Argüelles-Álvarez, I. (2025). Exploring Peer Assessment in Academic Writing With ChatGPT: Insights From the Legitimation Code Theory in Higher Education. *Teaching English with Technology*, *25*(1), 41–59.
https://doi.org/10.56297/vaca6841/OZFH1876/PMRR6657

ISSUES

Exploring Peer Assessment in Academic Writing With ChatGPT: Insights From the Legitimation Code Theory in Higher Education

Irina Argüelles-Álvarez[1]

[1] Department of Linguistics Applied to Science and Technology, Polytechnic University of Madrid

Keywords: Higher education, peer assessment, comparative judgement, Legitimation Code Theory, ChatGPT

https://doi.org/10.56297/vaca6841/OZFH1876/PMRR6657

Teaching English with Technology

Vol. 25, Issue 1, 2025

This study explores the use of ChatGPT in an Academic English course at a Spanish Polytechnic university, focusing on critical thinking, writing skills, and academic integrity. It proposes a workshop demonstrating how Artificial Intelligence tools can be integrated into the classroom routine. Involving 101 engineering students, ChatGPT was used to write cover emails for a professional communication assignment. Students used comparative judgment to rank their peers' emails from high to low quality. Their evaluations were analysed using the Legitimation Code Theory (LCT), with a focus on the Specialization dimension, examining how knowledge and personal attributes are valued within this specific context. The findings highlight the effectiveness of peer assessment in writing tasks where the use of ChatGPT is explicitly permitted as a resource. Students demonstrated the ability to differentiate quality in AI-assisted writing and provided structured evaluations of their peers' work. The study emphasizes the need for balancing AI use with fostering originality and critical thinking. LCT provides a useful framework for understanding students' perceptions of quality in AI-mediated writing tasks, offering insights into the integration of AI tools in educational settings.

1. Introduction

This study explores the integration of ChatGPT into an Academic English course for higher education in Spain. Traditionally, students were asked to complete individual or collaborative writing assignments at home, as these activities require considerable time and effort. However, in recent years, educators worldwide have observed that many assignments are suspiciously well-written, often indicating the involvement of Artificial Intelligence (AI). Concerns regarding the use of tools like ChatGPT centre on the potential for students to over-rely on AI, which may hinder the development of critical thinking skills (Çobanoğulları, 2024), facilitate academic dishonesty (Levine et al., 2024; Sajawal & Kittur, 2024), and evade detection by plagiarism software (Alexander et al., 2023). This growing trend presents challenges for educators, who must devise innovative strategies to ensure students engage in authentic learning processes, and for students, as over-reliance on AI can impede the development of essential writing skills.

To address this issue, some educators have implemented in-class writing sessions. While this approach ensures authenticity, it has drawbacks, such as consuming valuable class time for what is often a personal and introspective activity. Additionally, it deprives students of legitimate aids, including ChatGPT, online dictionaries, corpora, and grammar-checking tools, which can be especially beneficial when writing in a foreign language.

Other educators and researchers argue that AI is here to stay (Atlas, 2023; Barrot, 2023; LatinCALL Conference, 2024). Given the growing prevalence of AI tools in professional contexts, such as report writing, it has become essential for university students to integrate these technologies into their academic routines while critically reflecting on their appropriate usage. From this perspective, this paper explores the incorporation of such a tool into a higher education course that, like many others, was not initially designed to accommodate this technological evolution. For the purposes of this study, one of the course's writing tasks was adapted to include the use of ChatGPT. This modification aimed not only to integrate the tool into the course routine and evaluate its outcomes and value for both the course and the specific task but also to prompt students to engage in deeper reflection on the writing process and assess the quality of outputs generated with AI assistance.

The task of writing a cover letter holds particular significance in this study due to its dual demand for mastery of formal writing aspects—such as grammar, vocabulary, genre-specific moves, phraseology, and appropriate tone—and personal engagement. Beyond demonstrating their mastery of written language, students must also present themselves convincingly in a way that effectively positions them as ideal candidates for a specific job. Moreover, this genre inherently requires a high degree of personal involvement, as it necessitates the ability to reflect on one's qualifications, experiences, and aspirations and articulate them in a way that attracts a potential employer's attention. The added complexity lies in balancing formal writing conventions with persuasive and personalized communication, making it a uniquely challenging exercise that tests both their more technical writing skills and their ability to connect authentically with their audience.

The peer assessment methodology employed aimed to add a distinct competitive dimension to the writing process. Students were tasked with evaluating and ranking their peers' cover letters from best to worst, introducing an opportunity for critical observation and reflection. To achieve a higher ranking and, consequently, a better grade, students' cover emails needed to stand out, highlighting the importance of originality, authenticity, and personal engagement. This process not only encouraged students to critically assess their own writing and that of their peers but also challenged the over-reliance on ChatGPT to complete the task. By fostering active reflection through peer evaluation and competition, the study seeks to explore

how AI tools like ChatGPT can be meaningfully integrated into language learning and professional communication tasks, while maintaining the personal and creative aspects crucial to such genres.

For the purposes of this study, peer assessment methodology is informed by the Legitimation Code Theory (LCT), particularly its Specialization dimension, alongside comparative judgment. This theoretical framework provides a lens to analyse how students negotiate and balance the formal aspects and personal dimensions of cover email writing—a genre that inherently demands both precision and authentic self-representation. The integration of ChatGPT further complicates this balance, offering opportunities for reflection on the role of AI in professional communication.

Through the process of ranking and evaluating their peers' cover emails, students engage in legitimation practices, revealing their perceptions of what constitutes quality writing within this context. By systematically categorizing and interpreting peer feedback, the study seeks to illuminate how students enact these legitimation practices, particularly as they relate to originality, authenticity, and the appropriate use of AI tools. This approach not only advances our understanding of how ChatGPT can be integrated into language learning but also contributes to broader discussions on how specialization codes can inform the design and assessment of AI-mediated writing tasks in higher education.

2. Theoretical framework

2.1. ChatGPT in language learning

Generative Artificial Intelligence (GenAI) has significantly influenced language teaching and learning. At its core, AI enables machines to independently process vast external data and achieve outcomes, imitating aspects of human cognition (Callanan, 2024). While it offers immense potential, AI also raises challenges, such as lesson planning, homework integrity, and authentic assessment (Rangelov, 2024). Academic integrity is particularly pressing, necessitating AI integration into teaching.

The integration of Generative Artificial Intelligence (GenAI) tools like ChatGPT into language learning and higher education has gathered significant attention due to their potential to transform traditional educational practices. ChatGPT, developed by OpenAI, is a powerful language model capable of generating coherent and contextually appropriate text, making it a valuable tool for personalized language instruction and the creation of authentic language materials. Baskara and Mukarto (2023) emphasize ChatGPT's potential for personalized language instruction and the generation of authentic materials but caution against ethical concerns, such as bias and over-reliance on AI, which could undermine critical thinking and originality. Its ability to provide real-time feedback and tailored content aligns with the principles of individualized learning, enhancing students'

engagement and comprehension (Atlas, 2023). However, the use of ChatGPT also raises ethical concerns, such as the potential for bias in generated content and the risk of over-reliance on AI, which could hinder the development of critical thinking skills (Çobanoğulları, 2024).

The benefits of ChatGPT in educational settings are multifaceted. It offers personalized support, helping students improve their writing skills through immediate feedback on grammar, vocabulary, and structure (Levine et al., 2024). Additionally, ChatGPT can serve as a supplemental learning tool, providing interactive and engaging practice opportunities that enhance the overall learning experience (Çobanoğulları, 2024). Despite these advantages, challenges such as the potential for generating biased or inappropriate content and the limitations in handling complex or abstract ideas must be addressed (Deng & Lin, 2023). Furthermore, the ethical implications of using AI in education, including issues related to academic integrity and the authenticity of student work, necessitate careful consideration and responsible use (Barrot, 2023).

2.2. Comparative Judgement

Comparative Judgment (CJ) is an innovative assessment method in which evaluators compare two pieces of work and decide which is better, rather than assigning absolute scores based on predefined criteria. Rooted in holistic evaluation, CJ leverages human expertise to generate reliable rankings without requiring detailed rubrics or extensive training. By aggregating multiple judgments, CJ consistently achieves high reliability and mitigates individual biases. It minimizes rater bias through relative comparisons and accommodates diverse perspectives by synthesizing multiple judgments via statistical modelling (Thwaites & Paquot, 2024). Unlike single-marker systems, CJ ensures that scripts are reviewed multiple times by different judges, enhancing the fairness and robustness of the evaluation process. Studies have reported reliability indices (SSR) ranging from .73 to .99, surpassing those of traditional rubric-based methods for essay marking (Pollitt, 2012; Steedle & Ferrara, 2016).

In addition to its validity, CJ has demonstrated exceptional effectiveness in evaluating constructs that are both complex and creative. These include areas like creativity, writing quality, problem-solving, and conceptual understanding, which are difficult to assess with conventional rubrics (Thwaites & Paquot, 2024). By utilizing holistic judgments, CJ enables experts to apply their intuitive understanding of quality without being restricted by rigid rubrics. This approach captures construct validity more effectively, as it directly aligns judgments with the constructs being evaluated, enhancing their inherent validity (Pollitt & Crisp, 2004).

Its versatility also extends to evaluating diverse outputs, such as design portfolios and second-language writing, where its adaptability promotes reflective learning and detailed evaluations (Jones, 2015; Jones & Davies, 2024). By addressing the limitations of traditional rubrics, CJ offers a robust and innovative method for assessing outputs that require a blend of technical and creative insights.

2.3. Peer assessment

Peer assessment has proved to enhance writing skills, critical thinking, and self-awareness by fostering active engagement and collaborative learning (Xiao & Lucking, 2008). In the context of peer assessment, CJ enhances the learning experience by encouraging critical evaluation and reflection. Students engaging in comparative judgment (CJ) are exposed to diverse examples of work, helping them internalize standards of quality and develop higher-order thinking skills. Peer assessments using CJ have demonstrated measurable benefits in fostering evaluative skills, reflection, and understanding (Hendry et al., 2017). Peer-based CJ improves self-assessment and critical thinking, benefiting both the feedback recipient and provider. Even novices apply comparative judgments reliably when guided appropriately (Jones, 2015). Repeated evidence suggests that learning benefits occur even when peers lack expertise (van Daal et al., 2023).

This method is particularly valuable in educational settings due to its scalability and ability to provide meaningful feedback, making it a compelling alternative to criteria-based approaches (Jones, 2015; van Daal et al., 2023). Hendry et al. (2017) demonstrate high inter-rater reliability in peer feedback, especially when judgments are frequent and distributed. Validity-adaptive judgment allows students to identify and internalize quality, leading to improved outcomes in open-ended assessments. Furthermore, the aggregation of diverse expert judgments produces a consensus that is robust against individual variability in bias or interpretation. Aggregation eliminates leniency and severity bias across experts (Bramley, 2007).

In conclusion, CJ is effective for peer assessment in diverse contexts and age groups, especially for assessing complex or open-ended tasks. Peer assessment using CJ could enhance scalability in contexts like online education or MOOCs (Jones, 2015; van Daal et al., 2023).

2.4. Legitimation Code Theory

Maton's Legitimation Code Theory (LCT) is a sociological framework developed to analyse the underlying principles that legitimize knowledge and practices in different contexts. It offers tools for understanding the "rules of the game" in different fields—what is considered legitimate or valuable and why (Maton, 2014; Maton et al., 2018).

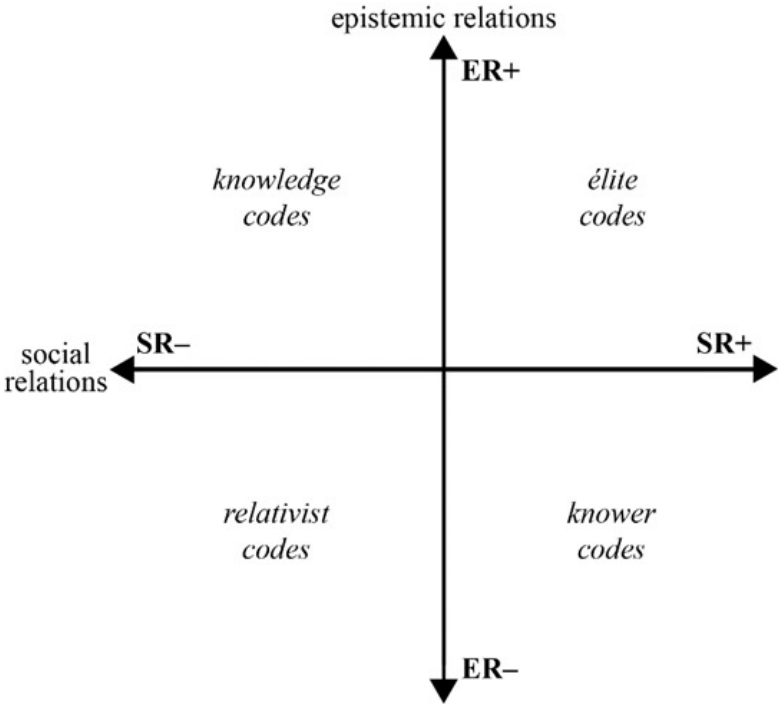

Figure 1. The specialization plane (Maton, 2014, p. 30)

Specialization (Maton & Chen, 2019) refers to how knowledge and practices are distinguished and valued. It involves two dimensions: epistemic relations (ER), the degree to which legitimacy is tied to specialized knowledge, skills, and expertise (e.g., technical rigor, theoretical frameworks); and social relations (SR), the degree to which legitimacy depends on the dispositions, experiences, or personal traits of the knower (e.g., creativity, lived experiences).

These dimensions create four codes of legitimation:

- Knowledge code (ER+, SR–): Legitimacy is based on strong knowledge and technical expertise (e.g., STEM fields).
- Knower code (ER–, SR+): Legitimacy is based on the characteristics of the individual (e.g., creativity, identity in arts).
- Elite code (ER+, SR+): Legitimacy combines both strong knowledge and personal characteristics (e.g., leadership roles requiring both expertise and charisma).
- Relativist code (ER–, SR–): Legitimacy is weak in both knowledge and personal traits (e.g., unstructured or ad hoc approaches).

LCT is highly versatile and has been largely applied across various domains such as education including assessment practices (Hindhede & Højbjerg, 2024; Morton & Nashaat-Sobhy, 2024; Nashaat-Sobhy, 2022); professions,

understanding how different fields (medicine, law, arts) define legitimacy and professional standards; research, analysing how academic disciplines prioritize certain kinds of knowledge or ways of knowing and culture and identity, exploring how communities or groups construct legitimacy based on shared knowledge or identity (Maton, 2014; Maton et al., 2018).

By examining the legitimation codes, LCT helps unpack dynamics, providing practical insights into how to address inequities or improve practices. Specialization codes in LCT analyse how legitimacy in a field or task is constructed based on: epistemic relations (ER) (the role of knowledge, skills, and technical expertise) and social relations (SR) (the role of personal attributes, dispositions, or creative expression).

3. Method

3.1. Research questions

Based on the objectives of this study and the insights from the theoretical framework, the following research questions have been formulated:

1. How effective is peer assessment in evaluating the quality of AI-assisted writing tasks, and how does it foster students' ability to critically assess writing beyond technical correctness?
2. How does the use of ChatGPT in writing tasks influence students' perceptions of quality, particularly in balancing technical proficiency (epistemic relations) and personal engagement (social relations)?
3. To what extent does engagement in comparative peer assessment enhance students' awareness of originality and critical thinking, reducing over-reliance on AI-generated content?
4. How can the Legitimation Code Theory (LCT) help researchers make visible and explain the students' perceptions of writing quality in AI-mediated tasks?

3.2. Setting and participants

This study involves 101 students from a Polytechnic University of Madrid studying degrees in Telecommunications. The participants are enrolled in one of the following programs: Telecommunication Systems Engineering, Sound and Image Engineering, Communications Electronics Engineering, Telematics Engineering, or a double degree in Electronics and Telematics. All participants are taking the English for Professional and Academic Communication course during the Autumn semester of the 2024-2025 academic year.

This compulsory course, valued at 6 European Credit Transfer and Accumulation System (ECTS) credits, is scheduled for the final year of their degree programs (semestres 7 and 8). To enrol, students must demonstrate a B2 level of English proficiency as defined by the Common European Framework of Reference for Languages (CEFR) through certification by an official entity. Exceptionally, students with a B1 certification from recognized examiners may enroll provisionally but must present a B2 certificate before the final exam (ordinary call) to qualify for course evaluation.

The course follows a blended learning methodology supported by the Virtual Learning Environment (VLE) Moodle. Students are required to attend four hours of class per week, where both theoretical and practical aspects of communication are covered. During these sessions, teachers introduce theoretical concepts, while students are encouraged to interact and take on a participative role. For homework, students complete individual or collaborative exercises and tasks related to the in-class content, utilizing the diverse resources available on the platform.

3.3. The task on Moodle

In the sixth week of the course, as part of the continuous assessment for the first module, Job Seeking and Job Interviews, students must complete a final task worth 5% of the course grade. This task involves writing a 250-word covering email in response to a mock job offer. The job description is tailored to appeal to the students and is adapted to their current undergraduate status, eliminating the need to fabricate work experience. The Moodle "Workshop" tool is configured to provide students access to the job offer, upload their covering email, and assess their peers' submissions within a week.

For this task, students are encouraged to use ChatGPT to assist in writing the covering email. The reason for requiring students to use ChatGPT is to ensure fairness and transparency: if its use is openly permitted by the instructor, all students have the same opportunity to benefit from it. In the past, some students have used ChatGPT as a writing aid, while others, either due to fear of being penalized or unfamiliarity with the tool, have not. This has created dilemmas for instructors when evaluating submissions, as they must consider how to fairly assess work that has clearly benefited from AI assistance versus work that has not. Moreover, as companies are increasingly encouraging their employees to use tools like ChatGPT for various purposes, it is considered essential to incorporate such technologies transparently into the course. This approach aims to prepare students for professional environments while fostering equitable opportunities to engage with emerging tools.

Prior to this task, students have received explicit instruction in class on writing a covering email, including discussions on layout, content, organization, style, and register. However, no specific guidance is given on

how extensively or in what manner they should use AI assistance. Students are informed in advance that they will need to evaluate five emails written by their peers and that their own email will be assessed in the same way. This encourages them to strategically decide how to use ChatGPT most effectively. Although not central to this study, students completed surveys before and after the workshop to share their predictions and perceptions about using ChatGPT (Di Sarno-García & Argüelles-Álvarez, under review).

3.4. The peer assessment

Rather than assigning traditional marks or grades, students were asked to rank a set of five peers' emails within their randomly assigned Moodle groups, using a scale from 10 (the highest mark) to 6 (the lowest mark). This ranking method has been used in previous studies, where graders evaluate the quality of a small set of responses (Attali, 2014; Waters et al., 2015). It employs the comparative judgment approach based on subjective evaluations, relative rankings within a limited subset, and preference-based judgments.

This method was estimated most appropriate for two primary reasons. First, the goal of the peer assessment is not merely to rank essays but, more importantly, to foster meaningful feedback and reflection among students. While feedback can be included in standard Comparative Judgment (CJ) practices, it is not intrinsic to CJ's design, which primarily focuses on efficient and reliable ranking, often in large-scale assessments. In this case, students were required to provide written justifications for the rankings they assigned, offering detailed feedback to their peers. Second, although Moodle lacks built-in tools for peer ranking, the scale from 10 to 6 was adapted to serve as the grade for the covering email (worth 20% of the final mark). The remaining 80% of the grade was based on the quality of the students' assessments and feedback provided to peers. Moodle's algorithm adjusts the final scores based on the agreement between a student's assigned rankings and those given by or received from others in the group, ensuring fairness and consistency in the evaluation process.

Although students learn how to write a covering email in class, focusing on aspects such as writing, content, and organization, they do not receive specific instruction on how to provide feedback to their peers. Instead, they are instructed to justify their rankings by explaining why they consider certain covering emails better or worse than others. As is typical in this type of assessment, students do not use a rubric or guide to inform their evaluations, as outlined in the earlier section on Comparative Judgment.

This approach allows students the freedom to prioritize the aspects they find most relevant during evaluation encouraging evaluators to articulate their reasoning. Consequently, it provides valuable insights for both the writers and the evaluators themselves. Furthermore, the task of writing covering

emails involves balancing technical writing skills (ER) and persuasive communication (SR), and peer assessment comments offer a rich source of data for analysing how students evaluate these dimensions.

3.5. The research methodology

Fifteen cover emails, each with feedback from five student evaluators (totalling 75 comments), were selected to represent "high-quality standards," "mid-quality standards," and "low-quality standards." This categorization was based on consistent agreement among evaluators, with most raters assigning high (10), mid (8), or low (6) rankings to the cover emails. Statistical analysis confirmed interrater reliability, ensuring the robustness of the ranking process.

Peer comments were categorized and analysed to assess their validity and to identify the dominant relational focus with the aid of ChatGPT. The peer-assessment system, operating within a competitive environment, would potentially incentivise students to pursue "elite codes" by excelling both technically (epistemic relations) and creatively (social relations). A translation device was developed to classify comments based on their emphasis on skills and knowledge (ER) or on creativity and personal involvement (SR), providing insights into the interplay of these dimensions in the evaluation process.

3.6. The translation device

Using LCT's specialization codes, peer comments were analysed to determine whether they emphasized epistemic relations or social relations. For epistemic relations (ER), a distinction was made between two key aspects: references to technical knowledge (e.g., domain-specific expertise in telecommunications engineering, such as understanding systems, projects, or field-specific skills) and writing conventions (e.g., grammar, structure, vocabulary, or adherence to rhetorical moves).

A rubric or set of indicators (translation device in LCT) was developed for identifying comments aligned with each code. A translation device is a conceptual tool used to link theoretical concepts to empirical data. It offers a systematic approach to analysing and interpreting the underlying principles that structure knowledge or practices within a specific context. Translation devices bridge theory and practice by enabling to application of abstract LCT concepts, such as epistemic relations (ER) and social relations (SR), to tangible data (Maton & Chen, 2015).

In this case, the indicators established for each code are derived from the following general ideas, which are informed by the students' data:

- ER+: References to either technical knowledge or writing conventions that reflect mastery of required skills.

- SR+: References to creativity, originality, or authenticity.
- ER– and SR–: Absence of meaningful feedback or vague critiques.

In alignment with LCT guidelines for developing a translation device, the examples provided are drawn directly from the corpus of students' comments.

Knowledge Code	Description	Example
ER+, SR+ Elite code	Emphasis on both strong technical knowledge or writing conventions and a compelling personal touch or creative element.	"Mario balances technical skills with soft skills and genuine motivation. His tone is inviting, and his emphasis on growth, learning, and real-world impact makes this a highly memorable email."
ER+, SR– Knowledge code	Focus on either technical knowledge (e.g., domain-specific expertise) or writing conventions (e.g., adherence to formal structure and grammar) without attention to personal or creative elements.	"The cover email is concise and well-structured. It provides relevant details about his background in communication systems and project management, with a concrete example from his internship experience."
ER–, SR+ Knower code	Focus on personal involvement, creativity, originality, or the perceived authenticity of the author, with little attention to technical or writing aspects.	"This email feels genuine and unique, showing the candidate's personality and motivation."
ER–, SR– Relativist code	Little emphasis on either technical skills (content or writing) or personal attributes, resulting in generic or unremarkable feedback.	"This email is very general and doesn't showcase skills or exact achievements. It uses the exact same phrases as others."

4. Results

For the results, the fifteen cover emails selected have been labelled alphabetically, while their corresponding comments have been numbered. Each comment has been referenced using the email's letter and the comment's number (e.g., "Email C, Comment 3" is coded as C3). Also, students' names have been changed to hide their identity.

4.1. Quantitative analysis

This section presents the quantitative analysis of email quality, summarizing score ranges, mean scores, and the level of agreement among raters (Table 1).

Table 1. Quantitative analysis.

Email quality	Scores and mean	Comments on scores and quality
Top-quality (A–E)	Scores between 9 and 10 (ranking 1–2), with a mean of 10	Universal agreement on their high quality. The raters demonstrated a high level of consistency in their evaluations, indicating strong interrater reliability.
Low-quality (F–J)	Scores between 6 and 7 (ranking 5–4), with a mean of 6	Clear consensus on their lower quality.
Mid-quality (K–O)	Scores ranging from 7 to 8 (ranking 4–3), with means of 7 or 8	Slight variations likely attributable to differences in subjective interpretation of quality.

4.2. Qualitative analysis

ChatGPT (OpenAI, 2025) was used as a supplementary tool to assist in identifying and organizing key themes during the analysis of qualitative data. In Tables 2 to 4, the characteristics defining each of the three quality levels are summarized, along with examples drawn from the peer evaluation comments.

Table 2. High-quality emails.

Comments overview: The high-quality cover emails (A–E) exhibit several key strengths that reinforce their effectiveness, with weaknesses mostly restricted to minor aspects, such as phrasing.	
Features	**Examples**
1. The emails are consistently praised for their well-organized and professional presentation. Comments highlight concise structure or emphasize clarity.	"The letter is concise and well-structured." (A1) "It is clear, well-structured and he has presented a perfect balance of technical and soft skills." (C4)
2. They demonstrate a clear understanding of the job requirements, effectively highlighting relevant technical expertise and aligning with the advertised role. Comments emphasize the balance of technical and soft skills.	"I liked how you didn't only tell your technical and soft skills, but also related them to the position." (A2) "This cover email best combines relevant experience, adaptability, and a clear focus on educational impact." (B1)
3. High-quality emails include specific examples that substantiate their claims, making them more credible. Comments praise that they provide concrete examples, such as internship experience, or reinforce technical credibility through technical project mentions.	"It shows professionalism, including a specific project related to mobile networks, showing relevant skills in telecommunications." (D3) "Some specific projects are included, which demonstrates initiative and technical skills." (D1)
4. The emails incorporate an engaging tone and a clear sense of enthusiasm, creating a strong connection with the reader.	"Mario balances technical skills with soft skills and genuine motivation. His tone is inviting, and his emphasis on growth, learning, and real-world impact makes this a highly memorable letter." (C2)

Table 3. Low-quality emails.

Comments overview: The low-quality cover emails (F–J) share several notable weaknesses that undermine their effectiveness, with some strengths offering limited compensatory qualities.	
Features	**Examples**
1. Poor grammar, awkward phrasing, and structural problems are recurring issues, making the emails difficult to read and less professional.	Comments F1 and F4 highlight issues with long sentences and incorrect punctuation: "the hardest one to finish even though he connects the different points he is making." (F4) Comments I1, I2, and I5 criticize awkward phrases like "this opportunity strongly interests to me."
2. Although some emails exhibit basic structure, others are criticized for their brevity, lack of detail, or formatting issues.	"Could have add more information about your skills or interests." (H4). Comment J3 observes that the format is "not visually attractive."
3. These emails fail to provide specific examples of skills, achievements, or connections to the job requirements, reducing their credibility.	"Very general: no showcasing skills or exact achievements - uses the exact same phrases as others - no personal connection or story related to the job." (F3) "Adding more examples of his skills or experience would make it stronger." (J4)
4. The emails come across as impersonal, failing to show enthusiasm or create a connection with the reader.	"Next time try and add some personal touches to the email, with your own experiences." (G3) There is "no personal connection or story related to the job." (F3)
5. Despite their weaknesses, some emails display occasional positive features, such as motivation or enthusiasm.	Comment F4 acknowledges the candidate's motivation Comment I4 praises the candidate's enthusiasm for green projects and relevant language skills Comments J1 and J2 highlight the candidate's interest in social impact and practical engineering skills.

Table 4. Mid-quality emails.

Comments overview: Mid-quality emails (K-O) exhibit a mix of strengths and weaknesses, which can be categorized into two distinct groups (emails K–M and emails N–O).	
Features	**Examples**
Emails K–M (strengths)	
1. Comments across these emails praise their formal tone and clear organization. Some comments highlight the concise presentation of qualifications.	"Pablo's cover letter is professional and highlights relevant qualifications concisely." (K2). "The letter is correct, nice structured and well carried out." (M2)
2. Each email references relevant qualifications, skills, or technical backgrounds.	"He clearly states his technical foundation and emphasizes a strong drive for teamwork and contributing positively to the project." (K2) "You mention your double degree in engineering, showing you have the necessary technical background." (L5)
Emails K–M (weaknesses)	
3. Lack of specific examples. All three emails fail to include concrete examples or detailed achievements, reducing their technical credibility.	"By including specific examples and a personal connection, you can make it even more impactful." (K5) "He could strengthen his letter by adding a concrete example of an experience where he applied his knowledge." (L4)
4. These emails show limited personal engagement, lack warmth or a compelling narrative to connect the candidate emotionally to the role.	"I think it should be more personal and warm." (M3) "By including specific examples and a personal connection, you can make it even more impactful" (K5).
Emails N–O (strengths)	
5. Both emails are praised for expressing enthusiasm, motivation, and a clear desire to contribute to the role.	"Additionally, she has a great passion for innovation and wants to grow in her professional career with this internship." (N5) "I think this email shows a related to the topic background; motivation to participate on the project and the willingness of causing a meaningful impact." (O1)
Emails N–O (weaknesses)	
6. While enthusiasm is evident, both emails could better express passion and desire in their conclusions.	"I think that the conclusion could show more enthusiasm and desire to get the job. I think that this email would be better if you comment how your experience and the independent projects that you did, can help the rural area and show more enthusiasm in the conclusion." (O5)
7. Both emails lack technical depth and specific examples or details connecting skills and experiences to the job.	"In your cover letter, you explain why do you like the job, but I miss some hard skills and why you are the perfect candidate for the role" (N1) "I think it lacks relating his skills and personal projects to the rural field." (O5)

5. Discussion

Comments A-E provide detailed and specific feedback, offering a strong basis for evaluating both technical rigor (epistemic relations, ER) and personal engagement (social relations, SR). They often include actionable suggestions, such as rephrasing or adding specific examples demonstrating thoughtful engagement with the content: "Consider refining phrases to make the letter more concise, such as rephrasing 'I want to create a better life for people' with a more focused impact statement related to the program." (E2) or "Your letter is engaging and well-structured. Including specific examples would further enhance your strong candidacy." (C5)

Some comments, while valid, focus more on subjective preferences or secondary aspects. For example, Comment E5 critiques the opening line and the inclusion of a question but acknowledges the email's overall effectiveness: "I really liked the fact that you presented different ways to contact you, such

as telephone number and linkedin, not only email. The structure of the email is perfect. I didn't like the first line: 'My name is Duna and I want to create a better life for people.' It seems to me that you are trying to sell me something more than aplying for the job." These comments do not significantly detract from the overall consensus about the emails' high quality.

In summary, emails A–E consistently rank first or near the top across all evaluations, indicating broad consensus among evaluators. Comments are aligned in identifying key strengths, such as structure, relevance, and engaging tone, further supporting the consistency of rankings. Variability is limited to subjective preferences or secondary concerns, such as format or phrasing and these critiques do not significantly impact the overall assessment. Emails A–E consistently demonstrate a strong balance of technical content (ER+) and personal engagement (SR+). Critiques are minor and do not undermine the overall classification of these emails as elite codes (ER+, SR+).

Comments F-J provide thoughtful critiques, offering detailed feedback about grammar, structure, and missing content. For instance, Comments F3, G4 and H2 emphasize the lack of evidence for skills: "It's more beneficial to provide examples that demonstrate your skills. Anyone can claim to possess certain skills, but providing examples adds credibility by showing where you acquired those skills and that you genuinely possess them" (F3). These comments align well with specialization analysis, identifying deficiencies in epistemic relations (ER) and social relations (SR).

All emails F-J rank fourth or fifth across evaluations, reflecting strong agreement about their deficiencies. Variability in rankings is minimal, with some comments emphasizing positive traits (e.g., enthusiasm in emails G and J) but agreeing on the overall weaknesses. Feedback consistently identifies poor grammar, lack of detail, and minimal personal engagement as key issues. Variability in the depth of comments (e.g., detailed critiques like Comment F4 vs. brief remarks like Comment H5) does not affect the consensus about the emails' low quality. In summary, emails F–J consistently exhibit significant weaknesses in both technical content (epistemic relations, ER–) and personal engagement (social relations, SR–), situating them in the relativist code quadrant.

Across all emails (K-O), detailed feedback provides actionable suggestions for improving technical content and engagement: "Good expression, but it should include personal experience and explain more about why is she the one for the offer" (N3). Comments focused solely on structure or tone contribute less depth to the analysis: "Good structure. I like it. Well structure" (O2)

Mostly mid rankings across all K-O emails are consistently ranked in the third or fourth positions, reflecting their mid-quality status. Rankings reflect each email's strengths and weaknesses: Emails K–M demonstrate technical

potential and professionalism (ER+) undermined by weaknesses in social relations (SR−) as they do not connect personally with the reader (SR−). These emails are situated in the knowledge code (ER+, SR−). Meanwhile, emails N–O shine in enthusiasm and emotional connection but strengths in social relations (SR+) are counterbalanced by weaknesses in epistemic relations (ER−). They emphasize enthusiasm and passion for the role (SR+) but lack technical specificity and examples (ER−). These emails are situated in the knower code (ER−, SR+).

By analysing the peer comments through the lens of LCT's specialization codes, we can draw conclusions about how students conceptualize what makes a cover email legitimate. These conclusions are summarized in Table 5.

Table 5. LCT specialization codes and students' criteria.

Quality and LCT codes	Students' criteria
High-quality cover emails align with elite codes (ER+, SR+).	• Clarity and structure: Comments for high-ranked emails frequently highlight clear organization, concise language, and proper adherence to the structure discussed in class. • Engagement and relevance: Effective emails demonstrate personal relevance and enthusiasm for the position. Peer evaluators value when candidates align their skills and experiences directly with the job requirements. • Balance of skills and personal touch: High-ranking comments emphasize a balance between technical qualifications (e.g., specific projects or internships) and personal attributes (e.g., enthusiasm or motivation).
Low-quality cover emails align with relativist codes (ER−, SR−).	• Grammar and format issues: Basic writing errors and unpolished language detract from credibility. • Lack of specificity: Poorly ranked emails are often critiqued for generic content or insufficient examples to support claims about skills or experiences. • Weak connection to role: Evaluators note when emails fail to explain how the applicant's skills or experiences align with the job.
Mid-quality emails can reflect a knowledge code (ER+, SR−) or a knower code (ER−, SR+).	• Knowledge code (ER+, SR−), are technically correct, show a formal tone, clear structure, and references to relevant qualifications or skills. However, they lack personal engagement, coming across as impersonal and failing to establish a meaningful emotional connection with the reader. • Knower code (ER−, SR+) emphasize enthusiasm, passion, and motivation for the role. While these emails successfully engage the reader on a personal level, they lack concrete examples of technical skills, professional achievements, or relevant experiences to substantiate their claims.

Finally, regarding the role of AI, comments such as "I think it is missing quite a lot of creativity and freshness, it is too noticeable that it is written by ChatGPT", which address whether students overly relied on AI tools, highlight perceptions of weak social relations. These concerns are similar to those raised in Di Sarno-García and Argüelles-Álvarez (under review) about the diminishing sense of personal involvement.

Emails consistently ranked first or ranked highly by most evaluators reflect clear agreement among peers regarding their technical and personal strengths. Moreover, detailed and constructive comments tend to reinforce rankings and align well with specialization analysis. Superficial or narrowly focused comments (e.g., emphasizing grammar or structure alone) are less consistent and some introduce subjectivity but do not significantly affect the overall trends.

The quantitative analysis reinforces the qualitative findings. High-quality and low-quality emails were universally recognized as best or worst, while mid-quality emails showed more variability in scores. The overall consistency of ratings suggests a reasonable level of interrater reliability, though some subjectivity was evident in scoring mid-quality emails.

6. Conclusions

This study has explored the application of peer assessment using a Comparative Judgement technique based on a ranking method in a learning context where higher education students in the area of engineering overtly used ChatGPT to complete a professional writing task. The Legitimation Code Theory (LCT) and the Specialization dimension, in particular, have been used to analyse and interpret peer students' feedback and evaluations.

The findings confirm that peer assessment is an effective method for evaluating the quality of AI-assisted writing tasks, fostering students' ability to critically assess writing beyond technical correctness. Students demonstrated consistency in their assessments, revealing the implicit and explicit criteria that define a 'high-quality' cover email. These high-quality texts exhibit a balance between strong technical skills and engaging, authentic narratives, exemplifying the characteristics of elite codes (ER+, SR+). Mid-quality emails display distinct patterns of strengths and weaknesses, aligning with knowledge codes (ER+, SR−) and knower codes (ER−, SR+). In contrast, low-quality emails primarily exhibit weaknesses and correspond to relativist codes (ER−, SR−).

Furthermore, the study highlights that the use of ChatGPT influences students' perceptions of writing quality, with higher-rated texts reflecting a balance between technical proficiency and personal engagement. Comparative peer assessment has proven to enhance students' awareness of originality and critical thinking, mitigating the risk of over-reliance on AI-generated content. The application of LCT effectively makes students' evaluative criteria visible, revealing how specialization codes shape the aspects they value in writing and how these relate to the final grade in AI-mediated tasks.

From a pedagogical perspective, the peer assessment process fostered critical engagement by requiring students to provide feedback, which deepened their understanding of effective writing. The ranking methodology proved effective in highlighting strengths and areas for improvement, with the added dimension of peer feedback enriching the learning process. However, the critique of emails as "too generic" or "obviously written by ChatGPT" underlines concerns about over-reliance on AI, which can diminish authenticity and personal involvement, as noted in prior research (Di Sarno-García & Argüelles-Álvarez, under review).

In conclusion, the study validates the peer assessment CJ model in the context of higher education for evaluating AI-assisted writing and confirms the usefulness of specialization codes for categorizing quality. This framework not only identifies what makes a "successful" email but also can provide guidance to help students meet those standards effectively.

Finally, while the rankings were largely consistent, practical measures such as addressing outliers and ensuring transparency in handling discrepancies are recommended for pedagogical in-class application. Future research could explore scaling this approach in diverse educational contexts or further examine the interplay of AI assistance and personal engagement in professional writing.

Acknowledgements

The author acknowledges the use of ChatGPT (OpenAI, version January 2025) as a tool to assist in the qualitative data analysis and revise linguistic correctness of this paper. The primary work and interpretation, however, were carried out solely by the author.

Competing Interests

The author reports there are no competing interests to declare.

Published: June 16, 2025 EEST.

REFERENCES

Alexander, K., Savvidou, C., & Alexander, C. (2023). Who wrote this essay? Detecting AI-generated writing in second language education in higher education. *Teaching English with Technology*, *23*(2), 25–43. https://doi.org/10.56297/BUKA4060/XHLD5365

Atlas, S. (2023). *ChatGPT for higher education and professional development: A guide to conversational AI*. University of Rhode Island. https://digitalcommons.uri.edu/cba_facpubs/548

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement*, *74*(5), 795–808. https://doi.org/10.1177/0013164414527450

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Baskara, F. X. R., & Mukarto, F. X. (2023). Exploring the implications of ChatGPT for language learning in higher education. *Indonesian Journal of English Language Teaching and Applied Linguistics*, *7*(2), 1–15. https://doi.org/10.21093/ijeltal.v7i2.1387

Bramley, T. (2007). Quantifying marker agreement: Terminology, statistics and issues. *Research Matters*, *4*, 22–27. https://doi.org/10.17863/CAM.100457

Callanan, L. (2024). Artificial intelligence: Panacea or Pandora's box? *TLN Journal*, *31*(2), 2.

Çobanoğulları, F. (2024). Learning and teaching with ChatGPT: Potentials and applications in foreign language education. *The EuroCALL Review*, *31*(1), 4–15. https://doi.org/10.4995/eurocall.2024.19957

Deng, J., & Lin, Y. (2023). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, *2*(2), 81–83. https://doi.org/10.54097/fcis.v2i2.4465

Di Sarno-García, S., & Argüelles-Álvarez, I. (under review). *Behind my students' cover email: the role of ChatGPT in homework writing tasks*.

Hendry, G. D., Bromfield, A. J., & Smith, K. E. (2017). ComPAIR: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry: The ISSOTL Journal*, *5*(2), 89. https://doi.org/10.20343/teachlearninqu.5.2.8

Hindhede, A. L., & Højbjerg, K. (2024). Disciplinary knowledge, pedagogy, and assessment in non-university marine engineering education – consequences for student academic success. *Teaching in Higher Education*, *29*(6), 1521–1536. https://doi.org/10.1080/13562517.2022.2067746

Jones, I. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation*, *47*, 93–101. https://doi.org/10.1016/j.stueduc.2015.09.004

Jones, I., & Davies, B. (2024). Comparative judgement in education research. *International Journal of Research & Method in Education*, *47*(2), 170–181. https://doi.org/10.1080/1743727X.2023.2242273

LatinCALL Conference. (2024, November 9–10). *Challenges and new horizons in the age of AI*. http://www.latincall.org

Levine, S., Beck, S. W., Mah, C., Phalen, L., & Pittman, J. (2024). How do students use ChatGPT as a writing support? *Journal of Adolescent & Adult Literacy*, *67*(5), 473–483. https://doi.org/10.1002/jaal.1373

Maton, K. (2014). *Knowledge and knowers: Towards a realist sociology of education*. Routledge. https://doi.org/10.4324/9780203885734

Maton, K., & Chen, R. T.-H. (2015). LCT in qualitative research: Creating a translation device for studying constructivist pedagogy. In K. Maton, S. Hood, & S. Shay (Eds.), *Knowledge-building: Educational studies in Legitimation Code Theory* (pp. 45–67). Routledge.

Maton, K., & Chen, R. T.-H. (2019). Knowledge, knowers, and student success. In J. R. Martin, K. Maton, & Y. J. Doran (Eds.), *Accessing academic discourse: Systemic functional linguistics and legitimation code theory* (pp. 35–58). Routledge. https://doi.org/10.4324/9780429280726

Maton, K., Hood, S., & Shay, S. (Eds.). (2018). *Knowledge-building: Educational studies in Legitimation Code Theory*. Routledge. https://doi.org/10.4324/9780429441919

Morton, T., & Nashaat-Sobhy, N. (2024). Exploring bases of achievement in content and language integrated assessment in a bilingual education program. *TESOL Quarterly*, *58*(1), 5–31. https://doi.org/10.1002/tesq.3207

Nashaat-Sobhy, N. (2022). Promoting and assessing knowledge building in the writing of English-medium instruction students. In O. Z. Barnawi, M. S. Alharbi, & A. A. Alzahrani (Eds.), *Transnational English Language Assessment Practices in the Age of Metrics* (pp. 130–152). Taylor & Francis. https://doi.org/10.4324/9781003252382-12

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Pollitt, A., & Crisp, V. (2004, September). Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions. *BERA Annual Conference*.

Rangelov, M. (2024). AI: Revolutionising education. *TLN Journal*, *31*(2), 3.

Sajawal, M. F., & Kittur, J. (2024). Examining students' beliefs on the use of ChatGPT in engineering. *2024 ASEE Annual Conference & Exposition*. https://doi.org/10.18260/1-2--47373

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, *29*(3), 211–223. https://doi.org/10.1080/08957347.2016.1171769

Thwaites, P., & Paquot, M. (2024). Comparative judgement for advancing research in applied linguistics. *Research Methods in Applied Linguistics*, *3*(3), 100142. https://doi.org/10.1016/j.rmal.2024.100142

van Daal, T., Snajder, M., Nijs, K., & Van Dyck, H. (2023). Peer assessment using criteria or Comparative Judgement? A replication study on the learning effect of two peer assessment methods. In O. Noroozi & B. De Wever (Eds.), *The Power of Peer Learning. Social Interaction in Learning and Development*. Springer. https://doi.org/10.1007/978-3-031-29411-2_4

Waters, A. E., Tinapple, D., & Baraniuk, R. G. (2015). BayesRank: A Bayesian approach to ranked peer grading. *Proceedings of the 2nd ACM Conference on Learning @ Scale (L@S 2015)*, 177–183. https://doi.org/10.1145/2724660.2724672

Xiao, Y., & Lucking, R. (2008). The impact of two types of peer assessment on students' performance and satisfaction within a Wiki environment. *The Internet and Higher Education*, *11*(3–4), 186–193. https://doi.org/10.1016/j.iheduc.2008.06.005