	Researcher; one specialist per content, assessment, and language domain	Self-review, separate domain appraisal, qualitative revision	Appraised prototype; no inter-rater validity coefficient
Formative try-out and field test	ADDIE Implementation; Tessmer individual/small-group/field procedures	Individual review n=3; informal usability check n not recorded; field n=60	Review clarity and usability; administer 20-item pool; collect separate user feedback	Exploratory item and perception data
Finalization	ADDIE Evaluation	Item statistics and qualitative feedback	Flag weak items; revise claims; package provisional bank	Retain 17 items provisionally; revise and retest Items 6, 11, and 12; no dissemination claim

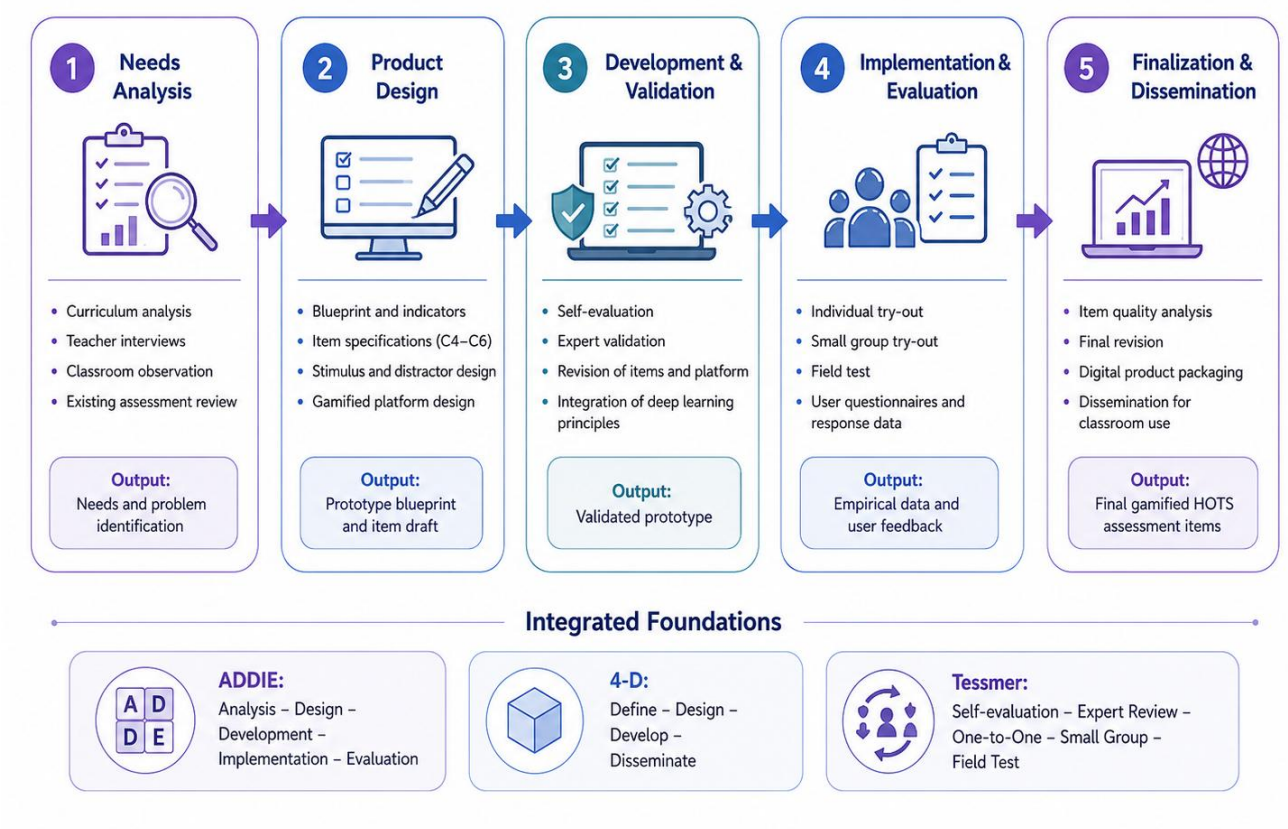

Figure 1. Development framework and operational stages

The initial product comprised 20 four-option multiple-choice items delivered through Quizizz. Ten items targeted C4 analysis and ten targeted C5 evaluation. All items were restricted to the Digital-Society element and to four content clusters: digital ethics, digital footprints and privacy, cyberbullying, and data security. The principles of mindful, meaningful, and joyful learning were treated as design intentions rather than validated product dimensions.

Research Setting and Participants

The study was conducted from January to March 2026 during the 2025-2026 academic year in UPT SMPN 2 Duampanua, UPT SMPN 2 Suppa, and UPT SMPN 4 Mattirobulu in Pinrang Regency, South Sulawesi, Indonesia. Each school implemented the Merdeka Curriculum and taught Informatics in Grade VIII.

The field-test sample comprised 60 Grade VIII students, with 20 students from each school. Because students were drawn from available classes in the participating schools rather than through probability sampling, the sample is treated as a school-based convenience sample. The ratio of three respondents per item is adequate only for exploratory classroom screening and is not sufficient for stable dimensionality, invariance, or item-response modelling. Age, sex, prior Informatics exposure, and prior Quizizz experience were not systematically reported; no subgroup score interpretation was attempted.

Expert appraisal involved three specialists: one in Informatics content, one in educational assessment, and one in language. Each specialist evaluated only the domain corresponding to his or her role. Standardized qualification thresholds, years of experience, independence from the research team, and multiple-rater agreement within each domain were not documented. Consequently, the results are reported as domain-specific expert appraisal rather than as inter-rater content-validity evidence, and Aiken's V or a content-validity index was not calculated.

Informatics teachers from the participating schools contributed to the needs analysis and completed a practicality questionnaire after implementation. The number of teacher respondents was not preserved in the manuscript record and is therefore reported as not documented in the practicality table. This prevents estimation of dispersion, uncertainty, or questionnaire reliability for the teacher group.

Ethical and Access Conditions

The manuscript record did not include an institutional ethics-approval number or sufficiently detailed documentation of school permission, parental consent, student assent, confidentiality procedures, or accommodations. No personally identifying student information is reported in this article, but the absence of formal ethics reporting limits evaluation of research governance and must be corrected from the original study files before publication if such documentation exists.

Development Procedures

The needs-analysis stage examined the official Phase D curriculum, local assessment practices, and student access conditions. Data sources included curriculum documents, teacher interviews, classroom observations, and existing assessment materials from the participating schools. Because exact counts for observation events and reviewed tests were not retained, the needs analysis is interpreted as formative design input rather than a reproducible prevalence study.

The analysis identified tasks in which students would need to interpret online interactions, distinguish public and private information, identify ethical or security risks, infer consequences of digital behavior, compare possible responses, and select a defensible action. Student characteristics considered during design included reading demands, familiarity with digital devices, experience with online quizzes, and ability to interpret visual stimuli; these characteristics were not measured as respondent variables in the field test.

The needs analysis produced five requirements: the item pool had to (1) remain within the Digital-Society element, (2) target C4 and C5 rather than C6, (3) use contextual stimuli that were necessary for answering, (4) employ plausible response options, and (5) be deliverable through a digital platform with automatic scoring and feedback. Mindful, meaningful, and joyful learning were included as intended design characteristics, not as measured constructs.

The product-design stage translated these requirements into an assessment blueprint. The blueprint linked the official Digital-Society outcome, the four sampled content clusters, observable C4-C5 processes, stimulus type, item number, response options, key, and dichotomous scoring rule. Table 2 provides the blueprint summary available for this report.

Item-level allocation across the four content clusters was not preserved in the manuscript record, so claims are limited to the documented total of 10 C4 and 10 C5 items.

Table 2. Summary blueprint for the initial item pool

Blueprint component	Documented representation	Item allocation	Evidence claim
Curriculum scope	Digital-Society element: openness of information, public/private information, ethical conduct, and personal digital safety	20 items total	Restricted to sampled Digital-Society content
Content clusters	Digital ethics; digital footprints/privacy; cyberbullying; data security	Cluster-specific counts not preserved	No claim of full Phase D coverage
C4 analysis	Interpret stimulus; identify violation/risk; infer consequence; distinguish relevant information	10 items	Selected-response evidence of analysis
C5 evaluation	Compare alternatives; judge defensibility; select the most appropriate response	10 items	Selected-response evidence of evaluation
Format and scoring	Contextual four-option multiple choice; 20-minute total administration; 1 = correct, 0 = incorrect	20 scored responses	No C6 creation claim

Each item used an essential stimulus such as a social-media interaction, digital case, table, graph, image, or security scenario. Students had to interpret the stimulus, identify relationships or risks, infer consequences, compare responses, or judge the most defensible action. The correct response received 1 point and an incorrect response received 0 points.

Distractors were written to represent plausible misconceptions, incomplete reasoning, or inappropriate digital actions. Item-writing review addressed grammatical clues, overlapping alternatives, inconsistent option length, ambiguous wording, and irrelevant reading difficulty (Haladyna et al., 2002; Scully, 2017). Option-level response frequencies were not retained in the available analysis output; distractor functioning is therefore not claimed empirically in this report.

The Quizizz administration contained 20 items with a total worksheet time of 20 minutes, as shown in Figure 2. Questions were presented sequentially with four options, automatic scoring, progress information, and feedback after completion. The records did not preserve exact settings for leaderboard visibility, item or option randomization, retry permissions, or whether time limits varied by item. These unrecorded settings may have introduced speededness, competition, or other construct-irrelevant variance and are treated as a limitation.

During development, the researcher conducted self-evaluation of curriculum alignment, item indicators, cognitive level, answer keys, stimulus clarity, distractor plausibility, language, and digital display. The item set was then converted into a functional Quizizz prototype.

The three specialists appraised the prototype using separate forms. The Informatics form addressed alignment with the Digital-Society outcome, conceptual accuracy, contextual relevance, and content depth. The assessment form addressed C4-C5 demand, stimulus function, item construction, scoring, and response-option quality. The language form addressed grammar, sentence effectiveness, readability, ambiguity, spelling, punctuation, and consistency of terminology.

Each appraisal form used a four-point response scale and included space for written recommendations. Scores were normalized as percentages of the maximum domain score. Because the forms differed by domain and only one specialist completed each form, the percentages were not pooled to estimate inter-rater agreement. Raw scores, denominators, and

domain-specific item counts were not preserved in the manuscript record; the percentages are therefore descriptive rather than precise estimates of content validity.

Formative review began with three students representing high, moderate, and low classroom performance. They commented on instructions, stimulus clarity, readability, response options, images, time allocation, and navigation. Their feedback informed sentence simplification and visual adjustment.

An informal small-group usability check was conducted before the field test. However, the number of participants, selection criteria, and school origin were not preserved in the available record. No quantitative result from this phase is therefore reported, and its observations are treated only as undocumented formative input rather than reproducible evidence.

The field test involved the 60 Grade VIII students. Students used available digital devices in their regular school context and completed the 20 items within the 20-minute total limit. Device type, screen size, internet quality, practice opportunities, supervision conditions, and accommodations were not standardized or logged. Quizizz recorded responses, total scores, completion information, and selected options, which were exported for analysis.

After the assessment, students completed a locally developed practicality questionnaire addressing ease of use, time efficiency, usefulness, and user suitability. Teachers completed a separate locally developed questionnaire covering related but not necessarily equivalent dimensions. The instruments were analyzed separately because they involved different respondent populations and potentially different item compositions.

Research Instruments and Data Collection

Data-collection tools comprised curriculum-review notes, observation sheets, semi-structured teacher-interview guides, three domain-specific expert-appraisal forms, the 20 HOTS items, student and teacher practicality questionnaires, and classroom implementation records. The exact number of items in each non-test instrument and evidence of their reliability were not preserved in the study record. Therefore, questionnaire findings are interpreted as descriptive user perceptions rather than validated scale scores.

The expert-appraisal forms used four-point ratings. The HOTS items used dichotomous scoring. The student and teacher questionnaires produced normalized percentage scores for the four reported aspects. Missing-data rules were not documented, and no dispersion, confidence interval, or internal-consistency estimate was available for either questionnaire.

Quantitative evidence included normalized expert-appraisal scores, item responses, item difficulty, item-total correlations, discrimination indices, Cronbach's alpha, and user-perception percentages. Qualitative evidence included teacher interviews, observation notes, expert comments, and student feedback from formative review. Qualitative evidence was used to guide revision rather than to establish prevalence or causal effects.

Data Analysis

Expert-appraisal and practicality percentages were reported descriptively without applying unsupported labels such as 'highly valid' or 'highly practical.' Item functioning was examined using the uncorrected Pearson item-total correlations reported in the original SPSS output. Because these coefficients include part-whole overlap and their p-values depend on sample size, they were treated as exploratory screening rather than as sufficient evidence of item validity. Corrected point-biserial correlations, confidence intervals, and alpha-if-item-deleted values could not be reported because the item-level response matrix was not available for this revision. Cronbach's alpha was reported only for the unrevised 20-item pool and was interpreted in relation to test length, score use, and multidimensional content rather than as a pass-fail criterion.

Item difficulty was the proportion of students answering correctly; values of .00-.30 were described as difficult, .31-.70 as moderate, and .71-1.00 as easy (Ebel & Frisbie, 1991). The available output also contained upper-lower discrimination indices, interpreted as very good (>=.70), good (.40-.69), adequate (.20-.39), or poor (<.20). The exact upper- and lower-group

proportions, group sizes, formula, and treatment of tied scores were not preserved, so these indices are interpreted cautiously. Option-level distractor frequencies were unavailable and no numerical distractor-functioning criterion was applied. Items were flagged for revision when weak item-total association and poor discrimination converged.

RESULT AND DISCUSSION

Product Development Results

The development process produced an initial 20-item digital assessment for the Digital-Society element of Phase D Informatics. The pool covered digital ethics, digital footprints and privacy, cyberbullying, and data security. Ten items targeted C4 analysis and ten targeted C5 evaluation. No item was classified as C6. Stimuli included short cases, social-media interactions, tables, graphs, images, digital-security situations, and online-communication problems. The Quizizz product included identity fields, instructions, four response options, a 20-minute total time, automatic scoring, feedback, and progress information.

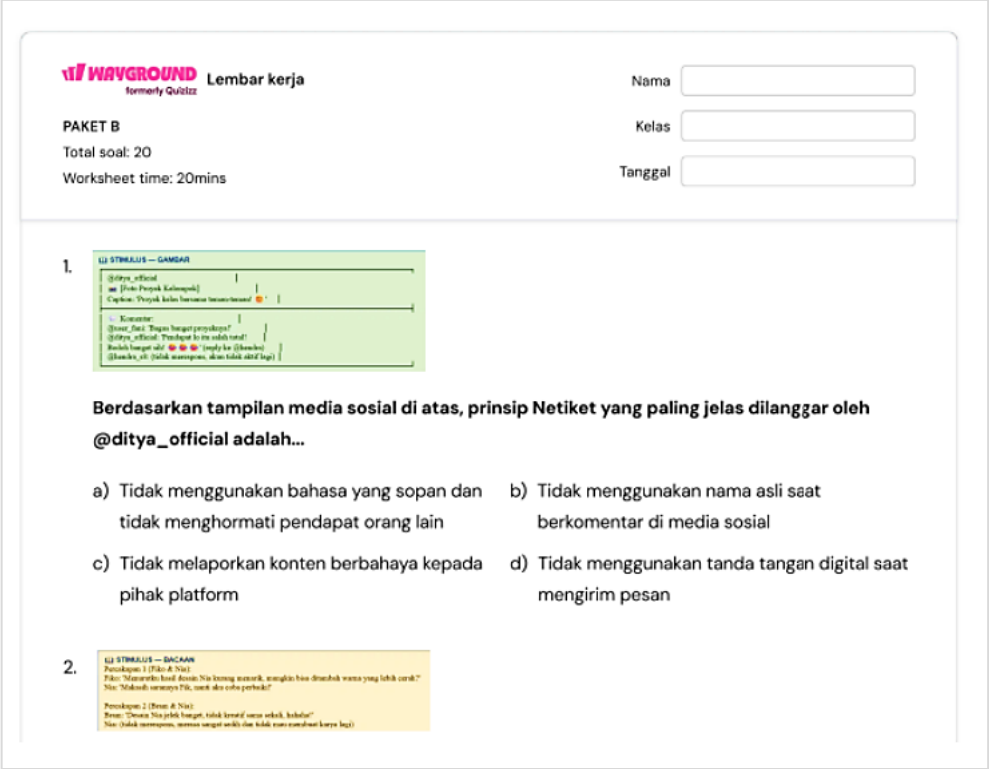

Figure 2. Design of the gamified HOTS assessment interface

The questions were displayed sequentially so that students could inspect textual or visual information before selecting a response. Figure 3 illustrates an item that requires interpretation of a graph and evaluation of a digital-behavior problem. The figures demonstrate platform delivery and task presentation; they do not establish that gamification improved performance or engagement.

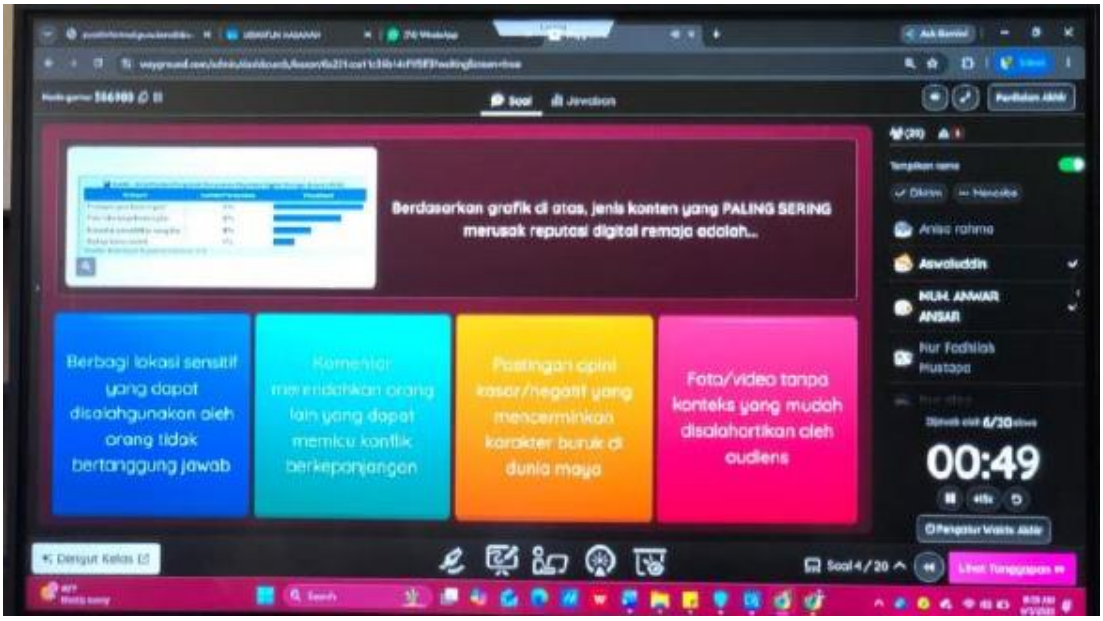

Figure 3. Example of a HOTS item and its contextual stimulus

Formative revisions addressed lengthy wording, image size, terminology consistency, weak response options, and time allocation. Because the small-group record was incomplete, only the documented individual review, expert comments, and field-test evidence are used in evaluating the product. The resulting pool remained provisional pending empirical screening.

Expert Validation

One specialist in each domain appraised the assessment. The Informatics appraisal addressed curriculum alignment, conceptual accuracy, contextual relevance, and correspondence with the Digital-Society outcome. The assessment appraisal addressed C4-C5 demand, item construction, stimulus function, scoring, and response-option quality. The language appraisal addressed clarity, readability, grammar, spelling, punctuation, and terminology. Because one specialist represented each domain, the findings are descriptive appraisals rather than evidence of inter-rater content validity.

Table 3. Domain-specific expert-appraisal scores

Validation area	Percentage	Category
Informatics content	90.77%	Highly valid
Educational assessment	93.75%	Highly valid
Language and readability	98.19%	Highly valid
Overall mean	94.24%	Highly valid

The normalized appraisal scores were 90.77% for Informatics content, 93.75% for assessment construction, and 98.19% for language and readability. These values indicate favorable judgments by the three specialists, but their precision should not be overstated because raw denominators, domain item counts, and multiple-rater agreement were unavailable. The three percentages were therefore not pooled into an overall validity coefficient.

Empirical Item Quality

Exploratory field testing involved 60 students. The available SPSS output reported uncorrected Pearson item-total correlations. Seventeen items had positive coefficients with $p < .05$ under the original screening rule, while Items 6, 11, and 12 did not. Item 11 showed the weakest association ($r = .046$), and Items 9 and 16 showed the largest reported coefficients ($r = .726$). Because the coefficients were uncorrected and the sample was small, these values are used only to flag items for further investigation, not to establish item validity.

Table 4. Exploratory item-total, difficulty, and discrimination statistics

Item	Item-total r	p	Screening	Difficulty	Level	Discrimination	Category
1	.605	.001	Provisionally retain	.40	Moderate	.43	Good
2	.555	.001	Provisionally retain	.36	Moderate	.50	Good
3	.441	.001	Provisionally retain	.30	Difficult	.37	Adequate
4	.392	.002	Provisionally retain	.20	Difficult	.27	Adequate
5	.632	.001	Provisionally retain	.33	Moderate	.53	Good
6	.208	.110	Flag for revision	.25	Difficult	.17	Poor
7	.549	.001	Provisionally retain	.41	Moderate	.40	Good
8	.533	.001	Provisionally retain	.30	Difficult	.53	Good
9	.726	.001	Provisionally retain	.31	Moderate	.67	Good
10	.591	.001	Provisionally retain	.40	Moderate	.50	Good
11	.046	.726	Flag for revision	.45	Moderate	.07	Poor
12	.121	.355	Flag for revision	.63	Moderate	.07	Poor
13	.577	.001	Provisionally retain	.41	Moderate	.57	Good
14	.658	.001	Provisionally retain	.48	Moderate	.63	Good
15	.634	.001	Provisionally retain	.50	Moderate	.60	Good
16	.726	.001	Provisionally retain	.31	Moderate	.60	Good
17	.318	.013	Provisionally retain	.65	Moderate	.27	Adequate
18	.455	.001	Provisionally retain	.60	Moderate	.30	Adequate
19	.674	.001	Provisionally retain	.63	Moderate	.60	Good
20	.504	.001	Provisionally retain	.53	Moderate	.37	Adequate

Items 6, 11, and 12 combined weak item-total association with discrimination below .20 and were therefore flagged for substantial revision. The provisional usable form contains 17 items; the three flagged items are not described as part of a final functioning form until they are rewritten and retested. Corrected point-biserial correlations, confidence intervals, alpha-if-item-deleted values, and reliability for the 17-item form were not available without the original item-level response matrix.

Table 5. Internal-consistency estimate for the unrevised item pool

Method	Number of items	Coefficient	Interpretation
Cronbach's alpha	20	.735	Reliable

Cronbach's alpha for the unrevised 20-item pool was .735. This estimate includes the three malfunctioning items and should not be interpreted as a pass-fail confirmation of reliability or unidimensionality. For a short, early-stage classroom item pool, the coefficient provides limited information about score consistency; it does not establish that HOTS and the sampled Digital-Society content form a single dimension. Difficulty indices classified four items as difficult and 16 as moderate, with no easy items. After correcting the boundary error for Item 7 (discrimination = .40), 12 items were classified as good, five as adequate, and three as poor under the stated categories.

Product Practicality

Teacher and student questionnaires addressed perceived ease of use, time efficiency, usefulness, and user suitability. The two groups were analyzed separately because the instruments represented different populations and their equivalence was not established.

Table 6. Separate teacher and student practicality scores

Respondent	Aspect	Percentage	Category
Teacher	Ease of use	93.75%	Highly practical
	Time efficiency	91.67%	Highly practical
	Usefulness	87.50%	Practical

	User suitability	95.83%	Highly practical
	Overall teacher response	91.96%	Highly practical
Student	Ease of use	75.69%	Practical
	Time efficiency	83.33%	Highly practical
	Usefulness	80.55%	Highly practical
	User suitability	87.03%	Highly practical
	Overall student response	81.65%	Highly practical
	Combined mean	86.81%	Highly practical

The reported overall teacher score was 91.96%, while the student score was 81.65%. The teacher respondent count, dispersion, missing-data treatment, questionnaire item counts, and questionnaire reliability were not documented; the student results likewise lacked dispersion and reliability estimates. The percentages therefore describe perceived practicality in this implementation only. No combined mean is reported, and the ratings are not evidence that gamification caused better usability or learning.

Discussion

This study provides preliminary, context-bound evidence concerning the development and functioning of a gamified C4-C5 item bank for the Digital-Society element of Phase D Informatics. Its contribution is narrower than the original manuscript claimed. The study documents curriculum-bounded design, three domain-specific expert appraisals, exploratory item statistics from 60 students, an alpha estimate for the initial pool, and separate user-perception scores. It does not provide a complete validity argument or evidence of instructional effectiveness.

The favorable expert-appraisal percentages suggest that each specialist generally endorsed the aspect reviewed. Nevertheless, one rater per domain cannot support inter-rater agreement, Aiken's V, or a content-validity index, and the absence of transparent raw scores and denominators limits interpretability. Future development should use multiple independent experts per domain, predefined qualification criteria, item-level ratings, and explicit decision rules.

The narrowed curricular scope is essential. The four sampled topics align most directly with the official Digital-Society element, but they do not represent the full Phase D Informatics domain. Consequently, the item bank should not be described as a general measure of Phase D computational thinking. Prior computational-thinking instruments demonstrate the value of broader construct mapping, dimensionality analysis, item-response modelling, and subgroup fairness testing (Kong & Lai, 2022; El-Hamamsy et al., 2022a, 2022b, 2025; Zhang & Wong, 2023).

The convergence of weak item-total association and poor discrimination for Items 6, 11, and 12 provides a practical basis for revision. Possible causes include ambiguity, an inaccurate key, an implausible or overly obvious distractor, excessive reading demand, or mismatch between the intended and elicited reasoning. Because corrected point-biserial correlations and option-level frequencies were unavailable, the specific cause cannot be diagnosed from the current output.

The identification of problematic items is a useful formative result, but the remaining 17 items should be regarded as provisional rather than final. Their functioning should be re-examined after removal or revision of the three flagged items. A larger sample is needed to estimate internal structure, item parameters, score precision, and differential functioning across schools and student subgroups (Ezeamuzie & Leung, 2022; El-Hamamsy et al., 2025).

The alpha coefficient of .735 must be interpreted in relation to the 20-item length and the heterogeneous content and cognitive demands. Alpha does not show that all items are interchangeable, does not establish unidimensionality, and does not justify high-stakes use. A future study should report corrected point-biserial correlations, alpha and omega for the

retained form, dimensionality evidence, and uncertainty estimates before proposing score-based decisions.

The difficulty distribution contained no easy items, which may reduce information for lower-performing students. The discrimination results were mostly adequate or good after correcting the category boundary for Item 7, but the exact upper-lower grouping procedure and treatment of tied scores were not recorded. These values should therefore be treated as descriptive signals rather than stable item parameters.

The original manuscript described distractor effectiveness but did not report option-level response frequencies or a numerical criterion for a functioning distractor. That claim has been removed. Future analysis should report the frequency and ability distribution for each option and examine whether distractors capture identifiable misconceptions about privacy, security, ethics, or online conduct.

Teachers rated practicality more positively than students, but the percentages cannot be compared as equivalent scale scores because questionnaire composition, respondent numbers, dispersion, and reliability were not fully documented. Teacher enthusiasm, familiarity with the scoring dashboard, novelty, and automatic reporting could all influence ratings. The student score may also reflect platform access, reading load, device constraints, or timing rather than gamification itself.

Unequal devices, internet conditions, screen sizes, practice opportunities, and accommodations were not standardized or logged. These factors can affect response accuracy independently of HOTS and may explain part of the lower student ease-of-use score. A future administration should standardize or record access conditions, provide practice items, document accommodations, and analyze response time and missingness.

Gamification should therefore be understood only as the delivery architecture used in this study. Meta-analytic evidence suggests that gamification effects are variable and generally depend on design, learner characteristics, autonomy, competence, and context (Sailer & Homner, 2020; Li et al., 2024). Personalized or visually rich game features do not automatically improve motivation or performance (Oliveira et al., 2022). Without an identical non-gamified comparison condition, no causal conclusion about gamification is warranted.

Mindful, meaningful, and joyful learning were likewise design intentions rather than independently measured dimensions. Contextual stimuli may support relevance, careful information processing may support awareness, and interactive feedback may support a positive experience, but these features do not demonstrate that pedagogical deep learning occurred. Experimental and process-based studies are needed to test these dimensions directly.

The main limitations are the small convenience sample; clustering of students within three schools without multilevel analysis; restricted coverage of one Phase D element; incomplete documentation of needs-analysis counts, sampling characteristics, expert qualifications, small-group procedures, platform settings, access conditions, questionnaire structure, missing-data treatment, and ethics approval; use of uncorrected item-total correlations; absence of corrected correlations, confidence intervals, alpha-if-item-deleted, 17-item reliability, dimensionality analysis, distractor frequencies, external criteria, and subgroup fairness evidence; and the lack of a non-gamified comparison. These limitations require cautious, local interpretation and substantial further validation.

CONCLUSION

This study developed and preliminarily evaluated a gamified 20-item C4-C5 pool for the Digital-Society element of Phase D Informatics. Three specialists provided favorable domain-specific appraisals, and exploratory testing with 60 Grade VIII students identified 17 provisionally functioning items. Items 6, 11, and 12 showed weak item-total associations and poor discrimination and require revision and retesting. Cronbach's alpha of .735 applies only to the unrevised 20-item pool. Teacher and student practicality percentages were reported separately and are interpreted as context-specific perceptions. The evidence supports only a provisional,

context-bound item bank; it does not demonstrate improved HOTS, computational reasoning, sustained engagement, pedagogical deep learning, or superiority over non-gamified administration. Further work should document ethical procedures and platform conditions, use multiple independent experts, retain complete item-level data, retest the revised pool in larger multisite samples, and examine internal structure, external relations, subgroup fairness, and consequences of score use.

REFERENCE

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

Anggraena, Y., Ginanto, D. E., Kesuma, A. T., & Setiyowati, D. (2025). *Panduan pembelajaran dan asesmen pendidikan anak usia dini, jenjang pendidikan dasar, dan jenjang pendidikan menengah* (Edisi revisi 2025). Badan Standar, Kurikulum, dan Asesmen Pendidikan.

Badan Standar, Kurikulum, dan Asesmen Pendidikan. (2025). *Keputusan Kepala Badan Standar, Kurikulum, dan Asesmen Pendidikan Nomor 046/H/KR/2025 tentang capaian pembelajaran pada pendidikan anak usia dini, jenjang pendidikan dasar, dan jenjang pendidikan menengah*. Kementerian Pendidikan Dasar dan Menengah.

Bai, S., Hew, K. F., & Huang, B. (2020). Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review, 30*, Article 100322. https://doi.org/10.1016/j.edurev.2020.100322

Branch, R. M. (2009). *Instructional design: The ADDIE approach*. Springer. https://doi.org/10.1007/978-0-387-09506-6

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification." In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning future media environments* (pp. 9–15). Association for Computing Machinery. https://doi.org/10.1145/2181037.2181040

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

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.

El-Hamamsy, L., Zapata-Cáceres, M., Marcelino, P., Bruno, B., Dehler Zufferey, J., Martín-Barroso, E., & Román-González, M. (2022). Comparing the psychometric properties of two primary school computational thinking assessments for Grades 3 and 4: The Beginners' CT Test and the Competent CT Test. *Frontiers in Psychology, 13*, Article 1082659. https://doi.org/10.3389/fpsyg.2022.1082659

El-Hamamsy, L., Zapata-Cáceres, M., Martín-Barroso, E., Mondada, F., Dehler Zufferey, J., & Bruno, B. (2022). The Competent Computational Thinking Test: Development and validation of an unplugged computational thinking test for upper primary school. *Journal of Educational Computing Research, 60*(7), 1818–1866. https://doi.org/10.1177/07356331221081753

El-Hamamsy, L., Zapata-Cáceres, M., Martín-Barroso, E., Mondada, F., Dehler Zufferey, J., Bruno, B., & Román-González, M. (2025). The Competent Computational Thinking Test: A valid, reliable, and gender-fair test for longitudinal CT studies in Grades 3–6. *Technology, Knowledge and Learning, 30*, 1607–1661. https://doi.org/10.1007/s10758-024-09777-8

Ezeamuzie, N. O., & Leung, J. S. C. (2022). Computational thinking through an empirical lens: A systematic review of literature. *Journal of Educational Computing Research, 60*(2), 481–511. https://doi.org/10.1177/07356331211033158

Göksün, D. O., & Gürsoy, G. (2019). Comparing success and engagement in gamified learning experiences via Kahoot and Quizizz. *Computers & Education, 135*, 15–29. https://doi.org/10.1016/j.compedu.2019.02.015

Grover, S., & Pea, R. (2013). Computational thinking in K–12: A review of the state of the field. *Educational Researcher, 42*(1), 38–43. https://doi.org/10.3102/0013189X12463051

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334. https://doi.org/10.1207/S15324818AME1503_5

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

Kong, S. C., & Lai, M. (2022). Validating a computational thinking concepts test for primary education using item response theory: An analysis of students' responses. *Computers & Education, 187*, Article 104562. https://doi.org/10.1016/j.compedu.2022.104562

Lewis, A., & Smith, D. (1993). Defining higher-order thinking. *Theory Into Practice, 32*(3), 131–137. https://doi.org/10.1080/00405849309543588

Li, L., Hew, K. F., & Du, J. (2024). Gamification enhances student intrinsic motivation, perceptions of autonomy and relatedness, but minimal impact on competency: A meta-analysis and systematic review. *Educational Technology Research and Development, 72*, 765–796. https://doi.org/10.1007/s11423-023-10337-7

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

OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Oliveira, W., Hamari, J., Joaquim, S., Toda, A. M., Palomino, P. T., Vassileva, J., & Isotani, S. (2022). The effects of personalized gamification on students' flow experience, motivation, and enjoyment. *Smart Learning Environments, 9*, Article 16. https://doi.org/10.1186/s40561-022-00194-x

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st-century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*(2), 65–77. https://doi.org/10.1016/j.pse.2014.10.001

Redecker, C., & Johannessen, O. (2013). Changing assessment—Towards a new assessment paradigm using ICT. *European Journal of Education, 48*(1), 79–96. https://doi.org/10.1111/ejed.12018

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

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research & Evaluation, 22*, Article 4. https://doi.org/10.7275/swgt-rj52

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

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review, 22*, 142–158. https://doi.org/10.1016/j.edurev.2017.09.003

Tessmer, M. (1993). *Planning and conducting formative evaluations: Improving the quality of education and training*. Routledge. https://doi.org/10.4324/9780203061978

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. https://eric.ed.gov/?id=ED090725

Tikva, C., & Tambouris, E. (2021). Mapping computational thinking through programming in K–12 education: A conceptual model based on a systematic literature review. *Computers & Education, 162*, Article 104083. https://doi.org/10.1016/j.compedu.2020.104083

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

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology, 25*(1), 127–147. https://doi.org/10.1007/s10956-015-9581-5

Wing, J. M. (2006). Computational thinking. *Communications of the ACM, 49*(3), 33–35. https://doi.org/10.1145/1118178.1118215

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*, Article 100326. https://doi.org/10.1016/j.edurev.2020.100326

Zhang, S., & Wong, G. K. W. (2023). Development and validation of a computational thinking test for lower primary school students. *Educational Technology Research and Development, 71*, 1595–1630. https://doi.org/10.1007/s11423-023-10231-2