Evaluating Stakeholder-Integrated Authentic Assessment Design in an Undergraduate Web Engineering Course

Erny Arniza Ahmad

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

DOI: https://doi.org/10.47772/IJRISS.2026.101000102

Received: 09 October 2026; Accepted: 14 October 2026; Published: 30 October 2026

ABSTRACT

This study examines WAVE (Work-Integrated Authentic Validation and Entrepreneurship) as a stakeholder-integrated assessment design for undergraduate Web Engineering. The study aimed to evaluate how WAVE organises stakeholder engagement around assessable course outputs, examine students' perceptions of the learning experience, describe performance in the Requirements Analysis Document (RAD) and Web Application (WebApp) prototype, and identify the educational value of integrating stakeholder judgement at key stages of project development. A descriptive multi-source evaluation was conducted during one 14-week implementation involving 58 Year 2 Netcentric Computing students. Evidence comprised a post-implementation survey completed by 52 students (89.7%), author-reviewed open-ended responses, and formal assessment records for the RAD and WebApp prototype components. Student evaluations were consistently positive across all five survey domains, with mean scores ranging from 4.75 to 4.87. Full-cohort mean performance was 70.69% for RAD and 77.17% for prototype components. Qualitative findings showed that students particularly valued stakeholder communication, interpreting real requirements, applying technical skills, professional exposure, teamwork, and opportunities to refine their work through feedback. Students also identified practical implementation considerations relating to stakeholder availability, workload, and team coordination. Overall, the findings demonstrate that WAVE provides a coherent course-level assessment architecture that connects authentic stakeholder engagement with requirements analysis, solution development, iterative feedback, and formal assessment evidence. The model offers a structured approach for positioning stakeholder judgement at identifiable points within the assessment process while supporting professionally relevant learning experiences. The study concludes that WAVE has strong potential as an authentic assessment design for Web Engineering and provides a practical foundation for further implementation and evaluation across broader cohorts and educational contexts.

Keywords: Authentic Assessment, Stakeholder Engagement, Work-Integrated Learning, Web Engineering, Assessment Design

INTRODUCTION

Authentic assessment is more than placing a project in a realistic setting. It requires students to interpret a consequential problem, make defensible disciplinary decisions, produce evidence of those decisions, and respond to feedback that can change the work. Constructive alignment provides the underlying logic: assessment evidence should align with the learning outcomes, and performance should reflect what a course intends to develop (Biggs, 1996). In Web Engineering, this matters because a technically functional solution can still be educationally weak if students have misunderstood the stakeholder need, cannot justify their design decisions, or treat feedback as an optional final comment rather than part of the development process.

Research on authentic assessment emphasises realistic challenge, judgement, and a credible relationship between assessment and practice (Ashford-Rowe et al., 2014; Villarroel et al., 2018). Work-integrated learning (WIL) adds a further complication: practice-linked projects can expose students to ambiguity,

negotiation, collaboration, and professional responsibility, but they can also create uneven access to stakeholders, inconsistent expectations, and different opportunities to demonstrate learning (Luk & Chan, 2024). The key design question is therefore not simply whether a project has a stakeholder, but where stakeholder judgement enters the assessment process, what evidence it is expected to generate, and how teams govern variation.

The WAVE (Work-Integrated Authentic Validation and Entrepreneurship) model organises six connected components: real-world context, authentic engagement, needs analysis, solution development, authentic validation, and evidence-based assessment. Students interpret and document a need, return that interpretation to the stakeholder before substantial development, develop a response, and seek a later stakeholder judgement of the solution. These activities draw on established project and feedback practices. WAVE's proposed contribution is their explicit organisation around two course-assessment artefacts, with stakeholder judgement positioned at identifiable decision points. This study examines that context-specific design proposition rather than claiming the invention of client involvement or iterative validation. The entrepreneurship and continuation pathway is part of the wider model but falls outside the outcomes evaluated here.

Evaluating this design requires evidence about both student experience and assessed work. Ratings and written responses describe perceived authenticity, usefulness, fairness, and professional relevance. Requirements Analysis Document (RAD) and Web Application (WebApp) prototype marks describe performance on course tasks. Neither source establishes learning gain or course learning outcome (CLO) attainment. Together, they can identify implementation strengths and difficulties that a satisfaction survey alone would overlook. This evaluation therefore offers a bounded account of WAVE's assessment sequence, one cohort's experience, and the conditions that warrant closer investigation.

The evaluation was guided by four research questions:

RQ1. How was WAVE organised at the course-design level to connect stakeholder judgement with assessable Web Engineering work?

RQ2. How did students evaluate the authenticity, capability development, engagement, feedback, fairness, and professional relevance of the WAVE assessment experience?

RQ3. What patterns of assessed performance were observed in the RAD and WebApp prototype aligned with the practical and interpersonal outcomes emphasised in the WAVE implementation?

RQ4. How do student ratings, qualitative responses, and formal assessment evidence together inform the strengths, limitations, and implementation requirements of WAVE?

LITERATURE REVIEW

WAVE draws on authentic assessment, constructive alignment, work-integrated learning, formative feedback, and professional development. These traditions support realistic tasks and stakeholder engagement; they also require clarity about what is assessed and how judgements inform learning. Reviews link authentic assessment to engagement and professionally relevant capabilities (Sokhanvar et al., 2021) and identify implementation demands, including stakeholder participation, training, and resources (Vlachopoulos & Makri, 2024). These findings justify attention to assessment conditions, but they do not establish that any particular local model produces those benefits.

Authenticity is not a property of a task label. It depends on how closely an assessment reflects the activities, purposes, constraints, and judgement practices that matter in the domain being learned. Gulikers et al. (2004) frame authentic assessment as multidimensional, while Ashford-Rowe et al. (2014) and Villarroel et al. (2018) translate authenticity into design features that make meaningful performance visible. For Web Engineering, that means more than producing a working WebApp prototype. Students must interpret requirements,

represent them through disciplinary artefacts, justify design choices, and respond when stakeholder feedback exposes a mismatch between the intended need and the proposed solution.

Constructive alignment also highlights the evidence used to judge whether students have demonstrated the intended learning outcomes. Student perceptions can indicate whether an assessment feels relevant, challenging, or professionally useful, but they do not replace assessed performance. Conversely, course marks show how work was judged against assessment requirements but do not reveal how students experienced the process or why particular stages were difficult. A stronger evaluation therefore distinguishes perceived learning from demonstrated performance and uses each evidence source for the question it can reasonably answer.

Work-integrated learning makes this distinction especially important because educational goals must be negotiated alongside the realities of practice. Billett (2009) argues that work experience becomes educationally valuable only when it is deliberately integrated with learning. A systematic review by Luk and Chan (2024) similarly found that WIL assessment often serves developmental purposes, including self-regulation and future-oriented capability, rather than merely certifying performance. In software education, client-connected projects expose students to ambiguous requirements, coordination demands, and responsibilities that are difficult to reproduce in a lecturer-written scenario (Chen & Chong, 2011; Dong et al., 2022; Kauppinen et al., 2022). However, those benefits depend on stakeholder availability, clarity, feedback quality, and the relationship between academic and workplace judgement (Pažur Aničić et al., 2026).

The closest comparison for WAVE is an existing client-connected software project, rather than an assessment with no authentic activity. For example, Kauppinen et al. (2022) describe a university–industry model involving real customers and a software provider, with facilitation and mentoring needed to support students. Pažur Aničić et al. (2026) examine employers' assessment roles and the importance of shared expectations and criteria.

WAVE addresses a narrower design question: how stakeholder judgement can be located around the RAD and WebApp prototype components within a single course. Its first checkpoint concerns stakeholder endorsement of the documented requirements and preliminary design artefacts before WebApp prototype development; its second concerns stakeholder evaluation of the developed solution against those requirements. This is an explicit configuration of established practices, not evidence that earlier approaches lack comparable checkpoints. The proposed value lies in making the relationship between judgement, revision, and assessable artefacts visible enough to examine.

The educational rationale for the two checkpoints rests on feedback use. Stakeholder responses can support learning when students use them to reconsider requirements or revise a solution, consistent with formative feedback and self-regulation (Nicol & Macfarlane-Dick, 2006). Academic and industry perspectives may also provide different forms of feedback during WIL assessment (Twyford & Dean, 2024).

In WAVE, RAD marks relate to requirements analysis and design documentation, while WebApp prototype marks relate to applied development and stakeholder-facing performance. These components support the practical and interpersonal outcomes identified in CLO2 and CLO3. Establishing the proposed mechanism would nevertheless require records showing what feedback was received, what changed, and how the change was assessed.

WAVE also includes a continuation pathway for work that has value beyond the immediate grade. Sustainable assessment focuses on students' capacity to make judgements and continue learning beyond a single task (Boud & Soler, 2016), while practice-oriented entrepreneurship education emphasises learning through action (Neck & Greene, 2011).

WAVE draws on both ideas, but continuation remains selective. Portfolio use, deployment, adoption, or entrepreneurial activity may follow later, but they do not automatically follow from completing an authentic

project. The present evaluation therefore treats professional readiness and continuation as perceived intermediate outcomes rather than evidence of employment or venture creation.

Table 1 summarises the relationship between WAVE's design intentions and the evidence examined in this study. The distinction is deliberate: student evaluation data describe experience, formal marks provide course performance context, and neither source by itself verifies implementation fidelity, stakeholder benefit, or causal learning gain.

Table 1. WAVE Design Intentions and the Scope of the Evaluation

Proposed element	Intended assessment role	Evidence examined in this evaluation
Real-world context and engagement	Connect work to a stakeholder need and establish interaction.	A items and written accounts describe perceived authenticity and stakeholder contact; stakeholder benefit itself is not measured.
Needs analysis and first validation	Document the interpreted need and obtain stakeholder response before WebApp prototype development.	B1-B3, written responses, and RAD performance provide evidence related to requirements analysis and CLO2-aligned practical work; checkpoint fidelity is not independently audited.
Solution development and later validation	Develop an artefact and use stakeholder evaluation to inform revision.	B4-B8, C items, written responses, and WebApp prototype performance provide evidence related to applied development and CLO2/CLO3-aligned performance; quality improvement is not independently verified.
Evidence-based assessment	Combine disciplinary artefacts, process evidence, and appropriate assessor judgements.	D items and formal RAD and WebApp prototype marks provide perception and course-performance evidence; criterion-level assessor agreement and direct CLO attainment are not analysed.
Continuation and entrepreneurship pathway	Allow selected outputs to progress beyond grading when appropriate.	E4-E5 and written reflections address portfolio potential, readiness, and understanding of deployment; the study does not measure actual continuation, adoption, or entrepreneurship.

METHODOLOGY

Evaluation design and setting

The study used a descriptive, multi-source evaluation of one WAVE implementation in an undergraduate Web Engineering course. The study analysed numerical survey responses, open-ended responses from the same survey, and formal course marks separately, then integrated them at the interpretive stage. The course-design account addressed RQ1; student evaluations addressed RQ2; assessment records addressed RQ3; and a comparison of patterns and tensions across these sources addressed RQ4. The design supports an account of experience and course performance, rather than an estimate of WAVE's causal effectiveness.

The implementation took place during the March to August 2026 semester in ITT545 Web Engineering, a Year 2 course in the Bachelor of Computer Science (Netcentric Computing) programme at Universiti Teknologi MARA (UiTM) Shah Alam, Selangor. The cohort comprised 58 students working in project teams over the full 14-week semester. The WAVE project accounted for 50% of the course assessment, consisting of a RAD report (30%) and a WebApp prototype component (20%). Teams comprised three to four students.

The lecturer and stakeholder assigned the assessment marks, which each project team shared among its members.

The RAD report covered requirements analysis, documentation, modelling, and translating stakeholder needs into design decisions. The WebApp prototype component addressed applied development and stakeholder-facing demonstration, with communication and collaboration relevant to its interpersonal demands. These are intended links to CLO2 and CLO3, rather than a criterion-level attainment analysis. The evaluation used component totals; it did not establish which particular rubric criteria accounted for the observed marks.

Assessment sequence and stakeholder roles

The project context was established with the Business Innovation & Technology Commercialisation Centre (BITCOM), UiTM, which provided the institutional stakeholder setting. Each team worked with a designated stakeholder, while a director served as the overall stakeholder representative across the projects. Students first identified and clarified stakeholder needs through project briefings and requirements elicitation. They subsequently translated these requirements into design specifications and preliminary Web Engineering artefacts, including system diagrams and wireframes. They compiled these artefacts into the Requirements Analysis Document (RAD) and presented it to stakeholders for endorsement before WebApp prototype development. This constituted the first stakeholder-validation checkpoint.

Following RAD endorsement, teams developed their WebApp prototypes and demonstrated the proposed solutions to stakeholders for evaluation against the agreed requirements. This formed the second stakeholder-validation checkpoint, during which stakeholders assessed the prototypes and provided feedback for potential refinement. Table 2 summarises the reported course sequence and the limits of the evaluation evidence. The sequence describes the course design and implementation account; dated checkpoint records and artefact-level revision histories were not independently audited. Consequently, the evaluation cannot establish that every team experienced equivalent validation or acted on stakeholder feedback to the same extent.

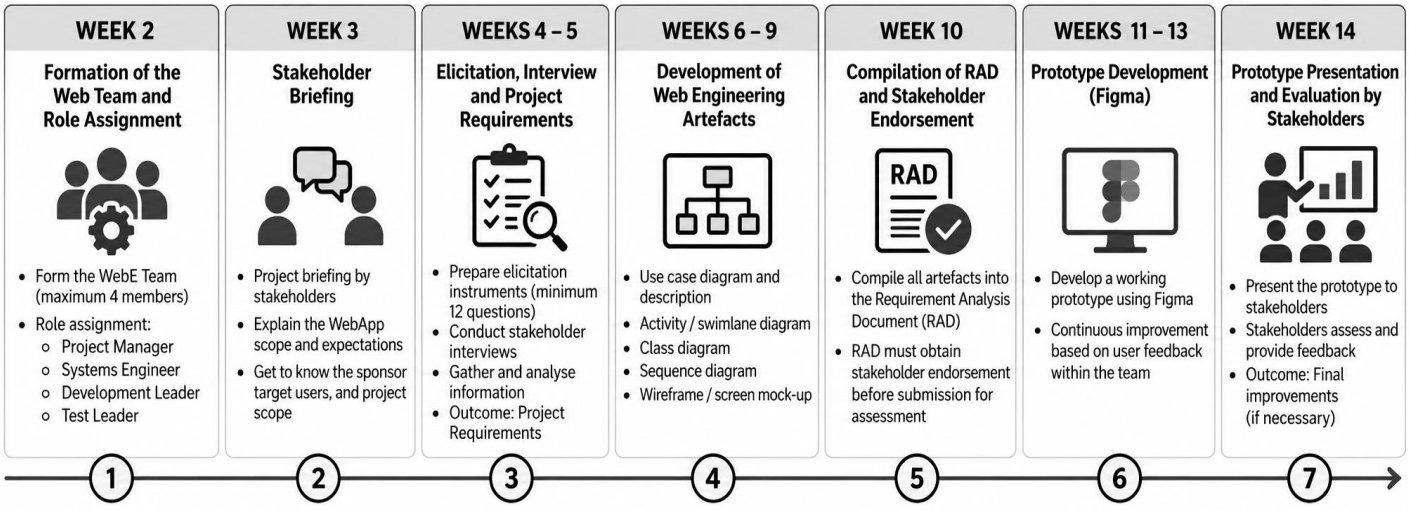

Figure 1. Progressive semester implementation of WAVE.

Figure 1 presents the semester-long operational sequence used in the implementation, from team formation and stakeholder briefing through requirements elicitation, Web Engineering artefact development, RAD endorsement, WebApp prototype development, and final stakeholder evaluation. The figure makes the course-design sequence underlying RQ1 visible. Table 2 complements the figure by distinguishing the corresponding assessment stages from the limits of the evidence that the present evaluation can establish.

Table 2. Reported Assessment Sequence and Evidence Boundaries

Stage	Timing and stakeholder role	Course output	Evaluation boundary
Needs interpretation	Before substantial development; teams clarify a designated stakeholder need.	Requirements and design specifications.	Student accounts describe experience; initial needs were not independently audited.
First validation	Following requirements analysis and preliminary design artefact development, but before WebApp prototype development; stakeholders review and endorse the RAD documentation.	RAD report, worth 30% of the course.	RAD totals provide performance context; checkpoint completion and resulting revisions were not independently verified.
Development and second validation	Following RAD endorsement, teams develop WebApp prototypes and present them to stakeholders for evaluation against the agreed requirements and feedback-informed refinement.	Web artefacts and WebApp prototype, worth 20% of the course.	Component marks and student feedback reports do not isolate the effect of the second checkpoint.
Continuation pathway	Beyond course assessment, where further development is justified.	Potential deployment, adoption, portfolio use, or entrepreneurial activity.	Not evaluated as an observed post-course outcome.

Participants and data collection

All 58 students were eligible to participate in the post-implementation evaluation. Participation was voluntary; 52 students completed the evaluation, representing 89.7% of the cohort. The reasons for the six non-responses were not recorded. All 52 respondents completed the numerical section, with no missing values across the 32 five-point Likert items, two global ratings, or the comparative item. All five open-ended questions received responses, although some entries were too brief to support substantive interpretation.

Responses were collected on 8 and 9 August 2026, following completion of the WAVE assessment but before the release of final course grades. The evaluation was administered by the course lecturer as part of routine course evaluation and reflection. Students could decline participation without any effect on their assessment marks or course grades. The lecturer had access to identifiable responses at the time of collection. Email addresses were used solely to identify potential duplicate submissions and were excluded from the analysis.

The WAVE implementation was conducted as part of routine teaching, learning, and assessment activities. The evaluation data were originally collected for educational evaluation and course improvement rather than as part of a separately commissioned research intervention. No formal research ethics approval was sought for the original educational evaluation. The present manuscript reports a retrospective descriptive analysis of existing course-evaluation and assessment data using only de-identified, aggregated findings. No identifiable student information or verbatim student comments are disclosed.

Questionnaire and descriptive analysis

The questionnaire was developed specifically for this WAVE implementation. Thirty-two positively worded Likert items were organised into five descriptive domains: authenticity and relevance (A1-A5), capability

development (B1-B8), engagement, feedback and refinement (C1-C6), assessment effectiveness and fairness (D1-D7), and professional readiness, continuation and recommendation (E1-E5 and E7). Responses were recorded on a 1 (Strongly Disagree) to 5 (Strongly Agree) scale. The missing E6 code reflects a numbering error in the original instrument rather than an omitted item.

The capability items addressed requirements elicitation and documentation, translation of stakeholder requirements into design decisions, WebApp prototype development, problem solving, technical communication, teamwork, and justified Web Engineering decision-making. Engagement items addressed responsibility, stakeholder feedback, revision, and contribution to project decisions. These items were interpreted as perception-based evidence related to the capabilities targeted by the RAD and WebApp prototype assessments, not as direct measures of CLO attainment.

Three additional closed questions provided broader judgements. Students rated the assessment's overall effectiveness in supporting learning and the value of working with a real stakeholder on separate 10-point scales. A comparative item asked them to judge WAVE relative to a conventional lecturer-provided project scenario using five categories from Much less effective to Much more effective. The questionnaire included five open-ended questions about valuable learning, challenges, improvements, professionally useful skills, and overall reflections.

Because the questionnaire was created for this implementation, it had not undergone formal expert content validation or pilot testing before administration. It is therefore treated as an implementation-evaluation instrument rather than a validated psychometric scale. Internal consistency is reported only to describe how the item groupings behaved within this cohort.

For each respondent, domain scores were calculated as the unweighted mean of the items assigned to that domain. Domain means and sample standard deviations were then calculated across the 52 respondents. Item-level means, standard deviations, and response frequencies were also examined. No imputation was required and no reverse scoring was applied because all numerical items were complete and positively worded.

Internal consistency was summarised using unstandardised Cronbach's alpha. Alpha was treated as an index of consistency within the observed data, not as evidence of dimensionality, construct validity, or WAVE's effectiveness. Ceiling concentration was examined through item frequencies, the number of respondents reaching the maximum score in each domain, and the number selecting 5 on all 32 items. Global ratings were summarised using means, standard deviations, medians, and observed ranges, while the comparative item was reported as counts and percentages.

Assessment records and units of analysis

Formal RAD and WebApp prototype records were examined for all 58 enrolled students. Component scores were summarised on their original 0–100 scales. The combined course contribution was calculated as $0.30 \times$ RAD + $0.20 \times$ WebApp prototype, giving a possible range of 0–50. The record count refers to students, not necessarily to independent project performances: marks were shared within project teams. Repeated marks for members of the same team therefore do not constitute independent assessments of individual performance. The summaries therefore describe recorded student marks without assuming independence between team members. Survey responses were not linked to individual marks, and no association between perceptions and performance was estimated.

Two assessment records contained zero in both WAVE components and no project-group code in the supplied records. The available data did not allow us to determine why. Results are therefore reported for the full cohort and, as a sensitivity check, for the 56 non-zero records. This distinction prevents two unexplained zero records from being silently treated as ordinary completed performance while retaining the full-cohort result as the primary descriptive summary.

The marks provide descriptive performance context. The analysis did not apply criterion-level CLO mappings, a separately specified attainment threshold, or independent external moderation. RAD and WebApp prototype components were different assessment tasks, not repeated measurements on an equivalent test. Differences between their means or standard deviations were therefore not interpreted as learning gains or as evidence of the effect of either validation checkpoint.

Qualitative analysis and integration

The 260 open-ended entries were analysed using descriptive qualitative content analysis. The five prompts provided an organising frame covering valuable learning, challenges, improvements, professionally useful skills, and overall reflections. Within these areas, recurring explicit content included stakeholder communication, requirements interpretation, technical application, revision, teamwork, professional exposure, and implementation difficulties.

Generative AI helped preliminarily organise recurring response patterns. The author then reviewed all 260 entries against the proposed classifications and made final decisions on inclusion, categorisation, interpretation, and reporting. Only author-accepted classifications were included in the frequency summaries. In this study, author-reviewed denotes this checking process; it does not imply an independent second analysis, inter-rater agreement, or thematic saturation.

Eight entries contained only a dash, 'nothing', or 'no' and were retained in the audit record but excluded from substantive interpretation. This left 252 substantive entries: 52 each for G1, G2, and G4, 45 for G3, and 51 for G5. The reporting unit was the respondent mentioning a category within a particular prompt, rather than the number of times a word appeared. One response could contribute to several categories when it contained several explicit ideas. Frequencies should therefore be read within prompts; category counts cannot be added to obtain numbers of distinct students across prompts.

Integration occurred at the interpretive stage. Survey ratings described student experience, formal marks provided performance context, and written responses explained what favourable or less favourable ratings meant in practice. Agreement across sources was treated as complementary evidence, not independent corroboration. Particular attention was given to points of tension, such as high ratings of fairness alongside reports of uneven stakeholder access, and high capability ratings alongside greater RAD variability among the non-zero records.

FINDINGS AND DISCUSSION

Assessment organisation

At the course-design level (RQ1), WAVE connected a stakeholder-facing requirements stage with a later demonstration of the developed solution. The RAD and WebApp prototype components provided corresponding assessed outputs. This establishes the intended assessment organisation while leaving the consistency, quality, and consequences of checkpoint participation unverified. The remaining findings address student experience (RQ2), recorded performance (RQ3), and their integrated implications (RQ4).

Student evaluations

Student evaluation data were strongly positive across all five descriptive domains. Means ranged from 4.75 for capability development to 4.87 for authenticity and relevance (Table 3), with sample standard deviations from 0.26 to 0.40 and Cronbach's alpha from .842 to .942. Each domain had a median of 5.00, and 32 to 38 respondents achieved the maximum possible score. These results indicate clear endorsement of the experience, but the concentration at the upper end also limits the instrument's ability to discriminate among dimensions.

Table 3. Domain Scores and Internal Consistency

Domain	Items	Mean	SD	Alpha	Maximum n (%)
Authenticity and relevance	5	4.87	0.26	.842	36 (69.2)
Capability development	8	4.75	0.40	.942	32 (61.5)
Engagement, feedback and refinement	6	4.86	0.28	.848	38 (73.1)
Assessment effectiveness and fairness	7	4.85	0.28	.892	37 (71.2)
Readiness, continuation and recommendation	6	4.81	0.34	.900	37 (71.2)

That lack of discrimination is visible at the item level. Of the 1,664 five-point responses, 1,387 (83.4%) were 5, 263 (15.8%) were 4, and only 14 (0.8%) were 3; no respondent selected 1 or 2. More strikingly, 28 students (53.8%) selected 5 for all 32 items. The high domain means therefore describe strong reported approval, but they do not provide fine-grained evidence about relative strengths across the assessment or demonstrate that the instrument distinguishes different levels of capability.

The few lower item means are nevertheless informative as descriptive signals. Translating stakeholder requirements into system design decisions (B3) received the lowest mean (M = 4.67, SD = 0.58). Developing a WebApp prototype from actual requirements (B4) and making and justifying Web Engineering decisions (B8) each averaged 4.71. The fairness item across stakeholders (D6) averaged 4.79 (SD = 0.46), with one response of 3, nine of 4, and 42 of 5. These small differences reflect descriptive variation within an otherwise highly concentrated response pattern. No inferential comparisons were conducted, and the observed differences should not be interpreted as evidence of statistically established differences between assessment dimensions.

The two global ratings showed the same strong approval pattern. Learning support averaged 9.33 out of 10 (SD = 0.83, median = 10, range = 8-10), while working with a real stakeholder averaged 9.44 (SD = 0.89, median = 10, range = 7-10). Twenty-nine respondents (55.8%) gave the maximum rating for learning support, and 35 (67.3%) did so for stakeholder value.

The comparative item was similarly favourable: 27 respondents (51.9%) selected 'Much more effective', 22 (42.3%) selected 'More effective', and three (5.8%) selected 'About the same'. Overall, 49 of 52 respondents (94.2%) made a favourable comparison. That result should be read carefully. It records students' judgement of WAVE relative to the lecturer-provided scenario described in the survey. The study did not establish equivalent prior exposure to that conventional scenario, and the item is not evidence of a measured difference in learning outcomes between conditions.

Recorded assessment performance

Formal assessment records provide a second perspective on the implementation. Across all 58 enrolled students, the RAD component averaged 70.69% (SD = 15.75), and the WebApp prototype component averaged 77.17% (SD = 16.17). The combined WAVE contribution averaged 36.64 out of 50. Because two records had zeros in both components, Table 4 also reports a sensitivity summary for the 56 non-zero records, with mean performance of 73.21% in RAD and 79.92% in WebApp prototype.

Table 4. Formal WAVE Assessment Performance

Component	Weight	Full cohort Mean (SD)	Median	Range	Non-zero records Mean (SD)
RAD report	30%	70.69% (15.75)	75.00%	0-82%	n=56; 73.21% (8.29)
WebApp Prototype	20%	77.17% (16.17)	80.25%	0-90%	n=56; 79.92% (6.83)
Combined WAVE mark	50%	36.64/50 (7.55)	38.25/50	0-41.65	n=56; 37.95/50 (2.92)

WebApp prototype had a higher mean mark than RAD in both summaries. Variability depended on the denominator: in the full cohort, the WebApp prototype SD was 16.17 compared with 15.75 for RAD; among the 56 non-zero records, the corresponding SDs were 6.83 and 8.29. Thus, lower WebApp prototype variability applies only to the non-zero subset. These differences describe two distinct tasks and cannot establish improvement over time. The low relative rating of requirements-to-design translation (B3) and the written reports of requirements ambiguity identify a plausible area for additional support, but the unlinked records do not establish an individual-level relationship between these experiences and RAD marks.

Student accounts of value and difficulty

The written responses help explain what students appeared to value. For G1, the most frequently author-reviewed categories were stakeholder communication (n = 37), real-world or professional exposure (n = 30), requirements interpretation (n = 27), technical application (n = 14), and feedback and refinement (n = 10). These responses suggest that authenticity was experienced less as a label attached to the project and more as the need to communicate, interpret another person's requirements, and apply technical knowledge in response.

The same data also show where authenticity became difficult. In G2, the most common challenges were ambiguity or interpretation of requirements (n = 31) and stakeholder communication or availability (n = 26). Time or workload pressure (n = 16), team coordination or participation (n = 14), and technical development challenges (n = 13) followed.

Students' improvement suggestions in G3 were correspondingly practical: more structured stakeholder engagement or checkpoints (n = 12), better-spaced timelines or milestones (n = 12), clearer guidance or requirements (n = 7), more lecturer consultation or support (n = 5), and better team management (n = 4). These issues do not argue against authentic assessment. They remind us that authentic variability must be managed to avoid becoming inequity.

Students also connected the assessment to capabilities they expected to use beyond the course. For G4, communication with stakeholders (n = 33) and technical or web development (n = 27) were the most common author-reviewed categories, followed by teamwork and collaboration (n = 19), professional confidence or adaptability (n = 19), requirements elicitation or analysis (n = 10), and problem solving or critical thinking (n = 8). G5 showed a similar balance between endorsement and qualification.

Positive authentic or professional experience (n = 37), stakeholder engagement value (n = 33), confidence or readiness (n = 27), and learning or skill development (n = 26) appeared alongside suggestions for improvement (n = 18) and concerns about challenge or workload (n = 14). In other words, students could value the experience while still identifying conditions that made it uneven.

Table 5. Author-Reviewed Recurring Categories in Open-Ended Responses

Prompt	Author-reviewed recurring categories (respondents mentioning)
G1 – Most valuable learning	Stakeholder communication (37); real-world/professional exposure (30); requirements interpretation (27); technical application (14); feedback and refinement (10)
G2 – Most challenging	Requirements ambiguity/interpretation (31); stakeholder communication/availability (26); time/workload pressure (16); team coordination/participation (14); technical/development challenge (13)
G3 – Improvement priorities	More stakeholder engagement/checkpoints (12); more structured timeline/milestones (12); clearer guidance/requirements (7); more lecturer support/consultation (5); team management improvements (4)
G4 – Professionally useful skills	Communication with stakeholders (33); technical/web development (27); teamwork/collaboration (19); professional confidence/adaptability (19); requirements elicitation/analysis (10); problem solving/critical thinking (8)
G5 – Overall reflections	Positive authentic/professional experience (37); stakeholder engagement value (33); confidence/readiness (27); learning and skill development (26); suggestion for improvement (18); teamwork/collaboration (16); challenge/workload concern (14)

Integration and educational implications

Table 6 brings the three evidence streams together around the main stages and implementation questions in WAVE. The purpose is not to manufacture agreement between sources, but to show where student ratings, assessed performance, and written accounts converge or create useful tension.

Table 6. Integrated Interpretation Across Student Evaluation and Performance Evidence

Evaluation focus	Evidence across sources	Interpretation
Authenticity and stakeholder engagement	Survey: authenticity M = 4.87; stakeholder value = 9.44/10. G1: communication (37), professional exposure (30), requirements interpretation (27).	Students associate authenticity with stakeholder interaction and interpreting a real need. Stakeholder benefit was not directly assessed.
Needs analysis and first validation	RAD: full-cohort M = 70.69%; non-zero M = 73.21%. Survey: B3 M = 4.67. G2: requirements ambiguity/interpretation (31).	Requirements-to-design translation warrants additional support. The unlinked sources do not establish an association at the individual level.
Solution development and later validation	WebApp prototype: full-cohort M = 77.17%; non-zero M = 79.92%. Feedback domain M = 4.86. G1: feedback/refinement (10). G4: technical development (27).	The component marks and reported feedback value describe different aspects of the experience; neither isolates the effect of the second checkpoint.

Evaluation focus	Evidence across sources	Interpretation
Fairness and implementation conditions	Survey: effectiveness/fairness M = 4.85. G2: stakeholder availability (26), time/workload (16), team coordination/participation (14).	High overall ratings coexist with uneven project conditions. Authentic variation requires active management to avoid becoming inequity.
Readiness and continuation	Survey: readiness grouping M = 4.81. G5: confidence/readiness (27). G4: technical development (27). The survey did not measure deployment, adoption, employment, or entrepreneurship.	Students reported professionally relevant experience, but perceived readiness should not be equated with demonstrated workplace competence or post-course impact.

Across the sources, stakeholder communication and requirements interpretation were prominent features of the reported experience. The written responses locate difficulties in requirements ambiguity, stakeholder access, workload, and team participation; the marks add context about performance on the associated tasks. Together, they help identify priorities for course improvement. It does not show that WAVE's two checkpoints generated the favourable responses or the recorded performance.

WAVE's principal contribution is its structured integration of stakeholder judgement into a course-level assessment sequence. Rather than treating stakeholder participation as a general feature of project-based learning, WAVE positions stakeholder validation at two identifiable assessment points: endorsing documented requirements before WebApp prototype development and evaluating the developed solution against those requirements. This configuration explicitly connects authentic stakeholder engagement with disciplinary artefacts, assessment evidence, and opportunities for feedback-informed refinement. While these practices are individually established in authentic assessment and work-integrated learning, WAVE provides an integrated assessment architecture that makes their relationships visible within undergraduate Web Engineering. The present evaluation offers an empirically informed account of this implementation, providing a transferable design proposition for further investigation in comparable educational settings (Luk & Chan, 2024; Pažur Aničić et al., 2026; van der Merwe et al., 2026).

Several alternative explanations remain compatible with the findings. Students may have valued client contact and project work generally, rather than the dual-validation sequence specifically. Positive wording, general satisfaction, and perceived lecturer expectations may also have contributed to favourable survey responses. Component marks may reflect differences in task demands, marking criteria, or group performance. The higher WebApp prototype mean cannot distinguish these explanations from a beneficial role for stakeholder feedback. The evidence therefore supports further examination of the sequence, rather than attributing outcomes to its mechanism.

A direct test of the proposed mechanism would link the initial requirements, dated stakeholder feedback, the student's or team's response, and the later assessment judgement. Such records would show whether feedback changed the work and whether that change addressed a relevant criterion. Criterion-level evidence mapped to CLO2 and CLO3 would also separate overall marks from the attainment of particular outcomes. These are priorities for subsequent evaluation, not evidence provided by the current component totals.

The ceiling effect limits the survey's contribution. More than half the respondents selected the maximum response on every five-point item, and every domain median was 5.00. High alpha values do not establish that the domains measure distinct constructs. The small differences between domain and item means should therefore be treated as descriptive signals, not as a ranking of WAVE's strengths or a measure of capability gain.

A revised questionnaire should therefore do more than add negatively worded items. The deeper issue is construct clarity. Experience, satisfaction, perceived change, and demonstrated capability are different and

should be measured separately. Cognitive interviewing could test how students interpret the items, while behaviour-specific questions could ask when students received feedback, what changed as a result, and who made the decision. Where feasible, separate administration from grading and lecturer access to reduce perceived pressure. Validation would then need a larger and more varied sample.

The fairness findings show a similar tension. Students rated the assessment highly even though the written responses referred to ambiguous requirements, variable stakeholder access, team coordination, workload, and uneven support. These observations do not directly contradict D6, which asked about fairness across groups working with different stakeholders. They broaden the fairness question by pointing to differences in opportunity, responsiveness, and contribution within teams. Recent WIL research makes the same point from another angle: authentic settings can produce competing expectations and different ideas about what counts as valid evidence (van der Merwe et al., 2026). A project can feel worthwhile and still be unevenly experienced.

For future implementations, the practical response is not to remove authentic variation but to manage it more deliberately. Reasonable safeguards include agreed expectations for stakeholder responsiveness, early scope checks, planned feedback opportunities, records of individual contribution, and a fallback process when stakeholder access breaks down. The written responses also point to the importance of lecturer support. Authentic assessment depends as much on the climate inside the course as on external engagement; students need to be able to ask for clarification and question feedback without feeling that doing so puts their assessment position at risk.

Professional confidence, portfolio potential, and understanding of deployment are intermediate perceptions. Sustainable assessment provides a rationale for continued judgement and learning (Boud & Soler, 2016), while entrepreneurship education highlights learning through action (Neck & Greene, 2011). Neither perspective equates confidence with workplace competence, stakeholder adoption, or venture creation. The entrepreneurship element of WAVE should therefore remain explicitly outside the empirical claims of this evaluation.

The continuation pathway in WAVE should therefore be studied as a separate outcome. Follow-up work could document whether WebApp prototypes were refined, adopted, deployed, entered into further development, or abandoned, and why. Combined stakeholder and student accounts would then allow examination of whose needs were served and whether continuation created educational or social value, consistent with McArthur's (2023) broader account of authenticity.

LIMITATIONS AND FUTURE RESEARCH

The findings should be interpreted within the scope of this course-level evaluation. The study examined one 14-week WAVE implementation involving 58 students at a single institution, with 52 completing the post-implementation evaluation. This provided substantial coverage of the cohort and enabled examination of student perceptions, qualitative accounts, and formal assessment performance together. However, because the evaluation focused on one implementation, broader studies involving additional cohorts and institutional contexts would be valuable for examining the consistency and transferability of the observed patterns.

The evaluation instrument captured students' perceptions across authenticity, capability development, engagement and feedback, assessment effectiveness, fairness, and professional readiness. Responses showed consistently strong endorsement of the WAVE experience. The concentration of ratings at the upper end of the scale suggests that refining the instrument could better distinguish between dimensions of the learning experience.

Therefore, expert review, cognitive interviewing, and pilot testing are recommended before subsequent use. Future studies could also link de-identified perception data with criterion-level assessment evidence to examine more closely how students' reported experiences relate to demonstrated performance.

Formal assessment records provided additional evidence through the RAD and WebApp prototype components. These records reflect authentic team-based work and therefore represent the collaborative nature of the WAVE implementation. Future evaluations could strengthen this evidence by explicitly mapping assessment criteria to CLO2 and CLO3 and including records of individual contributions. This would allow examination of both team achievement and individual capability development while preserving the assessment's collaborative character.

The qualitative responses added contextual depth by identifying what students valued, the challenges they encountered, and the improvements they considered important. Author-reviewed content analysis provided a structured account of these responses, while future studies could further strengthen qualitative evidence through independent coding and follow-up interviews with students and stakeholders. Stakeholder perspectives would be particularly valuable for examining whether requirements were interpreted appropriately, whether proposed solutions addressed the identified needs, and how stakeholder feedback influenced subsequent development decisions.

Future research should also examine WAVE implementation more directly. Maintaining dated records of stakeholder-validation checkpoints, the feedback provided, and subsequent revisions would make the feedback-to-revision process more visible and allow more systematic investigation of the proposed dual-validation sequence. Comparative studies with an alternative assessment design could further clarify WAVE's distinctive contribution, while multi-cohort and multi-institution studies would establish its applicability across different teaching and learning environments.

Finally, the wider WAVE model includes a continuation pathway beyond formal course assessment. Longitudinal follow-up could examine whether students' WebApp prototypes progress to further refinement, portfolio use, deployment, stakeholder adoption, or entrepreneurial development. Such evidence would extend the present course-level findings and provide a stronger basis for understanding WAVE not only as an authentic assessment architecture but also as a potential bridge between university project work, professional practice, and continued innovation.

CONCLUSION

This study evaluated WAVE as a stakeholder-integrated authentic assessment design for undergraduate Web Engineering and demonstrated how stakeholder judgement can be systematically positioned around assessable course outputs. Over the 14-week implementation, students reported consistently positive experiences and particularly valued stakeholder communication, interpreting real requirements, applying technical skills, professional exposure, teamwork, and opportunities to refine their work through feedback. Formal assessment records provided complementary descriptive evidence of performance across the Requirements Analysis Document (RAD) and WebApp prototype components, reflecting students' assessed work on the intended disciplinary tasks without establishing individual course learning outcome attainment.

The findings indicate that WAVE offers a coherent course-level assessment architecture that connects requirements analysis, stakeholder validation, solution development, iterative feedback, and formal assessment within a single authentic project sequence. Its contribution is making stakeholder judgement visible at key points in the assessment process and linking those interactions directly to disciplinary artefacts and course assessment evidence. At the same time, the implementation highlighted practical considerations involving stakeholder availability, workload, and team coordination, suggesting that authentic assessment requires deliberate support and management to ensure consistent student experience.

Overall, WAVE shows strong potential as a structured authentic assessment approach for Web Engineering and provides a practical model for integrating real stakeholder engagement into project-based learning. Further studies across broader cohorts, institutions, and educational contexts would help extend the evidence base and examine the model's applicability, implementation consistency, and contribution to learning outcomes over time.

REFERENCES

1. Ashford-Rowe, K., Herrington, J., & Brown, C. (2014). Establishing the critical elements that determine authentic assessment. Assessment & Evaluation in Higher Education, 39(2), 205–222. https://doi.org/10.1080/02602938.2013.819566
2. Biggs, J. (1996). Enhancing teaching through constructive alignment. Higher Education, 32(3), 347–364. https://doi.org/10.1007/BF00138871
3. Billett, S. (2009). Realising the educational worth of integrating work experiences in higher education. Studies in Higher Education, 34(7), 827–843. https://doi.org/10.1080/03075070802706561
4. Boud, D., & Soler, R. (2016). Sustainable assessment revisited. Assessment & Evaluation in Higher Education, 41(3), 400–413. https://doi.org/10.1080/02602938.2015.1018133
5. Chen, C.-Y., & Chong, P. P. (2011). Software engineering education: A study on conducting collaborative senior project development. Journal of Systems and Software, 84(3), 479–491. https://doi.org/10.1016/j.jss.2010.10.042
6. Dong, D., Butler, R., & Herbert, J. (2022). Evaluation of the effectiveness of developing real-world software projects as a motivational device for bridging theory and practice. Journal of Further and Higher Education, 46(9), 1275–1289. https://doi.org/10.1080/0309877X.2022.2070727
7. Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. Educational Technology Research and Development, 52(3), 67–86. https://doi.org/10.1007/BF02504676
8. Kauppinen, R., Lagstedt, A., Lindstedt, J. P., & Rainio, O. (2022). Software engineering education with industry-mentored projects. PACIS 2022 Proceedings, 134. https://aisel.aisnet.org/pacis2022/134/
9. Luk, L. Y. Y., & Chan, C. K. Y. (2024). Assessing student learning in work-integrated learning (WIL): A systematic review on assessment methods and purposes. Studies in Educational Evaluation, 81, 101345. https://doi.org/10.1016/j.stueduc.2024.101345
10. McArthur, J. (2023). Rethinking authentic assessment: Work, well-being, and society. Higher Education, 85(1), 85–101. https://doi.org/10.1007/s10734-022-00822-y
11. Neck, H. M., & Greene, P. G. (2011). Entrepreneurship education: Known worlds and new frontiers. Journal of Small Business Management, 49(1), 55–70. https://doi.org/10.1111/j.1540-627X.2010.00314.x
12. 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
13. Pažur Aničić, K., Divjak, B., & Ferns, S. J. (2026). The role of employers in the assessment process in WIL: Students' and employers' perspectives. Industry and Higher Education, 40(1), 59–70. https://doi.org/10.1177/09504222251316340
14. Sokhanvar, Z., Salehi, K., & Sokhanvar, F. (2021). Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. Studies in Educational Evaluation, 70, 101030. https://doi.org/10.1016/j.stueduc.2021.101030
15. Twyford, E. J., & Dean, B. A. (2024). Examining the contribution of diverse feedback perspectives during a work-integrated learning intervention. Journal of Accounting Education, 69, 100931. https://doi.org/10.1016/j.jaccedu.2024.100931
16. van der Merwe, T., de Jager, T., & White, C. J. (2026). Reconciling work-integrated learning assessment: A collaborative employer–coordinator approach. Transformation in Higher Education, 11(0), a752. https://doi.org/10.4102/the.v11i0.752
17. Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. Assessment & Evaluation in Higher Education, 43(5), 840–854. https://doi.org/10.1080/02602938.2017.1412396
18. Vlachopoulos, D., & Makri, A. (2024). A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations. Studies in Educational Evaluation, 83, 101425. https://doi.org/10.1016/j.stueduc.2024.101425