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

Cuevas-Aburto, JA; Fuentes-Rifo, KA; Tamarín-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 2 September 2026.

INTRODUCTION

Review question / Objective Objectives : This review pursues three interrelated objectives. First, to synthesise the available empirical evidence on the effect of hybrid human-generative AI feedback systems on the dialogic quality of faculty-to-faculty feedback within formal reciprocal peer observation programmes in higher education, understood as the umbrella category defined in section Condition being studied below. Second, to synthesise the available empirical evidence on the effect of hybrid human-generative AI feedback systems on evaluator temporal efficiency, operationalised separately as objective time required per feedback cycle, number of iterations required to produce final feedback, and perceived evaluator workload. Third, to examine whether the magnitude or direction of observed effects varies according to sociogeographic context, particularly Latin America, and according to prespecified implementation characteristics, including hybrid configuration type, number of feedback cycles (k = 1 versus k ≥ 2), comparator type, GenAI model or version, and degree of human oversight.

PICOT-M research question

Among in-service university faculty participating in formal reciprocal peer observation (RPO) programmes that include a structured teaching-observation or teaching-review process followed by faculty-to-faculty feedback (P), are hybrid human–generative AI feedback systems (I), compared primarily with human-only feedback within otherwise comparable formal pedagogical-support programmes (C), associated with improvements in the dialogic quality of feedback and/or evaluator temporal efficiency (O), and do these effects vary according to sociogeographic and implementation context, particularly in Latin America (M), among studies published from 1 January 2018, and 2 September 2026 (T)? For causal interpretation, conclusions will be restricted

to evidence from randomised and eligible quasi-experimental studies.

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 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, supported, or mediated by hybrid human and generative AI systems. In this context, formal programs are operationalized as institutionally documented structures comprising one or more complete feedback cycles. Each cycle must include teaching observation or review followed by a systematized and verifiable feedback mechanism. This definition allows single-cycle studies to be included when they provide sufficient documentation of the observation or review process and its associated feedback procedure, while still excluding informal, spontaneous, or undocumented exchanges. The number of complete cycles reported in each study, k=1 versus k≥2, will be extracted as a data item and examined as a potential moderating variable in subgroup and sensitivity analyses, rather than being used as an eligibility threshold. This approach preserves the sensitivity of the review, reduces the risk of excluding informative studies, and allows variation in program intensity to be analysed empirically.

Eligible modalities include reciprocal peer observation, peer review of teaching, institutional pedagogical coaching, and faculty support structures with an explicit evaluative component. The condition of interest is twofold. First, the review examines the dialogic quality of peer feedback, operationalized across five subdimensions, specificity, personalization, knowledge-building markers, orientation toward self-regulation, and dialogic tone. These dimensions address the documented tendency of unsupported human feedback to remain generic, weakly personalized, or insufficiently dialogic. Second, the review examines the temporal efficiency of faculty evaluators, including time per cycle, number of iterations, and perceived workload. This dimension responds to the structural time and scaffolding constraints that limit the scalability of high-quality feedback when it depends exclusively on human evaluators.

Both dimensions are examined in relation to in-service university faculty acting as feedback providers, feedback recipients, or both. This focus is necessary because the existing literature on generative AI-mediated feedback has concentrated mainly on students, leaving faculty-to-faculty pedagogical feedback comparatively underexamined. Accordingly, the review centres on formal institutional feedback practices in higher education and assesses whether hybrid human and generative AI systems can improve feedback quality, reduce evaluator workload, or reshape the dialogic conditions under which university teachers engage in pedagogical improvement.

METHODS

Search strategy A comprehensive search will be conducted in ERIC and Education Source via EBSCO, APA PsycINFO via APA PsycNet, Scopus, Web of Science, ACM Digital Library, IEEE Xplore, Redalyc, SciELO, ProQuest Dissertations & Theses Global, EdArXiv, OSF Preprints, SSRN, and Google Scholar. For Google Scholar, the first 200 relevance-ranked records will be retrieved using Publish or Perish and exported for deduplication and eligibility screening.

Guided by the prespecified PICO-TM framework, the strategy will comprise three conceptual blocks combined with AND: (1) Technology/Intervention, (2) Programme/Pedagogical Context, and (3) Higher-Education Faculty/Setting. Within each block, database-specific controlled vocabulary, where available, and English- and Spanish-language free-text terms, synonyms, spelling variants, and appropriate truncation or proximity operators will be combined using OR. Each strategy will then be adapted to the source-specific syntax, indexing conventions, and field structure, including TITLE-ABS-KEY in Scopus and the Topic field (TS) in Web of Science. Eligible studies will be published in English or Spanish between 1 January 2018 and 2 September 2026.

Before definitive execution, an independent reviewer will assess the master strategy and its source-specific adaptations using the PRESS framework, while the research team will test each strategy against a prespecified Known-Item and Boundary-Study Search Validation Set. If an indexed or otherwise discoverable validation record is not retrieved, the relevant blocks and syntax will be systematically reviewed, with any justified modifications documented and retested. However, retrieval will not establish eligibility, which will be determined independently against the preregistered criteria. Searches will be conducted and reported according to PRISMA-S and applicable Cochrane MECIR standards, with the complete strategies, search dates, limits, and retrieval yields retained.

Retrieved records will first be exported to Mendeley and then imported into Rayyan, where automated deduplication will be supplemented by manual verification of records with discrepant metadata or indexing. Subsequently, two reviewers will independently screen titles and abstracts, followed by potentially eligible full texts, using Rayyan's blind mode. Before formal screening, both reviewers will independently assess a prespecified calibration sample, with agreement quantified using Cohen's κ. Disagreements will be resolved through discussion and consensus or, when necessary, adjudication by a third reviewer. Finally, all calibration decisions and operational clarifications will be documented to preserve a transparent audit trail without changing the substantive scope of the preregistered criteria.

Participant or population Participant or population: The broader population of interest comprises in-service university faculty working in higher education institutions. Within this population, eligible participants are defined as faculty members who actively engage as peer evaluators, feedback providers, and/or observees, feedback recipients, in formal pedagogical support programs. For the purposes of this review, these programs are operationalized as institutional initiatives with a documented structure that includes at least one complete cycle of teaching observation or review, followed by a systematized and verifiable faculty-to-faculty feedback mechanism. This criterion is intended to ensure that included studies examine structured pedagogical feedback processes, while avoiding the unnecessary exclusion of relevant evidence from programs that report a single documented cycle. Accordingly, informal, spontaneous, or undocumented exchanges will be excluded, as they do not provide sufficient methodological traceability for analysis. Eligible modalities include reciprocal peer observation, peer review of teaching, institutional pedagogical coaching, and faculty support structures with an explicit evaluative feedback component. Rather than being used as an eligibility threshold, the number of complete cycles documented in each study will be extracted as a data item and explored as a potential moderator of effect size. This approach preserves the sensitivity of the review, reduces the risk of excluding informative studies, and allows variation in program intensity to be examined analytically. Studies centered on students, preschool, primary or secondary educators, pre-service teachers, or professional development programs lacking a formal feedback component will be excluded.

Intervention The eligible intervention is any hybrid human and generative artificial intelligence (GenAI) system implemented within a formal reciprocal peer observation (RPO) programme in higher education. Here, RPO is an umbrella term encompassing structured teaching observation, peer review of teaching, pedagogical coaching, and other documented evaluative feedback processes. Eligible interventions include any hybrid workflow in which human agency and GenAI jointly contribute to feedback production. GenAI Eligibility will be determined by the functional role of GenAI in producing or substantively supporting the feedback delivered to faculty members. Four eligible configurations form a descriptive taxonomy rather than a closed eligibility framework: (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 human evaluator develops into feedback; (c) parallel-synthesis, in which AI and the human evaluator independently generate feedback that is subsequently integrated; and (d) real-time

assistance, in which AI provides live, adaptive scaffolding during observation.
Other documented hybrid workflows are also eligible, provided that human agency and GenAI jointly contribute to feedback production. Systems combining generative and non-generative components are eligible only when the generative component produces or substantively supports the textual feedback delivered to the faculty member. This hybrid feature will be coded as an intervention variable for heterogeneity analyses. Finally, included studies must explicitly report the GenAI technology and model used, including the version where available.

Comparator The primary comparator is faculty-to-faculty feedback produced exclusively by human evaluators within a formal pedagogical support programme, thereby providing the hybrid versus human-only contrast. As the principal comparison, it is the only contrast eligible for pooling in the quantitative synthesis. In addition, two comparators will be treated as secondary and exploratory, analysed separately from the primary contrast and never pooled with it: (1) feedback generated exclusively by a GenAI system without human involvement in content elaboration, although non-editorial oversight, such as selecting the observation submitted, may occur; and (2) the absence of a formal support programme for in-service faculty in higher education. Similarly, comparisons between hybrid systems and non-generative AI, such as rule-based automated scoring systems, will constitute a further exploratory secondary comparator. However, these comparisons will be analysed independently of both the primary and GenAI-only contrasts. Where a study includes multiple comparators, each will be analysed separately, while preserving the original unit of randomisation or comparison and applying the procedures for dependent effect sizes specified under Data Management.

Study designs to be included Eligible designs include randomised controlled trials (RCTs); quasi-experimental studies with an equivalent control group or statistical adjustment; longitudinal observational studies with between-group comparisons; mixed-methods studies with an evaluable primary quantitative component; and rigorous qualitative studies. Single-case designs will be analysed separately as exploratory evidence. By contrast, prior systematic reviews are used exclusively for theoretical comparison and are not treated as primary units of analysis. Following the initial eligibility assessment of the indexed search results, backward and forward citation searching will be conducted for all included studies and, where relevant, for key reports identified during full-text assessment. To ensure transparency and reproducibility, the search log will record the source used for forward citation tracking, the date of each citation search, the records identified, and those subsequently assessed for eligibility.
Randomised, quasi-experimental, and eligible longitudinal comparative studies will primarily inform the effectiveness questions addressed in Objectives 1 and 2. Meanwhile, qualitative and qualitative-dominant mixed-methods evidence will inform the implementation, acceptability, contextual mechanisms, and sociogeographic interpretation addressed in Objective 3. Where sufficient comparable studies are available, this evidence will complement the formal quantitative moderator analyses described under Subgroup Analysis; however, qualitative evidence will not be treated as an independent estimator of intervention effects.
Because this review addresses both intervention effectiveness and implementation context, eligible quantitative, qualitative, and mixed-methods primary studies will be included. Nevertheless, the review remains a systematic review rather than a mixed studies review because quantitative and qualitative evidence will be synthesised separately and integrated only during the final interpretive stage.

Eligibility criteria Eligibility criteria: Included: studies examining in-service university faculty within formal pedagogical support programs comprising one or more 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. Single-cycle studies are eligible where at least one complete cycle of observation or review of teaching performance followed by a systematized, verifiable feedback mechanism is documented. The number of complete cycles reported in each study, $k=1$ versus $k\geq 2$, will be extracted as a data item and explored as a potential moderating variable in subgroup and sensitivity analyses, rather than applied as an eligibility threshold. 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 under-representation in Anglophone indexing.

Main outcome(s) The primary outcome domain is dialogic feedback quality, comprising five prespecified components: content specificity, personalisation to the observed practice, knowledge-building markers, orientation towards self-regulation, and dialogic tone. Each component will be analysed as a separate outcome unless a study reports a validated multidimensional instrument supporting total-score interpretation. Accordingly, no composite score will be created solely for quantitative synthesis.

Given the absence of a standardised instrument for AI-mediated peer observation, these components will be operationalised using a low-inference qualitative codebook based on directed qualitative content analysis (Hsieh & Shannon, 2005). Inter-rater reliability will be assessed using Cohen's Kappa and the codebook will be piloted on a sample of textual feedback materials before full implementation, and this will incorporate documented positive and negative examples, together with decision rules for ambiguous cases.

Dialogic-quality data will subsequently be classified by source as either (a) outcomes reported by study authors using validated or study-specific instruments or (b) outcomes coded directly by the review team from primary feedback texts, where sufficient information is available. These sources will be analysed separately and combined into a single effect estimate only when conceptual equivalence and measurement comparability can be justified a priori. Likewise, they will not be merged into a composite score without explicit justification and separate reporting. Finally, this outcome will be synthesised primarily through narrative methods. Meta-analysis will be considered only when studies are sufficiently comparable in terms of population, intervention, comparator, outcome construct, study design, follow-up period, and available quantitative data. Thus, $k \geq 5$ is treated as a necessary but not sufficient prudential threshold; pooling additionally depends on the availability of a sufficient body of conceptually, methodologically, and statistically comparable evidence, as specified under Strategy of data synthesis.

Additional outcome(s) Temporal efficiency is retained as a secondary domain comprising three conceptually and statistically distinct outcome measures, analysed separately rather than pooled into a single construct: (i) objective time per feedback cycle (continuous, e.g., minutes); (ii) number of iterations per cycle (count data); and (iii) perceived workload (ordinal/Likert or continuous self-report, depending on instrument). Each measure will be synthesised using the effect statistic appropriate to its measurement scale. Additional secondary outcomes, where available, include recipients' trust in the feedback, observable changes in pedagogical practice, faculty perceptions of agency and autonomy, and indicators of over-reliance on the GenAI system.

Data management All retrieved records will be imported into Rayyan QCRI (web and mobile, v. 1.7.7). Automated duplicate detection will be supplemented by manual verification of uncertain matches, and duplicates will be reported in the PRISMA 2020 flow diagram. Thereafter, both reviewers will independently assess an approximately 10% proportionate, source-stratified calibration sample, after which all sampled records will be returned to the definitive screening pool. Cohen's κ will be reported descriptively, while identified discrepancies will inform a decision-rules manual finalised before two-stage independent duplicate screening begins. Throughout screening, Blind Mode will remain enabled until disagreements are resolved through consensus or, where necessary, adjudicated by a third reviewer.

Following completion of the study-selection process, critical outcomes will be extracted independently and in duplicate using a standardised form piloted on five studies. Where relevant data are unavailable, the review team will contact the corresponding authors and allow 21 days for a response. The study will constitute the unit of analysis; accordingly, multiple reports will

be linked, while dependencies among effect sizes will be addressed using robust variance estimation or a three-level meta-analytic model. Finally, records identified through citation searching will undergo the same screening process and be reported in the PRISMA 2020 flow diagram.

Quality assessment / Risk of bias analysis Two independent reviewers will assess risk of bias using design-specific tools: RoB 2 for randomised controlled trials; ROBINS-I for non-randomised comparative studies; the CASP Qualitative Checklist for qualitative studies; MMAT (2018) for mixed-methods studies, without calculating an overall score, in accordance with its user guide; and the What Works Clearinghouse Design and Evidence Standards for Single-Case Designs (2022) for single-case studies.
Subsequently, certainty of evidence for each priority outcome, namely dialogic feedback quality, objective time per cycle, number of iterations, and perceived workload, will be assessed using GRADE across five domains: risk of bias, inconsistency, indirectness, imprecision, and publication bias. Findings will be presented in GRADE Summary of Findings tables. In parallel, confidence in qualitative findings contributing to the narrative and joint-display syntheses will be assessed using GRADE-CERQual across its four components: methodological limitations, coherence, adequacy of data, and relevance. Finally, beyond assessing missing data within individual studies, the review will separately evaluate risk of bias arising from missing results across the overall body of evidence.

Strategy of data synthesis Meta-analysis will be conducted only when at least five studies ($k \geq 5$) share the primary comparator and are conceptually and statistically comparable, as determined by similar PICOT-M profiles, plausible mechanisms and directions of effect, and heterogeneity estimates (τ^2 and I^2) that do not suggest implausible pooling. However, $k \geq 5$ is a necessary rather than sufficient prudential threshold. Where these conditions are not met, the relevant objective will be addressed narratively or, for Objective 3, through structured descriptive and qualitative cross-study synthesis.
Random-effects models will be used for all pooled analyses. Effect measures will be selected by outcome type: Hedges' g for continuous outcomes, including dialogic quality subdimensions and objective time per feedback cycle; rate ratios for count outcomes, including the number of iterations; and odds ratios for dichotomous outcomes. Between-study variance (τ^2) will be estimated using restricted maximum likelihood (REML), with the Hartung-Knapp-Sidik-Jonkman (HKSJ) adjustment applied to pooled-effect confidence intervals. When $k = 5$ or 6, prediction intervals will also be reported because of the substantial imprecision in τ^2 estimation with small samples. Analyses will be conducted in R using the metafor package (Viechtbauer, 2010).
Finally, quantitative findings and qualitative evidence will be synthesised independently using methods appropriate to each evidence type and integrated only at the interpretive stage through joint displays. Qualitative findings will provide explanatory and contextual interpretation of the quantitative evidence and will not contribute to pooled effect estimates.

Subgroup analysis Prespecified subgroup analyses examine sociogeographic context, hybrid configuration type, comparator type, GenAI model/version, degree of human oversight, and number of feedback cycles ($k = 1$ versus $k \geq 2$), where sufficient information is available. Crucially, subgroup definitions and classification rules will be prespecified before analysis. Furthermore, when quantitative subgroup comparisons are undertaken, differences between subgroups will be evaluated using an appropriate formal test of interaction or subgroup differences. However, differences in statistical significance within individual subgroups will not, by themselves, be interpreted as evidence of effect modification. Conversely, when the available evidence is insufficient to support a stable quantitative subgroup comparison, the moderating role of sociogeographic context will be examined through structured descriptive and qualitative cross-study synthesis. In all cases, quantitative subgroup comparisons will be undertaken only when the number, comparability, and precision of the contributing studies permit meaningful estimation.

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
Operational review start date: 2 September 2026.
Anticipated completion date: 15 December 2026.

Any protocol amendment will be dated, justified, registered in INPLASY, and documented in an amendment log.

Protocol amendment (v2.0, 14 August 2026): The protocol and draft search strategy were refined following methodological review. These refinements concerned the operational definition and reporting of the review methods and the development of the search strategy. No pilot screening was undertaken before this amendment. No records from a definitive review search had been imported, processed, screened, assessed for eligibility, or used in a screening-calibration exercise.

Protocol amendment (v3.0, 2 September 2026): The operational review start date was revised to 2 September 2026. The search strategy was further refined and specified as a three-block framework combining: (1) GenAI/LLMs and hybrid human–AI feedback; (2) formal faculty peer-feedback, reciprocal peer observation, mentoring, coaching, and related institutional pedagogical-feedback programmes; and (3) higher-education faculty populations and settings. Source-specific adaptations, controlled vocabulary and free-text terms, field codes, Boolean and proximity logic, truncation, and limits were specified for each information source.

The amendment also prespecified independent PRESS peer review of the master strategy and its database-specific adaptations; Known-Item and Boundary-Study validation before the definitive searches; separate documentation of any exploratory terminology-development or validation searches; backward and forward citation searching; and transparent reporting and archiving of complete executable strategies, search dates, limits, retrieval yields, validation records, PRESS materials, and the search audit trail in the Open Science Framework.

Procedures for record management and study selection were clarified, including export to Mendeley, import into Rayyan, automated deduplication supplemented by manual verification, independent duplicate title/abstract and full-text screening in blind mode, a prespecified screening-calibration sample with Cohen's κ, and resolution of disagreements through discussion, consensus, or third-reviewer adjudication. The analysis plan was also clarified before review conduct.

These amendments did not alter the registered title, review question, PICOT-M framework, core eligibility logic, comparator set, outcome domains, or underlying purpose of the review. They strengthened operational clarity, transparency, reproducibility, and methodological rigour. At the time v3.0 was finalised, no definitive searches had been executed and no records had entered the review workflow for deduplication, screening, eligibility assessment, data extraction, risk-of-bias assessment, or synthesis.

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; review coordination; INPLASY registration; protocol development; management of author correspondence; and final manuscript correspondence.

Author 2 – Karina Andrea Fuentes-Rifo: Methodological supervision; third-reviewer adjudication of unresolved disagreements; guidance on review design; critical manuscript revision; and approval of the final version.

Author 3 – Nicolás Andrés Tamarín-Tanarín: Independent title/abstract and full-text screening; study selection; data extraction and verification; risk-of-bias assessment; participation in discrepancy resolution; and contribution to manuscript development.

Author 4 – Rafael Zapata-Larmana. Contribution: methodological adviser on PRISMA-S, PRISMA-P, MECIR, and Cochrane Handbook compliance; led the PRESS review of the master search strategy and the correction of database-specific search syntax; contributed to revision of the comparator hierarchy and outcome disaggregation; reviewed and approved the final version of the amended protocol.

Author 5- Alejandra Robre-Campos. Contribution: methodological adviser on PRISMA-S, PRISMA-P, MECIR, and Cochrane Handbook compliance; contributed to revision of the risk-of-bias and certainty-of-evidence plan (GRADE/GRADE-CERQual) and the data-management procedures; conducted general critical revision of the amended manuscript. Email: alejandrarobles@udec.cl

Author 6- Carolina Fernández-Chávez. Contribution: Conducted the PRESS search strategy evaluation and validated syntax expansions to maximise retrieval sensitivity. Contributed to qualitative codebook calibration and the classification of hybrid intervention configurations. Email: carofernandezc@udec.cl
All authors will approve the final manuscript and are accountable for their respective contributions.

Affiliations

Author 1 - Jorge Artidoro Cuevas Aburto
Email: jorgecuevas@santotomas.cl
Affiliation: Learning Center, National Undergraduate Studies Office, Vice-Rectory for Academic Affairs, Research and Postgraduate Studies, Universidad Santo Tomás, Los Ángeles campus, Chile; Doctoral Program in Education, Universidad de Concepción, Concepción, Chile.
Author 2: Nicolás Andrés Tamarín Tanarín. Affiliation: Universidad del Bío-Bío, Chillán, Chile.
Author 3: Karina Fuentes-Rifo, Faculty of Education, Universidad de Concepción, Concepción, Chile.
Author 4: Rafael Zapata-Lermana,. Afilation: School of Kinesiology, Faculty of Health, Universidad Santo Tomás, Los Ángeles, Chile; School of Education, Universidad de Concepción, Los Ángeles, Chile.
Author 5: Alejandra Robles-Campos. Afiliation: School of Education, Universidad de Concepción, Los Ángeles, Chile.
Author 6: Carolina Fernández-Chávez Afiliation: Department of Educational Sciences, Faculty of Education, Universidad de Concepción, Concepción, Chile.