INPLASY

Hybrid Human–Generative AI Feedback in Formal Reciprocal Peer Observation Programs in Higher Education: A Systematic Review and Potential Meta-Analysis

INPLASY202680039

doi: 10.37766/inplasy2026.8.0039

Received: 10 August 2026

Published: 10 August 2026

Aburto, JAC; Rifo, KAF; Tamarín, NA.

Corresponding author:
Jorge Cuevas Aburto

jorgecuevas@santotomas.cl

Author Affiliation:
Learning Center, National Undergraduate Studies Office, Vice-Rectory for Academic Affairs, Research and Postgraduate Studies, Universidad Santo Tomás, campus Los Ángeles, Chile; Doctoral Program in Education, Universidad de Concepción, Concepción, Chile.

ADMINISTRATIVE INFORMATION

Support - No external funding is declared. The study is conducted using the research team's own resources within the scope of their institutional academic duties.

Review Stage at time of this submission - The review has not yet started.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY202680039

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 10 August 2026 and was last updated on 10 August 2026.

INTRODUCTION

Review question / Objective Among in-service university faculty participating in formal Reciprocal Peer Observation programs (P), do hybrid human–generative AI feedback systems (I), compared with exclusively human feedback, exclusively automated feedback, or the absence of a formal program (C), produce improvements in the dialogic quality of feedback (primary outcome) and reductions in evaluator time burden (secondary outcome) (O), in studies published between January 1, 2018, and August 17, 2026 (T), and to what extent are these effects moderated by sociogeographic context, particularly in Latin America (M)?
Objectives:
1. To synthesize available evidence on the effect of hybrid human–generative AI systems on the dialogic quality of feedback within formal Reciprocal Peer Observation programs in higher education.
2. To estimate quantitatively the effect of these systems on the evaluator's temporal efficiency.
3. To examine whether sociogeographic context, with an emphasis on Latin America, moderates the effects identified in the preceding objectives.

Rationale Reciprocal peer observation (RPO) is a core component of formal pedagogical support programs in higher education because, given optimal conditions such as structured design, institutional support, flexible leadership, and training, it remains an effective modality for generating sustained change in pedagogical practice (Johnston et al., 2022; Zeng, 2020). However, most interventions omit elements indispensable for a high-quality feedback cycle due to time constraints, lack of conceptual scaffolding, and a tendency toward generic comments (Flores et al., 2025; Kerman et al., 2023; Qi et al., 2026; Ribosa et al., 2024). Consequently, scaling high-quality RPO exclusively through human effort is structurally unfeasible (Maheshi et

al., 2024). To address this limitation, hybrid human–generative AI systems offer a scalable, personalizable sociotechnical response (Kaliisa et al., 2026), provided they are properly calibrated (C.-Q. Huang et al., 2026; Gu & Yan, 2025; Y. Huang et al., 2026). Despite this potential, empirical evidence reveals a critical gap; existing literature predominantly focuses on students, systematically excluding in-service university faculty as feedback recipients or providers (Bond, 2024; Nedrehagen et al., 2025). Furthermore, pedagogical support programs in Latin America operate under heterogeneous digital infrastructures and weak regulatory frameworks compared to the Global North (UNESCO-IESALC, 2025), and are severely underrepresented in the literature (Bond, 2024; Crompton & Burke, 2023). Therefore, broadening the search horizon to include regional databases is required to counteract indexing bias and capture these qualitatively distinct realities (Hartuique et al., 2025).

Condition being studied This review examines the dialogic quality and temporal efficiency of faculty-to-faculty feedback exchanged within formal pedagogical support programs in higher education, specifically where such feedback is produced or mediated by hybrid human–generative AI systems. Formal programs are operationalized as institutionally documented structures comprising at least two complete, iterative cycles of teaching observation or review followed by systematized feedback, encompassing reciprocal peer observation, peer review of teaching, institutional pedagogical coaching, and faculty support with an explicit evaluative component.

The condition of interest is twofold. First, the dialogic quality of peer feedback, operationalized across five subdimensions: specificity, personalization, knowledge-building markers, orientation toward self-regulation, and dialogic tone, addresses the documented tendency of unsupported human feedback toward generic, non-dialogic commentary. Second, the faculty evaluator's temporal efficiency (time per cycle, number of iterations, perceived workload) addresses the structural time and scaffolding constraints that limit scalable, high-quality feedback provision under exclusively human models.

Both conditions are examined specifically in relation to in-service university faculty acting as both feedback recipients and providers, a population largely absent from the existing generative AI feedback literature, which predominantly concentrates on students. Furthermore, the review examines whether the sociogeographic context, particularly the structural conditions documented for Latin America, such as heterogeneous digital infrastructures and weak regulatory frameworks for AI in faculty evaluation, moderates the magnitude or direction of these effects.

METHODS

Search strategy The search strategy structures three conceptual blocks combined with the Boolean operator AND. The first is a technology block incorporating generative AI systems, large language models, and hybrid tools, specifically including AI-mediated feedback, generative AI-assisted formative assessment, dialogic AI feedback, and automated feedback system. The second is a pedagogical context block focusing on faculty support programs, peer observation, and pedagogical feedback. The third is a population block targeting in-service university faculty and higher education. Searches will be conducted in English and Spanish, with Spanish-language terms incorporated into the technology and context blocks where database fields permit. Database-specific syntaxes were developed for Scopus, Web of Science SSCI and ESCI, ERIC, PsycINFO, Education Source, ACM Digital Library, IEEE Xplore, SciELO, Redalyc, and Dialnet, adapting field codes, controlled descriptors, and Boolean complexity to the technical requirements of each interface. The complete database-specific search syntaxes will be annexed to the INPLASY registration rather than reproduced here. Coverage of Latin American production is ensured through SciELO, Redalyc, and Dialnet as primary sources. ProQuest Dissertations and Theses Global will be searched for graduate-level gray literature, alongside EdArXiv and SSRN for discipline-specific preprints, to mitigate publication bias in this emerging field. An additional search will be conducted in Google Scholar, limited to the first 200 records ranked by relevance per thematic block, with the access date documented. The search period covers January 1, 2018, to the effective closing date of the search in each database, not later than the close of the seven-calendar-day window beginning August 10, 2026, recorded specifically per source in the search log.

Participant or population The broader population of interest comprises in-service university faculty operating within higher education institutions. Within the eligible primary studies, the participants are defined specifically as faculty members who actively engage as peer evaluators (feedback providers) and/or observees (feedback recipients) in formal pedagogical support programs. These formal programs are operationalized as institutional

initiatives with a documented structure comprising at least two complete, iterative cycles of teaching observation or review, followed by a systematized, verifiable faculty-to-faculty feedback mechanism; single cycles and informal or undocumented exchanges are strictly excluded. Eligible modalities encompass reciprocal peer observation, peer review of teaching, institutional pedagogical coaching, and faculty support structures with an explicit evaluative component. Excluded are studies centered on students, preschool or primary/secondary educators, pre-service teachers, and professional development programs lacking a formal feedback component. Studies combining multiple populations are included only if disaggregated data for the in-service university faculty subpopulation are explicitly reported.

Intervention The eligible intervention is any hybrid human–generative AI system implemented within a formal program of observation of teaching practice in higher education. Generative AI (GenAI) is based on Large Language Models (LLMs) or other architectures grounded in transformer mechanisms and their associated attention mechanisms, designed to infer and produce novel natural-language text (OECD, 2024; UNESCO, 2023). Four eligible configurations are considered: (a) draft-review, in which AI generates an initial feedback draft that the human evaluator reviews and adjusts; (b) detection-elaboration, in which AI identifies markers in the observed practice that the instructor elaborates into feedback; (c) parallel-synthesis, in which AI and human generate independent feedback subsequently integrated; and (d) real-time assistance, in which AI provides live, adaptive scaffolding during observation. Systems combining generative and non-generative components are eligible only where the generative component effectively produces the textual feedback delivered to the instructor; this hybrid feature is coded as an intervention variable for heterogeneity exploration. Studies must explicitly document the GenAI technology, model, and, where available, version used.

Comparator Three valid comparators are defined: (1) faculty-to-faculty feedback produced exclusively by human evaluators within a formal pedagogical support program; (2) feedback generated exclusively by a GenAI system without human intervention in content elaboration, though non-editorial oversight (e.g., selecting which observation is submitted) may exist; and (3) absence of a formal support program for in-service teachers in higher edcucation. Studies comparing the hybrid system with non-generative AI systems (e.g., rule-based automated scoring) are recorded separately as exploratory secondary comparators and not combined with pure-GenAI comparisons. A single study may include more than one comparator, each analyzed independently respecting the original unit of randomization or comparison.

Study designs to be included Eligible designs include randomized controlled trials, quasi-experimental studies with an equivalent control group or statistical adjustment, longitudinal observational studies with between-group comparison, mixed-methods studies with an evaluable primary quantitative component, and rigorous qualitative studies. Single-case designs will be analyzed separately as exploratory evidence. Prior systematic reviews will be utilized exclusively for theoretical contrast and backward citation searching, not as primary units of analysis.

Eligibility criteria Included: studies examining in-service university faculty within formal pedagogical support programs (minimum two documented feedback cycles), evaluating hybrid human-GenAI feedback systems against exclusively human feedback, exclusively GenAI feedback, or absence of a formal program, on dialogic feedback quality and/or evaluator temporal efficiency, published in Spanish or English between January 1, 2018, and the search closure date. Eligible designs: RCTs, quasi-experimental studies, longitudinal observational comparisons, mixed-methods studies, and qualitative studies with explicit rigor criteria; single-case designs analyzed separately.

Excluded: studies centered on students, non-university-level or pre-service teachers, professional development without a formal feedback component; interventions using non-generative AI as the sole comparator without a qualifying primary comparator; studies not reporting disaggregated data for in-service university faculty when populations are mixed; studies in languages other than Spanish or English.

Information sources Searches will be executed in Scopus (Elsevier), Web of Science Core Collection with emphasis on SSCI and ESCI (Clarivate), ERIC via EBSCO, PsycINFO via APA PsycNET, Education Source (EBSCO), ACM Digital Library, IEEE Xplore, SciELO, Redalyc, Dialnet, and ProQuest Dissertations & Theses Global for graduate-level gray literature. To mitigate publication bias, discipline-specific preprint repositories EdArXiv and SSRN will also be searched. An additional Google Scholar search will be conducted, limited to the first 200 records ranked by relevance per thematic block, with

access date documented, acknowledging that platform's reproducibility limitations. Regional Latin American databases (SciELO, Redalyc, Dialnet) are treated as primary, not complementary, sources given documented Latin American underrepresentation in Anglophone indexing.

Main outcome(s) The primary outcome is dialogic quality of feedback, operationalized across five subdimensions: specificity of content, personalization to the observed practice, knowledge-building markers (adapted exploratorily from Scardamalia & Bereiter, 2005), orientation to self-regulation, and dialogic tone. In the absence of a standardized instrument for this construct in AI-mediated peer observation, these dimensions will be operationalized through a low-inference qualitative codebook based on directed qualitative content analysis (Hsieh & Shannon, 2005), calibrated via Cohen's Kappa. This outcome will be synthesized primarily through narrative synthesis; meta-analysis is exploratory, contingent on standardized or transformable metrics and $k \geq 5$ studies.

Additional outcome(s) The secondary outcome is the faculty evaluator's temporal efficiency, measured continuously as time per feedback cycle, number of iterations, and perceived workload. Additional secondary outcomes, extracted when available, include the recipient's trust in the feedback, observable changes in pedagogical practice, faculty perceptions of agency and autonomy, and indicators of over-reliance on the generative AI system.

Data management Study selection follows independent double screening in two phases using Rayyan QCRI (web and mobile v. 1.7.7) with Blind Mode active until discrepancy resolution; access date and version are documented in the methodological log. Screening and discrepancy resolution run between August 17 and September 15, 2026. Inter-rater agreement is assessed via Cohen's Kappa (acceptance threshold $\kappa \geq 0.70$); persistent discrepancies are resolved by a third arbitrating reviewer. Data extraction uses a standardized form, piloted on five studies (September 11–15, 2026), by two independent reviewers extracting in duplicate, with discrepancies resolved by consensus. When data required for meta-analysis are missing, corresponding authors are contacted by email with a 21-day response window. For studies with multiple publications from the same sample, data are extracted once from the most complete publication. The complete process is reported via a PRISMA 2020 flow diagram.

Quality assessment / Risk of bias analysis Risk of bias is assessed by two independent reviewers using tools selected by study design: RoB 2 (Cochrane) for randomized controlled trials; ROBINS-I (Cochrane) for non-randomized studies with a comparator; CASP Qualitative Checklist for qualitative studies; MMAT (2018 version) for mixed-methods studies; and the Design and Evidence Standards for Single-Case Designs, What Works Clearinghouse (2022) Standards Handbook, for single-case designs. Prior systematic reviews are assessed with AMSTAR 2 solely for descriptive purposes regarding the state of the prior synthesis field; this does not constitute their inclusion as an analytic unit. Assessment uses Rayyan QCRI's Risk of Bias Tool in blinded mode, distinct from the screening module.

Strategy of data synthesis Structured narrative synthesis, following Popay et al. (2006), is the central integration method for the entire corpus, organized by outcome, comparator type, and hybrid configuration, using evidence summary tables. Meta-analysis proceeds only when $k \geq 5$ studies share a comparator, the outcome is standardizable or convertible to a common metric, and sufficient data are available; when $k < 5$, the corresponding objective is addressed narratively or, for Objective 3, qualitatively. A random-effects model is adopted given expected clinical and methodological heterogeneity; Hedges' g (continuous outcomes) and odds ratios (dichotomous outcomes) are calculated with 95% CIs. Heterogeneity variance (τ^2) is estimated via REML, with Hartung-Knapp-Sidik-Jonkman (HKSJ) adjustment for pooled-effect confidence intervals; DerSimonian-Laird is reported as sensitivity analysis. When $k \in \{5,6\}$, results are accompanied by prediction intervals. Analyses use R with the metafor package (Viechtbauer, 2010). Quantitative and qualitative evidence are integrated via joint displays (Guetterman et al., 2015), mapping effect sizes and temporal-efficiency metrics against analytic themes and Latin American structural conditions.

Subgroup analysis Predefined subgroup analyses examine sociogeographic context (Latin America versus other regions, per UNESCO-IESALC classification), hybrid configuration type (draft-review, detection-elaboration, parallel-synthesis, real-time assistance), comparator type, and implementation characteristics, conducted where study volume permits. When the Latin American subgroup yields $k < 5$, moderation is addressed through qualitative, descriptive, cross-study synthesis integrated within the joint displays rather than formal subgroup statistics.

Sensitivity analysis Sensitivity analyses will compare REML and Hartung-Knapp-Sidik-Jonkman estimates against conventional DerSimonian-Laird estimates. Further analyses will exclude studies assessed as having a high risk of bias and will evaluate results according to the availability of complete versus imputed data. When $k \in \{5, 6\}$, results will be interpreted with strict caution and accompanied by prediction intervals, given the high imprecision inherent in τ^2 estimation with small samples.

Language restriction The review includes exclusively Spanish and English studies. Portuguese and Mandarin are excluded due to linguistic constraints and database access barriers, constituting a formally declared coverage limitation.

Country(ies) involved Chile (review team). No geographic restrictions apply to the eligibility of primary studies; the country or region of origin will be extracted and incorporated as a moderating variable.

Other relevant information
Review not yet started.
Registration: INPLASY, registration number [INSERT INPLASY REGISTRATION NUMBER].
Review start date: 10-08-2026.
Anticipated completion date: 15-12-2026.
Any protocol amendment will be dated, justified and registered.

Keywords Human-AI hybrid feedback; generative artificial intelligence; reciprocal peer observation; dialogic feedback; higher education; Latin America.

Dissemination plans The final manuscript will be submitted for open-access publication in an evidence-synthesis or education sciences journal. Findings will additionally be disseminated through presentations at academic congresses, scientific seminars, and specialized institutional lectures. Synthesis data and analytic materials will be deposited in an open data repository, such as the Open Science Framework, to ensure reproducibility. The complete protocol will remain available as supplementary material appended to the INPLASY registration.

Contributions of each author
Author 1 - Jorge Cuevas Aburto. - Corresponding author; led conception, design, and methodology; coordinated the review team; managed INPLASY registration; developed protocol, eligibility criteria, search, extraction, and synthesis plans; contributed to screening, full-text review, extraction, and verification; supervised risk-of-bias assessment, synthesis, and meta-analysis; drafted, revised, coordinated feedback, managed correspondence, and ensured integrity transparency, and process accountability.
Email: jorgecuevas@santotomas.cl
Author 2 - Karina Fuentes Rifo - Third arbitrating reviewer for unresolved discrepancies; methodological oversight and support in consensus decisions; doctoral supervisor of the corresponding author; academic guidance on the review design and manuscript development; support in author correspondence when required; critical manuscript revision; and approval of the final version.
Email: kafuente@udec.cl
Author 3 - Nicolás Tamarín - Contribution: Independent co-reviewer; title, abstract, and full-text screening using Rayyan QCRI; independent selection of eligible papers; data extraction and verification; risk-of-bias assessment; participation in discrepancy resolution; methodological and theoretical input on GenAI-mediated feedback, pedagogical support programs, temporal efficiency, and sociogeographic moderation; contribution to evidence synthesis, manuscript drafting, critical revision.
Email: nicotamarin2026@udec.cl