EsyGrade: An AI-Based Essay Assessment System for Economics Education

Ramadzan Defitri Pratama[1], Khresna Bayu Sangka[2] and Cicilia Dyah Sulistyaningrum Indrawati[3]

[1]Master of Economics Education, Faculty of Teachers Training and Education, Sebelas Maret University, Indonesia

[2]Economics Education, Faculty of Teachers Training and Education, Sebelas Maret University, Indonesia

[3]Office Administration Education, Faculty of Teachers Training and Education, Sebelas Maret University, Indonesia

ramadp_24@student.uns.ac.id

b.sangka@staff.uns.ac.id (corresponding author)

ciciliadyah@staff.uns.ac.id

https://doi.org/10.34190/ejel.24.3.4475

Abstract: This study aims to design, develop, and evaluate EsyGrade, a web-based essay grading system integrated with ChatGPT to support the assessment of conceptual understanding in economics education. This study is motivated by the prevalence of multiple-choice questions in Indonesian high schools due to efficiency considerations, while essay assessments better suited to measuring conceptual reasoning are still rarely used because of high assessment load. The study employed a simplified Research and Development (R&D) approach consisting of seven stages. Data were collected through expert validation using Aiken's V, a user response questionnaire, and pretest–post-test, and were later analysed using N-Gain and effect size (Cohen's d). The research findings indicate that EsyGrade has a high level of validity based on the expert evaluation and an excellent acceptance rate in the initial pilot test. In the main field trial, the results indicated a moderate improvement in conceptual understanding based on N-Gain, with a "large" effect size. These findings suggest that the overall change in learning outcomes indicates a fairly strong impact, although the degree of improvement varied among students. In particular, the system's key value lies in its ability to generate structured, reflective feedback, so that it serves not only as a summative assessment tool but also as a means for formative one. This study contributes to the development of e-learning by demonstrating that AI-based essay grading systems can be designed to improve efficiency while supporting deeper learning through the integration of pedagogical principles and technology.

Keywords: Automated essay scoring, ChatGPT, Conceptual understanding, Economics education, Formative assessment, Educational technology

1. Introduction

Economics education at vocational or public high schools emphasizes the importance of in-depth mastery of concepts so that students are able to analyse economic phenomena, understand the interrelationships between variables, and evaluate policies in a real-world context. In Indonesian context, the implementation of the Freedom to Learn curriculum (*Kurikulum Merdeka*) developed in response to the disruption of learning caused by the COVID-19 pandemic marks a significant turning point in reevaluating student learning outcomes. As the pandemic has led to learning loss and gaps in understanding, this curriculum was designated as a solution to emphasize learning recovery through the strengthening of essential competencies, particularly critical thinking skills and conceptual understanding. Thus, it serves as not only a learning framework but also a new benchmark for assessing learning outcomes that are more oriented toward understanding rather than mere mastery of facts.

In economics education, this approach is particularly relevant because economic concepts are abstract and interconnected, requiring advanced analytical and reasoning skills. This approach aligns with the concept of deep learning, which demands that students not only memorize information but also understand, analyse, and apply concepts in context (Marton and Hounsell, 1984; Biggs and Tang, 2011). Therefore, the success of this curriculum's implementation depends heavily on the assessment system's ability to measure deep conceptual understanding, rather than merely superficial learning outcomes.

However, the results of the 2022 Program for International Student Assessment (PISA) indicate that Indonesian students' performance remains relatively low, with scores ranging from 359 to 371 far below the OECD average

Cite this article: Pratama, R.D., Sangka, K.B. and Indrawati, C.D.S. 2026. "EsyGrade: An AI-Based Essay Assessment System for Economics Education", *Electronic Journal of e-Learning*, 24(3), pp 16-32, https://doi.org/10.34190/ejel.24.3.4475

of 472 to 476 (OECD, 2023). Although PISA does not specifically measure economics, the indicators used such as reasoning skills, problem solving, and the application of concepts in real-world situations reflect key aspects of conceptual understanding that are also central to the learning objectives of economics. These low scores indicate that Indonesian students still struggle to connect concepts with real-world practice and to develop higher order reasoning.

Although essay questions are considered more representative for measuring conceptual understanding because they allow students to express their arguments and reasoning (Elliott and Balasubramanyam, 2016; Beldar, 2025), assessment practices in the classroom are still dominated by multiple-choice questions. A study by Kamalia (2023) found that the majority of high school economics teachers in Indonesia prefer objective questions for their efficiency, despite these instruments' limitations in measuring higher-order thinking skills (Petersen, Craig, and Denny, 2016). Preliminary research conducted by researchers in Surakarta reveals that 73% of teachers tend to use multiple-choice questions more frequently, while less than a third of them employ essays, and the main obstacle is that correcting one essay may take an average of six to eight minutes to assess (Pasaribu, Budiman, and Indrarini, 2024), making it difficult for teachers to provide quick and reflective feedback. This situation slightly creates a research gap between curriculum policies that emphasize authentic assessment and the actual practices of classroom assessment.

On the other hand, advancements in digital technology particularly in the context of e-learning and artificial intelligence have opened up new opportunities in educational assessment practices. AI-based systems enable automated assessment, real-time feedback, and personalized learning, all of which are key elements in the modern digital learning ecosystem (Mizumoto and Eguchi, 2023). Previous research from Indonesia has developed an essay assessment system based on Latent Semantic Analysis (LSA), but this system exhibits limitations in comprehending context, argument logic, and language flexibility (Kinanti and Qoiriah, 2020). Meanwhile, international research has begun to utilize Large Language Models (LLM) for automatic assessment, including GPT-3.5 and GPT-4, which have proven to be superior in assessing students' answers semantically and contextually (Pack, Barrett, and Escalante, 2024; Smerdon, 2024). Mizumoto and Eguchi (2023) even show that the integration of ChatGPT with learning analytics can offer reflective feedback and personalize learning experiences. However, research in Indonesia that specifically focuses on developing ChatGPT-based automatic assessment systems for high school and vocational school economics subjects is still quite limited.

This study introduces two novel aspects in the field of economics education. Firstly, it shifts the focus from merely assessing learning outcomes based on facts to evaluating students' conceptual understanding in line with the Merdeka Curriculum's requirements. Secondly, the automated assessment system developed using ChatGPT not only generates scores but also provides reflective feedback that supports formative assessment and differentiated learning. This innovation aims to bridge the gap between curriculum expectations and the limitations of traditional assessment practices.

Based on the background and research gaps, this study aims to (1) design and develop a ChatGPT-based automatic assessment system that is in line with learning outcome indicators, (2) test the feasibility and practicality of the developed system according to experts, teachers, and students, and (3) analyse the development of students' conceptual understanding after using the system through a pretest–post-test design with Normalized Gain (N-Gain) analysis.

2. Literature Review

2.1 Conceptual Understanding in Economics Education

Conceptual understanding is a foundation for meaningful learning, especially in economics where students must integrate abstract concepts and apply them in real contexts. Rittle-Johnson, Schneider, and Star (2015) define conceptual understanding as the ability to connect ideas and use them to solve problems, while Anderson and Krathwohl, (2001) position it as a key stage in the revised Bloom's taxonomy. Within Bloom's framework, conceptual understanding spans from remembering (C1) and understanding (C2), to applying (C3), analysing (C4), evaluating (C5), and creating (C6) (Bloom, 1956; Anderson, and Krathwohl, 2001). These indicators are central to assessing deep learning outcomes in economics education, as students are expected not only to recall definitions but to analyse market dynamics, evaluate policy impacts, and design alternative solutions. However, studies highlight persistent difficulties among Indonesian students in achieving higher-order thinking skills, with many remaining at the level of factual recall (Arini, Dewi, and Wibawa, 2024). Despite its importance, the development of conceptual understanding is often constrained by existing assessment practices. Thus,

examining how assessment is implemented in economics education becomes crucial to determine whether it aligns with the goal of fostering higher-order thinking skills.

2.2 Assessment Practices in Economics Education

Despite the centrality of conceptual understanding, classroom assessment practices remain dominated by multiple choice tests. Teachers favour these instruments because they are efficient, quick to grade, and objective (Kamalia, 2023). Yet, such tools are limited in capturing higher-order thinking skills (Petersen et al., 2016). Essay assessments, by contrast, are more representative for evaluating students' reasoning and conceptual mastery (Elliott and Balasubramanyam, 2016; Beldar, 2025). This is consistent with the findings of Walstad (2006) and Al-Obaydi, Pikhart, and Tawafak (2023), which show that essay questions are more effective in measuring the depth of economic understanding than multiple-choice questions. They allow students to articulate arguments, demonstrate critical thinking, and connect theories with socio-economic realities (Van Wyk, 2015). However, the manual grading process is time-consuming, with each essay requiring six to eight minutes on average (Pasaribu, Budiman, and Indrarini, 2024), and in a class of 30–32 students, teachers can spend three to five hours marking a single task (Rafi, Ramadhani, and Sanjaya, 2025). This workload seems to create delays in providing feedback and reduces the potential for formative assessment.

2.3 Teacher Workload and Assessment Challenges

Teacher workload is a recurring issue in the implementation of authentic assessment. According to Runtuwene and Tangkawarow, (2020), teacher responsibilities span lesson planning, instruction, assessment, and administrative duties. Although Minister of Elementary and Secondary Education Regulation No. 11 of 2025 stipulates a standard weekly workload of 37.5 hours (2025), empirical studies show that assessment tasks, especially essay grading, often exceed this limit (Morris *et al.*, 2024, 2026). This heavy workload not only affects teachers' ability to provide timely and reflective feedback but also impacts instructional quality and teacher well-being (Sabon, 2020; Johnson and Coleman, 2025). These challenges underscore the imperative of embracing technology to mitigate administrative burdens while maintaining the integrity of assessment processes.

2.4 Automated Essay Scoring and Artificial Intelligence in Education

To address such challenges, Automated Essay Scoring (AES) systems have been developed, relying on Natural Language Processing (NLP). Early models such as Latent Semantic Analysis (LSA) have been applied in Indonesian contexts (Kinanti and Qoiriah, 2020), yet remain limited in handling complex argumentation and contextual reasoning (Alief, Irawati, and Mude, 2023). The emergence of Large Language Models (LLMs), including GPT-3.5 and GPT-4, introduces greater capability in semantic and logical analysis. International studies have shown that ChatGPT can effectively evaluate essays, provide formative feedback, and support personalised learning (Mizumoto and Eguchi, 2023; Foung, Lin, and Chen, 2024; Latif and Zhai, 2024; Smerdon, 2024). This suggests that the feedback generated by language models can help explain students' conceptual errors and support a deeper learning process. Thus, this development aligns with the goals of the Merdeka Curriculum, which promotes personalized, reflective, and meaningful learning experiences.

2.5 Research Gap and Contribution

Despite the development of various automated grading systems, including GradedPro, CoGrader, and Automark, most of these systems primarily prioritize efficiency through the automation of scoring. However, these systems generally lack explicit integration of pedagogical aspects, such as reflective feedback to rectify conceptual errors, alignment with Bloom's taxonomy, the utilization of structured rubrics, and prompt flexibility tailored to individual learning requirements.

In accordance with this, while international research increasingly emphasizes the potential of artificial intelligence (AI) in essay grading, its application in economics education in Indonesia remains limited. Existing studies have not adequately addressed how AI-based systems can assess conceptual understanding aligned with Bloom's taxonomy while simultaneously reducing teachers' workload. Therefore, this study addresses this gap by developing a ChatGPT-based essay grading system (EsyGrade) specifically designed for high school economics education in Indonesia, and by evaluating its impact on students' conceptual understanding through a pre-test–post-test design with N-Gain analysis, paired Cohen's d, and distribution analysis of N-Gain categories.

3. Methodology

This study employed a Research and Development (R&D) design adapted from the model of Borg and Gall (Bennett, Borg, and Gall, 1984). Educational R&D aims to develop and validate products in iterative cycles of

design, expert validation, revision, and field testing (Plomp and Nieveen, 2013). While Borg and Gall originally proposed ten steps, this study streamlined the process into seven stages due to contextual and time constraints, focusing on the development of an automated essay assessment system using ChatGPT for evaluating students' conceptual understanding. The research phases include: (1) needs analysis, (2) design planning, (3) prototype development, (4) expert validation, (5) product revision, (6) field testing, and (7) evaluation and finalization. This simplification aims to maintain a balance between the depth of development and the feasibility of implementation in an educational context.

Besides, students' conceptual understanding in this study was assessed using essay-based tests administered in pre-test and post-test formats. The items were designed based on the revised Bloom's taxonomy (C1–C6) to capture different levels of cognitive processes. Student responses were evaluated using a rubric consisting of four criteria: conceptual mastery, argumentation, relevance of examples, and structure, ensuring alignment between automated scoring and pedagogical standards.

A validation instrument, meticulously crafted by experts, comprising 15 items, was employed to assess the appropriateness of the indicators, the clarity of the rubric, and the feasibility of the system from both an assessment and a technological standpoint. This evaluation was conducted by educational assessment experts and IT experts. The instrument underwent expert evaluation using a Likert scale (ranging from 1 to 4) and was subsequently analysed employing Aiken's V.

The pilot study was then conducted in two phases, a preliminary pilot study (with 10 participants) and a main pilot study (72 respondents). These sample sizes were chosen to represent the exploratory and confirmatory phases of the development research, in which the preliminary pilot study was used to identify initial improvements, while the main pilot study was used to evaluate the system's performance on a broader scale. During the initial pilot phase, questionnaires were distributed to teachers and students to assess usability and practicality, while observations were conducted during implementation to document technical and pedagogical challenges.

Data were collected using a pre-test and post-test design with equivalent question characteristics to measure changes in conceptual understanding after exposure to AI-generated feedback. In addition, questionnaires and observations were employed to capture user responses and implementation challenges.

Data analysis combined quantitative and qualitative approaches. Quantitative analysis included Normalized Gain (N-Gain) to measure relative improvement (Hake, 1998), paired Cohen's d to assess the magnitude of change (Cohen, 1988), and distribution analysis of N-Gain categories (low, medium, high) to examine individual variation. Qualitative data from observations and open-ended responses were analysed descriptively to provide contextual insights into system usability and classroom integration.

4. Results

4.1 Needs Analysis

A previous study worked on a systematic literature review (SLR) on automated essay scoring (AES) between 2020 and 2025, which revealed a notable surge in research in this field, particularly with the adoption of GPT-based large language models (LLMs) such as GPT-3.5, GPT-4, and ChatGPT (Pratama and Sangka, 2025). These models demonstrate high accuracy and reliability, especially when combined with prompt designs based on rubrics (Tang *et al.*, 2024; Yavuz, Çelik and Yavaş Çelik, 2025). However, the literature also highlights challenges related to validity, consistency, and domain-specific appropriateness (Steiss *et al.*, 2024; Gandolfi, 2025). Importantly, existing studies are still concentrated on language education, with limited application in economics education, indicating a gap in domain-specific development (Lee *et al.*, 2024).

Complementing these findings, the needs analysis involving local schools in Surakarta revealed that essay assessments remain underutilised in economics classrooms. Teachers reported a strong preference for multiple-choice questions due to efficiency concerns, with essay assessments accounting for less than one-third of practice. On average, grading a single essay required six to eight minutes, which created a heavy workload in classes of 30 or more students. This burden limited the ability of teachers to provide timely and reflective feedback. These results confirm the broader international literature, which recognises manual essay scoring as labour-intensive, subjective, and inconsistent (Mendonça, Quintal, and Mendonça, 2025).

In the context of e-learning, these limitations become increasingly relevant, as the effectiveness of digital learning depends heavily on the availability of rapid and meaningful feedback. Thus, from both theoretical and practical perspectives, there is a compelling need for AI-based assessment systems that not only enhance the

efficiency of assessment but are also capable of facilitating the development of conceptual understanding through reflective and contextual feedback.

4.2 Design Planning and Prototype Development

Drawing inspiration from both international literature and field requirements, the system was meticulously designed to encompass essay input, automated scoring facilitated by ChatGPT, and structured feedback output. A rubric aligned with Bloom's revised taxonomy (C1–C6) was constructed to ensure the evaluation captured different levels of conceptual understanding, from recall to higher-order reasoning (Anderson and Krathwohl, 2001). This choice directly addresses the gap identified in prior research, where AES systems often emphasise surface-level correctness but lack explicit alignment with pedagogical standards (Poole and Coss, 2023).

The prototype system, named EsyGrade, was developed as a web-based platform integrated with the ChatGPT API to support automated essay assessment in economics education. EsyGrade was designed with dual user roles: teachers and students. Those taking the role as teachers can create question packages, design rubrics, embed GPT-based scoring prompts, monitor submissions, and export results. On the other side, students can log in, answer essay questions, and directly access scores and rubric-based feedback.

The system architecture was implemented through four main modules: (1) Essay Input Module, where teachers add essay tasks and students submit answers; (2) Rubric and Prompt Module, which encodes assessment criteria into GPT-compatible instructions; (3) Scoring and Feedback Engine, powered by the GPT-3.5 Turbo API to produce scores and reflective feedback; and (4) Result Display Module, which presents individual results, summary analytics, and export options for teachers and students.

The early prototype of EsyGrade successfully demonstrated real-time essay scoring and transparent rubric-based reporting. Figure 1 presents the homepage, while Figure 2 shows the teacher dashboard with features for question package management, monitoring student answers, and tracking token usage.

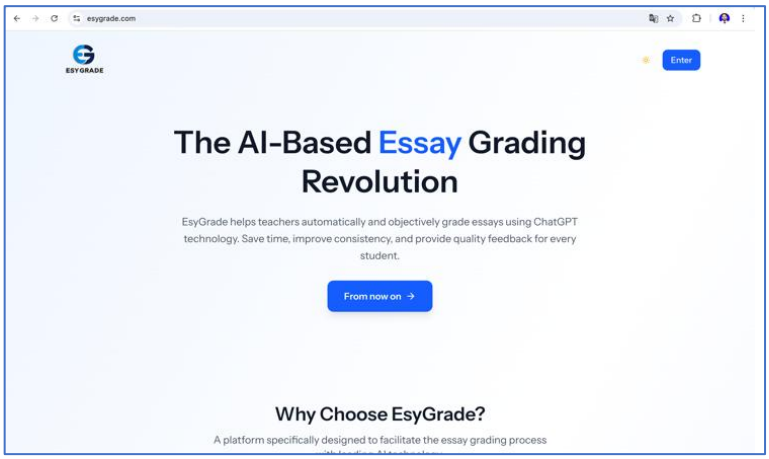

Figure 1: Early Prototype Snapshot of EsyGrade Homepage

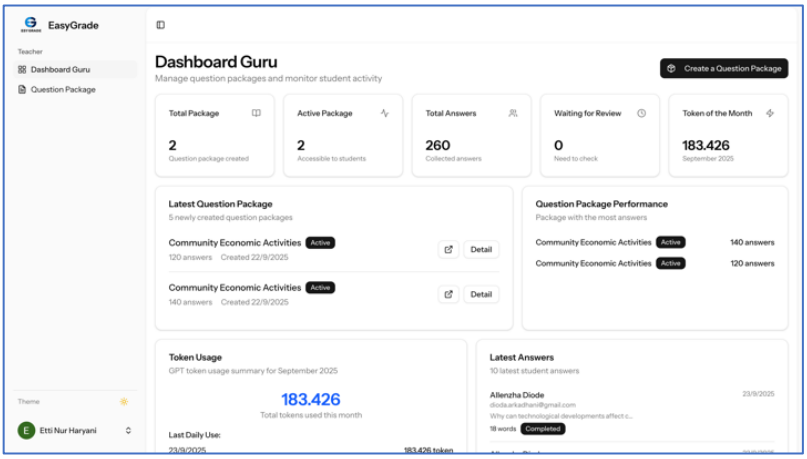

Figure: 2 Teacher Dashboard on EsyGrade

The development of this prototype demonstrates that the system can grade essays in *real-time* and present structured results based on a rubric. This capability indicates the system's potential to improve grading efficiency while providing faster and more transparent feedback. From an e-learning perspective, this feature is relevant because timely, criteria-based feedback is one of the key factors in supporting meaningful and reflective learning.

4.3 Expert Validation

The essay assessment system developed in this study went to a validation phase, involving a total of six experts to scrutinize the items' assessment and technological aspects. The validation instrument comprises 15 items that cover two main aspects: (1) assessment design quality (appropriateness of indicators, clarity of rubrics, and alignment with learning objectives) and (2) technological aspects (system functionality, ease of use, and clarity of the interface). Each aspect is rated on a 1–4 Likert scale. Example items include "The system aligns with the curriculum and applicable assessment standards" and "The system interface is easy for users to understand". Table 1 presents the content validity coefficient (V value) for each item obtained using the Aiken formula.

Table 1: Content validity coefficient (V-value) of the items

Assessment Expert		Technology Expert	
Items	V	Items	V
1	1	1	0.88888889
2	1	2	0.88888889
3	1	3	1
4	0.77777778	4	0.88888889
5	0.88888889	5	0.88888889
6	0.88888889	6	0.77777778
7	1	7	0.88888889
8	0.88888889	8	0.88888889
9	1	9	0.88888889
10	0.88888889	10	0.88888889
11	0.88888889	11	1
12	1	12	0.88888889
13	1	13	1
14	1	14	1
15	1	15	1

The validation results conducted by six experts, comprising three assessment experts and three technology experts, revealed that the Aiken's V values exhibited a range of 0.78 to 1. Notably, most of the items fulfilled the feasibility criteria, with values exceeding the V-table threshold of 0.79, thereby demonstrating high content validity. However, certain items demonstrated slightly below-threshold scores, particularly in the domains of system usability and interface clarity, suggesting a need for enhancement.

Overall, the results confirm that most items are valid and feasible for use, especially in the domains of assessment content and rubric clarity. Nevertheless, improvements were made to items with lower coefficients in response to expert suggestions. In addition to the quantitative data, experts provided written feedback focusing on rubric flexibility, and user interface enhancement. These comments were used to revise and strengthen the platform before field testing. Thus, the validation phase serves not only as an evaluation process but also as the basis for refining the system before it enters the field-testing phase.

4.4 Product Revision

Following expert validation, several items with Aiken's V values below the threshold (0.79) required revision, particularly in the aspects of system usability and interface clarity. Experts recommended improvements in two key areas: (1) rubric flexibility, to allow teachers to customise criteria and scoring according to learning objectives, and (2) user interface enhancements, to improve navigation and provide short guidance for first-time users.

These suggestions were accommodated in the revised version of EsyGrade. As illustrated in Figure 3, the rubric and scoring criteria are now fully customisable, enabling teachers to set different criteria weights (e.g., conceptual understanding, argumentation, relevance, language) and determine maximum scores per criterion. This change ensures greater alignment with classroom needs and curriculum indicators.

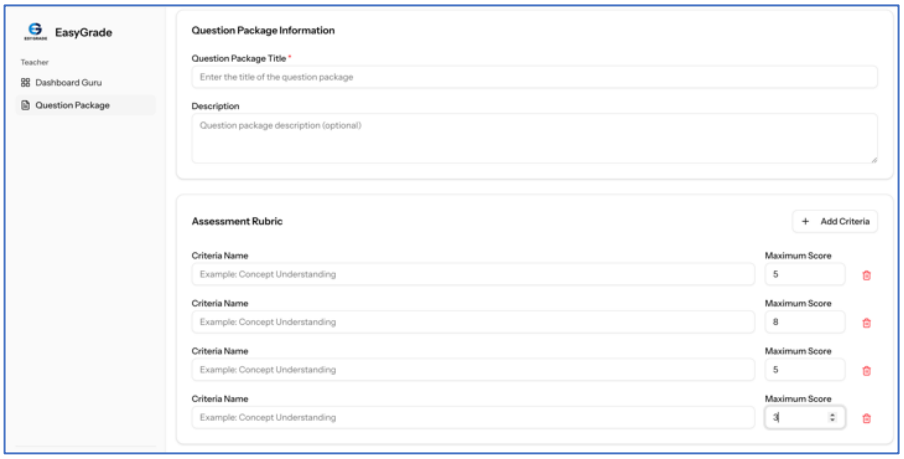

Figure 3:Revision Rubric Flexibility

Furthermore, Figure 4 demonstrates improvements in user-friendliness, where a simplified workflow (Create Question Package – Students Working – Automatic Results) is presented on the landing page, supported by a short usage guide to assist new users.

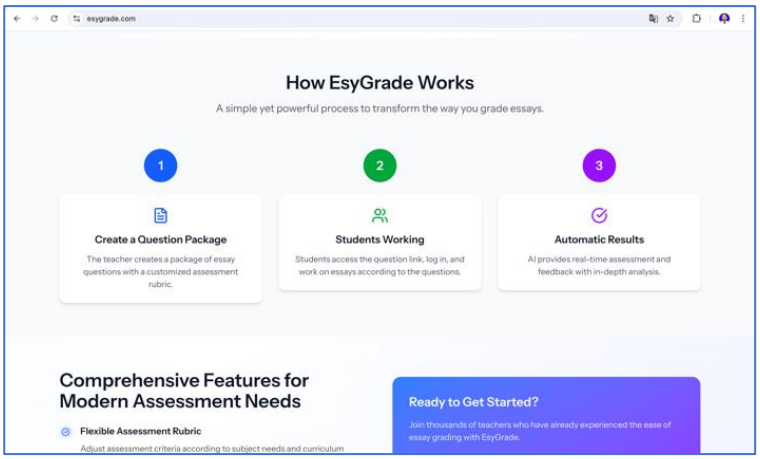

Figure 4: Revised User Interface

These changes indicate that the quality of a system is determined not only by the accuracy of its assessments but also by its ease of use and the clarity of user interactions. In the context of e-learning, the aspect of usability is crucial because it influences users' acceptance of the technology as well as the effectiveness of its implementation in learning. Consequently, the product revision phase serves as a technical enhancement and an endeavour to ensure that the developed system is pedagogically pertinent and adaptable to user requirements prior to entering the field-testing phase.

4.5 Preliminary Testing

A preliminary field test was conducted with ten students utilizing a Likert scale (1 = poor, 4 = very good) to assess the feasibility and practicality of the revised concept after expert input, as shown in Table 2.

Table 2: Preliminary Testing Results

Preliminary Testing		
Aspect	**Mean**	**Category**
Content Suitability & Validity	3.73333333	Very Good
Reliability & Fairness of Assessment	3.73333333	Very Good
Alignment with Educational Assessment Principles	3.46666667	Very Good
Clarity of Reporting Results	3.56666667	Very Good
Data Ethics & Privacy	3.55	Very Good
Integration with the Curriculum	3.9	Very Good

The questionnaire assessed the usability, interface clarity, rubric clarity, and feedback usefulness. All items as shown in Table 2, achieved the "Very Good" category (mean ≥ 3.26), with no item falling below the cut-off. These results designate strong acceptability and ease of use from the student side, confirming that the revisions flexible rubric customisation and a clear, step-by-step usage guide successfully addressed earlier expert feedback. In the context of e-learning, user acceptance is a key factor in determining the success of learning technology implementation. Systems that are easy to use and provide a clear user experience tend to be adopted more readily by users, making this initial pilot phase a crucial foundation before conducting further testing on a larger scale. Consistent "Very Good" ratings across aspects likely suggest that EsyGrade is ready to proceed to wider implementation in the main field test.

4.6 Main Field Testing

4.6.1 Normalized gain (n-gain) analysis

The main field testing involved 72 students across two grade levels: Grade 10 (36 students) who answered essay questions on Economic Activities and Grade 11 (36 students) who worked on Business and Management. Both groups completed the pre-test and post-test each of which has five essay items designed with equivalent type and weight according to Bloom's revised taxonomy. This standard is intended to ensure that changes in learning outcomes reflect improvements in conceptual understanding, rather than differences in the difficulty level of the questions.

The analysis employed the Normalised Gain (N-Gain) method, categorising improvement as *High* (≥ 0.7), *Medium* (0.3–0.69), and *Low* (< 0.3). Results showed that the average N-Gain score for both classes was in the Medium category, as shown in Table 3. These findings suggest that the use of a ChatGPT-based essay grading system contributes to an improvement in students' conceptual understanding. However, the fact that the improvement falls into the "moderate" category indicates that the resulting impact is moderate, meaning that not all students experienced optimal improvement.

Table 3: Main Field Testing

Main Field Testing							
Economic Activities (10th Grade)				**Business and Management (11th Grade)**			
Pre-Test Mean	**Post-test Mean**	**N-Gain Mean**	**Category**	**Pre-Test Mean**	**Post-Test Mean**	**N-Gain Mean**	**Category**
73.9166667	85.6388889	0.44358861	Medium	74.6111111	82.6944444	0.30379371	Medium

In this context, the observed improvement can be attributed to the role of the reflective feedback generated by the system. By using equivalent items in both the pre-test and post-test, changes in learning outcomes are more likely to result from students' ability to recognize their previous errors and correct them based on the feedback provided. Therefore, the system functions not only as an assessment tool but also as a formative learning tool that supports students' reflection on their conceptual understanding.

4.6.2 Effect size analysis

To complete the analysis, effect sizes were measured using paired Cohen's d. The results as shown in Table 4, show that Grade 10 had an effect size of 1.68, while Grade 11 had an effect size of 1.08, both of which fall into the "Large" category.

Table 4: Effect Size Testing

Class	Mean Difference	SD Difference	Cohen's d	Category
Economic Activities (10th Grade)	11.72	6.98	1.68	Large
Business and Management (11th Grade)	8.08	7.47	1.08	Large

*The interpretation of effect sizes follows Cohen's (1988) criteria, namely small (0.2), moderate (0.5), and large (0.8).

These findings suggest that, although the relative improvement measured using N-Gain falls into the moderate category, the magnitude of the change in learning outcomes is quite significant. This indicates that the system has a strong practical impact, while the degree of improvement varies among students.

4.6.3 Distribution of n-gain

To obtain a more detailed picture of the variation in improvement at the individual level, an analysis of the N-Gain distribution was conducted based on low, moderate, and high categories, as shown in Figure 5.

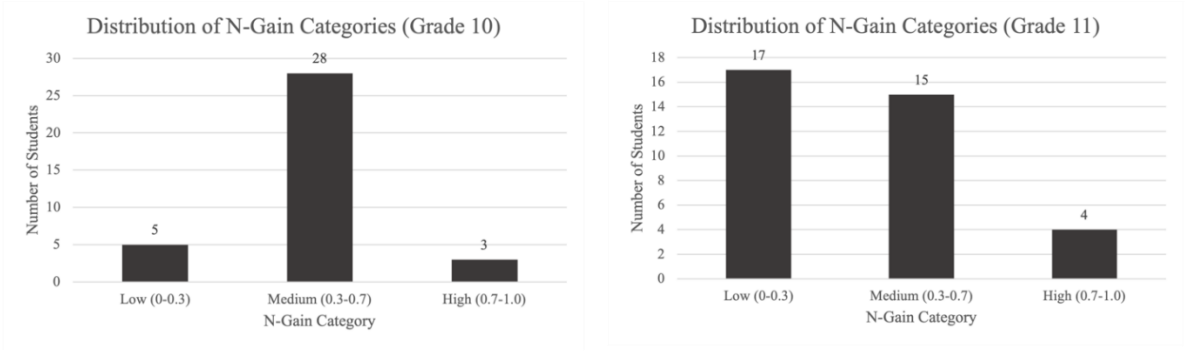

Figure 5: Distribution of N-Gain Categories

Based on the distribution of N-Gain shown in Figure 5, in 10th grade, the majority of students fell into the moderate category (28 students), with a small number of students in the low category (5 students) and the high category (3 students). This indicates that improvement occurred relatively evenly among most students. In contrast, in 11th grade, the distribution shows greater variation, with more students in the low category (17 students) than in the moderate (15 students) and high (4 students) categories. These findings suggest that the system's impact is not uniform across all students. This variation is likely influenced by factors such as initial ability, readiness to learn, and the ability to utilize the feedback provided.

4.7 Evaluation and Finalisation

The evaluation phase integrated all results from expert validation, pilot testing, and the main field trial. Overall, the results showed a moderate improvement in students' conceptual understanding based on N-Gain analysis, supported by a large effect size. These findings indicate that the system has a strong practical impact, although the relative level of improvement among students still varies. Beyond quantitative improvements, the system's primary value lies in its ability to provide reflective feedback on each student's essay response. This feedback not only presents a score but also specifically identifies the strengths and weaknesses of the response, thereby serving as a formative learning tool.

As shown in Figure 6, for partially correct answers, the system not only displays a score but also provides guidance on how to improve specific aspects such as reasoning and depth of analysis.

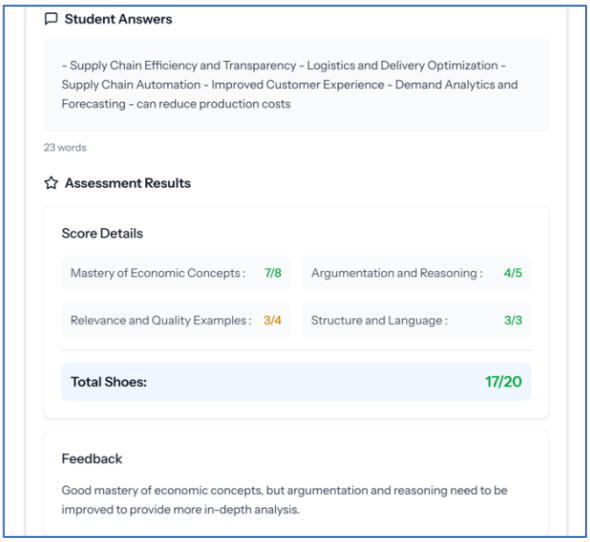

Figure 6: Feedback Partially Correct

Conversely, as shown in Figure 7, for correct and well-structured answers, the system reinforces the conceptual understanding that has been achieved.

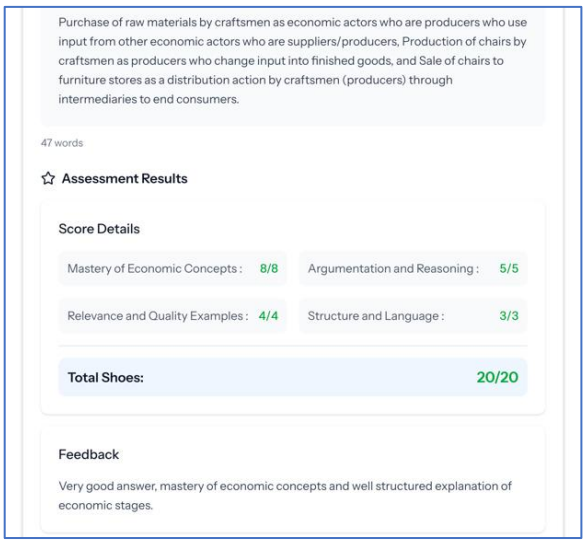

Figure 7: Feedback Completely Correct

These dual functions evaluative (scoring) and corrective (providing the right conceptual explanation) ensure that students not only recognise their errors but also learn the accurate concepts, thereby supporting formative and meaningful learning (Marton et al., 1984). The integration of reflective and corrective feedback is particularly important in economics education, where conceptual reasoning is essential. In the context of e-learning, specific and timely feedback is a crucial component in enhancing the quality of learning and student engagement.

Overall, this evaluation shows that EsyGrade functions effectively in line with its objectives: reducing teachers' workload, providing timely and structured feedback, and supporting the development of students' conceptual understanding. The system has been refined into a tool that is feasible, practical, and pedagogically sound for integrating AI-based automated essay grading into high school economics education and is ready for implementation in broader learning contexts.

5. Discussion

The results of this study indicate that the development of EsyGrade as a ChatGPT-based essay grading system can be successfully implemented and contributes positively to learning. Expert validation revealed a high level of validity, while preliminary trials indicated excellent levels of acceptance and practicality. In the main field trial phase, the analysis results showed a moderate improvement in students' conceptual understanding based on N-Gain, indicating an improvement in their abilities following the use of the system.

However, when combined with effect size results indicating a large category (Cohen's d > 1), a more comprehensive understanding emerges, indicating that the changes in learning outcomes are of substantial magnitude, even though the relative rate of improvement among students was uneven. This is reinforced by the N-Gain distribution analysis, which shows variation in improvement among individuals, particularly in 11th grade, where a larger proportion of students exhibited low improvement compared to 10th grade. Thus, the results of this study are more accurately interpreted as a moderate improvement on average, but with a strong overall impact.

These findings highlight that the primary contribution of the system lies not only in the automation of assessment, but in its ability to provide reflective feedback that supports formative learning. The feedback generated does not merely indicate right or wrong but also offers specific explanations regarding the strengths and weaknesses of students' answers. From the perspective of formative assessment theory, clear, specific, and timely feedback is a key factor in improving the quality of learning (Hattie and Timperley, 2007). Therefore, the integration of AI-based feedback in EsyGrade can be understood as a mechanism that fosters a continuous process of reflection and improvement in conceptual understanding.

These findings are consistent with previous research showing that the use of ChatGPT in essay grading can improve the quality of feedback and support students' independent learning (Mizumoto and Eguchi, 2023). Additionally, a study by Yavuz, Çelik and Yavaş Çelik, (2025) indicates that rubric-based grading assisted by ChatGPT has reliability comparable to that of human raters. This study expands on these findings by demonstrating that a similar approach can also be applied in the context of economics education, which has different conceptual and analytical characteristics compared to language learning.

In the Indonesian context, these findings are particularly relevant because assessment practices in schools still tend to be dominated by multiple-choice questions that emphasize rote memorization. This differs from trends in the broader literature, which is increasingly promoting the use of essay-based assessments and higher-order thinking as part of a more in-depth learning process (Maryani *et al.*, 2021; Zhao and Fu, 2025). Furthermore, teachers' heavy workloads and time constraints pose major obstacles to the widespread implementation of essay-based assessment. This contextual difference implies that automated grading systems like EsyGrade need to be designed adaptively, focusing not only on score automation but also on providing clear, structured, and user-friendly feedback. Therefore, features such as Bloom's taxonomy-based rubrics, flexibility in assessment customization, and a simple interface are crucial to ensuring alignment with the needs of users in Indonesia.

Nevertheless, several limitations should be noted. The system still relies on a stable internet connection, and the quality of the AI's responses sometimes requires verification by a teacher. Furthermore, this study is limited to a sample in the Surakarta region, so generalizing the findings requires caution.

Despite these limitations, this study offers valuable contributions both practically and theoretically. Practically, EsyGrade can serve as a tool to enhance the quality of formative assessment while helping to reduce teachers' workload in the essay grading process. In addition, this study contributes to e-learning research by demonstrating how AI-based assessment systems can be designed to support not only efficiency but also meaningful learning processes. Although this study was conducted in the context of economics education in Indonesia, the design principles developed such as the integration of rubrics based on Bloom's taxonomy, the use of customizable prompts, and the provision of reflective feedback have the potential to be applied to various other fields of learning. Theoretically, these findings suggest that AI-based essay grading systems can be conceptualised as tools not only to improve efficiency but also to support deep learning through structured reflective feedback. Thus, this study expands the e-learning literature by emphasizing the importance of integrating AI technology with pedagogical principles in the development of digital assessment systems.

6. Conclusion and Implications

This study aims to design, develop, and evaluate EsyGrade, a web-based automated essay grading system integrated with the ChatGPT API to assess conceptual understanding in economics education. Using a modified version of the Bennett, Borg and Gall, (1984) R&D model, the findings indicate that the developed system is feasible, user-friendly, and capable of supporting improvements in students' conceptual understanding. Expert validation indicates a high level of content validity, while preliminary trials confirm a strong level of user acceptance. Results from the main field trial show a moderate improvement based on N-Gain, supported by a large effect size, indicating a change in learning outcomes of considerable magnitude, although the level of improvement varies among students.

These findings indicate that the primary value of EsyGrade lies not only in the automation of the grading process, but also in its ability to provide structured, reflective feedback to support formative learning. Thus, the assessment is no longer viewed merely as a measurement tool, but as a means of supporting conceptual understanding and meaningful learning.

From a practical perspective, EsyGrade has the potential to serve as a tool to help teachers manage essay grading more efficiently without compromising the quality of feedback. This is particularly relevant in classes with large student populations and limited time for manual grading. From a policy perspective, these findings suggest that AI-based assessment systems can help bridge the gap between curriculum demands that emphasize higher-order thinking skills and the practical constraints faced by teachers.

More broadly, this study contributes to the development of e-learning by demonstrating how AI-based essay grading systems can be integrated with pedagogical principles such as rubric-based assessment and formative feedback. Although this system was developed in the context of economics education in Indonesia, the design principles employed such as the use of customizable prompts, alignment with rubrics, and a focus on feedback have the potential to be adapted across various fields and other educational contexts.

However, several limitations should be noted. The system still relies on a stable internet connection, and in some cases, the quality of the AI-generated feedback still requires verification by a teacher to ensure contextual appropriateness. Furthermore, this study was conducted within a limited geographical context, so the findings should be generalized with caution.

Future research could focus on several areas. First, studies with a broader and more diverse sample both in terms of geographic region and educational level are needed to enhance the generalizability of the findings. Second, further research could explore the long-term impact of AI-based feedback on the development of students' conceptual understanding. Third, integrating the system with Learning Management Systems (LMS) and evaluating its implementation over a longer period can provide a more comprehensive picture regarding the system's scalability and sustainability.

Acknowledgement

The author sincerely thanks the academic supervisors, family, and friends for their guidance and support during this study. Appreciation is also extended to the BIMA Master's Thesis Research Program of the Ministry of Higher Education, Science, and Technology of Indonesia (Fiscal Year 2025, Contract No. 105/C3/DT.05.00/PL/2025) for their valuable support.

AI Statement: No AI was used at any point in the research, writing, or creating of this paper.

Ethical Approvals: Ethical review and approval were not required for this study in accordance with the local legislation and institutional requirements. Participation in the questionnaire was entirely voluntary, and all respondents provided informed consent before participating. No personal identifying information was collected, and data were analysed anonymously to ensure participants' confidentiality.

Declaration of Conflict of Interest: All authors have read and approved the manuscript and take full responsibility for its content. The authors have no conflicts of interest regarding this research and its funding.

Data availability: Data will be made available upon request.

References

Alief, S., Irawati, I. and Mude, Muh.A. (2023) 'Impelementasi metode latent semantic analysis pada sistem informasi ujian online berbasis web', *Buletin Sistem Informasi dan Teknologi Islam*, 4(2), pp. 106–111. Available at: https://doi.org/10.33096/busiti.v4i2.1639.

Al-Obaydi, L.H., Pikhart, M. and Tawafak, R.M. (2023) 'Online Assessment in Language Teaching Environment through Essays, Oral Discussion, and Multiple-Choice Questions', *CALL-EJ*, 24(2), pp. 175–197. Available at: https://doi.org/callej.org/index.php/journal/article/view/7.

Anderson, L.W. and Krathwohl, D.R. (eds) (2001) '*A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*'. Complete ed. New York: Longman.

Arini, Y.A.Y, Dewi, T.A. and Wibawa, F.A. (2024) 'Pengembangan e-modul ekonomi berbasis kurikulum merdeka kompetensi 4C semester ganjil kelas X SMAN 2 Metro', *EDUNOMIA: Jurnal Ilmiah Pendidikan Ekonomi*, 5(1), pp. 70–81. Available at: https://doi.org/10.24127/edunomia.v5i1.6600.

Beldar, P. (2025) Exploring the efficacy of open-ended questions in theory-based subjects', *Journal of Engineering Education Transformations*, 38(4), pp. 117–127. Available at: https://doi.org/10.16920/jeet/2024/v38i4/25101.

Bennett, N., Borg, W.R. and Gall, M.D. (1984) 'Educational research: An Introduction', *British Journal of Educational Studies*, 32(3), p. 274. Available at: https://doi.org/10.2307/3121583.

Biggs, J.B. and Tang, C.S. (2011) '*Teaching for quality learning at university: What the student does.*' 4th edition. Maidenhead: Open University Press (Society for Research into Higher Education).

Bloom, B.S. (1956) '*Taxonomy of educational objectives: The classification of educational goals*'. Longmans, Green (Taxonomy of Educational Objectives: The Classification of Educational Goals). Available at: https://books.google.co.id/books?id=1WjuAAAAMAAJ.

Cohen, J. (1988) '*Statistical power analysis for the behavioral sciences*'. 2nd ed. Hillsdale, N.J: L. Erlbaum Associates. Available at: https://doi.org/https://doi.org/10.4324/9780203771587.

Elliott, C. and Balasubramanyam, V. (2016) 'Assessing students: Real-world analyses underpinned by economic theory', *Cogent Economics & Finance*. Edited by S. Cook, 4(1), p. 1151171. Available at: https://doi.org/10.1080/23322039.2016.1151171.

Foung, D., Lin, L. and Chen, J. (2024) 'Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices', *Computers and Education: Artificial Intelligence*, 6, p. 100250. Available at: https://doi.org/10.1016/j.caeai.2024.100250.

Gandolfi, A. (2025) 'GPT-4 in education: Evaluating aptness, reliability, and loss of coherence in solving calculus problems and grading submissions', *International Journal of Artificial Intelligence in Education*, 35(1), pp. 367–397. Available at: https://doi.org/10.1007/s40593-024-00403-3.

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

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

Johnson, M. and Coleman, V. (2025) 'Teaching in uncertain times: Exploring links between the pandemic, assessment workload, and teacher wellbeing in England', *Research in Education*, 121(1), pp. 69–92. Available at: https://doi.org/10.1177/00345237231195270.

Kamalia, P.U.K., Yonisa; P. R., Fiky, A., Ghofur, M.A., Ginanjar, A.E. (2023) ' Pengembangan asesmen digital berbasis hots pada kurikulum merdeka bagi guru ekonomi, *SELAPARANG: Jurnal Pengabdian Masyarakat Berkemajuan*, (Vol 7, No 4 (2023): December), pp. 2886–2893. Available at: https://doi.org/https://journal.ummat.ac.id/index.php/jpmb/article/view/17046/8298.

Kinanti, N.L. & Qoiriah, A. (2020) ' Sistem penilaian otomatis jawaban esai Bahasa Indonesia berdasarkan kemiripan kalimat menggunakan syntactic-semantic similarity 02. Available at: https://ejournal.unesa.ac.id/index.php/jinacs/article/view/37555/33283.

Latif, E. & Zhai, X. (2024) ' Fine-tuning ChatGPT for automatic scoring', *Computers and Education: Artificial Intelligence*, 6, p. 100210. Available at: https://doi.org/10.1016/j.caeai.2024.100210.

Lee, G. G. *et al.* (2024) ' Applying large language models and chain-of-thought for automatic scoring', *Computers and Education: Artificial Intelligence*, 6, p. 100213. Available at: https://doi.org/10.1016/j.caeai.2024.100213.

Marton, F. & Hounsell, D. (eds) (1984) '*The experience of learning*'. Edinburgh: Scottish Academic Press. Available at: https://doi.org/doi.org/10.1111/b.9780631211860.1998.00025.x.

Maryani, I. *et al.* (2021) ' HOTs multiple choice and essay questions: A validated instrument to measure higher-order thinking skills of prospective teachers', *Turkish Journal of Science Education*, p. 4. Available at: https://doi.org/10.36681/tused.2021.97.

Mendonça, P.C., Quintal, F. & Mendonça, F. (2025) ' Evaluating LLMs for automated scoring in formative assessments', *Applied Sciences*, 15(5), p. 2787. Available at: https://doi.org/10.3390/app15052787.

Mizumoto, A. & Eguchi, M. (2023) ' Exploring the potential of using an AI language model for automated essay scoring, *Research Methods in Applied Linguistics*, 2(2), p. 100050. Available at: https://doi.org/10.1016/j.rmal.2023.100050.

Van Wyk, M.M. (2015) ' Teaching Economics, *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, pp. 83–88. Available at: https://doi.org/10.1016/B978-0-08-097086-8.92072-5.

Morris, C. *et al.* (2026) ' The components and implications of teacher workload: a review, *Educational Research*, pp. 1–22. Available at: https://doi.org/10.1080/00131881.2026.2629284.

Morris, R. *et al.* (2024) Can a code-based approach to marking and feedback reduce teachers' workload? An evaluation of the FLASH marking intervention, *Oxford Review of Education*, 50(4), pp. 552–569. Available at: https://doi.org/10.1080/03054985.2023.2258779.

OECD (2023) *PISA 2022* '*Results (Volume I): The State of Learning and Equity in Education*. OECD Publishing (PISA). Available at: https://doi.org/10.1787/53f23881-en.

Pack, A., Barrett, A. & Escalante, J. (2024) ' Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability, *Computers and Education: Artificial Intelligence*, 6, p. 100234. Available at: https://doi.org/10.1016/j.caeai.2024.100234.

Pasaribu, N.G., Budiman, G. & Indrarini, D.I. (2024) ' Auto evaluation for essay zssessment using a 1D convolutional neural network, *IEEE Access*, 12, pp. 188217–188230. Available at: https://doi.org/10.1109/ACCESS.2024.3515837.

Permendikdasmen (2025) '*Peraturan menteri pendidikan dasar dan menengah Nomor 11 Tahun 2025 tentang Pemenuhan Beban Kerja Guru*. Available at: https://peraturan.bpk.go.id/Details/322487/permendikdasmen-no-11-tahun-2025 (Accessed: 25 November 2025).

Petersen, A., Craig, M. & Denny, P. '*ITiCSE '16: Innovation and Technology in Computer Science Education Conference 2016*, Arequipa Peru: ACM, pp. 252–253. Available at: https://doi.org/10.1145/2899415.2925503.

Plomp, T. & Nieveen, N. (2013) '*Educational design research: An introduction*. Available at: https://slo.nl/publish/pages/2904/educational-design-research-part-a.pdf (Accessed: 23 November 2025).

Poole, F.J. and Coss, M. (2023) ' Can ChatGPT reliably and accurately apply a rubric to L2 writing assessments? The devil is in the prompt(s). EdArXiv. Available at: https://doi.org/10.35542/osf.io/3r2zb.

Pratama, R.D. & Sangka, K.B. (2025) ' Assessing economic essays with ChatGPT: Systematic review and preliminary design', *International Conference on Teaching and Learning*, 1(1), pp. 337–345. Available at: https://doi.org/conference.ut.ac.id/index.php/ictl/article/view/1753/2343.

Rafi, R.R., Ramadhani, R.A. and Sanjaya, A. (2025) ' Pemanfaatan algoritma cosine similarity untuk mengkoreksi ujian esai", *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, 10(3), pp. 2242–2254. Available at: https://doi.org/10.29100/jipi.v10i3.8058.

Rittle-Johnson, B., Schneider, M. and Star, J.R. (2015) 'Not a one-way street: Bidirectional relations between procedural and conceptual knowledge of Mathematics', *Educational Psychology Review*, 27(4), pp. 587–597. Available at: https://doi.org/10.1007/s10648-015-9302-x.

Runtuwene, J.P.A. and Tangkawarow, I.R.H.T. (2020) 'The quality classification of professional teacher using fuzzy-analytical hierarchy process', *IOP Conference Series: Materials Science and Engineering*, 830(2), p. 022099. Available at: https://doi.org/10.1088/1757-899X/830/2/022099.

Sabon, S.S. (2020) 'Problematik pemenuhan beban kerja guru dan alternatif pemenuhannya (Studi kasus di kota depok provinsi jawa barat)', *Jurnal Penelitian Kebijakan Pendidikan*, 13(1), pp. 27–44. Available at: https://doi.org/10.24832/jpkp.v13i1.345.

Smerdon, D. (2024) 'AI in essay-based assessment: Student adoption, usage, and performance', *Computers and Education: Artificial Intelligence*, 7, p. 100288. Available at: https://doi.org/10.1016/j.caeai.2024.100288.

Steiss, J. *et al.* (2024) 'Comparing the quality of human and ChatGPT feedback of students' writing', *Learning and Instruction*, 91, p. 101894. Available at: https://doi.org/10.1016/j.learninstruc.2024.101894.

Tang, X. *et al.* (2024) 'Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments', *Heliyon*, 10(14), p. e34262. Available at: https://doi.org/10.1016/j.heliyon.2024.e34262.

Walstad, W.B. (2006) 'Testing for depth of understanding in Economics using essay questions', *The Journal of Economic Education*, 37(1), pp. 38–47. Available at: https://doi.org/10.3200/JECE.37.1.38-47.

Yavuz, F., Çelik, Ö. and Yavaş Çelik, G. (2025) 'Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments', *British Journal of Educational Technology*, 56(1), pp. 150–166. Available at: https://doi.org/10.1111/bjet.13494.

Zhao, N. and Fu, Q. (2025) 'Under the framework of deep learning cognitivetheory and Bloom's Taxonomy: Investigating the Role of artificial intelligence in fostering higher-order thinking in engineering education', *2025 7th International Conference on Computer Science and Technologies in Education (CSTE)*. *2025 7th International Conference on Computer Science and Technologies in Education (CSTE)*, Wuhan, China: IEEE, pp. 892–896. Available at: https://doi.org/10.1109/CSTE64638.2025.11092073.

Appendix

This appendix presents the research instruments, including expert validation instruments, preliminary testing instruments, and pre-test and post-test questions to measure students' conceptual understanding.

A1. Instruments for educational assessment experts

	Indicator & Items
	A. Content Relevance & Validity
1.	The questions entered into the system can be tailored to the intended learning objectives for economics
2.	The automatic grading rubric reflects the competency indicators being assessed
3.	Test instructions, rubrics, and prompts are presented clearly and are easy for students and teachers to understand
	B. Reliability & Fairness in Assessment
4.	Automated essay grading is free from bias (student-neutral)
5.	The system assigns consistent scores to essay answers of the same quality
6.	The grading scale and weighting used are in accordance with educational assessment standards
	C. Alignment with Educational Assessment Principles
7.	The system supports the assessment of higher-order thinking skills (analysis, evaluation, synthesis)
8.	The system provides feedback in accordance with the principles of formative assessment

	Indicator & Items
9.	The system allows teachers to customize assessment rubrics and prompts as needed
	D. Clarity in Reporting Results
10.	The assessment reports are easy for teachers to read and understand
11.	Assessment results can be communicated to students in a transparent manner
12.	The system provides analysis of assessment results (such as score distributions and averages) to help teachers evaluate students
	E. Data Ethics & Data Privacy
13.	Student assessment data is stored and managed with due regard for confidentiality
14.	The system promotes student discipline and honesty (e.g., time limits)
	F. Alignment with the Curriculum
15	The system is aligned with the applicable curriculum and competency standards

A2. Instruments for educational technology experts

	Indicator & Items
	A. Functional Suitability
1.	The system's core features (question set creation, automated grading, and results reports) meet teachers' needs
2.	The automatic assessment results displayed according to the planned specifications
	B. Performance Efficiency
3.	The system's response time (for example, when processing essays) is fast and consistent
4.	The system can handle large volumes of students or questions without a significant drop in performance
	C. Compatibility
5.	The system can create custom rubrics and assessment prompts
	D. Usability
6.	The system's user interface is easy for teachers to understand and use
7.	The menu navigation (create question sets, rubrics, prompts, view results) is easy to follow
8.	The system provides sufficient guidance and help for new users
	E. Reliability
9.	The system rarely experiences errors or crashes during use
10.	User data (questions, student answers, assessment results) remains securely stored even in the event of an outage
	F. Security
11.	The system includes user data security features (secure login, protection of assessment results)
	G. Maintainability
12.	The system is easy to update (feature updates, bug fixes) without disrupting users
13.	The system's technical documentation is sufficient to support further development
	H. Portability
14.	The system can be easily moved or installed on another server, device, or operating system if necessary
15.	Assessment results can be exported as PDF or CSV

A3. Preliminary Testing Instrument for Students

	Indicator & Items
A. Ease of Access & System Navigation	
1.	It's easy for me to log in and access this system
2.	I found the menu and interface of this system easy to understand
B. Clarity of Instructions & Display	
3.	The instructions for answering the questions provided by this system are clear to me
4.	The layout of the questions and the scoring criteria in this system are engaging and easy to understand
C. Fairness & Transparency in Evaluation	
5.	The score I received feels fair based on my answers
6.	This system clearly displays the grading criteria, so I know the basis for the grading
D. Speed & Accuracy of Feedback	
7.	This system provides assessment results immediately after I finish answering the questions
8.	The feedback I received helped me understand my mistakes
E. Privacy & Security	
9.	I feel secure about my personal data when using this system
10.	My assessment results are securely stored and not misused
F. Support for Learning	
11.	This system helps me understand the economics material covered on the exam
12.	This system makes it easier for me to prepare for the next question
13.	This system motivates me to improve my academic performance
G. Overall Experience	
14.	Overall, this system is easy to use for solving problems
15.	I am satisfied with this automated grading system as a tool for evaluating the learning process

A4. Pre-test and post-test questions

Questions class X

Phase 1 (pre-test)

1. Why is production the core of a society's economic activities? (Level C4)
2. What are the four types of factors of production needed to produce a good or service? (Level C1)
3. What does distribution mean in economic activities? (Level C2)
4. A farmer grows rice. The rice is then sold to a rice merchant. The rice merchant sells it to small shops. Consumers buy rice at the shops to cook. Outline the sequence of economic activities! (Level C4)
5. Why can lifestyle influence a person's level of consumption? (Level C4)

Phase 2 (post-test)

1. Give three examples of consumption activities carried out by households on a daily basis! (Level C1)
2. Explain the difference between production and consumption activities! (Level C2)
3. What is the role of distribution institutions in delivering goods and services to consumers? (Level C2)
4. Look at the following illustration: A rattan craftsman buys raw materials from a supplier, produces chairs, and then sells them to a furniture store. Order the stages of these economic activities based on the roles of the economic actors! (Level C4)
5. Why can technological developments affect how producers distribute their goods? (Level C4)

Questions class XI

Phase 1 (pre-test)

1. Name three types of business entities in Indonesia! (Level C1)
2. Briefly explain the fundamental differences between state-owned enterprises (SOEs) and private enterprises! (Level C2)
3. A cooperative has a Surplus (SHU) of Rp100,000,000. Of this amount, 25% is distributed to members based on their deposits. If Member A's total deposit is Rp2,000,000 out of the total deposits of all members amounting to Rp20,000,000, how much of the Surplus will Member A receive? (Level C3)
4. Consider the following scenario: A new retail company is opening branches in five cities. Determine the planning and organizing steps the company should take! (Level C3)
5. Analyze the three main areas of management within a business entity (e.g., marketing, finance, and production), and explain how they are related! (Level C4)

Phase 2 (post-test)

1. Give two examples of cooperative-style business entities in Indonesia! (Level C1)
2. Briefly explain the difference between top management and middle management! (Level C2)
3. A cooperative receives a net surplus of Rp50,000,000. Thirty percent is distributed based on savings. If Member B's savings are Rp1,500,000 out of total savings of Rp15,000,000, calculate the net surplus received by Member B! (Level C3)
4. Given the following scenario: A logistics company is opening a new branch. Identify the "Actuating" and "Controlling" steps that need to be taken to ensure smooth operations! (Level C3)
5. Analyze the interrelationship between production, marketing, and finance in supporting the success of a startup! (Level C4)

