Chapter 06 / Capítulo 06

New literacies in the age of AI: Ethics, teaching, and writing (English Version)
ISBN: 978-9915-9854-5-9
DOI: 10.62486/978-9915-9854-5-9.ch06
Pages: 81-93

©2025 The authors. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY) 4.0 License.

Generative feedback: causal effects of LLM's on writing quality and evaluative equity in higher education

Retroalimentación generativa: efectos causales de los LLM en la calidad de la escritura y la equidad evaluativa en educación superior

Carluys Suescum Coelho[1], Car-Emyr Suescum Coelho[2], Carlysmar Suescum Coelho[1], Carelys Suescum Coelho[1]

[1]Centro de Estudios Gerenciales Avanzados (CEGA). Caracas, Venezuela.

[2] Universidad Metropolitana, Universidad Central de Venezuela. Caracas, Venezuela.

Corresponding author: Carluys Suescum Coelho

ABSTRACT

This chapter examines, within the framework of SDG 4, how large-scale generative *feedback* using large language models (LLMs) impacts the quality of scholarly writing and evaluative equity in higher education. The aim is to estimate improvements in coherence, argumentation, use of evidence, and clarity, and to determine whether gaps between subgroups (L1/L2, first generation, and baseline performance) are reduced. An experimental or quasi-experimental design (*stepped-wedge* cluster trial) with blinded assessment based on analytical rubrics and secondary metrics such as time to *feedback*, revision iterations, editing distance, self-efficacy, and teaching load is recommended. A synthesis of recent evidence suggests consistent gains when LLM *feedback* is effectively integrated into explicit rubrics and revision microtasks; furthermore, it increases engagement and can decrease inter-rater variability. However, risks of dependency, stylistic homogenization, and algorithmic bias are noted. Therefore, this chapter proposes a framework for didactic governance with traceability (registration of *prompts*, versions, and logs), open science principles (pre-registration, repositories), and a dual human-AI evaluation scheme with bias audits and human-in-the-loop thresholds. Among the limitations acknowledged are the heterogeneity of contexts, sensitivity to model versions, and external validity. The conclusion is that, under ethical safeguards and with teacher training, generative *feedback* can raise writing standards and broaden inclusion; further research is suggested on differential effects, disciplinary transferability, and institutional sustainability of scaling up.

Keywords: Evaluative Equity; Academic Writing; Large-Scale Language Models (LLMS); SDG 4 (Quality Education); Generative *Feedback*; Algorithmic Biases.

RESUMEN

Este capítulo examina, en el marco del ODS 4, cómo la retroalimentación generativa a escala mediante modelos de lenguaje grandes (LLM) incide en la calidad de la escritura académica y en la equidad evaluativa en educación superior. El objetivo es estimar mejoras en coherencia, argumentación, uso de evidencia y claridad, y determinar si se reducen brechas entre subgrupos (L1/L2, primera generación y rendimiento basal). Se recomienda un diseño experimental o cuasiexperimental (ensayo por clúster tipo *stepped-wedge*) con evaluación ciega basada en rúbricas analíticas, y métricas secundarias como tiempo a *feedback*, iteraciones de revisión, distancia de edición, autoeficacia y carga docente. La síntesis de evidencia reciente sugiere

ganancias consistentes cuando el *feedback* del LLM se integra efectivamente a rúbricas explícitas y microtareas de revisión; además, aumenta el involucramiento y puede disminuir la variabilidad interevaluador. Sin embargo, se advierten riesgos de dependencia, homogeneización estilística y sesgos algorítmicos. Por ello, el capítulo plantea un marco de gobernanza didáctica con trazabilidad (registro de *prompts*, versiones y logs), principios de ciencia abierta (pre-registro, repositorios) y un esquema de evaluación dual humano-IA con auditorías de sesgo y umbrales human-in-the-loop. Entre las limitaciones, se reconocen la heterogeneidad de contextos, la sensibilidad a versiones de modelos y la validez externa. Se concluye que, bajo salvaguardas éticas y formación docente, la retroalimentación generativa puede elevar estándares de escritura y ampliar la inclusión; se sugieren además, líneas futuras sobre efectos diferenciales, transferibilidad disciplinar y sostenibilidad institucional del escalamiento.

Palabras clave: Equidad Evaluativa; Escritura Académica; Modelos de Lenguaje a Gran Escala (LLM); ODS 4 (Educación de Calidad); Retroalimentación Generativa; Sesgos Algorítmicos.

INTRODUCTION

The phenomenon of mass higher education and the growing diversity of students has highlighted a classic structural problem, namely the provision of timely, consistent, and highly personalized academic feedback (Wambsganss et al., 2022; Zhuang et al., 2025). It is precisely in the face of this practical impossibility of attending to a large cohort of students individually that the emergence of generative artificial intelligence (Gen AI) based on large language models (LLMs) offers a significant advance.

These systems promise to provide generative feedback at scale, meaning they generate automated comments rich in examples and concrete explanations of students' texts (Benner, 2024; Thomas et al., 2025). Such immediate and timely feedback can substantially improve the quality of student writing by clearly pointing out criteria for coherence, evidence, and clarity of argument. In addition, the uniformity of the criteria used by AI promotes evaluative equity, avoiding exclusions or inequalities, as all students receive the same level of attention regardless of their origin, gender, or social status, reducing traditional gaps (González et al., 2024; Jovic et al., 2025). This is where generative feedback represents a phase change, combining scalability and traceability of the evaluation process without sacrificing pedagogical specificity.

The relevance of this approach lies in two significant potential contributions. The first is that it can improve students' writing quality. In this regard, recent research has found positive effects on writing tasks when detailed automated feedback is incorporated. Wambsganss et al. (2022) observed that students who received automated argumentation feedback, along with a social comparison nudge, wrote much more convincing, better-structured texts.

Similarly, Glüsing et al. (2025) found in a randomized controlled trial (RCT) that individualized feedback generated by GPT-4 improved the quality of academic abstract reviews and increased students' review time. In contrast, in language learning environments, Zhuang et al. (2025) demonstrated that an adaptive AI system significantly raised the writing scores of hundreds of high school students, findings that are consistent and in harmony with those of Mekheimer (2025), who found that graduate students of English as a foreign language obtained significantly higher post-test scores-test scores and greater improvement in content, organization, and cohesion after using AI-assisted feedback.

Secondly, the contribution to evaluative equity is emphasized: by applying consistent

criteria, AI can level the playing field for students with less linguistic proficiency or who are first-generation students, giving them access to the same constructive feedback as their more experienced peers, thereby reducing dispersion in results. This generates multiple effects, such as greater self-confidence and commitment reported by students after using AI tools, suggesting that these technologies can serve as uniform scaffolding.

This chapter aims to analyze in depth how generative feedback influences the quality of academic writing and evaluative equity, all within the framework of SDG 4 (Quality Education). To this end, we aim to evaluate how this type of feedback can not only assess but also improve writing quality, and then examine its impact on equity, particularly by reducing gaps associated with native language, initial performance, or first-generation status.

To this end, research questions are posed that focus on the impact on the quality of the written product, the timeliness and usefulness of the feedback generated, and the heterogeneity of effects among subgroups. Key concepts are operationally defined: generative feedback from AI-generated personalized feedback; writing quality, with the ability to argue coherently, with evidence and clarity; and evaluative equity, understood as the distributive justice of feedback in terms of access, timeliness, and uniform usefulness (Jovic et al., 2025). Notions of scaffolding (progressive educational support) and traceability (technical record of the feedback generation process) are also introduced to frame the methodological discussion.

To this end, a roadmap of the chapter is provided, developing the theoretical framework and empirical evidence on automated feedback; then a methodological design with recommended experimental approaches is proposed; a synthesis of results is presented; pedagogical and institutional implications are discussed; ethical considerations and academic integrity are addressed; and finally, limitations and the future agenda are outlined.

DEVELOPMENT

Conceptual framework

From a theoretical perspective, the literature highlights the importance of self-regulated learning and iterative writing to improve writing performance. The self-regulated learning model (Zimmerman, 2000) implies that students plan, monitor, and evaluate their own writing process. In feedback environments, the authors point out that feedback-based revision tasks are highly self-regulated, as students must use cognitive and metacognitive strategies to incorporate suggestions (Alcudia Ólan & Morales Vázquez, 2025). Tian & Liu (2022) show in their research that when receiving feedback, whether automatic, from peers, or from teachers, writers deploy various regulation strategies (planning or self-revision) to improve their texts. This suggests that well-designed, immediate, and actionable feedback stimulates the self-regulated cycle of; conversely, insufficient feedback greatly hinders self-revision.

In line with Hattie & Timperley (2007), feedback is most effective when it is immediate, accurate, and elaborative. For example, feedback that not only points out errors but also explains their correction and provides concrete examples produces greater learning gains. In this sense, the generative feedback of LLMs can act as a cognitive scaffold in that the models can provide specific examples and explanations, which help students compare their writing with established criteria in terms of argumentation, evidence, and clarity, but above all, to rewrite more effectively (Quesada & Salinas, 2021).

Furthermore, as it is an automated process, it provides actionable, consistent feedback, ensuring that all students receive guidance based on an explicit analytical rubric that defines

observable indicators. This methodological consistency of automated feedback also points to equity, as applying the same standards uniformly seeks to minimize human bias.

The concept of evaluative equity is understood here as the distributive justice of feedback, grounded in equal accessibility and utility. This implies that all students, regardless of their profile or sociocultural conditions, have equal access to relevant and timely comments. In practical terms, equity metrics include reducing both absolute and relative gaps in written results between these groups. The literature on assessment has suggested that equity does not always imply equal results, but rather a fair application of criteria (Romero-González, 2024).

Jovic et al. (2025) define equity as the fairness and consistency of feedback across different groups of students, whether native or non-native speakers. Based on the precepts of these authors, an automated feedback system must record indicators such as response time to feedback and variability in subgroup improvements, which are crucial for identifying whether existing disparities are effectively reduced.

LLMs also function as educational technology with their own potential and limitations. Among their advantages is the ability to dynamically scale feedback to fit the student's text. However, recent studies warn that these systems require human supervision. For example, evidence indicates that conversational prompting (guiding the LLM with specific questions) can improve the quality of feedback. However, LLMs frequently exhibit flaws such as hallucinations (inventing nonexistent content) or generic responses when faced with more complex arguments (Sandoval et al., 2024).

Therefore, given this scenario, it is necessary to integrate clear rubrics, guided review tasks, and validation by human experts so that AI serves to complement, rather than replace, teachers' work (Valenzuela & Pérez, 2025). Similarly, the traceability of the process through the recording of prompts, model versions, and usage logs is key to auditing and correcting algorithmic biases, as well as adjusting thresholds or levels of human intervention. In this context, generative feedback is an innovative pedagogical resource. However, in order to guarantee educational quality, it must be embedded in supervisory structures to ensure relevance and fairness in assessment (Xia et al., 2024).

Empirical research on generative feedback in higher education has recently expanded. To date, several studies have shown that automated feedback improves writing quality and the efficiency of the feedback process. In their study of 124 students, Wambsganss et al. (2022) found that those who received automated argumentation feedback combined with a social comparison incentive significantly improved the quality of their argumentative texts compared to control groups (Afrin et al., 2021). Quasi-experimental studies, such as the one developed by Glüsing et al. (2025), have confirmed that students who received feedback from GPT-4 achieved higher-quality revisions of research abstracts and showed greater behavioral engagement in the task (more revision time, more editing distance) than a group without automated feedback.

In the context of language teaching, the study developed by Zhuang et al. (2025) reports statistically significant improvements in the writing scores of secondary school students after using a multimedia AI system for adaptive feedback. Similarly, Mekheimer (2025) found that English as a foreign language (EFL) students who received AI-assisted feedback from Grammarly outperformed the control group on writing tests. In addition, they revised their texts more frequently and reported greater satisfaction and confidence in the process.

This body of evidence, as suggested by Gombert et al. (2024), demonstrates that generative feedback, when properly designed, can significantly improve the revision cycle and enhance writing quality. Furthermore, immediate feedback availability reduces students' waiting time, which is crucial in large-scale courses with high enrollment. It should be noted that, in general, students have rated AI-generated feedback positively, especially when it is perceived as specific and relevant to their learning process (Thomas et al., 2025).

In contrast, research on heterogeneous effects is still in its infancy, as few studies have systematically examined how these interventions benefit subgroups differently. However, Jovic et al. (2025) observe that AI-generated responses tend to be more consistent across students from different backgrounds, though they lack depth in analyzing complex arguments.

According to the evidence presented in this research, the gaps to be closed include validating the fairness of feedback algorithms, given concerns about the replicability of data biases, and ensuring explicit traceability to audit for possible favoritism. Precisely in this context, the state of the art suggests that generative feedback can improve learning, reducing existing disparities in a scalable manner. However, more causal and analytical research by subgroups is still needed. Therefore, the research will add value by explicitly focusing on a causal and equitable approach, as well as on the institutional dimensions of scaling and the governance of automated feedback, aspects that have been little explored until now.

Recommended methodological design for generative feedback

To rigorously evaluate the effects of generative feedback, an empirical experimental or quasi-experimental design in real higher education settings is recommended. Specifically, a cluster trial or stepped-wedge design balances internal validity and logistical feasibility (Chen et al., 2021). For example, several university courses or subjects with similar written assignments (taught by cohort or class) could be selected and randomized by groups or sequences to receive the intervention.

The control condition would correspond to the usual feedback method (professor or manual corrections). At the same time, the treatments would include at least two variants, namely: 1) LLM + structured rubric, and 2) LLM + rubric + guided review microtasks (explicit scaffolding). The target population would be undergraduate students in writing courses or subjects with substantial written components, preferably large and diverse samples to ensure greater statistical power and better subgroup analysis.

The primary results would be writing quality metrics obtained through blind evaluation with an analytical rubric, which would include key criteria such as argumentation (coherence of the thesis and its corresponding logical structure), textual coherence, use of evidence, and expressive clarity, which must be consistent with pedagogical standards (Glüsing et al., 2025). These rubrics should be developed based on previous literature on good writing practices and, as far as possible, adapted by expert linguists or teaching professionals prior to the study.

In addition, secondary outcome indicators will be collected, such as: time from text submission to receipt of feedback (to measure timeliness), number of revision iterations performed by the student, editing distance (quantitative measure of how much the text changes after feedback), writing self-efficacy (student surveys on their confidence), and perception of teaching load (questionnaires to teachers about supervision effort); these outcomes will cover both effectiveness and efficiency dimensions.

To assess equity, an analysis of effect heterogeneity would be performed. The sociodemographic data to be included would include native language, first-generation status, baseline performance decile or quartile, and gender, among others deemed relevant. Average effects (ATE) and conditional effects (CATE) will be estimated using statistical models that allow for interactions, such as regressions with subgroup dummy variables and their interactions with treatment (Holgersson et al., 2015).

In particular, we will examine whether treatment with generative feedback significantly reduces the score gap between groups relative to the control, as would be expected with changes in the absolute and relative differences in means between students at one level (L1) and students at a higher level (L2). The statistical analysis will use multilevel models (e.g., with students at level 1 and sections/classes at level 2) to control for hierarchical dependence in the data (Impey et al., 2025). Statistical correction methods for multiple comparisons (e.g., Bonferroni or FDR) will be applied when testing multiple outcomes, and effect sizes and confidence intervals will be calculated to assess practical magnitudes (Ren & Ren, 2025).

As a good practice in open science, it is recommended to pre-register the research protocol, for example by following CONSORT-AI for AI interventions (Chen et al., 2025), and to provide a public repository with anonymized materials and data such as prompts, LLM model version, hyperparameters used, complete rubrics, analysis scripts, and interaction logs (sequences of prompts and responses). This transparency facilitates reproducibility and future bias auditing.

On the presentation of results and synthesis of evidence

The results section would present both a descriptive and inferential synthesis of the findings. First, the researcher must describe the sample: the number of students, distribution by key characteristics (native language, gender, among others), and the baseline or reference point used to determine writing quality (initial scores). Previous studies have observed initial variability that justifies baseline controls (Wambsganss et al., 2022).

Next, the main effects of LLM treatment versus control on writing quality and time would be shown. For example, statistically significant and practically relevant increases in blind quality scores would be expected; the effect size (Cohen's d) would be interpreted in pedagogical terms, considering how much the average writing score improves (Sullivan & Feinn, 2012). Likewise, the change in feedback time would be evaluated; ideally, students in the LLM group would receive almost instantaneous corrections (close to zero), in contrast to the days or weeks in the control group.

Regarding differential effects, we would report how feedback influences each subgroup. In this case, we would examine whether the absolute gap between students at level 1 (L1) and level 2 (L2) in writing scores is reduced after receiving generative feedback (indicating greater equity). Differences across initial quartiles would also be analyzed to determine whether some groups obtain greater benefits or, conversely, whether the intervention maintains uniform improvements (a desirable situation for equity).

In similar studies (Jovic et al., 2025), greater relative gains are sometimes observed in traditionally disadvantaged groups, as AI can raise the lowest base. However, the possibility of reverse bias will also be considered when LLMs privilege the linguistic patterns of native speakers, for example, in second-language learning. Complementarily, the robustness of these results would be evaluated through a sensitivity analysis and controls for contamination between groups to prevent the control group from receiving additional feedback.

Subsequently, complementary indicators would be presented, such as when, in the experiment by Glüsing et al. (2025), feedback from the LLM increased editing time (an indicator of engagement), which in turn mediated higher revised quality. Student assessments of the usefulness of feedback (using Likert scales) would also be reported to understand the perception of usability. All tables would include 95 % confidence intervals for estimated effects and statistical criteria (p-value, effect size). In this way, the section would summarize the key findings, namely changes in blinded quality scores and response times, along with the reduction or increase in gaps between subgroups, based on the interpretation of practical magnitudes.

The hypothetical findings described would have several pedagogical and, obviously, institutional implications. First, a positive effect on writing quality would indicate that generative feedback acts as effective cognitive scaffolding. This can be interpreted through self-regulated learning mechanisms when receiving explanations and examples, whereby students assimilate explicit criteria (metacognition) and improve their writing in successive iterations.

The results of Wambsganss et al. (2024) show that the motivational component (social nudge) combined with explicit feedback triggered favorable psychological processes (self-efficacy, motivation). Similarly, the use of specific examples in feedback (content generated by an LLM) provides direct scaffolding, as the student sees models of argument structure or appropriate citations, which facilitate rewrites. In practice, this suggests that AI-based pedagogical interventions should include prompts that generate clarifications and concrete examples, rather than merely making general observations.

In terms of equity, the results indicate the extent to which technology leveled the playing field. A finding of reduced gaps between L1 and L2 students, or greater improvements in the first deciles, would support the idea that AI can promote distributive justice. However, this should be interpreted with caution: the internal validity of the experiment (randomization, control) strengthens the causal inference from these average effects, but generalizability (external validity) will depend mainly on the institutional context. For example, if the study were conducted in universities with high technological infrastructure, its transfer to universities with fewer resources would require adaptations. Similarly, consistency in causal inference is reinforced by the multilevel design and rigorous analysis applied (Impey et al., 2025), but the variability of LLMs, whether due to the emergence of new versions or changes in prompts, may limit exact replicability.

In any case, the results would have obvious implications: if generative feedback demonstrates transparency and robustness, its institutional deployment would be recommended. This implies, first of all, the implementation of AI-based teaching co-pilots to assist students and teachers, such as writing chatbots, review assistants, or AI agents. In addition, to enhance these results, institutions will need to invest in teacher training, teaching teachers how to interpret and complement automated feedback and how to use traceability dashboards to see what feedback was given.

Secondly, from a governance perspective, it would be necessary to define continuous improvement protocols that periodically review prompts and adjust the system based on actual performance data. In short, this is a systemic approach to decision-making that, being data-driven, would be key, leveraging feedback log analytics to refine teaching policies and thus ensure writing excellence through deeply responsible AI.

Ethical considerations, integrity, and traceability

The use of AI in education poses significant ethical and academic integrity challenges that must be explicitly addressed. In terms of ethics and privacy, students' informed consent must be obtained to process their writing with AI, ensuring, above all, that the data is anonymized. In addition, the principle of minimization must be applied, meaning that only the information strictly necessary to evaluate the system's effectiveness should be collected, and personal identifiers should be removed. Similarly, students who choose not to participate must be guaranteed equivalent support with manual feedback, without any reprisals or sanctions.

In terms of academic integrity, it is crucial to emphasize students' active role in creating their writing. While it is true that AI will provide feedback, the final intellectual work belongs to the student and is their intellectual production. Human authorship is therefore promoted, and AI systems should not be listed as authors in any derivative product. This starts with requiring students to declare the use of AI tools in their processes to safeguard transparency. In addition, it is advisable to integrate training modules on the responsible use of AI, emphasizing that AI is an assistant, not a substitute for critical-analytical thinking. Mekheimer (2025) found that students emphasize the need to avoid over-reliance on AI and, for this reason, the study documentation will require reinforcing digital and ethical literacy, prioritizing critical thinking skills even when AI is used.

Regarding technical traceability, detailed metadata must be recorded for each feedback cycle, including the prompts used with the LLMs, the model version, the generation parameters, and interaction logs, without including personal information. This will facilitate auditing for biases, allowing scenarios to be rerun if the model shows unfair responses.

This starts with establishing a review threshold: when the AI detects ambiguities or inconsistencies in a text, it automatically activates human intervention ("human-in-the-loop"), so that a teacher can review or supplement the feedback. In addition, at this point, it would be desirable to apply periodic bias audits to the system, as suggested by guidelines for responsible AI in education, to identify discriminatory patterns (UNESCO, 2024), since transparency and traceability are what guarantee that the use of AI does not compromise the equity or quality of teaching, in line with international ethical standards.

Like any study, the proposed research has limitations. First, specific data and contexts, such as courses, subjects, institutions, and cohorts, may limit generalizability. The results will depend on the LLM model version available during the experiment phase, and future updates to the model could alter the effectiveness of the feedback (Saini et al., 2024). In addition, the measurement of writing quality would depend on the reliability of the analytical rubrics. Although the grading will be shielded, there will always be room for subjectivity. Moreover, in terms of equity, the categories analyzed for L1 and L2 student levels may be correlated with unobserved factors, such as socioeconomic status, which would complicate the interpretation of the gaps.

It is important to emphasize that the chapter proposes a future agenda to deepen these explorations. Therefore, longitudinal follow-ups are recommended to assess whether writing improvements persist over time or accumulate with continuous feedback. It would also be valuable to extend the study to other disciplines where writing has very different functions. The integration of this feedback with broader learning analytics can enrich the understanding of the mechanisms of change. In view of this, it is suggested that the cost, benefits, and institutional effectiveness be evaluated, because although the technology promises scalability,

its implementation requires investment in infrastructure and teacher training, a factor of great importance for sustainable adoption within the framework of SDG 4.

CONCLUSIONS

The evidence gathered shows that LLM-powered generative feedback has considerable potential to improve writing quality in secondary and higher education by providing detailed, timely comments that leverage students' self-regulated learning. Likewise, this approach can contribute to greater evaluative equity by applying consistent criteria, thereby reducing baseline disparities between groups and expanding access to meaningful feedback. However, it is important to note that these positive effects depend on careful design that requires human oversight and transparent architecture to avoid automatic biases, as the current literature still lacks robust research on the precise differential effects by subgroups and on large-scale institutional deployment.

As recommendations, teachers and curricula are urged to adopt AI-assisted teaching gradually; this involves defining clear adoption criteria, such as explicit rubrics, and training teachers to interpret automated feedback. At the institutional level, scaling should be planned with continuous auditing by educational AI ethics committees, and the AI strategy should be linked to the equity goals established in SDG 4. Above all, the latent need to continue generating scientific evidence is emphasized, stemming from the creation of university research networks that share data in a protected manner and document their lessons learned.

Finally, and looking ahead to the roadmap for the remaining years to reach the 2030 agenda deadline, this chapter envisions a future in which writing excellence and equity are enhanced by responsible AI. To achieve this, iterative improvement paths must be mapped out that monitor new LLM developments, integrate automated feedback into curricula, and train the entire academic community in its critical use. Only then can the quality education envisaged in SDG 4 achieve its goal of relevant and effective learning for all students, driven by scalable and equitable feedback.

REFERENCES

Afrin, T., Kashefi, O., Olshefski, C., Litman, D., Hwa, R., & Godley, A. (2021). Effective interfaces for student-driven revision sessions for argumentative writing. In Proceedings of the 2021 CHI conference on human factors in computing systems. https://doi.org/10.1145/3411764.3445683

Alcudia Olán, Z. V., & Morales Vázquez, E. (2025). Aprendizaje Autorregulado y Aplicaciones Digitales. Estudios Y Perspectivas Revista Científica Y Académica. 5(3), 3815–3830. https://doi.org/10.61384/r.c.a.v5i3.1434

Benner, D. (2024). Introducing Generative *Feedback* to Chatbots in Digital Higher Education. Proceedings of WI2024, 99. https://aisel.aisnet.org/wi2024/99

Chen, D-G., Pan, W., & Kainz, K. (2021). *Stepped-wedge* Cluster Randomized Controlled Trial for Intervention Research: Design and Analysis. Journal of the Society for Social Work and Research. 12 (2). https://doi.org/10.1086/714135

Chen, D., Arnold, K., Sukhdeo, R., Farag J., & Raman, S. (2025). Concordance with CONSORT-AI guidelines in reporting of randomised controlled trials investigating artificial intelligence in oncology: a systematic review. BMJ Oncology. 2025;4:e000733. https://

doi.org/10.1136/bmjonc-2025-000733

Glüsing, R., Fleckenstein, J., Schmidt, F. T. C., & Möller, J. (2025). LLM *feedback* for academic writing: Effects on students' performance and engagement. https://doi.org/10.2139/ssrn.5445319

Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J., Frey, A., & Drachsler, H. (2024). From the automated assessment of student essay content to highly informative *feedback*: A case study. International Journal of Artificial Intelligence in Education, 34(3). https://doi.org/10.1007/s40593-023-00387-6

González, L., O'Neil-Gonzalez, K., Eberhardt-Alstot, M., McGarry., M., & Van Tyne, G. (15 de agosto de 2024). Leveraging Generative AI for Inclusive Excellence in Higher Education. EDUCASE REVIEW. https://n9.cl/85lrnr

Hattie, J., & Timperley, H. (2007). El poder de la retroalimentación. Revista de Investigación Educativa , 77 (1), 81-112. https://doi.org/10.3102/003465430298487

Holgersson, T., Nordström, L., & Öner, Ö. (2015). On regression modelling with dummy variables versus separate regressions per group: comment on Holgersson et al. Journal of Applied Statistics, 43(8), 1564–1565. https://doi.org/10.1080/02664763.2015.1092711

Impey, C., Wenger, M., Garuda, N., Golchin, S., & Stamer, S. (2025). Using Large Language Models for Automated Grading of Student Writing about Science. International Journal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-024-00453-7

Jovic, M., Papakonstantinidis, S., & Kirkpatrick, R. (2025). From red ink to algorithms: investigating the use of large language models in academic writing *feedback*. Language Testing in Asia. 15 (59). https://doi.org/10.1186/s40468-025-00389-2

Mekheimer, M. (2025). Generative AI-assisted *feedback* and EFL writing: a study on proficiency, revision frequency and writing quality. Discover Education 4, 170. https://doi.org/10.1007/s44217-025-00602-7

Quezada, S., & Salinas, C. (2021). Modelo de retroalimentación para el aprendizaje: Una propuesta basada en la revisión de literatura. Revista mexicana de investigación educativa, 26(88).225-251. https://ojs.rmie.mx/index.php/rmie/article/view/219

Ren, F., & Ren, C. (2025). Optimizing physical education strategies through circular intuitionistic Fuzzy Bonferroni based school policy formulation. Scientific reports, 15(1), 21616. https://doi.org/10.1038/s41598-025-03363-3

Romero-González, L. (2024). La evaluación inclusiva para una evaluación justa. Know and share psychology, 5(3), 16-22. https://doi.org/10.25115/kasp.v5i3.10137

Saini, A. K., Cope, L., Kalantzis, M., & Zapata, R. (2024). The future of *feedback*: Integrating peer and generative AI reviews to support student work. https://doi.org/10.35542/osf.io/x3dct

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough.

Journal of graduate medical education, 4(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Thomas, S. L., Penn, A., Hauser, M., Severyn, A., Radev, D., & Liu, R. (2025). LLM-Generated *Feedback* Supports Learning if Learners Choose to Use It. Arvix. https://doi.org/10.48550/arXiv.2506.17006

Tian, L., & Liu, Q. (2022). Self-regulated writing strategy use when revising upon automated, peer, and teacher *feedback* in an online English as a foreign language writing course. Frontiers in Psychology, 13, 873170. https://doi.org/10.3389/fpsyg.2022.873170

UNESCO (2024). Guía para el uso de IA generativa en educación e investigación. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000389227_spa

Valenzuela, R., & Pérez, A. . (2025). Inteligencia artificial en educación superior: ¿un reemplazo para los profesores o una herramienta de apoyo?. Revista Iberoamericana De Investigación En Educación, (9). https://doi.org/10.58663/riied.vi9.221

Wambsganss, S., Günther, A., Ketelhut, D. J., & Scheiter, K. (2022). Enhancing argumentative writing with automated *feedback* and social comparison nudging. Computers & Education, 191, 104644. https://doi.org/10.1016/j.compedu.2022.104644

Xia, Q., Weng, X., Fan, O., Lin, T., & Chiu, T.K. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. International Journal of Educational Technology in Higher Education, 21, 1-22. https://doi.org/10.1186/s41239-024-00468-z

Zhuang, Y., Zhao, R., Xie, Z., & Yu, P. L. H. (2025). Enhancing language learning through generative AI *feedback* on picture-cued writing tasks. Computers & Education: Artificial Intelligence, 9, Art. 100450. https://doi.org/10.1016/j.caeai.2025.100450

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. En M. Boekaerts, P.R. Pintrich y M. Zeidner (Eds.), Handbook of self-regulation.13-40. https://doi.org/10.1016/B978-012109890-2/50031-7

CONFLICTS OF INTEREST

There are no conflicts of interest of any kind.

FUNDING

The work has not received any funding from public, commercial, or non-profit organizations.

USE OF ARTIFICIAL INTELLIGENCE

AI was used for the translation of the abstract and style correction.

AUTHORSHIP CONTRIBUTION

Conceptualization: Carluys Suescum Coelho.

Data curation: Car-Emyr Suescum Coelho, Carluys Suescum Coelho.

Formal analysis: Carelys Suescum Coelho, Carlysmar Suescum Coelho.

Research: Carelys Suescum Coelho, Car-Emyr Suescum Coelho, Carluys Suescum Coelho, Carlysmar Suescum Coelho.

Methodology: Car-Emyr Suescum Coelho, Carluys Suescum Coelho.
Project management: Carluys Suescum Coelho.
Supervision: Carluys Suescum Coelho.
Validation: Carluys Suescum Coelho, Car-Emyr Suescum Coelho.
Writing - original draft: Carluys Suescum Coelho, Car-Emyr Suescum Coelho.
Writing – proofreading and editing: Carluys Suescum Coelho, Carlysmar Suescum Coelho.