



Sistema de intervención didáctica con inteligencia artificial y realidad virtual para contribuir a la mejora del desempeño en competencias transversales de comunicación y solución de problemas mediante ciclos de práctica situada, retroalimentación criterial y reflexión en estudiantes de pregrado del Tecnológico de Monterrey, México, durante el periodo 2024 a 2026

TESIS DOCTORAL

que, para obtener el Grado de Ph.D

DOCTOR EN EDUCACIÓN E INNOVACIÓN

PRESENTA

Roberto Gómez Tobías

ASESOR

Dr. Agustín Villarreal

México, 2026

La presente Tesis Doctoral debe ser citada como:

Gómez Tobías, R. (2026). *Sistema de intervención didáctica con inteligencia artificial y realidad virtual para contribuir a la mejora del desempeño en competencias transversales de comunicación y solución de problemas mediante ciclos de práctica situada, retroalimentación criterial y reflexión en estudiantes de pregrado del Tecnológico de Monterrey, México, durante el periodo 2024 a 2026* [Tesis doctoral, Universidad de Investigación e Innovación de México]. Repositorio Institucional UIIX.

Esta tesis cuenta con un paquete reproducible depositado en *Open Science Framework* (OSF), que incluye: (i) *dataset* cuantitativo desidentificado y archivos analíticos; (ii) documentación técnica (*codebook* e instrumentos); y (iii) código de análisis (*scripts/notebooks*) para reproducir tablas y figuras reportadas en la tesis. El repositorio se encuentra disponible mediante DOI: <https://doi.org/10.17605/OSF.IO/56GPE>



Esta obra está bajo una Licencia Creative Commons Atribución-NoComercial-SinDerivar 4.0 Internacional

Se permite la reproducción y distribución total o parcial, siempre que se reconozca la autoría y la fuente. No se permite el uso comercial ni la creación de obras derivadas.

RESUMEN

La educación superior enfrenta el reto de desarrollar competencias transferibles y demostrar su logro mediante evidencias objetivas en contextos auténticos. Entre ellas destacan la Comunicación y la Solución de problemas/toma de decisiones, cuyo desarrollo exige práctica situada, retroalimentación formativa y evaluación criterial. Este estudio evaluó el efecto de una arquitectura didáctica que integra retroalimentación asistida por inteligencia artificial (IA) y simulaciones en realidad virtual (RV), organizada en ciclos de *briefing*, simulación y *debriefing*, sobre el desempeño en dichas competencias bajo condiciones reales de aula. Se implementó un diseño cuasi-experimental por conglomerados con métodos mixtos en el Tecnológico de Monterrey durante 2024-2025, con seis clases en condición IA+RV y condición de comparación ($N = 180$). Con base en un criterio preespecificado de asistencia mínima $\geq 75\%$, se excluyeron 20 estudiantes; el análisis principal se realizó con $n = 160$. Los resultados se midieron mediante rúbricas analíticas con doble calificación ciega; la fiabilidad interjueces fue alta ($ICC [2, k] = 0.788-0.855$) y la validez de contenido fue máxima ($S-CVI/Ave = 1.00$). El efecto se estimó mediante modelos lineales mixtos con interacción Tiempo $\times$ Grupo y tamaños de efecto d_{ppc2} con IC95%. Los resultados mostraron mejoras diferenciales a favor de IA+RV en Comunicación y Solución de problemas, con tamaños de efecto muy grandes. A partir de estos hallazgos, la tesis propone una arquitectura IA+RV replicable, auditable y escalable bajo condiciones de calidad, seguridad, equidad, trazabilidad y gobernanza. En conclusión, la participación en la arquitectura IA+RV se asoció con mejoras sustantivas en desempeño observable.

Palabras clave: inteligencia artificial; realidad virtual; retroalimentación formativa; comunicación; solución de problemas.

ABSTRACT

Higher education faces the challenge of developing transferable competencies and demonstrating their attainment through objective evidence in authentic contexts. Among these competencies, Communication and Problem-solving/Decision-making require situated practice, formative feedback, and criterion-referenced assessment. This study evaluated the effect of an instructional architecture that integrates artificial intelligence (AI)-assisted feedback and virtual reality (VR) simulations, organized through briefing, simulation, and *debriefing* cycles, on student performance in these competencies under real classroom conditions. A cluster quasi-experimental mixed-methods design was implemented at Tecnológico de Monterrey during 2024–2025, with six classes assigned to either the AI+VR condition or the comparison condition ($N = 180$). Based on a prespecified minimum attendance criterion of $\geq 75\%$, 20 students were excluded; the primary analysis was conducted with $n = 160$. Outcomes were assessed using analytic rubrics with blinded double scoring; interrater reliability was high ($ICC [2, k] = 0.788–0.855$), and content validity was maximal ($S-CVI/Ave = 1.00$). Effects were estimated through linear mixed models with a Time $\times$ Group interaction and d_{ppc2} effect sizes with 95% confidence intervals. Results showed differential gains favoring the AI+VR condition in Communication and Problem-solving, with very large effect sizes. Based on these findings, the dissertation proposes a replicable, auditable, and scalable AI+VR architecture under conditions of quality, safety, equity, traceability, and governance. In conclusion, participation in the AI+VR architecture was associated with substantial improvements in observable performance.

Keywords: artificial intelligence; virtual reality; formative feedback; communication; problem solving.

AGRADECIMIENTOS

A la Universidad de Investigación e Innovación de México (UIIX), por brindarme la oportunidad de cursar el Doctorado en Educación e Innovación y por proporcionarme herramientas clave para mi desarrollo académico. Expreso un reconocimiento particular al Dr. Cristian F. Torres Salvador, director académico de posgrado, a la Dra. Odalys Peñate López, directora de la unidad de investigación de posgrado y a la Dra. Lyzzi Coromoto Davalillo Bolívar, coordinadora de investigación, por su acompañamiento institucional y apoyo durante el proceso.

Al Tecnológico de Monterrey, por facilitar la implementación de las intervenciones desarrolladas en el marco de esta investigación, mediante el acceso a instalaciones de primer nivel y a equipos de realidad virtual indispensables para la ejecución del estudio. Asimismo, agradezco de manera especial al Summit AI del CEDDIE por su apoyo a través de la provisión de la licencia utilizada.

A mis colegas de las escuelas de negocios e ingeniería y ciencias, cuyo respaldo constante durante las fases de intervención y evaluación contribuyó de manera significativa al fortalecimiento metodológico del estudio y a la calidad de los resultados obtenidos.

A los alumnos participantes, cuyo entusiasmo y compromiso hicieron posible la realización de esta investigación, aportando su tiempo, esfuerzo y valiosas perspectivas al proceso.

A mi esposa, Claudia, por su inquebrantable apoyo, paciencia y aliento constante, siendo un pilar fundamental durante este arduo pero gratificante camino académico.

Finalmente, a mi asesor de tesis, Dr. Agustín Villarreal, por su guía experta, retroalimentación invaluable y motivación continua, elementos esenciales para la culminación exitosa de este proyecto doctoral.

A todos ustedes, mi más sincero agradecimiento.

DEDICATORIA

A Dios, por ser mi guía, mi fuerza y mi luz, brindándome sabiduría y fortaleza en cada paso de este recorrido.

A mi esposa Claudia, por su amor, paciencia y apoyo incondicional en cada paso de este camino, por ser mi fuente constante de inspiración y fortaleza.

A mis hijos, Alejandra y Roberto, por ser mi mayor motivación, y recordarme cada día la importancia de perseguir los sueños con dedicación y amor.

A mis padres, por inculcarme el valor del conocimiento y el esfuerzo, y por su apoyo inquebrantable a lo largo de mi vida.

A todos aquellos que, con su ejemplo y guía, han dejado una huella imborrable en mi formación académica y personal.

A ustedes, dedico con gratitud y cariño el fruto de este esfuerzo.

TABLA DE CONTENIDOS

INTRODUCCIÓN.....	33
CAPÍTULO 1. PROTOCOLO DE INVESTIGACIÓN.....	37
1.1 Tema de investigación.....	37
1.2 Planteamiento del problema.....	38
1.3 Pregunta de investigación.....	40
1.4 Justificación.....	41
1.5 Objeto de estudio.....	41
1.6 Campo de acción.....	43
1.7 Objetivos.....	45
1.7.1 Objetivo general.....	45
1.7.2 Objetivos específicos.....	45
1.8 Hipótesis.....	46
1.9 Alcance y limitaciones.....	46
1.9.1 Alcances del estudio.....	46
1.9.2 Limitaciones del estudio.....	47
1.10 Síntesis visual del protocolo.....	49
CAPÍTULO 2. FUNDAMENTO TEÓRICO.....	51
2.1 Estado del arte.....	51
2.2 Marco teórico.....	55
2.2.1 Constructivismo y aprendizaje experiencial.....	55
2.2.2 Carga cognitiva y aprendizaje multimedia.....	59
2.3 Marco conceptual.....	63
2.4 Marco histórico	66
2.5 Marco contextual y actual.....	70
2.6 Marco normativo y legal.....	74
CAPÍTULO 3. MARCO METODOLÓGICO, OPERACIONALIZACIÓN Y DISEÑO DE MEDICIÓN.....	79
3.1 Cuadro de operacionalización de variables.....	79
3.2 Diseño metodológico y estrategia de medición.....	81

3.2.1 Perspectiva epistemológica y lógica de inferencia.....	81
3.2.2 Enfoque, diseño y tipología de investigación.....	81
3.2.3 Estrategia de medición y temporalidad del diseño.....	82
3.2.4 Determinación del tamaño de muestra y efecto de diseño.....	83
3.2.5 Coherencia entre diseño, hipótesis y métrica de desempeño.....	85
3.3 Variables del estudio: independiente, interviniente y dependiente.....	86
3.3.1 Variable independiente.....	86
3.3.2 Variables intervinientes.....	87
3.3.3 Variables dependientes.....	89
3.3.4 Variables moderadoras y covariables para ajuste.....	90
3.3.5 Ligadura explícita entre variables e hipótesis.....	90
3.4 Métodos y técnicas de recolección de datos.....	91
3.4.1 Técnica principal: evaluación de desempeño mediante productos auténticos y rúbricas analíticas.....	92
3.4.2 Técnica de implementación: registros de dosificación y fidelidad de IA y RV.....	92
3.4.3 Técnica complementaria: autoinformes breves de proceso y experiencia de Aprendizaje.....	93
3.4.4 Técnicas cualitativas: entrevistas semiestructuradas y observación participante.....	93
3.4.5 Técnica documental y trazabilidad: bitácora analítica y control de decisiones de datos.....	94
3.4.6 Integración de técnicas y correspondencia con el sistema de variables.....	95
3.5 Instrumentos.....	96
3.5.1 Rúbricas analíticas de desempeño para variables dependientes.....	96
3.5.2 Instrumentos para variables intervinientes de implementación dosificación y fidelidad.....	97
3.5.3 Instrumentos breves de proceso y experiencia.....	98
3.5.4 Instrumentos cualitativos: guías de entrevista y protocolos de observación.....	99

3.5.5 Instrumentos de control de calidad y documentación técnica.....	99
3.6 Aseguramiento de calidad.....	100
3.7 Síntesis visual del capítulo metodológico.....	101
CAPÍTULO 4. ANÁLISIS DE RESULTADOS.....	103
4.1 Prueba Piloto.....	103
4.1.1 Ajustes previos al trabajo de campo y cronograma.....	103
4.1.2 Calibración de evaluadores y pilotaje de instrumentos.....	104
4.1.3 Prueba de campo y verificación operativa previa.....	105
4.2 Aplicación de Instrumentos.....	110
4.2.1 Método de aplicación y levantamiento de la información.....	111
4.2.2 Implementación de la intervención IA+RV: fidelidad, dosis y control ... de variabilidad operacional.....	115
4.2.3 Seguridad, contingencia y condiciones operativas para RV.....	121
4.2.4 Fidelidad de implementación y <i>quality gates</i> como evidencia de adopción controlada.....	123
4.3 Procesamiento y análisis de la Información.....	125
4.3.1 Gobernanza de datos, documentación y reglas de trazabilidad.....	126
4.3.2 Evidencia de implementación y trazabilidad del flujo de datos: de la ejecución presencial a la base analítica final.....	126
4.3.3 Propósito, estructura y anclaje del análisis.....	129
4.3.4 Criterios de preparación, estimación y robustez.....	131
4.3.5 Especificación de modelo.....	134
4.3.6 Base diagnóstica para la interpretación de resultados.....	138
4.3.7 Descriptivos, flujo de muestra y verificaciones previas.....	141
4.3.8 Aseguramiento de la calidad de los instrumentos.....	142
4.3.9 Calidad de medición.....	143
4.3.10 Pruebas de hipótesis y estimaciones principales.....	145
4.3.10.1 Modelos multinivel (efecto Tiempo × Grupo).....	145
4.3.10.2 Tamaños de efecto (diferencia-en-diferencias).....	147
4.3.11 Análisis complementarios y robustez.....	151
4.3.11.1 Dosis-respuesta y moderación (OLS).....	151

4.3.11.2 Fidelidad de implementación (adherencia).....	152
4.3.11.3 Contraste explícito de hipótesis (síntesis).....	154
4.4 Graficación.....	155
4.4.1 Ganancia PRE–POST por grupo.....	157
4.4.2 Distribuciones PRE–POST y <i>boxplots</i>	158
4.4.3 Gráficas radiales PRE-POST por criterio.....	160
4.4.4 Graficación complementaria de dosis–respuesta.....	162
4.5 Redacción de los resultados.....	164
4.5.1 Interpretación por hipótesis.....	165
4.5.2 Mecanismo plausible: fidelidad, dosificación y experiencia de aprendizaje.....	167
4.5.3 Comparación con la literatura.....	169
4.5.4 Límites y amenazas a validez.....	170
4.6 Integración de hallazgos empíricos, calidad de medición e implicaciones Inmediatas.....	136
CAPÍTULO 5. PROPUESTA DE TRANSFORMACIÓN.....	140
5.1 Denominación, fundamentación de la propuesta.....	176
5.2 Fundamentación teórico conceptual.....	176
5.3 Objetivo general de la propuesta.....	181
5.4 Objetivos específicos instruccionales de la propuesta.....	182
5.5 Representación teórica y práctica de la propuesta mediante esquema lógico..	183
5.6 Fases o etapas de implementación de la propuesta.....	186
5.7 Acciones y actividades vinculadas a las fases de implementación.....	189
5.8 Selección de métodos, técnicas e instrumentos para su aplicación.....	192
5.9 Recursos para el funcionamiento de la propuesta.....	196
5.10 Valoración de la pertinencia de la propuesta.....	201
5.10.1 Factibilidad técnica, operativa e institucional.....	203
5.10.2 Criterios de adopción y escalamiento responsable.....	204
5.10.3 Validación complementaria mediante Delphi modificado de dos rondas.....	205
5.10.3.1 Propósito y panel de expertos.....	206

5.10.3.2 Primera ronda: valoración inicial de la propuesta.....	207
5.10.3.3 Ajustes incorporados entre rondas.....	208
5.10.3.4 Segunda ronda: revaloración y convergencia.....	209
5.10.3.5 Síntesis interpretativa del proceso Delphi.....	210
5.11 Evaluación de la efectividad de la propuesta.....	211
CONCLUSIONES	217
RECOMENDACIONES.....	226
BIBLIOGRAFÍA.....	229
APÉNDICE A.....	248
APÉNDICE B.	277
APÉNDICE C.	289
ANEXO 1.....	303
ANEXO 2.....	304

ÍNDICE DE FIGURAS

Figura 1.1	Acceso a internet y disponibilidad de computadora en hogares con estudiantes por nivel educativo, México, 2024.....	39
Figura 1.2	Mapa conceptual del Protocolo de Investigación.....	50
Figura 2.1	Mapa de conexiones por citación (2020–2025).....	18
Figura 3.1	Síntesis del marco metodológico y ruta de análisis del estudio.....	102
Figura 4.1	Diagrama operativo del trabajo de campo para la construcción del diagnóstico.....	110
Figura 4.2	Implementación en aula de la intervención IA+RV durante una sesión de práctica inmersiva en el laboratorio de negocios internacionales.....	112
Figura 4.3	Lógica operacional de la intervención presencial IA+RV.....	115
Figura 4.4A	Panel ilustrativo de resultados y analítica automatizada en VirtualSpeech.....	118
Figura 4.4B	Mapa ilustrativo de habilidades practicadas y desglose de subcompetencias en VirtualSpeech.....	119
Figura 4.4C	Escenario ilustrativo de práctica guiada en VirtualSpeech.....	120
Figura 4.5	Árbol de decisión operativa ante incidentes en RV y activación de alternativa equivalente.....	123
Figura 4.6	Mapa de aseguramiento de calidad: de riesgos a evidencia verificable.....	125
Figura 4.7	Mapa del proceso de trazabilidad del estudio.....	129
Figura 4.8	Mapa metodológico y plan analítico.....	137
Figura 4.9	Pipeline de datos del diagnóstico: de evidencia primaria a base analítica.....	139
Figura 4.10	Diagrama del trabajo de campo y del análisis: responsabilidades, insumos y puntos de control.....	140
Figura 4.11	Diagrama tipo CONSORT de inclusión/exclusión y muestra analizada.....	156
Figura 4.12A	Ganancia PRE–POST por grupo, Comunicación.....	157

Figura 4.12B	Ganancia PRE–POST por grupo, Solución de problemas.....	158
Figura 4.12C	Distribución PRE–POST por grupo, Comunicación (0–100).....	159
Figura 4.12D	Distribución PRE–POST por grupo, Solución de problemas (0–100).....	160
Figura 4.13A	Perfil PRE–POST por criterio, Comunicación (rúbrica analítica, 6 criterios).....	161
Figura 4.13B	Perfil PRE–POST por criterio, Solución de problemas (rúbrica analítica, 7 criterios).....	162
Figura 4.14A	Scatter dosis–respuesta, Comunicación.....	163
Figura 4.14B	Scatter dosis–respuesta, Solución de problemas.....	164
Figura 4.15	Síntesis de resultados, calidad de medición y trazabilidad de implementación.....	175
Figura 5.1	Modelo lógico de la arquitectura IA+RV: insumos, procesos y resultados.....	186
Figura 5.2	Estrategia de implementación por fases: mínimo viable, consolidación, expansión controlada e integración institucional....	189
Figura 5.3	Escenario de uso de RV como entorno de desempeño en la arquitectura didáctica IA+RV.....	194
Figura 5.4	Sistema mínimo viable para sostener y escalar la arquitectura IA+RV.....	198
Figura 5.5	Síntesis del Capítulo 5: de la evidencia empírica a una arquitectura IA+RV auditable, validada y escalable.....	216
Figura A1.1	Mapa de instrumentos, control de calidad y salvaguardas operativas.....	250
Figura A7.1	Gobernanza operativa y responsabilidades RACI en el ciclo de datos del estudio.....	269
Figura A7.2	Quality gates para continuidad y escalamiento del modelo IA y RV.....	270
Figura A7.3	Diagrama de respuesta ante incidente durante práctica en Realidad Virtual y activación de alternativa equivalente.....	275

Figura A7.4	Ciclo operativo pre–durante–post para estandarización de sesiones con RV.....	276
Figura B2.1	Estructura mínima del proyecto reproducible y organización de rutas del pipeline analítico.....	279
Figura B3.1	Estructura de carpetas del proyecto reproducible y ubicación de salidas analíticas.....	283
Figura B3.2	Pipeline analítico reproducible del cuaderno en Python (Colab): etapas, controles y artefactos de salida.....	284
Figura B4.1	Depósito del paquete reproducible de la tesis en Open Science Framework (OSF).....	286
Figura B4.2	Estructura del paquete reproducible en OSF y correspondencia con apéndices de la tesis.....	288
Figura C2.1	Procedimiento de validación complementaria mediante Delphi modificado de dos rondas.....	293
Figura C3.1	Modelo de invitación enviada al panel de expertos.....	294
Figura Anexo 2.1	Diagrama mínimo de cumplimiento para uso de materiales de terceros (licencias y atribución).....	306

ÍNDICE DE TABLAS

Tabla 1.1	Disponibilidad de internet y computadora en hogares (total nacional), México 2023–2024 (ENDUTIH).....	39
Tabla 1.2	Definición Operacional de las Variables de Estudio.....	43
Tabla 2.1	Estudios y revisiones recientes sobre IA/RV en educación superior.....	53
Tabla 2.2	Matriz de cumplimiento normativo y controles implementados....	77
Tabla 3.1	Operacionalización de variables.....	80
Tabla 4.1	Cronograma operativo del trabajo de campo, responsables, recursos y evidencia verificable (ciclo 2024–2025).....	107
Tabla 4.2	Ciclo instruccional estandarizado de la intervención y evidencia verificable.....	114
Tabla 4.3	Quality gates de implementación.....	124
Tabla 4.4	Evidencias de implementación y trazabilidad.....	128
Tabla 4.5	Matriz de congruencia del estudio.....	131
Tabla 4.6A	Índices de calidad de medición, rúbricas de desempeño (doble calificación ciega).....	143
Tabla 4.6B	Índices psicométricos, escalas de autoinforme experiencial (POST).....	143
Tabla 4.7	Fiabilidad interevaluador por criterio (ICC [2,2]), doble calificación ciega.....	145
Tabla 4.8	Validez de contenido (S-CVI/Ave) por dimensión (acuerdo absoluto).....	145

Tabla 4.9A	Efecto diferencial del modelo multinivel: interacción Tiempo × Grupo.....	147
Tabla 4.9B	Tamaños de efecto para diseños PRE–POST con comparación (<i>d_ppc2</i>).....	148
Tabla 4.9C	Amenazas a la validez y mitigaciones en un cuasi-experimento por conglomerados.....	149
Tabla 4.9D	Descriptivos PRE–POST por grupo y porcentaje de puntajes en techo (0–100).....	150
Tabla 4.10	Modelo OLS para ganancia en Comunicación (Δ Comunicación), n = 180.....	152
Tabla 4.11	Modelo OLS para ganancia en Solución de problemas (Δ Solución de problemas), n = 180.....	152
Tabla 4.12	Adherencia por componente (promedio y rango).....	154
Tabla 5.1	Fases, acciones y criterios de avance para la implementación de la arquitectura IA+RV.....	191
Tabla 5.2	Matriz de componentes, métodos, técnicas e instrumentos de aplicación de la arquitectura IA+RV.....	195
Tabla 5.3	Recursos mínimos para el funcionamiento y sostenibilidad de la arquitectura IA+RV.....	200
Tabla 5.4	Valoración de pertinencia de la arquitectura IA+RV.....	202
Tabla 5.5	Matriz de factibilidad técnica, operativa e institucional de la arquitectura IA+RV.....	204
Tabla 5.6	Criterios de adopción y escalamiento responsable de la arquitectura IA+RV.....	205
Tabla 5.7	Caracterización del panel participante en el Delphi modificado....	207

Tabla 5.8	Matriz de valoración experta, ronda 1.....	208
Tabla 5.9	Observaciones recurrentes de la ronda 1 y ajustes incorporados en la propuesta.....	209
Tabla 5.10	Matriz de revaloración experta, ronda 2.....	210
Tabla 5.11	Esquema de evaluación de la efectividad de la arquitectura IA+RV.....	213
Tabla 5.12	Matriz de decisión para seguimiento de efectividad y mejora continua.....	214
Tabla 6.1	Síntesis de cumplimiento de objetivos e hipótesis.....	225
Tabla A2.1	Rúbrica analítica de comunicación oral (escala 0–4; puntaje global 0–100).....	252
Tabla A2.2	Rúbrica analítica de solución de problemas (escala 0–4; puntaje global 0–100).....	253
Tabla A3.1	Formato de control de doble calificación y discrepancias entre jueces).....	254
Tabla A3.2	Resumen de consistencia interjueces e índice ICC (2,2).....	255
Tabla A4.1	Checklist de integridad y coherencia de datos por corte temporal (PRE/POST)).....	256
Tabla A4.2	Registro de imputación múltiple (si MAR; m=20).....	256
Tabla A4.3	Plantilla de cambios al plan analítico y justificación (prerregistro vs. ejecución).....	256
Tabla A5.1	Inventario de datos y clasificación.....	258
Tabla A5.2	Matriz RACI de gobernanza y responsabilidades.....	259
Tabla A5.3	Reglas de resguardo, retención y trazabilidad.....	260

Tabla A5.4	Formato de registro mínimo de accesos (plantilla).....	261
Tabla A6.1	Aviso de privacidad simplificado para participantes (plantilla).....	263
Tabla A6.2	Lineamientos de uso responsable de IA (estudiantes).....	264
Tabla A6.3	Inducción operativa (IA+RV) al inicio del curso: checklist de comprensión.....	265
Tabla A6.4	Declaración breve para materiales del curso.....	266
Tabla A7.1	Matriz de riesgos (registro mínimo).....	268
Tabla A7.2	Quality gates para escalamiento (umbral–evidencia–decisión).....	269
Tabla A7.3	Inducción breve al inicio de práctica.....	271
Tabla A7.4	Checklist operativo por sesión (pre–durante–post).....	272
Tabla A7.5	Protocolo ante ciber-mareo (ciberenfermedad).....	273
Tabla A7.6	Formato de reporte de incidente (plantilla).....	274
Tabla A7.7	Lineamientos de higiene y cuidado del equipo.....	274
Tabla B1.1	Paquetes y versiones (entorno Python).....	278
Tabla B2.1	Estructura del proyecto y contenido esperado por carpeta.....	280
Tabla B2.2	Variables de rutas y parámetros mínimos (Google Colab).....	281
Tabla B3.1	Flujo de ejecución del cuaderno reproducible en Python (Google Colab): etapas del pipeline analítico.....	282
Tabla B4.1	Inventario mínimo del paquete reproducible en OSF: artefactos, propósito y condición de privacidad.....	287
Tabla C1.1	Criterios de inclusión del panel de expertos.....	290
Tabla C1.2	Caracterización del panel participante.....	291
Tabla C5.1	Matriz de valoración experta de la propuesta. Ronda 1.....	296

Tabla C6.1	Síntesis agregada de resultados de la ronda 1.....	298
Tabla C7.1	Observaciones recurrentes de la ronda 1 y ajustes incorporados en la propuesta.....	299
Tabla C8.1	Matriz de valoración experta. Ronda 2.....	300
Tabla C9.1	Comparación entre la ronda 1 y la ronda 2.....	301
Tabla AN2.1	Registro de materiales de terceros utilizados.....	304
Tabla AN2.2	Texto estándar de atribución para figuras de plataforma.....	305

APÉNDICES Y ANEXOS

Sección	Descripción
APÉNDICES	Material elaborado específicamente para esta investigación y utilizado como soporte metodológico, operativo, analítico o reproducible del estudio.
Apéndice A. Instrumentos, control de calidad y documentación de campo	Conjunto de instrumentos, formatos, protocolos, matrices y registros elaborados para la implementación, monitoreo y evaluación de la intervención.
A1. Síntesis de instrumentos, control de calidad y salvaguardas operativas	Mapa integrador que sintetiza la relación funcional entre instrumentos, control de calidad, trazabilidad y salvaguardas operativas utilizadas en la implementación y evaluación del estudio.
A2. Rúbricas y guías de aplicación	Rúbricas analíticas utilizadas para evaluar el desempeño, junto con guías de aplicación, criterios operativos y anclajes de calificación.
A3. Control de calidad de la calificación e ICC (consistencia interjueces)	Formatos y tablas de control para documentar doble calificación, discrepancias, resolución de diferencias y verificación de consistencia interjueces mediante ICC.
A4. Bitácora analítica, integridad y de calidad de datos	Registro de limpieza, transformaciones, decisiones analíticas, exclusiones, controles de integridad, <i>checklist</i> de coherencia y trazabilidad del procesamiento.
A5. Gobernanza de datos, roles y resguardo	Reglas de gobernanza, roles de acceso, inventario, retención, resguardo y controles mínimos para asegurar trazabilidad, confidencialidad y auditabilidad del estudio.
A6. Avisos de privacidad y lineamientos de uso de IA y RV	Textos operativos elaborados para informar a participantes y establecer reglas de uso responsable de IA y RV, con énfasis en privacidad, integridad académica y transparencia.
A7. Matriz de riesgos, <i>quality gates</i> y operación segura para la IA+RV	Registro de riesgos, mitigaciones y umbrales operativos para implementar IA y RV con seguridad, equidad y calidad, incluyendo <i>quality gates</i> para continuidad o escalamiento.

Apéndice B. Cuaderno de análisis reproducible (Python/Colab)

B1. Entorno computacional

Material técnico elaborado para replicar el procesamiento, control de calidad, análisis estadístico y exportación de resultados del estudio.

Especificación de versiones, librerías, configuraciones y dependencias utilizadas para la ejecución reproducible del cuaderno analítico.

B2. Estructura del proyecto, rutas y criterios de reproducibilidad

Organización de carpetas, variables de ruta, convenciones de almacenamiento, verificación de ejecución, versionado, trazabilidad y estructura mínima del proyecto reproducible.

B3. Flujo de ejecución del cuaderno reproducible

Secuencia de ejecución del *pipeline* analítico en Python/Colab, desde la carga y validación de datos hasta la estimación de modelos, exportación de salidas y registro de bitácora y hash.

B4. Depósito en Open Science Framework (OSF)

Información del repositorio OSF, DOI del paquete reproducible, inventario mínimo de artefactos depositados y correspondencia entre datos, cuaderno, *scripts* y salidas.

APÉNDICE C

Validación complementaria de la propuesta mediante delphi modificado de dos rondas

Incluye el panel experto, el procedimiento Delphi, los instrumentos aplicados, los resultados de ambas rondas y los ajustes incorporados a la propuesta.

C1. Criterios de selección y caracterización del panel de expertos

Define los criterios de inclusión del panel y presenta la caracterización sintética de los expertos participantes en la validación Delphi.

C2. Procedimiento general del Delphi modificado

Describe la secuencia metodológica de la validación en dos rondas, incluyendo envío, valoración, síntesis de resultados, ajuste de la propuesta y revaloración.

C3. Formato de invitación de expertos

Incluye el modelo de invitación remitido al panel para solicitar su participación en la validación complementaria de la propuesta.

C4. Resumen ejecutivo enviado a expertos	Síntesis estructurada de la propuesta remitida al panel, incluyendo problema, hallazgo central, mecanismo instruccional, componentes, sistema mínimo viable, estrategia de implementación y criterios de gobernanza
C5. Instrumento de valoración experta. Ronda 1	Matriz de evaluación aplicada en la primera ronda Delphi para valorar claridad, coherencia empírica, pertinencia pedagógica, factibilidad, escalabilidad, gobernanza, equidad y aporte doctoral, junto con preguntas abiertas de retroalimentación.
C6. Síntesis de resultados de primera ronda	Concentrado de valoraciones cuantitativas y observaciones cualitativas de la ronda 1, con identificación de fortalezas, áreas de mejora y acciones de ajuste derivadas.
C7. Matriz de ajustes realizados entre rondas	Registro de observaciones recurrentes del panel experto y de las modificaciones incorporadas a la propuesta antes de la segunda ronda.
C8. Instrumento de valoración experta. Ronda 2	Formato de revaloración aplicado sobre la versión revisada de la propuesta, orientado a estimar suficiencia de ajustes, convergencia y aceptación del modelo refinado.
C9. Resultados finales del Delphi modificado	Comparación sintética entre ambas rondas y síntesis interpretativa del proceso, mostrando convergencia, mejora de valoraciones y respaldo experto a la propuesta revisada.
ANEXOS	Documentos externos, institucionales o de terceros que respaldan el estudio, pero cuya autoría o procedencia no corresponde directamente al desarrollo original de esta investigación.
Anexo 1. Formato institucional de consentimiento informado y aviso de privacidad	Documento base institucional utilizado como referencia o soporte formal para el consentimiento y la protección de datos de los participantes.

**Anexo 2. Licencias,
atribuciones y materiales de
terceros**

Registro de materiales de terceros, capturas de plataforma, formatos de atribución y compatibilidad de licencias o permisos de uso.

GLOSARIO (A–Z).¹

Término	Definición
Agencia (<i>agency</i>)	Capacidad y sentido percibido de actuar intencionalmente para influir en el propio desempeño y resultados; en educación se vincula con autoría, control y autoeficacia al ejecutar tareas. (Bandura, 1986, 2001; Ryan & Deci, 2000).
Análisis de robustez	Conjunto de análisis complementarios para evaluar si los resultados se sostienen bajo supuestos alternativos (p. ej., modelos distintos, exclusiones, controles adicionales).
Anclaje del análisis	Decisión explícita de qué resultados responden a qué preguntas/hipótesis y con qué evidencia (tablas/figuras/estimadores).
Andamiaje	En el ámbito Pedagógico o Educativo (Teoría de Vygotsky) es un término técnico que describe el apoyo temporal que un tutor da a un estudiante para que adquiriera una nueva habilidad.
Analítica formativa	Uso de datos del proceso de aprendizaje para retroalimentar y mejorar desempeño durante el curso, no solo para calificar al final.
Anglicismo	Término tomado del inglés que se incorpora al texto académico (p. ej., <i>briefing</i> , <i>debriefing</i> , <i>feedback</i>).
Asignación por conglomerados (<i>cluster assignment</i>)	Asignación de condiciones (intervención/comparación) por unidades naturales preexistentes (p. ej., clases) en lugar de individuos, con implicaciones para la dependencia intraclase y el análisis. (Donner & Klar, 2000).
Auditabilidad	Capacidad de rastrear decisiones, datos y procedimientos para verificar cómo se obtuvieron resultados (bitácoras, versiones, reglas QA).

¹ Nota. Las definiciones de este glosario fueron elaboradas por el autor con fines de claridad, consistencia conceptual y alineación con la operacionalización de variables del estudio. Cuando una entrada corresponde a un constructo con definición canónica o proviene/adapta marcos y estándares ampliamente reconocidos, se indica la fuente de referencia.

Término	Definición
Autorregulación	Proceso cíclico mediante el cual el estudiante establece metas, planifica, monitorea su progreso, ajusta estrategias y evalúa resultados para optimizar su aprendizaje. (Zimmerman, 2000).
Autoinforme (<i>self-report</i>)	Medición basada en respuestas del participante sobre su experiencia (percepción, carga, presencia, utilidad).
<i>Bias</i> / sesgo	Distorsión sistemática que puede afectar medición o inferencia (selección, desempeño, expectativas, instrumentación).
Bitácora analítica	Registro de decisiones de limpieza, transformación, exclusiones, modelos y criterios de reporte para reproducibilidad.
<i>Briefing</i>	Preparación previa a la simulación: objetivos, reglas, criterios de éxito, advertencias y enfoque atencional.
CAMIL (<i>Cognitive-Affective Model of Immersive Learning</i>)	Modelo que explica cómo variables de inmersión e interacción influyen en presencia/agencia, afecto y carga cognitiva, y cómo estos mecanismos median resultados de aprendizaje en entornos inmersivos. (Makransky & Petersen, 2021).
Carga cognitiva	Demanda sobre la memoria de trabajo durante una tarea; suele distinguirse en intrínseca (complejidad inherente), extrínseca (distractores/diseño ineficiente) y germana (esfuerzo orientado a construir esquemas útiles). (Sweller, 1988; Sweller <i>et al.</i> , 1998).
Carga extrínseca	Esfuerzo mental inducido por elementos no esenciales (interfaz confusa, distractores irrelevantes).
Carga germana	Esfuerzo orientado a construir esquemas útiles (procesamiento significativo).
Carga intrínseca	Complejidad inherente de la tarea (número de elementos y su interacción).
<i>Checklist</i> / lista de cotejo	Instrumento para verificar adherencia a pasos o criterios de implementación.

Término	Definición
Ciber-mareo <i>(cybersickness)</i>	Conjunto de síntomas de malestar (p. ej., náusea, mareo, desorientación, fatiga) asociados a entornos virtuales inmersivos; relevante para criterios de seguridad, tolerancia y exposición. (LaViola, 2000).
Clúster (clase)	Unidad agrupada (p. ej., una clase) dentro de un diseño por conglomerados.
Competencias transversales <i>(soft skills)</i>	Habilidades aplicables en múltiples contextos (comunicación, solución de problemas, colaboración).
Condición de comparación	Grupo sin la arquitectura IA+RV (o sin su integración completa) usado como referencia para estimar efectos diferenciales.
CONSORT (diagrama de flujo)	Estándar de reporte para describir de forma transparente el flujo de participantes (elegibilidad, inclusión, exclusión y análisis), frecuentemente adaptado a diseños no clínicos. (Schulz <i>et al.</i> , 2010)
Control de línea base (PRE)	Ajuste por el nivel inicial del resultado para estimar efectos netos más interpretables.
Covariable	Variable adicional incluida para ajustar comparaciones (p. ej., puntaje PRE, programa, experiencia tecnológica).
Criterio (rúbrica)	Componente evaluado en desempeño (p. ej., claridad, estructura, evidencia).
Críterial <i>(criterial feedback)</i>	Retroalimentación referida a criterios explícitos (no a impresiones generales).
Debriefing	Reflexión posterior a la simulación: análisis de decisiones, transferencia, lecciones y plan de mejora.
Dependencia intraclase	Similitud de resultados entre estudiantes de la misma clase, relevante para potencia e inferencia.
Desempeño observable	Resultado medido sobre productos o actuaciones (presentaciones/casos) con criterios explícitos.

Término	Definición
Diferencia-en-diferencias (DID)	Estimación del efecto neto como la diferencia entre el cambio PRE–POST del grupo intervención y el cambio PRE–POST del grupo comparación, bajo supuestos de identificación pertinentes. (Angrist & Pischke, 2009).
Dosificación (dosis)	Cantidad/intensidad de exposición (minutos en RV, escenarios, eventos de <i>feedback</i> , frecuencia).
Dosis–respuesta	Relación entre mayor exposición y mayor cambio; puede existir o no incluso si hay efecto promedio.
Efecto diferencial	Diferencia de cambios entre condiciones (no solo diferencia en un momento).
Efecto neto	Parte del cambio atribuible a la intervención una vez descontado el cambio del grupo comparación.
Eficiencia instruccional	Logro de aprendizaje en relación con esfuerzo/carga/tiempo; idealmente logro alto con carga contenida.
Equidad de implementación	Capacidad de sostener la intervención sin amplificar brechas por acceso, soporte o condiciones de uso.
Feedback	Retroalimentación. Puede ser humana, asistida por IA o combinada.
Fidelidad de implementación	Grado en que lo ejecutado coincide con el diseño previsto (<i>briefing</i> , <i>debriefing</i> , rúbrica, señalización).
Flujo metodológico	Representación del proceso de campo y análisis (responsables, insumos, puntos de control).
Funcionalidad (realismo funcional)	Realismo relevante para el objetivo de aprendizaje (no solo realismo gráfico).
Gobernanza de datos	Políticas y controles para recolección, uso, resguardo, acceso, retención y destrucción de datos.
Hedges' g	Tamaño de efecto estandarizado con corrección por sesgo de muestras pequeñas/medias, utilizado para comparar magnitudes entre estudios o condiciones. (Hedges, 1981).
Hipótesis	Proposición contrastable sobre efectos (H1–H4 en la tesis).

Término	Definición
Heterogeneidad de efectos	Variación del impacto según perfil (programa, experiencia previa, etc.).
ICC (<i>Intraclass Correlation Coefficient</i>)	Coeficiente de correlación intraclase para cuantificar consistencia/acuerdo entre evaluadores o mediciones dentro de clústeres, y para describir dependencia intraclase. (Shrout & Fleiss, 1979).
ICC (2, k)	Variante de ICC correspondiente al modelo de dos vías, efectos aleatorios para el promedio de k evaluadores, bajo convención psicométrica ampliamente utilizada. (Shrout & Fleiss, 1979).
Imputación múltiple	Estrategia para manejo de datos faltantes que genera múltiples bases imputadas y combina estimaciones para reflejar incertidumbre por faltantes bajo supuestos explícitos. (Rubin, 1987).
Inmersión	Características técnicas del entorno (p. ej., RV) que aumentan sensación de estar dentro.
Interacción Tiempo $\times$ Grupo	Término de modelo que captura el efecto diferencial del cambio PRE–POST entre condiciones.
Intervalo de confianza (IC95%)	Rango plausible del estimador; expresa precisión, no solo significancia.
IA (inteligencia artificial)	Sistemas computacionales que realizan tareas asociadas con inteligencia humana (predicción, clasificación, generación de texto, analítica).
IA generativa	IA capaz de producir texto/imágenes/código; en educación, requiere lineamientos de uso responsable para evitar sustitución de autoría.
IA+RV (arquitectura)	Integración de retroalimentación/analítica asistida por IA y práctica inmersiva en RV, organizada como ciclo didáctico.

Término	Definición
<i>Implementation science</i>	Campo de investigación aplicada que estudia cómo introducir, sostener y escalar intervenciones (programas, prácticas o políticas) en contextos reales, cerrando la brecha entre lo que funciona y lo que se adopta efectivamente. Enfatiza el diseño de estrategias de implementación (capacitación, soporte, monitoreo, retroalimentación, adaptación controlada), la medición de fidelidad y resultados de implementación (p. ej., aceptabilidad, factibilidad, adopción, sostenibilidad), y la gestión de contexto y gobernanza para que una innovación mantenga su calidad al pasar de piloto a práctica recurrente.
<i>Quality gates</i>	Puntos de control predefinidos dentro de un proceso de despliegue o escalamiento en los que se decide si avanzar, pausar o corregir con base en criterios verificables. En esta tesis, los <i>quality gates</i> operan como umbrales que combinan: (i) evidencia de implementación (dosis y fidelidad), (ii) evidencia de medición (fiabilidad/estabilidad de instrumentos), y (iii) evidencia de seguridad y ética (incidentes, sesgos, accesibilidad, trazabilidad). Su propósito es asegurar adopción con fidelidad y evidencia, evitando que el escalamiento ocurra por presión institucional o entusiasmo tecnológico sin control de calidad.
Modelo lineal mixto	Familia de modelos que incorporan efectos fijos (promedios) y aleatorios (variación por clúster/individuo), apropiados para datos jerárquicos o repetidos. (Raudenbush & Bryk, 2002).
MAR / MCAR	Supuestos formales sobre mecanismos de datos faltantes: <i>Missing At Random</i> y <i>Missing Completely At Random</i> , relevantes para decidir métodos de análisis e imputación. (Little & Rubin, 2019).
Metodología de métodos mixtos	Enfoque que integra evidencia cuantitativa y cualitativa mediante diseños explícitos de convergencia/explicación/expansión para responder cuánto cambia y por qué/cómo cambia. (Creswell & Plano Clark, 2018).

Término	Definición
Modelo lógico	Representación insumos → procesos → resultados para ju escalamiento y evaluación.
Moderación	Efecto que cambia según otra variable (p. ej., experiencia tecnológica previa).
MR (Realidad mixta)	Integración de contenidos digitales con el entorno físico, con anclaje espacial e interacción bidireccional (más allá de AR).
N (tamaño de muestra)	Número de participantes analizados; la tesis menciona N=180 con análisis principal n=160.
<i>Novelty effect</i> (efecto novedad)	Mejora transitoria por novedad tecnológica, no necesariamente por aprendizaje estable.
OLS (mínimos cuadrados ordinarios)	Método estándar de estimación para regresión lineal que minimiza la suma de residuos al cuadrado bajo supuestos clásicos, con inferencia dependiente de dichos supuestos o de ajustes robustos. (Wooldridge, 2016).
Operacionalización	Traducción de constructos (p. ej., comunicación) a indicadores y métricas medibles (rúbricas, escalas).
p-valor	Probabilidad de observar un resultado igual o más extremo bajo la hipótesis nula; no mide tamaño del efecto.
Presencia	Constructo que describe la experiencia subjetiva de estar en un entorno virtual, influida por inmersión, contingencia sensorimotora y coherencia del entorno. (Slater & Wilbur, 1997).
Presencia social	Percepción de estar con otros y de interacción social significativa mediada por tecnología, incluyendo señales de copresencia, involucramiento e intencionalidad. (Biocca <i>et al.</i> , 2003).
PRE–POST	Medición antes y después de la intervención.
Práctica deliberada	Entrenamiento estructurado con metas específicas, retroalimentación, repetición y corrección focalizada para mejorar desempeño, más allá de la simple práctica. (Ericsson <i>et al.</i> , 1993).
Proposición Teórica Integradora (PTI)	Síntesis que conecta marcos teóricos en una cadena plausible y contrastable.

Término	Definición
QA (<i>Quality Assurance</i>)	Aseguramiento de calidad del proceso (reglas, verificaciones, trazabilidad).
<i>Quasi-experimental</i> (cuasi-experimental)	Diseño con grupos de comparación sin aleatorización individual estricta.
RV (Realidad virtual)	Entorno inmersivo generado por computadora, usualmente con visor HMD, que sustituye el campo visual del mundo físico.
Realismo funcional	Realismo que importa para el desempeño (restricciones, <i>trade-offs</i> , consecuencias).
Reproducibilidad	Capacidad de replicar análisis con mismos datos/procedimientos y obtener resultados consistentes.
Rúbrica analítica	Instrumento de evaluación con criterios desagregados y niveles de desempeño; permite retroalimentación específica.
SAMR	Marco que clasifica el uso tecnológico desde sustitución hasta redefinición de tareas.
S-CVI/Ave	Índice promedio de validez de contenido a nivel escala/dimensión (consenso de expertos). La tesis reporta valores máximos.
Seudonimización	Sustitución de identificadores directos por códigos para proteger identidad.
Señalización	Estrategia instruccional que dirige atención a elementos relevantes (reduce carga extrínseca).
Significancia estadística	Evidencia contra hipótesis nula; no equivale a relevancia educativa.
<i>Soft skills</i>	Habilidades blandas; competencias transversales.
Sostenibilidad operativa	Capacidad institucional para mantener intervención (equipos, soporte, protocolos, formación).
TPACK	Marco que plantea que el diseño instruccional eficaz integra coherentemente conocimiento tecnológico, pedagógico y disciplinar, y sus intersecciones. (Mishra & Koehler, 2006).
Tamaño de efecto (<i>effect size</i>)	Medida de magnitud del cambio o diferencia (p. ej., d, g).

Término	Definición
<i>d_{ppc2}</i>	Tamaño de efecto estandarizado para diseños PRE–POST con grupo comparación, útil para expresar la diferencia-en-diferencias en unidades de desviación estándar con correcciones recomendadas. (Morris, 2008)
Trazabilidad	Capacidad de seguir el hilo desde pregunta → hipótesis → instrumento → análisis → resultado.
Validez ecológica	Grado en que el estudio representa condiciones reales (curso regular, operación natural).
Validez de contenido	Grado en que el instrumento cubre el dominio del constructo (evaluación por expertos).
Viabilidad	Factibilidad técnica y organizacional de implementar la arquitectura (infraestructura, soporte, tiempos).
VirtualSpeech	Plataforma de práctica y analítica para entrenamiento comunicativo

INTRODUCCIÓN

*“Quien concibe cursos de acción para transformar
lo existente en lo deseado, diseña.”*

— *Herbert A. Simon.*²

La integración de la inteligencia artificial (IA) y la realidad virtual (RV) se ha expandido en la educación superior como un conjunto de prácticas que, cuando se diseñan y evalúan con criterios verificables, pueden reconfigurar la enseñanza más allá de una adopción tecnológica superficial. En México y América Latina, diversas instituciones han acelerado esta transición al articular innovación educativa con evaluación rigurosa y enfoque competencial. En ese marco, esta tesis examinó, en condiciones reales de aula, la influencia de una arquitectura didáctica que combina retroalimentación formativa asistida por IA y simulaciones en RV sobre dos competencias transversales decisivas para la empleabilidad y el liderazgo profesional: Comunicación y Solución de problemas. El estudio se desarrolló con estudiantes de pregrado del Tecnológico de Monterrey durante el ciclo 2024–2025, en un ecosistema institucional que favoreció la experimentación responsable, la analítica formativa y la transferencia de prácticas docentes efectivas (Chen & Lee, 2024; Di Palma et al., 2025). Desde el inicio se aseguró la trazabilidad entre el problema, las decisiones de diseño y el análisis, de modo que cada componente de la intervención quedara vinculado a indicadores observables de exposición y a métricas de desempeño, permitiendo evaluar la contribución de la arquitectura propuesta con criterios comparables y auditables.

El Tecnológico de Monterrey ofreció un entorno particularmente propicio para implementar y valorar, con evidencia, el uso de IA y RV en cursos regulares. En el periodo 2024–2025 coincidieron tres condiciones habilitadoras: la maduración de la IA generativa en entornos educativos, una mayor accesibilidad a dispositivos inmersivos y la consolidación de prácticas mediadas por datos. Este punto de inflexión permitió

² Nota. Traducción propia del epígrafe, con base en Simon, H. A. (1996), *The Sciences of the Artificial* (3rd ed.). MIT Press.

observar el potencial de IA y RV más allá de pilotos aislados o evidencias estrictamente de laboratorio y someterlo a evaluación con criterios de validez interna y externa propios de la investigación educativa. En términos operativos, las intervenciones se organizaron como una arquitectura instruccional que combinó (i) retroalimentación mediada por IA y (ii) práctica inmersiva en RV, articuladas mediante un ciclo *briefing–simulación–debriefing* integrado a las evaluaciones y productos académicos habituales (Peláez-Sánchez *et al.*, 2024). La fidelidad de implementación se documentó mediante listas de cotejo y registros de dosificación (minutos de uso, número de escenarios, eventos de *feedback* y presencia percibida), con el fin de sostener comparabilidad entre grupos y apoyar la interpretación de los resultados.

La literatura reciente ha documentado beneficios potenciales asociados con aprendizaje activo, presencia/inmersión, personalización de la retroalimentación y transferencia a contextos auténticos, con variabilidad sustantiva según dominio, población y calidad del diseño instruccional (Trisnawati *et al.*, 2023; Pupic *et al.*, 2023; Radianti *et al.*, 2020; Makransky & Petersen, 2021). Estudios aplicados en dominios específicos han reportado impactos positivos en desempeño o rehabilitación en función de cómo se implementa la tecnología (Cinar, 2024; Gómez Portes, 2023). Sin embargo, persistía una brecha relevante para el campo: evidencia focalizada sobre el efecto combinado de IA y RV en habilidades blandas en educación superior en México, particularmente en comunicación y solución de problemas, competencias altamente vinculadas con empleabilidad, liderazgo y desempeño profesional (Chen & Lee, 2024). Esta investigación atendió dicha brecha con un diseño empírico contextualizado y mediciones alineadas a estándares psicométricos, de manera que los hallazgos fueran auditables y transferibles.

Este dinamismo también se refleja en tesis doctorales recientes que exploran RV/MR, analítica multimodal del aprendizaje y automatización de la retroalimentación en contextos formativos, lo que refuerza la vigencia y relevancia del problema abordado (Barzola, 2020; Ding, 2023; Etienne, 2025; Ouhaichi, 2024; Suresh, 2022).

Sobre esta base, metodológicamente, el estudio adoptó un enfoque de métodos mixtos con propósito explicativo. El componente cuantitativo utilizó mediciones PRE y

POST con grupos asignados por conglomerados (clases), con y sin intervención, preservando validez ecológica y permitiendo estimar efectos diferenciales en condiciones de curso. El componente cualitativo (entrevistas semiestructuradas y observación participante) aportó densidad interpretativa para comprender la experiencia estudiantil y docente, así como factores que mediaron o moderaron los resultados observados. La muestra se conformó por aproximadamente 180 estudiantes, estratificada por programa y organizada a nivel clase, con el objetivo de sostener heterogeneidad y potencia analítica para evaluar la intervención en un contexto real (Peláez-Sánchez *et al.*, 2024). El protocolo incluyó un plan explícito para el manejo de datos faltantes (diagnóstico MCAR/MAR e imputación múltiple cuando procedió) y una bitácora analítica que registró decisiones de depuración y modelado.

El estudio no se limitó a describir asociaciones; puso a prueba mecanismos de cambio plausibles y verificables: (1) retroalimentación mediada por IA que favorece autorregulación y mejora iterativa del desempeño; (2) práctica deliberada en RV que incrementa presencia y permite ensayos repetibles de desempeño y toma de decisiones en condiciones de menor costo social; y (3) gestión de carga cognitiva mediante secuencias instruccionales que equilibran desafío y apoyo. Estos mecanismos se operacionalizaron mediante rúbricas analíticas, registros de dosificación IA/RV y evidencia cualitativa de proceso. Sobre esa base se derivaron hipótesis específicas sobre la incidencia de la intervención en la calidad de la comunicación (claridad, estructura, evidencia, adecuación a la audiencia) y en la solución de problemas (definición del problema, generación y evaluación de alternativas, anticipación de consecuencias y justificación). Asimismo, la especificación de mediaciones (p. ej., presencia/autoeficacia) y moderaciones (p. ej., experiencia tecnológica previa) se dejó explícita para su contraste en los modelos estadísticos y la triangulación cualitativa.

La contribución de la tesis es doble. En el plano teórico, propone un andamiaje conceptual que vincula tecnologías emergentes, principios de diseño instruccional y mecanismos de aprendizaje con competencias transversales, enfatizando que el efecto educativo depende de una arquitectura auditable (rúbricas explícitas, ciclos de práctica deliberada y retroalimentación criterial), y no de la tecnología en abstracto (Hattie &

Timperley, 2007; Shute, 2008; Wisniewski *et al.*, 2020; Makransky & Petersen, 2021). En el plano empírico, ofrece evidencia evaluada con estándares de calidad, mediciones pre-post por conglomerados, control de fidelidad de implementación, análisis comparativos con inferencia apropiada y triangulación, capaz de orientar decisiones curriculares y el escalamiento responsable de intervenciones tecnopedagógicas en educación superior. De forma transversal, el estudio incorporó lineamientos éticos y de gobernanza de datos, resguardando privacidad, equidad y seguridad de la experiencia estudiantil, y sosteniendo la reproducibilidad mediante bitácoras, plantillas y versiones de materiales analíticos.

El proyecto se condujo conforme a principios de autonomía, equidad, transparencia y uso responsable de tecnologías, en alineación con la normativa institucional y las disposiciones nacionales vigentes en protección de datos personales. Todos los participantes otorgaron consentimiento informado; los datos se trataron con minimización, seudonimización y resguardo seguro. La IA no sustituyó la autoría estudiantil ni generó productos evaluables; su uso se limitó a analítica formativa y retroalimentación pedagógica. Además, se definieron periodos de retención, protocolos de destrucción segura y registros de versiones y decisiones para asegurar auditabilidad. Las consideraciones éticas completas, incluyendo manejo de datos, accesibilidad y uso responsable de IA y RV, se describen en la sección metodológica correspondiente.

La tesis se organiza en cinco capítulos. El Capítulo 1 presenta el problema de investigación, los objetivos, las hipótesis y las delimitaciones. El Capítulo 2 desarrolla el fundamento teórico, el marco conceptual y la lógica de mecanismos que sustenta la arquitectura IA+RV. El Capítulo 3 expone la operacionalización de variables, el diseño metodológico y la estrategia de medición. El Capítulo 4 presenta el análisis de resultados, integrando evidencia de calidad de medición, fidelidad de implementación, estimaciones principales y análisis complementarios. El Capítulo 5 transforma los hallazgos empíricos en una propuesta de arquitectura didáctica IA+RV, replicable, auditable y escalable bajo condiciones de calidad, seguridad y gobernanza. Finalmente, se presentan las conclusiones generales, recomendaciones, limitaciones y líneas futuras de investigación.

CAPÍTULO 1. PROTOCOLO DE INVESTIGACIÓN

El presente capítulo expondrá el protocolo de investigación que dará coherencia al estudio, articulando de manera consistente el problema de investigación, las hipótesis, el método, la medición y el análisis, bajo criterios de validez interna y externa propios de la investigación educativa aplicada. En primer término, se definirá el tema y se delimitará el problema en torno a la integración didáctica de inteligencia artificial (IA) y realidad virtual (RV) en cursos regulares de educación superior, con foco en su contribución a dos competencias transversales: comunicación y solución de problemas. En segundo término, se formulará la pregunta central, se justificará la pertinencia académica y social del trabajo y se precisará el objeto de estudio dentro del ecosistema institucional. Posteriormente, se establecerán los objetivos (general y específicos) y las hipótesis que guían las pruebas empíricas, junto con el alcance y las delimitaciones necesarias para interpretar con prudencia los hallazgos. Finalmente, se presentarán las decisiones metodológicas clave (diseño cuasi-experimental por conglomerados, métodos mixtos, medición del desempeño mediante rúbricas analíticas con doble calificación e índices de acuerdo interjueces, registros de dosificación IA/RV y salvaguardas éticas y de datos), dejando trazable el marco que estructurará los capítulos siguientes.

1.1 Tema de investigación

Esta investigación se centra en la integración didáctica de inteligencia artificial (IA) y realidad virtual (RV) para fortalecer dos competencias transversales en estudiantes de pregrado del Tecnológico de Monterrey durante el ciclo 2024–2025: Comunicación y Solución de problemas. El tema se ubica en la innovación educativa con tecnologías emergentes y aborda la necesidad de diseñar experiencias formativas que articulen práctica situada, retroalimentación criterial, mediación docente y evaluación mediante evidencias observables. En este marco, la arquitectura IA+RV concibe la IA como apoyo formativo y la RV como entorno de desempeño, bajo condiciones reales de aula, con criterios de calidad, trazabilidad y responsabilidad educativa.

1.2 Planteamiento del problema

La incorporación de inteligencia artificial (IA) y realidad virtual (RV) en educación superior ha ampliado el repertorio de estrategias instruccionales para promover aprendizaje activo, retroalimentación más oportuna y experiencias situadas de práctica deliberada. En particular, la IA, cuando se configura como tutor formativo, puede incrementar la precisión y la accionabilidad de la retroalimentación, mientras que la RV, cuando se diseña como entorno de desempeño, puede elevar la autenticidad de la tarea y sostener ensayos repetibles en condiciones controladas. No obstante, la evidencia indica que los beneficios educativos no dependen de la tecnología por sí misma, sino de su integración como arquitectura didáctica con criterios explícitos, secuencias estables y mecanismos de reflexión que traduzcan la experiencia en mejora observable (Trisnawati *et al.*, 2023; Chen & Lee, 2024; Di Palma *et al.*, 2025; Makransky & Petersen, 2021).

Pese a la expansión acelerada de iniciativas institucionales y experiencias de implementación, persiste en la educación superior mexicana una brecha evaluativa específica: faltan estimaciones robustas, bajo condiciones reales de curso, que determinen si y en qué magnitud la integración conjunta de IA y RV mejora competencias transversales decisivas para el desempeño profesional, particularmente comunicación y solución de problemas, utilizando mediciones de desempeño con criterios estables y análisis consistentes con estructuras jerárquicas (clases) (Shumeiko & Osadcha, 2024). Una proporción relevante de la literatura disponible se concentra en dominios disciplinares o en diseños pre-experimentales, lo que limita la atribución del efecto y la transferencia a escenarios curriculares regulares.

El problema adquiere mayor relevancia al considerar las condiciones estructurales de acceso y de la capacidad instalada para el uso educativo de tecnologías en México. Bajo la premisa, documentada por la literatura sobre brecha digital en el país, de que la adopción y los patrones de uso de internet se asocian con la escolaridad y el estatus económico del hogar, y difieren sistemáticamente por lugar de residencia (urbano–rural), persisten asimetrías por región y estrato socioeconómico, así como desigualdades en conectividad y en infraestructura institucional (Martínez-Domínguez & Fierros-González, 2022; Martínez-Domínguez & Mora-Rivera, 2020). En consecuencia,

la adopción sostenible de arquitecturas IA+RV enfrenta restricciones estructurales que condicionan su viabilidad y su equidad. Para contextualizar este marco, la Tabla 1.1 y la Figura 1.1 sintetizan indicadores nacionales de acceso a internet y disponibilidad de computadora en hogares, utilizados aquí como referencia interpretativa (INEGI, 2024).

Tabla 1.1

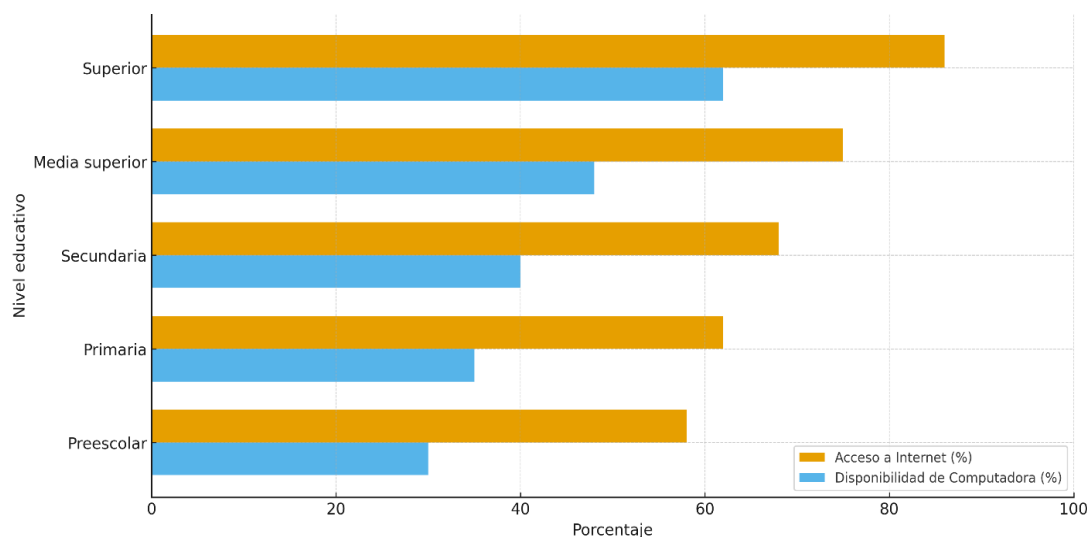
Disponibilidad de internet y computadora en hogares (total nacional), México 2023–2024 (ENDUTIH)

Año	Hogares con internet (%)	Hogares con computadora (%)
2023	71.7	44.7*
2024	73.6	43.9

Nota. Cifras porcentuales sobre el total de hogares; ENDUTIH/INEGI.

Figura 1.1

Acceso a internet y disponibilidad de computadora en hogares con estudiantes por nivel educativo, México, 2024.



Nota: La visualización muestra los porcentajes de hogares con estudiantes que cuentan con acceso a internet y con computadora, por nivel educativo. Gómez Tobías (2026), con base en microdatos de ENDUTIH/INEGI³ (2024).

³ En esta investigación, los indicadores ENDUTIH se utilizan con fines contextuales para delimitar condiciones estructurales que inciden en la viabilidad, escalabilidad y equidad de intervenciones tecno-pedagógicas. Aunque el acceso a internet en hogares puede ser alto, la disponibilidad de computadoras y la conectividad institucional no son homogéneas; por ello, la arquitectura IA+RV exige no solo conectividad básica, sino disponibilidad operativa sostenida, soporte técnico, condiciones de uso seguro (p. ej., protocolos ante ciber-mareo) y consistencia de implementación (registro de dosificación y fidelidad).

En contraste con ese panorama, el Tecnológico de Monterrey ofrece un entorno con infraestructura, soporte académico y cultura de innovación que permite implementar y evaluar con mayor control una arquitectura IA+RV en cursos regulares, asegurando comparabilidad de experiencias, registro de dosificación y evidencia de fidelidad. Esto habilita una prueba empírica en condiciones operativas auténticas, donde el foco no es usar tecnología, sino estimar su contribución cuando se integra como un ciclo instruccional verificable (*briefing–simulación–debriefing*) y con retroalimentación criterial.

Desde la perspectiva de la formación profesional, comunicación y solución de problemas se reconocen como competencias críticas en escenarios de trabajo complejos, automatizados y colaborativos, donde la calidad del desempeño depende de claridad argumental, lectura de contexto, generación y evaluación de alternativas y justificación de decisiones (Ngo, 2024; Ferreira *et al.*, 2024). Bajo este encuadre, la tesis plantea si la arquitectura IA+RV produce mejoras significativas y educacionalmente relevantes en dichas competencias frente a cursos equivalentes sin integración, controlando la línea base y los efectos de clúster (clases). Asimismo, se consideran mecanismos plausibles (autorregulación, presencia/inmersión y gestión de carga cognitiva) como marcos explicativos, sin sobredimensionar inferencias causales cuando la medición no las sustente.

En suma, el problema de investigación no reside en la disponibilidad de tecnologías, sino en la falta de evidencia local, rigurosa y transferible sobre qué cambia, cuánto cambia y bajo qué condiciones mínimas cambia el desempeño en competencias transversales cuando IA y RV se articulan como una arquitectura didáctica auditable en cursos regulares de educación superior (Shumeiko & Osadcha, 2024; Makransky & Petersen, 2021).

1.3 Pregunta de investigación

Se formuló la siguiente pregunta central:

¿Cómo se puede contribuir a la mejora del desempeño en competencias transversales de comunicación y solución de problemas mediante ciclos de práctica

situada, retroalimentación criterial y reflexión en estudiantes de pregrado del Tecnológico de Monterrey, México, durante el periodo 2024 a 2026?

1.4 Justificación

La investigación se justifica por la necesidad de aportar nuevo conocimiento sobre cómo la inteligencia artificial (IA) y la realidad virtual (RV), integradas como arquitectura didáctica y no como recursos aislados, pueden contribuir al desempeño observable en Comunicación y Solución de problemas en educación superior. Desde el punto de vista teórico, el estudio fortalece la discusión sobre tecnologías emergentes, práctica situada y retroalimentación criterial al analizar su articulación en competencias transversales, campo donde la evidencia en contextos reales de curso en México aún es limitada (Trisnawati et al., 2023; Shumeiko & Osadcha, 2024).

Desde el punto de vista práctico, la investigación aporta una propuesta aplicable para diseñar, implementar y evaluar experiencias IA+RV mediante rúbricas, indicadores de fidelidad, dosificación y trazabilidad, lo que puede apoyar decisiones curriculares, mejora instruccional y escalamiento responsable (Carroll et al., 2007; Creswell & Plano Clark, 2018). Desde el punto de vista social, beneficia directa e indirectamente a estudiantes, docentes e instituciones al orientar el desarrollo de competencias valoradas por empleadores y organismos profesionales, como comunicación efectiva y solución de problemas complejos (Ngo, 2024; Robles, 2012). Finalmente, desde el punto de vista metodológico y personal, el estudio ofrece un modelo evaluativo replicable en contextos semejantes y responde al interés del investigador por mejorar la formación universitaria mediante innovación educativa sustentada en evidencia, ética y responsabilidad institucional.

1.5 Objeto de estudio

El objeto de estudio es la contribución pedagógica de la integración de inteligencia artificial (IA) y realidad virtual (RV) al desarrollo y desempeño de dos competencias transversales en estudiantes de pregrado del Tecnológico de Monterrey durante el ciclo 2024–2025: comunicación y solución de problemas. En esta tesis, dichas competencias

se entienden como capacidades demostrables en tareas auténticas del curso, evaluables mediante criterios explícitos y estables.

En el caso de la comunicación oral, el objeto se precisa como el desempeño del estudiante para construir y sostener un mensaje con hilo conductor, argumentación y ajuste a un público. En términos evaluativos, este desempeño se concreta en criterios observables: claridad y coherencia del discurso, estructura argumental, uso de evidencia para respaldar afirmaciones, adecuación a la audiencia (registro, tono y ejemplos), fluidez verbal (incluyendo control de muletillas) y manejo del lenguaje no verbal como soporte del mensaje. Esta definición permite delimitar la comunicación no como percepción general, sino como un conjunto de componentes evaluables en niveles de calidad progresivos.

En el caso de la solución de problemas, el objeto se delimita como el desempeño del estudiante para comprender un problema situado, formular alternativas y justificar decisiones en función de criterios y consecuencias. De forma operacional en tareas del curso, este desempeño se expresa en criterios observables: definición precisa del problema (con supuestos y límites), generación de alternativas plausibles, evaluación de alternativas con criterios explícitos y consideración de trade-offs, anticipación de consecuencias en distintos horizontes temporales, justificación de la decisión con evidencia, contextualización (restricciones, recursos, actores y viabilidad) e incorporación de principios y análisis de impacto social de la decisión.

La integración de IA y RV se considera en el objeto de estudio como una arquitectura didáctica que combina, por un lado, apoyos de retroalimentación formativa y, por otro, oportunidades de desempeño en escenarios que demandan comunicación y toma de decisiones. Para sostener una interpretación rigurosa del objeto, el estudio contempla el registro de la exposición efectiva a dichas experiencias y el control de condiciones de base que pueden influir en el desempeño. La evidencia de desempeño se deriva de productos auténticos evaluados con rúbricas analíticas, con procedimientos orientados a consistencia entre evaluadores y trazabilidad analítica para respaldar la auditabilidad de los resultados.

Tabla 1.2

Definición Operacional de las Variables de Estudio

Variable	Definición operacional	Indicadores observables	Unidad de análisis e instrumentos / Métrica
IA (exposición)	Herramientas que proveyeron analítica formativa y retroalimentación criterial durante las sesiones.	Frecuencia (sesiones con IA / total); intensidad (minutos por sesión); n.º de eventos de <i>feedback</i> IA–estudiante; calidad percibida (1–5).	Sesión/estudiante; registros de plataforma y bitácora docente; minutos (continuo), conteos (eventos) y Likert (1–5).
RV (exposición)	Entornos inmersivos para práctica deliberada y simulación de escenarios profesionales con <i>debriefing</i> .	Minutos por sesión; n.º de escenarios completados; presencia percibida (1–5); incidentes (ciber-mareo: sí/no).	Sesión/estudiante; <i>logs</i> y <i>checklist</i> de fidelidad; continuo y dicotómico.
Comunicación (desempeño)	Capacidad para expresar ideas claras, estructuradas y adecuadas a la audiencia en productos orales.	Claridad/coherencia; estructura argumental; evidencia; entrega oral; manejo de audiencia; lenguaje corporal.	Producto/estudiante (presentaciones, debates). Rúbrica analítica (criterios 0–4) escalada a 0–100; doble calificación; ICC interjueces.
Solución de problemas (desempeño)	Capacidad para identificar problemas, generar/evaluar alternativas, anticipar consecuencias y justificar decisiones.	Identificación; alternativas; evaluación; consecuencias; justificación/síntesis (opcional: criterios éticos).	Producto/estudiante (casos/dilemas). Rúbrica analítica (criterios 0–4) escalada a 0–100; doble calificación; ICC interjueces.

Nota. En el análisis inferencial se controlaron covariables y se modelaron efectos de clúster conforme al diseño. Los registros de dosificación IA/RV alimentaron los análisis de dosis–respuesta, distinguiendo exposición (cantidad) de desempeño (resultado).

1.6 Campo de acción

El campo de acción comprendió el contexto académico del Tecnológico de Monterrey durante el ciclo 2024–2025. La intervención se implementó en seis clases de pregrado, estratificadas por programa académico y seleccionadas para preservar la operación natural del curso. Cada clase mantuvo su currículo, docente titular y calendario, mientras que el equipo de investigación coordinó lineamientos comunes de implementación, seguimiento y control de calidad, con el fin de asegurar consistencia operativa entre grupos y trazabilidad de las decisiones.

La muestra fue de 180 estudiantes (alrededor de 30 por clase), provenientes de seis grupos impartidos en el periodo señalado. Para sostener comparabilidad entre grupos y mantener condiciones reales de curso, la selección de clases se realizó mediante un esquema estratificado por programa, utilizando listas oficiales de grupos elegibles y

criterios de equivalencia curricular, incluyendo asignatura o asignaturas comparables en objetivos formativos, nivel y carga de trabajo. Una vez definido el marco de clases elegibles, la inclusión de grupos se efectuó con base en muestreo aleatorio estratificado, de modo que la representación por programa responda a la distribución real de la oferta y minimice sesgos por composición.

La asignación a condición de intervención o comparación se efectuó a nivel de clase (conglomerado), con el propósito de reducir el riesgo de contaminación entre estudiantes y evitar interferencias derivadas del intercambio de materiales, prácticas o retroalimentación entre condiciones. Este enfoque permitió preservar la validez ecológica del estudio al mantener el funcionamiento regular del curso, al tiempo que facilitó el tratamiento analítico de la estructura jerárquica inherente a la implementación por grupos. Para sostener la interpretación de diferencias entre condiciones, el estudio contempló la medición de línea base y el registro de covariables relevantes del perfil estudiantil y académico, las cuales se utilizaron para describir la equivalencia inicial y, cuando corresponda, ajustar los análisis.

La fidelidad de implementación de la arquitectura IA+RV se monitoreó de manera sistemática durante el periodo de intervención, con el fin de distinguir entre ausencia de efecto pedagógico y variación atribuible a ejecución heterogénea. En particular, la fidelidad se documentó mediante tres componentes complementarios: dosificación efectiva de IA y RV por sesión y por estudiante (incluyendo minutos de exposición, escenarios utilizados y eventos de retroalimentación formativa), adherencia a la secuencia instruccional prevista (cumplimiento de pasos, tiempos y productos esperados) y calidad percibida por el alumnado mediante un instrumento estandarizado en escala Likert de 1 a 5. Estos registros se integraron en una bitácora analítica de trazabilidad para documentar decisiones operativas, incidencias de implementación y criterios de control de calidad, con el objetivo de fortalecer la auditabilidad del estudio y respaldar interpretaciones válidas sobre la contribución pedagógica de la arquitectura IA+RV al desempeño en comunicación y solución de problemas.

1.7 Objetivos

1.7.1 Objetivo general

Diseñar un sistema de intervención didáctica con inteligencia artificial y realidad virtual para contribuir a la mejora del desempeño en competencias transversales de comunicación y solución de problemas mediante ciclos de práctica situada, retroalimentación criterial y reflexión en estudiantes de pregrado del Tecnológico de Monterrey, México, durante el periodo 2024 a 2026.

1.7.2 Objetivos específicos

1. Estimar el cambio pre–post en comunicación y solución de problemas en intervención vs. comparación, controlando línea base y efectos de clúster.
2. Cuantificar fidelidad de implementación y dosis IA/RV, y evaluar su asociación con resultados (patrones dosis–respuesta).
3. Examinar la consistencia interjueces (ICC) y la evidencia psicométrica de las métricas de desempeño basadas en rúbricas.
4. Explorar la experiencia de estudiantes y docentes respecto de mecanismos plausibles (autorregulación, presencia/inmersión y carga cognitiva) mediante entrevistas y observación participante.
5. Identificar implicaciones instruccionales para el escalamiento responsable de IA y RV en cursos regulares, considerando infraestructura, formación docente y criterios de gobernanza.
6. Elaborar una propuesta de implementación escalable de la arquitectura didáctica IA+RV para cursos regulares, integrando evidencia cuantitativa y cualitativa, criterios de fidelidad de implementación y salvaguardas de equidad y gobernanza.
7. Validar la propuesta de implementación escalable de la arquitectura didáctica IA+RV mediante juicio de expertos, considerando criterios de pertinencia, validez, factibilidad, aplicabilidad y generalización.

1.8 Hipótesis

Hipótesis general (HG). La integración instruccional de IA y RV, implementada como arquitectura didáctica en cursos regulares, se asociará con un desempeño superior en competencias transversales (comunicación y solución de problemas) respecto de cursos equivalentes sin dicha integración. Adicionalmente, la evidencia cuantitativa y cualitativa del estudio permitirá identificar condiciones de implementación, dosificación y mecanismos plausibles, suficientes para derivar una propuesta proactiva de adopción y escalamiento responsable en educación superior.

Hipótesis específicas.

H1 (Comunicación). El grupo IA+RV presentará una ganancia pre–post en comunicación mayor que la observada en el grupo de comparación.

H2 (Solución de problemas). El grupo IA+RV presentará una ganancia pre–post en solución de problemas mayor que la del grupo de comparación.

H3 (Dosis–respuesta). Una mayor dosificación de IA y/o RV se asociará con mejoras superiores en ambas competencias.

H4 (Moderación). La experiencia tecnológica previa y el programa académico moderarán los efectos de la intervención.

Las hipótesis se contrastarán mediante modelos consistentes con la estructura jerárquica de los datos y con mediciones repetidas, incorporando la dosificación como predictor y covariables definidas a priori.

1.9 Alcance y limitaciones

1.9.1 Alcances del estudio

El estudio alcanzó la evaluación de una arquitectura didáctica que integra inteligencia artificial (IA) y realidad virtual (RV) para el desarrollo de dos competencias transversales en estudiantes de pregrado: comunicación y solución de problemas, por su relevancia para la empleabilidad y el desempeño profesional en contextos intensivos en colaboración y toma de decisiones (Ngo, 2024; Robles, 2012). El alcance de medición se

concentró en desempeños observables mediante productos auténticos evaluados con rúbricas analíticas y doble calificación, incorporando evidencia de consistencia interjueces y trazabilidad del proceso de evaluación.

El alcance institucional se circunscribió a cursos regulares del Tecnológico de Monterrey durante el ciclo 2024–2025, manteniendo condiciones reales de operación académica (currículo, docente titular y calendario) y procedimientos estandarizados de implementación para las actividades con IA y RV. En términos de entregables, el estudio produjo: (i) estimaciones del efecto pre–post de la arquitectura IA+RV en ambas competencias, comparadas contra una condición equivalente sin integración; (ii) evidencia técnica de calidad de medición mediante rúbricas (consistencia interjueces y documentación de criterios); (iii) un registro sistemático de implementación y dosificación de IA y RV; (iv) un análisis integrado cuantitativo y cualitativo orientado a identificar mecanismos plausibles y condiciones operativas; y (v) una propuesta proactiva para implementación y escalamiento responsable en cursos regulares, fundamentada en la evidencia del estudio y en restricciones realistas de infraestructura, formación docente y gobernanza institucional.

El horizonte de trascendencia se orientó a proveer criterios aplicables para la toma de decisiones curriculares y el diseño de estrategias de adopción responsable de IA y RV en educación superior, con transferibilidad preferente hacia instituciones con condiciones tecnológicas y organizacionales comparables, y con lineamientos explícitos para adaptación en escenarios con restricciones (Trisnawati *et al.*, 2023; INEGI, 2024). En coherencia con los criterios de delimitación del alcance en investigación aplicada, este apartado explicita el nivel de generalización esperable a partir del contexto y del diseño de implementación (Hernández Sampieri *et al.*, 2022).

1.9.2 Limitaciones del estudio

Las limitaciones del estudio se relacionan principalmente con condiciones de selección, estructura de implementación, disponibilidad de información y restricciones de tiempo y tecnología. En términos de población y muestra, la población de referencia estuvo conformada por estudiantes inscritos en cursos regulares de pregrado elegibles

durante el ciclo 2024–2025, y la muestra aproximada fue de 180 estudiantes distribuidos en seis clases de alrededor de 30 participantes por grupo. La construcción de la muestra fue realizada de manera secuencial para asegurar transparencia y minimizar sesgos de selección, iniciando con la definición del marco muestral a partir de listas oficiales de grupos elegibles; posteriormente se estableció la estratificación por programa académico con el fin de asegurar representación proporcional; después fueron seleccionadas aleatoriamente las clases dentro de cada estrato; y finalmente se asignaron las clases a condición de intervención o comparación a nivel de conglomerado, preservando la operación natural del curso. Este procedimiento corresponde a un muestreo probabilístico estratificado aplicado a unidades agrupadas, adecuado cuando la unidad operativa real es el grupo y cuando se busca reducir la contaminación entre participantes, aunque introduce limitaciones propias de la ausencia de aleatorización individual y de la posibilidad de diferencias no observadas entre clases (Otzen & Manterola, 2017).

Dado que la implementación se realizó por conglomerados, la dependencia intraclase constituye una fuente potencial de sesgo si no se tratara de forma explícita, ya que estudiantes dentro de una misma clase comparten exposiciones, dinámicas instruccionales y prácticas docentes que pueden correlacionar los resultados y afectar la precisión de las inferencias. En consecuencia, el análisis consideró la estructura jerárquica inherente a los datos mediante modelación multinivel y estimación de efectos asociados a la clase, reconociendo que la variación atribuible a diferencias entre grupos puede coexistir con la variación individual. Aun con estos ajustes, se mantiene la limitación típica de los diseños por conglomerados relacionada con potencia estadística y sensibilidad a variación entre docentes y grupos, especialmente cuando el número de clústeres es acotado.

Adicionalmente, el control de diferencias iniciales y de heterogeneidad entre grupos dependió de la disponibilidad y calidad de covariables preespecificadas. En este estudio se incluyeron covariables definidas a priori que pueden influir en el desempeño observado, particularmente el puntaje basal en comunicación y solución de problemas obtenido en la medición PRE, la experiencia tecnológica previa entendida como

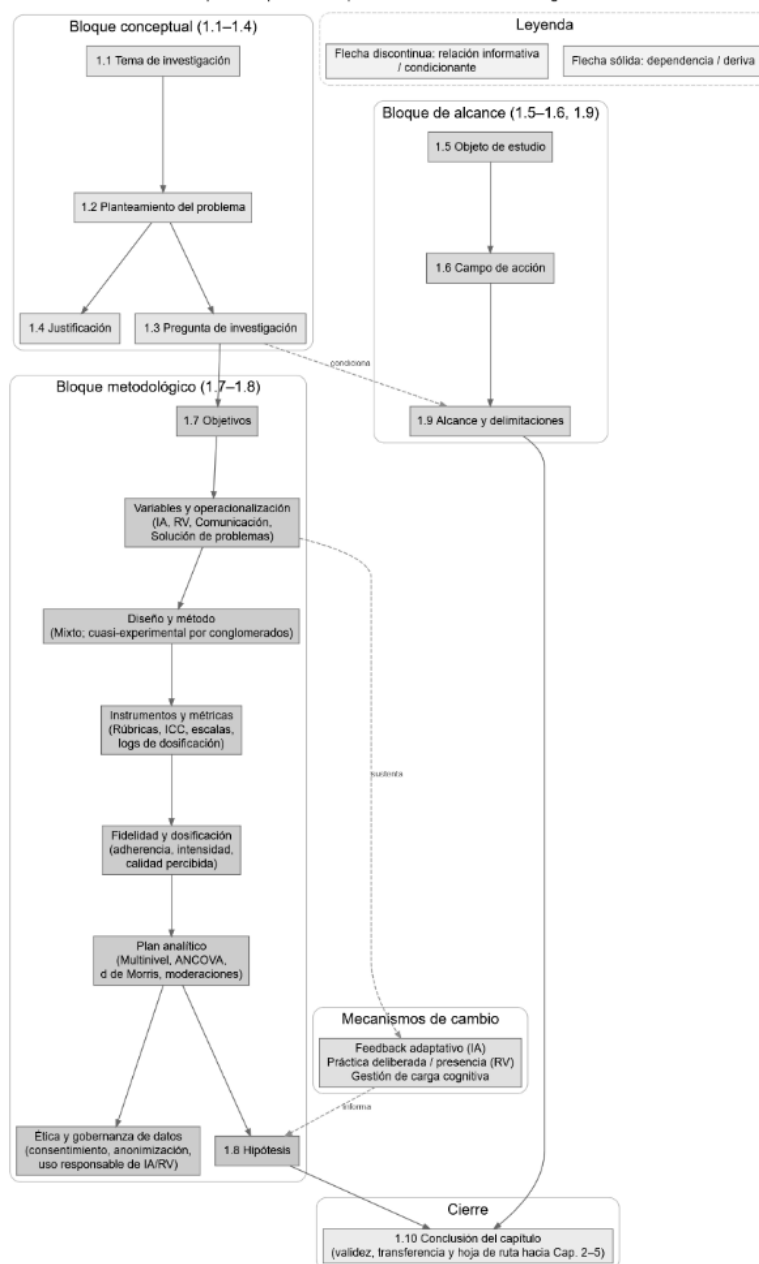
familiaridad y uso habitual de herramientas digitales relevantes para el curso, y el programa académico como aproximación a trayectorias formativas y demandas curriculares distintas. Estas variables se utilizaron para describir la equivalencia inicial y, cuando correspondió, ajustar los modelos comparativos; sin embargo, su inclusión no elimina la posibilidad de factores no medidos que también afecten el desempeño, como diferencias en estilos de enseñanza o motivación diferencial entre grupos.

En cuanto al tiempo, la delimitación al ciclo 2024–2025 permitió observar cambios pre–post durante un periodo académico completo con exposición repetida a IA y RV, pero no se incorporaron mediciones de seguimiento longitudinal a 6–12 meses, por lo que la evidencia se concentró en el cambio observado dentro del periodo de intervención. Esta restricción temporal limitó inferencias sobre estabilidad y transferencia en el mediano plazo, las cuales se reservaron como línea futura de investigación.

Finalmente, existen limitaciones vinculadas a recursos, tecnología e información de implementación. La documentación de la exposición a IA y RV se sostuvo principalmente en registros de plataforma y bitácoras docentes estructuradas por sesión.

1.10 Síntesis visual del protocolo

Este capítulo presenta el protocolo de investigación y organiza los componentes que estructuran el estudio: delimitación del problema, pregunta de investigación, justificación, objeto de estudio, campo de acción, objetivos, hipótesis, alcances y limitaciones. Asimismo, ubica la investigación en literatura reciente sobre IA y RV en educación superior, con énfasis en las condiciones de implementación y en la variación de resultados cuando la tecnología se integra como arquitectura instruccional con criterios explícitos (Trisnawati et al., 2023; Chen & Lee, 2024; Di Palma et al., 2025; Makransky & Petersen, 2021). Como cierre visual, la Figura 1.2 presenta un mapa conceptual que sintetiza la arquitectura lógica del estudio, mostrando la relación entre problema, objetivos, hipótesis, métricas, procedimientos de medición, criterios de análisis, control de calidad y consideraciones éticas.

Figura 1.2*Mapa conceptual del Protocolo de Investigación*

Nota: Flecha sólida = dependencia/derivación; flecha discontinua = relación informativa/ condicionante.
 Bloques: conceptual (1.1–1.4), alcance (1.5–1.6, 1.9), metodológico (1.7–1.8 y piezas) y cierre (1.10).
 Gómez Tobías (2026). Tesis doctoral UIIX.

CAPÍTULO 2. FUNDAMENTOS TEÓRICOS REFERENCIALES

Este capítulo presenta los fundamentos conceptuales, empíricos y referenciales que sustentan el análisis de la relación entre inteligencia artificial (IA), realidad virtual (RV) y el desarrollo de competencias transversales, en particular comunicación y solución de problemas, en estudiantes de pregrado en una institución universitaria mexicana de carácter multicampus. Se privilegia literatura especializada publicada principalmente entre 2019 y 2025, incorporando también fuentes previas cuando aportan fundamentos teóricos, definiciones consolidadas o antecedentes necesarios para interpretar el panorama contemporáneo. La sección 2.1 sintetiza el estado del arte. La sección 2.2 desarrolla el marco teórico en prosa a partir de dos ejes, constructivismo y aprendizaje experiencial, y carga cognitiva y aprendizaje multimedia, integrando aportes complementarios sobre aprendizaje inmersivo, práctica deliberada, retroalimentación formativa y criterios de integración tecnopedagógica, y culminando en una proposición teórica integradora. La sección 2.3 precisa el marco conceptual operativo. A partir de 2.4 se desarrolla el marco referencial: 2.4 presenta el marco histórico, 2.5 delimita el marco contextual y actual, y 2.6 establece el marco legal y normativo aplicable. En conjunto, el capítulo sustenta las decisiones metodológicas y de medición que se desarrollan en los capítulos posteriores.

2.1 Estado del arte

La evidencia contemporánea situó a la RV inmersiva y a la IA educativa como palancas relevantes para el aprendizaje en educación superior. La revisión sistemática de Radianti *et al.* (2020) sintetizó usos de RV en múltiples disciplinas y reportó beneficios en práctica segura, participación y transferencia, junto con retos de diseño instruccional (alineación tarea–objetivo), carga cognitiva y sostenibilidad operativa. Metaanálisis y revisiones recientes convergieron en mejoras de magnitud pequeña–moderada en desempeño y retención frente a métodos tradicionales, con efectos mayores cuando el diseño controló la carga extrínseca, ofreció andamiajes explícitos y aseguró fidelidad funcional de la simulación (Parong & Mayer, 2018; Makransky & Petersen, 2021; Çeken & Taşkın, 2024).

En habilidades de comunicación, la evidencia reciente en RV apuntó a mejoras en desempeño de presentación y variables asociadas (p. ej., autoeficacia y ajuste a la audiencia) cuando la práctica situada incorporó criterios explícitos y ciclos estructurados de ensayo–retroalimentación–revisión, mediando el rol de la presencia en el rendimiento (Di Palma *et al.*, 2025). En paralelo, revisiones focalizadas en entrenamiento de habilidades blandas con RV sintetizaron tendencias y brechas metodológicas, destacando la necesidad de definir operacionalmente las competencias, describir la intervención con precisión y fortalecer los enfoques de evaluación (Dubiel *et al.*, 2025). Este patrón reforzó que la RV no operó por sí sola, sino en conjunción con diseño instruccional explícito, criterios observables y evaluación criterial.

Respecto a IA en educación superior, mapeos de alcance y directrices (UNESCO, 2023; Varsik & Vosberg, 2024) documentaron impactos consistentes en personalización, retroalimentación rápida y analítica del aprendizaje, a la vez que advirtieron sobre privacidad, seguridad, sesgos y formación docente como condiciones habilitantes. En particular, estos marcos subrayan que el beneficio pedagógico no proviene de la automatización *per se*, sino de su integración en un diseño instruccional con criterios explícitos y mecanismos de rendición de cuentas. Asimismo, recomiendan transparencia sobre el rol de la IA (qué hace y qué no hace) y delimitación de usos de alto impacto, para evitar sustitución indebida del juicio docente y efectos adversos en equidad. Finalmente, enfatizan que la evaluación debe considerar no solo desempeño, sino también de calidad de datos, trazabilidad y riesgos operativos como parte del ciclo de mejora. En América Latina, los informes de política educativa reconocieron el potencial para reducir brechas siempre que la adopción fuera acompañada de conectividad, capacitación y gobernanza (UNESCO, 2021).

A nivel institucional, el Tecnológico de Monterrey acumuló experiencias con IA y RV aplicadas a comunicación y toma de decisiones, con rúbricas analíticas y doble evaluación; esta tesis retomó y escaló dichas prácticas en un diseño cuasi-experimental por conglomerados (Gómez, 2026; Robles, 2012; Ngo, 2024).

Como síntesis operativa de esta revisión, se organizó la evidencia más pertinente en una matriz comparativa que concentró: (a) el contexto y tipo de estudio, (b) los

hallazgos principales sobre aprendizaje y desempeño, (c) las condiciones bajo las cuales los efectos fueron mayores (gestión de carga, andamiajes, rúbricas y *debriefing*, agencia del estudiante, gobernanza de datos) y (d) las limitaciones declaradas (heterogeneidad de diseños, *novelty effect*, métricas de presencia, sesgos y brechas de conectividad). La Tabla 2.1 resumió estas fuentes anclando priorizando trabajos de alto nivel de síntesis (revisiones, metaanálisis, directrices) y estudios empíricos recientes y explicitó cómo informaron decisiones de esta tesis: uso de rúbricas analíticas y doble evaluación, incorporación de *briefings/debriefings* y señalización para controlar carga, registro de dosificación IA/RV para modelar dosis–respuesta, y adopción de salvaguardas éticas alineadas con UNESCO/OCDE.

Tabla 2.1

Estudios y revisiones recientes sobre IA/RV en educación superior

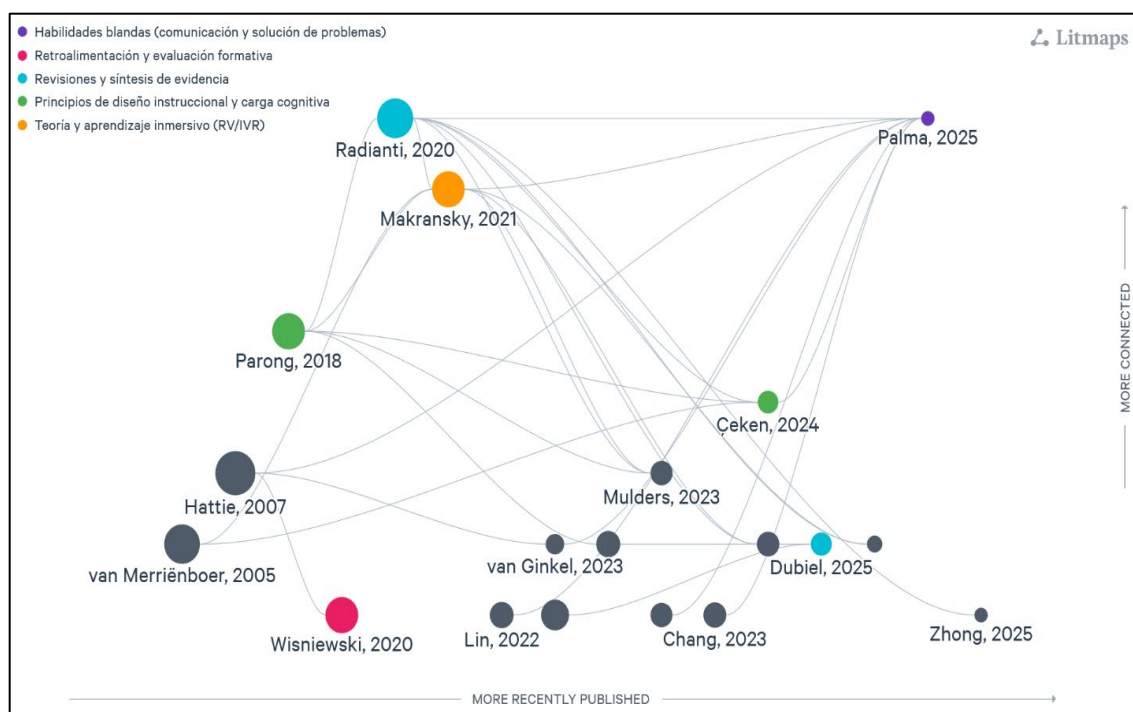
Referencia	Contexto / muestra	Tipo de evidencia	Hallazgos principales	Condiciones de mayor efecto	Limitaciones señaladas
Radianti <i>et al.</i> (2020)	Educación superior (multidisciplinaria)	Revisión sistemática	RV apoya práctica segura, compromiso y transferencia.	Tareas con propósito claro; evaluación criterial; <i>debriefing</i> .	Carga cognitiva; heterogeneidad de diseños; sostenibilidad.
Parong & Mayer (2018)	Aprendizaje de ciencias	Experimentos controlados	IVR eleva presencia/motivación; el aprendizaje mejora con estrategias generativas.	Gestión de carga extrínseca; guías antes/durante/después.	<i>Novelty effect</i> ; variabilidad en medición de logro.
Makransky & Petersen (2021)	IVR y aprendizaje	Modelo teórico + evidencia	CAMIL: inmersión/control → presencia/agencia → interés, autoeficacia, carga → aprendizaje.	Control de interfaz; <i>feedback</i> ; agencia del estudiante.	Falta de métricas estándar de presencia útil.
Çeken & Taşkın (2024)	AR/VR en entornos de aprendizaje	Estudio/Análisis empírico (principios multimedia)	Examina principios de aprendizaje multimedia aplicados a AR/VR; aporta criterios de diseño cognitivo para reducir carga y orientar aprendizaje.	Señalización; segmentación; coherencia; diseño guiado por principios multimedia.	Dependencia del contexto/tarea; variabilidad de implementaciones AR/VR.
Dubiel <i>et al.</i> (2025)	Entrenamiento de <i>soft skills</i> con RV (33 estudios)	Revisión (PRISMA) + análisis cualitativo	Identifica tendencias, enfoques de intervención y métodos de evaluación; mapea brechas y oportunidades en RV para habilidades blandas.	Definición operacional de habilidades; diseño explícito de intervención; evaluación robusta y transparente.	Heterogeneidad; limitaciones en evaluación; necesidad de mayor validez interna/replicabilidad.
Di Palma <i>et al.</i> (2025)	Habilidades de presentación en estudiantes	Estudio empírico	Evalúa RV para habilidades de presentación; destaca rol de <i>feedback</i> y presencia en el desempeño.	Ciclos de práctica deliberada; <i>feedback</i> accionable; criterios explícitos.	Muestras acotadas; posible sesgo de selección; dependencia de implementación.
UNESCO (2023); Varsik & Vosberg, 2024	Política educativa	Directrices	IA responsable: privacidad, transparencia, formación docente y gobernanza.	Marcos éticos; gobernanza de datos; alfabetización en IA.	Implementación desigual; brechas de conectividad/capacitación.

Nota. La tabla sintetiza fuentes que guiaron decisiones de esta tesis: integración IA+RV con rúbricas, andamiajes y gestión de carga; además, salvaguardas éticas y de datos para adopción responsable.

Con el fin de visualizar la estructura de conexión por citación entre la literatura ancla sobre aprendizaje inmersivo (RV/IVR), diseño instruccional, retroalimentación formativa y habilidades blandas, se elaboró un mapa de literatura conectada en Litmaps (Figura 2.1). El mapa se construyó a partir de artículos seminales y revisiones clave (p. ej., Radianti *et al.*, 2020; Makransky & Petersen, 2021; Dubiel *et al.*, 2025) y se complementó con literatura reciente en Journal of Computer Assisted Learning, incluyendo evidencia sobre habilidades de presentación mediadas por RV y el rol de la retroalimentación y la presencia (Di Palma *et al.*, 2025), así como evidencia sobre la aplicación de principios de aprendizaje multimedia en entornos AR/VR (Çeken & Taşkın, 2024).

Figura 2.1

Mapa de conexiones por citación (2020–2025).



Nota. Visualización generada con Litmaps a partir de artículos ancla del marco teórico y ampliada con artículos relacionados según los filtros temporales y de búsqueda aplicados. Los colores representan etiquetas temáticas definidas por el autor y los nodos grises, artículos conectados dentro del rango temporal. La figura se incluye con propósito descriptivo y no se presenta como análisis bibliométrico formal. Gómez Tobías (2026). *Tesis doctoral UIIX*.

2.2 Marco teórico

Este apartado desarrolla el cuerpo teórico que permite interpretar la integración didáctica de la inteligencia artificial (IA) y la realidad virtual (RV) como una arquitectura instruccional orientada al desarrollo de competencias transversales, específicamente comunicación y solución de problemas, en estudiantes de pregrado. El marco se construye desde una premisa explícita: la tecnología no opera como causa automática del aprendizaje. Su efecto es contingente y depende del diseño instruccional, de la calidad de la práctica, de las condiciones cognitivas y afectivas que el entorno induce y de la consistencia con la que se sostienen criterios, secuencias y salvaguardas institucionales. En consecuencia, el marco teórico no se presenta como un inventario de teorías, sino como un argumento articulado en dos ejes que permiten explicar por qué una intervención IA+RV puede producir mejora, bajo qué condiciones, y por qué esos resultados pueden ser heterogéneos cuando dichas condiciones no se cumplen.

2.2.1 Constructivismo y aprendizaje experiencial

Desde una perspectiva constructivista, el aprendizaje experiencial y la construcción social del desempeño ofrecen un fundamento para comprender el desarrollo de competencias transversales. En esta tradición, el aprendizaje no se concibe como una transferencia lineal de información del docente al estudiante, sino como un proceso de construcción activa que se produce cuando la persona enfrenta situaciones problemáticas, contrasta sus comprensiones previas con las demandas de la tarea y reorganiza sus esquemas mediante ajustes progresivos. En la perspectiva de Piaget, el cambio conceptual se activa cuando surge un desequilibrio cognitivo: se reconoce que el esquema disponible no es suficiente para resolver el problema y, en consecuencia, se realizan procesos de asimilación y acomodación que reestructuran la comprensión (Piaget, 1970). Este planteamiento resulta especialmente pertinente para competencias complejas, porque la comunicación efectiva y la solución de problemas no dependen de memorizar reglas aisladas, sino de reorganizar criterios, estrategias y decisiones en función de contextos variables y restricciones reales.

Vygotsky amplía este fundamento al subrayar que el aprendizaje es inseparable de la mediación social y cultural. Las herramientas simbólicas, el lenguaje, las normas de interacción y la participación guiada permiten internalizar formas de razonamiento y desempeño que el estudiante no alcanzaría de manera aislada. Desde esta mirada, la zona de desarrollo próximo explica por qué la presencia del docente y de pares no es un elemento accesorio, sino un mecanismo constitutivo del avance: el aprendizaje se acelera cuando el estudiante recibe apoyo oportuno para operar por encima de su nivel actual de independencia (Vygotsky, 1978). En competencias transversales, este principio adquiere un matiz operativo. La mejora ocurre cuando el estudiante dispone de criterios explícitos para distinguir calidad, cuando recibe retroalimentación que reduce ambigüedad y cuando puede ensayar nuevamente bajo guía, en lugar de enfrentar evaluaciones finales sin ciclos formativos.

Este fundamento social y de criterio es central para el tipo de competencias que se analizan en esta tesis. Comunicación y solución de problemas son desempeños multicomponente: su calidad depende de la coherencia interna del argumento, la selección y uso de evidencia, la adecuación a la audiencia, la claridad bajo presión, así como de la capacidad para identificar supuestos, generar alternativas plausibles, anticipar consecuencias y justificar decisiones. En ese sentido, no basta con participar o hablar más para mejorar comunicación, ni con resolver para mejorar solución de problemas. Se requiere que el estudiante aprenda a reconocer estándares de excelencia y a traducir esos estándares a acciones observables durante la práctica. Por ello, el constructivismo, aplicado a un enfoque por competencias, se expresa como una exigencia de diseño: tareas auténticas, oportunidades de interacción y mediación, y mecanismos explícitos para reconstruir el desempeño con base en criterios, evidencias y reflexión.

El aprendizaje experiencial de Kolb profundiza este argumento al describir el desarrollo de competencias como un ciclo dinámico que integra experiencia concreta, observación reflexiva, conceptualización y experimentación activa (Kolb, 1984). Esta formulación permite comprender por qué entornos de simulación, como la RV, pueden ser relevantes para competencias transversales: habilitan experiencias situadas que

exigen desempeño, no solo comprensión declarativa. No obstante, el propio marco experiencial establece un límite crítico: la experiencia no garantiza aprendizaje si no se convierte en reflexión estructurada y conceptualización. En otras palabras, el ciclo se rompe si se privilegia la vivencia inmersiva sin mecanismos para observarse, recibir retroalimentación, abstraer principios y reintentar con intención formativa. Esta advertencia es particularmente importante en RV porque el realismo percibido puede inducir la ilusión de competencia: el estudiante puede sentirse eficaz por estar inmerso y activo, aun cuando su desempeño no cumpla criterios de calidad. Por tanto, el valor pedagógico de la inmersión se comprende mejor como la combinación entre autenticidad de la tarea, explicitación de estándares, oportunidades de ensayo y espacios de reflexión guiada que convierten la experiencia en conocimiento operativo y transferible (Parong & Mayer, 2018; Slater & Sanchez-Vives, 2016).

Desde este eje, la RV se interpreta como un dispositivo que puede potenciar principios constructivistas siempre que funcione como escenario para el aprendizaje situado y no como actividad de alta intensidad sin guía. Cuando se diseña apropiadamente, la RV puede aproximar demandas profesionales de comunicación y toma de decisiones: obliga a organizar un mensaje para una audiencia, ajustar el tono, seleccionar evidencia, gestionar el tiempo y responder a contingencias. A la vez, puede crear situaciones problemáticas con ambigüedad realista, donde el estudiante deba definir el problema, identificar restricciones, evaluar alternativas y justificar trade-offs. Sin embargo, el mismo potencial inmersivo puede volverse contraproducente si la tarea no incorpora mediaciones que orienten el aprendizaje. En consecuencia, este marco exige que la experiencia inmersiva se estructure con *briefings* que clarifiquen objetivos y criterios, con oportunidades de interacción guiada y con *debriefings* que articulen reflexión y reconstrucción del desempeño, evitando que la inmersión se reduzca a exposición.

En esta lógica, la integración de IA y RV se interpreta como una forma de intensificar condiciones constructivistas, pero bajo un principio rector: toda tecnología debe estar subordinada a procesos formativos verificables. La IA, en particular, puede ampliar la mediación y el andamiaje cuando se utiliza para sostener retroalimentación

oportuna, identificar patrones de desempeño y guiar la autorregulación, siempre y cuando lo haga con criterios transparentes y con límites claros. La mediación no consiste en dar respuestas, sino en ayudar al estudiante a construir mejores decisiones. Por ello, la IA es compatible con el constructivismo cuando favorece la reflexión criterial, la revisión iterativa y la toma de conciencia sobre la brecha entre desempeño actual y esperado. Si, por el contrario, la IA se emplea como mecanismo que sustituye el trabajo de elaboración o que entrega productos listos sin exigir reconstrucción, puede debilitar el aprendizaje al reducir el esfuerzo cognitivo significativo y la agencia del estudiante.

La implicación metodológica de este eje es relevante para esta tesis: si la comunicación y la solución de problemas son competencias complejas, entonces el aprendizaje debe evidenciarse en cambios de desempeño observables, no solo en percepciones de utilidad o satisfacción. Por ello, el enfoque constructivista y experiencial converge con la necesidad de evaluación criterial mediante rúbricas analíticas y con la lógica de doble evaluación cuando se busca robustez. La evaluación no se concibe solo como medición final, sino como mecanismo de aprendizaje: hace explícitos los criterios, habilita retroalimentación y orienta la reconstrucción del desempeño. En esa medida, el constructivismo no solo justifica el uso de tareas auténticas en RV, sino que obliga a diseñarlas de modo que el estudiante reciba señales consistentes de calidad, pueda interpretar esas señales y tenga oportunidades reales de reintentos y mejora.

En síntesis, el constructivismo, la mediación sociocultural y el aprendizaje experiencial fundamentan el primer pilar de este marco: la arquitectura IA+RV es pedagógicamente plausible cuando activa aprendizaje situado y social, promueve reflexión guiada y sostiene criterios explícitos para mejorar desempeño. Este pilar establece que la tecnología solo tiene valor educativo si intensifica condiciones formativas esenciales: autenticidad, mediación, criterios y ciclos de reconstrucción del desempeño. Sobre esa base, el siguiente apartado especifica las condiciones cognitivo-afectivas y de diseño que determinan si esa arquitectura se traduce efectivamente en aprendizaje o si queda limitada por sobrecarga, baja agencia útil o falta de coherencia de integración.

2.2.2 Carga cognitiva y aprendizaje multimedia

La efectividad de una arquitectura IA y RV depende de condiciones de diseño cognitivo y afectivo, de la calidad de la práctica deliberada y de la coherencia de integración pedagógica. Si el primer eje fundamenta por qué la experiencia situada y la mediación social favorecen el aprendizaje de competencias complejas, el segundo eje especifica las condiciones bajo las cuales una arquitectura IA y RV puede traducirse en mejora verificable o, en su defecto, producir resultados modestos o inconsistentes. Este eje integra tres componentes complementarios: el diseño cognitivo orientado a gestionar la carga cognitiva en entornos ricos en estímulos, la dinámica de mejora gradual mediante práctica deliberada y retroalimentación formativa, y la coherencia pedagógica y responsable de la integración tecnológica en función de objetivos, tareas, criterios de evaluación y salvaguardas.

La Teoría de la Carga Cognitiva aporta una distinción central para interpretar intervenciones con RV y experiencias multimedia: la mente humana opera con una memoria de trabajo limitada, por lo que el aprendizaje se ve afectado por la cantidad y el tipo de carga que una tarea impone. Sweller distingue carga intrínseca, asociada a la complejidad inherente del contenido; carga extrínseca, generada por demandas irrelevantes inducidas por el formato, la interfaz o la navegación; y carga germana, vinculada con el esfuerzo invertido en construir y automatizar esquemas útiles (Sweller, 1988; Sweller *et al.*, 2011). En entornos de RV, la promesa pedagógica se acompaña de un riesgo: la riqueza sensorial, la exploración espacial, el control de dispositivos y la interacción con múltiples estímulos pueden elevar carga extrínseca si el diseño no es funcional. En esos casos, el estudiante puede sentirse altamente involucrado y, aun así, aprender menos, porque la energía mental se consume en orientarse o operar la tecnología en vez de procesar el razonamiento requerido por la competencia.

La investigación en aprendizaje multimedia ofrece principios complementarios para reducir ese riesgo. Mayer plantea que el aprendizaje mejora cuando el diseño elimina información irrelevante, organiza el material en segmentos manejables, utiliza señalización para guiar la atención y considera el preentrenamiento para familiarizar al estudiante con elementos esenciales antes de tareas complejas (Mayer, 2021). En

experiencias inmersivas, estos principios son especialmente relevantes porque el costo cognitivo de la interacción suele ser mayor que en formatos tradicionales. Por ello, el aporte del enfoque cognitivo a esta tesis es interpretar la RV como un entorno que debe ser instruccionalmente gobernado: su efectividad depende de cómo contiene la carga extrínseca y habilita que la atención se dirija a procesos de argumentación, evaluación de evidencia, análisis de alternativas y justificación de decisiones, en lugar de dispersarse en estímulos de baja funcionalidad (Makransky *et al.*, 2019; Parong & Mayer, 2018). En este sentido, la IA también puede operar como un apoyo cognitivo, pero su valor depende de su diseño: la retroalimentación extensa, inespecífica o contradictoria puede incrementar carga extrínseca, mientras que el *feedback* focalizado y accionable puede favorecer carga germana, promoviendo elaboración, autoexplicación y revisión estratégica.

El modelo CAMIL integra esta dimensión cognitiva con variables afectivas propias de la inmersión. Makransky y Petersen proponen que características del entorno inmersivo, como la inmersión, el control y el realismo funcional, favorecen estados psicológicos como presencia y agencia. Estos estados, a su vez, influyen en variables de interés, autoeficacia, autorregulación y carga cognitiva, que en conjunto afectan el aprendizaje (Makransky & Petersen, 2021). CAMIL es particularmente pertinente para competencias transversales porque comunicación y solución de problemas implican un componente emocional y de control percibido: claridad bajo presión, manejo de incertidumbre, ajuste a la audiencia, persistencia ante errores y regulación de atención. La RV puede intensificar esos estados, elevando interés y compromiso; sin embargo, CAMIL también explica por qué el resultado puede ser ambivalente. Si la inmersión incrementa presencia e interés, pero también eleva carga extrínseca, el desempeño puede deteriorarse. Por ello, CAMIL converge con la teoría de carga cognitiva en un principio operativo: el aprendizaje inmersivo exige equilibrio entre presencia útil, agencia controlable y demanda cognitiva contenida (Makransky *et al.*, 2019; Parong & Mayer, 2018). Esta convergencia es clave para interpretar variabilidad de resultados, porque permite comprender por qué dos intervenciones con RV pueden diferir: no basta con usar RV; importa cómo se diseña la experiencia para que el involucramiento psicológico se traduzca en aprendizaje y no solo en novedad tecnológica.

La segunda dimensión de este eje es la lógica de mejora gradual sustentada en práctica deliberada y retroalimentación. Ericsson define la práctica deliberada como entrenamiento orientado a metas específicas, centrado en subcomponentes críticos, con estándares claros, retroalimentación correctiva oportuna y repetición sistemática (Ericsson *et al.*, 1993/2006). En competencias transversales, esta perspectiva es especialmente robusta, porque el desempeño experto no emerge por exposición pasiva, sino por refinamiento progresivo de componentes como estructura argumental, selección y uso de evidencia, anticipación de consecuencias, consistencia lógica y calidad de justificación. La práctica deliberada exige condiciones que rara vez se sostienen de manera espontánea: criterios explícitos, oportunidades reiteradas de ensayo, *feedback* que apunte a acciones concretas y tiempo para iterar.

La literatura sobre retroalimentación refuerza este punto: el *feedback* es potente cuando especifica qué mejorar, por qué y cómo, y cuando reduce la brecha entre desempeño actual y esperado al alinearse con estándares claros y comprensibles para el estudiante (Hattie & Timperley, 2007; Shute, 2008; Wiggins, 2012). Desde esta perspectiva, la IA puede ampliar la frecuencia y oportunidad del *feedback*, particularmente en escenarios donde el docente no puede retroalimentar cada microintento. Sin embargo, el marco también introduce cautelas: si la IA se utiliza de forma opaca, si produce recomendaciones plausibles pero incorrectas, o si induce dependencia, puede debilitar competencias complejas al reducir la deliberación del estudiante. Por ello, los marcos internacionales recientes recomiendan transparencia sobre el rol de la IA, supervisión humana, delimitación de usos de alto impacto y alfabetización para que estudiantes comprendan límites, riesgos y criterios de uso responsable (UNESCO, 2023; Varsik & Vosberg, 2024). En esta tesis, ello implica que la retroalimentación asistida por IA se conceptualiza como apoyo formativo sujeto a diseño y gobernanza, no como sustituto del juicio docente ni como mecanismo para automatizar evaluación.

La tercera dimensión del eje es la coherencia de integración pedagógica. TPACK sostiene que una integración educativa robusta requiere consistencia entre conocimiento disciplinar, pedagógico y tecnológico; la tecnología debe amplificar intenciones

instruccionales, no dictarlas (Mishra & Koehler, 2006). SAMR, por su parte, distingue entre adopciones instrumentales y niveles de transformación donde la tecnología modifica o redefine tareas, habilitando experiencias que no serían viables sin ella (Puentedura, 2010; Hamilton *et al.*, 2016). Aplicados al campo de IA y RV, estos marcos permiten interpretar por qué intervenciones con herramientas avanzadas pueden generar efectos modestos si se integran como accesorio, y por qué otras intervenciones muestran mejoras cuando la tecnología se convierte en arquitectura instruccional completa, con secuencias, criterios, práctica y revisión. Esta coherencia también incorpora responsabilidad institucional: gobernanza de datos, transparencia, mitigación de sesgos, protección de bienestar y mecanismos de rendición de cuentas, tal como lo recomiendan UNESCO y OCDE en sus directrices para adopción responsable (UNESCO, 2021, 2023; Varsik & Vosberg, 2024). En suma, el criterio no es la presencia de IA o RV, sino la congruencia entre objetivos de aprendizaje, diseño de tareas, evaluación criterial, retroalimentación y salvaguardas.

Desde la integración de los componentes anteriores se deriva la Proposición Teórica Integradora de este estudio. La PTI organiza una cadena explicativa plausible sobre cómo una arquitectura didáctica basada en IA y RV puede influir en el desarrollo de comunicación y solución de problemas. El punto de partida es la coherencia pedagógica: cuando la integración tecnológica se diseña de manera consistente con TPACK y se orienta a niveles transformacionales en SAMR, las tareas, criterios y secuencias se alinean de forma verificable. Bajo esa coherencia, el diseño instruccional puede contener carga extrínseca mediante principios de aprendizaje multimedia y decisiones de interfaz y navegación, manteniendo la demanda cognitiva en rangos manejables, como lo anticipa la teoría de carga cognitiva (Sweller, 1988; Sweller *et al.*, 2011; Mayer, 2021). En paralelo, la RV puede incrementar presencia y agencia y, con ello, favorecer interés, autoeficacia y autorregulación, como plantea CAMIL, siempre que la inmersión sea funcional y esté alineada con objetivos y criterios, evitando que el involucramiento psicológico se convierta en distracción (Makrasky & Petersen, 2021; Makrasky *et al.*, 2019). En ese contexto, la IA puede aportar retroalimentación más oportuna y específica, facilitando ciclos iterativos de revisión y mejora, consistentes con la práctica deliberada y con modelos contemporáneos de *feedback* efectivo, siempre bajo

supervisión humana, transparencia y límites para proteger integridad y autonomía del estudiante (Ericsson *et al.*, 1993/2006; Hattie & Timperley, 2007; Shute, 2008; UNESCO, 2023; Varsik & Vosberg, 2024).

La PTI, por tanto, no atribuye el aprendizaje a la tecnología por sí misma; atribuye cambios en desempeño a la interacción entre diseño pedagógico, calidad de práctica, condiciones cognitivas y afectivas del entorno, y gobernanza responsable. Bajo esta proposición, las mejoras en comunicación deberían reflejarse en componentes como claridad y coherencia, estructura argumental, uso de evidencia, adecuación a la audiencia y calidad de entrega. En solución de problemas, los cambios deberían observarse en definición rigurosa del problema, generación y evaluación de alternativas, anticipación de consecuencias y justificación consistente de decisiones. La magnitud y estabilidad de esos cambios se interpretan como dependientes de la fidelidad de implementación, de la dosificación efectiva de práctica y retroalimentación, y del equilibrio entre presencia útil y carga cognitiva. Asimismo, la PTI permite una lectura crítica de resultados heterogéneos: efectos nulos o modestos pueden ser consistentes con sobrecarga extrínseca, baja agencia útil, retroalimentación poco alineada con criterios o integración tecnológica de bajo nivel transformacional. Con ello, el marco teórico cumple una función explicativa y evaluativa adecuada para sustentar el diseño del estudio, y preparar la transición hacia el marco conceptual.

2.3 Marco conceptual

En esta tesis, la inteligencia artificial (IA) se entiende como un campo de métodos computacionales orientados a automatizar y ampliar capacidades asociadas con percepción, predicción, generación de lenguaje y apoyo a la decisión, incluyendo aprendizaje automático, modelos generativos y analítica. En educación superior, su relevancia se vincula con aplicaciones para personalización instruccional, retroalimentación formativa, tutoría asistida y seguimiento del progreso del aprendizaje, siempre bajo criterios de transparencia, supervisión humana e integridad académica (Russell & Norvig, 2021; Chen & Lee, 2024; UNESCO, 2023; Varsik & Vosberg, 2024). Desde una perspectiva contemporánea, la IA educativa no se reduce a la automatización de tareas, sino que se sitúa en un ecosistema sociotécnico donde su valor

depende de la alineación con objetivos pedagógicos, de la calidad del diseño instruccional y de la gobernanza de datos y riesgos. En consecuencia, la IA se concibe aquí como un componente de una arquitectura didáctica más amplia, cuya función es favorecer ciclos de práctica y revisión a partir de criterios explícitos y retroalimentación interpretable.

La literatura doctoral reciente ha ampliado esta comprensión al documentar el uso de enfoques de analítica multimodal y de sistemas de apoyo basados en datos para monitorear procesos de aprendizaje y sostener intervenciones formativas, especialmente cuando se requiere observar trayectorias de desempeño más allá del resultado final. Estos trabajos han contribuido a consolidar la plausibilidad de mecanismos educativos donde la IA actúa como soporte para la autorregulación, la toma de conciencia criterial y la mejora iterativa del desempeño, más que como sustituto del trabajo intelectual del estudiante (Barzola, 2020; Suresh, 2022; Ouhaichi, 2024; Schmucker, 2024). Este enfoque resulta congruente con marcos recientes de política educativa que advierten que el uso responsable de IA debe incorporar límites funcionales, rendición de cuentas y prácticas explícitas de alfabetización y supervisión (UNESCO, 2021, 2023; Varsik & Vosberg, 2024).

De forma complementaria, la realidad virtual (RV) se conceptualiza como un conjunto de entornos inmersivos que generan una experiencia de presencia y agencia mediante interacción en escenarios tridimensionales, permitiendo ensayo y desempeño en situaciones cercanas a demandas profesionales. En educación, la RV se ha asociado con oportunidades de práctica segura, repetible y evaluable en contextos situados, con potencial para incrementar involucramiento y transferencia, siempre que las simulaciones sostengan fidelidad funcional y un diseño instruccional que contenga carga extrínseca y propicie reflexión guiada (Slater & Sanchez-Vives, 2016; Radianti *et al.*, 2020; Parong & Mayer, 2018; Makransky & Petersen, 2021). Esta conceptualización enfatiza que la inmersión no es un fin en sí mismo. Su valor depende de la estructura pedagógica que la acompaña, de la claridad de metas y criterios, y de la articulación entre experiencia, *debriefing* y oportunidades de mejora.

En cuanto al objeto formativo, esta tesis se enfoca en dos competencias transversales críticas para el desempeño profesional: comunicación oral y solución de problemas. La comunicación oral se comprende como un desempeño complejo que integra claridad y coherencia, estructura argumentativa, uso de evidencia, adecuación a la audiencia y componentes de entrega verbal y no verbal. Este enfoque se vincula con marcos de evaluación de competencias ampliamente utilizados en educación superior y con literatura sobre empleabilidad que reconoce la comunicación como un predictor relevante de efectividad profesional en entornos colaborativos y de toma de decisiones (AAC&U, 2010; Robles, 2012; Ngo, 2024). La solución de problemas, por su parte, se conceptualiza como la capacidad de definir con precisión una situación problemática, generar alternativas plausibles, evaluarlas con criterios explícitos, anticipar consecuencias y justificar decisiones de manera consistente. Este constructo es central en contextos organizacionales donde se enfrentan restricciones reales, incertidumbre y trade-offs, y su desarrollo suele requerir práctica deliberada en tareas situadas, con criterios estables de calidad (AAC&U, 2010; Di Palma *et al.*, 2025; Trisnawati *et al.*, 2023).

Finalmente, el concepto articulador del estudio es el de arquitectura didáctica IA+RV. En esta tesis, se entiende como una integración instruccional en la que IA y RV se combinan de manera coherente con objetivos y criterios de evaluación para sostener ciclos de aprendizaje basados en práctica situada, retroalimentación formativa y mejora iterativa del desempeño. Esta definición se alinea con la idea de que la tecnología aporta valor educativo cuando redefine tareas, intensifica oportunidades de práctica y retroalimentación, y mantiene estándares de responsabilidad institucional, más que cuando se incorpora como recurso accesorio o por novedad tecnológica (Mishra & Koehler, 2006; Hamilton *et al.*, 2016; UNESCO, 2023; Varsik & Vosberg, 2024). Con ello, el marco conceptual fija un vocabulario común para interpretar el diseño del estudio y sustenta, en los capítulos metodológicos posteriores, la especificación operativa de variables, instrumentos y procedimientos de análisis.

2.4 Marco histórico

La integración educativa de la inteligencia artificial (IA) y la realidad virtual (RV) se comprende mejor cuando se sitúa en su trayectoria histórica, ya que la disponibilidad de capacidades técnicas, la estabilidad de los dispositivos y la madurez conceptual del campo han delimitado, en distintos periodos, qué usos pedagógicos eran viables y qué promesas resultaban prematuras. Este marco histórico describe la evolución de ambas tecnologías y explica por qué su convergencia ha desplazado gradualmente la lógica de intervención educativa desde recursos de apoyo periférico hacia entornos y sistemas que pueden reconfigurar tareas, intensificar la práctica y ampliar posibilidades de retroalimentación, siempre que el diseño instruccional mantenga control sobre criterios, secuencias y calidad de implementación (Russell & Norvig, 2021; Slater & Sanchez-Vives, 2016).

En el caso de la IA, su desarrollo suele caracterizarse por fases dominantes que han condicionado el tipo de aplicaciones disponibles para educación y, en consecuencia, el modo en que se ha conceptualizado su valor pedagógico. Una primera etapa, asociada con enfoques simbólicos y sistemas expertos, enfatizó la representación explícita del conocimiento mediante reglas, heurísticas y estructuras lógicas. Desde una perspectiva histórica, esta fase alimentó la idea de que parte del razonamiento experto podía modelarse y, por extensión, que podrían diseñarse sistemas capaces de guiar al estudiante mediante secuencias instruccionales programadas. En el ámbito educativo, este enfoque se materializó en tradiciones tempranas de sistemas tutoriales inteligentes y software de instrucción asistida por computadora, efectivos en dominios acotados y con respuestas relativamente determinísticas, pero con limitaciones para competencias transversales, donde el desempeño depende de juicio contextual, argumentación y evaluación de matices (Russell & Norvig, 2021; Woolf, 2010; VanLehn, 2011).

Con el avance de enfoques estadísticos y de aprendizaje automático, la IA transitó hacia un paradigma donde el conocimiento deja de codificarse principalmente como reglas y pasa a inferirse desde datos. Este giro contribuyó a la expansión de aplicaciones orientadas a predicción, clasificación y recomendación, y abrió el camino para el surgimiento y consolidación de la analítica del aprendizaje como un campo que

busca describir, predecir y optimizar procesos educativos a partir de trazas y registros de interacción. Históricamente, este cambio fue relevante porque desplazó parte de la toma de decisiones educativas desde la intuición hacia modelos probabilísticos, aunque con el costo de requerir infraestructura, gobernanza y capacidad institucional para capturar y traducir datos en decisiones pedagógicas consistentes (Siemens & Long, 2011; Ferguson, 2012). Este desplazamiento también se refleja en la consolidación de AI in Education como un campo que articula teoría del aprendizaje, modelado computacional y diseño instruccional, con debates persistentes sobre personalización, agencia y límites pedagógicos de la automatización (Luckin *et al.*, 2016; Zawacki-Richter *et al.*, 2019). En esta fase, la retroalimentación automatizada y la personalización se hicieron más plausibles, pero a menudo permanecieron como apoyos indirectos o diferidos, dependientes de la calidad del diseño y de la madurez operativa del ecosistema educativo.

Desde mediados de la década de 2010, el avance del aprendizaje profundo reconfiguró de manera sustantiva el panorama, al incrementar la capacidad de los modelos para procesar patrones complejos y lenguajes naturales con mayor flexibilidad. Este cambio de paradigma consolidó aplicaciones más robustas en visión, lenguaje y reconocimiento de patrones, con efectos indirectos en educación por la expansión de sistemas capaces de generar explicaciones, sugerencias y apoyo en formatos más cercanos a la comunicación humana. En términos históricos, el aprendizaje profundo funcionó como un habilitador de escala: muchos desarrollos que antes dependían de ingeniería manual de características comenzaron a apoyarse en representaciones aprendidas por redes neuronales, ampliando las posibilidades de automatización y soporte adaptativo (LeCun *et al.*, 2015). Posteriormente, la introducción de arquitecturas tipo transformer intensificó esta tendencia en el procesamiento de lenguaje natural, habilitando modelos que aprenden relaciones de atención a gran escala, lo cual aceleró el desarrollo de sistemas conversacionales y de generación de texto (Vaswani *et al.*, 2017). En educación superior, esta transición ha sido discutida como un punto de inflexión por el potencial de apoyo lingüístico y retroalimentación, pero también por riesgos de validez, dependencia y superficialidad si no se delimitan finalidades y controles, lo que

refuerza la necesidad de supervisión humana y alineación con criterios (Kasneci *et al.*, 2023; UNESCO, 2021; Russell & Norvig, 2021).

En paralelo, la RV siguió una trayectoria tecnológica distinta, marcada por avances en dispositivos, interacción y realismo funcional. Sus antecedentes se remontan a desarrollos experimentales de simulación sensorial y visualización inmersiva, cuyo uso estuvo durante décadas limitado por costos, complejidad técnica y requerimientos de infraestructura. Aun así, desde etapas tempranas se fue consolidando un vocabulario conceptual en torno a la inmersión y la presencia, entendida como la experiencia subjetiva de estar ahí en un entorno mediado. Este concepto se convirtió en un pilar para interpretar el potencial y las restricciones de los entornos inmersivos, no como simple visualización, sino como experiencia psicológica con implicaciones para atención, emoción y control percibido (Steuer, 1992; Slater & Sanchez-Vives, 2016).

Históricamente, la evolución de la RV también se articuló con la discusión sobre el continuo entre entornos reales y virtuales, base conceptual para comprender configuraciones híbridas que combinan elementos físicos y digitales. La formulación del continuo realidad-virtualidad permitió ordenar conceptualmente la realidad aumentada, la realidad mixta y la RV, y clarificar que la diferencia no es únicamente de dispositivo, sino de grado de virtualidad, integración y control perceptivo (Milgram & Kishino, 1994). Esta distinción se volvió especialmente relevante conforme los dispositivos se democratizaron: la disponibilidad de visores comerciales con mejores capacidades de seguimiento y control redujo barreras de acceso y permitió que la RV pasará de ser un recurso de laboratorio a un entorno de entrenamiento con posibilidades de repetición, estandarización y evaluación.

Con la expansión de dispositivos más accesibles y con mejoras en seguimiento de movimiento, controladores e interacción, la RV dejó de funcionar solo como recurso de demostración y comenzó a consolidarse como entorno de desempeño. Este cambio tuvo implicaciones pedagógicas concretas, pues habilitó prácticas seguras, repetibles y evaluables en escenarios donde el ensayo real sería costoso o riesgoso. En educación superior, ello favoreció aplicaciones en simulación de procedimientos, laboratorios virtuales y entrenamiento de situaciones profesionales, incluyendo tareas comunicativas

y toma de decisiones. Sin embargo, la literatura sobre aprendizaje inmersivo y presencia ha sido consistente en advertir que el realismo perceptual y la intensidad sensorial no garantizan aprendizaje si la experiencia no se gobierna instruccionalmente, y que los efectos tienden a depender del diseño, el tipo de tarea y las condiciones de implementación, particularmente por la interacción entre presencia, agencia y carga cognitiva (Merchant *et al.*, 2014; Parong & Mayer, 2018; Makransky *et al.*, 2019; Slater & Sanchez-Vives, 2016).

La convergencia histórica entre IA y RV puede entenderse como la articulación de dos trayectorias tecnológicas complementarias. Mientras la RV aporta un entorno de experiencia situada que puede exigir desempeño bajo condiciones controladas, la IA aporta capacidades para observar patrones, ofrecer apoyo, y sostener retroalimentación y secuencias de práctica. Esta convergencia desplaza la unidad de análisis instruccional desde la sesión como bloque relativamente homogéneo hacia ciclos más cortos de ejecución, retroalimentación y ajuste, con potencial para intensificar aprendizaje activo y deliberado. En términos históricos, este desplazamiento refuerza una idea clave: la tecnología aporta valor educativo cuando permite reorganizar el tiempo pedagógico y aumentar oportunidades de práctica y revisión, más que cuando se incorpora como un accesorio de presentación. Aun así, la historia de la adopción tecnológica en educación muestra que la disponibilidad de herramientas no produce automáticamente mejoras; sin una arquitectura pedagógica explícita, la innovación tiende a derivar en tecnologismo, es decir, uso por novedad sin alineación suficiente con objetivos, criterios y evaluación (Mishra & Koehler, 2006; Hamilton *et al.*, 2016).

Este marco histórico también ayuda a justificar por qué la discusión contemporánea sobre IA y RV se ha acompañado de una creciente atención a ética, gobernanza y responsabilidad institucional. A medida que las tecnologías se vuelven más accesibles y más capaces de intervenir en procesos formativos, aumentan los riesgos asociados al tratamiento de datos, a la transparencia de sistemas y a la integridad académica. En consecuencia, el desarrollo histórico no solo se expresa como progreso técnico, sino como expansión de exigencias de control, supervisión y rendición de cuentas, particularmente en contextos educativos donde la asimetría de poder y la

obligación de protección al estudiante son relevantes (UNESCO, 2021; Varsik & Vosberg, 2024).

En síntesis, la IA ha transitado desde enfoques simbólicos hacia paradigmas basados en datos, aprendizaje profundo y procesamiento de lenguaje a gran escala; la RV ha evolucionado desde prototipos costosos hacia entornos inmersivos más accesibles con mayores posibilidades de interacción y desempeño. La convergencia de ambas trayectorias sustenta la plausibilidad de una arquitectura didáctica IA+RV orientada a competencias transversales, al combinar experiencia situada con apoyo y retroalimentación potencialmente más frecuente. Sobre esta base histórica, el apartado siguiente presenta el marco contextual y actual, donde se describen las condiciones contemporáneas de adopción, gobernanza e implementación en educación superior y en el entorno institucional específico donde se desarrolla el estudio.

2.5 Marco contextual y actual

En esta tesis, el término entorno actual se delimita al periodo 2020–2025, con énfasis en el punto de inflexión observado desde 2023 por la expansión de sistemas de IA generativa basados en modelos de lenguaje y por la mayor disponibilidad operativa de experiencias inmersivas en educación superior. En consecuencia, este marco contextual describe condiciones contemporáneas de adopción, implementación y evaluación de IA y RV en universidades, con atención a requisitos de coherencia pedagógica, calidad de diseño instruccional y gobernanza institucional. Aunque el enfoque temporal corresponde a la ventana definida, se incorporan referencias previas cuando aportan fundamentos canónicos, definiciones consolidadas o hitos necesarios para interpretar el panorama contemporáneo (Radianti *et al.*, 2020; UNESCO, 2021; UNESCO, 2023; Varsik & Vosberg, 2024).

Durante el periodo analizado, la RV inmersiva ha transitado desde usos predominantemente exploratorios hacia implementaciones más sistemáticas en educación superior, especialmente en ámbitos donde la práctica situada, la repetición deliberada y la simulación segura son relevantes. Revisiones sistemáticas y trabajos de síntesis reportan efectos en aprendizaje y desempeño de magnitud pequeña a moderada,

con heterogeneidad sustantiva explicada por diferencias en diseño instruccional, fidelidad funcional de la simulación, estructura de apoyo y calidad de evaluación. La evidencia converge en que los resultados tienden a ser más consistentes cuando la experiencia inmersiva incorpora señalización, segmentación, criterios explícitos de desempeño y secuencias de preparación y reflexión posterior (*briefing* y *debriefing*), evitando que la inmersión se traduzca únicamente en involucramiento subjetivo o en efectos transitorios asociados a novedad (Radianti *et al.*, 2020; Makransky *et al.*, 2019; Makransky & Petersen, 2021; Parong & Mayer, 2018). En competencias transversales, y particularmente en comunicación, estudios empíricos recientes sugieren que la mejora en desempeño se relaciona con ciclos estructurados de práctica, retroalimentación accionable y revisión, y que variables como presencia, autoeficacia y autorregulación pueden mediar el rendimiento, lo que refuerza la necesidad de diseños que articulen componentes cognitivos y afectivos sin incrementar carga extrínseca (Di Palma *et al.*, 2025; Makransky & Petersen, 2021). Asimismo, revisiones focalizadas en habilidades blandas con RV han identificado avances, pero también brechas persistentes en operacionalización de competencias, transparencia metodológica y robustez de evaluación, lo que incrementa la relevancia de describir con precisión la intervención y los criterios de medición cuando se busca generar evidencia útil para escalamiento (Dubiel *et al.*, 2025).

En el mismo ciclo reciente, la discusión sobre IA en educación superior se ha intensificado por la expansión de aplicaciones orientadas a personalización, analítica del aprendizaje y retroalimentación formativa. Una síntesis relevante del campo muestra que, aunque existen múltiples aplicaciones reportadas, la participación explícita de marcos pedagógicos y el rol del profesorado en el diseño e implementación han sido históricamente inconsistentes, lo que sugiere que los resultados dependen menos de la capacidad técnica y más de decisiones instruccionales, condiciones institucionales y calidad de la integración didáctica (Zawacki-Richter *et al.*, 2019).

Desde 2023, la disponibilidad extendida de modelos de lenguaje de gran escala ha reconfigurado el entorno de uso por su capacidad de interactuar en lenguaje natural, generar retroalimentación textual y apoyar procesos de revisión. La literatura reciente

identifica oportunidades para tutoría asistida, explicación, andamiaje y *feedback*, pero también advierte riesgos de validez, sesgos, dependencia, y superficialidad del aprendizaje si no se delimitan propósitos, se comunican límites operativos y se mantiene supervisión humana como condición de uso responsable (Kasneci *et al.*, 2023; UNESCO, 2023). En consecuencia, la IA se vuelve pedagógicamente defendible cuando fortalece ciclos de mejora con criterios explícitos, preserva la autoría del estudiante, promueve elaboración cognitiva y se inserta en tareas donde el juicio académico y la evaluación criterial continúan siendo centrales.

El entorno contemporáneo se caracteriza también por una presión institucional creciente para sostener decisiones educativas con evidencia, trazabilidad y evaluación consistente de competencias. Esta tendencia se expresa en un desplazamiento desde métricas de percepción hacia mediciones de desempeño mediante instrumentos y rúbricas que permitan comparabilidad y control de calidad. Bajo este enfoque, IA y RV adquieren relevancia potencial cuando amplían oportunidades de práctica observable, retroalimentación oportuna y registro de procesos, sin degradar consistencia evaluativa. No obstante, revisiones contemporáneas han enfatizado que los beneficios dependen de condiciones habilitantes frecuentemente subestimadas: capacitación docente, soporte técnico, tiempo para diseño instruccional, estabilidad de implementación y sostenibilidad operativa, además de criterios claros sobre el rol de la tecnología en la secuencia didáctica y en la integridad académica (Radianti *et al.*, 2020; UNESCO, 2023; Varsik & Vosberg, 2024). Por ello, esta tesis no asume efectos automáticos; conceptualiza la integración IA+RV como una arquitectura didáctica cuya plausibilidad depende de la calidad del diseño, la gestión de carga y la estructura de práctica y retroalimentación.

En América Latina, el contexto de adopción tecnológica en educación superior se caracteriza por asimetrías institucionales: coexisten universidades con infraestructura avanzada y capacidad de soporte con entornos que enfrentan brechas de conectividad, equipamiento y desarrollo profesional docente. Los marcos internacionales recientes han enfatizado que reducir brechas no se limita a adquisición de tecnologías, sino que requiere alfabetización digital y en IA, gobernanza de datos, criterios de transparencia, y

capacidad institucional para sostener procesos responsables y basados en evidencia (UNESCO, 2021; UNESCO, 2023; Varsik & Vosberg, 2024). En este escenario, el riesgo de adopciones instrumentales aumenta: incorporar tecnologías por disponibilidad o presión externa sin rediseño de tareas, criterios y mecanismos de evaluación suele traducirse en resultados modestos y heterogéneos. En consecuencia, el contexto regional refuerza la necesidad de explicitar condiciones de implementación y estándares de medición que permitan distinguir entre uso de tecnología e impacto formativo, con especial atención a equidad de acceso y consistencia evaluativa.

En el plano institucional, el estudio se implementa en el Tecnológico de Monterrey, una universidad privada mexicana de carácter multicampus, donde existen condiciones organizacionales para diseñar y evaluar intervenciones tecnopedagógicas con estandarización de criterios de desempeño. Para los fines del estudio, esta referencia institucional opera como descripción del contexto de implementación: disponibilidad de infraestructura para experiencias inmersivas, soporte tecnológico para el despliegue de herramientas, y prácticas de evaluación que pueden apoyarse en rúbricas analíticas y procedimientos de calibración. Estas condiciones habilitan un diseño evaluable con control relativo de la implementación y medición criterial de competencias, aunque también introducen exigencias metodológicas: asegurar comparabilidad entre grupos y sedes, documentar variabilidad docente, resguardar integridad académica en escenarios con acceso a asistentes generativos, y sostener procedimientos de calidad para entrenamiento y consistencia de evaluadores.

En suma, este marco contextual establece que IA y RV se han vuelto tecnologías operativamente disponibles para intervenir en la formación por competencias en educación superior, pero su efectividad depende de condiciones específicas: coherencia pedagógica, diseño instruccional explícito, control de carga cognitiva, ciclos deliberados de práctica y retroalimentación, y gobernanza responsable. Este entorno justifica la relevancia de estudiar la integración IA+RV bajo un diseño evaluable, ya que permite aportar evidencia sobre el comportamiento de una arquitectura didáctica en condiciones reales de implementación y sobre los componentes que explican variación en resultados, además de identificar cautelas necesarias para sostenibilidad, equidad e integridad

académica (Radianti *et al.*, 2020; Dubiel *et al.*, 2025; Kasneci *et al.*, 2023; Varsik & Vosberg, 2024). Con este encuadre, el apartado siguiente desarrolla el marco legal y normativo, donde se precisan criterios de protección, responsabilidad y cumplimiento aplicables al estudio (UNESCO, 2021; UNESCO, 2023).

2.6 Marco normativo y legal

El diseño y la implementación de este estudio se sustentan en un marco normativo y legal que prioriza la protección de derechos de las personas participantes, la integridad académica, la gobernanza responsable de datos y la mitigación de riesgos asociados al uso de inteligencia artificial (IA) y realidad virtual (RV) en contextos formativos. Esta sección no se limita a describir principios, sino que explicita su traducción operativa en controles verificables a lo largo del ciclo de vida del dato y de la intervención, en consonancia con recomendaciones internacionales sobre ética de la IA y con el marco legal aplicable en México.

En el plano normativo internacional, se adoptaron como referentes los principios de la *Recommendation on the Ethics of Artificial Intelligence* de la UNESCO, que establece exigencias de supervisión humana, transparencia, protección de privacidad, equidad y rendición de cuentas para sistemas de IA en contextos sociales, incluyendo educación (UNESCO, 2021). De manera complementaria, se consideró la guía de la UNESCO sobre IA generativa para educación e investigación, que traslada dichos principios a prácticas institucionales concretas, tales como alfabetización en IA, trazabilidad de uso, delimitación de finalidades, gestión de riesgos y salvaguardas para usos de alto impacto (UNESCO, 2023). Asimismo, se incorporaron orientaciones de política educativa de la OCDE que enfatizan la necesidad de ecosistemas habilitantes para adopción sostenible, particularmente en materia de gobernanza de datos, formación docente y uso basado en evidencia (Varsik y Vosberg, 2024). Finalmente, para fortalecer un enfoque sistemático de gestión de riesgo tecnológico, se tomaron como referencia lineamientos técnicos de gestión de riesgos en sistemas de IA, en la medida en que contribuyen a formalizar identificación de riesgos, controles y trazabilidad de mitigaciones (NIST, 2023). En conjunto, estos referentes normativos proporcionan el estándar de diligencia debida para evitar un uso meramente instrumental de la IA y la

RV, y asegurar que su integración se mantenga alineada con criterios éticos, pedagógicos y de auditabilidad.

En el plano legal, dado que el estudio se implementó en una institución privada mexicana y con personas participantes mayores de edad, el tratamiento de datos personales se estructuró conforme a la Ley Federal de Protección de Datos Personales en Posesión de los Particulares (LFPDPPP; Cámara de Diputados del H. Congreso de la Unión, 2025) , observando los principios de licitud, consentimiento, información, calidad, finalidad, lealtad, proporcionalidad y responsabilidad. En congruencia con este marco, se contemplaron mecanismos para el ejercicio de derechos ARCO, así como criterios de minimización, seudonimización, control de accesos y retención limitada, con eliminación o anonimización al cierre del análisis, según corresponda. Adicionalmente, por tratarse de investigación con participación humana, el protocolo adoptó salvaguardas reforzadas de protección a participantes mediante consentimiento informado, identificación y mitigación de riesgos previsibles, y mecanismos de atención a incidentes. En los aspectos donde fue pertinente, se tomaron como referencia criterios de buenas prácticas del Reglamento de la Ley General de Salud en Materia de Investigación para la Salud (Cámara de Diputados del H. Congreso de la Unión, 1987) , únicamente como estándar de protección y rigor procedimental en consentimiento y gestión de riesgo, sin reclasificar el estudio como investigación clínica.

En materia de propiedad intelectual, el uso de materiales de terceros, capturas, recursos audiovisuales o artefactos empleados en actividades se condujo bajo criterios de atribución y uso legítimo conforme a la Ley Federal del Derecho de Autor (LFDA; Cámara de Diputados del H. Congreso de la Unión, 1996), privilegiando insumos con licencias compatibles, citas adecuadas y elaboración de material propio cuando fue necesario. Dado que la intervención incluye tecnologías emergentes y potencial uso de plataformas externas, se consideraron también implicaciones de confidencialidad y seguridad asociadas a servicios de terceros, con atención a condiciones de tratamiento, almacenamiento y cierre del procesamiento una vez concluido el proyecto.

Estos referentes se tradujeron en un esquema de cumplimiento verificable organizado por fases de implementación y por riesgos específicos de IA y RV. En

primer lugar, el consentimiento informado y la transparencia se operacionalizaron mediante documentación explícita de propósitos, procedimientos, riesgos y beneficios previsibles asociados a IA y RV, tratamiento de datos, posibilidad de desistimiento sin consecuencias académicas, y canales para ejercer derechos ARCO. En segundo lugar, se aplicó minimización y seudonimización de datos en las bases analíticas para asegurar que el análisis se realizara sobre identificadores no directos, manteniendo por separado y bajo protección reforzada cualquier llave de identificación, con controles de acceso por roles. En tercer lugar, se establecieron prácticas de gobernanza y auditoría para asegurar trazabilidad de accesos, modificaciones y versiones de bases analíticas, así como reglas de retención y disposición final, con medidas de seguridad acordes al tipo de dato y al nivel de sensibilidad.

En cuarto lugar, se adoptó un enfoque específico de gestión de riesgos de IA, delimitando el rol de la IA al ámbito formativo y estableciendo salvaguardas para evitar su uso como sustituto de evaluación o de juicio docente. Este enfoque incluyó comunicación clara sobre lo que la IA puede y no puede hacer, supervisión humana en decisiones de alto impacto, atención a riesgos de sesgo, error y sobreconfianza, y registro de configuraciones y pautas de uso cuando resultó aplicable, con el propósito de sostener trazabilidad. En quinto lugar, se contemplaron medidas de seguridad y bienestar para experiencias de RV, incluyendo inducción previa, pausas programadas, monitoreo de malestar, registro de incidentes asociados a ciber-mareo u otras molestias, y alternativas no inmersivas cuando fuera necesario para asegurar participación segura y equitativa. En sexto lugar, se previó supervisión ética institucional y un mecanismo de reporte de incidentes, con acciones correctivas documentadas, de modo que el cumplimiento no fuera solo declarativo, sino operativo y trazable.

En séptimo lugar, se reforzó la integridad académica mediante reglas explícitas de originalidad, citación y declaración de apoyos tecnológicos, cuidando que la incorporación de IA en actividades formativas no derivara en sustitución de autoría ni en degradación de habilidades de elaboración y argumentación. En octavo lugar, cuando se emplearon plataformas o servicios externos, se verificaron condiciones de confidencialidad y seguridad, y se consideraron prácticas de cierre del procesamiento,

borrado o limitación de retención al término del proyecto, conforme a criterios de minimización y responsabilidad.

Con el fin de facilitar verificación interna y externa, la correspondencia entre exigencias normativas, controles aplicados y evidencia documental se consolidó en una matriz de cumplimiento (Tabla 2.2). Esta matriz funciona como un instrumento de auditabilidad: permite rastrear de manera directa qué principio o norma fundamenta cada control, y dónde se encuentra el soporte documental dentro de los apéndices y secciones del protocolo.

Tabla 2.2

Matriz de cumplimiento normativo y controles implementados

Norma/Principio	Exigencia clave	Control aplicado en la tesis	Evidencia/Anexo
UNESCO (2021) Ética en IA	Supervisión humana; equidad; transparencia	Delimitación del rol de la IA a retroalimentación formativa; supervisión humana en decisiones relevantes; comunicación de finalidad y límites; alternativas equivalentes cuando procede	Apéndice A6 (avisos y lineamientos de uso de IA y RV); Apéndice A7 (riesgos, <i>quality gates</i> y operación segura); sección de ética del protocolo
UNESCO (2023) Guía sobre IA generativa	Gestión de riesgos; alfabetización en IA; trazabilidad	Evaluación de riesgos y mitigaciones; registro de configuraciones/pautas de uso; inducción operativa a estudiantes sobre uso permitido/prohibido y expectativas de integridad académica	Apéndice A7 (matriz de riesgos y controles); Apéndice A6 (lineamientos de uso e inducción operativa)
OCDE (2024)	Gobernanza de datos; uso basado en evidencia; formación	Roles y reglas de uso/resguardo; trazabilidad de decisiones y control de calidad; inducción operativa y lineamientos de uso para asegurar aplicación consistente	Apéndice A5 (gobernanza, roles y resguardo); Apéndice A4 (bitácora analítica, QA y trazabilidad); Apéndice A6 (lineamientos de uso e inducción)
LFPDPPP (2010)	Principios y derechos ARCO	Consentimiento informado; mecanismos ARCO; minimización; seudonimización; resguardo seguro; periodos de retención y eliminación segura (cuando aplique)	Anexo 1 (formato institucional de consentimiento y aviso de privacidad); Apéndice A5 (resguardo y gobernanza); sección de ética del protocolo
Reglamento de investigación en salud (referencial)	Consentimiento; gestión de riesgo; dictamen	Salvaguardas de protección a participantes; documentación de consentimiento; resguardo documental conforme a revisión institucional	Anexo 1 (consentimiento institucional, si aplica); sección de ética del protocolo; evidencia de revisión institucional correspondiente
Ley Federal del Derecho de Autor	Uso legítimo y atribución	Preferencia por recursos propios o bajo licencias compatibles; atribución y citación; inventario de materiales cuando aplica	Anexo 2 (licencias, atribuciones y materiales de terceros)
Seguridad en RV	Salud y ergonomía	Inducción; pausas; criterios de suspensión; registro de incidentes (ciber-mareo) y alternativas equivalentes	Apéndice A7 (inducción, <i>checklist</i> , protocolo ante ciber-mareo, reporte de incidente y operación segura)

Nota. La matriz vincula cada exigencia normativa con el control aplicado y la evidencia documental del estudio.

Con los referentes teóricos, conceptuales, históricos y normativos establecidos en este capítulo, el manuscrito queda en condiciones de traducir los constructos centrales

del estudio a definiciones operativas verificables. En el capítulo siguiente se presenta la operacionalización de las variables de exposición y de resultado, incluyendo la lógica de medición, los instrumentos y rúbricas, los procedimientos de entrenamiento y calibración de evaluadores, los criterios de calidad psicométrica, la construcción de los indicadores de dosificación IA y RV, así como las reglas de integración de evidencias que sostienen la trazabilidad entre teoría, diseño y análisis.

CAPÍTULO 3. MARCO METODOLÓGICO, OPERACIONALIZACIÓN Y DISEÑO DE MEDICIÓN

El presente capítulo presenta el marco metodológico que sostiene la evaluación de la arquitectura didáctica IA+RV y define, con criterios explícitos, la lógica de medición que permitirá contrastar las hipótesis del estudio bajo condiciones reales de curso. En primer término, se precisará la naturaleza metodológica de la investigación, considerando el paradigma, el enfoque, el diseño y la tipología, de modo que el marco metodológico resulte coherente con el problema, la pregunta y el objeto de estudio establecidos en el Capítulo 1. En segundo término, se establecerá la operacionalización de variables, articulando de manera verificable la variable independiente, las variables intervinientes vinculadas a la implementación, y las variables dependientes asociadas al desempeño en comunicación y solución de problemas, junto con sus dimensiones e indicadores. Posteriormente, se describirán los métodos, técnicas e instrumentos de recolección de datos, así como las decisiones de aseguramiento de calidad, incluyendo procedimientos de doble calificación con rúbricas analíticas, estimación de acuerdo interjueces y criterios psicométricos mínimos para integrar las métricas al análisis. Finalmente, se explicitará el puente entre medición y derivaciones analíticas, dejando trazable la forma en que el diseño cuasi-experimental por conglomerados, la medición pre y post, y los registros de dosificación IA y RV se integrarán para interpretar resultados con validez y auditabilidad.

3.1 Cuadro de operacionalización de variables

Para asegurar trazabilidad entre la pregunta de investigación, los objetivos, las hipótesis y la estrategia de medición, se presenta el cuadro de operacionalización de variables. Su finalidad es traducir los constructos del estudio a dimensiones e indicadores observables, precisando qué evidencia se recabará, bajo qué lógica se interpretarán los resultados y cómo se distinguirá entre disponibilidad nominal, exposición efectiva, dosificación y fidelidad de implementación. Esta organización fortalece la coherencia interna del marco metodológico y permite documentar variaciones de entrega relevantes para interpretar los efectos observados, conforme a

enfoques de validez basada en evidencia y evaluación de intervenciones complejas (Messick, 1995; Kane, 2006; AERA, APA, & NCME, 2014; Bellg et al., 2004; Carroll et al., 2007; Campbell et al., 2012). La Tabla 3.1 sintetiza la alineación entre pregunta, objetivos, hipótesis, variables, dimensiones e indicadores del estudio.

Tabla 3.1

Operacionalización de variables

Pregunta de investigación	Objetivo general	Objetivos específicos	Hipótesis	Variables estudiadas	Dimensiones	Indicadores
¿Cómo se puede contribuir a la mejora del desempeño en competencias transversales de comunicación y solución de problemas mediante ciclos de práctica situada, retroalimentación criterial y reflexión en estudiantes de pregrado del Tecnológico de Monterrey, México, durante el periodo 2024 a 2026?	Analizar la influencia de una intervención instruccional basada en IA y RV en el desempeño de comunicación y solución de problemas en estudiantes de pregrado del Tecnológico de Monterrey (2024–2025), mediante mediciones pre y post, evaluación de desempeño e integración de evidencia cualitativa.	O1. Estimar el cambio pre a post en Comunicación y Solución de problemas entre intervención y comparación, controlando línea base y efectos de clúster. O2. Cuantificar la dosificación y fidelidad de implementación de IA y RV, y evaluar su asociación con los resultados. O3. Documentar la consistencia interjueces y la evidencia psicométrica mínima de las medidas utilizadas. O4. Explorar la experiencia de estudiantes y docentes, así como mecanismos plausibles mediante entrevistas y observación. O5. Identificar implicaciones instruccionales para el escalamiento responsable de IA y RV en cursos regulares. O6. Elaborar una propuesta de implementación escalable de la arquitectura didáctica IA+RV, integrando evidencia cuantitativa y cualitativa, criterios de fidelidad y salvaguardas de equidad y gobernanza. O7. Validar la propuesta de implementación escalable de la arquitectura didáctica IA+RV mediante juicio de expertos, considerando criterios de pertinencia, validez, factibilidad, aplicabilidad y generalización.	HG. La intervención basada en IA y RV se asociará con desempeño superior en comunicación y solución de problemas respecto de cursos sin integración, bajo un diseño pre y post con control de línea base y clúster. H1. El cambio pre a post en Comunicación será mayor en intervención que en comparación. H2. El cambio pre a post en Solución de problemas será mayor en intervención que en comparación. H3. Mayor dosificación de IA y/o RV se asociará con mejoras superiores. H4. La experiencia tecnológica previa y el programa académico moderarán los efectos. Los objetivos O6 y O7 tienen carácter propositivo y de validación experta, por lo que se derivan de la evidencia empírica y no se formulan como hipótesis estadísticas adicionales.	Variable independiente (VI): condición de estudio (intervención basada en IA y RV vs comparación sin integración). Variables intervinientes: dosificación IA; dosificación RV; fidelidad de implementación; experiencia/proceso del estudiante. Variables dependientes (VD): desempeño en comunicación; desempeño en solución de problemas.	VI condición: asignación por clase a intervención o comparación. Intervinientes: dosificación IA; dosificación RV; fidelidad; experiencia del estudiante. VD comunicación: desempeño por criterios. VD solución de problemas: desempeño por criterios.	Desempeño (rúbricas analíticas): Comunicación: claridad y coherencia; estructura argumental; uso de evidencia; adecuación a audiencia fluidez y manejo de muletillas; lenguaje no verbal. Solución de problemas: definición del problema; generación de alternativas; evaluación con criterios; anticipación de consecuencias; justificación de la decisión. Implementación (registros/bitácoras): Dosificación IA: frecuencia de uso; minutos de interacción; eventos de retroalimentación. Dosificación RV: minutos de uso; escenarios completados; incidentes reportados. Fidelidad: adherencia a secuencia instruccional; completitud por sesión. Proceso/experiencia (autoinforme y cualitativo): calidad percibida del <i>feedback</i>; presencia percibida; autoeficacia situacional; hallazgos de entrevistas y observación.

Nota. La tabla organiza la alineación lógica del estudio entre pregunta, objetivos e hipótesis y su operacionalización en variables, dimensiones e indicadores. Los indicadores de desempeño se obtienen mediante rúbricas analíticas (Apéndice A2) con control de consistencia interjueces (Apéndice A3). Los indicadores de implementación, integridad y trazabilidad se documentan mediante bitácoras y controles de calidad de datos (Apéndice A4), mientras que las condiciones de gobernanza y resguardo se establecen en los apartados correspondientes (Apéndice A5). Los lineamientos de privacidad y uso responsable de IA y RV se presentan en el Apéndice A6, y la evidencia relativa a riesgos, *quality gates* y protocolos de seguridad operativa en entornos inmersivos se concentra en el Apéndice A7. El consentimiento informado institucional y los materiales externos o de terceros se presentan por separado en los anexos, manteniendo en el capítulo la narrativa metodológica evaluable.

3.2 Diseño metodológico y estrategia de medición

Este apartado define, de manera precisa y coherente con el problema, los objetivos y las hipótesis del estudio, el marco metodológico que orienta la investigación. En particular, establece la perspectiva epistemológica asumida, el enfoque, el diseño y la tipología de investigación, y delimita la lógica de medición que permitirá estimar cambios en el desempeño de comunicación y solución de problemas asociados a la implementación de una arquitectura didáctica IA+RV en condiciones reales de curso. La prioridad del diseño es sostener inferencias plausibles y auditables bajo restricciones propias del contexto educativo aplicado, preservando validez ecológica, control de calidad de medición y trazabilidad entre hipótesis, variables e indicadores (Shadish *et al.*, 2002; AERA/APA/NCME, 2014).

3.2.1 Perspectiva epistemológica y lógica de inferencia

La investigación se ubica en una perspectiva pragmática, orientada a responder una pregunta aplicada sobre efectividad educativa bajo condiciones operativas reales. Esta postura prioriza la coherencia entre el tipo de evidencia recabada y el tipo de inferencia que se pretende sostener, así como la utilidad del conocimiento para la toma de decisiones instruccionales e institucionales. En este marco, el estudio combina evidencia cuantitativa y cualitativa de manera complementaria, sin asumir que una sola fuente agota la explicación del fenómeno. La inferencia principal es de tipo explicativo-comparativo en desempeño, y se acompaña por una inferencia interpretativa de proceso orientada a comprender condiciones de implementación y mecanismos plausibles de ocurrencia del efecto (Creswell & Plano Clark, 2018; Tashakkori & Teddlie, 2010).

3.2.2 Enfoque, diseño y tipología de investigación

El estudio adopta un enfoque de métodos mixtos con predominio cuantitativo integrado por un componente cuantitativo orientado a estimar el efecto diferencial de la arquitectura IA+RV y un componente cualitativo dirigido a contextualizar la experiencia de implementación y los mecanismos plausibles de aprendizaje. Esta distinción permite precisar que el diseño cuasi-experimental corresponde al componente cuantitativo, mientras que el enfoque mixto define la estrategia general de integración de evidencias.

El componente cuantitativo se estructura como un diseño cuasi-experimental por conglomerados con medición pre y post, donde la unidad de asignación operativa es la clase y la unidad principal de análisis es el estudiante, reconociendo la estructura jerárquica de los datos. La comparación se realiza entre una condición con implementación de la arquitectura IA+RV y una condición equivalente sin dicha integración, preservando el funcionamiento natural del curso y reduciendo el riesgo de contaminación entre participantes cuando comparten espacios, materiales o prácticas instruccionales (Campbell & Stanley, 1963; Shadish et al., 2002).

En términos de tipología, la investigación es aplicada y evaluativa, dado que su objetivo no es solo describir el fenómeno, sino estimar la contribución asociada a una intervención instruccional implementada en un contexto auténtico. La elección de un diseño por conglomerados responde a restricciones prácticas y éticas del entorno educativo, donde la asignación individual suele ser inviable y el control de contaminación requiere operar a nivel grupo. En este tipo de diseños, la calidad de inferencia depende de la medición de línea base, la documentación de equivalencia inicial, el registro de implementación y el uso de modelos analíticos consistentes con la dependencia intraclase (Murray, 1998; Donner & Klar, 2000; Hayes & Moulton, 2017).

El componente cualitativo se incorpora con una función explicativa y contextual. Su propósito es interpretar los hallazgos cuantitativos, identificar condiciones de implementación y aportar evidencia de proceso vinculada a mecanismos plausibles como autorregulación, presencia percibida y gestión de carga cognitiva. En coherencia con un enfoque mixto, la integración de evidencia se plantea a nivel de interpretación y discusión, de modo que la convergencia o divergencia entre fuentes se utilice para fortalecer explicaciones y precisar límites de generalización (Creswell & Plano Clark, 2018).

3.2.3 Estrategia de medición y temporalidad del diseño

La estrategia de medición se fundamenta en un esquema pre y post que permite estimar el cambio en desempeño dentro del periodo académico 2024–2025. La medición pre establece línea base y permite controlar diferencias iniciales en comunicación y

solución de problemas. La medición post captura el desempeño tras la exposición acumulada a las actividades del curso, distinguiendo entre disponibilidad tecnológica y exposición efectiva. Esta estructura es consistente con hipótesis de ganancia diferencial entre condiciones y con análisis que incorporan ajuste por línea base, práctica recomendada para mejorar precisión y reducir sesgo por diferencias iniciales en diseños cuasi-experimentales (Shadish *et al.*, 2002; Van Breukelen, 2006).

Dado que la implementación ocurre a nivel clase, el diseño asume que estudiantes dentro de una misma clase comparten experiencias instruccionales y condiciones docentes que tienden a correlacionar resultados. Por ello, la medición se acompaña por un registro sistemático de implementación. En particular, se prioriza evidencia verificable de dosificación y fidelidad para sostener interpretaciones válidas cuando existan variaciones de efecto, incluyendo escenarios donde la intervención se ejecute con heterogeneidad operativa (Bellg *et al.*, 2004; Carroll *et al.*, 2007).

3.2.4 Determinación del tamaño de muestra y efecto de diseño

La determinación del tamaño de muestra se aborda considerando tres elementos que condicionan la potencia en diseños por conglomerados: el número de clases, el tamaño promedio de clase y el grado de dependencia intraclase, expresado mediante el coeficiente de correlación intraclase. En estudios educativos donde la asignación operativa es el grupo, la potencia se reduce respecto a un diseño con asignación individual equivalente, debido a la correlación entre participantes del mismo conglomerado. Por ello, la justificación del tamaño muestral requiere incorporar el ajuste por efecto de diseño, comúnmente expresado como:

$$DEFF = 1 + (m - 1) ICC$$

donde m es el tamaño promedio del conglomerado e ICC representa el grado de similitud intragrupo. Este ajuste permite estimar el tamaño efectivo para inferencia y clarifica por qué, en contextos educativos, el número de conglomerados suele ser un determinante crítico de potencia, incluso cuando el total de estudiantes parece elevado (Donner & Klar, 2000; Murray, 1998; Hayes & Moulton, 2017).

En coherencia con el diseño propuesto, la planeación del estudio se apoya en un tamaño de seis clases y 180 estudiantes, con ajuste analítico por dependencia intraclase. En este marco, la suficiencia del tamaño se interpreta en términos del efecto mínimo detectable en el desempeño, la estabilidad de la medición basada en rúbricas con doble calificación, y la capacidad de los modelos multinivel para incorporar covariables de línea base y estimar efectos promedio con errores estándar robustos. La estimación formal de potencia puede complementarse mediante herramientas estándar de análisis de potencia y supuestos explícitos de tamaño del efecto, nivel alfa y potencia objetivo, pero siempre interpretando resultados a la luz del número de conglomerados y del ICC esperable en el contexto educativo (Cohen, 1988; Faul *et al.*, 2007; Snijders & Bosker, 2012).

En términos operativos, este encuadre de potencia y efecto de diseño se definió sobre un marco muestral específico, con criterios explícitos de elegibilidad y unidades de asignación y análisis, para asegurar coherencia entre planeación estadística y viabilidad del trabajo de campo.

La población accesible se definió como estudiantes inscritos en grupos elegibles de cursos regulares de pregrado en el ciclo 2024–2025, bajo condiciones académicas y operativas comparables que permitieran implementar la arquitectura didáctica basada en IA y RV sin alterar la estructura formativa del curso. El marco muestral se construyó a partir de listados oficiales de grupos elegibles, filtrando por equivalencia de objetivos formativos, productos evaluables y esquema de evaluación. La unidad de asignación fue la clase como conglomerado, con el propósito de reducir contaminación entre condiciones; la unidad de análisis primaria fue el estudiante, con mediciones repetidas pre y post y ajuste explícito por dependencia intraclase. Se consideraron criterios de inclusión consistentes con matrícula activa, participación en el curso durante el periodo de medición y completitud mínima de evidencias pre y post; se establecieron criterios de exclusión relacionados con no participación por decisión informada, ausencia prolongada o falta de evidencia de desempeño evaluable conforme a la rúbrica. Este marco de población, selección y unidades analíticas sustenta la estimación del tamaño

muestral bajo potencia para diseños por conglomerados y su correspondiente ajuste por efecto de diseño, manteniendo trazabilidad entre elegibilidad, implementación y análisis.

3.2.5 Coherencia entre diseño, hipótesis y métrica de desempeño

La coherencia diseño–hipótesis–medición se asegura mediante tres decisiones: primero, la medición de desempeño se basa en productos auténticos evaluados con rúbricas analíticas alineadas a criterios explícitos, lo que permite tratar comunicación y solución de problemas como constructos observables en niveles de calidad progresivos. Segundo, la puntuación se realiza con doble calificación y estimación de consistencia interjueces, fortaleciendo fiabilidad y auditabilidad de la medición. Tercero, la estrategia analítica se define para ser consistente con mediciones repetidas y estructura jerárquica, evitando suposiciones de independencia que no se cumplen cuando el conglomerado es la clase (AERA/APA/NCME, 2014; Kane, 2006; Raudenbush & Bryk, 2002; Koo & Li, 2016).

En conjunto, el diseño metodológico y la estrategia de medición se plantean para sostener inferencias razonables sobre diferencias de ganancia pre y post entre condiciones, y para documentar condiciones de implementación que permitan distinguir entre ausencia de efecto pedagógico y fallas o heterogeneidad de ejecución. Este encuadre deja establecidos los elementos necesarios para desarrollar, en los apartados siguientes, la definición operativa de variables, los métodos y técnicas de recolección, la descripción de instrumentos, el aseguramiento de calidad y las derivaciones analíticas alineadas con la rúbrica institucional.

Con el enfoque, el diseño y la estrategia de medición ya definidos, el siguiente paso metodológico es precisar el sistema de variables que operacionaliza las hipótesis y delimita con claridad qué componente se manipula o implementa, qué factores median o condicionan su efecto durante la intervención y qué resultados se observan como evidencia de cambio. Esta delimitación es indispensable para mantener trazabilidad entre el planteamiento inferencial del estudio, las fuentes de datos y las decisiones analíticas posteriores, particularmente en un diseño por conglomerados con mediciones pre y post. En consecuencia, el apartado siguiente establece de manera explícita la variable

independiente, las variables intervinientes asociadas a la implementación y el proceso, y las variables dependientes de desempeño, dejando claro su rol lógico dentro del modelo de investigación y su relación directa con H1 a H4.

3.3. Variables del estudio: independiente, interviniente y dependiente

La validez interpretativa de una tesis aplicada en educación depende, en parte, de que el sistema de variables refleje con precisión la lógica causal propuesta en las hipótesis y, al mismo tiempo, sea medible con fuentes verificables y criterios de calidad explícitos. En este estudio, el sistema de variables se definió para sostener inferencias consistentes con un diseño cuasi experimental por conglomerados y mediciones pre y post, distinguiendo con claridad entre el componente implementado por el equipo docente e investigador, los factores de proceso que intervienen en la ocurrencia del efecto y los resultados de desempeño que expresan el cambio esperado en competencias transversales. Esta estructura busca mantener trazabilidad entre hipótesis, medición y análisis, conforme a enfoques contemporáneos de argumentación de validez y de interpretación de resultados en investigación educativa (Messick, 1995; Kane, 2006; AERA, APA, & NCME, 2014).

3.3.1 Variable independiente

La variable independiente del estudio es la condición de implementación instruccional, definida como la presencia o ausencia de una arquitectura didáctica integrada de IA y RV dentro del curso.

En términos operativos, la condición se codifica a nivel de clase, ya que el diseño por conglomerados utiliza al grupo como unidad de asignación para reducir contaminación entre estudiantes y preservar la operación natural del curso. Se establecen dos niveles:

La condición experimental corresponde a un curso implementado con la arquitectura instruccional basada en IA y RV, entendida como una secuencia que integra práctica situada en RV y retroalimentación formativa criterial mediada por IA, organizada mediante una estructura de *briefing*, ejecución de tareas y *debriefing*,

mientras que la condición de comparación se define como un curso equivalente sin dicha arquitectura integrada, manteniendo los mismos contenidos, productos y esquema de evaluación conforme a la operación regular establecida por el programa.

Esta variable independiente se vincula de manera directa con las hipótesis H1 y H2, ya que representa el componente que se implementa para explicar diferencias en ganancia pre post en las competencias de comunicación y solución de problemas. Dado que se trata de un estudio aplicado en condiciones reales de aula, la variable independiente no se interpreta como simple disponibilidad tecnológica, sino como una condición instruccional planificada que debe documentarse mediante evidencias de entrega y exposición efectiva (Campbell *et al.*, 2012; Carroll *et al.*, 2007; Bellg *et al.*, 2004).

3.3.2 Variables intervinientes

En esta tesis, las variables intervinientes se conceptualizan como factores de proceso e implementación que se sitúan entre la condición de estudio y el desempeño observado. Su función no es sustituir el contraste principal entre condiciones, sino explicar por qué una intervención puede producir un efecto verificable en algunos escenarios y un efecto modesto o heterogéneo en otros, aun cuando la condición experimental esté formalmente presente. Este enfoque es consistente con marcos de fidelidad de implementación y evaluación de intervenciones complejas, donde la interpretación del efecto requiere distinguir entre asignación, exposición real y calidad de ejecución (Bellg *et al.*, 2004; Carroll *et al.*, 2007; Campbell *et al.*, 2012).

Las variables intervinientes se agrupan en tres conjuntos coherentes con el marco teórico del Capítulo 2 y con la lógica de análisis de esta tesis.

a) Exposición y dosificación de la intervención

La exposición efectiva es la traducción medible de la intervención en experiencia real del estudiante. En un estudio mediado por tecnologías, el efecto no se atribuye a la presencia nominal de una herramienta, sino a la dosis y forma en que el estudiante interactúa con la arquitectura instruccional.

Para IA, la dosificación se define mediante indicadores auditables vinculados a frecuencia e intensidad de interacción y a eventos de retroalimentación, ya que el mecanismo plausible implica ciclos de revisión sostenidos por *feedback* accionable (Hattie & Timperley, 2007; Shute, 2008; Wiggins, 2012).

Para RV, la dosificación se define mediante minutos de exposición, escenarios completados y registros de ejecución, en tanto la práctica situada requiere repetición suficiente y condiciones comparables de experiencia.

Estas variables intervinientes se articulan con H3, que plantea una relación de tipo dosis respuesta dentro del rango observado. El análisis de dosificación busca distinguir mejora atribuible a exposición real frente a mejora atribuible a otros componentes del curso o a variaciones no controladas.

b) Fidelidad de implementación y calidad de entrega

La fidelidad de implementación se refiere al grado en que la intervención se ejecuta conforme a su secuencia y criterios previstos, sin reemplazos sustantivos que cambien su naturaleza. En este estudio, la fidelidad se entiende como un constructo operativo que incluye adherencia a la secuencia instruccional, consistencia en la aplicación de criterios y completitud de los ciclos de práctica y retroalimentación.

Dado que la evidencia en intervenciones educativas muestra que la fidelidad modula la magnitud y estabilidad del efecto, se considera una variable interviniente crítica para interpretar resultados nulos o heterogéneos, evitando atribuciones simplistas a la tecnología o al estudiantado (Bell *et al.*, 2004; Carroll *et al.*, 2007; Campbell *et al.*, 2012).

c) Variables de proceso cognitivo afectivo asociadas a la experiencia de aprendizaje

Aunque el estudio no está diseñado como prueba causal de mecanismos psicológicos, se incorporan variables de proceso que permiten contextualizar el funcionamiento de la intervención en términos coherentes con el marco teórico. En particular, se consideran medidas situacionales como presencia percibida y percepción de utilidad del *feedback*, porque la literatura sobre aprendizaje inmersivo y

retroalimentación sugiere que estos estados y percepciones pueden intervenir en la traducción de experiencia en mejora de desempeño (Makransky & Petersen, 2021; Radianti *et al.*, 2020; Hattie & Timperley, 2007; Shute, 2008).

Estas medidas se interpretan como evidencia complementaria para explicar variación del efecto, no como desenlaces principales ni como sustitutos del desempeño.

3.3.3 Variables dependientes

Las variables dependientes del estudio son dos medidas de desempeño vinculadas a competencias transversales consideradas críticas para el desempeño profesional y coherentes con el objeto de estudio definido en el Capítulo 1.

a) Desempeño en comunicación

La variable dependiente comunicación se define como desempeño observable del estudiante en tareas orales auténticas del curso, evaluado mediante rúbrica analítica con criterios estables. Se expresa en componentes como claridad y coherencia, estructura argumental, uso de evidencia, adecuación a audiencia y calidad de entrega. La medición se realiza en pre y post para estimar cambio y comparar ganancia entre condiciones, conforme a H1.

b) Desempeño en solución de problemas

La variable dependiente solución de problemas se define como desempeño observable del estudiante para definir un problema situado, generar alternativas, evaluarlas con criterios explícitos, anticipar consecuencias y justificar una decisión. Se evalúa mediante rúbrica analítica aplicada a productos auténticos del curso. La medición se realiza en pre y post para estimar cambio y comparar ganancia entre condiciones, conforme a H2.

En ambos casos, la elección de rúbricas analíticas responde a un criterio de coherencia entre constructo y medición, ya que las competencias transversales requieren evaluación criterial basada en desempeño, y no solo autoinforme. Para fortalecer confiabilidad de la medición, se aplica doble calificación y se estima consistencia interjueces mediante coeficientes de correlación intraclase, con especificaciones acordes

a mediciones por jueces y a su interpretación en investigación educativa (Shrout & Fleiss, 1979; Koo & Li, 2016). La argumentación de validez se sostiene en la trazabilidad entre constructo, criterios de rúbrica, procedimiento de puntuación e interpretación prevista del puntaje (Messick, 1995; Kane, 2006; AERA, APA, & NCME, 2014).

3.3.4 Variables moderadoras y covariables para ajuste

Además de las variables intervinientes, el estudio contempla variables cuya función es evaluar heterogeneidad del efecto o controlar diferencias iniciales relevantes, sin redefinir la naturaleza del tratamiento.

En particular, H4 plantea moderación por experiencia tecnológica previa y por programa académico. Estas variables se incorporan como moderadoras en el análisis mediante términos de interacción, siguiendo criterios metodológicos para modelos con moderación e interpretación de interacciones (McClelland & Judd, 1993; West *et al.*, 1996). El objetivo es determinar si la magnitud del efecto de la condición IA+RV difiere sistemáticamente según perfiles de experiencia o trayectorias formativas.

Asimismo, los puntajes pre de comunicación y solución de problemas se utilizan como control de línea base, dado que el interés inferencial se centra en diferencias de ganancia pre post entre condiciones.

3.3.5 Ligadura explícita entre variables e hipótesis

El sistema de variables se alinea con las hipótesis del estudio de la siguiente forma:

H1 se contrasta mediante diferencias de ganancia pre post en comunicación entre condición IA+RV y condición de comparación.

H2 se contrasta mediante diferencias de ganancia pre post en solución de problemas entre condición IA+RV y condición de comparación.

H3 se explora mediante la asociación entre indicadores de dosificación y las ganancias observadas, interpretando la dosificación como variable interviniente de

implementación dentro de la condición experimental y, cuando sea metodológicamente defendible, como predictor de variación del efecto.

H4 se examina mediante moderación, evaluando si la relación entre condición y desempeño cambia en función de experiencia tecnológica previa y programa académico.

Con este planteamiento, el capítulo conserva coherencia con el marco teórico y establece una base metodológica trazable para describir técnicas de recolección, instrumentos y aseguramiento de calidad en los apartados siguientes.

3.4 Métodos y técnicas de recolección de datos

La recolección de datos se organizó mediante un método mixto de carácter convergente y complementario, orientado a integrar evidencia cuantitativa de desempeño con evidencia cualitativa y documental sobre la implementación de la arquitectura IA+RV. En el componente cuantitativo, las técnicas principales fueron la evaluación de productos auténticos mediante rúbricas analíticas, la aplicación de instrumentos breves de experiencia y el registro sistemático de dosificación y fidelidad. En el componente cualitativo, se utilizaron entrevistas semiestructuradas, observación participante y bitácoras de campo para contextualizar la experiencia de aprendizaje, identificar condiciones de implementación y comprender mecanismos plausibles asociados al cambio observado. Esta combinación permitió mantener trazabilidad entre condición de estudio, exposición efectiva, calidad de implementación y desempeño en competencias transversales, en congruencia con enfoques de métodos mixtos aplicados a la evaluación educativa, donde la evidencia cualitativa fortalece la interpretación sin sustituir el contraste principal definido por el diseño y las hipótesis (Creswell & Plano Clark, 2018).

3.4.1 Técnica principal: evaluación de desempeño mediante productos auténticos y rúbricas analíticas

La técnica central de recolección corresponde a la evaluación de desempeño en comunicación y solución de problemas, operacionalizada mediante rúbricas analíticas aplicadas a productos auténticos del curso. La lógica de esta técnica es medir

competencias transversales como desempeño observable en tareas situadas, evitando depender exclusivamente de percepciones o autoinformes, lo cual es metodológicamente consistente con la evaluación criterial de habilidades complejas. La recolección se implementa en dos momentos, pre y post, para estimar ganancia y comparar cambios entre condiciones conforme a H1 y H2.

Para sostener consistencia y auditabilidad, la calificación se realiza mediante doble evaluación y se documentan procedimientos de calibración, así como métricas de consistencia interjueces. La elección de esta técnica se alinea con criterios contemporáneos de validez en mediciones educativas, donde la interpretabilidad del puntaje exige evidencia sobre el constructo, el proceso de puntuación y la estabilidad del procedimiento evaluativo (Messick, 1995; Kane, 2006; AERA, APA, & NCME, 2014). La consistencia entre evaluadores se estima con coeficientes de correlación intraclase bajo especificaciones apropiadas para rúbricas y calificación por jueces, con interpretación en términos de acuerdo y fiabilidad del puntaje utilizado para análisis (Shrout & Fleiss, 1979; Koo & Li, 2016).

3.4.2 Técnica de implementación: registros de dosificación y fidelidad de IA y RV

Dado que la variable independiente se interpreta como una condición instruccional y no como disponibilidad nominal de tecnología, se incorporan técnicas de recolección orientadas a documentar exposición real y fidelidad de implementación. Esta evidencia permite distinguir entre asignación a condición, entrega efectiva de la intervención y experiencia real del estudiante, lo cual es esencial para interpretar efectos modestos, heterogéneos o nulos, y para sostener el análisis de H3 sobre relaciones dosis respuesta dentro del rango observado.

La recolección de implementación utiliza registros auditables provenientes de dos fuentes principales. La primera corresponde a analíticas de uso o registros de plataforma, cuando la infraestructura lo permite, para capturar indicadores de interacción con IA y RV, tales como tiempos de uso, eventos de retroalimentación, escenarios ejecutados y completitud de actividades. La segunda fuente corresponde a bitácoras docentes y listas de cotejo de fidelidad por sesión, diseñadas con campos estructurados

para registrar adherencia a la secuencia instruccional, incidencias de operación y evidencia mínima de entrega. El diseño de esta técnica se fundamenta en literatura de fidelidad de implementación y evaluación de intervenciones complejas, que destaca la necesidad de medir adherencia, exposición y calidad de entrega para interpretar resultados y sostener escalamiento responsable (Bellg *et al.*, 2004; Carroll *et al.*, 2007; Campbell *et al.*, 2012).

3.4.3 Técnica complementaria: autoinformes breves de proceso y experiencia de aprendizaje

Con fines de contextualización y explicación adicional, se incluyen instrumentos breves de autoinforme vinculados a experiencia en la intervención, tales como utilidad percibida del *feedback*, presencia percibida en RV y autoeficacia situacional asociada a la tarea. Estas medidas se interpretan como evidencia de proceso y no como desenlaces principales. Su papel metodológico es describir condiciones subjetivas plausibles que pueden intervenir en la traducción de experiencia en mejora del desempeño, particularmente en entornos inmersivos y en ciclos de retroalimentación formativa.

Cuando estos instrumentos se aplican en más de un momento, se considera la necesidad de asegurar comparabilidad temporal en términos de estructura y consistencia, bajo criterios habituales para instrumentos psicométricos. En escalas aplicadas solo en post, la interpretación se limita a evidencia situacional de proceso, evitando inferencias de cambio longitudinal. Estas decisiones se sostienen en recomendaciones de evaluación de ajuste y comparabilidad en modelos de medición, así como en buenas prácticas de consistencia interna que consideran limitaciones del alfa y el uso complementario de omega cuando sea pertinente (Hu & Bentler, 1999; Cheung & Rensvold, 2002; Chen, 2007; Dunn, Baguley, & Brunsden, 2014).

3.4.4 Técnicas cualitativas: entrevistas semiestructuradas y observación participante

La recolección cualitativa se incorpora para fortalecer la interpretación del efecto y documentar mecanismos plausibles sin sobredimensionar inferencias causales. Se emplean entrevistas semiestructuradas y observación participante como técnicas

orientadas a describir la experiencia del estudiantado y del docente, identificar condiciones de implementación que favorecen o limitan resultados y complementar la lectura de la evidencia cuantitativa.

Las entrevistas semiestructuradas se seleccionan por su capacidad para producir información comparable entre participantes y, al mismo tiempo, permitir profundización en aspectos emergentes relevantes para la arquitectura IA+RV, tales como comprensión del *feedback*, gestión de carga mental, participación en *debriefing* y percepción de utilidad de la práctica situada. La observación participante se utiliza para registrar aspectos no capturados por los registros de plataforma o por el puntaje, incluyendo dinámicas de clase, adherencia real a la secuencia, incidencias operativas y calidad del cierre reflexivo posterior a la actividad. En conjunto, estas técnicas permiten construir una narrativa analítica de implementación y experiencia, coherente con un enfoque de métodos mixtos orientado a explicación y mejora de intervención (Creswell & Plano Clark, 2018).

3.4.5 Técnica documental y trazabilidad: bitácora analítica y control de decisiones de datos

Adicionalmente, se integra una bitácora analítica como técnica documental para asegurar trazabilidad de decisiones durante el ciclo de recolección y preparación de datos. Esta bitácora registra definiciones operativas, decisiones de limpieza, criterios de exclusión, tratamiento de datos faltantes, transformaciones de puntajes, verificación de coherencia y control de versiones de insumos analíticos. Su finalidad es fortalecer la auditabilidad metodológica y reducir ambigüedades interpretativas sobre cómo se construyeron las variables utilizadas en análisis. Esta práctica se alinea con estándares contemporáneos de transparencia en investigación aplicada y con criterios de validez que exigen evidencia clara del proceso de generación del puntaje y su uso inferencial (AERA, APA, & NCME, 2014; Kane, 2006).

3.4.6 Integración de técnicas y correspondencia con el sistema de variables

La estrategia de recolección mantiene una correspondencia explícita con el sistema de variables del estudio. La condición de intervención se registra a nivel clase.

Las variables intervinientes de dosificación y fidelidad se sostienen con registros auditables de plataforma y bitácoras estructuradas. Las variables dependientes de desempeño se recogen mediante rúbricas analíticas en pre y post con doble calificación. Las variables de proceso se recogen con autoinformes breves de experiencia. La dimensión cualitativa se recoge mediante entrevistas y observación para explicar condiciones y aportar evidencia contextual, sin sustituir el contraste principal entre condiciones.

Con esta integración, la recolección de datos queda metodológicamente preparada para el desarrollo del apartado siguiente, en el que se precisan instrumentos, remisiones a apéndices y decisiones de calidad para garantizar consistencia, comparabilidad y trazabilidad del estudio.

3.5 Instrumentos

Para asegurar que los instrumentos de recolección de datos sean auditables y cumplan con la estructura solicitada por la rúbrica institucional, cada instrumento se presenta con una declaración mínima de cinco componentes: (a) tema de medición; (b) objetivo del instrumento; (c) consigna o instrucción de aplicación; (d) instrumentación o desarrollo; y (e) notas de procedencia, cuando corresponda. La versión completa de los instrumentos se ubica en los apéndices y anexos respectivos: las rúbricas analíticas de desempeño se presentan en el Apéndice A2; los formatos de doble calificación, discrepancias e ICC en el Apéndice A3; la bitácora analítica y los controles de integridad de datos en el Apéndice A4; los instrumentos de gobernanza, resguardo y trazabilidad en el Apéndice A5; los avisos de privacidad y lineamientos de uso responsable de IA y RV en el Apéndice A6; los protocolos de riesgo, seguridad, incidentes y quality gates en el Apéndice A7; el consentimiento informado institucional en el Anexo 1; y los registros de licencias, atribuciones y materiales de terceros en el Anexo 2. Los materiales técnicos de reproducibilidad se documentan en el Apéndice B y los instrumentos de validación experta de la propuesta en el Apéndice C.

La selección y el diseño de los instrumentos se orientaron por tres criterios operativos complementarios: medir competencias transversales mediante desempeño

observable, documentar la implementación real de la arquitectura IA+RV y asegurar trazabilidad entre hipótesis, variables, procedimientos de captura y decisiones analíticas. Este enfoque permite distinguir entre desempeño, exposición efectiva, fidelidad de implementación, gobernanza y evidencia técnica de replicabilidad, en congruencia con estándares vigentes de medición educativa, que enfatizan evidencia sobre el constructo, el proceso de puntuación y el uso inferencial de los puntajes (AERA, APA, & NCME, 2014; Kane, 2006; Messick, 1995).

3.5.1 Rúbricas analíticas de desempeño para variables dependientes

Las variables dependientes del estudio, desempeño en comunicación y desempeño en solución de problemas, se midieron mediante rúbricas analíticas aplicadas a productos auténticos situados en actividades del curso. En este instrumento, el tema de medición es el desempeño observable, el objetivo es estimar cambio pre y post con criterios estandarizados, la consigna se operacionaliza mediante la entrega del producto auténtico y la instrumentación se concreta en criterios, niveles y reglas de puntuación. La selección de este tipo de rúbrica responde a la naturaleza compleja y multicomponente de estas competencias, donde la medición exige criterios explícitos y niveles de logro observables. El uso de rúbricas como instrumentos de medición de desempeño permite estandarizar la evaluación, sostener consistencia entre evaluadores y fortalecer interpretabilidad pedagógica del cambio pre y post (AERA, APA, & NCME, 2014).

Las rúbricas empleadas se incluyen íntegramente en el Apéndice A2, con sus criterios, descriptores de niveles y reglas de puntuación, para facilitar verificación y replicabilidad. En el capítulo metodológico se establece que la calificación se realiza mediante doble evaluación ciega respecto a condición y momento, y que el puntaje operativo se utiliza como promedio de evaluadores; esta decisión se fundamenta en la literatura sobre evaluación por jueces y consistencia interjueces (Shrout & Fleiss, 1979; Koo & Li, 2016). El detalle del procedimiento de calibración y el reporte de acuerdo se desarrollan en el apartado 3.6, con el fin de concentrar la evidencia de calidad de medición en un solo lugar y evitar redundancia.

3.5.2 Instrumentos para variables intervinientes de implementación: dosificación y fidelidad

Para sostener H3 y para interpretar con rigor resultados heterogéneos o modestos, se incorporaron instrumentos de implementación centrados en dos componentes. El primero es la dosificación de IA y RV, definida como exposición efectiva y auditable. El segundo es la fidelidad de implementación, definida como la adherencia a la secuencia instruccional y a las condiciones mínimas de entrega previstas. La inclusión de estos instrumentos se sustenta en marcos de fidelidad de intervención y evaluación de procesos, que recomiendan medir exposición, adherencia y calidad de entrega para interpretar efectos y sostener escalamiento responsable (Bellg *et al.*, 2004; Carroll *et al.*, 2007; Campbell *et al.*, 2012). En este estudio, estos componentes se tratan como variables de implementación y proceso que permiten interpretar variaciones en el efecto observado y evaluar la validez de la entrega, sin asumir por sí mismos una relación mediacional causal.

En estos instrumentos, el tema es la exposición y adherencia, el objetivo es documentar implementación real para interpretar resultados, y la instrumentación se operacionaliza mediante registros, bitácoras y listas de cotejo con reglas de completitud y uso. En la práctica, la dosificación se documenta mediante dos tipos de evidencia. Por un lado, registros de uso o analíticas disponibles en las plataformas, cuando la plataforma provee métricas consistentes con la unidad de análisis requerida (p. ej., tiempo de uso, número de interacciones o completitud de actividades). Por otro lado, una bitácora docente estructurada por sesión para registrar tiempos, actividades efectivamente realizadas, eventos de retroalimentación y observaciones operativas relevantes. Los formatos estandarizados de bitácora y control de implementación se integran en el Apéndice A4, junto con criterios mínimos de completitud y trazabilidad.

De forma complementaria, la fidelidad de implementación se documenta con listas de cotejo por sesión que verifican cumplimiento de pasos instruccionales clave, incluyendo preparación operativa, *briefing*, ejecución de actividad con IA o RV según condición, y cierre reflexivo. Los instrumentos de cotejo y sus reglas de uso forman

parte del conjunto operativo de campo incluido en el Apéndice A1 y el Apéndice A7, según corresponda a actividades de IA o sesiones con visores de RV.

3.5.3 Instrumentos breves de proceso y experiencia

Con el fin de contextualizar mecanismos plausibles de efecto y describir condiciones de experiencia del estudiante, el estudio emplea escalas breves de autoinforme asociadas a variables de proceso, tales como utilidad percibida del *feedback*, presencia percibida y autoeficacia situacional. En estos instrumentos, el tema de medición se concentra en percepciones situacionales relevantes para el proceso, el objetivo es aportar evidencia explicativa complementaria, la consigna solicita valoración breve posterior a actividades específicas, y la instrumentación se define por escalas tipo Likert con reglas explícitas de aplicación por momento. Estas medidas se interpretan como evidencia de proceso y apoyo explicativo, evitando convertirlas en desenlaces principales cuando su naturaleza no lo justifica. En instrumentos aplicados en más de un momento, se considera la necesidad de comparabilidad temporal y consistencia; en instrumentos aplicados solo al final, su lectura se acota a evidencia situacional del periodo post. Este enfoque es coherente con prácticas habituales de medición y verificación de ajuste y comparabilidad cuando se busca interpretar cambios o estructuras latentes (Hu & Bentler, 1999; Cheung & Rensvold, 2002; Chen, 2007; Dunn, Baguley, & Brunsden, 2014). En consecuencia, cuando la evidencia y el tamaño muestral lo permiten, se verifica consistencia interna y, en instrumentos repetidos, se examina comparabilidad a un nivel acorde con el propósito explicativo de estas medidas, evitando sobreinterpretaciones más allá de su función de proceso.

Los instrumentos y reactivos exactos, con sus instrucciones y escalas de respuesta, se presentan en los apéndices conforme a su función operativa. Los formatos de aplicación y resguardos éticos (incluida la comunicación de consentimiento y condiciones de participación) se integran en el Apéndice A1 y en el Anexo 1. Los instrumentos vinculados a la interacción con IA y RV, junto con lineamientos de uso y comunicación operativa asociados a dichas actividades, se incorporan en el Apéndice A6.

3.5.4 Instrumentos cualitativos: guías de entrevista y protocolos de observación

La evidencia cualitativa se recaba mediante una guía de entrevista semiestructurada y un protocolo de observación en aula con registro estructurado. En estos instrumentos, el tema se centra en mecanismos y condiciones de implementación, el objetivo es aportar evidencia explicativa complementaria, la consigna se concreta en preguntas y focos de observación definidos, y la instrumentación se operacionaliza mediante guías semiestructuradas y un protocolo de registro sistemático. La guía de entrevista se orienta a capturar percepciones y descripciones sobre cómo se interpretó el *feedback*, cómo se gestionó la carga mental en actividades complejas, qué elementos del ciclo *briefing*, práctica y *debriefing* resultaron más útiles y qué condiciones facilitaron o limitaron la mejora. La observación se centra en condiciones de implementación, dinámica del trabajo en aula, incidencias operativas y calidad del cierre reflexivo, con el objetivo de complementar la lectura de registros y puntajes. Esta integración responde a un enfoque de métodos mixtos que busca fortalecer la interpretación y la explicación del funcionamiento real de la intervención en contexto, para explicar variaciones observadas en el componente cuantitativo y describir condiciones de implementación, sin desplazar el contraste principal definido por el diseño y las hipótesis (Creswell & Plano Clark, 2018).

Las guías y protocolos cualitativos se incluyen en el Apéndice A1 como parte del conjunto de documentación de campo y control de implementación, en coherencia con la exigencia institucional de trazabilidad y evidencia operativa. Esta organización permite que los insumos cualitativos queden vinculados al paquete operativo de campo y a las evidencias de implementación, manteniendo trazabilidad entre recolección, resguardo y uso analítico.

3.5.5 Instrumentos de control de calidad y documentación técnica

Finalmente, el estudio incorpora instrumentos orientados a control técnico y trazabilidad del proceso de evaluación y análisis. Estos instrumentos se organizaron para controlar, de manera documentada, la calidad de la puntuación, la integridad del *dataset*, la auditabilidad de decisiones analíticas y la gobernanza del resguardo, de modo que

cada etapa deje evidencia verificable. La hoja de control para cálculo y verificación de consistencia interjueces se incluye en el Apéndice A3. La bitácora analítica y de calidad de datos se incluye en el Apéndice A4 y se utiliza para registrar decisiones de preparación y depuración de datos y controles de coherencia. Los elementos de gobernanza, roles de acceso, resguardo, retención y evidencia de cumplimiento se integran en el Apéndice A5. La matriz de riesgos y puntos de control de calidad para la operación de IA y RV se integra en el Apéndice A7, como soporte a decisiones de operación y escalamiento con base en evidencia.

En conjunto, estos instrumentos permiten documentar de forma verificable y auditable la cadena medición, implementación, control de calidad y trazabilidad, lo cual refuerza la interpretabilidad de los resultados y su alineación con estándares metodológicos de medición y validez en investigación educativa (AERA, APA, & NCME, 2014; Kane, 2006; Messick, 1995).

3.6 Aseguramiento de calidad

El sistema de aseguramiento de calidad del estudio se definió como un conjunto de decisiones previas y controles durante la ejecución orientados a proteger la validez de las inferencias, la confiabilidad de la medición y la trazabilidad del proceso completo. En un diseño por conglomerados, estos controles son especialmente relevantes para mitigar riesgos de contaminación entre condiciones, sesgos de calificación, variación en la entrega de la intervención y dependencia intraclase, preservando la interpretabilidad de las diferencias observadas. En términos metodológicos, se asumió que la calidad no se reduce a reportar coeficientes al final, sino que depende de la coherencia entre constructos, procedimientos de puntuación, control de sesgos previsibles y documentación de decisiones analíticas, en línea con estándares contemporáneos de medición educativa y argumentación de validez (AERA, APA, & NCME, 2014; Kane, 2006; Messick, 1995). Por ello, el diseño de calidad se organizó alrededor de cuatro ejes: la calibración y consistencia de la evaluación por rúbricas, la consistencia interna de instrumentos de autoinforme cuando aplica, la verificación de consistencia y comparabilidad a un nivel acorde con el propósito inferencial de cada medida cuando se

pretende interpretar cambios, y la trazabilidad de la implementación y del tratamiento de datos.

3.7 Síntesis visual del capítulo metodológico

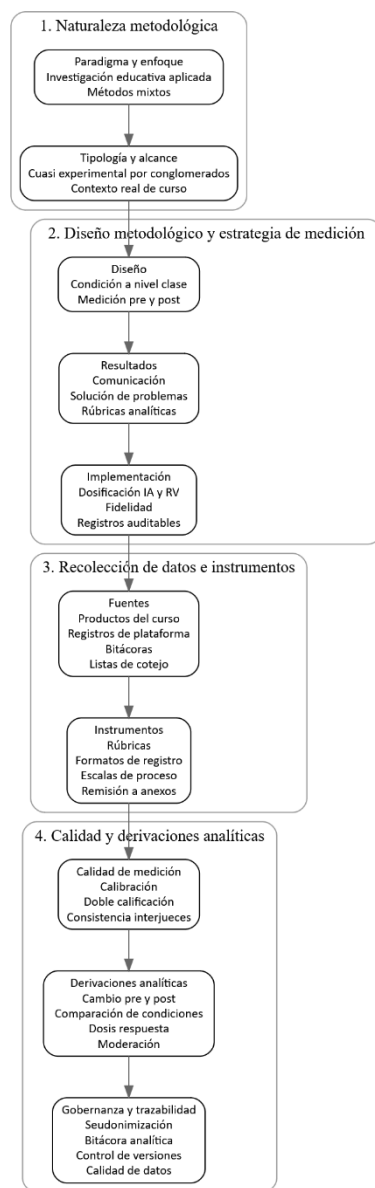
Para facilitar la lectura integral del marco metodológico, se presenta una síntesis visual que integra, en una sola vista, la lógica del diseño, la estrategia de medición, el aseguramiento de calidad y las derivaciones analíticas. La figura no agrega contenido nuevo, sino que organiza de forma compacta lo establecido en las secciones 3.1 a 3.6, permitiendo verificar la coherencia entre la naturaleza metodológica del estudio, la definición de variables, la recolección de datos y la ruta de análisis. Esta forma de síntesis es consistente con buenas prácticas de presentación en estudios aplicados con intervención educativa, donde la trazabilidad entre diseño, medición y análisis resulta necesaria para sostener inferencias y facilitar auditabilidad (Shadish et al., 2002; AERA et al., 2014; Kane, 2006).

Este capítulo estableció el marco metodológico para evaluar, en condiciones reales de curso, la contribución de una arquitectura didáctica basada en IA y RV al desempeño en Comunicación y Solución de problemas. Se precisaron la naturaleza metodológica, el diseño, la estrategia de medición, el sistema de variables, la ruta de recolección de datos y los criterios de aseguramiento de calidad. Con ello, el estudio quedó preparado para pasar a la fase operativa del trabajo de campo y, posteriormente, a la presentación de resultados bajo una lógica analítica coherente con el diseño por conglomerados, la evidencia de implementación y la prudencia interpretativa requerida en los análisis secundarios de dosificación y moderación.

Como cierre visual del capítulo, la Figura 3.1 sintetiza la secuencia diseño–medición–aseguramiento de calidad–derivaciones analíticas, así como los puntos donde los apéndices y anexos documentan instrumentos, bitácoras y evidencia técnica para auditoría y replicación.

Figura 3.1

Síntesis del marco metodológico y ruta de análisis del estudio



Nota. La figura integra, en una sola vista, la secuencia diseño–medición–aseguramiento de calidad–derivaciones analíticas, e indica los puntos donde los apéndices y anexos contienen instrumentos, bitácoras y evidencia técnica para auditoría y replicación.

CAPÍTULO 4. ANÁLISIS DE RESULTADOS

Este capítulo presenta los resultados del estudio y su interpretación pedagógica, manteniendo trazabilidad entre pregunta, hipótesis, variables, instrumentos y estimadores. En primer lugar, se documenta la integridad de la muestra y se verifican condiciones previas para la inferencia, incluyendo el flujo de participantes y la calidad de medición, siguiendo buenas prácticas de reporte para estudios experimentales y por conglomerados (Schulz *et al.*, 2010; Campbell *et al.*, 2012). En segundo lugar, se estiman los efectos principales de la intervención mediante modelos multinivel/mixtos acordes con la estructura jerárquica del diseño (mediciones repetidas anidadas en estudiantes y estudiantes anidados en clases), enfatizando la interacción Tiempo×Grupo como parámetro central del contraste tipo diferencia-en-diferencias (Fitzmaurice, Laird, & Ware, 2011; Snijders & Bosker, 2012). Para complementar la interpretación, se reportan intervalos de confianza al 95% y tamaños de efecto para diseños pre–post con grupo control mediante d_{ppc2} , facilitando la comparación de magnitudes entre *outcomes* más allá del valor p (Morris, 2008). Posteriormente, se exploran análisis complementarios orientados a implementación y mecanismo (dosificación/fidelidad; seguridad; evidencia cualitativa), así como pruebas de sensibilidad y robustez. Finalmente, se sintetizan implicaciones para la mejora instruccional y el escalamiento responsable del modelo IA+RV.

4.1 Prueba Piloto

4.1.1 Ajustes previos al trabajo de campo y cronograma

En continuidad con lo desarrollado en el capítulo anterior, donde se definieron el diseño metodológico, los instrumentos y los criterios generales de aseguramiento de calidad, antes del inicio del trabajo de campo se estableció un conjunto de ajustes operativos y de control metodológico orientados a llevar esas definiciones a condiciones reales de aplicación. Este tramo cumple una función distinta a la del capítulo 3: no se centra en justificar el diseño del estudio, sino en documentar cómo dicho diseño fue estabilizado y traducido en procedimientos comparables, verificables y ejecutables en aula antes del levantamiento formal. Los ajustes tuvieron dos propósitos

complementarios. El primero fue asegurar que la implementación de la arquitectura IA y RV fuera suficientemente estable y comparable entre clases, de modo que las diferencias observadas no se expliquen por variación evitable en la entrega. El segundo fue fortalecer la calidad de medición y la gobernanza de datos, garantizando consistencia en los criterios de evaluación, trazabilidad de transformaciones y resguardo conforme a principios éticos y de protección de información. En estudios aplicados con intervención educativa, este tramo previo suele ser determinante para sostener inferencias válidas. Por ejemplo, Bellg *et al.* (2004) mostraron que la estandarización de procedimientos, la capacitación de implementadores y el monitoreo de fidelidad son condiciones necesarias para distinguir entre falla de implementación y ausencia real de efecto; de forma complementaria, Carroll *et al.* (2007) subrayaron que la fidelidad de la intervención no depende solo de aplicar una estrategia, sino de asegurar adherencia, calidad y documentación del proceso; y Dane y Schneider (1998) advirtieron que, en contextos educativos, una proporción relevante de las amenazas a la inferencia puede originarse en variaciones no estandarizadas, la variabilidad entre sesiones y otras fuentes de heterogeneidad operativa, más que en el análisis estadístico en sí. Desde esta perspectiva, los ajustes previos realizados en esta tesis buscaron reducir precisamente esas fuentes de variación y fortalecer la interpretabilidad de los resultados.

4.1.2 Calibración de evaluadores y pilotaje de instrumentos

A diferencia del apartado 3.6, que describe el aseguramiento de calidad durante la ejecución y el análisis, esta sección documenta las acciones previas necesarias para iniciar el trabajo de campo con instrumentos y criterios estabilizados. La calidad de medición se fortaleció mediante un proceso previo de calibración de evaluadores para las rúbricas analíticas, orientado a mejorar consistencia y a reducir deriva de criterio durante el periodo de evaluación. En términos prácticos, la calibración consideró el uso de productos ancla, discusión criterial y una ronda piloto de puntuación, con revisión de discrepancias y ajuste de criterios operativos. Esta práctica es consistente con estándares contemporáneos de argumentación de validez y con el uso de coeficientes de acuerdo como parte del aseguramiento de calidad del puntaje derivado de rúbricas (AERA *et al.*, 2014; Koo & Li, 2016; Shrout & Fleiss, 1979).

Además, los instrumentos de proceso y autoinforme, cuando aplican, se sometieron a revisión de claridad, carga de respuesta y adecuación contextual, con un pilotaje breve para identificar ambigüedades y tiempos reales de aplicación. En consonancia con Díaz Muñoz (2020), este pilotaje no se concibió como una prueba de hipótesis, sino como una estrategia para verificar la adecuación metodológica del estudio principal, particularmente en dimensiones de proceso (pasos de aplicación), recursos (tiempos y requerimientos operativos), gestión (manejo de datos y consistencia de registros) y aspectos científicos vinculados con la medición y la seguridad de los participantes. Desde esta lógica, la fase previa permitió detectar fallas de comprensión, ajustar instrucciones, afinar criterios operativos y anticipar contingencias antes del despliegue formal del trabajo de campo, reduciendo así fuentes evitables de error de medición y variación no controlada (Creswell & Plano Clark, 2018; DeVellis, 2017; Díaz Muñoz, 2020).

4.1.3 Prueba de campo y verificación operativa previa

La prueba piloto se ejecutó en la ventana inmediatamente previa al inicio formal del trabajo de campo, correspondiente a las semanas -2 a -1 del cronograma referencial del estudio; es decir, en el periodo previo a la medición basal (PRE) y antes del despliegue regular de la intervención en aula. Esta ubicación temporal permitió verificar en condiciones presenciales equivalentes la claridad de instrucciones, la factibilidad logística y la estabilidad operativa de los instrumentos, preservando la secuencia lógica del diseño y estandarizando condiciones de aplicación comparables entre clases.

Esta fase se utilizó como mecanismo de verificación previa para asegurar que los instrumentos y procedimientos fueran comprensibles, aplicables en el tiempo disponible y consistentes con el contexto presencial de implementación. En esta fase se evaluó la claridad de instrucciones, la carga asociada al formato de respuesta, la factibilidad logística en aula y la estabilidad de los criterios de calificación. Asimismo, se registraron los tiempos reales requeridos para la lectura de instrucciones, la resolución de la tarea y el cierre de la aplicación, con el fin de verificar su compatibilidad con la ventana disponible de clase. De manera paralela, se identificaron ambigüedades concretas vinculadas con la extensión esperada del entregable, la secuencia de respuesta, la

interpretación de algunos criterios de rúbrica y la comprensión de instrucciones operativas, por su potencial para introducir error sistemático o heterogeneidad no deseada en la medición de línea base.

Los ajustes derivados del piloto se documentaron en una bitácora de cambios y se integraron en una versión final controlada del instrumento y sus guías de aplicación. Metodológicamente, los cambios se limitaron a refinamientos de forma (redacción, orden, precisión de instrucciones, consistencia de escalas y criterios) sin alterar los constructos ni los dominios evaluados, preservando la equivalencia conceptual de la medición. La trazabilidad de dichos ajustes se vincula con la bitácora del piloto y ajustes previos (Tabla A4.3), así como con la documentación de anclajes y estandarización de rúbricas (Apéndices A2.1 y A2.2).

Como resultado del piloto, se estandarizaron las instrucciones operativas para la tarea PRE/POST, se corrigieron y verificaron los componentes de cómputo y reescalamiento de las rúbricas, y se definieron reglas de ceguera, resolución de discrepancias y tiempos máximos de administración, quedando estos ajustes documentados y verificables en la Tabla A4.3.

En correspondencia con estos ajustes previos, el trabajo de campo se organizó en fases secuenciales con puntos de control (*quality gates*), cuidando su compatibilidad con el ciclo académico 2024–2025 y con la lógica temporal del diseño PRE–implementación–POST. Este orden no se modificó en ninguna clase, ya que constituye la condición mínima para asegurar comparabilidad entre condiciones y preservar la interpretabilidad del contraste de cambio. Por ello, los periodos se presentan como referencias relativas al calendario del módulo (semanas previas, semanas de implementación y semanas posteriores), dado que la calendarización concreta depende del calendario oficial de cada curso y de la disponibilidad institucional de recursos de realidad virtual; aun así, la estructura del cronograma garantiza que los hitos metodológicos críticos se ejecuten en el mismo orden y bajo reglas comunes.

La Tabla 4.1 sintetiza el cronograma operativo por fase e incorpora, para cada tramo, el responsable primario, los recursos mínimos, el control operativo y la evidencia verificable. Esta evidencia se vincula de forma explícita con los apéndices de control de

calidad y gobernanza (Apéndice A) y con la bitácora técnica y el *pipeline* reproducible (Apéndice B), con el propósito de fortalecer la trazabilidad del proceso desde la preparación metodológica y el resguardo de información hasta la congelación de la base final y la preparación de insumos auditables.

Además, esta organización permite diferenciar con mayor claridad las actividades de preparación, aplicación, control de calidad y cierre analítico, evitando que el trabajo de campo se interprete como una secuencia meramente administrativa. Por el contrario, cada fase cumple una función metodológica específica dentro del estudio: estabilizar los instrumentos, proteger la comparabilidad entre grupos, documentar la exposición efectiva a la intervención y asegurar que la base final utilizada para el análisis responda a criterios explícitos de integridad, trazabilidad y reproducibilidad.

Tabla 4.1

Cronograma operativo del trabajo de campo, responsables, recursos y evidencia verificable (ciclo 2024–2025)

Fase	Periodo (según cronograma académico)	Responsable principal	Recursos mínimos	Control operativo	Evidencia verificable
Preparación metodológica y estandarización	Semanas –6 a –4	Investigador principal	Instrumentos finales, guías de sesión, formatos de registro	Control de versiones; criterios de resguardo; definición de componentes esenciales y reglas de dosificación/registro	A2 (rúbricas); A5.3 (resguardo/trazabilidad); A6.4 (declaración); B3 (<i>pipeline</i>)
Inducción y seguridad para RV	Semanas –4 a –3	Investigador principal y docente	Protocolo de uso del visor; <i>checklist</i> operativo; plan de contingencia; formato de incidentes	Criterios de pausa/tiempos máximos cuando aplique; registro de incidentes; condiciones mínimas del espacio	A6.3 (inducción/seguridad); A7.1 (riesgos/contingencias); A7.4 (<i>checklist</i> sesión)
Calibración de evaluadores	Semanas –3 a –2	Evaladores y investigador principal	Productos ancla; guía de calificación; rúbricas analíticas	Ronda piloto de puntuación; resolución de discrepancias; preparación de hoja de control ICC; evidencia de ceguera	A3.2 (ICC y decisión); A2.2 (guía/ancajes); A3.1 (discrepancias Δ)
Pilotaje de instrumentos de proceso (si aplica)	Semanas –2 a –1	Investigador principal	Instrumentos de autoinforme/proceso; registro de tiempos; bitácora	Verificación de claridad y factibilidad; ajustes menores documentados; consistencia operativa entre grupos	A4.3 (bitácora de ajustes); A2.2 (nota de aplicación)
Medición PRE (línea base)	Semana 1	Docente e investigador principal	Instrucciones estandarizadas; tiempos; resguardo físico/digital	<i>Checklist</i> de integridad y coherencia; verificación de completitud inicial	A4.1 (integridad/QA); A5.1 (inventario); A5.3 (trazabilidad)

Fase	Periodo (según cronograma académico)	Responsable principal	Recursos mínimos	Control operativo	Evidencia verificable
Implementación IA y RV	Semanas 2 a 8	Docente	Material didáctico; protocolo de implementación; formatos de registro	Registro de exposición efectiva y fidelidad por sesión; control de incidencias; <i>quality gates</i>	A7.2 (<i>quality gates</i>); A7.4 (<i>checklist</i> por sesión); A7.1 (incidencias/riesgos)
Medición POST (cierre)	Semana 9 o 10	Docente e investigador principal	Instrucciones estandarizadas; control de faltantes	Reglas de reposición justificada; coherencia entre identificadores (clase, estudiante, condición, momento)	A4.1 (faltantes/QA); A5.3 (trazabilidad); B3 (pasos 1–3)
Calificación doble ciega	Posterior a la aplicación	Dos evaluadores	Rúbricas analíticas; guía de anclajes	Ceguera entre evaluadores; protocolo de discrepancias; tercer dictamen si aplica	A3.1 (Δ); A3.2 (ICC y decisión); A2 (rúbricas)
Depuración y base final	Semanas +1 a +2	Analista e investigador principal	Diccionario de variables; <i>scripts</i> /versionado	Validación de rangos/duplicados/coherencias; congelamiento de base; exportación reproducible	A4.1–A4.3 (QA/bitácora); A5.3 (trazabilidad); B3 (exportación)

Nota. La tabla resume la secuencia operativa del trabajo de campo y los puntos de control que aseguran trazabilidad (bitácoras, resguardo y *quality gates*) desde la preparación hasta la congelación de la base. N = 180 describe el flujo de implementación; n = 160 corresponde a la muestra analítica per-protocol usada para la inferencia principal.

Con el fin de traducir el cronograma y los controles operativos de la Tabla 4.1 a una representación secuencial fácil de auditar, la Figura 4.1 presenta un diagrama operativo del trabajo de campo, organizado como una secuencia de etapas con puntos de verificación explícitos. El diagrama permite visualizar, en un solo trazo, la lógica de preparación, pilotaje, medición PRE, implementación, medición POST y aseguramiento final, así como los *gates* de calidad y trazabilidad que conectan cada fase con su evidencia verificable en apéndices. Esta figura cumple una función de integración: no sustituye el detalle tabular, sino que facilita la lectura del proceso completo y refuerza la coherencia entre calendarización, gobernanza de datos y base analítica utilizada para el diagnóstico.

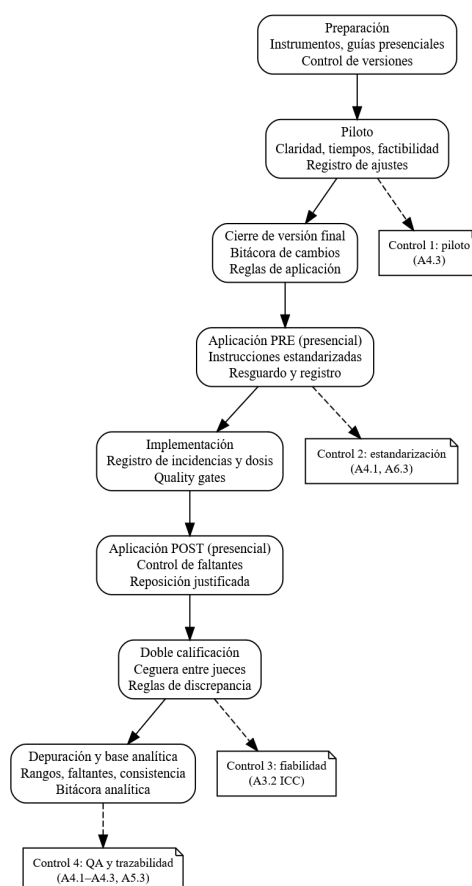
En términos metodológicos, esta representación visual también permite verificar que el diagnóstico no se construyó a partir de una aplicación aislada de instrumentos, sino mediante una cadena organizada de decisiones, controles y evidencias sucesivas. Esta distinción es relevante porque el valor del trabajo de campo no depende únicamente de haber aplicado mediciones PRE y POST, sino de haber asegurado condiciones mínimas de comparabilidad, resguardo, consistencia operativa y documentación durante todo el proceso. Por ello, la figura funciona como un puente entre la planeación descrita

en la tabla y la lógica de aseguramiento que sostiene la interpretación posterior de los resultados.

Esta forma de representación también responde a una exigencia metodológica relevante en estudios de intervención educativa: documentar no solo qué se aplicó, sino cómo se aplicó, con qué grado de adherencia y bajo qué condiciones operativas. La literatura sobre fidelidad de implementación advierte que la interpretación de los efectos de una intervención depende de distinguir entre el diseño previsto, la entrega real y la exposición efectiva de los participantes, especialmente cuando se trabaja en contextos educativos naturales y con variabilidad entre grupos (Bellg et al., 2004; Carroll et al., 2007; Campbell et al., 2012). Por ello, la Figura 4.1 contribuye a hacer visible la cadena de implementación que sostiene la validez del diagnóstico y permite interpretar los resultados posteriores como producto de un proceso controlado, documentado y metodológicamente auditable.

Figura 4.1

Diagrama operativo del trabajo de campo para la construcción del diagnóstico



Nota. El mapa conceptual resume la secuencia operativa y los puntos de control (piloto, estandarización, resguardo y bitácoras) que sustentan la verificabilidad del diagnóstico.

4.2 Aplicación de Instrumentos

La aplicación de los instrumentos se organizó en congruencia con el diseño cuasi experimental por conglomerados con mediciones PRE y POST, con el propósito de garantizar consistencia operativa, control de calidad y trazabilidad documental desde la preparación hasta el cierre del levantamiento. La implementación fue presencial y se estructuró por fases con responsables definidos (investigador principal, docente, evaluadores y analista), recursos mínimos necesarios y evidencia verificable asociada a cada hito, con el fin de reducir variaciones no deseadas en la aplicación y asegurar comparabilidad entre participantes. La planeación consideró ventanas de aplicación integradas al calendario del curso, resguardo de instrumentos, reglas de contingencia

ante ausencias o incidencias y criterios de documentación del proceso, manteniendo integridad del diagnóstico y auditabilidad del trabajo de campo.

4.2.1 Método de aplicación y levantamiento de la información

La medición de línea base (PRE) constituye la evidencia empírica central para caracterizar el estado inicial del problema y fundamentar el diagnóstico. Esta medición corresponde al levantamiento inicial realizado antes de la exposición a la arquitectura didáctica basada en IA y RV, y su función metodológica fue establecer un punto de referencia comparable para valorar posteriormente los cambios observados en los momentos subsiguientes. La aplicación fue presencial y se realizó bajo condiciones estandarizadas, con instrucciones unificadas, ventanas temporales definidas y registro de incidencias. Este control es crítico, ya que el diagnóstico no debe reflejar variaciones de administración, sino el fenómeno medido. En consecuencia, se documentó integridad de captura, consistencia básica y verificación de rangos conforme al *checklist* operativo de integridad y coherencia (Apéndice A4.1), y se mantuvo resguardo y control de acceso bajo las reglas de gobernanza descritas en el Apéndice A5.

La aplicación de los instrumentos no se limitó a una administración puntual, sino que se articuló con una estructura operativa común para la condición de intervención, con el fin de asegurar comparabilidad entre sesiones, grupos y participantes. En consecuencia, además del registro de la medición y de las condiciones de ejecución, se definieron componentes esenciales de implementación, reglas mínimas de entrega por sesión y mecanismos de verificación de adherencia, de modo que la evidencia empírica generada pudiera interpretarse a la luz de condiciones de aplicación homogéneas y no de variaciones evitables en la entrega. Este soporte operativo se respaldó con lineamientos de uso y privacidad (Apéndice A6) y con controles de riesgo, seguridad y continuidad de implementación (Apéndice A7).

Como complemento a esta documentación operativa, resulta pertinente mostrar el entorno real en que se desarrolló la intervención. La Figura 4.2 se incorpora como evidencia contextual verificable del despliegue presencial en el laboratorio, mostrando la ecología de aplicación en la que se articularon la inducción operativa, el uso de visores y

la práctica inmersiva. Su propósito no es aportar un dato analítico ni funcionar como instrumento de medición, sino ofrecer transparencia sobre las condiciones materiales y pedagógicas bajo las cuales se ejecutó la intervención y facilitar la lectura posterior de la fidelidad, la dosificación y el control de variabilidad operacional.

Figura 4.2

Implementación en aula de la intervención IA+RV durante una sesión de práctica inmersiva en el laboratorio de negocios internacionales



Nota. Registro fotográfico de una sesión real en el Laboratorio de Negocios Internacionales durante la condición de intervención (IA+RV). Se observa al autor (investigador) durante la fase de inducción operativa al uso de visores; en pantalla se muestra el entorno de simulación (*VirtualSpeech*). Para proteger la confidencialidad, los participantes fueron anonimizados. La figura se incluye como evidencia documental del contexto de implementación y no se utilizó como instrumento de medición ni se codificó como variable en los análisis.

Sobre esta base contextual y operativa, la intervención se organizó mediante criterios explícitos de estandarización orientados a asegurar comparabilidad entre sesiones, grupos y participantes.

Esta estandarización se tradujo en una secuencia instruccional común para la condición de intervención, con especificación de componentes esenciales y reglas

mínimas de entrega por sesión, independientemente del canal previsto (IA o RV). En términos operativos, se fijaron criterios de adherencia, dosificación y registro con el fin de diferenciar la mera disponibilidad tecnológica de la exposición efectiva. Este enfoque se alinea con la literatura sobre fidelidad en intervenciones complejas, que recomienda definir con precisión los componentes esenciales, documentar la variación de implementación y establecer procedimientos de verificación para evitar que la ausencia de efecto se confunda con fallas de entrega (Bellg *et al.*, 2004; Campbell *et al.*, 2012). En paralelo, se definieron mecanismos de seguimiento mediante bitácoras por sesión y listas de cotejo de ejecución, así como reglas de contingencia para situaciones en las que una sesión con RV no pudiera realizarse por causas técnicas o logísticas, preservando en esos casos una alternativa pedagógicamente equivalente y debidamente documentada para su consideración analítica.

La intervención se operacionalizó mediante un ciclo instruccional presencial estandarizado compuesto por tres momentos: *briefing*, práctica (inmersiva o equivalente) y *debriefing*. El *briefing* estableció objetivos de la sesión, criterios de desempeño y reglas de seguridad, integrando explícitamente lineamientos de privacidad e integridad académica para el uso de IA. La práctica se centró en la realización de tareas dentro del entorno RV, asegurando condiciones mínimas comparables (tiempo efectivo, funcionamiento técnico, pausas y monitoreo). El *debriefing* funcionó como instancia pedagógica de cierre y metaaprendizaje, donde se revisaron decisiones, evidencia del desempeño y criterios de mejora, reforzando que la IA opera como apoyo formativo y no como sustituto del desempeño evaluable.

Esta arquitectura reduce el riesgo de heterogeneidad didáctica entre sesiones al fijar un ciclo mínimo común (*briefing*–práctica–*debriefing*) y reglas de control que estandarizan la exposición efectiva, la seguridad operativa y el uso responsable de IA. En particular, la estandarización preserva la comparabilidad entre grupos incluso ante incidencias técnicas o condiciones individuales que obliguen a una alternativa pedagógicamente equivalente, porque el componente esencial no es el medio (RV vs. sustitución), sino la secuencia de objetivos, criterios explícitos y retroalimentación formativa anclada en rúbrica. La evidencia operativa del ciclo y su verificación se

documentó mediante *checklists* y bitácoras, conforme a los protocolos de inducción y operación de RV (Apéndice A7) y a los lineamientos de uso de IA (Apéndice A6).

Tabla 4.2

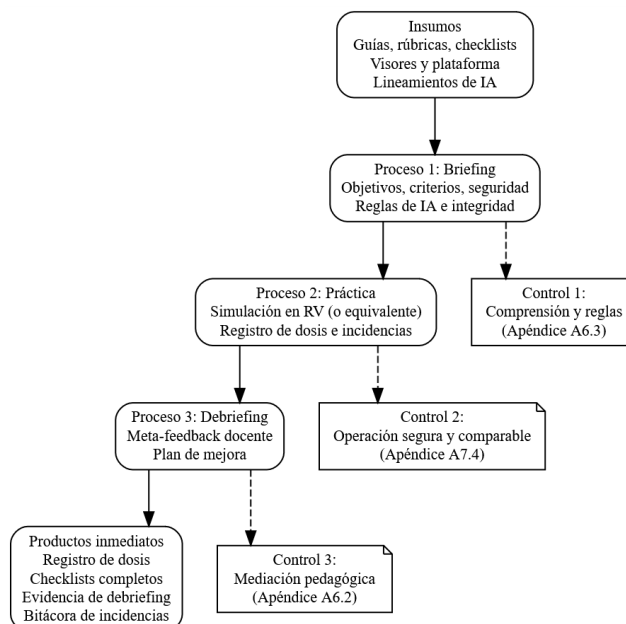
Ciclo instruccional estandarizado de la intervención y evidencia verificable

Momento	Propósito operativo	Acciones mínimas estandarizadas	Control de variabilidad	Evidencia verificable
Briefing	Unificar objetivos, criterios y reglas	Objetivo de sesión, criterios de rúbrica, reglas de seguridad RV, reglas de uso responsable de IA	Guion mínimo común y verificación de comprensión	<i>Checklist</i> de inducción y comprensión (Apéndice A6.3); inducción breve (Apéndice A7.3)
Práctica (RV o equivalente)	Ejecutar tarea y registrar dosis real	Configuración, calibración, pausas, monitoreo de malestar, registro de tiempo efectivo	<i>Checklist</i> por sesión y registro de incidencias	<i>Checklist</i> operativo por sesión (Apéndice A7.4); bitácora/incidencias (Apéndice A7.6)
Debriefing	Consolidar aprendizaje y orientar iteración	Revisión de desempeño, retroalimentación formativa, acuerdos de mejora, recordatorio de integridad académica	Uso de rúbrica como lenguaje común, mediación docente	Evidencia de <i>debriefing</i> en <i>checklist</i> por sesión; lineamientos de IA (Apéndice A6.2)

Nota. La tabla define el ciclo mínimo por sesión y la evidencia documental que respalda su aplicación (*checklists* y registros). La estandarización busca asegurar fidelidad de implementación y comparabilidad.

Para facilitar la lectura integrada del ciclo y explicitar su lógica causal-operativa, la Figura 4.3 representa el encadenamiento entre insumos (recursos y reglas), proceso (*briefing*–*práctica*–*debriefing*) y productos inmediatos (registros, evidencias y acuerdos de mejora) que sostienen la fidelidad de implementación. Mientras la Tabla 4.2 detalla acciones mínimas, controles y evidencia verificable por momento, el diagrama permite visualizar de forma sintética cómo cada fase se traduce en un punto de control documentable y cómo se preserva la comparabilidad aun cuando se active una alternativa equivalente.

Esta representación resulta especialmente útil porque muestra que la fidelidad no se limita al cumplimiento formal de actividades, sino que depende de la conexión entre recursos disponibles, decisiones docentes, participación estudiantil y evidencia generada en cada momento del ciclo. Al hacer explícita esta secuencia, el diagrama permite verificar que la intervención mantuvo una lógica común entre grupos y sesiones, aun cuando se realizaron ajustes operativos o se recurrió a alternativas equivalentes para preservar la continuidad pedagógica y la seguridad de los participantes.

Figura 4.3*Lógica operacional de la intervención presencial IA+RV*

Nota. El diagrama articula insumos, secuencia didáctica y productos inmediatos verificables (*checklists*, bitácoras, acuerdos de mejora) que funcionan como puntos de control para fidelidad y trazabilidad. Su propósito es mostrar cómo la estandarización preserva comparabilidad entre sesiones y grupos, aun ante incidencias que requieran una alternativa pedagógicamente equivalente.

4.2.2 Implementación de la intervención IA+RV: fidelidad, dosis y control de variabilidad operacional

Una vez definida la arquitectura instruccional común de la intervención, la implementación se documentó en términos de fidelidad, dosificación y control de variabilidad operacional. En esta fase, el interés metodológico no fue únicamente describir la experiencia, sino demostrar cómo se registró la exposición efectiva, cómo se controlaron las desviaciones de ejecución y cómo se preservó la trazabilidad necesaria para sostener la validez interna del contraste posterior. Con este fin, se integraron registros de dosis, controles de integridad académica, evidencia operativa verificable y documentación reproducible de la implementación.

La dosificación de la intervención se definió como una variable operacional crítica, dado que diferencias en la exposición efectiva pueden sesgar el cambio observado y debilitar la comparabilidad entre participantes y grupos. Por ello, en cada

sesión se documentó el tiempo efectivo de práctica como medida de exposición real, distinguiéndolo del tiempo planificado, y se registraron eventos capaces de alterar la dosis recibida, como interrupciones por fallas técnicas, tiempos de calibración, pausas por seguridad y activación de alternativas pedagógicamente equivalentes. Este registro cumple dos funciones metodológicas complementarias: en primer lugar, opera como control de fidelidad, al verificar que la intervención mantuvo una exposición mínima comparable y que cualquier desviación relevante quedó identificada; en segundo lugar, habilita análisis complementarios, como exploraciones dosis–respuesta o análisis de sensibilidad, evitando confundir la variación en la exposición con cambios atribuibles al diseño instruccional.

El registro de dosis se alineó con los *checklists* operativos por sesión (Apéndice A7) y se integró a la bitácora analítica y de calidad (Apéndice A4), de manera que toda desviación con potencial impacto en comparabilidad quedara documentada con fecha, descripción del evento y decisión operativa. En términos de gobernanza, este procedimiento asegura trazabilidad entre la evidencia de implementación y los insumos analíticos, y permite distinguir entre (i) diferencias de desempeño atribuibles al aprendizaje y (ii) variación no deseada asociada a exposición desigual, contingencias técnicas o condiciones de seguridad.

En la condición IA+RV, la plataforma VirtualSpeech operó como entorno de práctica deliberada y retroalimentación asistida por IA dentro del ciclo *briefing–ejecución–debriefing*. Tras cada interacción, el sistema genera un reporte de desempeño con indicadores conductuales y comunicacionales, por ejemplo, contacto visual, lenguaje corporal, listenability, ritmo del habla y uso de muletillas, además de retroalimentación accionable y rutas de entrenamiento o microcursos sugeridos para corregir áreas específicas. Estos reportes se utilizaron como insumo inmediato de *debriefing* y como guía de práctica entre sesiones, constituyendo un componente material explícito de la intervención; es decir, no solo exposición a RV, sino un ciclo estructurado de práctica, retroalimentación e iteración. Este diseño es consistente con principios de aprendizaje autorregulado y práctica deliberada, donde el *feedback*

específico y oportuno facilita calibración, ajuste de estrategias y mejora progresiva del desempeño (Zimmerman, 2002; Ericsson, 2008; Hattie & Timperley, 2007).

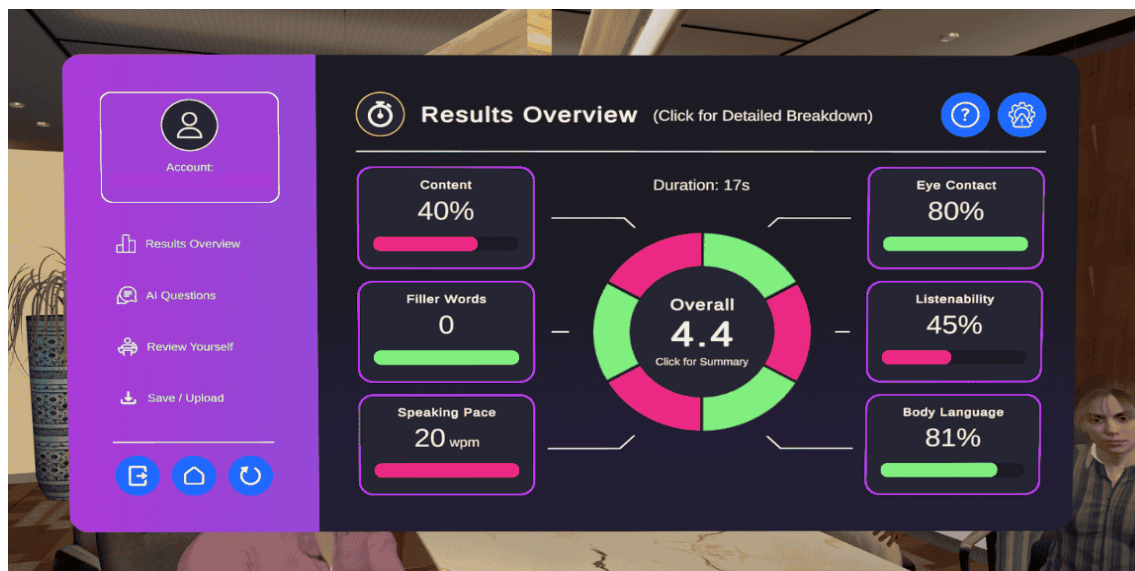
Adicionalmente, el rol del avatar y el nivel de dificultad de la interacción se parametrizaron mediante *prompts* definidos por el docente, lo que permitió estandarizar escenarios, controlar el tipo de demanda conversacional y graduar la complejidad del reto. En consecuencia, la intervención combinó (i) control instruccional del escenario, (ii) práctica situada en RV y (iii) retroalimentación inmediata basada en métricas, reduciendo la ambigüedad típica de ejercicios de role-play no estructurados y favoreciendo consistencia de implementación entre grupos y sesiones. Las capturas de pantalla incluidas (Figuras 4.4A–4.4C) se presentan únicamente como evidencia ilustrativa del tipo de reportes y paneles generados por el sistema utilizado durante el trabajo de campo.

Las escalas experienciales se aplicaron en POST, inmediatamente tras la exposición, con el propósito de reducir sesgo de recuerdo y capturar la experiencia situacional con alta validez ecológica; por ello, se interpretan como evidencia de proceso y mecanismo, no como *outcomes* longitudinales equivalentes al desempeño en rúbricas (Shadish *et al.*, 2002).

Esta evidencia visual resulta pertinente porque permite documentar el tipo de información que recibía el estudiante después de la práctica inmersiva y, al mismo tiempo, ayuda a distinguir la intervención de una actividad tecnológica meramente experiencial. En la lógica del estudio, los paneles generados por la plataforma funcionaron como insumos de retroalimentación formativa, ya que ofrecieron señales inmediatas sobre dimensiones observables del desempeño comunicativo y facilitaron la reflexión posterior. Esta función es consistente con la literatura sobre retroalimentación efectiva, en la medida en que el valor pedagógico del feedback depende de su oportunidad, especificidad y capacidad para orientar acciones de mejora (Hattie & Timperley, 2007; Shute, 2008). Así, las Figuras 4.4A–4.4C no se incorporan como resultados analíticos, sino como evidencia contextual del entorno tecnológico que apoyó la práctica, la autorregulación y el cierre reflexivo de la sesión.

Figura 4.4A

Panel ilustrativo de resultados y analítica automatizada en VirtualSpeech.

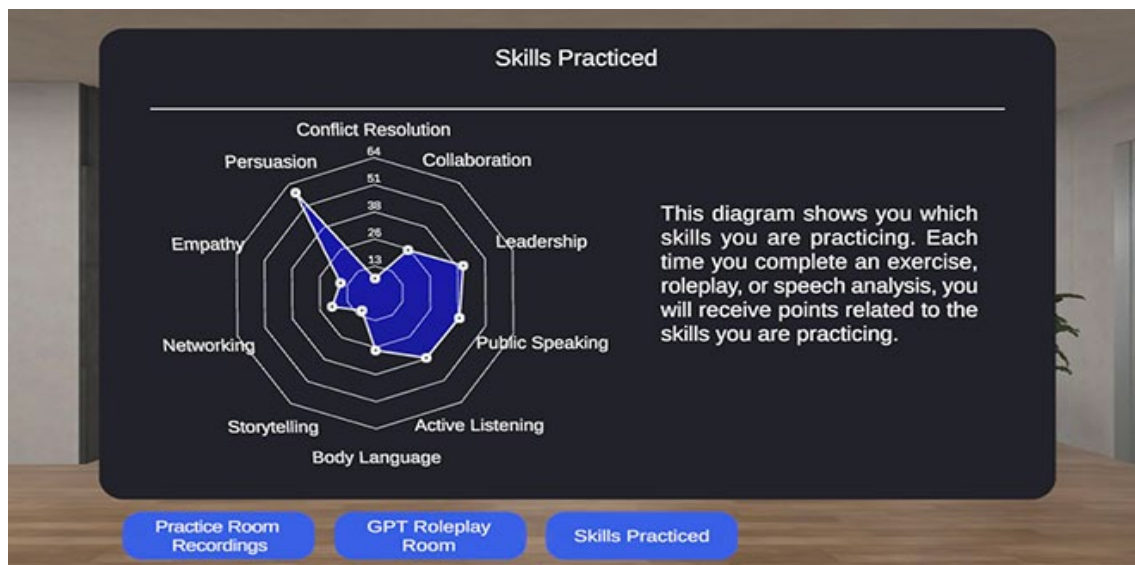


Nota. Captura de pantalla ilustrativa del tipo de métricas y retroalimentación automatizada que la plataforma despliega tras una práctica (p. ej., contacto visual, escucha, lenguaje corporal, muletillas y duración). No corresponde a resultados de estudiantes del estudio ni contiene datos del piloto. Captura de pantalla de *VirtualSpeech* (© *VirtualSpeech Ltd.*, usada con fines educativos bajo licencia institucional del Tecnológico de Monterrey). La imagen se emplea exclusivamente con fines ilustrativos; no forma parte de los análisis estadísticos.

De forma complementaria, la siguiente captura permite observar cómo la plataforma organiza la práctica comunicativa en habilidades y subcompetencias específicas, lo cual resulta relevante para comprender la lógica formativa de la intervención. Mientras la Figura 4.4A ilustra el tipo de retroalimentación automatizada posterior a una práctica, la Figura 4.4B muestra la estructura de habilidades que orienta dicha retroalimentación y permite vincular la experiencia inmersiva con criterios concretos de desempeño. Esta desagregación contribuye a reducir la ambigüedad evaluativa, ya que transforma una actividad general de comunicación oral en componentes observables que pueden ser practicados, monitoreados y discutidos durante el *debriefing*. En consecuencia, la imagen ayuda a contextualizar cómo el entorno tecnológico apoyó la práctica deliberada y la reflexión criterial dentro de la intervención IA+RV.

Figura 4.4B

Mapa ilustrativo de habilidades practicadas y desglose de subcompetencias en VirtualSpeech.



Nota. Captura de pantalla referencial que muestra una visualización agregada de habilidades y registros de práctica dentro de la plataforma (p. ej., categorías de habilidades y conteos asociados a sesiones/actividades). La figura se incorpora para contextualizar el tipo de seguimiento interno y reportes disponibles en el sistema como apoyo a la práctica deliberada y al *debriefing*. En la intervención, este tipo de analítica se integró operativamente a un ciclo de retroalimentación en dos capas: (i) *feedback* automatizado e inmediato generado por la plataforma tras cada práctica, y (ii) retroalimentación docente posterior durante el *debriefing*, que reafirmó criterios de rúbrica, modeló estrategias y orientó revisiones específicas. Esta combinación de inmediatez + mediación pedagógica se describe como un mecanismo instruccional plausible que puede facilitar iteraciones de mejora entre PRE y POST; sin embargo, la evidencia del efecto se sustenta en los estimadores inferenciales (p. ej., interacción Tiempo×Grupo en modelos multinivel) y en las rúbricas analíticas, no en la presente captura. No representa resultados empíricos del estudio ni evidencia individual de participantes. Captura de pantalla de la plataforma *VirtualSpeech* (© *VirtualSpeech Ltd.*, usada con fines educativos y bajo licencia institucional del Tecnológico de Monterrey). La imagen se emplea exclusivamente con fines ilustrativos; no forma parte de los análisis estadísticos.

Figura 4.4C

Escenario ilustrativo de práctica guiada en VirtualSpeech



Nota. Captura de pantalla de referencia que muestra un ejemplo de escenario/actividad y su interfaz de navegación. Se presenta exclusivamente para contextualizar el tipo de interacción y la ruta de práctica; no corresponde a evidencia empírica del estudio ni a desempeños individuales. Captura de pantalla de la plataforma *VirtualSpeech* (© *VirtualSpeech Ltd.*, usada con fines educativos y bajo licencia institucional del Tecnológico de Monterrey). La imagen se emplea exclusivamente con fines ilustrativos; no forma parte de los análisis estadísticos.

Además del soporte material de la plataforma, la implementación del estudio requirió una articulación explícita entre técnicas de obtención de información y procedimientos de validación, con el fin de sostener una lectura convergente entre desempeño, proceso e implementación. En términos de Rojas Crotte (2011), la técnica debe entenderse como un procedimiento orientado a obtener y transformar información útil, siempre vinculado con el método, el instrumento, la aplicación, el procesamiento y la validación. Bajo esta lógica, en la presente investigación se integraron, al menos, cuatro componentes técnicamente diferenciados: (i) evaluación de desempeño mediante rúbricas analíticas con doble calificación ciega, como técnica principal para los *outcomes* observables; (ii) escalas de autoinforme post-exposición, como técnica complementaria para captar procesos y mecanismos percibidos; (iii) registros de implementación, dosis y fidelidad mediante *checklists*, bitácoras e incidencias, como soporte de trazabilidad operativa; y (iv) entrevistas y observación con función

cualitativa-explicativa para matizar mecanismos, barreras y facilitadores. En consecuencia, la evidencia del estudio debe leerse como resultado de una estrategia de integración técnica y analítica, y no como producto de una única fuente o procedimiento aislado.

4.2.3 Seguridad, contingencia y condiciones operativas para RV

Para las sesiones con RV se establecieron protocolos de inducción, seguridad y contingencia orientados a reducir incidentes, estandarizar la experiencia y proteger al estudiante sin penalización académica. Estos protocolos incluyeron reglas de uso del visor, límites operativos de uso continuo, criterios de pausa, registro de incidentes y condiciones mínimas del espacio físico. Esta previsión responde a dos necesidades metodológicas complementarias: proteger a los participantes y evitar que incidentes o incomodidad alteren el desempeño medido de manera diferencial entre clases. En investigación educativa aplicada, este tipo de control forma parte del aseguramiento de la validez de interpretación, porque disminuye fuentes de varianza no deseada vinculadas al contexto de aplicación (AERA, APA, & NCME, 2014; Kane, 2006). Con ello, se reduce la varianza no deseada asociada al contexto físico y a la tolerancia individual a la inmersión, preservando la comparabilidad del desempeño evaluado.

Estos controles respondieron, en particular, al riesgo de malestar asociado a la inmersión o al ciber-mareo, ya que un incidente no gestionado puede traducirse en pérdida diferencial de exposición y, con ello, en sesgos por deserción operativa o por reducción sistemática de dosis en subgrupos específicos. Por esta razón, se estableció un procedimiento estandarizado de contingencia que prioriza la seguridad del estudiante y, al mismo tiempo, preserva la comparabilidad pedagógica del diseño.

Cuando un estudiante reportó malestar o se observó afectación funcional durante la práctica, se activó el protocolo de contingencia con los siguientes componentes mínimos: (a) interrupción segura e inmediata de la actividad inmersiva, (b) periodo de recuperación y verificación de condiciones para continuar, (c) registro del incidente con descripción del evento, momento y decisión operativa, y (d) determinación documentada de una de dos rutas: reincorporación controlada (si procede) o activación de alternativa

equivalente. La alternativa equivalente se diseñó bajo el principio de seguridad con equivalencia, manteniendo constantes tres elementos: el objetivo instruccional de la sesión, el criterio de evaluación (rúbrica aplicable) y el cierre formativo (*debriefing*), de modo que el ajuste no altere la intención pedagógica ni introduzca una ventaja o desventaja sistemática para el estudiante.

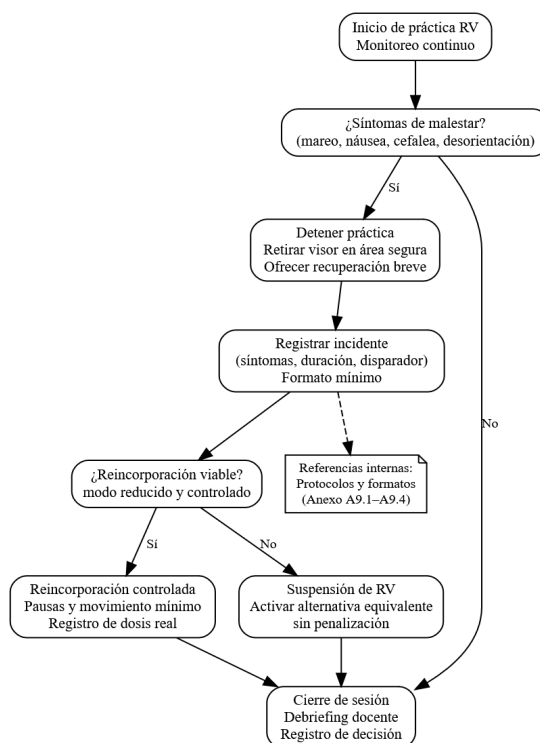
Los protocolos y plantillas de reporte asociados a este control se resguardan en el Apéndice A7, incluyendo inducción breve, *checklist* operativo por sesión, protocolo ante ciber-mareo y formato de incidente. Este paquete documental permite distinguir entre la exposición planificada y la exposición efectivamente recibida, y garantiza que cualquier contingencia quede trazable como evento de implementación, evitando interpretaciones erróneas de los resultados atribuibles a variación operativa no registrada.

Para explicitar la lógica de decisión y asegurar replicabilidad, la Figura 4.5 sintetiza el árbol operativo ante incidentes en RV, así como los puntos mínimos de registro y la activación de alternativas equivalentes bajo reglas consistentes. Con ello, la gestión de seguridad deja de ser un elemento implícito y se convierte en un componente metodológico verificable del trabajo de campo.

La utilidad metodológica de este árbol de decisión radica en que convierte la gestión de incidentes en un procedimiento previamente definido y no en una reacción discrecional durante la intervención. Esto es especialmente importante en experiencias inmersivas, donde la seguridad, la continuidad pedagógica y la comparabilidad entre participantes deben preservarse aun cuando existan eventos de incomodidad, ciber-mareo o interrupción técnica. En este sentido, el protocolo no solo protege al participante, sino que también resguarda la validez del estudio, al documentar cuándo se mantuvo la exposición prevista, cuándo se suspendió y cuándo fue necesario activar una alternativa equivalente bajo criterios homogéneos.

Figura 4.5

Árbol de decisión operativa ante incidentes en RV y activación de alternativa equivalente



Nota. El diagrama explicita decisiones operativas orientadas a proteger la seguridad y sostener comparabilidad, incorporando el registro mínimo requerido para trazabilidad.

4.2.4 Fidelidad de implementación y *quality gates* como evidencia de adopción controlada

La fidelidad de implementación se operacionalizó mediante un conjunto de criterios verificables que permiten distinguir entre una adopción nominal de la intervención y una adopción con evidencia. En este marco, los *quality gates* funcionan como umbrales mínimos de decisión: establecen condiciones operativas, de medición, de integridad y de equidad que deben cumplirse antes de considerar continuidad, escalamiento o replicación en nuevas cohortes. Metodológicamente, esta lógica fortalece la validez interna y la interpretabilidad de los resultados, porque documenta que la intervención no dependió de improvisación ni de variaciones idiosincráticas entre sesiones, sino de una arquitectura de ejecución controlada y auditable.

En coherencia con el registro de riesgos y mitigaciones (Apéndice A7.1), los *quality gates* se definieron en cinco dominios: (G1) operación segura y logística, (G2) fidelidad del ciclo *briefing*–práctica–*debriefing*, (G3) calidad de medición y fiabilidad interjueces, (G4) integridad académica en el uso de IA, y (G5) equidad mediante alternativas equivalentes cuando la RV no fuese viable. La Tabla 4.3 sintetiza estos umbrales y explicita la evidencia mínima requerida para su verificación, remitiendo a los apéndices donde se documentan formatos operativos (A6 y A7), gobernanza y resguardo (A5), y trazabilidad analítica (A4 y Apéndice B). Con ello, cada gate queda asociado a un rastro documental suficiente para sostener decisiones de adopción controlada sin ambigüedad.

Tabla 4.3

Quality gates de implementación

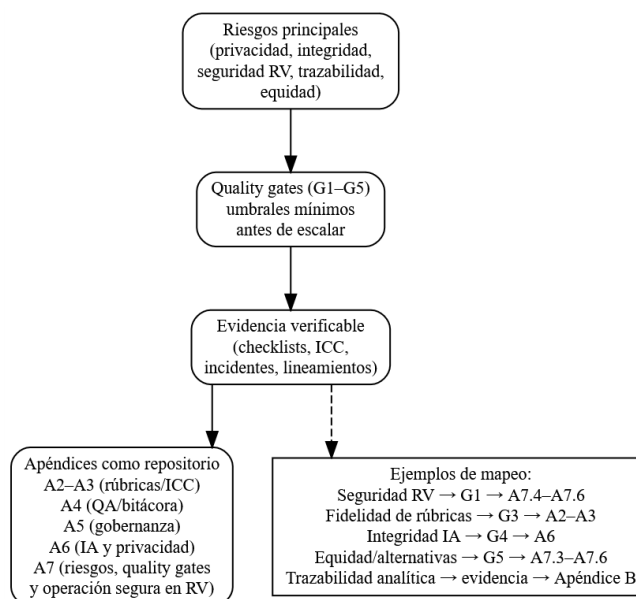
<i>Gate</i>	<i>Dominio</i>	<i>Criterio</i>	<i>Umbral mínimo</i>	<i>Evidencia verificable</i>	<i>Decisión</i>
G1	Operación	Seguridad y logística RV	0 incidentes con suspensión no gestionada y registro completo; $\geq 90\%$ <i>checklist</i> pre-sesión completo	<i>Checklist</i> operativo por sesión (Apéndice A7.4); reporte de incidentes si aplica (Apéndice A7.6)	Continuar / Ajustar operación
G2	Fidelidad	Ciclo <i>briefing</i> –práctica– <i>debriefing</i> documentado	$\geq 80\%$ de sesiones con <i>briefing</i> y <i>debriefing</i> registrados	Inducción y comprensión (Apéndice A6.3); inducción breve (Apéndice A7.3); evidencia de <i>debriefing</i> en registros de sesión	Continuar / Reforzar guion didáctico
G3	Medición	Fiabilidad interjueces en rúbricas	$ICC(2,k) \geq .75$ y estabilidad por criterio	Hoja de control ICC y calibración (Apéndice A3); rúbricas y guía de aplicación (Apéndice A2)	Aprobar medición / Recalibrar
G4	Integridad	Uso responsable de IA sin sustitución de desempeño	0 casos confirmados de sustitución en evidencias evaluables	Lineamientos de IA (Apéndice A6.2); declaración breve (Apéndice A6.4); trazabilidad mínima (Apéndice A6)	Escalar / Reforzar reglas
G5	Equidad	Alternativas equivalentes cuando RV no sea viable	100% de estudiantes con alternativa equivalente sin penalización	Protocolos y criterio de suspensión (Apéndice A7.3–A7.5); registro de incidente/alternativa (Apéndice A7.6)	Escalar / Corregir brecha

Nota. La tabla sintetiza umbrales operativos y de calidad para documentar fidelidad y reducir ambigüedad en decisiones de continuidad o escalamiento. La evidencia detallada (formatos, *checklists* y bitácoras) se encuentra en los apéndices referidos.

Para complementar la lectura *gate* por *gate* y hacer explícita la lógica de aseguramiento, la Figura 4.6 representa el mapeo entre riesgos, mecanismos de mitigación y evidencia verificable.

Figura 4.6

Mapa de aseguramiento de calidad: de riesgos a evidencia verificable



Nota. El diagrama muestra cómo los riesgos principales se traducen en *quality gates* y en evidencia verificable documentada en los apéndices del estudio. Su propósito es hacer explícita la relación entre controles de implementación, criterios de decisión y soportes documentales mínimos para continuidad o escalamiento.

En conjunto, los *quality gates* delimitan el alcance interpretativo de los resultados: el cambio observado puede leerse como asociado a una implementación presencial con fidelidad mínima y controles explícitos, y no como efecto de variaciones de ejecución o de riesgos no gestionados. Esta documentación, además, establece un criterio replicable para futuras cohortes, ya que explicita qué condiciones deben cumplirse para considerar la adopción como metodológicamente comparable entre clases y cohortes.

4.3 Procesamiento y análisis de la Información

Antes de presentar los resultados inferenciales, fue necesario explicitar cómo se organizaron, depuraron, documentaron y resguardaron los datos que sustentan el análisis. En consecuencia, esta sección describe el procesamiento de la información como un conjunto de decisiones metodológicas orientadas a asegurar integridad, verificabilidad y consistencia entre la evidencia primaria, la base analítica final y los procedimientos de estimación. Bajo esta lógica, el apartado se estructura en torno a la

gobernanza de datos, la documentación del flujo desde la implementación presencial hasta el *dataset* depurado, el anclaje del análisis y la especificación del modelo, con el fin de mostrar que los resultados posteriores descansan en una base técnicamente controlada y metodológicamente auditable.

4.3.1 Gobernanza de datos, documentación y reglas de trazabilidad

Previo al trabajo de campo se definieron reglas explícitas de gobernanza y trazabilidad: roles de acceso, almacenamiento, retención, anonimización o seudonimización, y criterios de control de calidad para *logs*, bitácoras y bases de rúbricas. Estas decisiones se justifican porque, en estudios con componentes digitales, la integridad analítica depende de que la ruta desde la fuente primaria hasta la variable derivada sea verificable. La documentación sistemática de transformaciones y exclusiones, así como la preservación de versiones, se vincula directamente con estándares de medición, transparencia y auditabilidad en investigación educativa (AERA *et al.*, 2014; Kane, 2006). En esta tesis, los apéndices resguardan plantillas, bitácoras y evidencia de versiones, mientras que el cuerpo del capítulo conserva la lógica metodológica evaluable.

4.3.2 Evidencia de implementación y trazabilidad del flujo de datos: de la ejecución presencial a la base analítica final

La evidencia de implementación se documentó con un enfoque de trazabilidad que integra dos planos complementarios: (i) el plano operativo, que registra condiciones de ejecución presencial, incidencias, dosis y activación de alternativas equivalentes cuando aplicó, y (ii) el plano analítico, que registra el tránsito desde evidencia primaria hasta la base depurada que alimenta los modelos estadísticos. Esta separación es intencional: permite demostrar que la intervención se ejecutó bajo controles definidos y, al mismo tiempo, que el procesamiento de datos siguió reglas explícitas de depuración, consistencia y gobernanza. Con ello se evita confundir cómo se implementó con cómo se procesó, y se fortalece la coherencia interna del capítulo al mantener diferenciadas la evidencia operacional y la evidencia analítica.

En el plano operativo, los registros por sesión constituyen evidencia mínima de fidelidad y control de variación no deseada. En particular, el *checklist* operativo por sesión, los reportes de incidente y el protocolo de respuesta ante malestar en RV (Apéndice A7) permiten documentar que las prácticas se realizaron bajo estándares de seguridad y continuidad pedagógica, y que cualquier desviación relevante se registró en el momento en que ocurrió. Esta evidencia se articula con la matriz de riesgos y mitigaciones para sostener que la implementación presencial no dependió de condiciones tácitas, sino de procedimientos verificables, consistentes entre sesiones y con capacidad de auditoría.

En el plano analítico, la trazabilidad se asegura mediante mecanismos mínimos pero suficientes: diccionario de variables, control de versiones, bitácora de calidad de datos, reglas de resguardo y acceso, y registro de cambios al plan analítico cuando fue estrictamente necesario. La bitácora analítica y el *checklist* de integridad (Apéndice A4) documentan decisiones sobre rangos válidos, consistencia, faltantes, codificación y reglas de tratamiento, evitando limpieza ad hoc no auditable. Complementariamente, la gobernanza (Apéndice A5) delimita responsabilidades y salvaguardas de resguardo (inventario, roles y reglas de acceso), lo que reduce el riesgo de modificaciones no controladas y sostiene la integridad del ciclo de vida de los datos desde su captura hasta su congelamiento.

Finalmente, la reproducibilidad se operacionaliza mediante el cuaderno de análisis y el *pipeline* computacional (Apéndice B). En lugar de describir procedimientos de forma exclusivamente narrativa, se documenta un flujo ejecutable que produce salidas estandarizadas y auditables (tablas y figuras) a partir de insumos definidos, con control de versiones, pasos replicables y registro mínimo de ejecución. Esta arquitectura cierra el circuito de trazabilidad: evidencia de implementación, datos depurados, *scripts* reproducibles y outputs publicables alineados al diagnóstico reportado.

Con el propósito de presentar esta lógica de manera compacta y verificable, la Tabla 4.4 organiza los artefactos mínimos por componente, explicita su finalidad metodológica, identifica el riesgo que mitigan y remite al apéndice correspondiente donde se encuentra la evidencia completa. La tabla funciona como un índice de

auditabilidad: permite localizar, para cada afirmación operativa o analítica, el soporte documental asociado.

Tabla 4.4

Evidencias de implementación y trazabilidad

Componente	Artefacto mínimo	Propósito metodológico	Riesgo que mitiga	Evidencia documental
Operación presencial	<i>Checklist</i> por sesión (pre–durante–post)	Estandarizar condiciones y registrar variaciones relevantes	Variabilidad operativa; fallas técnicas; dosis no comparable	Apéndice A7.4
Seguridad RV	Protocolo ante ciber-mareo + reporte de incidente	Respuesta estandarizada y trazabilidad de incidencias	Riesgo de salud; interrupciones no documentadas	Apéndices A7.5–A7.6
Integridad de uso IA	Lineamientos y declaración breve en materiales	Delimitar IA como apoyo formativo, no sustitución del desempeño	Uso indebido; sesgo por dependencia de IA en productos evaluables	Apéndices A6.2–A6.4
Medición	Rúbricas y guía de aplicación	Estandarizar criterios y anclajes conductuales	Error de medición; discrepancias por interpretación	Apéndice A2
Fiabilidad interjueces	Control de doble calificación + ICC (2,k)	Verificar acuerdo absoluto y activar recalibración si aplica	Inconsistencia entre jueces; ruido en resultados	Apéndice A3
Calidad de datos	<i>Checklist</i> de integridad + bitácora analítica	Registrar decisiones de depuración, rangos, faltantes y consistencia	Errores por limpieza ad hoc; pérdida de trazabilidad	Apéndice A4
Gobernanza	Inventario de datos + RACI + reglas de resguardo	Definir responsabilidades y salvaguardas del ciclo de vida	Accesos no controlados; reidentificación; cambios no auditables	Apéndice A5
Reproducibilidad	Cuaderno Python/Colab + <i>pipeline</i> + outputs estandarizados	Garantizar réplica de tablas/figuras y registro de ejecución	Imposibilidad de réplica; discrepancias por entorno	Apéndice B (B1–B4)

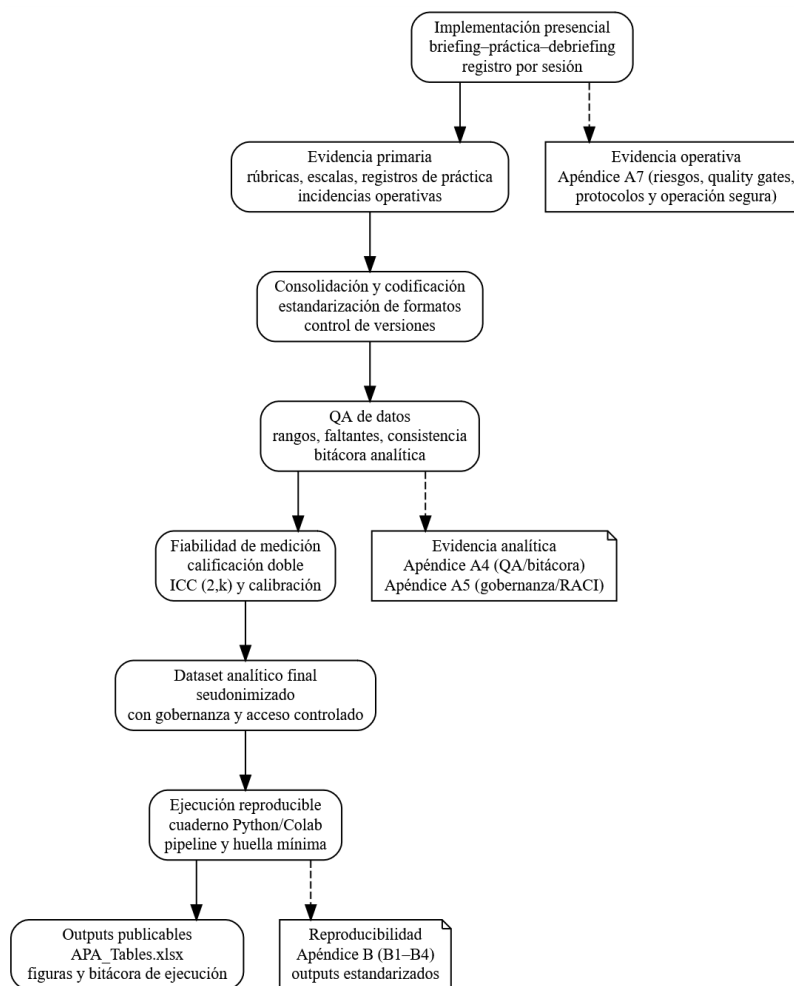
Nota. La tabla organiza la evidencia por tipo de artefacto y su función metodológica (fidelidad, seguridad, QA, gobernanza y reproducibilidad), remitiendo a los apéndices donde se encuentran los formatos completos.

Para ofrecer una visión integrada de extremo a extremo, la Figura 4.7 sintetiza el proceso de trazabilidad del estudio desde la ejecución presencial hasta la base analítica final. A diferencia de la Tabla 4.4, que indexa artefactos por componente, el diagrama permite ubicar en qué tramo del flujo interviene cada evidencia (captura, resguardo,

depuración, congelamiento y generación de outputs), y explicita los puntos de control que sostienen verificabilidad y reproducción del diagnóstico.

Figura 4.7

Mapa del proceso de trazabilidad del estudio



Nota. El mapa explicita el proceso de extremo a extremo y ubica, en cada tramo, la evidencia documental mínima que sustenta verificabilidad y reproducción.

4.3.3 Propósito, estructura y anclaje del análisis

El marco analítico se definió por el encadenamiento explícito entre la pregunta central del estudio y las hipótesis (H1–H4), asegurando trazabilidad directa hacia variables, instrumentos, criterios de calidad de medición y estimadores inferenciales. En coherencia con la asignación por conglomerados y el esquema pre–post, el parámetro inferencial primario se especifica como la interacción Tiempo×Grupo en modelos

multinivel, estimada sobre la muestra analítica definida para inferencia principal. Este parámetro se complementa con tamaños de efecto para diseños pre–post con control (d_{ppc2}) y con verificaciones de calidad de medición (fiabilidad interjueces e indicadores psicométricos cuando proceda), de acuerdo con estándares metodológicos reconocidos (Hu & Bentler, 1999; Cheung & Rensvold, 2002; Chen, 2007; Raudenbush & Bryk, 2002; Morris, 2008).

Con el propósito de evitar saltos lógicos entre diseño, medición e inferencia, la Matriz de congruencia (Tabla 4.5) se utiliza como un plano de navegación del capítulo. Su función no es descriptiva, sino estructural: hace visibles las dependencias problema, método, análisis y explicita las decisiones técnicas que sostienen la interpretabilidad del diagnóstico (por ejemplo, doble calificación ciega e ICC (2, k) para rúbricas; registro de implementación para análisis complementario de gradiente dosis–respuesta; y tratamiento de faltantes bajo supuestos MCAR/MAR con procedimientos consistentes con la gobernanza y la trazabilidad). De este modo, cada resultado reportado en las secciones subsecuentes se interpreta sobre métricas consistentes y reglas previamente justificadas, reduciendo el riesgo de inferencias post hoc o de decisiones analíticas no ancladas al diseño.

En complemento, el mapa metodológico y el plan analítico (Figura 4.8) ubican temporalmente las mediciones (PRE/POST), distinguen explícitamente los niveles de anidación (PRE/POST), distinguen explícitamente los niveles de anidación (mediciones repetidas en estudiantes; estudiantes en clases) y colocan, en cada tramo, los puntos de control de calidad que delimitan el alcance interpretativo de los resultados (psicometría y fiabilidad, fidelidad/dosis, seguridad, integridad y gobernanza). A diferencia de la matriz, que indexa elementos del diseño y de la inferencia, el diagrama proporciona una vista integrada del proceso de extremo a extremo, conectando insumos (rúbricas, registros de implementación, autoinformes post-exposición y evidencia cualitativa) con salidas analíticas (β de interacción, d_{ppc2} e indicadores de varianza explicada a nivel marginal/condicional), de forma compacta, auditable y pedagógicamente interpretable.

Tabla 4.5

Matriz de congruencia del estudio

Elemento	Contenido
Pregunta de investigación	¿En qué medida la incorporación instruccional de IA y RV produjo mejoras significativas y educacionalmente relevantes en comunicación y solución de problemas en estudiantes de pregrado del Tecnológico de Monterrey (2024–2025), frente a cursos sin dicha integración, controlando por línea base y efectos de clúster?
Objetivo general	Estimar el efecto de la implementación didáctica de IA y RV en comunicación y solución de problemas mediante mediciones pre–post, productos auténticos evaluados con rúbricas, registros de implementación (dosificación/fidelidad) y evidencia cualitativa.
Objetivos específicos	(1) Cuantificar cambios pre–post en comunicación; (2) Cuantificar cambios pre–post en solución de problemas; (3) Caracterizar la implementación (dosificación, fidelidad y seguridad) y explorar mecanismos percibidos mediante entrevistas/observación y autoinformes post-exposición.
Hipótesis	H1: IA+RV > comparación en ganancia pre–post de comunicación. H2: IA+RV > comparación en ganancia pre–post de solución de problemas. H3 (dosificación): mayor exposición/fidelidad (IA/RV) se asoció con mayor mejora. H4 (moderación): experiencia tecnológica y programa modularon los efectos.
Variables independientes (exposición/implementación)	Asignación por condición: intervención (IA+RV) vs. comparación. Dosificación IA: frecuencia/duración, tipo de <i>feedback</i> , eventos de retroalimentación. Dosificación RV: minutos de simulación, número de escenarios. Fidelidad: cumplimiento de <i>briefing–simulación–debriefing</i> , señalización, aplicación de rúbrica, etc. Seguridad: incidentes (p. ej., ciber-marco).
Variables dependientes principales (<i>outcomes</i> de desempeño)	Comunicación (rúbrica 0–100): claridad, coherencia/estructura, evidencia, fluidez, audiencia, componentes no verbales. Solución de problemas (rúbrica 0–100): definición, alternativas, evaluación, consecuencias, justificación, contextualización, principios/impacto social (subescala ética opcional).
Variables secundarias (proceso/mecanismo; no- <i>outcomes</i> longitudinales duros)	Autoinformes post-exposición (Likert 1–5): presencia percibida, utilidad/claridad del <i>feedback</i> , autoeficacia situacional, carga cognitiva, engagement. Propósito: caracterizar experiencia y aportar evidencia de mecanismo/aceptabilidad; se interpretan principalmente en POST y como triangulación (no como cambio PRE–POST cuando PRE se aplicó antes de la experiencia).
Indicadores	<i>Outcomes</i> principales: puntajes de rúbricas 0–100 (promedio de dos evaluadores; ICC interjueces con IC95%). Implementación: índice de dosificación (IA/RV), <i>checklists</i> de fidelidad (% adherencia), registro de incidentes. Proceso/mecanismo: promedios y perfiles POST de escalas Likert; evidencia cualitativa (temas).
Instrumentos	Rúbrica de Comunicación (6 criterios), Rúbrica de Solución de problemas/Decisión (7 criterios). Escalas de proceso/experiencia (presencia, utilidad de <i>feedback</i> , autoeficacia situacional, carga, engagement). Guía de entrevista semiestructurada. Listas de cotejo de fidelidad y bitácoras/incidentes.
Técnicas de obtención	Observación estructurada de productos auténticos (presentaciones/casos) videograbados; entrevistas a muestra teórica; registros de plataforma y <i>checklists</i> de sesión (dosificación/fidelidad); encuestas (expectativas en PRE cuando aplique; experiencia post-exposición en POST).
Plan de análisis	Modelos multinivel (medidas repetidas anidadas en estudiante y estudiantes en clase) con interacción Tiempo×Grupo como parámetro central; tamaños de efecto d_{ppc2} (Morris, 2008) e IC95%; análisis de dosificación/fidelidad (gradiente dosis–respuesta) y moderación; psicometría: α/ω y, cuando proceda, CFA/invariancia (Hu & Bentler, 1999; Cheung & Rensvold, 2002; Chen, 2007), explicitando que escalas experienciales aplicadas antes de la exposición no se interpretan como medidas longitudinales equivalentes; análisis temático (Braun & Clarke, 2006) y triangulación.

Nota. La matriz asegura la trazabilidad problema–método–análisis y fundamenta cuatro decisiones: (i) doble calificación ciega con ICC(2, k); (ii) registro de dosificación y fidelidad de IA/RV para análisis de gradiente dosis–respuesta; (iii) verificación psicométrica de instrumentos (piloto; α/ω ; y análisis factorial/CFA e invariancia únicamente cuando las mediciones sean comparables); y (iv) protocolo de faltantes bajo supuestos MCAR/MAR, con imputación múltiple cuando corresponda.

4.3.4 Criterios de preparación, estimación y robustez

El procesamiento de la información se organizó en ocho frentes secuenciales: (1) preparación y depuración (repositorio seguro, verificación de integridad, rangos,

consistencia y diccionario de variables); (2) tratamiento de datos faltantes, evaluando el mecanismo de ausencia (p. ej., prueba de Little para MCAR) y definiendo la estrategia según magnitud y patrón (Little, 1988; Enders, 2010); (3) evaluación de calidad de medición en escalas mediante consistencia interna (α y ω) y estructura factorial (CFA) y, cuando existió comparabilidad conceptual y temporal entre mediciones, pruebas de invariancia PRE/POST configural, métrica y escalar (Cronbach, 1951; McDonald, 1999; Hu & Bentler, 1999; Cheung & Rensvold, 2002; Chen, 2007; Brown, 2015); (4) fiabilidad de rúbricas mediante ICC(2,k) con IC95% bajo doble calificación ciega (Shrout & Fleiss, 1979; McGraw & Wong, 1996; Koo & Li, 2016); (5) construcción de puntajes (rúbricas: promedio de evaluadores por criterio y puntaje global 0–100; escalas: suma o promedio de ítems en rango 1–5 y estandarización cuando fue requerida por comparabilidad analítica); (6) estimación de modelos multinivel o mixtos (mediciones repetidas anidadas en estudiantes y estudiantes anidados en clases); (7) análisis de implementación (dosificación y fidelidad) y exploración de moderación; y (8) verificación de supuestos, sensibilidad y robustez mediante especificaciones alternativas.

Con el fin de evitar ambigüedad entre flujo de implementación y base inferencial, se definieron dos niveles de base de datos. En primer lugar, la base de flujo e implementación (N = 180) incluyó a todos los estudiantes que participaron en el trabajo de campo y para quienes existió evidencia operativa del proceso (calendarización, aplicación presencial, registros de sesión e incidencias), útil para describir organización, cobertura y ejecución. En segundo lugar, la base analítica per-protocol (n = 160) se definió como el subconjunto con evidencia mínima necesaria y comparable para la inferencia principal del diseño pre–post con estructura jerárquica: identificación válida y consistente, pertenencia a clúster, medición PRE y POST disponible para al menos un *outcome* principal, y calificación válida de rúbricas conforme al protocolo de doble evaluación. Los casos que no cumplieron estos criterios (por ejemplo, no completar la secuencia de intervención y medición, ausencia de un punto temporal clave, o evidencia incompleta no recuperable bajo reglas de reposición documentadas) se conservaron para el reporte del flujo (N = 180) y se excluyeron únicamente de la inferencia principal, conforme a las reglas de integridad y trazabilidad establecidas (Apéndice A4.1 y A4.3; Apéndice B).

En escalas, se definió como criterio operativo de aceptabilidad psicométrica α y $\omega \geq .70$, interpretándolo como consistencia interna adecuada para fines de investigación aplicada. Cuando los coeficientes se ubicaron por debajo de dicho umbral, los puntajes se interpretaron con cautela y se restringieron a su rol de evidencia de proceso o mecanismo, evitando sobregeneralizaciones como *outcomes* longitudinales equivalentes (Cronbach, 1951; McDonald, 1999). La validez estructural se examinó mediante CFA reportando índices globales de ajuste (CFI, TLI, RMSEA, SRMR) y, cuando procedió por comparabilidad longitudinal, se evaluó invariancia utilizando criterios operativos ampliamente empleados (p. ej., $\Delta CFI \leq .01$) para sostener interpretaciones de cambio sin contaminación por variación del instrumento (Hu & Bentler, 1999; Cheung & Rensvold, 2002; Chen, 2007).

La fiabilidad de las rúbricas se estimó mediante ICC(2,k) con IC95% bajo doble calificación ciega, y se estableció un procedimiento de resolución de discrepancias conforme al protocolo (anclajes, revisión dirigida y decisión documentada), de modo que los puntajes finales reflejaran consistencia interjueces y trazabilidad en los criterios evaluados (Shrout & Fleiss, 1979; McGraw & Wong, 1996; Koo & Li, 2016). La construcción de puntajes en rúbricas se realizó como promedio de jueces por criterio y puntaje global en escala 0–100, mientras que en escalas se utilizó suma o promedio de ítems en rango 1–5, con estandarización únicamente cuando se requirió para comparabilidad analítica entre medidas o modelos.

Para fortalecer la auditabilidad y la reproducibilidad, el *pipeline* analítico se versionó con *scripts*, semillas fijadas y bitácoras de ejecución; asimismo, se documentaron decisiones analíticas y se compararon resultados bajo escenarios de robustez (base analítica per-protocol, imputación múltiple, modelos con y sin covariables, especificaciones alternativas y reglas de agregación de rúbricas), siguiendo prácticas recomendadas de investigación reproducible (Peng, 2011; Wilson *et al.*, 2014). En el manejo de datos faltantes, la decisión se aplicó dentro de la base analítica ($n=160$): se utilizó lista completa cuando la proporción fue mínima y el patrón no sugirió sesgo; cuando la magnitud y el patrón lo justificaron bajo supuestos MAR plausibles, se implementó imputación múltiple ($m = 20$) incorporando predictores auxiliares (línea

base, programa e indicadores de implementación) para reducir sesgo y preservar potencia (Rubin, 1987; Enders, 2010; White *et al.*, 2011; Graham *et al.*, 2007). Finalmente, la comunicación de resultados se alineó con recomendaciones de reporte con énfasis en tamaños de efecto e intervalos de confianza, complementando la inferencia con métricas de varianza explicada en modelos mixtos (Wilkinson & APA Task Force on Statistical Inference, 1999; Cumming, 2014; Nakagawa & Schielzeth, 2013; Johnson, 2014).

4.3.5 Especificación de modelo

Para cada variable de desempeño (Y = Comunicación, Solución de problemas) se estimó un modelo lineal mixto para medidas repetidas (LMM), con dos mediciones (PRE y POST) anidadas en estudiantes y estudiantes anidados en clases (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; West *et al.*, 1996). La especificación se formuló como un modelo jerárquico con interceptos aleatorios a nivel estudiante y clase:

Nivel 1 (medición t del estudiante i en la clase k):

$$Y_{tik} = \beta_{0ik} + \beta_1 Time_t + \beta_3 (Time_t \times Group_k) + \varepsilon_{tik}$$

Nivel 2 (estudiante):

$$\beta_{0ik} = \gamma_{00k} + \gamma_{01} Baseline_{ik} + \gamma_{02} X_{ik} + u_{0ik}$$

Nivel 3 (clase):

$$\gamma_{00k} = \delta_{000} + \delta_{001} Group_k + v_{00k}$$

donde $Time_t$ indica el momento de medición (0 = PRE, 1 = POST), $Group_k$ identifica la condición asignada a nivel clase (IA+RV vs. comparación), $Baseline_{ik}$ corresponde al puntaje basal del *outcome* del estudiante (medido en PRE), y X_{ik} representa covariables individuales (experiencia tecnológica previa y programa/semestre). Se incluyó el puntaje baseline (PRE) como covariable a nivel estudiante para incrementar precisión y reducir posibles desequilibrios basales entre conglomerados; en formato largo, esta parametrización se implementa como una variable constante por estudiante y *outcome*, distinta del indicador Time, siguiendo el espíritu ANCOVA dentro de un marco mixto. Los términos u_{0ik} y v_{00k} representan

interceptos aleatorios a nivel estudiante y clase, respectivamente, y ε_{tik} es el error a nivel de medición.

El parámetro central de inferencia fue β_3 , correspondiente a la interacción Tiempo $\times$ Grupo, que operacionaliza el cambio diferencial PRE–POST entre condiciones (contraste tipo diferencia-en-diferencias) dentro del marco multinivel. El modelo incorporó covariables (línea base del *outcome*, experiencia tecnológica y programa/semestre) para mejorar precisión y reducir confusión residual, en línea con recomendaciones para ensayos por conglomerados y análisis longitudinales (Donner & Klar, 2000; Fitzmaurice et al., 2011). β_3

Se reportaron intervalos de confianza al 95% utilizando aproximaciones de Kenward–Roger o Satterthwaite para los grados de libertad, según correspondió (Kenward & Roger, 1997; Satterthwaite, 1946; Luke, 2017). Adicionalmente, se reportaron R^2 marginal y condicional para cuantificar la varianza explicada por efectos fijos y por el modelo completo (fijos + aleatorios), respectivamente (Nakagawa & Schielzeth, 2013). Para facilitar la interpretación comparativa de magnitud, se estimaron tamaños de efecto para diseños PRE–POST con grupo de comparación mediante d_{ppc2} (diferencia-en-diferencias), con corrección tipo Hedges, reportando su intervalo de confianza al 95% (Morris, 2008; Hedges, 1981).

Las estimaciones inferenciales principales (H1–H2) se realizaron sobre la base analítica per-protocol ($n = 160$), definida por contar con evidencia comparable PRE y POST para al menos un *outcome* principal y con calificaciones válidas de rúbrica conforme al protocolo de doble evaluación. Los análisis complementarios de implementación (H3–H4) exploraron gradientes de dosificación y fidelidad, así como posibles moderaciones mediante términos de interacción con experiencia tecnológica previa y programa/semestre. Estos análisis se estimaron sobre la muestra ampliada con información de implementación disponible (hasta $N = 180$), con el propósito de maximizar variación en dosificación, manteniendo su interpretación como evidencia de mecanismo y heterogeneidad, no como sustituto de la inferencia primaria del efecto (Fitzmaurice et al., 2011).

El componente cualitativo se abordó mediante análisis temático (Braun & Clarke, 2006), con doble codificación parcial del corpus y triangulación con patrones cuantitativos para evaluar convergencias, complementariedades y discrepancias explicativas. Finalmente, se realizaron verificaciones de supuestos y análisis de sensibilidad (especificaciones alternativas, inclusión o exclusión de covariables y comparación entre lista completa vs imputación múltiple cuando aplicó), documentando cualquier desviación del plan analítico para preservar trazabilidad y auditabilidad (Rubin, 1987; White *et al.*, 2011).

En conjunto, el plan analítico (Figura 4.8) combinó modelos lineales mixtos con el contraste Tiempo $\times$ Grupo, tamaños de efecto d_{ppc2} (IC95%) y análisis de implementación (dosis–respuesta y moderación).

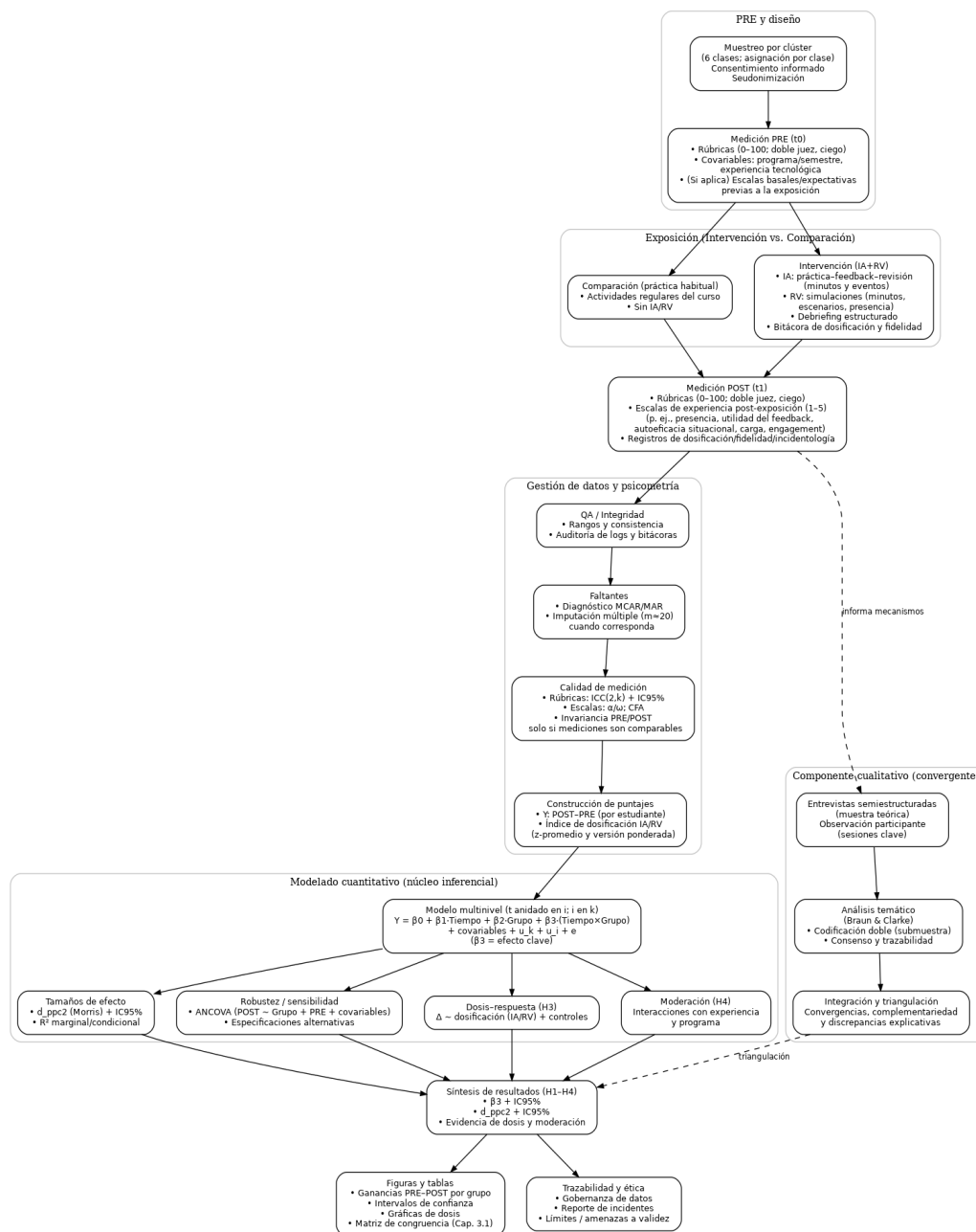
La utilidad de este mapa metodológico radica en que permite mostrar, de manera integrada, cómo las decisiones de modelación, estimación de efectos, análisis de implementación y triangulación cualitativa se articularon dentro de una misma estrategia analítica. Esta representación es particularmente relevante porque el estudio no se apoyó en un único contraste estadístico, sino en una cadena de evidencias complementarias orientadas a responder la pregunta central con prudencia inferencial. En este sentido, la figura ayuda a diferenciar la inferencia principal, basada en el contraste Tiempo $\times$ Grupo y en la muestra analítica per-protocol, de los análisis secundarios de dosificación, fidelidad y moderación, cuya función fue ampliar la interpretación de los resultados sin sustituir el contraste primario.

Asimismo, el mapa metodológico permite hacer explícita la separación entre evidencia confirmatoria, evidencia explicativa y evidencia de apoyo operativo. Esta distinción resulta importante porque evita sobredimensionar los análisis complementarios y mantiene la jerarquía inferencial del estudio: primero se estima el efecto diferencial de la intervención sobre las competencias evaluadas; después se examinan patrones de implementación que ayudan a comprender la magnitud, consistencia o variación de dichos efectos; y finalmente se integra la evidencia cualitativa para contextualizar los mecanismos observados. De este modo, la figura

contribuye a transparentar la lógica de análisis y a mostrar que cada fuente de evidencia cumple una función específica dentro del argumento empírico general.

Figura 4.8

Mapa metodológico y plan analítico



Nota. El diagrama sintetiza la estructura del modelo multinivel con mediciones PRE y POST, estudiantes anidados en clases ($k = 6$) y el parámetro central β_3 (Tiempo $\times$ Grupo), junto con el reporte de R^2 marginal/condicional y tamaños de efecto d_{ppc2} .

Con el diseño, los instrumentos y el plan analítico establecidos, los apartados siguientes presentan primero la integridad muestral y las verificaciones previas, posteriormente los efectos principales (H1–H2) mediante modelos multinivel y tamaños de efecto, y finalmente los análisis complementarios orientados a implementación y mecanismo (H3–H4), junto con pruebas de robustez y sensibilidad.

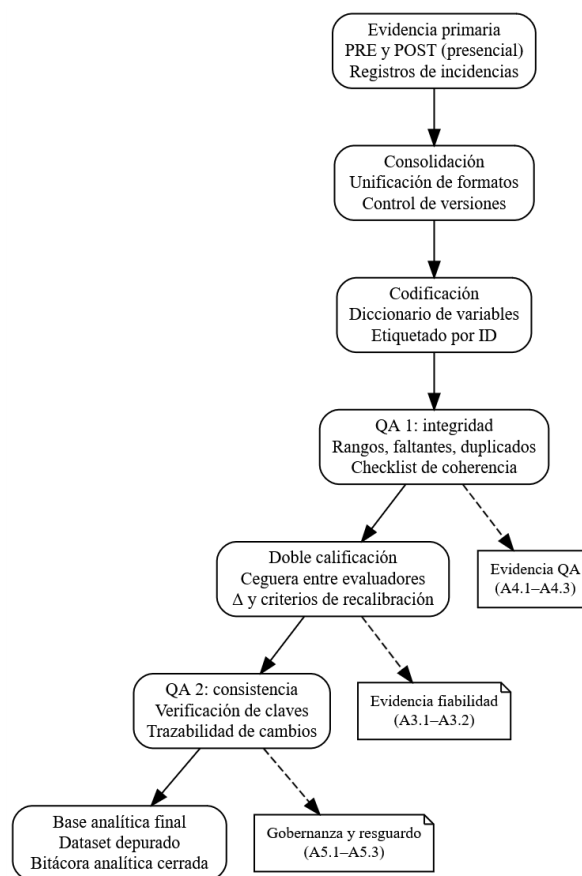
4.3.6 Base diagnóstica para la interpretación de resultados

La redacción del diagnóstico se construyó en dos capas complementarias. En la primera, descriptiva, se reportan regularidades empíricas del estado inicial mediante estadísticos de tendencia central, dispersión y distribución, junto con evidencia gráfica que permite identificar patrones por criterios y variabilidad entre participantes. En la segunda, analítica, se interpretan dichas regularidades con prudencia, articulando la evidencia cuantitativa con la trazabilidad de la implementación presencial, sin atribuir causalidad fuera del alcance del diseño y del control operacional documentado.

En consecuencia, el diagnóstico no se limita a describir resultados. También explicita las condiciones de confiabilidad bajo las cuales esos resultados deben leerse: consistencia de calificación con doble ceguera y verificación de ICC (Apéndice A3), integridad y coherencia de datos (Apéndice A4), y reglas de gobernanza y resguardo que sostienen la auditabilidad del flujo (Apéndice A5). Esta lógica fortalece la coherencia interna del capítulo, porque justifica por qué los contrastes posteriores pueden interpretarse sobre una base comparativa sólida y verificable. En este sentido, la Figura 4.9 sintetiza la ruta de transformación de la evidencia primaria hacia la base analítica final, haciendo visibles los controles de calidad, consistencia y trazabilidad que sostienen la verificabilidad del diagnóstico.

Figura 4.9

Pipeline de datos del diagnóstico: de evidencia primaria a base analítica

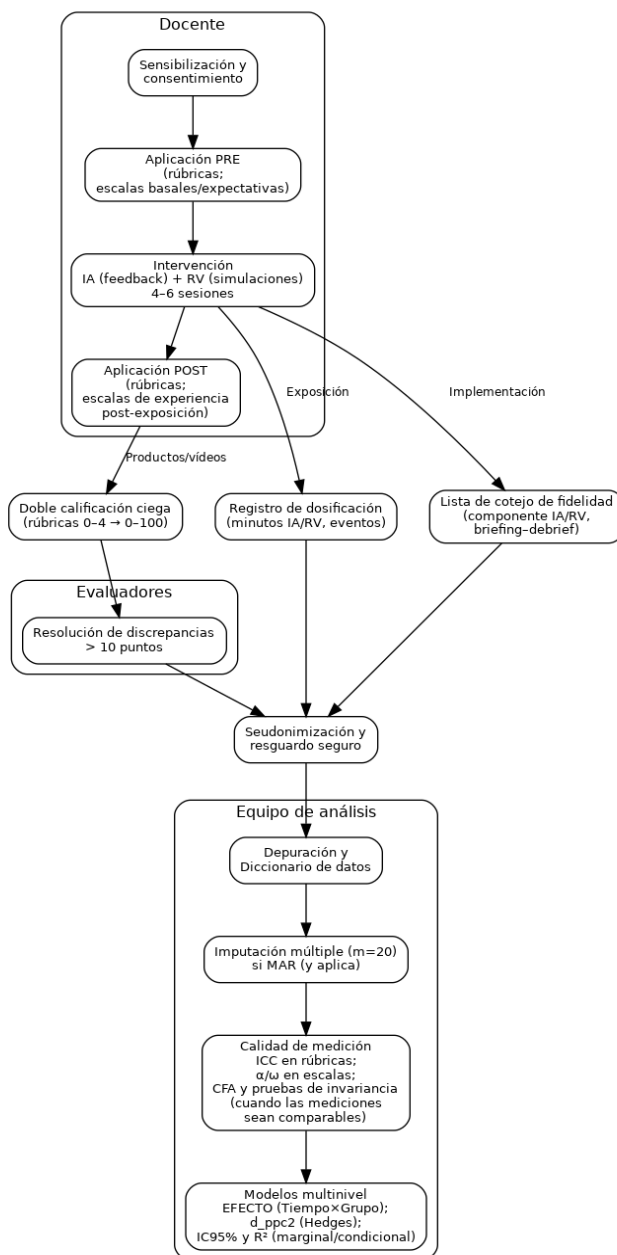


Nota. El diagrama muestra la secuencia de transformación de datos y los controles aplicados (rangos, faltantes, consistencia, reglas de exclusión y bitácora) para asegurar verificabilidad.

Esta subsección consolidó así el diagnóstico empírico del estado inicial del problema y su verificabilidad, documentando flujo de muestra, equivalencia basal y calidad de medición como condiciones necesarias para interpretar regularidades y cambios posteriores. De manera complementaria, la Figura 4.10 amplía esta base diagnóstica al representar la secuencia integral del trabajo de campo y del análisis, incorporando responsabilidades, insumos y puntos de control desde la medición PRE hasta la estimación final. Mientras la Figura 4.13 enfatiza la trazabilidad de los datos, la Figura 4.10 permite visualizar la articulación metodológica completa entre implementación, medición, aseguramiento de calidad y análisis inferencial.

Figura 4.10

Diagrama del trabajo de campo y del análisis: responsabilidades, insumos y puntos de control



Nota. La figura sintetiza la secuencia PRE, intervención, POST y análisis, con controles de calidad: registro de dosificación y fidelidad; doble evaluación ciega y resolución de discrepancias; seudonimización; psicometría (α , ω , ICC); análisis factorial/CFA y pruebas de invariancia cuando exista comparabilidad conceptual y temporal; e imputación múltiple cuando MAR; culminando en el análisis multinivel y el tamaño de efecto para diseños pre-post con control (Little, 1988; Rubin, 1987; Hu & Bentler, 1999; Cheung & Rensvold, 2002; Chen, 2007; Shrout & Fleiss, 1979; Koo & Li, 2016; Raudenbush & Bryk, 2002; Morris, 2008).

4.3.7 Descriptivos, flujo de muestra y verificaciones previas

La cohorte inicial estuvo conformada por 180 estudiantes distribuidos en seis clases, con asignación por conglomerados (a nivel de clase) a dos condiciones: tres clases a la intervención IA+RV y tres a la condición de comparación. Conforme al criterio preespecificado de asistencia mínima ($\geq 75\%$) definido para asegurar una exposición mínima a la intervención y preservar consistencia inferencial en los contrastes principales se excluyeron 20 estudiantes por ausentismo sistemático. En consecuencia, el análisis principal de resultados de desempeño se reporta sobre la muestra analítica $n = 160$, integrada por 79 en el grupo de comparación y 81 en el grupo de intervención. Para evitar ambigüedad entre cohorte completa y muestra analítica, este capítulo distingue dos universos: (i) inferencias principales sobre desempeño rubricado (Comunicación y Solución de problemas) estimadas sobre la muestra per-protocol ($n = 160$) definida por asistencia mínima y evidencia PRE y POST comparable; y (ii) evidencia complementaria de implementación y mecanismo (dosificación, fidelidad y autoinformes post-exposición) reportada sobre la cohorte con información operativa disponible (hasta $N = 180$). Esta distinción se mantiene de forma consistente en tablas, figuras y modelos.

Adicionalmente, antes de estimar los efectos PRE–POST, se verificó la comparabilidad basal entre condiciones en variables académicas y demográficas disponibles, así como en los puntajes PRE de Comunicación y Solución de problemas, con el fin de descartar desequilibrios sustantivos atribuibles a la asignación por conglomerados. Estas verificaciones no sustituyen el ajuste multinivel, pero permiten contextualizar la plausibilidad de las inferencias y anticipar la necesidad de covariables si se observaran diferencias iniciales relevantes. En paralelo, se inspeccionaron distribuciones, rangos válidos y consistencia de codificación (tiempo, grupo, clase) para asegurar integridad de la base analítica y trazabilidad entre registros y puntajes rubricados. Dado que la (exposición mínima $\geq 75\%$) para preservar interpretabilidad del contraste, mientras que la lógica intervención requiere exposición efectiva, el análisis principal se definió como *per-protocol intention-to-treat* se reconoce como referencia en ensayos, pero menos informativa aquí cuando el tratamiento no puede considerarse recibido sin asistencia suficiente. Finalmente, se documentaron las razones de exclusión

(asistencia <75%) como criterio operativo de exposición mínima, evitando dilución del efecto por no cumplimiento y sosteniendo consistencia inferencial en los contrastes principales.

4.3.8 Aseguramiento de la calidad de los instrumentos

Antes de interpretar los resultados, fue necesario verificar que los instrumentos utilizados ofrecieran evidencia suficiente de calidad de medición. Para ello, se implementó un proceso de aseguramiento en cuatro etapas concatenadas: ajuste fino de anclajes y descriptores con apoyo de juicio experto, pilotaje para revisar comprensión y logística de aplicación, calibración de evaluadores mediante productos ancla y revisión psicométrica de las escalas de autoinforme. Este proceso tuvo como propósito asegurar trazabilidad entre constructo, indicador y regla de puntuación, reducir ambigüedades interpretativas y fortalecer la comparabilidad PRE–POST, en coherencia con estándares contemporáneos de validez basada en evidencia (AERA/APA/NCME, 2014).

En el caso de las rúbricas de desempeño, la calidad de medición se sustentó en doble evaluación ciega, calibración previa y estimación de fiabilidad interjueces mediante ICC(2,2). En el caso de las escalas experienciales, se estimaron coeficientes de consistencia interna (α y ω) y ajuste factorial confirmatorio (CFA) en POST inmediato, dado su carácter situacional y su función principal como evidencia de proceso, mecanismo y aceptabilidad.

Esta verificación fue especialmente relevante porque los resultados principales del estudio dependen de puntajes contruados a partir de desempeño evaluado y no únicamente de respuestas autoinformadas. En consecuencia, la calidad de los instrumentos condiciona directamente la solidez de las inferencias sobre cambio, diferencia entre grupos y magnitud del efecto.

Las Tablas 4.6A y 4.6B resumen los índices obtenidos para ambos tipos de instrumentos y constituyen la base técnica que respalda la interpretación posterior de los resultados.

Tabla 4.6A*Índices de calidad de medición, rúbricas de desempeño (doble calificación ciega)*

Instrumento	Tipo	k (criterios)	n / Obs. (PRE+POST)	ICC(2,2)	IC95 % ICC	Rango ICC por criterio
Rúbrica de Comunicación	Desempeño (doble calificación ciega)	6	320	0.903	[0.882, 0.920]	0.802–0.849
Rúbrica de Solución de problemas/decisión	Desempeño (doble calificación ciega)	7	320	0.921	[0.907, 0.934]	0.813–0.855

Nota. Se reporta ICC (2,2) (dos vías, acuerdo absoluto) como estimador principal, dado que el puntaje operativo corresponde al promedio de dos evaluadores.

Tabla 4.6B*Índices psicométricos, escalas de autoinforme experiencial (POST)*

Instrumento	Tipo	k (ítems)	n	α (POST)	ω (POST)	CFI (POST)	TLI (POST)	RMSEA (POST)	SRMR (POST)
Presencia percibida	Autoinforme experiencial (Likert 1–5)	8	160	0.737	0.813	0.956	0.939	0.049	0.057
Utilidad/claridad del <i>feedback</i>	Autoinforme experiencial (Likert 1–5)	6	160	0.664	0.780	1.000	1.000	0.000	0.043
Autoeficacia situacional	Autoinforme experiencial (Likert 1–5)	6	160	0.676	0.787	1.000	1.000	0.000	0.034
Carga cognitiva	Autoinforme experiencial (Likert 1–5)	6	160	0.672	0.786	1.000	1.000	0.000	0.039
Engagement	Autoinforme experiencial (Likert 1–5)	6	160	0.597	0.749	1.000	1.000	0.000	0.029

Nota. Las escalas se interpretan como evidencia de proceso/mecanismo y aceptabilidad y se estiman en POST, inmediatamente tras la exposición, para maximizar validez ecológica del reporte. Los índices de ajuste corresponden al CFA en POST ($CFI/TLI \geq .90$; $RMSEA/SRMR \leq .08$ como criterios de referencia).

4.3.9 Calidad de medición

La interpretación de los efectos se sustenta en la calidad de medición de los instrumentos. En primer lugar, las rúbricas de desempeño presentan fiabilidad interevaluador buena tanto a nivel global (Tabla 4.6A) como por criterio (Tabla 4.7), con coeficientes ICC (2,2) en el rango 0.808–0.850 para Comunicación y 0.811–0.856 para Solución de problemas/decisión. El diseño de confiabilidad empleó doble calificación

independiente y ciega de los artefactos PRE y POST; previo a la calificación definitiva se realizó una sesión de calibración con artefactos piloto para alinear la interpretación de criterios y niveles de desempeño. Dado que el puntaje operativo se basó en el promedio de dos evaluadores, se estimó el ICC (2,2) bajo un modelo de dos vías de efectos aleatorios y criterio de acuerdo absoluto, consistente con la formulación clásica del ICC para generalizar a evaluadores comparables (Shrout & Fleiss, 1979; McGraw & Wong, 1996). Cuando se observaron discrepancias sustantivas, se aplicó un protocolo de recalibración y, de ser necesario, un tercer dictamen (Apéndice A7.2). En términos interpretativos, el rango observado se ubica en niveles buenos ($\geq .75$) para estudios aplicados, reduciendo la probabilidad de error de calificación y fortaleciendo la inferencia sobre cambios reales en desempeño (Koo & Li, 2016).

En segundo lugar, los índices de validez de contenido fueron consistentes con una alineación completa entre criterios y constructo: a nivel de ítem, los I-CVI = 1.00 se mantuvieron de forma uniforme, y a nivel de dimensión se obtuvieron S-CVI/Ave = 1.00 en Claridad, Coherencia y Relevancia para ambos instrumentos (Tabla 4.8). Esta evidencia es coherente con los procedimientos estándar de cuantificación de validez de contenido mediante CVI a nivel de ítem y promedio por dimensión/escala (Lynn, 1986; Polit & Beck, 2006).

Finalmente, las escalas de autoinforme (empleadas como evidencia de proceso/mecanismo) se interpretan principalmente en POST, donde la consistencia interna se ubicó en el rango $\alpha = 0.617\text{--}0.738$ ($n = 160$; Tabla 4.6B). Dado su carácter situacional y su brevedad, se espera que α sea moderada en comparación con instrumentos extensos; por ello, estas escalas se utilizan como soporte explicativo complementario y no como *outcomes* longitudinales equivalentes al desempeño rubricado, considerando además que α depende del número de ítems y de supuestos psicométricos específicos (Cronbach, 1951; Tavakol & Dennick, 2011).

Tabla 4.7*Fiabilidad interevaluador por criterio (ICC [2,2]), doble calificación ciega*

Instrumento	Criterio	ICC (2,2)	IC95%
Rúbrica de Comunicación	Adecuación a la audiencia	0.823	[0.78, 0.86]
Rúbrica de Comunicación	Claridad	0.810	[0.76, 0.85]
Rúbrica de Comunicación	Coherencia/estructura	0.808	[0.76, 0.85]
Rúbrica de Comunicación	Uso de evidencia	0.823	[0.78, 0.86]
Rúbrica de Comunicación	Fluidez / <i>listenability</i>	0.847	[0.81, 0.88]
Rúbrica de Comunicación	Componentes no verbales	0.850	[0.81, 0.88]
Rúbrica de Solución de problemas/decisión	Definición del problema	0.827	[0.78, 0.86]
Rúbrica de Solución de problemas/decisión	Generación de alternativas	0.830	[0.79, 0.87]
Rúbrica de Solución de problemas/decisión	Evaluación de alternativas	0.811	[0.76, 0.85]
Rúbrica de Solución de problemas/decisión	Consideración de consecuencias	0.827	[0.78, 0.86]
Rúbrica de Solución de problemas/decisión	Justificación	0.844	[0.80, 0.88]
Rúbrica de Solución de problemas/decisión	Contextualización	0.830	[0.79, 0.87]
Rúbrica de Solución de problemas/decisión	Principios éticos/impacto social	0.856	[0.82, 0.89]

Nota. Se reporta ICC (2,2) (promedio de dos evaluadores). IC95% = intervalo de confianza al 95%.

Tabla 4.8*Validez de contenido (S-CVI/Ave) por dimensión (acuerdo absoluto)*

Instrumento	Dimensión	n ítems	S-CVI/Ave
Rúbrica de Comunicación	Claridad	2	1.00
Rúbrica de Comunicación	Coherencia	2	1.00
Rúbrica de Comunicación	Relevancia	2	1.00
Rúbrica de Solución de problemas/decisión	Claridad	3	1.00
Rúbrica de Solución de problemas/decisión	Coherencia	2	1.00
Rúbrica de Solución de problemas/decisión	Relevancia	2	1.00

Nota. S-CVI/Ave = promedio del I-CVI dentro de cada dimensión. A nivel de ítem, los I-CVI = 1.00 fueron consistentes en ambos instrumentos (Lynn, 1986; Polit & Beck, 2006).

4.3.10 Pruebas de hipótesis y estimaciones principales

4.3.10.1 Modelos multinivel (efecto *Tiempo* × *Grupo*)

Dada la estructura jerárquica del estudio, mediciones repetidas (PRE/POST) anidadas en estudiantes y estudiantes anidados en clases, con asignación por conglomerados, se estimaron modelos lineales mixtos (modelos multinivel) para cada *outcome* de desempeño (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; West *et*

al., 1996). Esta especificación es adecuada por dos razones. Primero, respeta la dependencia estadística inducida tanto por el aula (efectos compartidos por clase) como por la repetición intraestudiante, evitando la subestimación de errores estándar y la inflación de significancia que ocurriría si las observaciones se trataran como independientes (Cameron & Miller, 2015; Donner & Klar, 2000; Murray, 1998). Segundo, permite estimar el efecto de la intervención como un contraste de cambio consistente con el diseño PRE–POST con comparación, integrando simultáneamente la variación intra e inter-sujeto y la estructura por conglomerados (Fitzmaurice *et al.*, 2011; Twisk, 2013). Adicionalmente, los modelos mixtos admiten heterogeneidad residual entre estudiantes y clases mediante efectos aleatorios, lo que mejora el ajuste y la validez de la inferencia en datos educacionales. Asimismo, esta estrategia facilita incorporar covariables o ajustes de línea base cuando es pertinente, sin romper la lógica de medidas repetidas del diseño.

El parámetro central de inferencia fue la interacción Tiempo $\times$ Grupo, que operacionaliza el efecto diferencial PRE–POST atribuible a la arquitectura IA+RV frente a la condición de comparación. En términos sustantivos, este coeficiente cuantifica cuánto se incrementa (o no) la ganancia en el grupo IA+RV por encima de la ganancia observada en el grupo comparación, ajustando por la anidación del diseño. Esta interacción se interpreta de forma natural como una diferencia-en-diferencias (cambio PRE–POST en intervención menos cambio PRE–POST en comparación), enfoque ampliamente utilizado para identificar efectos netos cuando existen medidas pre y post bajo condiciones comparativas (Angrist & Pischke, 2009; Wooldridge, 2010). Los coeficientes estimados muestran un efecto diferencial robusto y consistente en ambos *outcomes* (Tabla 4.9A). En consecuencia, la conclusión no descansa en diferencias simples de medias post, sino en el contraste de trayectorias entre grupos a través del tiempo, coherente con la lógica causal del diseño.

Tabla 4.9A

Efecto diferencial del modelo multinivel: interacción Tiempo $\times$ Grupo

Outcome	β (Tiempo $\times$ Grupo)	IC95%
Comunicación	22.122	[18.448, 25.796]
Solución de problemas	23.544	[19.880, 27.208]

Nota. β (Tiempo $\times$ Grupo) cuantifica la diferencia en la ganancia PRE–POST entre condiciones (intervención vs. comparación), ajustada por la estructura de anidación del diseño. Un intervalo de confianza que excluye 0 indica evidencia de un cambio diferencial a favor de la intervención.

Para complementar la interpretación inferencial con una métrica de magnitud comparable entre *outcomes*, se reportan a continuación tamaños de efecto d_{ppc2} con IC95%, recomendados para diseños PRE–POST con grupo de comparación al expresar el efecto neto como diferencia-en-diferencias estandarizada (Morris, 2008).

4.3.10.2 Tamaños de efecto (diferencia-en-diferencias)

Para complementar la inferencia del modelo multinivel con una métrica de magnitud fácilmente interpretable en términos educativos, se estimaron tamaños de efecto para diseños PRE–POST con grupo de comparación mediante d_{ppc2} . Este estimador es apropiado en estudios con medición repetida porque expresa el efecto neto de la intervención como una diferencia-en-diferencias, es decir, el cambio PRE–POST en la condición IA+RV menos el cambio PRE–POST en el grupo comparación, estandarizado en unidades de desviación estándar (Morris, 2008). En consecuencia, d_{ppc2} permite: (i) comparar magnitudes entre *outcomes* medidos en la misma escala, (ii) comunicar relevancia práctica más allá del valor p y (iii) contextualizar la precisión del estimador mediante intervalos de confianza al 95% (IC95%), recomendados como base para una interpretación sustantiva de tamaños de efecto (Lakens, 2013) (Tabla 4.9B).

En este estudio, d_{ppc2} se estimó como una diferencia-en-diferencias estandarizada usando como denominador la desviación estándar PRE combinada (*pooled*) de ambos grupos para cada *outcome*, calculada como: $d_{ppc2} = [(M_POST, IA+RV - M_PRE, IA+RV) - (M_POST, comparación -$

$M_{PRE,comparación}] / SD_{PRE,pooled}$. Esta definición ancla la estandarización en la dispersión inicial del desempeño y evita que el denominador dependa de la variabilidad POST potencialmente influida por la intervención, manteniendo coherencia con la familia de estimadores discutida por Morris (2008). El IC95% se obtuvo mediante aproximación normal (Wald) a partir del error estándar del estimador, $IC95\% = d_{ppc2} \pm 1.96 \times SE$, y el cálculo completo es reproducible en el *pipeline* del Apéndice

$$B.SD_{PRE} = \sqrt{(Var_{PRE,IA+RV} + Var_{PRE,comparación})/2}$$

Además, al basarse en el contraste de cambios entre grupos, d_{ppc2} es consistente con la lógica del parámetro Tiempo $\times$ Grupo del modelo mixto y facilita la triangulación entre evidencia estandarizada y evidencia en puntos (escala 0–100). Por ello, se reporta como complemento técnico para comparabilidad, sin sustituir la lectura sustantiva del cambio en la escala original de la rúbrica.

Tabla 4.9B

Tamaños de efecto para diseños PRE–POST con comparación (d_{ppc2}) d_{ppc2}

<i>Outcome</i>	d_{ppc2}	IC95%
Comunicación	2.598	[2.20, 3.00]
Solución de problemas	2.466	[2.08, 2.85]

Nota: d_{ppc2} se estimó como diferencia-en-diferencias estandarizada con la DE basal combinada, siguiendo a Morris (2008), y su IC95% se obtuvo mediante aproximación normal tipo Wald, conforme a la implementación reproducible del Apéndice B. La interpretación se realizó junto con el estimador multinivel Tiempo $\times$ Grupo, el efecto en escala original y la evidencia distributiva.

A la luz de la Tabla 4.9B, la intervención IA+RV muestra un cambio diferencial PRE–POST de magnitud elevada. Dado que, en diseños con grupo de comparación y rúbricas analíticas, la estandarización puede magnificar el tamaño de efecto cuando la dispersión basal es moderada en relación con un desplazamiento sustantivo en puntos, la lectura se plantea deliberadamente en tres capas: el estimador estandarizado, el efecto inferencial del modelo multinivel (β de Tiempo $\times$ Grupo) y el patrón descriptivo en la escala de la rúbrica, que ofrece una interpretación educacional directa (Morris, 2008; Lakens, 2013; Cumming, 2014). Así, d_{ppc2} opera como indicador técnico complementario, mientras que la plausibilidad del cambio se fundamenta en criterios explícitos, práctica deliberada y protocolos de implementación y fidelidad que reducen

variación no deseada y hacen el progreso verificable (Bellg *et al.*, 2004; Carroll *et al.*, 2007). En coherencia con este enfoque, la evidencia descriptiva en la escala 0–100 permite evaluar si el efecto estandarizado pudiera estar influido por acumulación en valores altos o por compresión de la dispersión. En particular, la Tabla 4.9D y los *boxplots* PRE–POST (Figuras 4.12C–4.12D) muestran un desplazamiento claro del POST en la condición IA+RV con variabilidad preservada y proporciones mínimas de puntajes en techo ($\geq 90/100$), lo que reduce la plausibilidad de un artefacto por *ceiling effects* y ancla la magnitud en puntos de desempeño. Bajo este marco, la Tabla 4.9C organiza las amenazas a la validez interna y sus mitigaciones para restringir explicaciones alternativas.

Tabla 4.9C

Amenazas a la validez y mitigaciones en un cuasi-experimento por conglomerados

Amenaza	Riesgo para la inferencia	Mitigación aplicada en esta tesis	Limitación residual / implicación
No independencia por conglomerados (clases/grupos)	Inferencias infladas si se ignora la anidación	Modelos multinivel para modelar dependencia intraclase y estimar el efecto diferencial PRE–POST con Tiempo $\times$ Grupo	Con pocos conglomerados, la heterogeneidad contextual y moderaciones puede estimarse con menor precisión
Efecto de novedad (<i>novelty effect</i>)	Mejora transitoria por motivación/agrado, no por aprendizaje estable	Se privilegian <i>outcomes</i> de desempeño criterial con rúbricas; consistencia interjueces elevada; criterios anclados en descriptores conductuales (la novedad se reconoce como habilitante de compromiso, no como explicación suficiente del desempeño)	No puede descartarse completamente; se declara como amenaza plausible y se recomienda seguimiento longitudinal/transferecia
Inestabilidad/limitaciones de autoinforme	Cambios percibidos pueden reflejar sesgos o falta de invariancia	Tratamiento conservador: autoinformes se interpretan como evidencia de proceso/aceptabilidad cuando la medición basal es inestable; las inferencias fuertes se reservan para medidas con soporte psicométrico	Se acota el alcance de conclusiones sobre cambio perceptual; se sugiere reforzar invariancia y estabilidad en futuros estudios
Dosis–respuesta no detectable	Riesgo de concluir erróneamente que más uso siempre implica más efecto o que el marco teórico falla	Interpretación acotada: dentro del rango observado, el efecto opera como cambio de arquitectura (participar/no participar) más que como función lineal de intensidad marginal	Agenda futura: diseños con intensidades deliberadamente diferenciadas y medición más fina de calidad de implementación (<i>debriefing</i> , especificidad del <i>feedback</i> , cumplimiento de andamiajes)
Contaminación / variación docente / Hawthorne	Diferencias no observadas entre grupos o influencia por estar siendo estudiados	Reconocimiento explícito como amenazas típicas de estudios de campo; triangulación del hallazgo en dos <i>outcomes</i> con dirección consistente y magnitudes sustantivas, reduciendo plausibilidad de artefacto único	Persisten como límites propios del trabajo en aula; se explicita como restricción de validez externa/causalidad estricta y se proponen controles más finos en futuras réplicas

Nota. Amenazas y estrategias de mitigación típicas en estudios cuasi-experimentales y ensayos por conglomerados; ver lineamientos y fundamentos metodológicos (Shadish *et al.*, 2002; Donner & Klar, 2000; Murray, 1998; Campbell *et al.*, 2012; Cameron & Miller, 2015). Consideraciones de fidelidad/implementación: (Bellg *et al.*, 2004; Carroll *et al.*, 2007; Dane & Schneider, 1998). Invariancia/estabilidad de autoinformes: (Chen, 2007; Cheung & Rensvold, 2002). Efecto de novedad en IVR: (Parong & Mayer, 2018; Radianti *et al.*, 2020; Makransky & Petersen, 2021).

Para complementar la lectura estandarizada y anclar la magnitud del cambio en una escala de interpretación directa, se reportan descriptivos PRE–POST por condición en la escala 0–100, incluyendo tendencia central, dispersión y proporción de puntajes cercanos al techo. La Tabla 4.9D resume estos estadísticos y permite verificar si el incremento observado corresponde a un desplazamiento sustantivo del desempeño con variabilidad preservada o a una concentración en valores altos.

Tabla 4.9D

Descriptivos PRE–POST por grupo y porcentaje de puntajes en techo (0–100)

<i>Outcome</i>	<i>Grupo</i>	<i>Tiempo</i>	<i>n</i>	<i>Media</i>	<i>DE</i>	<i>Mediana</i>	<i>IQR</i>	<i>% techo (≥90/100)</i>
Comunicación	IA+RV	PRE	90	50.53	8.81	52.00	12.00	0.0
	IA+RV	POST	90	73.09	9.05	73.00	8.00	4.4
	Comparación	PRE	90	49.62	8.21	48.00	10.00	0.0
	Comparación	POST	90	50.06	9.44	52.00	14.00	0.0
Solución de problemas	IA+RV	PRE	90	48.99	9.15	48.00	11.00	0.0
	IA+RV	POST	90	73.29	6.94	73.00	9.00	0.0
	Comparación	PRE	90	50.03	9.93	50.00	14.00	0.0
	Comparación	POST	90	50.79	9.19	50.00	12.00	0.0

Nota. Puntajes agregados por rúbrica en escala 0–100. DE = desviación estándar; IQR = rango intercuartílico (P75–P25). El % techo corresponde a la proporción de puntajes $\geq 90/100$ (umbral fijo). Los descriptivos se reportan sobre la cohorte de implementación con datos disponibles por condición y por momento (hasta $N = 180$; nominalmente 90 por condición), por lo que el n puede variar por celda según completitud PRE/POST y reglas operativas de reposición documentadas. Esta tabla es descriptiva (distribución global) y se presenta en formato largo (PRE/POST). Los contrastes inferenciales principales de desempeño se estiman sobre la muestra analítica per-protocol ($n = 160$; comparación $n = 79$; intervención IA+RV $n = 81$), definida por asistencia mínima ($\geq 75\%$) y evidencia PRE y POST comparable. Para el diagnóstico por dimensión, los perfiles por criterio se reportan en la escala original 1–4.

En términos operativos, la plataforma utilizada entrega, tras cada interacción, analítica de desempeño y retroalimentación estructurada, además de recomendaciones para mejorar. En un contexto de retroalimentación inmediata y rutas de práctica guiada, es plausible que los tamaños de efecto reflejen un ciclo de práctica–*feedback*–re–práctica que acelera el aprendizaje observable en la medición POST, consistente con la lógica de la práctica deliberada y con la evidencia sobre *feedback* formativo oportuno y accionable

como motor de mejora (Ericsson *et al.*, 1993; Hattie & Timperley, 2007; Shute, 2008; Wisniewski *et al.*, 2020).

Con el efecto principal (H1–H2) establecido en términos de magnitud y patrón por criterios, la subsección 4.5.8 presenta análisis complementarios orientados a implementación: (i) dosis–respuesta y moderación, y (ii) verificaciones de robustez mediante especificaciones alternativas y análisis de sensibilidad.

4.3.11 Análisis complementarios y robustez

4.3.11.1 Dosis–respuesta y moderación (OLS)

Para explorar la posibilidad de un gradiente dosis–respuesta dentro de la cohorte completa, se construyó un índice compuesto de dosificación que integró: (a) tiempo acumulado en simulaciones de RV, (b) intensidad de uso de IA para retroalimentación y (c) número de eventos de *feedback*. Con el fin de combinar componentes heterogéneos en una métrica común, el índice se estandarizó (z) y se inspeccionó su distribución para descartar valores extremos que pudieran dominar la estimación (Tabachnick & Fidell, 2019; Wooldridge, 2016). Posteriormente, el índice se incorporó como predictor en modelos OLS donde la variable dependiente fue la ganancia ($\Delta = \text{POST} - \text{PRE}$) en Comunicación y en Solución de problemas ($n = 180$). Dado que la asignación y parte de la variación de la exposición ocurren a nivel clase, estos modelos se reportan como exploración de implementación y no como inferencia primaria. Por la dependencia intraclase, los resultados se interpretan con cautela y, cuando corresponde, se reportan con ajustes por conglomerados a nivel clase; sin embargo, dada la cantidad reducida de conglomerados, estos hallazgos se consideran indicativos y su confirmación se plantea como agenda futura con diseños que introduzcan variación deliberada de intensidad por clase (Cameron & Miller, 2015; Donner & Klar, 2000).

En ambos modelos se incluyeron efectos principales del índice, experiencia tecnológica previa y grupo (intervención vs. comparación), así como el término de interacción Índice $\times$ Grupo para evaluar moderación. Esta parametrización responde a recomendaciones estándar para probar efectos condicionados por pertenencia a grupo en regresión, privilegiando la interpretación del término de interacción como evidencia de

heterogeneidad del efecto (McClelland & Judd, 1993; West *et al.*, 1996). La interpretación se apoya en la calidad de medición reportada previamente (Tablas 4.6A y 4.7), lo que reduce la probabilidad de que la ausencia de gradientes dosis-respuesta refleje primordialmente error de calificación y orienta la lectura hacia el rango efectivo de variación en la exposición observada.

Tabla 4.10

Modelo OLS para ganancia en Comunicación (Δ Comunicación), $n = 180$

Parámetro	β	p
Índice de dosificación (z)	-2.7870	0.644
Experiencia tecnológica previa	-1.2163	0.088
Grupo (intervención vs. comparación)	21.6721	0.047
Índice $\times$ Grupo	1.0569	0.150

Nota. Ajuste del modelo: $R^2 = 0.433$; R^2 ajustado = 0.420; $F(4,175) = 33.36$; $p < .001$.

Tabla 4.11

Modelo OLS para ganancia en Solución de problemas (Δ Solución de problemas), $n = 180$

Parámetro	β	p
Índice de dosificación (z)	3.3529	0.572
Experiencia tecnológica previa	0.0434	0.951
Grupo (intervención vs. comparación)	18.5730	0.081
Índice $\times$ Grupo	-0.2337	0.747

Nota. Ajuste del modelo: $R^2 = 0.465$; R^2 ajustado = 0.453; $F(4,175) = 38.06$; $p < .001$.

En conjunto, ambos modelos explican una proporción moderada de varianza ($R^2 \approx 0.43$ – 0.47), pero no aportan evidencia estadísticamente significativa de asociación entre el índice de dosificación y la mejora en desempeño, ni de moderación por condición (Índice $\times$ Grupo). En contraste, la pertenencia al grupo de intervención emerge como predictor estadísticamente significativo en Comunicación ($\beta = 21.6721$, $p = 0.047$), mientras que en Solución de problemas la estimación es positiva pero menos

precisa ($\beta = 18.5730$, $p = 0.081$). Interpretativamente, dentro del rango de implementación observado, el efecto parece operar predominantemente como un cambio cualitativo de arquitectura didáctica (participar o no en IA+RV) más que como un gradiente cuantitativo de más vs. menos dosis. Este patrón es consistente con dos explicaciones plausibles: (i) rango restringido de dosificación efectiva dentro de clase (lo cual reduce poder para detectar pendientes dosis–respuesta aun cuando existan), y/o (ii) un mecanismo en el que el rendimiento depende más de la estructura de la intervención (*briefing*–simulación–*debriefing* y calidad del *feedback*) que de incrementos marginales de exposición (Sackett & Yang, 2000; Schmidt *et al.*, 2006; Snijders & Bosker, 2012; West *et al.*, 1996).

4.3.11.2 Fidelidad de implementación (adherencia)

La fidelidad de implementación se operacionalizó como adherencia a componentes críticos del ciclo instruccional y de retroalimentación (*briefing*, *debriefing*, aplicación de rúbrica y señalización explícita de IA/RV). Esta estrategia responde a marcos de fidelidad que recomiendan documentar adherencia a componentes núcleo para interpretar efectos y distinguir ausencia de efecto por implementación deficiente vs. por ineficacia del diseño (Carroll *et al.*, 2007; Dane & Schneider, 1998). La matriz de adherencia por clase mostró cumplimiento elevado en componentes clave, con promedios de 70.0% en *briefing*, 93.3% en *debriefing*, 80.0% en aplicación de rúbrica y 86.7% en señalización explícita de IA/RV. Además, estos indicadores se definieron como mínimos operativos porque capturan los mecanismos teóricos del cambio (clarificación de metas, práctica situada, reflexión guiada y *feedback* criterial), por lo que su ausencia comprometería la interpretabilidad del contraste PRE–POST. En términos analíticos, la fidelidad se utilizó como evidencia de plausibilidad causal: si el efecto es consistente y la adherencia es alta, disminuye la probabilidad de que el resultado se explique por variación idiosincrática entre clases o por implementación incompleta. El mayor cumplimiento observado en *debriefing* y señalización es particularmente relevante, ya que ambos componentes estructuran la reflexión metacognitiva y conectan explícitamente la experiencia inmersiva con criterios de desempeño, condición central para que la práctica deliberada produzca mejora

observable. Asimismo, la tabla permite observar no solo los promedios generales de cumplimiento, sino también la variación entre clases, lo cual resulta importante para valorar la estabilidad operativa de la intervención. Esta lectura por rango ayuda a identificar si la adherencia fue homogénea o si existieron componentes con mayor dispersión, aspecto relevante para futuras decisiones de estandarización y escalamiento. Finalmente, el menor valor relativo en *briefing* (70%) se reporta de forma transparente como área de mejora, sugiriendo que estandarizar aún más los guiones de objetivos/criterios podría reducir variabilidad residual y fortalecer el escalamiento institucional.

Tabla 4.12

Adherencia por componente (promedio y rango)

Componente	Promedio (%)	Rango (%)
<i>Briefing</i>	70.0	60–80
<i>Debriefing</i>	93.3	80–100
Aplicación de rúbrica	80.0	40–100
Señalización explícita IA/RV	86.7	60–100

Desde una lectura de implementación, la alta adherencia en *debriefing* es congruente con el supuesto de que la reflexión guiada y la retroalimentación estructurada constituyen mecanismos centrales del modelo; simultáneamente, la variabilidad observada en *briefing* y, especialmente, en la aplicación de rúbrica (40–100) puede contribuir a explicar por qué un índice agregado de dosificación no produce un gradiente claro: si la calidad o completitud de componentes clave fluctúa entre clases, la exposición cuantitativa puede no capturar adecuadamente la dosis efectiva de aprendizaje (Carroll *et al.*, 2007; Dane & Schneider, 1998).

4.3.11.3 Contraste explícito de hipótesis (síntesis)

Con base en los resultados principales, H1 y H2 se consideran confirmadas por efectos Tiempo $\times$ Grupo de gran magnitud ($\beta \approx 22\text{--}24$ con IC95% excluyendo 0) y por tamaños de efecto de diferencia-en-diferencias $d_{ppc2} \approx 2.47\text{--}2.60$, consistentes entre *outcomes* (Tablas 4.11A y 4.11B). En contraste, H3 (gradiente dosis–respuesta) no se confirma: ni el índice de dosificación estandarizado ni su interacción índice $\times$ grupo

alcanzaron significancia estadística para la ganancia en Comunicación y Solución de problemas (Tablas 4.12–4.13). Finalmente, H4 (moderación por experiencia tecnológica previa y/o contexto) no muestra patrones robustos bajo la especificación actual; por ello, se plantea como agenda futura mediante diseños deliberadamente escalonados en intensidad y métricas más sensibles de implementación, distinguiendo con mayor precisión la dosis cuantitativa (tiempo/eventos) de la calidad de implementación (p. ej., calidad del *debriefing* y del *feedback*), con el fin de fortalecer la inferencia sobre heterogeneidad y mejorar la interpretabilidad de efectos condicionados (Aiken & West, 1991; Hayes, 2018; Carroll *et al.*, 2007).

4.4 Graficación

La estrategia de graficación se definió para representar el diagnóstico de manera informativa y congruente con el diseño, evitando inferencias basadas únicamente en promedios. Se priorizaron visualizaciones que permiten observar tendencia central y dispersión, comparar perfiles por criterios o dimensiones y detectar posibles efectos de techo o compresión de varianza que pudieran distorsionar la lectura del estado inicial. Bajo este criterio, las figuras incluidas cumplen una función diagnóstica: hacen visible la estructura de la línea base, la heterogeneidad entre participantes y la forma de las distribuciones, antes de introducir contrastes inferenciales. Los criterios de calidad de medición que condicionan la interpretación de dichas visualizaciones se sustentan en la fiabilidad interjueces de las rúbricas (Apéndice A3) y en los controles de integridad de datos (Apéndice A4).

El flujo de inclusión/exclusión se sintetiza en un diagrama tipo CONSORT (Figura 4.11), distinguiendo las etapas de ingreso, exclusión por asistencia e inclusión final en los análisis, en línea con recomendaciones para transparentar el movimiento muestral y documentar criterios de exclusión y pérdidas en estudios comparativos, particularmente cuando existe asignación por conglomerados (Schulz *et al.*, 2010; Campbell *et al.*, 2012).

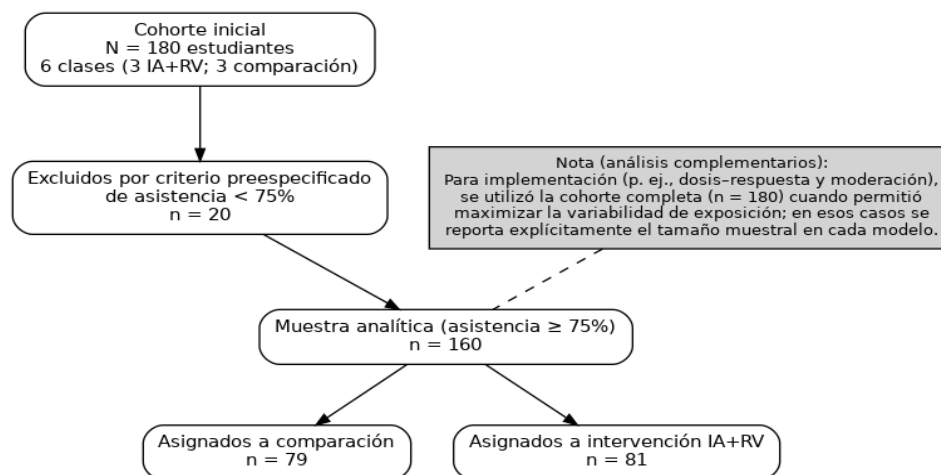
La inclusión de este diagrama resulta necesaria porque permite diferenciar entre la cohorte completa expuesta al proceso de implementación y la muestra analítica

utilizada para la inferencia principal. Esta distinción es metodológicamente relevante en el estudio, dado que $N = 180$ representa el flujo total de estudiantes considerados durante la implementación, mientras que $n = 160$ corresponde a la muestra *per-protocol* definida a partir del criterio mínimo de asistencia y completitud de evidencia. Al presentar esta información de forma visual, se evita que el lector interprete ambas cifras como intercambiables y se fortalece la transparencia sobre el tránsito desde la población participante hasta el conjunto final de análisis.

Además, el diagrama permite ubicar los criterios de exclusión como decisiones previamente justificadas y no como ajustes posteriores a los resultados. En estudios educativos aplicados, esta claridad es importante porque las pérdidas, ausencias o exclusiones pueden afectar la interpretación de los efectos si no se reportan de manera explícita. Por ello, la Figura 4.11 funciona como un punto de control visual que conecta la implementación del trabajo de campo con la validez de la base analítica, mostrando que los resultados posteriores se estimaron sobre una muestra definida mediante reglas consistentes y trazables.

Figura 4.11

Diagrama tipo CONSORT de inclusión/exclusión y muestra analizada



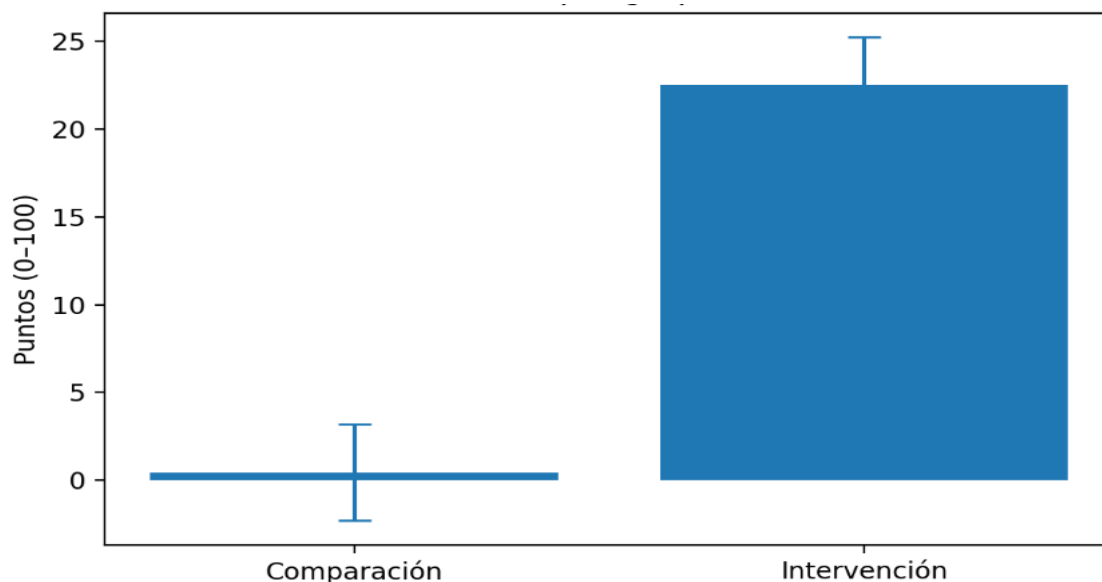
Nota. El diagrama presenta el flujo muestral desde la cohorte inicial ($N = 180$) hasta la base analítica principal de desempeño ($n = 160$) definida por asistencia mínima ($\geq 75\%$). Los conteos de $N = 180$ se reportan para transparencia del proceso y para análisis complementarios de implementación cuando exista información operativa disponible.

4.4.1 Ganancia PRE–POST por grupo

Para una lectura inmediata del contraste, las Figuras 4.12A y 4.12B muestran la ganancia media ($\Delta = \text{POST} - \text{PRE}$) por grupo en cada *outcome*. Estas visualizaciones no sustituyen los estimadores inferenciales del modelo multinivel ni los descriptivos de la Tabla 4.9D, pero ofrecen una síntesis intuitiva del cambio diferencial, haciendo explícita la magnitud relativa del avance en IA+RV frente a comparación y facilitando su interpretación pedagógica en la escala 0–100. Asimismo, estas figuras permiten identificar si el patrón observado es coherente con la dirección esperada por las hipótesis, sin adelantar conclusiones causales que corresponden al modelo estadístico. Por ello, su función es principalmente descriptiva y pedagógica: orientar al lector sobre el sentido del cambio antes de revisar la evidencia inferencial completa.

Figura 4.12A

Ganancia PRE–POST por grupo, Comunicación

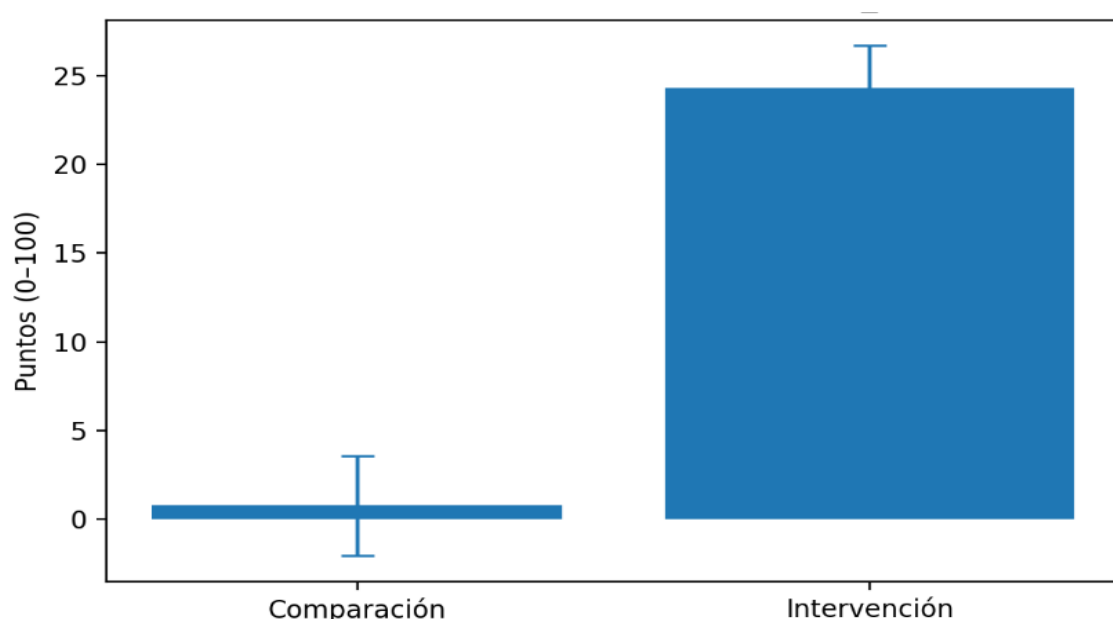


Nota. Se presenta la ganancia media ($\Delta = \text{POST} - \text{PRE}$) por condición en la escala 0–100. El contraste visual muestra una mejora mayor en la intervención IA+RV frente a comparación, coherente con el efecto diferencial estimado en el modelo multinivel y con la magnitud del efecto neto reportada para Comunicación (Tabla 4.9B). La figura se interpreta como anclaje descriptivo en puntos de rúbrica y no sustituye la inferencia basada en el término $\text{Tiempo} \times \text{Grupo}$. Para descartar interpretaciones por saturación de escala, la ganancia se triangula con la evidencia distributiva, los puntajes en techo reportados en la Tabla 4.9D y los *boxplots* correspondientes.

Con la misma lógica descriptiva, la Figura 4.12B presenta la ganancia observada en Solución de problemas, lo que permite mantener una lectura paralela entre ambos *outcomes*. Esta continuidad gráfica facilita comparar si el patrón de mejora diferencial se reproduce en una competencia distinta, conservando la misma escala, el mismo criterio de cambio y la misma interpretación pedagógica en puntos de rúbrica.

Figura 4.12B

Ganancia PRE–POST por grupo, Solución de problemas



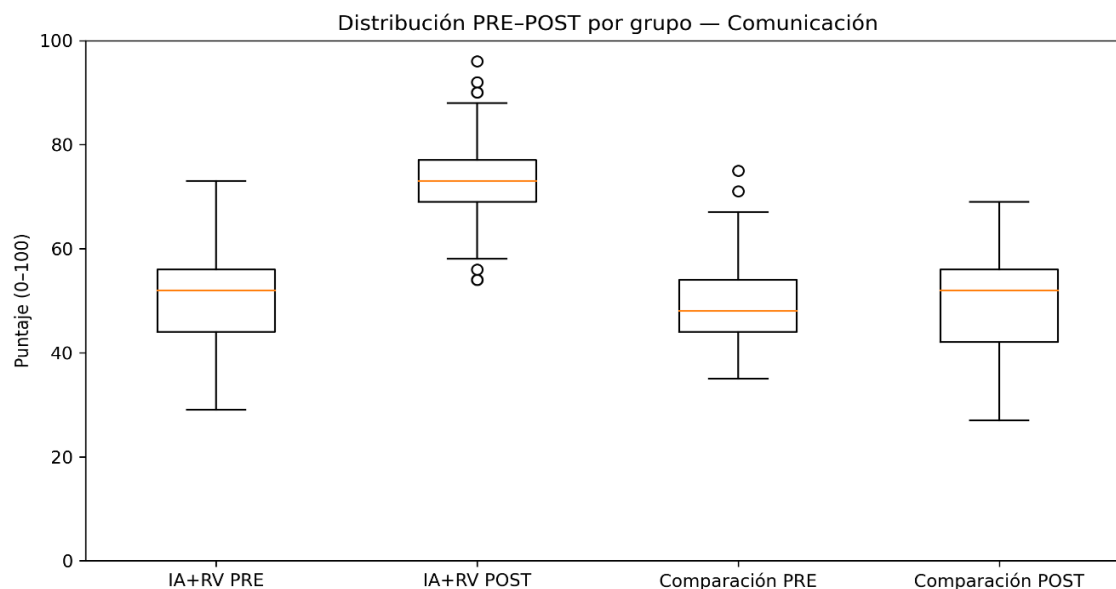
Nota. Se presenta la ganancia media ($\Delta = \text{POST} - \text{PRE}$) por condición. El patrón indica un incremento marcado en la condición IA+RV respecto a comparación y sugiere un efecto consistente entre *outcomes*, alineado con las estimaciones principales para Solución de problemas y con su tamaño de efecto de diferencia-en-diferencias (Tabla 4.9B). Esta visualización opera como anclaje descriptivo en la escala 0–100 para facilitar una lectura práctica de la magnitud del cambio. No sustituye la inferencia del modelo multinivel (β Tiempo $\times$ Grupo), sino que la complementa al hacer explícita la brecha de ganancias entre condiciones como aproximación al efecto neto de la intervención. Para robustecer la interpretación, se recomienda leerla en conjunto con la evidencia distributiva (mediana, IQR y ausencia de techo) reportada en la Tabla 4.9D y los *boxplots* correspondientes.

4.4.2 Distribuciones PRE–POST y *boxplots*

Complementariamente, las Figuras 4.12C y 4.12D presentan *boxplots* PRE–POST por grupo en la escala 0–100, con el fin de hacer visible la distribución (mediana, IQR y dispersión) y verificar la ausencia de acumulación generalizada en techo. En conjunto con la Tabla 4.9D, estos gráficos contextualizan la magnitud del cambio y fortalecen la plausibilidad del efecto observado.

Figura 4.12C

Distribución PRE-POST por grupo, Comunicación (0-100)

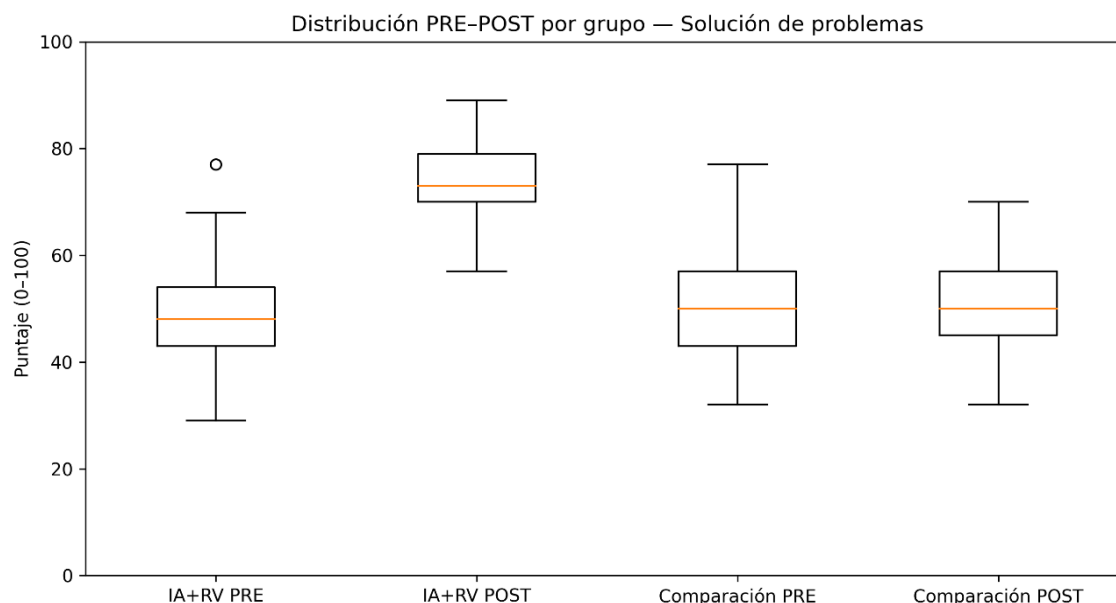


Nota. Se presenta la ganancia media ($\Delta = \text{POST} - \text{PRE}$) por condición en la escala 0-100. El contraste visual muestra una mejora mayor en la intervención IA+RV frente a comparación, coherente con el efecto diferencial estimado en el modelo multinivel y con la magnitud del efecto neto reportada para Comunicación (Tabla 4.9B). La figura se interpreta como anclaje descriptivo en puntos de rúbrica y no sustituye la inferencia basada en el término $\text{Tiempo} \times \text{Grupo}$. Para descartar interpretaciones por saturación de escala, la ganancia se triangula con la evidencia distributiva, los puntajes en techo reportados en la Tabla 4.9D y los boxplots correspondientes.

Con la misma lógica descriptiva, la Figura 4.12B presenta la ganancia observada en Solución de problemas, lo que permite mantener una lectura paralela entre ambos *outcomes*. Esta continuidad gráfica facilita comparar si el patrón de mejora diferencial se reproduce en una competencia distinta, conservando la misma escala, el mismo criterio de cambio y la misma interpretación pedagógica en puntos de rúbrica.

Figura 4.12D

Distribución PRE–POST por grupo, Solución de problemas (0–100)



Nota. *Boxplot* de puntajes PRE y POST por condición. La línea central representa la mediana; la caja corresponde al IQR (P75–P25) y los bigotes representan la dispersión fuera de la caja según el criterio estándar del *boxplot*. Se reporta con propósito descriptivo para anclar la magnitud del cambio en la escala 0–100 y facilitar la evaluación visual de desplazamiento y variabilidad, en coherencia con la evidencia inferencial (modelo multinivel) y la Tabla 4.9B.

En conjunto, aunque los tamaños de efecto estandarizados de la Tabla 4.9B son elevados, la evidencia descriptiva no sugiere una saturación artificial de la escala: la Tabla 4.9D y los *boxplots* (Figuras 4.12C–4.12D) muestran un desplazamiento claro del POST en IA+RV, con variabilidad preservada y proporciones mínimas de puntajes en techo ($\geq 90/100$). Bajo este encuadre, la interpretación prioriza la ganancia en puntos (0–100) y su coherencia con el estimador multinivel (Tiempo×Grupo), utilizando *d_ppc2* como estandarización complementaria.

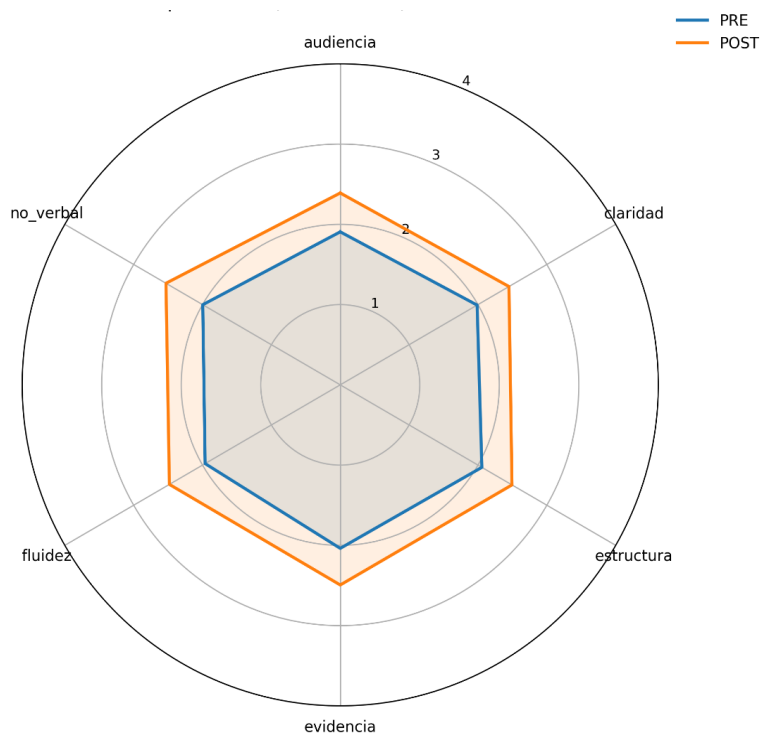
4.4.3 Gráficas radiales PRE–POST por criterio

Además de estimar el cambio global por variable de resultado (*outcome*), es pertinente examinar cómo se distribuye la mejora dentro de cada competencia, dado que Comunicación y Solución de problemas son constructos multidimensionales evaluados mediante criterios analíticos. Para la interpretación global, los puntajes se agregaron y

reescalaron a 0–100; para el diagnóstico por dimensión, los perfiles se reportan en la escala original 1–4. Con este propósito, las Figuras 4.13A y 4.13B presentan el perfil PRE–POST por criterio para cada rúbrica, lo que permite identificar si el avance es relativamente homogéneo o si se concentra en dimensiones específicas, aportando una lectura diagnóstica útil para discutir mecanismos instruccionales y áreas de ajuste.

Figura 4.13A

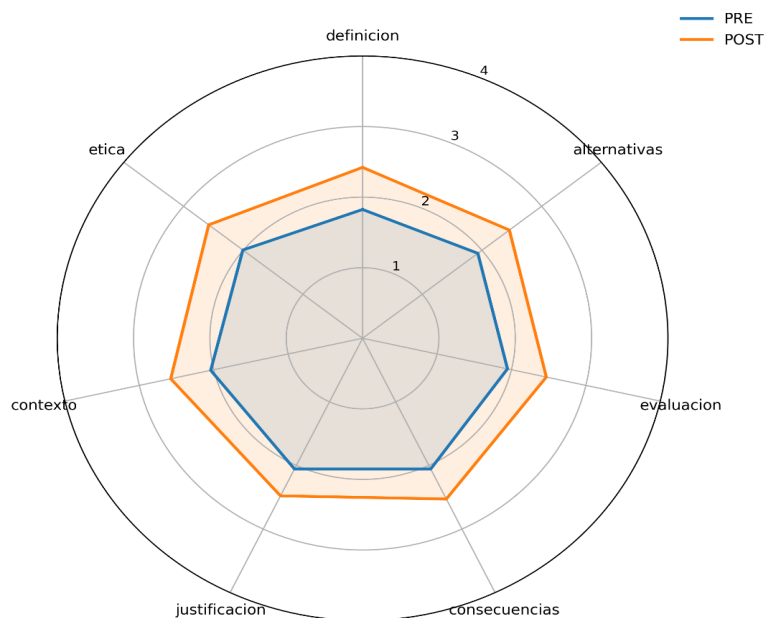
Perfil PRE–POST por criterio, Comunicación (rúbrica analítica, 6 criterios).



Nota. Se muestran los promedios PRE y POST por criterio en escala 1–4 (1 = desempeño incipiente; 4 = desempeño sobresaliente). Los perfiles son descriptivos; los contrastes inferenciales del efecto neto se reportan en el análisis PRE–POST y diferencia-en-diferencias. Los puntajes provienen de doble calificación independiente; ante discrepancias relevantes se aplicó el protocolo de recalibración/tercer dictamen descrito en el Apéndice A7.2.

Figura 4.13B

Perfil PRE–POST por criterio, Solución de problemas (rúbrica analítica, 7 criterios).



Nota. Se muestran los promedios PRE y POST por criterio en escala 1–4 (1 = desempeño incipiente; 4 = desempeño sobresaliente). Los perfiles son descriptivos; los contrastes inferenciales del efecto neto se reportan en el análisis PRE–POST y diferencia-en-diferencias. Los puntajes provienen de doble calificación independiente; ante discrepancias relevantes se aplicó el protocolo de recalibración/tercer dictamen descrito en el Apéndice A7.2.

En conjunto, los perfiles por criterio se reportan con un fin descriptivo y diagnóstico: no reemplazan los modelos multinivel ni los tamaños de efecto, sino que los complementan al caracterizar la huella multidimensional del cambio. Esta huella es útil para la interpretación de implementación, pues orienta dónde podrían estar operando con mayor intensidad los mecanismos del ciclo *briefing–simulación–debriefing* y del *feedback* criterial, y dónde convendría focalizar ajustes (señalización, andamiajes, *prompts* y prácticas de *debriefing*) en iteraciones futuras.

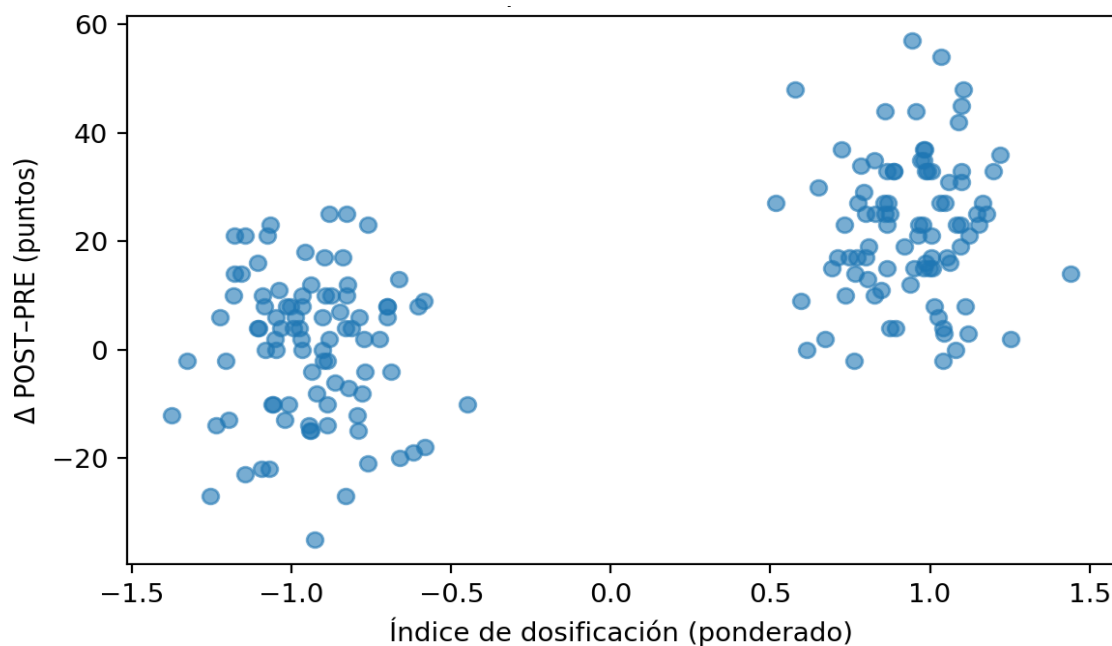
4.4.4 Graficación complementaria de dosis–respuesta

Para complementar los modelos OLS del gradiente dosis–respuesta (H3), se presentan las Figuras 4.14A y 4.14B como diagramas de dispersión (*Scatter*) entre el índice de dosificación estandarizado y las ganancias PRE–POST en Comunicación y Solución de problemas. Esta evidencia visual permite verificar si existe una tendencia

sistemática dentro del rango de exposición observado y, en conjunto con las Tablas 4.12–4.13, contextualiza la ausencia de un patrón dosis–respuesta robusta.

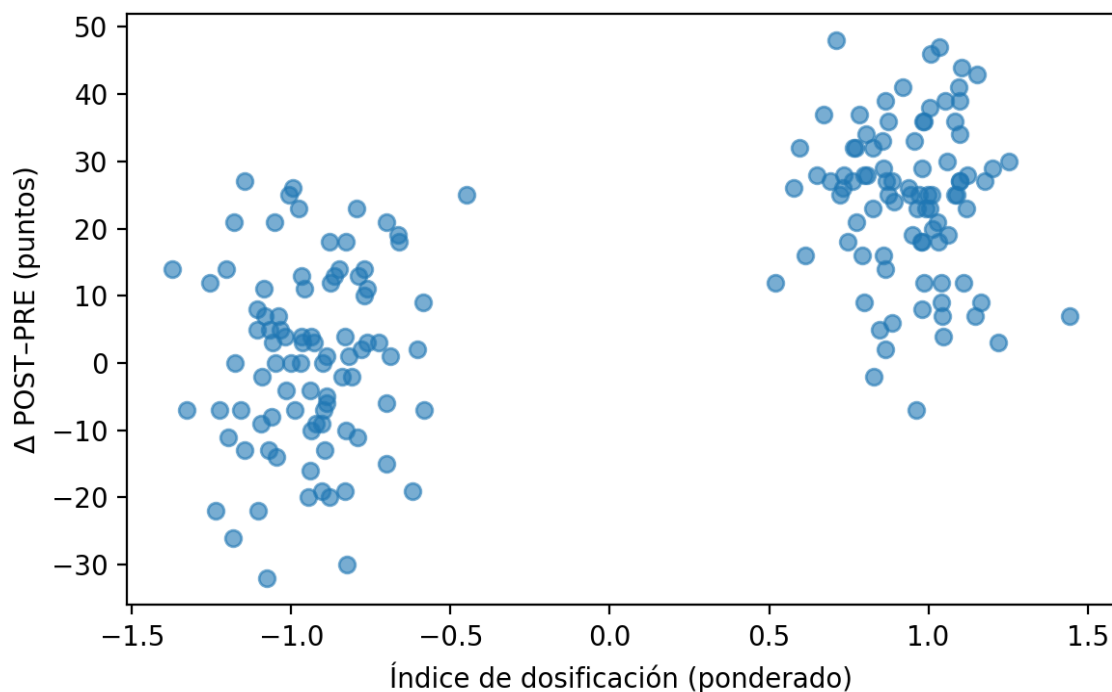
Figura 4.14A

Scatter dosis–respuesta, Comunicación



Nota. Cada punto representa un estudiante. El eje X corresponde al índice de dosificación estandarizado (z) y el eje Y representa la ganancia en Comunicación ($\Delta = \text{POST} - \text{PRE}$). La nube no sugiere una tendencia lineal marcada, en línea con el modelo OLS, donde el término del índice no resulta estadísticamente significativo.

De manera complementaria, la Figura 4.14B replica la misma lógica visual para Solución de problemas, manteniendo constante la escala del índice de dosificación y el criterio de ganancia PRE–POST. Esta comparación paralela permite valorar si la ausencia de una tendencia lineal clara se limita a Comunicación o si también se observa en el segundo *outcome* del estudio. Al presentar ambas visualizaciones bajo el mismo criterio gráfico, se facilita una interpretación más prudente del gradiente dosis–respuesta, evitando atribuir la mejora promedio únicamente a mayor exposición acumulada.

Figura 4.14B*Scatter dosis–respuesta, Solución de problemas*

Nota. Cada punto representa un estudiante. El eje X corresponde al índice de dosificación estandarizado (z) y el eje Y a la ganancia en Solución de problemas ($\Delta = \text{POST} - \text{PRE}$). El patrón visual es compatible con un gradiente dosis–respuesta débil o ausente dentro del rango de exposición observado, coherente con los resultados del modelo OLS (Tabla 4.11), incluyendo la falta de significación del índice y de su interacción con el grupo.

En suma, dentro del rango de implementación observado, la evidencia sugiere que el efecto opera principalmente como un cambio cualitativo de arquitectura didáctica (participar o no en IA+RV) más que como un gradiente cuantitativo de más vs. menos dosis. Este patrón es consistente con dos explicaciones plausibles: (i) un rango restringido de dosificación efectiva dentro de clase, y/o (ii) una dependencia del efecto en la estructura del ciclo *briefing*–simulación–*debriefing* y en la calidad del *feedback* más que en incrementos marginales de exposición.

4.5 Redacción de los resultados

La presente redacción se apoya en el supuesto metodológico de que los resultados solo pueden interpretarse con solidez cuando las variables del estudio han sido definidas, operacionalizadas y medidas de manera consistente. En este sentido, la

lectura de los hallazgos no parte únicamente de diferencias observadas entre grupos, sino de la correspondencia entre conceptos teóricos, dimensiones, indicadores y evidencias empíricas que estructuraron la medición de Comunicación, Solución de problemas, dosificación, fidelidad y experiencia de aprendizaje. Como señalan Martello y Fornetti (2021), las variables orientan la recolección y el análisis de datos, y su valor científico depende de la precisión con que se traduzcan en referentes observables; por ello, la discusión que sigue se fundamenta en resultados previamente asegurados por criterios de validez, consistencia de medición y trazabilidad analítica, lo que permite interpretar los hallazgos no como impresiones aisladas, sino como evidencia empíricamente sustentada.

4.5.1 Interpretación por hipótesis

En conjunto, los resultados apoyan con solidez la conclusión de que la arquitectura didáctica basada en IA + RV, implementada como un paquete instruccional estructurado (*briefing-práctica-debriefing* con criterios explícitos y retroalimentación mediada), produjo mejoras sustantivas en desempeño observable en *Comunicación y Solución de problemas*. La inferencia central se sostiene en dos pilares complementarios: (i) el contraste inferencial de la interacción Tiempo $\times$ Grupo en modelos multinivel (que operacionaliza el cambio diferencial PRE-POST entre condiciones) y (ii) la robustez de la medición mediante rúbricas con alta consistencia interjueces, lo cual reduce de manera importante la plausibilidad de que el patrón observado se explique por error de calificación o por fluctuación aleatoria (Raudenbush & Bryk, 2002; Shrout & Fleiss, 1979; Koo & Li, 2016). Desde el punto de vista pedagógico, la magnitud en escala 0–100 sugiere un cambio educativo relevante: no se trata únicamente de mejorar por exposición al curso, sino de un incremento diferencial compatible con una reconfiguración del régimen de práctica y retroalimentación.

Cabe enfatizar que el índice de dosificación operacionaliza principalmente cantidad de exposición/eventos y no captura directamente la calidad del *feedback* (p. ej., claridad, especificidad y accionabilidad), que podría ser un mediador más próximo del cambio observado.

Aunque los tamaños de efecto son inusualmente altos en investigación educativa, son pedagógicamente plausibles dada (i) la práctica deliberada altamente alineada con el desempeño evaluado mediante rúbricas auténticas, (ii) la retroalimentación más accionable, y (iii) la práctica distribuida. No obstante, la generalización del tamaño del efecto requiere replicación en nuevas cohortes y con variantes de medición.

Bajo H1 y H2, los hallazgos son coherentes con la literatura que establece que el aprendizaje mejora cuando la retroalimentación reduce explícitamente la brecha entre desempeño actual y esperado, está alineada a criterios, y favorece la autorregulación mediante información utilizable para revisar y reintentar (Hattie & Timperley, 2007; Shute, 2008; Wiggins, 2012). En su intervención, la rúbrica funciona como lenguaje común de calidad y el ciclo *briefing-debriefing* opera como andamiaje para que el estudiante convierta la experiencia en ajustes concretos, lo que coincide con enfoques de práctica deliberada: metas específicas, foco en subcomponentes críticos, *feedback* correctivo y reiteración (Ericsson *et al.*, 1993; Ericsson, 2006).

En contraste, H3 (dosis–respuesta) no se confirma con las métricas y el rango de variación disponibles en su implementación. El hecho de que el índice de dosificación no muestre una asociación significativa con la ganancia no invalida la teoría de dosis; más bien sugiere que, en este estudio, (a) el rango efectivo de variación dentro de la muestra pudo ser insuficiente para estimar un gradiente con precisión (rango restringido), (b) la dosificación capturada por minutos/conteos puede estar midiendo cantidad pero no calidad (p. ej., calidad del *debriefing* o accionabilidad del *feedback*), o (c) el efecto responde a un modelo de umbral: una vez alcanzadas condiciones mínimas de estructura instruccional y guía criterial, los retornos marginales de más minutos pueden ser decrecientes en el corto plazo. Esta interpretación es consistente con síntesis contemporáneas sobre retroalimentación, que muestran heterogeneidad considerable de efectos y dependencia crítica de condiciones de implementación, tarea y autorregulación del estudiante (Wisniewski *et al.*, 2020; Hattie & Timperley, 2007).

Finalmente, H4 (moderación por experiencia tecnológica previa/programa) no se observa de manera robusta. Este resultado es frecuente en diseños educativos por conglomerados: mientras el efecto principal puede detectarse con claridad cuando el

tratamiento representa un cambio sustantivo de arquitectura instruccional, la estimación de moderaciones requiere (i) mayor potencia efectiva, (ii) mayor variabilidad contextual y (iii) un número suficiente de unidades de agrupamiento para separar con precisión varianza entre clases/programas (McClelland & Judd, 1993; Raudenbush & Bryk, 2002). En términos interpretativos, lo más defendible es afirmar que la intervención opera como un cambio de régimen relativamente generalizable dentro del contexto estudiado, sin evidencia fuerte (aquí) de que su impacto dependa sistemáticamente del perfil tecnológico previo.

En conjunto, la evidencia indica que la intervención IA+RV produjo un efecto principal robusto sobre el desempeño observable en Comunicación y Solución de problemas, consistente con H1 y H2. A la vez, bajo el rango de implementación efectivamente alcanzado, no se identificaron gradientes estables atribuibles a incrementos marginales de dosificación (H3), ni patrones consistentes de heterogeneidad del efecto por perfil o programa (H4). En términos sustantivos, los hallazgos delimitan con precisión el tipo de inferencia que el estudio sostiene: la ganancia se explica mejor como un cambio cualitativo de arquitectura didáctica (criterios explícitos, práctica deliberada en tareas situadas y retroalimentación accionable) que como una función lineal de más minutos o más eventos dentro de la variabilidad observada.

Aunque los tamaños de efecto son inusualmente altos en investigación educativa, son plausibles dado que el tratamiento representa un cambio de régimen instruccional (arquitectura didáctica) más que un incremento marginal de exposición. No obstante, la generalización del tamaño del efecto requiere replicación en nuevas cohortes y con variantes de medición.

4.5.2 Mecanismo plausible: fidelidad, dosificación y experiencia de aprendizaje

Un aporte central de los resultados es que la explicación del efecto no descansa únicamente en que hubo RV o que hubo IA, sino en cómo se integraron pedagógicamente. La literatura sobre aprendizaje inmersivo subraya que la inmersión puede incrementar presencia e involucramiento, pero no garantiza aprendizaje por sí misma; el rendimiento depende de que la experiencia esté anclada a objetivos, criterios y

reflexión guiada, y de que el diseño contenga la carga extrínseca (Makransky & Petersen, 2021; Mayer, 2021; Sweller *et al.*, 2011). En otras palabras: la inmersión se vuelve pedagógica cuando se convierte en práctica evaluable (criterial) y no en exposición.

Desde la Teoría de Carga Cognitiva y el aprendizaje multimedia, el desempeño mejora cuando el diseño reduce demandas irrelevantes (carga extrínseca) y promueve procesamiento orientado a esquemas útiles (carga germana), mediante señalización, segmentación, preentrenamiento y *debriefing* estructurado (Mayer, 2021; Sweller *et al.*, 2011). En su intervención, el *briefing* define la meta y los criterios; la práctica inmersiva coloca al estudiante en una situación de desempeño; y el *debriefing* convierte la experiencia en aprendizaje al forzar explicación, justificación y planes de mejora. Esta secuencia es consistente con evidencia en simulación: el *debriefing* es un componente crítico para transformar hacer en aprender, especialmente cuando el objetivo es desempeño complejo (Fanning & Gaba, 2007; Cheng *et al.*, 2014).

Este marco también permite interpretar por qué no emerge un gradiente dosis–respuesta. Si la intervención aseguró alta fidelidad en componentes críticos (p. ej., *debriefing* consistente y uso criterial de rúbrica), la mayor parte de la varianza explicativa puede provenir de participar o no en esa arquitectura, más que de pequeñas diferencias marginales de minutos. Para capturar H3 con mayor sensibilidad, probablemente se requerirían niveles intencionalmente diferenciados de intensidad (p. ej., baja/media/alta) y, sobre todo, métricas de calidad de implementación (calidad del *debriefing*, accionabilidad del *feedback*, cumplimiento de andamiajes), no solo de tiempo o conteos.

En cuanto al componente de IA, su lectura mecánica debe formularse con precisión: la IA no explica el efecto por automatizar el aprendizaje, sino por acelerar ciclos de retroalimentación y apoyar la autorregulación, siempre que el diseño evite sustitución cognitiva. Esta cautela está alineada con evidencia más amplia sobre *automation bias* y dependencia en sistemas de apoyo a decisiones: cuando una herramienta reduce el esfuerzo, puede también desplazar procesamiento profundo si no se diseña con salvaguardas (Parasuraman & Riley, 1997; Skitka *et al.*, 1999). Su

decisión de acotar la IA al *feedback* formativo y mantener la evaluación en desempeño observable ayuda a blindar la inferencia: el tratamiento no es IA que resuelve, sino IA que orienta revisión.

4.5.3 Comparación con la literatura

Los tamaños de efecto observados en este estudio superan los promedios que suelen reportarse en síntesis de investigación sobre tecnologías inmersivas en educación, donde los efectos tienden a ubicarse en rangos pequeños a moderados y muestran heterogeneidad elevada por dominio, duración, población y calidad del diseño instruccional (Radianti *et al.*, 2020; Makransky & Petersen, 2021). Esta diferencia no necesariamente constituye una inconsistencia empírica; más bien, sugiere que el presente estudio se sitúa en una subclase de intervenciones donde la tecnología no opera como medio de exposición, sino como un entorno de desempeño organizado por criterios explícitos y ciclos de mejora observables.

Una fuente adicional de discrepancia con los promedios meta-analíticos es el tipo de *outcome*: varias síntesis reportan efectos más modestos cuando se evalúa conocimiento declarativo o pruebas estandarizadas, y efectos relativamente mayores cuando la RV se emplea para habilidades procedimentales o de desempeño bajo alta interactividad y tareas situadas, especialmente cuando el grupo de comparación mantiene prácticas habituales menos intensivas (Merchant *et al.*, 2014; Jensen & Konradsen, 2018). En este estudio, el uso de rúbricas analíticas, ciclos iterativos y *debriefing* sistemático sitúa la medición en desempeño observable, lo que hace más probable detectar cambios educativos significativos que en evaluaciones de recuerdo o comprensión superficial.

En esa subclase, la literatura sobre RV reporta resultados más favorables cuando el diseño controla la carga extrínseca, segmenta la complejidad, señaliza la información relevante y articula fases de preparación y reflexión (*briefing–debriefing*) que conectan la experiencia con criterios y transferencia (Parong & Mayer, 2018; Makransky & Petersen, 2021). De manera convergente, la evidencia sobre retroalimentación formativa muestra que el impacto depende críticamente de su calidad funcional: cuando el

feedback es específico, criterial y accionable, tiende a reducir la brecha entre desempeño actual y esperado; cuando es vago, excesivo o desalineado, puede producir efectos nulos o incluso adversos (Hattie & Timperley, 2007; Shute, 2008; Wisniewski *et al.*, 2020). Bajo esta lectura, el aporte del estudio se interpreta con mayor solidez como evidencia de una integración pedagógica en la que IA y RV se subordinan a una lógica de desempeño: rúbricas analíticas como lenguaje común, práctica deliberada en escenarios situados y ciclos reiterados de revisión y reflexión guiada.

Dentro de esta arquitectura, el componente de *virtual speech* (práctica de discurso oral con ajuste en tiempo real ante una audiencia simulada) aporta una explicación causal adicional: intensifica la presencia social y las demandas de comunicación auténticas, pero lo hace en un contexto de bajo costo social que permite ensayos repetibles y corrección inmediata orientada por rúbrica. En consecuencia, el efecto observado es coherente con marcos que sitúan la ganancia del aprendizaje inmersivo en la interacción entre presencia y agencia (CAMIL), y con evidencia que subraya que la inmersión solo se traduce en mejoras observables cuando se sostiene con andamiajes instruccionales explícitos (*briefing*–señalización–*debriefing*) y retroalimentación criterial (Cummings & Bailenson, 2016; Parong & Mayer, 2018; Makransky & Petersen, 2021).

4.5.4 Límites y amenazas a validez

Como intervención educativa aplicada en condiciones reales, este estudio está sujeto a amenazas de validez que deben declararse explícitamente y abordarse con razonamiento metodológico. En primer lugar, al tratarse de un diseño por conglomerados, la capacidad para estimar con alta precisión la heterogeneidad contextual (p. ej., variación entre clases, estilos docentes o microclimas de cohorte) y los efectos de moderación suele estar limitada cuando el número de unidades de nivel superior es reducido. No obstante, la estrategia analítica empleada atenúa este riesgo en el punto central de inferencia: los modelos multinivel permiten modelar la dependencia intraclase y estimar el efecto diferencial PRE–POST mediante la interacción Tiempo $\times$ Grupo, evitando inferencias infladas por no independencia y preservando validez inferencial en escenarios educativos naturalistas (Raudenbush & Bryk, 2002; Donner &

Klar, 2000). Además, la convergencia del patrón en dos *outcomes* (Comunicación y Solución de problemas), con dirección consistente y magnitudes sustantivas, reduce la plausibilidad de que el hallazgo principal sea un artefacto idiosincrático de una sola medida o de un único contexto.

En segundo término, la RV suele enfrentar la objeción del efecto de novedad: la inmersión puede incrementar motivación, interés y participación, especialmente en exposiciones iniciales. Esta amenaza se reconoce como teóricamente plausible; sin embargo, resulta menos parsimonioso atribuir exclusivamente a la novedad una mejora marcada en desempeño criterial cuando (i) la variable dependiente principal es desempeño observable evaluado con rúbricas, (ii) el proceso de calificación se soporta en consistencia interjueces elevada y estable, y (iii) los criterios de logro están anclados en descriptores conductuales. En otras palabras, la novedad puede operar como un facilitador del compromiso (condición habilitante), pero el resultado sustantivo se define por desempeño alineado a criterios explícitos, no por satisfacción o agrado con la experiencia (Makransky & Petersen, 2021; Parong & Mayer, 2018).

En tercer lugar, las escalas de autoinforme constituyen un riesgo clásico de validez cuando se interpretan como *outcomes* longitudinales robustos sin estabilidad psicométrica suficiente. En este estudio, dicho riesgo se maneja de forma deliberadamente conservadora: cuando la medición basal mostró inestabilidad, las escalas se interpretan principalmente como evidencia de proceso/mecanismo y aceptabilidad situacional, evitando derivar inferencias fuertes de cambio temporal perceptual sin garantías métricas. Esta decisión es consistente con estándares contemporáneos de evaluación educativa y psicológica, que enfatizan que las interpretaciones deben estar condicionadas por evidencia de fiabilidad, validez y comparabilidad de puntajes (AERA/APA/NCME, 2014).

En cuarto término, la ausencia de un gradiente dosis–respuesta detectable no debe leerse como refutación del marco teórico, sino como un límite empírico circunscrito al rango de implementación y a las métricas disponibles. Lo que el estudio permite afirmar con seguridad es que, dentro del rango observado y usando medidas como minutos, conteos y presencia percibida, el efecto se comporta predominantemente

como un cambio de arquitectura (participar o no en el régimen IA+RV) más que como una función lineal de intensidad marginal. Esto acota el tipo de inferencia posible y justifica una agenda de investigación futura basada en diseños con intensidades deliberadamente diferenciadas y medición más fina de calidad de implementación (p. ej., calidad del *debriefing*, especificidad/accionabilidad del *feedback*, cumplimiento de andamiajes críticos), dado que la literatura sobre implementación sugiere que la dosis efectiva no depende solo de cantidad, sino de fidelidad y calidad del componente entregado (Carroll *et al.*, 2007; Proctor *et al.*, 2011).

Finalmente, se reconoce que los estudios de campo enfrentan amenazas adicionales como contaminación entre condiciones, variaciones docentes no observadas o efectos Hawthorne. Sin embargo, el estudio incorpora salvaguardas que sostienen la interpretación principal: la estimación centrada en la interacción Tiempo $\times$ Grupo, el control por línea base (enfoque ANCOVA dentro del modelo), los análisis de sensibilidad y la trazabilidad de implementación reducen la probabilidad de que el efecto estimado sea atribuible a diferencias espurias o a decisiones analíticas ad hoc (Shadish *et al.*, 2002; Raudenbush & Bryk, 2002). En síntesis, pese a límites típicos de estudios aplicados, la convergencia entre inferencia multinivel, magnitudes observadas, fiabilidad interjueces elevada y evidencia de implementación consistente sostiene una conclusión robusta: una arquitectura IA+RV, implementada con criterios explícitos y *debriefing* estructurado, mejora de manera sustantiva el desempeño en comunicación y solución de problemas medidos por rúbricas.

4.6 Integración de hallazgos empíricos, calidad de medición e implicaciones inmediatas

Este capítulo aportó el núcleo empírico de la tesis al evaluar el impacto de una arquitectura didáctica que integra inteligencia artificial (IA) para retroalimentación formativa y simulaciones en realidad virtual (RV) sobre el desarrollo de dos competencias transversales de alta relevancia para la empleabilidad y el desempeño profesional: Comunicación y Solución de problemas/decisión. A partir de una cohorte inicial de $N = 180$, el criterio preespecificado de asistencia mínima ($\geq 75\%$) derivó en una muestra analítica final de $n = 160$ (comparación: $n = 79$; intervención IA+RV: $n =$

81). Este flujo, documentado en la Figura 4.15, reduce el riesgo de sesgo por atrición diferencial en la medida en que la exclusión obedeció a un criterio operativo uniforme (asistencia) y no a eventos adversos atribuibles a la exposición tecnológica.

El hallazgo central es la magnitud y consistencia del efecto diferencial PRE–POST a favor de la condición IA+RV. En los modelos multinivel, el parámetro clave de inferencia (la interacción Tiempo $\times$ Grupo) mostró estimaciones robustas para ambos *outcomes*: Comunicación ($\beta = 22.122$, IC95% [18.448, 25.796]) y Solución de problemas/decisión ($\beta = 23.544$, IC95% [19.880, 27.208]). Considerando la estructura jerárquica de los datos (mediciones repetidas anidadas en estudiantes y estudiantes anidados en clases) y el control por desempeño basal, estas estimaciones indican que el cambio del PRE al POST en la intervención excede ampliamente el cambio observado bajo docencia convencional, respaldando la existencia de una ganancia educativa sustantiva atribuible al paquete instruccional (Raudenbush & Bryk, 2002).

La conclusión sustantiva se refuerza con tamaños de efecto diferencia-en-diferencias de magnitud excepcional: $d_{ppc2} = 2.598$ (IC95% [2.20, 3.00]) para Comunicación y $d_{ppc2} = 2.466$ (IC95% [2.08, 2.85]) para Solución de problemas/decisión. Bajo criterios metodológicos habituales, estos valores corresponden a efectos muy grandes. No obstante, su interpretación se sostiene no por la etiqueta, sino por la coherencia entre estimación inferencial, precisión (intervalos de confianza), consistencia entre *outcomes* y trazabilidad del proceso de medición (Morris, 2008).

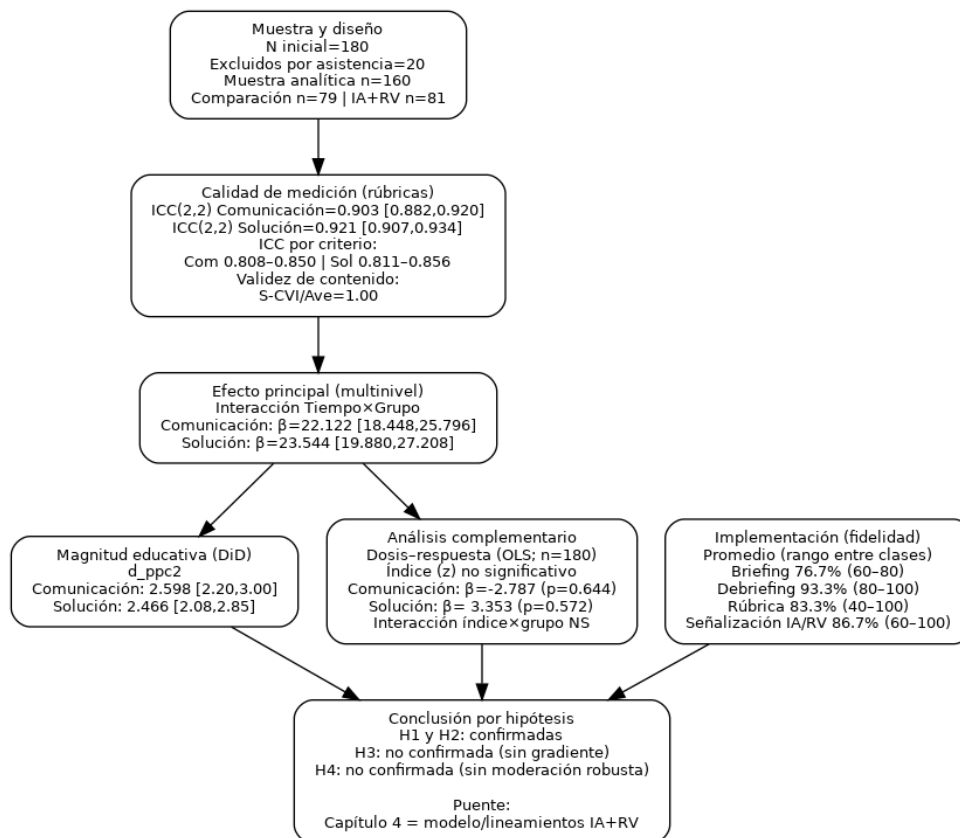
En términos de calidad de medición, el capítulo fundamenta su inferencia principalmente en desempeño observable debido a su elevada fiabilidad interjueces. El estimador operativo global mostró valores excelentes: ICC (2,2) = 0.903 (IC95% [0.882, 0.920]) para Comunicación y ICC (2,2) = 0.921 (IC95% [0.907, 0.934]) para Solución de problemas/decisión, con estabilidad adicional a nivel de criterios. Asimismo, la validez de contenido alcanzó niveles máximos (S-CVI/Ave = 1.00), reforzando que los descriptores capturan con pertinencia los constructos evaluados (Shrout & Fleiss, 1979; Koo & Li, 2016; AERA/APA/NCME, 2014). En contraste, las escalas de autoinforme se interpretaron de forma conservadora: dada la inestabilidad psicométrica basal, se utilizaron principalmente como evidencia de proceso/mecanismo en POST inmediato,

evitando sostener conclusiones longitudinales fuertes sobre percepciones sin estabilidad métrica suficiente (AERA/APA/NCME, 2014).

Los análisis complementarios añadieron un matiz relevante: el índice compuesto de dosificación no mostró asociaciones significativas con la ganancia ni moderación por condición, lo que sugiere que, dentro del rango de implementación observado, el efecto opera principalmente como adopción de una arquitectura instruccional (IA+RV) más que como función lineal de intensidad marginal medida por minutos, conteos o presencia percibida. Este resultado no invalida el efecto principal; más bien, acota el tipo de inferencia posible y delimita una agenda de investigación futura orientada a manipular intensidades de manera intencional y a medir con mayor resolución la calidad de los componentes críticos (p. ej., calidad del *debriefing* y accionabilidad del *feedback*) (Carroll *et al.*, 2007; Proctor *et al.*, 2011).

Finalmente, los indicadores de implementación respaldan la viabilidad operativa del modelo en un contexto universitario real: se observó adherencia alta en componentes clave, especialmente *debriefing*, y rangos razonables entre clases, lo cual reduce la objeción de que los efectos se expliquen por una entrega errática o por variaciones no controladas. En síntesis, el capítulo confirma con evidencia sólida H1 y H2 (efecto principal de la arquitectura IA+RV sobre desempeño), mientras que H3 (gradiente dosis–respuesta) y H4 (moderación por perfil/contexto) no se confirman en los términos estrictos planteados, aunque informan líneas de investigación específicas para diseños con variación deliberada de intensidad y medición más fina de implementación y mecanismos.

Sobre esta base, el Capítulo 5 retoma los resultados aquí documentados para traducirlos en una propuesta estructurada de implementación orientada a consolidación institucional, escalamiento responsable y evaluación continua, preservando las condiciones pedagógicas, operativas y de gobernanza que, en esta investigación, mostraron ser decisivas para la efectividad del modelo IA+RV.

Figura 4.15*Síntesis de resultados, calidad de medición y trazabilidad de implementación*

Nota. El diagrama integra muestra (N=180; n=160), calidad de medición (ICC global y por criterio; CVI),

CAPÍTULO 5. PROPUESTA DE TRANSFORMACIÓN

Este capítulo formula la propuesta derivada de los hallazgos empíricos del estudio: una arquitectura didáctica replicable para integrar inteligencia artificial (IA) y realidad virtual (RV) en educación superior con un propósito pedagógico verificable, orientado a elevar el desempeño observable en competencias complejas de Comunicación y Solución de problemas mediante ciclos de práctica deliberada sostenidos por *feedback* formativo criterial. A diferencia de aproximaciones centradas en incorporar tecnología, la propuesta se configura como un modelo instruccional auditable, en tanto especifica componentes, secuencias, criterios mínimos de calidad e indicadores de fidelidad que permiten documentar, comparar y mejorar su implementación entre cursos, cohortes y campus.

5.1 Denominación, fundamentación de la propuesta

La propuesta derivada de esta investigación se denomina “*Sistema de intervención didáctica con inteligencia artificial y realidad virtual para contribuir a la mejora del desempeño en competencias transversales de comunicación y solución de problemas mediante ciclos de práctica situada, retroalimentación criterial y reflexión en estudiantes de pregrado del Tecnológico de Monterrey, México, durante el periodo 2024 a 2025*”.

Esta denominación sintetiza la naturaleza, el propósito y el mecanismo central de la propuesta. En primer lugar, se define como una arquitectura didáctica porque no se limita a sugerir el uso de herramientas tecnológicas, sino que organiza componentes, secuencias, criterios de evaluación, registros de implementación y condiciones mínimas de calidad. En segundo lugar, integra inteligencia artificial y realidad virtual como recursos complementarios, no intercambiables ni autónomos. En tercer lugar, se orienta al desarrollo de competencias transversales, particularmente Comunicación y Solución de problemas, por tratarse de desempeños complejos, observables y relevantes para la formación universitaria. En cuarto lugar, se apoya en práctica deliberada, entendida como repetición intencional con metas específicas, retroalimentación y ajuste progresivo. Finalmente, utiliza la retroalimentación criterial y el desempeño situado

como mecanismos para conectar la experiencia de aprendizaje con evidencia observable de mejora.

La propuesta se diferencia de modelos centrados exclusivamente en la adopción tecnológica porque no asume que la presencia de IA o RV genere, por sí misma, mejores aprendizajes. Su supuesto central es que el efecto educativo depende de la calidad del diseño instruccional, la claridad de los criterios, la fidelidad de implementación, la calidad del *feedback*, la mediación docente y la posibilidad de convertir la experiencia en evidencia evaluable. En consecuencia, la arquitectura IA+RV se plantea como un modelo de adopción pedagógica responsable y no como una estrategia de modernización tecnológica superficial.

En términos de resultado científico propositivo, la arquitectura integra componentes conceptuales, instruccionales, metodológicos y operativos. Es conceptual porque se apoya en marcos consolidados sobre aprendizaje, desempeño, retroalimentación y autorregulación. Es instruccional porque organiza experiencias de aprendizaje mediante ciclos de preparación, ejecución, retroalimentación, reflexión y mejora. Es metodológica porque establece procedimientos, instrumentos e indicadores para implementar y documentar la intervención. Es operativa porque ofrece una ruta aplicable a cursos regulares, con potencial de adaptación a contextos institucionales comparables.

Por tanto, el tema de la propuesta no es la incorporación aislada de inteligencia artificial o realidad virtual en la enseñanza, sino el diseño de una arquitectura didáctica que articule ambas tecnologías con criterios pedagógicos verificables para fortalecer competencias transversales en educación superior.

5.2 Fundamentación teórico conceptual

La fundamentación teórico-conceptual de la propuesta se apoya en un conjunto de marcos que permiten explicar por qué una arquitectura didáctica IA+RV puede favorecer el desarrollo de competencias transversales cuando se diseña con criterios instruccionales explícitos. En particular, la propuesta articula aportes de la teoría de la carga cognitiva, el aprendizaje multimedia, el aprendizaje experiencial, la práctica

deliberada, la retroalimentación formativa, la evaluación criterial y la autorregulación como bases para diseñar experiencias de aprendizaje estructuradas, evaluables y orientadas a la mejora del desempeño (Sweller, 1988; Mayer, 2009; Kolb, 1984; Ericsson *et al.*, 1993; Hattie & Timperley, 2007; Shute, 2008; Zimmerman, 2000). A ello se integran estudios recientes sobre realidad virtual e inteligencia artificial en educación, que permiten justificar su uso como entorno de desempeño situado y como soporte de retroalimentación formativa bajo mediación docente (Radianti *et al.*, 2020; Makransky & Petersen, 2021; Kasneci *et al.*, 2023).

Desde la teoría de la carga cognitiva, el diseño de experiencias de aprendizaje debe considerar que la memoria de trabajo es limitada y que los entornos instruccionales pueden facilitar o dificultar el procesamiento de la información. En esta perspectiva, una experiencia con tecnología inmersiva solo tiene valor pedagógico si reduce estímulos irrelevantes, organiza la información, dirige la atención hacia los elementos importantes de la tarea y evita que la novedad tecnológica incremente la carga mental sin contribuir al aprendizaje (Sweller, 1988; Sweller *et al.*, 1998). Esta lógica es especialmente relevante para escenarios de RV, donde la riqueza sensorial puede favorecer la comprensión situada, pero también puede convertirse en distractor si no existe curaduría instruccional.

El aprendizaje multimedia complementa esta perspectiva al señalar que la combinación de información visual, auditiva e interactiva debe seguir principios de coherencia, señalización, segmentación y alineación con los objetivos de aprendizaje. En consecuencia, la arquitectura IA+RV debe evitar la acumulación indiscriminada de estímulos y privilegiar experiencias donde los recursos visuales, inmersivos y textuales favorezcan la comprensión de la tarea, la toma de decisiones y el desempeño observable (Mayer, 2009). Bajo esta lógica, la RV no se justifica por su realismo gráfico absoluto, sino por su capacidad para representar situaciones con valor funcional para la competencia.

El aprendizaje experiencial aporta una segunda base conceptual. Desde esta perspectiva, el aprendizaje se fortalece cuando el estudiante participa en experiencias concretas, reflexiona sobre ellas, conceptualiza lo aprendido y transfiere ese aprendizaje

a nuevas situaciones (Kolb, 1984). En la arquitectura propuesta, la RV cumple la función de generar experiencias de desempeño situado, mientras que el *debriefing* y la mediación docente permiten transformar la experiencia en reflexión, análisis y aprendizaje transferible. La experiencia, por sí misma, no garantiza aprendizaje; requiere estructuración, criterios, retroalimentación y reflexión posterior.

La práctica deliberada ofrece una tercera base conceptual. La mejora del desempeño no depende de repetir tareas de manera acumulativa, sino de practicar con metas claras, estándares explícitos, retroalimentación específica y oportunidades de corrección (Ericsson *et al.*, 1993). Por ello, la propuesta organiza la intervención en ciclos de *briefing*, simulación, *feedback*, *debriefing* y mejora. Cada ciclo busca que el estudiante comprenda el estándar esperado, actúe en un entorno situado, reciba retroalimentación criterial, reflexione sobre su desempeño y realice ajustes progresivos.

La retroalimentación formativa constituye otro fundamento central. Para que el *feedback* contribuya al aprendizaje, debe responder con claridad a tres preguntas: hacia dónde debe avanzar el estudiante, cómo se está desempeñando actualmente y qué acciones concretas puede realizar para mejorar (Hattie & Timperley, 2007). Asimismo, la retroalimentación debe ser específica, comprensible, oportuna, vinculada a criterios y orientada a la acción, evitando comentarios generales o ambiguos que no ayuden al estudiante a modificar su desempeño (Shute, 2008; Wisniewski *et al.*, 2020). En esta propuesta, la IA se concibe como apoyo para ampliar la disponibilidad, especificidad y granularidad del *feedback*, siempre bajo supervisión docente y con criterios explícitos.

La evaluación criterial y el uso de rúbricas analíticas permiten traducir competencias complejas en criterios observables. Las rúbricas no solo cumplen una función calificadora, sino que también clarifican expectativas, orientan la práctica, estructuran la retroalimentación y favorecen procesos de autorregulación del estudiante (Andrade, 2000; Jonsson & Svingby, 2007; Brookhart, 2013; Panadero & Jonsson, 2013). En la arquitectura IA+RV, la rúbrica funciona como lenguaje común entre estudiante, docente, IA y evidencia de desempeño. Sin rúbricas, la retroalimentación puede volverse genérica; con rúbricas, el *feedback* puede vincularse a criterios concretos de mejora.

La autorregulación también es un eje conceptual de la propuesta. El estudiante no solo debe recibir retroalimentación, sino aprender a interpretarla, relacionarla con criterios de desempeño, identificar brechas, tomar decisiones de mejora y monitorear su propio progreso. La autorregulación implica establecer metas, seleccionar estrategias, observar el desempeño propio y ajustar la acción con base en evidencia (Zimmerman, 2000). Por ello, la arquitectura no busca sustituir el juicio del estudiante mediante tecnología, sino fortalecer su capacidad de comprender y mejorar su propio desempeño.

La realidad virtual aporta el componente de desempeño situado. Su valor pedagógico radica en permitir que el estudiante practique en escenarios que representan situaciones profesionales, restricciones contextuales, interacción, incertidumbre y necesidad de toma de decisiones. La RV puede favorecer presencia, autenticidad funcional y práctica segura, siempre que el diseño instruccional controle la carga cognitiva y mantenga la alineación con los objetivos de aprendizaje (Radianti *et al.*, 2020; Makransky & Petersen, 2021). En consecuencia, la RV no se incorpora como recurso espectacular, sino como entorno de práctica donde el estudiante puede actuar, decidir, comunicar y producir evidencia observable.

La inteligencia artificial aporta el componente de apoyo formativo y retroalimentación criterial. Los sistemas de tutoría inteligente y las herramientas de IA educativa han mostrado potencial para ofrecer orientación, pistas graduadas, retroalimentación personalizada y apoyo al aprendizaje cuando se diseñan con criterios pedagógicos claros (VanLehn, 2011; Kulik & Fletcher, 2016). Sin embargo, la IA generativa también plantea riesgos relacionados con dependencia excesiva, opacidad, sesgos, integridad académica y sustitución del juicio humano. Por ello, su uso en esta propuesta se delimita como tutor formativo supervisado, no como evaluador autónomo ni como generador de productos finales del estudiante (Kasneci *et al.*, 2023; Holmes *et al.*, 2023).

A partir de estos fundamentos, la arquitectura IA+RV se concibe como una integración funcional entre desempeño situado, retroalimentación criterial, rúbricas, mediación docente y autorregulación. La RV genera una situación de desempeño; la rúbrica convierte ese desempeño en criterios observables; la IA ayuda a transformar la

ejecución en retroalimentación accionable; el *debriefing* docente reconstruye la experiencia y orienta la transferencia; y el estudiante realiza ajustes progresivos mediante ciclos de práctica deliberada. En consecuencia, el aprendizaje no se atribuye a la tecnología como artefacto, sino a la calidad de la secuencia instruccional que convierte experiencia, *feedback* y reflexión en mejora observable.

Esta fundamentación permite delimitar el alcance conceptual de la propuesta. No se plantea una solución universal para cualquier competencia, disciplina o institución, sino un modelo transferible bajo condiciones explícitas: tareas auténticas, rúbricas analíticas, escenarios de desempeño situado, tutoría IA supervisada, mediación docente, protocolos de seguridad, criterios de equidad, trazabilidad y evaluación continua. Su valor radica en articular principios de aprendizaje, evaluación y tecnología educativa dentro de una arquitectura didáctica aplicable y verificable.

5.3 Objetivo general de la propuesta

El objetivo general de la propuesta es diseñar una arquitectura didáctica IA+RV replicable, auditable y escalable para fortalecer el desempeño observable en Comunicación y Solución de problemas en estudiantes de educación superior, mediante ciclos instruccionales de práctica deliberada, retroalimentación criterial, desempeño situado, mediación docente y evaluación basada en evidencias.

Este objetivo general expresa la finalidad instruccional de la propuesta. No se orienta únicamente a incorporar inteligencia artificial o realidad virtual en cursos universitarios, sino a estructurar una experiencia de aprendizaje en la que cada componente cumpla una función pedagógica precisa. La IA se integra como apoyo formativo para ampliar la retroalimentación criterial; la RV se incorpora como entorno de desempeño situado; las rúbricas analíticas definen criterios observables de calidad; el docente conserva la mediación y el control pedagógico; y el estudiante participa activamente en ciclos de mejora progresiva.

5.4 Objetivos específicos instruccionales de la propuesta

Los objetivos específicos de la propuesta se formulan desde una lógica instruccional, es decir, como orientaciones para estructurar experiencias de aprendizaje, desempeño, retroalimentación, reflexión, evaluación y mejora. Su función es traducir el objetivo general en acciones pedagógicas observables y coherentes con el desarrollo de Comunicación y Solución de problemas en educación superior.

1. Estructurar una secuencia instruccional mínima para integrar IA y RV en cursos regulares mediante ciclos de *briefing*, simulación, *feedback*, *debriefing* y mejora, asegurando que cada momento cumpla una función pedagógica definida dentro del aprendizaje por desempeño.
2. Organizar los componentes didácticos no negociables de la arquitectura IA+RV, incluyendo rúbricas analíticas, escenarios de desempeño situado, tutoría formativa asistida por IA, mediación docente, trazabilidad de implementación y criterios de seguridad para sostener una experiencia de aprendizaje coherente.
3. Diseñar actividades de aprendizaje situadas que permitan a los estudiantes practicar Comunicación y Solución de problemas en escenarios de RV, con tareas auténticas, criterios explícitos, restricciones relevantes y evidencia observable de desempeño.
4. Integrar estrategias de retroalimentación formativa criterial mediante IA y mediación docente, de modo que el estudiante pueda identificar brechas, revisar decisiones, ajustar argumentos y mejorar progresivamente su desempeño con base en rúbricas analíticas.
5. Establecer mecanismos instruccionales de acompañamiento, reflexión y autorregulación mediante *debriefing*, revisión de evidencias, contraste con criterios de desempeño y formulación de planes de mejora transferibles a nuevas tareas.
6. Definir criterios instruccionales de evaluación de efectividad orientados a valorar si la arquitectura conserva su secuencia pedagógica, mejora el desempeño

observable, mantiene fidelidad de implementación y favorece condiciones de transferencia, equidad y sostenibilidad en futuras aplicaciones.

Estos objetivos específicos mantienen coherencia con la naturaleza de la propuesta porque se centran en el diseño de condiciones pedagógicas para el aprendizaje, y no solo en aspectos operativos o administrativos de implementación. Asimismo, permiten vincular la propuesta con la estructura posterior del capítulo: representación lógica, fases de implementación, acciones, métodos, técnicas, instrumentos, recursos, valoración de pertinencia y evaluación de efectividad.

5.5 Representación teórica y práctica de la propuesta mediante esquema lógico

La arquitectura didáctica IA+RV requiere una representación teórica y práctica que permita visualizar cómo se articulan sus componentes centrales, su secuencia instruccional y sus resultados esperados. Esta representación cumple una doble función. En el plano teórico, sintetiza la lógica causal de la propuesta al mostrar cómo los insumos pedagógicos, tecnológicos y organizativos se transforman en procesos de aprendizaje. En el plano práctico, ofrece una guía para implementar la arquitectura en cursos reales, documentar su funcionamiento y evaluar sus resultados.

El modelo lógico de la propuesta se organiza en tres niveles: insumos, procesos y resultados. Los insumos corresponden a los recursos pedagógicos, tecnológicos, humanos y normativos necesarios para activar la arquitectura. Los procesos describen la secuencia didáctica mediante la cual dichos insumos se convierten en experiencias de práctica deliberada, retroalimentación criterial, reflexión y mejora progresiva. Los resultados representan los efectos esperados en el desempeño del estudiante, en su autorregulación y en la capacidad institucional para sostener el modelo.

Los insumos centrales incluyen docencia capacitada, escenarios de realidad virtual curados, rúbricas analíticas validadas, tutoría IA configurada con criterios explícitos e infraestructura con protocolos de seguridad, privacidad y gobernanza. Estos insumos no operan de manera aislada. Su articulación permite que la propuesta funcione como una arquitectura didáctica y no como una suma de herramientas. La docencia capacitada asegura mediación pedagógica; los escenarios de RV generan desempeño

situado; las rúbricas definen criterios observables; la IA aporta retroalimentación formativa; y la infraestructura con protocolos permite operar con seguridad, trazabilidad y continuidad.

Los procesos se organizan en un ciclo didáctico compuesto por cinco momentos. El primero es el *briefing*, donde se explicitan objetivos, criterios de evaluación, reglas de interacción, restricciones de la tarea y activación de saberes previos. El segundo es la ejecución en RV, donde el estudiante enfrenta una situación auténtica, toma decisiones, comunica argumentos y deja trazas básicas de desempeño. El tercero es el *feedback* formativo criterial, apoyado por IA y mediación docente, mediante el cual se identifican brechas, fortalezas y acciones de mejora. El cuarto es el *debriefing*, en el que se reconstruye la experiencia, se contrasta el desempeño con la rúbrica y se orienta la transferencia del aprendizaje. El quinto es el registro y aseguramiento de calidad, que documenta dosis, fidelidad, seguridad, calidad del *feedback* y condiciones de implementación.

Los resultados esperados se organizan en tres planos. En el corto plazo, la propuesta busca mejorar el desempeño observable en Comunicación y Solución de problemas, medido mediante rúbricas analíticas. En el mediano plazo, se espera fortalecer la autorregulación del estudiante, su capacidad para usar estratégicamente el *feedback* y su transferencia hacia tareas nuevas. En el plano institucional, se espera construir capacidad instalada mediante bancos de escenarios, rúbricas compartidas, registros comparables, protocolos de seguridad y ciclos de mejora continua.

La representación lógica también permite distinguir entre resultados de implementación y resultados de aprendizaje. Los resultados de implementación indican si la arquitectura fue aplicada con fidelidad suficiente: exposición real a la RV, cumplimiento del ciclo instruccional, uso explícito de rúbricas, calidad del *feedback*, aplicación del *debriefing*, registro de incidentes y disponibilidad de alternativas equivalentes. Los resultados de aprendizaje indican si el estudiante mejora efectivamente en las competencias objetivo. Esta distinción es necesaria porque una intervención puede estar disponible tecnológicamente, pero no haberse aplicado con suficiente fidelidad pedagógica para atribuirle resultados educativos.

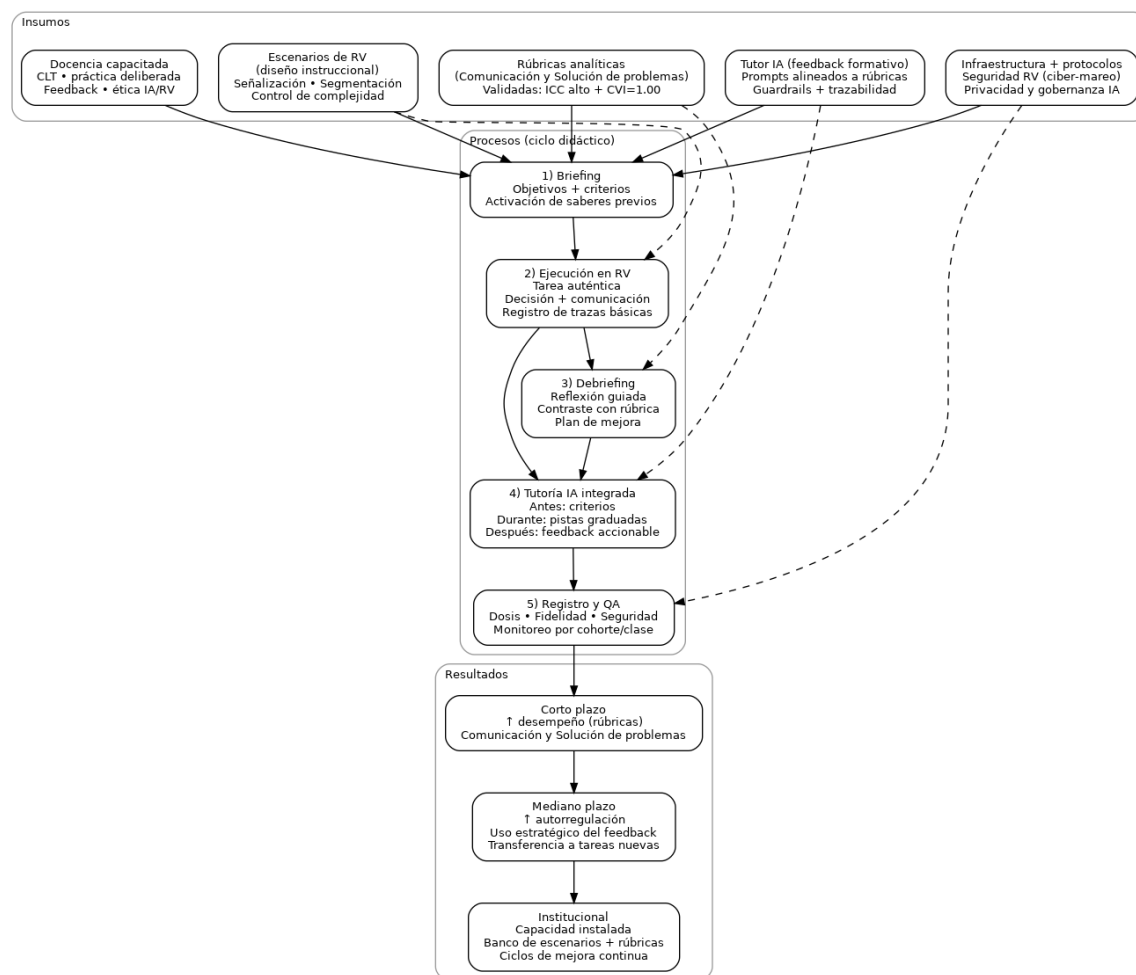
Asimismo, el modelo incorpora puntos de control o *quality gates*. Estos puntos permiten decidir si una implementación puede continuar, requiere ajuste o debe pausarse. Por ejemplo, no sería metodológicamente adecuado interpretar resultados de desempeño si no se completó el *debriefing*, si la IA produjo *feedback* genérico no vinculado a la rúbrica, si no se documentó la dosis real de uso o si hubo incidentes de seguridad sin registro adecuado. En consecuencia, el modelo lógico cumple una función analítica y de gobernanza: define qué debe ocurrir, cómo se verifica y qué evidencia es necesaria para sostener decisiones de mejora o escalamiento.

Desde esta perspectiva, la representación lógica no solo ordena la propuesta, sino que hace explícitas las relaciones entre condiciones de entrada, mecanismos pedagógicos y resultados esperados, permitiendo evaluar si la arquitectura conserva su coherencia interna al trasladarse a contextos distintos. Esta explicitación es clave porque facilita la replicabilidad del modelo, reduce ambigüedades en su implementación y ofrece a docentes e instituciones un marco común para interpretar evidencias, identificar cuellos de botella y decidir ajustes. De este modo, el esquema lógico actúa simultáneamente como mapa conceptual, guía operativa y herramienta de evaluación, al vincular la intención formativa de la propuesta con criterios observables de ejecución, calidad y mejora continua.

En síntesis, la Figura 5.1 permite cerrar la representación lógica de la arquitectura IA+RV y preparar su traducción operativa. Una vez establecida la relación entre insumos, procesos, resultados y condiciones de gobernanza, el siguiente paso consiste en definir cómo puede implementarse la propuesta de manera gradual, verificable y responsable en contextos reales de educación superior. Por ello, la sección siguiente presenta las fases de implementación necesarias para pasar del modelo lógico a una ruta institucional de aplicación y mejora continua.

Figura 5.1

Modelo lógico de la arquitectura IA+RV: insumos, procesos y resultados.



Nota. La figura sintetiza la lógica de funcionamiento de la propuesta. Los insumos, integrados por docencia capacitada, escenarios de RV, rúbricas analíticas, tutoría IA e infraestructura con protocolos, alimentan un ciclo didáctico compuesto por *briefing*, ejecución en RV, *feedback* formativo, *debriefing*, tutoría IA integrada y registro de calidad. Este proceso genera resultados esperados en tres niveles: mejora de corto plazo en el desempeño evaluado mediante rúbricas, fortalecimiento de la autorregulación y uso estratégico del *feedback* a mediano plazo, y capacidad institucional instalada para sostener escenarios, rúbricas y ciclos de mejora continua. Las líneas punteadas representan relaciones de soporte, trazabilidad y gobernanza que permiten mantener fidelidad, seguridad y comparabilidad durante la implementación.

5.6 Fases o etapas de implementación de la propuesta

La implementación de la arquitectura didáctica IA+RV se organiza en fases progresivas, con el propósito de asegurar que la propuesta pueda pasar de una aplicación inicial controlada a una adopción más amplia sin perder fidelidad pedagógica, seguridad, equidad ni trazabilidad. Esta secuencia responde al principio de que una innovación

educativa no debe escalar únicamente por disponibilidad tecnológica o interés institucional, sino por evidencia documentada de funcionamiento, calidad de implementación y resultados observables.

Desde la literatura sobre implementación, la fidelidad, la adaptación contextual y el monitoreo sistemático son condiciones necesarias para que una intervención conserve su integridad al trasladarse a contextos reales de operación (Dane & Schneider, 1998; Carroll *et al.*, 2007; Century *et al.*, 2010). Asimismo, los enfoques de mejora continua en educación enfatizan que el escalamiento debe apoyarse en ciclos iterativos de evidencia, aprendizaje institucional y ajuste progresivo, evitando convertir una innovación prometedora en una práctica superficial o difícil de sostener (Bryk *et al.*, 2015).

Bajo esta lógica, la propuesta se implementa mediante una ruta compuesta por un mínimo viable y tres fases sucesivas: consolidación en cursos ancla, expansión controlada a cursos afines e integración institucional. El mínimo viable establece las condiciones iniciales indispensables para iniciar la intervención. Las fases posteriores permiten estabilizar, transferir y sostener la arquitectura en condiciones reales de educación superior.

El mínimo viable corresponde al punto de partida de la implementación. Su función es asegurar que la arquitectura IA+RV cuente con los componentes esenciales antes de ser aplicada: rúbricas analíticas validadas o calibradas, escenarios de RV alineados con las competencias objetivo, tutoría IA configurada mediante criterios y *prompts* auditables, protocolo de seguridad para el uso de RV, guía docente de sesión y criterios mínimos de registro. Esta etapa evita que la propuesta se inicie como una experiencia tecnológica sin estándares pedagógicos suficientes.

La Fase 1, denominada consolidación en cursos ancla, tiene como finalidad estabilizar la arquitectura en cursos donde existan condiciones favorables de implementación. En esta fase se busca asegurar la aplicación completa del ciclo instruccional, que incluye *briefing*, simulación en RV, *feedback* criterial, *debriefing* y registro de evidencia. La prioridad no es ampliar cobertura, sino verificar que el diseño funcione con consistencia, que los docentes comprendan la secuencia, que los

estudiantes reciban criterios claros y que los indicadores mínimos de fidelidad y seguridad sean documentados.

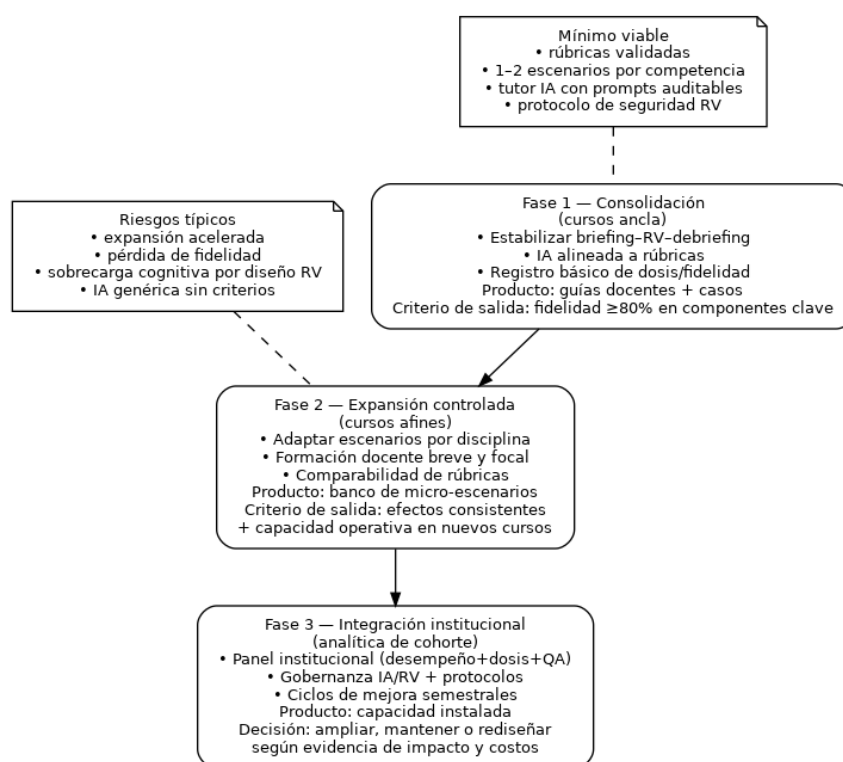
La Fase 2, denominada expansión controlada a cursos afines, tiene como propósito valorar la transferibilidad de la arquitectura a asignaturas, cohortes o contextos con competencias similares. En esta fase se permite adaptar escenarios, actividades y ejemplos al dominio disciplinar, pero no eliminar los componentes no negociables de la propuesta. La expansión debe mantenerse condicionada a la preservación de rúbricas, práctica situada, *feedback* criterial, mediación docente, trazabilidad, seguridad y alternativas equivalentes.

La Fase 3, denominada integración institucional, busca convertir la arquitectura IA+RV en una práctica recurrente, gobernada y mejorable. En esta etapa se formalizan bancos de escenarios, procesos de formación docente, protocolos de seguridad, criterios de uso responsable de IA, repositorios de evidencias, revisión semestral de indicadores y mecanismos de actualización. La finalidad es reducir la dependencia de esfuerzos docentes individuales y trasladar la continuidad del modelo hacia rutinas institucionales, roles definidos y procesos de mejora continua.

Esta secuencia gradual permite que la arquitectura IA+RV avance desde una aplicación inicial controlada hacia una adopción institucional sostenible, sin perder los componentes pedagógicos que le dan coherencia. Cada fase funciona como un punto de verificación antes de ampliar la cobertura, de modo que el escalamiento dependa de evidencia de funcionamiento, fidelidad y capacidad operativa, y no solo de disponibilidad tecnológica o interés institucional. La Figura 5.2 sintetiza esta ruta de implementación por fases, mostrando cómo el modelo progresa desde el mínimo viable hasta la consolidación, la expansión controlada y la integración institucional.

Figura 5.2

Estrategia de implementación por fases: mínimo viable, consolidación, expansión controlada e integración institucional.



Nota. La figura sintetiza la ruta gradual de implementación de la arquitectura IA+RV. El mínimo viable establece las condiciones básicas de operación: rúbricas, escenarios, tutoría IA, protocolo de seguridad y guías docentes. La Fase 1 consolida el modelo en cursos ancla; la Fase 2 permite expansión controlada hacia cursos afines; y la Fase 3 orienta la integración institucional mediante formación docente, gobernanza, analítica de cohorte y mejora continua. El avance entre fases se condiciona a evidencia de fidelidad, seguridad, equidad, calidad del *feedback* y resultados educacionalmente relevantes.

5.7 Acciones y actividades vinculadas a las fases de implementación

Las fases descritas requieren acciones específicas para convertirse en una ruta operativa de implementación. Por ello, esta sección traduce cada etapa de la propuesta en actividades concretas, responsables sugeridos, evidencias esperadas y criterios de avance. Esta organización permite que la arquitectura IA+RV no permanezca como una formulación conceptual, sino que pueda ser implementada, monitoreada y evaluada con criterios verificables.

La literatura sobre implementación educativa advierte que la fidelidad no depende únicamente de seguir un diseño original, sino de documentar qué componentes

se aplicaron, bajo qué condiciones, con qué adaptaciones y con qué evidencia de calidad (Dane & Schneider, 1998; Carroll et al., 2007; Century et al., 2010). En consecuencia, las acciones de esta propuesta se organizan para preservar sus componentes esenciales, permitir adaptaciones razonables al contexto y generar información útil para la mejora continua.

En el mínimo viable, las acciones se orientan a preparar las condiciones básicas de implementación: seleccionar competencias focales, confirmar rúbricas analíticas, elegir o diseñar escenarios de RV, configurar la tutoría IA, preparar guías docentes, establecer protocolos de seguridad y definir registros mínimos de dosis, fidelidad e incidentes. En la Fase 1, la prioridad es consolidar la arquitectura en cursos ancla mediante capacitación docente, inducción estudiantil, aplicación completa del ciclo instruccional, registro de evidencias, monitoreo de seguridad y revisión inicial de los productos generados por IA.

En la Fase 2, las acciones se dirigen a la expansión controlada hacia cursos afines, mediante adaptación de escenarios, formación de docentes adicionales, comparación de resultados, auditoría del *feedback*, revisión de incidentes y verificación de alternativas equivalentes. En la Fase 3, las actividades se orientan a la institucionalización, mediante bancos de escenarios, formación docente formalizada, integración de indicadores en reportes semestrales, responsables definidos, actualización de *prompts*, revisión de rúbricas, gobernanza de datos y decisiones de continuidad, expansión o rediseño. La Tabla 5.1 sintetiza las acciones, actividades, responsables, evidencias y criterios de avance por fase.

Tabla 5.1

Fases, acciones y criterios de avance para la implementación de la arquitectura IA+RV

Fase o etapa	Propósito	Acciones y actividades principales	Responsables sugeridos	Evidencia esperada	Criterio de avance
Mínimo viable	Establecer condiciones básicas antes de la implementación	Definir competencias focales, revisar rúbricas, seleccionar escenarios de RV, configurar tutoría IA, preparar guías docentes, establecer protocolo de seguridad y definir registros mínimos	Profesor responsable, coordinación académica y soporte tecnológico	Rúbricas disponibles, escenarios seleccionados, <i>prompts</i> auditables, guía docente, protocolo de seguridad y formato de registro	Iniciar implementación solo si los componentes esenciales están disponibles y documentados
Fase 1. Consolidación en cursos ancla	Estabilizar el modelo en condiciones controladas	Capacitar docentes, inducir estudiantes, aplicar ciclo <i>briefing</i> , simulación, <i>feedback</i> y <i>debriefing</i> , registrar dosis y fidelidad, documentar incidentes y ajustar guías	Profesor responsable, coordinación académica y soporte tecnológico	<i>Checklists</i> , registros de sesión, bitácora de incidentes, evidencias de <i>feedback</i> , rúbricas aplicadas y reporte de implementación	Continuar solo si existe fidelidad suficiente, ausencia de fallas críticas y evidencia de funcionamiento pedagógico
Fase 2. Expansión controlada a cursos afines	Probar transferibilidad sin perder fidelidad	Adaptar escenarios, formar docentes adicionales, monitorear calidad del <i>feedback</i> IA, comparar resultados entre cohortes, revisar seguridad, verificar alternativas equivalentes y documentar ajustes	Coordinación de programa, docentes participantes, soporte técnico y responsable de evaluación	Reportes por cohorte, auditoría de <i>feedback</i> , registros de uso, resultados PRE y POST, evidencias de adaptación y bitácoras	Avanzar si se conserva fidelidad, se mantiene seguridad y se observan resultados educacionalmente relevantes
Fase 3. Integración institucional	Convertir el modelo en práctica recurrente, gobernada y mejorable	Crear banco de escenarios, formalizar formación docente, integrar tablero o reporte de indicadores, revisar rúbricas, actualizar <i>prompts</i> , establecer gobernanza de datos y definir decisiones de continuidad	Dirección académica, innovación educativa, protección de datos, coordinación tecnológica y docentes líderes	Banco de escenarios, reportes semestrales, actas de revisión, lineamientos de gobernanza, actualización documentada de recursos	Institucionalizar, ampliar, mantener o rediseñar con base en evidencia acumulada

Nota. La tabla operacionaliza la implementación gradual de la arquitectura IA+RV. El avance entre fases no depende del número de estudiantes, cursos o dispositivos, sino de evidencia mínima de fidelidad pedagógica, seguridad, calidad del *feedback*, equidad, trazabilidad y resultados de aprendizaje.

Esta matriz permite traducir la propuesta en una ruta de acción verificable.

Además, facilita la toma de decisiones porque cada fase genera evidencia específica. Si

el mínimo viable no se cumple, la implementación debe aplazarse. Si la Fase 1 muestra baja fidelidad, no debe escalarse. Si la Fase 2 evidencia pérdida de calidad al adaptarse a nuevos cursos, debe reforzarse la capacitación o rediseñarse el escenario. Si la Fase 3 muestra capacidad instalada, la arquitectura puede integrarse a procesos institucionales de innovación educativa.

En conjunto, las acciones y actividades vinculadas a las fases permiten sostener una lógica de mejora continua. La implementación no se concibe como un evento único, sino como un proceso iterativo en el que cada ciclo produce información para ajustar la siguiente aplicación. Esta orientación es coherente con los enfoques de mejora educativa basados en evidencia, donde la innovación se fortalece mediante ciclos de prueba, revisión, aprendizaje institucional y ajuste progresivo (Bryk *et al.*, 2015).

5.8 Selección de métodos, técnicas e instrumentos para su aplicación

La aplicación de la arquitectura didáctica IA+RV requiere seleccionar métodos, técnicas e instrumentos congruentes con su finalidad instruccional. En esta propuesta, los métodos corresponden a la lógica pedagógica general que orienta la intervención; las técnicas son los procedimientos mediante los cuales se desarrolla la experiencia de aprendizaje; y los instrumentos son los recursos concretos que permiten aplicar, documentar, evaluar y mejorar la implementación. Esta distinción es relevante porque una propuesta doctoral no debe limitarse a describir actividades o tecnologías, sino precisar cómo se implementan, qué evidencias generan y bajo qué criterios pueden ser evaluadas.

El método central de la propuesta es la intervención didáctica basada en práctica deliberada con apoyo de IA y RV. Este método organiza la experiencia de aprendizaje en ciclos estructurados de preparación, ejecución, retroalimentación, reflexión y mejora. Su finalidad es producir desempeño observable y convertirlo en evidencia formativa. A diferencia de una clase expositiva o de una actividad tecnológica aislada, cada sesión debe incluir una tarea auténtica, criterios explícitos, instancia de práctica, retroalimentación accionable y cierre reflexivo.

Una segunda lógica metodológica es la evaluación criterial del desempeño. La propuesta no evalúa únicamente percepciones de satisfacción, intención de uso o asistencia a sesiones, sino evidencias de desempeño en Comunicación y Solución de problemas. Para ello, la rúbrica analítica opera como instrumento principal de valoración, tanto para evaluación formativa como para seguimiento de progreso. Las rúbricas permiten clarificar expectativas, estructurar la retroalimentación, orientar la práctica y favorecer procesos de autorregulación del estudiante (Andrade, 2000; Jonsson & Svingby, 2007; Brookhart, 2013; Panadero & Jonsson, 2013).

Una tercera lógica metodológica es el aseguramiento de calidad de implementación. La arquitectura solo puede interpretarse correctamente si existe evidencia de que fue implementada conforme al diseño previsto. Por ello, se incorporan técnicas de monitoreo de fidelidad, registro de dosis, revisión de *outputs* de IA, bitácora de incidentes y verificación de seguridad. La literatura sobre implementación educativa subraya que la fidelidad, la exposición, la calidad de entrega y la adaptación contextual deben documentarse para interpretar adecuadamente los efectos de una intervención (Dane & Schneider, 1998; Carroll et al., 2007; Century et al., 2010).

Las técnicas específicas de implementación incluyen *briefing* inicial, ejecución de escenarios de RV, interacción guiada con IA, *feedback* formativo, *debriefing* estructurado, revisión de rúbricas, observación docente, registro de *logs*, auditoría muestral del *feedback* IA y revisión periódica de incidentes. Cada técnica cumple una función dentro del ciclo: el *briefing* prepara la tarea; la RV genera la situación de desempeño; la IA ofrece apoyo formativo; el *feedback* orienta la mejora; el *debriefing* promueve reflexión y transferencia; la rúbrica ancla la retroalimentación; y los registros permiten verificar fidelidad.

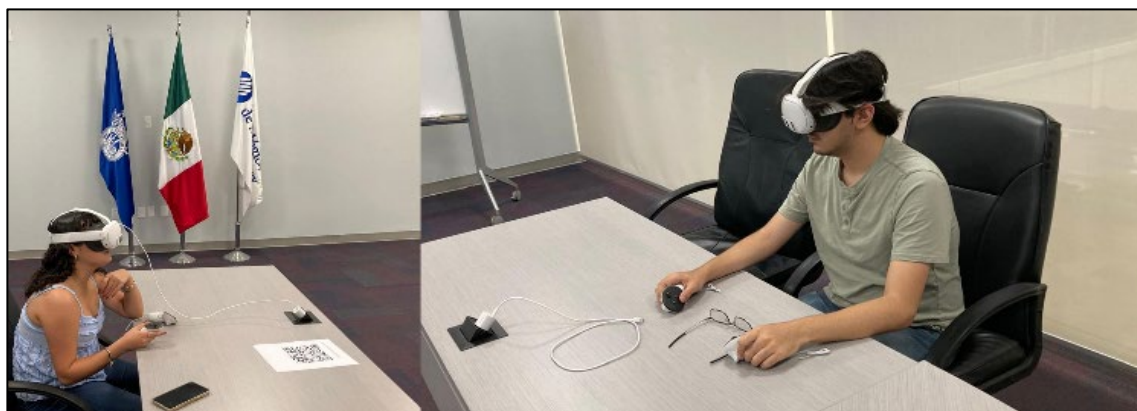
El *debriefing* ocupa un lugar especialmente importante porque convierte la experiencia en aprendizaje reflexivo. La simulación, por sí sola, no garantiza transferencia; requiere un espacio posterior donde el estudiante reconstruya sus decisiones, contraste su desempeño con criterios explícitos, identifique brechas y formule acciones de mejora. La literatura sobre simulación educativa ha señalado que el

debriefing es una de las condiciones centrales para transformar la experiencia en aprendizaje significativo (Fanning & Gaba, 2007; Rudolph et al., 2006).

La inteligencia artificial se integra como técnica de apoyo formativo y no como mecanismo autónomo de evaluación de alto impacto. Su función es ofrecer pistas, preguntas orientadoras, contraste de supuestos y retroalimentación vinculada a criterios de la rúbrica, bajo supervisión docente y criterios de transparencia para evitar dependencia, sesgos u opacidad (VanLehn, 2011; Kulik & Fletcher, 2016; Holmes et al., 2023; Kasneci et al., 2023). La realidad virtual, por su parte, se utiliza como técnica de simulación situada: no añade inmersión por sí misma, sino que representa tareas con autenticidad funcional, restricciones, interacción, incertidumbre y necesidad de decisión. En conjunto, IA y RV permiten que el estudiante practique competencias complejas en un entorno seguro, repetible, evaluable y alineado con objetivos de aprendizaje.

Figura 5.3

Escenario de uso de RV como entorno de desempeño en la arquitectura didáctica IA+RV.



Nota. La fotografía muestra una sesión real de implementación en la que estudiantes utilizan realidad virtual como entorno de práctica situada. Su función dentro del capítulo es ilustrar el componente operativo de RV en la arquitectura IA+RV: generar escenarios de desempeño en los que el estudiante enfrenta una tarea, toma decisiones, produce evidencia observable y posteriormente recibe retroalimentación criterial. La imagen no representa la totalidad del modelo, sino uno de sus componentes instruccionales centrales.

Los instrumentos asociados a la propuesta incluyen rúbricas analíticas de Comunicación y Solución de problemas, guías docentes de sesión, prompts o plantillas de tutoría IA, checklists de briefing y debriefing, registros de dosificación, bitácoras de incidentes, formatos de observación, evidencias de feedback y reportes de desempeño

PRE–POST. Estos instrumentos permiten que la arquitectura sea aplicable por otros docentes y comparable entre cursos, siempre que se preserven sus componentes núcleo. La Tabla 5.2 sintetiza la correspondencia entre componentes, métodos, técnicas, instrumentos y evidencias esperadas.

Tabla 5.2

Matriz de componentes, métodos, técnicas e instrumentos de aplicación de la arquitectura IA+RV

Componente de la arquitectura	Método asociado	Técnica de aplicación	Instrumento principal	Evidencia esperada
Rúbricas analíticas	Evaluación criterial del desempeño	Aplicación de criterios por competencia	Rúbricas de Comunicación y Solución de problemas	Puntajes por criterio, retroalimentación específica, progreso PRE–POST
Ciclo <i>briefing</i>, simulación y <i>debriefing</i>	Práctica deliberada estructurada	Preparación, ejecución, reflexión y ajuste	Guía docente de sesión y <i>checklist</i> de ciclo	Evidencia de cumplimiento del ciclo instruccional
Escenarios de RV	Simulación situada	Resolución de tareas en entorno inmersivo	Guía de escenario, <i>logs</i> de uso y observación docente	Desempeño situado, decisiones observables, exposición documentada
Tutoría IA	Retroalimentación formativa criterial	Preguntas diagnósticas, pistas, contraste de supuestos y sugerencias	<i>Prompts</i> , plantillas de <i>feedback</i> y muestras auditables	<i>Feedback</i> específico, accionable y vinculado a rúbrica
Trazabilidad y fidelidad	Aseguramiento de calidad de implementación	Registro, monitoreo y revisión periódica	Bitácora, <i>checklist</i> , tablero de QA y registros de incidentes	Dosis, adherencia, seguridad, comparabilidad y ajustes documentados

Nota. La tabla vincula cada componente de la arquitectura con los métodos, técnicas e instrumentos necesarios para su aplicación. Su finalidad es hacer visible que la propuesta no depende de la adopción aislada de tecnología, sino de una secuencia pedagógica documentada y evaluable.

Esta matriz cumple una función práctica dentro de la propuesta: permite que docentes, coordinadores académicos o unidades de innovación educativa identifiquen qué deben conservar para implementar el modelo sin diluirlo. También facilita la evaluación posterior, ya que cada componente genera una evidencia verificable. En consecuencia, la aplicación de la arquitectura no se define por la simple disponibilidad de IA o RV, sino por la presencia articulada de componentes, procedimientos e instrumentos.

En síntesis, la selección de métodos, técnicas e instrumentos permite convertir la arquitectura IA+RV en una intervención aplicable, documentable y evaluable. Los métodos orientan la lógica pedagógica, las técnicas organizan la experiencia de aprendizaje y los instrumentos permiten registrar, valorar y mejorar la implementación. Esta articulación es indispensable para conservar la integridad instruccional de la propuesta.

5.9 Recursos para el funcionamiento de la propuesta

El funcionamiento de la arquitectura didáctica IA+RV requiere recursos pedagógicos, humanos, tecnológicos, organizativos y financieros. Estos recursos deben concebirse como condiciones de operación del modelo y no como elementos accesorios. Una propuesta de integración tecnológica en educación superior solo puede sostenerse si existe correspondencia entre su ambición pedagógica, la capacidad institucional disponible y los mecanismos de gobernanza necesarios para operarla con calidad, seguridad y continuidad.

Desde la perspectiva de mejora continua, las innovaciones educativas requieren recursos, rutinas, roles definidos y mecanismos de aprendizaje institucional que permitan sostener la calidad más allá de una implementación inicial (Bryk *et al.*, 2015). Asimismo, las orientaciones internacionales sobre IA en educación enfatizan la necesidad de establecer condiciones de gobernanza, formación docente, protección de datos, equidad de acceso y supervisión humana para integrar tecnologías emergentes de manera responsable (UNESCO, 2023; Varsik & Vosberg, 2024).

Para organizar estos recursos, la propuesta contempla un sistema mínimo viable compuesto por tres subsistemas interdependientes: el subsistema didáctico y evaluativo, el subsistema tecnológico, y el subsistema organizativo y de gobernanza. Estos subsistemas permiten pasar de una experiencia piloto a una propuesta institucionalmente operable.

El subsistema didáctico y evaluativo integra las rúbricas analíticas, las guías docentes, los escenarios de aprendizaje, los criterios de retroalimentación, los instrumentos de evaluación y las evidencias de desempeño. Su función es asegurar que

la propuesta conserve una finalidad pedagógica clara. Este subsistema responde a la pregunta sobre qué se enseña, qué se practica, cómo se retroalimenta y cómo se evalúa. En ausencia de este subsistema, la arquitectura correría el riesgo de convertirse en una experiencia tecnológica atractiva, pero desconectada de competencias y criterios.

El subsistema tecnológico incluye la plataforma de RV, los visores, los espacios de práctica, el motor o herramienta de IA, el LMS institucional, los registros de interacción y, cuando sea posible, un tablero básico de monitoreo. Su función es habilitar la experiencia, registrar trazas mínimas y facilitar la retroalimentación. Este subsistema debe operar bajo criterios de confiabilidad, accesibilidad, soporte técnico, protección de datos y compatibilidad con las condiciones reales de aula. La tecnología debe ser suficientemente estable para no interrumpir la experiencia y suficientemente flexible para adaptarse a diferentes cursos o tareas.

El subsistema organizativo y de gobernanza contempla formación docente, protocolos de seguridad, reglas de uso de IA, lineamientos de privacidad, roles de responsabilidad, mecanismos de revisión periódica y criterios para la toma de decisiones. Su función es sostener la propuesta más allá de la iniciativa individual de un profesor. La arquitectura solo puede escalar si existe capacidad institucional para formar docentes, atender incidentes, revisar indicadores, actualizar escenarios, ajustar *prompts* y mantener consistencia entre cohortes. En este sentido, la definición de un sistema mínimo viable permite reducir la dependencia de esfuerzos docentes individuales, al desplazar la continuidad de la innovación hacia recursos documentados, rutinas organizativas, roles definidos y mecanismos institucionales de mejora continua (Bryk et al., 2015; UNESCO, 2023; Varsik & Vosberg, 2024).

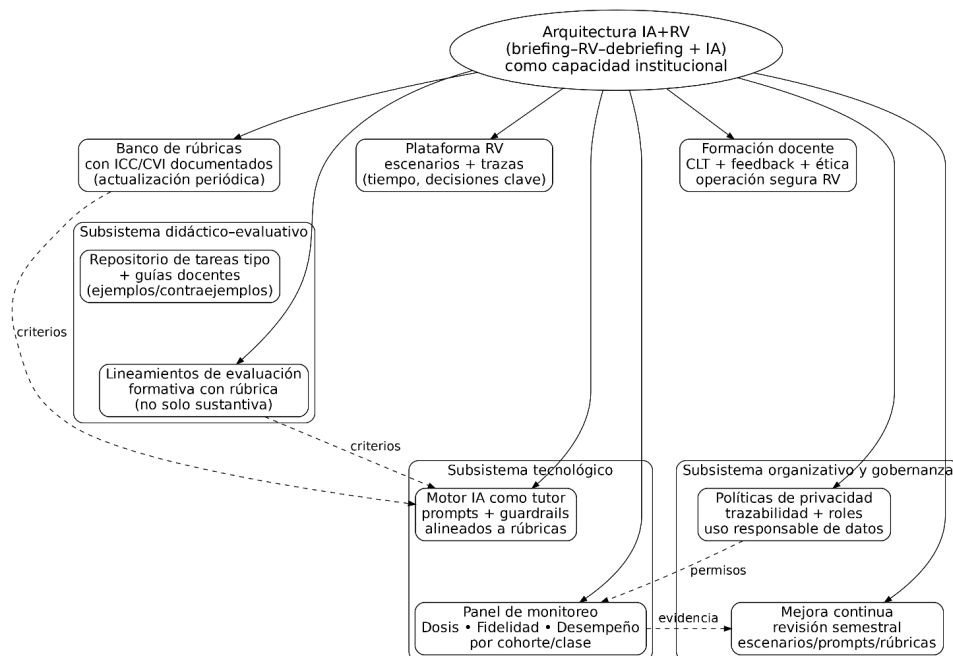
Adicionalmente, estos recursos deben entenderse como un sistema de soporte interdependiente: si uno de los subsistemas falla, la arquitectura pierde estabilidad operativa y capacidad de escalamiento. Por ejemplo, una plataforma tecnológica robusta no garantiza resultados si no existen rúbricas claras, formación docente o protocolos de uso responsable; del mismo modo, una propuesta pedagógicamente sólida puede volverse inviable si carece de soporte técnico, resguardo de datos o mecanismos institucionales de seguimiento. Por ello, la sostenibilidad de la arquitectura IA+RV

depende de la alineación entre recursos, gobernanza y mejora continua, más que de la simple disponibilidad de herramientas tecnológicas.

La Figura 5.4 sintetiza esta articulación entre recursos didácticos, tecnológicos y organizativos, mostrando cómo su interacción permite sostener la operación, el seguimiento y el escalamiento responsable de la arquitectura IA+RV.

Figura 5.4

Sistema mínimo viable para sostener y escalar la arquitectura IA+RV.



Nota. La figura representa tres subsistemas interdependientes: didáctico y evaluativo, tecnológico, y organizativo y de gobernanza. Las relaciones entre subsistemas muestran que la arquitectura IA+RV no depende de una herramienta aislada, sino de la articulación entre diseño pedagógico, infraestructura tecnológica, formación docente, protocolos de seguridad, gobernanza de datos y mecanismos de mejora continua.

A partir de este sistema, los recursos requeridos pueden organizarse en cinco categorías. Los recursos pedagógicos incluyen rúbricas analíticas, guías de sesión, bancos de escenarios, criterios de *feedback*, plantillas de *prompts*, instrumentos de observación y formatos de *debriefing*. Estos recursos aseguran que la implementación pueda ser replicada por distintos docentes sin depender exclusivamente de conocimiento tácito. También facilitan la formación docente y la comparación entre cohortes.

Los recursos humanos incluyen docentes capacitados, personal de soporte tecnológico, responsables de evaluación, especialistas en innovación educativa y, cuando sea necesario, apoyo en privacidad o gobernanza de datos. La formación docente es especialmente importante porque el modelo requiere que el profesor interprete rúbricas, conduzca *debriefings*, supervise el uso de IA, atienda incidentes y preserve el sentido pedagógico de la experiencia. La tecnología no reduce la importancia del docente; más bien aumenta la necesidad de mediación experta.

Los recursos tecnológicos incluyen visores de RV, espacios adecuados de práctica, conectividad, dispositivos de apoyo, plataforma de RV, herramienta de IA, LMS, repositorio de evidencias y registros de uso. La infraestructura debe garantizar continuidad operativa, seguridad física, acceso suficiente y resguardo mínimo de información. No se requiere necesariamente una infraestructura compleja desde el inicio, pero sí una infraestructura confiable y alineada al tamaño de la implementación.

Los recursos organizativos incluyen protocolos de seguridad, lineamientos de uso de IA, reglas de privacidad, calendario de implementación, criterios de asignación de espacios, soporte técnico programado, revisión de indicadores y mecanismos de escalamiento. Estos recursos permiten que el modelo funcione de manera institucional y no solo como iniciativa aislada. También facilitan la auditabilidad y la rendición de cuentas.

Los recursos financieros se relacionan con adquisición o mantenimiento de visores, licencias de plataformas, capacitación docente, soporte técnico, actualización de escenarios y posible desarrollo de tableros de seguimiento. Para favorecer la sostenibilidad, se recomienda una estrategia gradual: iniciar con cursos ancla y recursos existentes, documentar resultados, justificar inversión incremental y escalar solo cuando existan evidencias de fidelidad, aceptabilidad y efectividad.

La Tabla 5.3 sintetiza los recursos mínimos requeridos para el funcionamiento y sostenibilidad de la arquitectura IA+RV.

Tabla 5.3

Recursos mínimos para el funcionamiento y sostenibilidad de la arquitectura IA+RV

Tipo de recurso	Elementos requeridos	Función dentro de la propuesta	Condición mínima de sostenibilidad
Pedagógicos	Rúbricas, guías, <i>prompts</i> , escenarios, formatos de <i>debriefing</i>	Asegurar alineación entre competencia, práctica, <i>feedback</i> y evaluación	Materiales documentados y reutilizables
Humanos	Docentes capacitados, soporte técnico, responsable de evaluación	Mediar la experiencia, resolver incidentes y preservar calidad	Formación inicial y acompañamiento periódico
Tecnológicos	Visores RV, plataforma, IA, LMS, conectividad, repositorio	Habilitar práctica situada, <i>feedback</i> y trazabilidad	Equipos funcionales, soporte y acceso equitativo
Organizativos	Protocolos, calendario, roles, reglas de privacidad, revisión de indicadores	Coordinar implementación y gobernanza	Responsables definidos y revisión semestral
Financieros	Licencias, mantenimiento, capacitación, actualización de escenarios	Sostener operación y mejora continua	Inversión gradual vinculada a evidencia de uso y resultados

Nota. Los recursos se presentan como condiciones mínimas de funcionamiento. La propuesta puede adaptarse a distintos niveles de capacidad institucional, siempre que preserve sus componentes esenciales y documente las condiciones bajo las cuales se implementa.

La sostenibilidad de la propuesta no depende únicamente de contar con tecnología disponible, sino de asegurar que los recursos se integren en una rutina académica viable. Por ello, se recomienda que la institución adopte una lógica de mejora progresiva: comenzar con un número limitado de cursos, formar docentes, documentar resultados, revisar incidentes, ajustar escenarios y solo después ampliar cobertura. Esta estrategia permite reducir costos iniciales, evitar inversiones prematuras y demostrar valor pedagógico antes de escalar.

En síntesis, los recursos para el funcionamiento de la propuesta no son únicamente materiales o tecnológicos. Incluyen condiciones pedagógicas, capacidades humanas, infraestructura, gobernanza, financiamiento y mecanismos de mejora continua. Su articulación permite que la arquitectura IA+RV pueda operar con fidelidad, seguridad y sostenibilidad, evitando que dependa exclusivamente de la iniciativa individual de un docente o de la disponibilidad ocasional de herramientas tecnológicas.

5.10 Valoración de la pertinencia de la propuesta

La valoración de la pertinencia de la arquitectura didáctica IA+RV debe realizarse a partir de su coherencia con el problema de investigación, su fundamentación pedagógica, su viabilidad de aplicación y su contribución a la formación de competencias transversales en educación superior. Una propuesta es pertinente cuando responde a una necesidad formativa real, se vincula con evidencia y teoría relevante, puede implementarse bajo condiciones razonables y ofrece criterios para valorar sus resultados.

La propuesta es pertinente porque responde directamente al problema central identificado en la investigación: la necesidad de fortalecer competencias transversales complejas mediante experiencias auténticas, evaluables y pedagógicamente estructuradas. Comunicación y Solución de problemas no se desarrollan de manera suficiente mediante exposición conceptual, ejercicios aislados, satisfacción estudiantil o uso instrumental de tecnología; por el contrario, requieren desempeño observable, criterios explícitos, retroalimentación útil, reflexión guiada y oportunidades de transferencia (Hattie & Timperley, 2007; Shute, 2008; Brookhart, 2013). En consecuencia, la arquitectura IA+RV se plantea como una alternativa formativa para organizar práctica situada, *feedback* criterial y mediación docente en ciclos verificables de mejora progresiva.

Desde esta perspectiva, la pertinencia pedagógica de la propuesta radica en que articula competencia, tarea, retroalimentación y evaluación. La RV permite generar entornos de desempeño situado donde el estudiante practica bajo condiciones de mayor autenticidad funcional. La IA permite ampliar la disponibilidad y granularidad del *feedback* formativo, siempre que opere como tutor criterial y no como sustituto del juicio docente. Las rúbricas analíticas permiten traducir competencias complejas en criterios observables, mientras que el ciclo de *briefing*, simulación y *debriefing* convierte la experiencia en aprendizaje reflexivo.

La pertinencia metodológica se expresa en que la propuesta no depende de una noción general de innovación, sino de componentes, instrumentos, indicadores y criterios de trazabilidad. La arquitectura define qué debe ocurrir, cómo debe

documentarse y con qué evidencia puede interpretarse. Esta característica es importante porque permite distinguir entre una implementación completa y una adopción superficial de IA o RV.

La pertinencia institucional se manifiesta en que la propuesta puede implementarse de manera gradual mediante un mínimo viable, cursos ancla, expansión controlada e integración institucional. Esta ruta permite evitar el escalamiento prematuro y facilita que la institución tome decisiones con base en evidencia de fidelidad, seguridad, equidad y resultados. La propuesta no exige una transformación institucional inmediata, sino una adopción progresiva con criterios explícitos.

La pertinencia ética y de gobernanza radica en que la adopción de IA y RV exige reglas claras de privacidad, transparencia, equidad, seguridad y supervisión humana. En línea con orientaciones internacionales sobre IA educativa y gestión de riesgos tecnológicos, la propuesta incorpora salvaguardas de protección de datos, uso responsable de IA, alternativas equivalentes, seguridad en RV y revisión periódica de indicadores (UNESCO, 2021; UNESCO, 2023; NIST, 2023; Varsik & Vosberg, 2024). La Tabla 5.4 sintetiza la valoración de pertinencia de la propuesta.

Tabla 5.4
Valoración de pertinencia de la arquitectura IA+RV

Dimensión de pertinencia	Justificación dentro de la propuesta	Evidencia o soporte esperado
Pertinencia pedagógica	Responde a la necesidad de fortalecer Comunicación y Solución de problemas mediante práctica situada, rúbricas y <i>feedback</i> formativo.	Mejora observable en rúbricas, <i>debriefing</i> documentado y evidencia de progreso.
Pertinencia metodológica	Traduce una intervención educativa en componentes, instrumentos, indicadores y criterios de trazabilidad.	Registros de dosis, fidelidad, calidad del <i>feedback</i> y seguridad.
Pertinencia institucional	Permite adopción gradual mediante mínimo viable, cursos ancla, formación docente y gobernanza.	Reportes por cohorte, protocolos, revisión semestral e indicadores comparables.
Pertinencia tecnológica	Evita el uso instrumental de IA y RV al integrarlas en un mecanismo instruccional explícito.	IA como tutor criterial y RV como entorno de desempeño situado.
Pertinencia ética y de gobernanza	Incorpora privacidad, equidad, supervisión humana y seguridad en RV como condiciones de implementación.	Avisos, protocolos, alternativas equivalentes, bitácoras e intervención docente.

Nota. La pertinencia de la propuesta no se deriva de la novedad tecnológica de IA o RV, sino de su integración en una arquitectura pedagógica con criterios de aprendizaje, implementación, seguridad y evaluación.

5.10.1 Factibilidad técnica, operativa e institucional

La factibilidad de la propuesta debe analizarse en tres niveles: técnico, operativo e institucional. La factibilidad técnica se refiere a la disponibilidad y estabilidad mínima de herramientas, plataformas, dispositivos y registros. La factibilidad operativa se refiere a la posibilidad de implementar la arquitectura en cursos regulares sin generar cargas excesivas o interrupciones sistemáticas. La factibilidad institucional se refiere a la capacidad de sostener el modelo mediante formación docente, soporte, protocolos, gobernanza y mejora continua.

La factibilidad técnica de la arquitectura IA+RV no exige una infraestructura compleja desde el inicio. El modelo puede comenzar con un sistema mínimo viable compuesto por rúbricas validadas o calibradas, escenarios de RV alineados con competencias, una herramienta de IA configurada con *prompts* auditables, visores disponibles, guías docentes y un protocolo de seguridad. La condición crítica no es la sofisticación tecnológica, sino la confiabilidad de los recursos y su alineación con el ciclo didáctico.

La factibilidad operativa depende de que las sesiones puedan integrarse a la dinámica real del curso. Para ello, la propuesta requiere secuencias manejables, tiempos claros, roles definidos, *checklists* de aplicación, criterios de contingencia y alternativas equivalentes para estudiantes que no puedan usar RV. También requiere que el *feedback* asistido por IA se incorpore en momentos específicos del ciclo y no como actividad adicional desarticulada.

La factibilidad institucional depende de la capacidad de pasar de una implementación dependiente del profesor a una práctica con soporte organizativo. Esto exige formación docente, soporte técnico, lineamientos de uso de IA, criterios de privacidad, revisión de incidentes, actualización de escenarios y revisión periódica de indicadores. En consecuencia, la propuesta es factible si se implementa de manera gradual, iniciando en cursos ancla y avanzando hacia cursos afines solo cuando exista evidencia de fidelidad, aceptabilidad operativa y resultados educacionalmente relevantes. La Tabla 5.5 presenta una matriz de factibilidad de la propuesta.

Tabla 5.5

Matriz de factibilidad técnica, operativa e institucional de la arquitectura IA+RV

Nivel de factibilidad	Condición requerida	Riesgo principal	Estrategia de mitigación
Técnica	Visores funcionales, plataforma RV, herramienta IA, conectividad, LMS y registros básicos.	Fallas técnicas, falta de acceso o interrupción de sesión.	Prueba previa, soporte técnico, escenarios alternos y bitácora.
Operativa	Guías docentes, tiempos definidos, <i>checklists</i> , rúbricas y secuencia aplicable en clase.	Sobrecarga docente o implementación incompleta.	Ciclo mínimo viable, instrumentos breves y roles claros.
Pedagógica	Alineación entre competencia, tarea, rúbrica, <i>feedback</i> y <i>debriefing</i> .	Uso superficial de IA o RV.	Componentes no negociables y revisión de calidad.
Ética y seguridad	Protocolos de RV, privacidad, minimización de datos y supervisión humana.	Riesgos físicos, sesgos o uso indebido de información.	Protocolos, aviso, alternativas equivalentes y gobernanza.
Institucional	Formación docente, soporte, revisión de indicadores y mejora continua.	Dependencia de un solo docente o falta de sostenibilidad.	Cursos ancla, capacitación, reportes semestrales y escalamiento gradual.
Financiera	Recursos para licencias, mantenimiento, capacitación y actualización.	Inversión prematura sin evidencia de impacto.	Implementación gradual condicionada a resultados.

Nota. La factibilidad de la propuesta debe entenderse como capacidad de implementación con fidelidad, seguridad y sostenibilidad. No basta contar con tecnología disponible; se requiere un sistema pedagógico, operativo y organizativo que preserve la integridad del modelo.

5.10.2 Criterios de adopción y escalamiento responsable

La adopción de la arquitectura IA+RV debe realizarse mediante criterios explícitos que eviten dos errores frecuentes en innovación educativa: la expansión prematura y la dilución pedagógica. La expansión prematura ocurre cuando una práctica se escala antes de contar con evidencia suficiente de funcionamiento, seguridad y factibilidad. La dilución pedagógica ocurre cuando, al ampliar la cobertura, se conservan las herramientas, pero se pierden los componentes que explican el valor formativo del modelo.

Para prevenir estos riesgos, la propuesta establece que el escalamiento solo debe considerarse cuando se cumplan cinco condiciones. La primera es fidelidad mínima al ciclo de *briefing*, simulación, *feedback* y *debriefing*. La segunda es uso explícito de rúbricas para orientar práctica, retroalimentación y evaluación. La tercera es tutoría IA configurada con criterios auditables y supervisión docente. La cuarta es operación

segura de RV, con protocolos y alternativas equivalentes. La quinta es existencia de registros suficientes de dosis, fidelidad, incidentes, calidad del *feedback* y resultados.

Estos criterios ya fueron operacionalizados en el capítulo mediante indicadores, quality gates, matrices de recursos, fases de implementación y mecanismos de gobernanza. Por ello, la propuesta no requiere condiciones externas no previstas para justificar su adopción; requiere aplicar de manera consistente los estándares definidos. En otras palabras, el escalamiento responsable no depende de ampliar cobertura, sino de conservar integridad pedagógica y evidencia de funcionamiento. La Tabla 5.6 sintetiza los criterios de adopción y escalamiento responsable.

Tabla 5.6

Criterios de adopción y escalamiento responsable de la arquitectura IA+RV

Criterio	Evidencia mínima requerida	Decisión posible
Fidelidad instruccional	Cumplimiento documentado del ciclo <i>briefing</i> , RV, <i>feedback</i> , <i>debriefing</i> , IA y registro QA.	Avanzar, ajustar o repetir fase.
Calidad del <i>feedback</i> IA	Muestras auditadas de <i>feedback</i> específico, accionable y vinculado a rúbrica.	Mantener, ajustar <i>prompts</i> o limitar uso.
Seguridad RV	Protocolos aplicados, incidentes registrados y alternativas equivalentes disponibles.	Continuar, adaptar o suspender temporalmente.
Equidad de acceso	Evidencia de acceso efectivo o alternativa sin penalización.	Escalar solo si no amplifica brechas.
Calidad de medición	Rúbricas aplicadas con consistencia y criterios claros.	Interpretar resultados o recalibrar instrumentos.
Resultado educativo	Mejora PRE y POST educacionalmente relevante en competencias objetivo.	Ampliar, mantener o rediseñar.
Sostenibilidad operativa	Formación docente, soporte técnico y revisión semestral.	Institucionalizar o mantener como piloto.

Nota. Los criterios de adopción buscan asegurar que el escalamiento de la propuesta ocurra con fidelidad pedagógica, seguridad, equidad y evidencia de impacto. La expansión no debe interpretarse como aumento automático de cobertura, sino como una decisión condicionada por calidad y resultados.

5.10.3 Validación complementaria mediante Delphi modificado de dos rondas

Con el fin de fortalecer la solidez académica y la viabilidad de adopción institucional de la propuesta, se realizó una validación complementaria mediante un Delphi modificado de dos rondas con panel de expertos. Esta técnica resulta útil para obtener juicio estructurado, retroalimentación iterativa y convergencia progresiva en problemas complejos de diseño o validación profesional (Dalkey & Helmer, 1963;

Linstone & Turoff, 1975; Hasson *et al.*, 2000; Keeney *et al.*, 2011). El propósito de este procedimiento no fue volver a comprobar el efecto empírico de la arquitectura IA+RV, sino valorar la propuesta en términos de claridad conceptual, coherencia con la evidencia empírica, pertinencia pedagógica, factibilidad técnica y operativa, viabilidad institucional, condiciones de gobernanza y originalidad como aporte doctoral.

La validación por Delphi se concibió como una capa complementaria de refinamiento experto orientada al diseño y a la aplicabilidad del modelo. El proceso se organizó en dos rondas. En la primera, los expertos valoraron la propuesta y emitieron observaciones cualitativas para su mejora. En la segunda, se sometió a revaloración una versión ajustada de la propuesta, acompañada de una síntesis agregada de la ronda previa, con el propósito de estimar convergencia y suficiencia del refinamiento realizado.

Por razones de economía expositiva, el cuerpo del capítulo presenta la caracterización sintética del panel, los resultados agregados por ronda, los ajustes principales y la síntesis interpretativa. Los instrumentos completos, la carta de invitación, el resumen ejecutivo enviado al panel, los formatos de valoración, las matrices extendidas y la documentación detallada del proceso se presentan en el Apéndice C.

5.10.3.1 Propósito y panel de expertos

El panel se integró mediante muestreo intencional por criterio, considerando perfiles con experiencia en al menos una de las siguientes áreas: innovación educativa, diseño instruccional, evaluación por competencias y rúbricas, inteligencia artificial en educación, realidad virtual o simulación educativa, gobernanza tecnológica y formación docente en educación superior. Como criterios de inclusión se consideraron formación académica avanzada, trayectoria relevante y disponibilidad para participar en dos rondas sucesivas. La Tabla 5.7 presenta la caracterización sintética del panel participante.

Tabla 5.7*Caracterización del panel participante en el Delphi modificado*

Código del experto	Grado académico	Área principal de expertise	Años de experiencia	Ámbito institucional principal	Aporte esperado a la validación
E1	Doctorado	Diseño instruccional	18	Universidad	Valorar la coherencia del diseño pedagógico y la secuencia formativa.
E2	Doctorado	Evaluación por competencias	15	Universidad	Valorar congruencia entre resultados, criterios de desempeño y evaluación.
E3	Doctorado	Inteligencia artificial en educación	12	Investigación universitaria	Valorar la integración de IA como apoyo formativo.
E4	Maestría	Realidad virtual y simulación educativa	30	Tecnología educativa	Valorar factibilidad técnica y pertinencia del componente inmersivo.
E5	Doctorado	Innovación educativa	20	Gestión académica	Valorar viabilidad institucional, escalabilidad y adopción progresiva.
E6	Doctorado	Gobernanza y ética digital	14	Universidad y consultoría	Valorar privacidad, seguridad, equidad y gobernanza tecnológica.
E7	Maestría	Formación y desarrollo docente	16	Universidad	Valorar factibilidad operativa y acompañamiento docente.

Nota. Los integrantes del panel se reportan mediante códigos alfanuméricos para preservar una identificación funcional y homogénea. La composición buscó asegurar diversidad de expertise en dimensiones directamente vinculadas con la arquitectura IA+RV.

5.10.3.2 Primera ronda: valoración inicial de la propuesta

En la primera ronda, los expertos recibieron un resumen ejecutivo de la propuesta y una matriz estructurada de valoración. Las dimensiones evaluadas fueron claridad conceptual, coherencia empírica, pertinencia pedagógica, factibilidad operativa, escalabilidad institucional, gobernanza y seguridad, y originalidad o aporte doctoral.

Los resultados de la primera ronda mostraron una valoración favorable en coherencia empírica, pertinencia pedagógica y originalidad del aporte. Las dimensiones con valoración relativamente menor fueron factibilidad operativa, escalabilidad institucional y gobernanza, no porque los expertos consideraran inviable la propuesta, sino porque recomendaron hacer más explícitas las condiciones mínimas de adopción,

los criterios de salida, las responsabilidades institucionales y el papel delimitado de la IA. La Tabla 5.8 presenta la síntesis de la primera ronda.

Tabla 5.8

Matriz de valoración experta, ronda 1

Dimensión valorada	Valoración promedio	Lectura interpretativa	Implicación para la propuesta
Claridad conceptual	3.3	Adecuada, con necesidad de precisar alcance y secuencia.	Reforzar denominación, mecanismo e invariantes.
Coherencia empírica	3.8	Alta relación entre hallazgos y propuesta.	Mantener anclaje en resultados del estudio.
Pertinencia pedagógica	3.9	Muy alta pertinencia para competencias transversales.	Conservar práctica situada, rúbricas y <i>feedback</i> .
Factibilidad operativa	3.1	Viable bajo condiciones explícitas.	Precisar recursos, roles y mínimo viable.
Escalabilidad institucional	3.0	Posible, pero requiere fases y criterios.	Fortalecer estrategia gradual.
Gobernanza y seguridad	3.2	Relevante, pero requiere mayor delimitación.	Ampliar privacidad, equidad, seguridad y supervisión humana.
Originalidad y aporte doctoral	3.7	Aporte claro y pertinente.	Explicitar integración IA+RV como mecanismo doctoral.

Nota. La primera ronda permitió identificar fortalezas del diseño y áreas de mejora relacionadas con factibilidad, escalamiento y gobernanza. Estas observaciones sirvieron como base para refinar la propuesta antes de la segunda ronda.

5.10.3.3 Ajustes incorporados entre rondas

A partir de la primera ronda se realizaron ajustes orientados a fortalecer la claridad conceptual, la factibilidad institucional, los criterios de adopción y el papel de la IA dentro de la arquitectura. Los cambios no modificaron el núcleo de la propuesta, pero sí aumentaron su precisión, auditabilidad y potencial de adopción responsable.

Estos ajustes cumplieron una función de refinamiento metodológico, ya que permitieron convertir observaciones generales del panel en modificaciones concretas y verificables dentro de la propuesta. En particular, se buscó precisar qué componentes debían considerarse no negociables, qué elementos podían adaptarse según el contexto institucional y qué evidencias serían necesarias para justificar continuidad, ajuste o escalamiento. De esta forma, la retroalimentación experta no se incorporó como opinión aislada, sino como insumo para fortalecer la coherencia interna, la viabilidad operativa y

la gobernanza de la arquitectura IA+RV. La Tabla 5.9 sintetiza las observaciones recurrentes y los ajustes incorporados.

Tabla 5.9

Observaciones recurrentes de la ronda 1 y ajustes incorporados en la propuesta

Observación recurrente de expertos	Ajuste incorporado
Mayor claridad sobre las condiciones de escalamiento.	Se reforzó la estrategia por fases y los criterios de salida.
Necesidad de hacer visible la factibilidad institucional.	Se fortaleció el apartado de factibilidad técnica, operativa e institucional.
Criterios de adopción poco visibles.	Se consolidó el apartado de adopción y escalamiento responsable.
Necesidad de delimitar mejor el papel de la IA.	Se reforzó la noción de IA como tutor formativo con supervisión humana.
Mayor claridad sobre gobernanza y seguridad.	Se ampliaron criterios de privacidad, equidad, trazabilidad y seguridad RV.
Claridad del aporte doctoral.	Se hizo más explícita la originalidad del mecanismo integrado IA+RV.

Nota. La tabla sintetiza el vínculo entre observaciones expertas y ajustes concretos incorporados a la propuesta antes de la segunda ronda.

5.10.3.4 Segunda ronda: revaloración y convergencia

La segunda ronda se aplicó sobre la versión revisada de la propuesta y se acompañó de una síntesis agregada de la primera ronda. Los resultados mostraron mejora en las dimensiones previamente señaladas como susceptibles de fortalecimiento, especialmente claridad conceptual, factibilidad, escalabilidad y gobernanza. Aunque factibilidad operativa y escalabilidad institucional mantuvieron una valoración moderada, los expertos consideraron que los ajustes incorporados fueron suficientes para sostener la pertinencia y viabilidad de la propuesta como modelo de adopción gradual bajo evaluación continua.

Esta segunda ronda cumplió una función metodológica clave, ya que permitió verificar si las modificaciones incorporadas después de la valoración inicial habían reducido las ambigüedades señaladas por el panel y aumentado el nivel de consenso experto. En este sentido, la revaloración no se interpretó únicamente como una repetición del instrumento, sino como una evidencia de convergencia progresiva entre el juicio experto, los ajustes de diseño y la versión final de la propuesta. Esta lógica es

consistente con el uso del Delphi modificado como procedimiento iterativo de validación, en el que la calidad de la propuesta se fortalece mediante retroalimentación estructurada, revisión y nueva valoración.

Tabla 5.10

Matriz de revaloración experta, ronda 2

Dimensión	Valoración ronda 1	Nueva valoración	¿Ajuste suficiente?	Comentario final
Claridad conceptual	3.3	3.7	Sí	La propuesta quedó más clara en propósito, alcance y secuencia.
Coherencia empírica	3.8	3.9	Sí	La relación entre hallazgos y propuesta se mantiene sólida.
Pertinencia pedagógica	3.9	4.0	Sí	El mecanismo instruccional está bien fundamentado.
Factibilidad técnica y operativa	3.1	3.6	Sí	La viabilidad es suficiente bajo condiciones mínimas explícitas.
Escalabilidad institucional	3.0	3.5	Sí	La estrategia por fases mejora la claridad de adopción gradual.
Gobernanza y seguridad	3.2	3.7	Sí	Las salvaguardas y responsabilidades quedaron mejor delimitadas.
Originalidad y aporte doctoral	3.7	3.9	Sí	El aporte doctoral es claro y pertinente.

Nota. La segunda ronda permitió valorar la suficiencia de los ajustes incorporados. La mejora en claridad, factibilidad, escalabilidad y gobernanza indica mayor consistencia interna de la propuesta.

5.10.3.5 Síntesis interpretativa del proceso Delphi

La comparación entre ambas rondas muestra una mejora general en las dimensiones inicialmente observadas como susceptibles de fortalecimiento. La claridad conceptual aumentó al explicitar la denominación, el mecanismo instruccional y los objetivos instruccionales de la propuesta. La gobernanza se fortaleció mediante criterios de privacidad, supervisión humana, equidad, seguridad y trazabilidad. La adopción institucional se hizo más clara mediante fases, mínimo viable, criterios de avance y evaluación continua.

El Delphi modificado aportó evidencia complementaria de que la propuesta es clara, pertinente y viable bajo criterios académicos e institucionales razonables. Además, el proceso no solo funcionó como valoración externa, sino como mecanismo de refinamiento del diseño. Las observaciones expertas ayudaron a precisar condiciones de

factibilidad, evitar ambigüedad en el papel de la IA y reforzar el carácter auditable de la arquitectura.

Es importante subrayar que esta validación no sustituye la evidencia empírica del estudio. Su función es distinta y complementaria. Mientras el Capítulo 4 aporta evidencia de efecto sobre el desempeño estudiantil bajo condiciones reales de implementación, el Delphi aporta juicio experto sobre claridad, coherencia, factibilidad, gobernanza y potencial de adopción responsable. La convergencia entre evidencia empírica y validación experta fortalece la legitimidad de la propuesta como contribución doctoral.

En síntesis, la valoración de la pertinencia permite sostener que la arquitectura IA+RV responde a una necesidad formativa relevante, se apoya en criterios pedagógicos verificables, es factible bajo condiciones institucionales razonables, incorpora salvaguardas de gobernanza y fue refinada mediante juicio experto. Con ello, la propuesta queda preparada para ser evaluada en términos de efectividad, aspecto que se desarrolla en la siguiente sección.

5.11 Evaluación de la efectividad de la propuesta

La evaluación de la efectividad de la propuesta debe distinguirse de la validación empírica inicial presentada en el Capítulo 4. Aunque el estudio ya aportó evidencia de mejora en Comunicación y Solución de problemas bajo una implementación específica de la arquitectura IA+RV, una propuesta doctoral orientada a adopción y escalamiento requiere establecer cómo se verificará su funcionamiento en futuras aplicaciones, cursos, cohortes o contextos institucionales. Por ello, se propone evaluar la efectividad mediante un enfoque mixto, capaz de integrar evidencia cuantitativa sobre desempeño con evidencia cualitativa sobre experiencia, barreras, facilitadores y condiciones de implementación (Creswell & Plano Clark, 2018; Desalos et al., 2024; Universidad de Investigación e Innovación de México, s. f.-a). Esta evaluación debe diferenciar resultados de aprendizaje, calidad de implementación, seguridad tecnológica, sostenibilidad institucional y capacidad de escalamiento, dado que una intervención puede mostrar impacto pedagógico inicial y, al mismo tiempo, presentar debilidades en

fidelidad, equidad, seguridad o mantenimiento organizacional (Carroll et al., 2007; Century et al., 2010; Bryk et al., 2015).

En esta tesis, la efectividad no se define como uso de IA o RV, satisfacción estudiantil o número de sesiones realizadas, sino como la capacidad de conservar el mecanismo instruccional central, implementarse con fidelidad, producir mejoras observables, proteger a los estudiantes y sostenerse institucionalmente. En consecuencia, la evaluación se organiza en cinco niveles. La efectividad pedagógica se refiere a la mejora del desempeño observable en Comunicación y Solución de problemas, evaluada mediante rúbricas analíticas PRE y POST, con procedimientos de consistencia cuando sean viables y estimadores adecuados a la estructura de los datos, incluyendo modelos que consideren grupo, tiempo y agrupamiento por clase (AERA/APA/NCME, 2014; Raudenbush & Bryk, 2002). La efectividad de implementación alude al grado en que la arquitectura se ejecuta conforme al diseño previsto, incluyendo dosis, cumplimiento del ciclo instruccional, uso explícito de rúbricas, calidad del *feedback*, participación efectiva y documentación de incidentes, conforme a la literatura sobre fidelidad de implementación (Dane & Schneider, 1998; Carroll et al., 2007; Century et al., 2010).

La efectividad tecnológica y de seguridad se refiere a la operación estable del sistema sin interrupciones críticas, incidentes severos no gestionados o riesgos indebidos para la privacidad y el bienestar estudiantil; por ello, debe considerar disponibilidad de dispositivos, estabilidad de plataforma, activación de protocolos, registro de ciber marea o malestar, manejo responsable de datos y alternativas equivalentes (UNESCO, 2021; UNESCO, 2023; NIST, 2023; Varsik & Vosberg, 2024). La efectividad institucional se entiende como la capacidad de sostener la arquitectura más allá de una aplicación aislada, mediante formación docente, soporte técnico, revisión periódica de indicadores, actualización de escenarios, mejora de *prompts*, mantenimiento de rúbricas y continuidad de recursos (Bryk et al., 2015). Finalmente, la efectividad de escalamiento se refiere a la posibilidad de ampliar la propuesta hacia nuevos cursos o cohortes sin pérdida de fidelidad, seguridad, calidad del *feedback* o resultados de aprendizaje; por tanto, solo puede afirmarse cuando las adaptaciones disciplinares preservan los componentes esenciales del modelo y generan evidencia suficiente para justificar

continuidad, ajuste o expansión. La Tabla 5.11 presenta el esquema de evaluación de efectividad de la propuesta.

Tabla 5.11

Esquema de evaluación de la efectividad de la arquitectura IA+RV

Nivel de efectividad	Pregunta evaluativa	Indicadores principales	Fuente de evidencia	Criterio de interpretación
Pedagógica	¿Mejora el desempeño en Comunicación y Solución de problemas?	Ganancia PRE y POST, cambio por criterio, tamaño de efecto.	Rúbricas analíticas y base de resultados.	Mejora educacionalmente relevante y consistente.
Calidad de medición	¿La mejora se mide con instrumentos confiables?	Consistencia de rúbricas, acuerdo interevaluador, revisión de criterios.	Rúbricas, doble calificación o auditoría.	La inferencia debe basarse en medición estable.
Implementación	¿La arquitectura se aplicó con fidelidad?	<i>Briefing</i> , RV, <i>feedback</i> , <i>debriefing</i> , IA, QA y uso de rúbrica.	<i>Checklists</i> , observación, <i>logs</i> y bitácoras.	Interpretar resultados solo con fidelidad suficiente.
Feedback IA	¿La retroalimentación fue útil, criterial y accionable?	Especificidad, vínculo con rúbrica y acciones de mejora.	Auditoría de outputs y revisión docente.	Ajustar prompts si el <i>feedback</i> es genérico.
Seguridad y equidad	¿La propuesta fue segura y accesible?	Incidentes, alternativas equivalentes y acceso efectivo.	Bitácoras, registros de soporte y reportes.	No escalar si genera riesgos o brechas.
Institucional	¿Puede sostenerse en condiciones reales?	Formación docente, soporte, revisión semestral y actualización de recursos.	Actas, reportes y evidencias de capacitación.	Sostener solo si hay capacidad instalada.
Escalamiento	¿Puede ampliarse sin perder calidad?	Comparabilidad entre cohortes, fidelidad y resultados consistentes.	Reportes longitudinales y análisis por cohorte.	Escalar solo con evidencia acumulada.

Nota. La evaluación de efectividad no se limita a medir uso tecnológico. Integra resultados de aprendizaje, calidad de implementación, seguridad, equidad, sostenibilidad institucional y capacidad de escalamiento.

Este esquema permite evitar una interpretación reduccionista de la efectividad. Una implementación podría mostrar alto uso de RV, pero baja efectividad si no produce mejoras en rúbricas o si carece de *debriefing*. También podría mostrar mejora inicial, pero baja sostenibilidad si requiere recursos extraordinarios o depende exclusivamente de un docente. Del mismo modo, una implementación no debe considerarse escalable si

mejora resultados, pero aumenta brechas de acceso o presenta problemas de seguridad no resueltos.

Para valorar la efectividad en futuras cohortes, se recomienda aplicar una lógica de evaluación continua. En cada implementación deberán recolectarse indicadores mínimos de aprendizaje, fidelidad, seguridad, calidad del feedback y equidad. Al cierre de cada cohorte, estos indicadores deberán revisarse para decidir si la propuesta debe mantenerse, ajustarse, ampliarse o rediseñarse. Esta decisión no debe basarse en una sola métrica, sino en la convergencia de evidencias. La Tabla 5.12 presenta una matriz de decisión para el seguimiento de efectividad.

Tabla 5.12

Matriz de decisión para seguimiento de efectividad y mejora continua

Resultado observado	Interpretación	Decisión recomendada
Alta mejora en desempeño y alta fidelidad.	La arquitectura funciona conforme al diseño.	Mantener y considerar expansión controlada.
Alta mejora, pero baja fidelidad documentada.	El efecto es prometedor, pero difícil de atribuir.	Reforzar trazabilidad antes de escalar.
Baja mejora y alta fidelidad.	El modelo se aplicó, pero requiere rediseño pedagógico.	Revisar escenarios, <i>prompts</i> , rúbricas o <i>debriefing</i> .
Baja mejora y baja fidelidad.	No hay condiciones suficientes para evaluar efectividad.	Repetir fase de consolidación.
Incidentes de seguridad o inequidad de acceso.	La implementación no cumple condiciones éticas mínimas.	Pausar, adaptar o activar alternativas equivalentes.
Resultados positivos, pero alta carga operativa.	Impacto con sostenibilidad limitada.	Simplificar instrumentos o ajustar recursos.
Resultados consistentes entre cohortes.	Evidencia acumulada de replicabilidad.	Avanzar hacia institucionalización gradual.

Nota. La matriz permite traducir los indicadores de efectividad en decisiones de mejora. Su finalidad es evitar que la propuesta se escale por entusiasmo tecnológico sin evidencia suficiente de aprendizaje, fidelidad, seguridad y sostenibilidad.

En síntesis, la evaluación de efectividad de la arquitectura IA+RV debe responder a una pregunta más amplia que si se utilizó tecnología. Debe determinar si la propuesta conserva su mecanismo instruccional, si mejora competencias observables, si se implementa con calidad, si protege a los estudiantes, si puede sostenerse institucionalmente y si puede escalarse sin perder fidelidad. Bajo esta lógica, la

efectividad se entiende como una combinación de impacto pedagógico, calidad de implementación, seguridad, equidad y capacidad institucional de mejora continua.

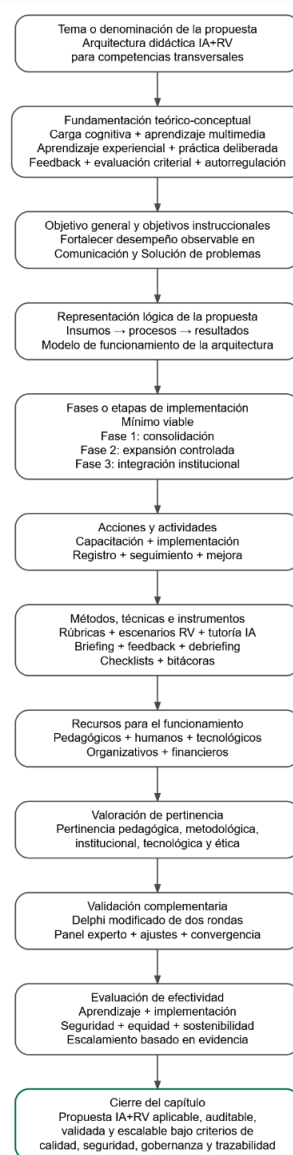
Esta concepción ampliada de efectividad es especialmente importante porque evita reducir la propuesta a una comparación entre grupos o a una valoración instrumental de herramientas tecnológicas. La arquitectura IA+RV se entiende, más bien, como un sistema pedagógico que debe preservar coherencia entre problema, fundamento teórico, componentes didácticos, condiciones de implementación, criterios de evaluación y mecanismos de mejora. Por ello, la valoración de efectividad no se limita a preguntar si hubo mejora en los puntajes, sino si dicha mejora puede explicarse a partir de una intervención implementada con fidelidad suficiente, documentada con evidencia trazable y protegida por condiciones mínimas de seguridad y equidad.

En consecuencia, el aporte del Capítulo 5 consiste en transformar los hallazgos del diagnóstico en una propuesta estructurada de intervención educativa. Esta transformación permite pasar de la evidencia empírica a un modelo aplicable, auditable y susceptible de mejora continua, capaz de orientar decisiones docentes e institucionales sobre el uso responsable de IA y RV en educación superior. Así, la propuesta no se presenta como una receta cerrada, sino como una arquitectura adaptable bajo condiciones explícitas de calidad, gobernanza y evaluación, lo que fortalece su pertinencia doctoral y su utilidad para contextos educativos comparables.

Como cierre del capítulo, la Figura 5.5 sintetiza la estructura interna de la propuesta desarrollada en el Capítulo 5, desde su denominación y fundamentación teórico-conceptual hasta su valoración de pertinencia y evaluación de efectividad.

Figura 5.5

Síntesis del Capítulo 5: de la evidencia empírica a una arquitectura IA+RV auditable, validada y escalable.



Nota. La figura sintetiza la estructura interna del Capítulo 5 conforme a la secuencia de resultados propositivos establecida por la universidad. Muestra el tránsito desde la denominación y fundamentación de la propuesta hasta su diseño operativo, validación complementaria y evaluación de efectividad. La síntesis permite observar que la arquitectura IA+RV se concibe como una propuesta aplicable, auditable y escalable, sustentada en condiciones explícitas de calidad, seguridad, equidad, gobernanza y trazabilidad.

CONCLUSIONES

El propósito de esta tesis fue estimar, con inferencia adecuada para datos jerárquicos, en qué medida una arquitectura didáctica que integra inteligencia artificial (IA) como tutor formativo criterial y realidad virtual (RV) como entorno de desempeño mejora el desarrollo de dos competencias transversales en educación superior: Comunicación y Solución de problemas. Para ello, se ejecutó un diseño con mediciones PRE y POST, grupo de comparación, asignación por conglomerados naturales, análisis principal mediante modelos lineales mixtos y una estrategia complementaria de métodos mixtos orientada a interpretar no solo si hubo mejora, sino bajo qué condiciones instruccionales dicha mejora resultó plausible (Raudenbush & Bryk, 2002; Creswell & Plano Clark, 2018).

El estudio partió de una cohorte inicial de $N = 180$ estudiantes y, tras aplicar criterios operativos preespecificados de asistencia mínima y elegibilidad analítica, definió una muestra principal de $n = 160$ participantes. Esta distinción entre cohorte inicial y muestra analítica permitió reportar el flujo de implementación sin confundirlo con el conjunto utilizado para la inferencia principal. La calidad de medición fue atendida mediante rúbricas analíticas, doble calificación ciega, estimación de fiabilidad interevaluador mediante ICC y valoración de contenido mediante S-CVI/Ave, en congruencia con estándares de medición educativa y psicométrica aplicados a evidencias de desempeño (AERA/APA/NCME, 2014; Shrout & Fleiss, 1979; Koo & Li, 2016; Polit & Beck, 2006).

Las conclusiones que siguen se organizan en función de los objetivos específicos que guiaron la investigación, con el propósito de explicitar de manera directa el grado en que cada uno fue alcanzado a partir de la evidencia empírica obtenida. Esta estructura mantiene la trazabilidad entre problema, objetivos, hipótesis, variables, instrumentos, resultados y propuesta, y recupera el contraste de hipótesis realizado en el Capítulo 4, diferenciando entre los supuestos confirmados por la evidencia principal y aquellos que no fueron confirmados en los términos estrictos inicialmente planteados, pero que abren líneas relevantes para investigación futura.

Conclusión vinculada al objetivo específico 1

El primer objetivo específico consistió en estimar el cambio en Comunicación y Solución de problemas entre la condición de intervención IA+RV y la condición de comparación, considerando la medición inicial, la medición posterior y la estructura agrupada de los datos. A partir de la evidencia obtenida, se concluye que este objetivo fue alcanzado de manera satisfactoria, ya que la arquitectura didáctica IA+RV se asoció con mejoras diferenciales sustantivas en ambas competencias.

El parámetro central de inferencia fue la interacción Tiempo $\times$ Grupo, interpretada como el cambio diferencial entre la condición IA+RV y la condición de comparación. Esta decisión analítica fue adecuada porque el estudio no buscó describir únicamente cambios absolutos dentro de cada grupo, sino estimar si la trayectoria de mejora fue mayor en los estudiantes expuestos a la arquitectura didáctica. El uso de modelos lineales mixtos permitió atender la dependencia derivada de la organización por clases y producir estimaciones más coherentes con la naturaleza del diseño por conglomerados (Raudenbush & Bryk, 2002).

Los resultados permiten concluir que la arquitectura IA+RV produjo un efecto favorable en Comunicación y Solución de problemas bajo condiciones reales de aula. Esta conclusión confirma el supuesto principal de que una secuencia instruccional integrada, basada en práctica situada, retroalimentación criterial, rúbricas analíticas y mediación docente, puede generar mejoras observables en competencias transversales complejas. La magnitud de los efectos estimados refuerza la relevancia educacional del hallazgo, aunque su interpretación debe mantenerse vinculada al contexto, al diseño y a las condiciones de implementación documentadas.

En consecuencia, el objetivo específico 1 se considera alcanzado y las hipótesis principales asociadas al efecto diferencial de IA+RV sobre Comunicación y Solución de problemas quedan confirmadas por la evidencia principal del estudio.

Conclusión vinculada al objetivo específico 2

El segundo objetivo específico consistió en asegurar que la medición del desempeño en Comunicación y Solución de problemas se realizara mediante

instrumentos suficientemente consistentes, interpretables y alineados con los constructos evaluados. Este objetivo también fue alcanzado, ya que las rúbricas analíticas mostraron niveles adecuados de fiabilidad interevaluador y validez de contenido.

La utilización de rúbricas analíticas permitió traducir competencias complejas en criterios observables, niveles de desempeño y evidencias evaluables. Esta decisión fue fundamental porque el estudio no se basó en satisfacción estudiantil ni en percepciones generales de aprendizaje, sino en evidencias de desempeño valoradas mediante criterios explícitos. En investigación educativa, la calidad de las inferencias depende en gran medida de la consistencia, pertinencia y validez de los instrumentos utilizados para representar el constructo evaluado (AERA/APA/NCME, 2014).

La fiabilidad interevaluador mostró que las calificaciones aplicadas por evaluadores independientes fueron suficientemente consistentes para sostener la interpretación de los resultados. Asimismo, la validez de contenido permitió afirmar que los criterios de las rúbricas estaban alineados con las dimensiones de Comunicación y Solución de problemas definidas en la operacionalización del estudio. Esta correspondencia entre constructo, criterio y puntuación fortalece la credibilidad de los hallazgos, porque reduce el riesgo de que las mejoras observadas sean producto de ambigüedad instrumental o variabilidad de calificación.

En consecuencia, el objetivo específico 2 se considera alcanzado. La calidad de medición sostiene la inferencia principal del estudio y permite afirmar que las mejoras reportadas corresponden a cambios observables en desempeño, no únicamente a percepciones subjetivas o a indicadores indirectos.

Conclusión vinculada al objetivo específico 3

El tercer objetivo específico consistió en documentar la implementación de la arquitectura IA+RV mediante indicadores de fidelidad, dosificación y condiciones operativas. Este objetivo fue alcanzado en términos de trazabilidad de implementación, aunque los análisis complementarios sugieren una interpretación prudente respecto de la relación dosis y respuesta.

La evidencia de implementación permitió verificar que la intervención no fue una exposición tecnológica indeterminada, sino una arquitectura didáctica con componentes identificables: *briefing*, simulación en RV, *feedback* formativo, *debriefing*, rúbricas, registros de dosis y criterios de seguridad. Esta documentación resulta esencial porque una intervención educativa solo puede interpretarse adecuadamente si se conoce en qué medida fue aplicada conforme al diseño previsto. La literatura sobre fidelidad de implementación advierte que la exposición, la calidad de entrega, la adherencia y la adaptación contextual deben considerarse para evitar atribuciones causales imprecisas (Dane & Schneider, 1998; Carroll *et al.*, 2007; Century *et al.*, 2010).

Los análisis de dosificación, sin embargo, no respaldaron de forma concluyente una relación lineal según la cual mayor exposición a IA o RV se tradujera automáticamente en mayor mejora. Esta conclusión es importante porque evita una lectura simplista del efecto tecnológico. Los resultados sugieren que el impacto observado depende más de la participación en una arquitectura instruccional integrada que de un gradiente acumulativo de tiempo o número de interacciones dentro del rango observado.

En consecuencia, el objetivo específico 3 se considera alcanzado en cuanto a documentación de fidelidad, dosificación y condiciones de implementación. No obstante, las hipótesis o expectativas asociadas a una relación dosis y respuesta deben considerarse no confirmadas en términos estrictos. Este resultado no debilita el estudio; más bien precisa su contribución: la calidad de la arquitectura parece ser más relevante que la cantidad aislada de exposición tecnológica.

Conclusión vinculada al objetivo específico 4

El cuarto objetivo específico consistió en interpretar la experiencia de aprendizaje y los mecanismos plausibles mediante los cuales la arquitectura IA+RV pudo favorecer el desarrollo de Comunicación y Solución de problemas. Este objetivo fue alcanzado mediante la integración de evidencia cuantitativa, evidencia de implementación y lectura cualitativa del proceso.

Los hallazgos permiten sostener que la mejora observada no debe atribuirse a la IA o a la RV como artefactos aislados, sino a la forma en que ambas tecnologías fueron integradas dentro de una secuencia pedagógica. La RV ofreció condiciones de desempeño situado, donde el estudiante enfrentó tareas con mayor autenticidad funcional; la IA aportó apoyo formativo y retroalimentación criterial; las rúbricas dieron lenguaje común al desempeño; y el *debriefing* permitió transformar la experiencia en reflexión y transferencia. Esta interpretación es coherente con la literatura sobre retroalimentación formativa, práctica deliberada, aprendizaje experiencial y aprendizaje inmersivo (Kolb, 1984; Ericsson *et al.*, 1993; Hattie & Timperley, 2007; Shute, 2008; Makransky & Petersen, 2021).

La evidencia cualitativa y de proceso también permitió identificar que la mediación docente siguió siendo un componente central. La IA no sustituyó el acompañamiento académico ni la evaluación humana; funcionó como apoyo para ampliar la disponibilidad y especificidad del *feedback*. La RV tampoco operó como estímulo visual autónomo, sino como entorno de práctica cuando estuvo alineada con tareas, criterios y reflexión posterior. Esta conclusión coincide con la idea de que las tecnologías educativas producen valor formativo cuando se integran en diseños instruccionales coherentes, no cuando se adoptan por novedad o disponibilidad (Hattie & Timperley, 2007; Shute, 2008; Radianti *et al.*, 2020; Kasneci *et al.*, 2023).

En consecuencia, el objetivo específico 4 se considera alcanzado. La tesis identifica un mecanismo plausible de mejora basado en la interacción entre desempeño situado, retroalimentación criterial, mediación docente, rúbricas y reflexión posterior.

Conclusión vinculada al objetivo específico 5

El quinto objetivo específico consistió en integrar los hallazgos empíricos, la calidad de medición, la fidelidad de implementación y sus implicaciones pedagógicas. Este objetivo fue alcanzado porque la investigación no se limitó a reportar diferencias estadísticas, sino que articuló resultados, instrumentos, condiciones de implementación y lectura pedagógica del efecto.

La principal implicación es que la arquitectura IA+RV debe entenderse como un cambio de diseño instruccional y no como una simple incorporación de recursos digitales. La evidencia muestra que el aprendizaje mejoró cuando IA y RV se integraron con criterios de desempeño, práctica deliberada, retroalimentación y mediación docente. Esta conclusión es relevante porque desplaza la discusión desde el entusiasmo tecnológico hacia la calidad del diseño pedagógico. En este sentido, la tesis aporta una advertencia institucional: no basta adquirir dispositivos, plataformas o herramientas de IA; es necesario construir secuencias didácticas, rúbricas, criterios de calidad, protocolos de seguridad y mecanismos de trazabilidad.

Desde el punto de vista metodológico, la tesis demuestra que la evaluación de innovaciones tecnopedagógicas requiere combinar estimaciones de efecto con evidencia de implementación. Una mejora en desempeño solo es interpretable si se conoce cómo se aplicó la intervención, qué instrumentos se utilizaron, qué grado de fidelidad se alcanzó y bajo qué condiciones operó la experiencia. Este principio es congruente con enfoques de implementación y mejora continua que recomiendan interpretar los resultados en función de fidelidad, contexto, adaptación y sostenibilidad (Carroll *et al.*, 2007; Century *et al.*, 2010; Bryk *et al.*, 2015).

En consecuencia, el objetivo específico 5 se considera alcanzado. La tesis integra resultados, medición, implementación y significado pedagógico en una explicación coherente y útil para orientar decisiones futuras.

Conclusión vinculada al objetivo específico 6

El sexto objetivo específico consistió en derivar, a partir de los hallazgos del estudio, una propuesta de arquitectura didáctica IA+RV para su implementación, monitoreo y mejora en educación superior. Este objetivo fue alcanzado mediante el desarrollo del Capítulo 5, en el que se formuló una propuesta estructurada conforme a la secuencia solicitada para resultados científicos de carácter propositivo.

La propuesta no se presenta como recomendación general de uso de IA o RV, sino como una arquitectura didáctica compuesta por denominación, fundamentación teórico conceptual, objetivo general, objetivos específicos instruccionales,

representación lógica, fases de implementación, acciones, métodos, técnicas, instrumentos, recursos, valoración de pertinencia y evaluación de efectividad. Esta estructura permite que el resultado de la investigación trascienda el reporte empírico y se convierta en una alternativa aplicable, auditable y susceptible de mejora.

Además, la propuesta incorpora criterios de gobernanza, seguridad, equidad y escalamiento responsable. Esto es relevante porque el uso educativo de IA y RV plantea desafíos relacionados con privacidad, sesgos, bienestar, accesibilidad, supervisión humana y legitimidad institucional. Las orientaciones internacionales sobre IA educativa y gobernanza tecnológica insisten en que la adopción responsable exige transparencia, control humano, protección de datos, equidad y monitoreo continuo (UNESCO, 2021; UNESCO, 2023; NIST, 2023; Varsik & Vosberg, 2024).

En consecuencia, el objetivo específico 6 se considera alcanzado. La tesis culmina en una propuesta doctoral que transforma los hallazgos empíricos en un modelo de intervención educativa con potencial de transferencia, siempre que conserve sus componentes no negociables: rúbricas, práctica situada, *feedback* formativo, mediación docente, trazabilidad, seguridad y evaluación continua.

Conclusión vinculada al objetivo específico 7

El séptimo objetivo específico consistió en validar la propuesta de implementación escalable de la arquitectura didáctica IA+RV mediante juicio de expertos, considerando criterios de pertinencia, validez, factibilidad, aplicabilidad y generalización. Este objetivo fue alcanzado mediante un proceso de valoración experta que permitió revisar la coherencia interna de la propuesta, su correspondencia con los hallazgos empíricos del estudio y sus posibilidades de adopción en contextos educativos comparables, en congruencia con el uso del juicio experto como estrategia de validación y refinamiento de propuestas aplicadas (Hsu & Sandford, 2007; Okoli & Pawlowski, 2004).

La validación permitió fortalecer la propuesta al incorporar observaciones orientadas a precisar sus componentes esenciales, sus condiciones de implementación y sus criterios de escalamiento responsable. En particular, el juicio experto confirmó la

pertinencia de concebir la IA y la RV como componentes de una arquitectura didáctica mediada por criterios, práctica situada, retroalimentación, trazabilidad y gobernanza, y no como herramientas tecnológicas aisladas. Esta interpretación es consistente con enfoques de fidelidad de implementación y adopción responsable, que recomiendan distinguir entre diseño previsto, entrega real, condiciones institucionales y salvaguardas de calidad (Carroll et al., 2007; Century et al., 2010; UNESCO, 2023; Varsik & Vosberg, 2024). En consecuencia, el objetivo específico 7 se considera alcanzado, ya que la propuesta fue sometida a revisión externa y ajustada con base en criterios de calidad, viabilidad y aplicabilidad institucional.

Síntesis de cumplimiento de objetivos e hipótesis

La síntesis final del cumplimiento de objetivos e hipótesis permite cerrar la sección de conclusiones con una lectura integrada del estudio. Después de presentar cada conclusión vinculada a los objetivos específicos, resulta necesario mostrar de manera compacta cómo la evidencia empírica respondió al planteamiento inicial de la investigación y qué nivel de respaldo obtuvo cada supuesto formulado. Esta integración evita que las conclusiones queden como afirmaciones aisladas y permite verificar la trazabilidad entre problema, objetivos, hipótesis, variables, instrumentos, resultados e interpretación final.

En este sentido, la síntesis distingue entre tres niveles de evidencia: los objetivos alcanzados con respaldo empírico directo, las hipótesis principales confirmadas mediante los modelos de estimación y los componentes secundarios que requieren una interpretación más prudente. Esta distinción es relevante porque la tesis no sostiene que todos los mecanismos hayan sido demostrados con el mismo grado de certeza. Más bien, confirma el efecto principal de la arquitectura IA+RV sobre el desempeño observable y delimita con claridad aquellos aspectos que deben continuar evaluándose en investigaciones posteriores.

La Tabla 6.1 sintetiza el grado de cumplimiento de los objetivos específicos y el resultado general del contraste de hipótesis. Su finalidad es hacer explícita la relación entre el planteamiento inicial, la evidencia obtenida y las conclusiones derivadas.

Tabla 6.1

Síntesis de cumplimiento de objetivos e hipótesis

Elemento evaluado	Resultado de cumplimiento	Evidencia principal	Interpretación final
Objetivo específico 1: estimar cambios en Comunicación y Solución de problemas	Alcanzado	Interacción Tiempo $\times$ Grupo y tamaños de efecto a favor de IA+RV	La arquitectura IA+RV produjo mejoras diferenciales sustantivas en ambas competencias
Objetivo específico 2: asegurar calidad de medición	Alcanzado	Rúbricas analíticas, doble calificación, ICC y S-CVI/Ave	Los instrumentos fueron suficientemente consistentes y pertinentes para sostener la inferencia principal
Objetivo específico 3: documentar dosificación y fidelidad	Alcanzado, con matiz interpretativo	Registros de implementación, fidelidad, dosis y condiciones de seguridad	La intervención fue trazable; la dosis no mostró una relación lineal concluyente con la mejora
Objetivo específico 4: interpretar mecanismos plausibles	Alcanzado	Evidencia de proceso, experiencia de aprendizaje y lectura pedagógica de la intervención	El efecto se explica mejor como cambio de arquitectura instruccional que como exposición tecnológica aislada
Objetivo específico 5: integrar resultados e implicaciones	Alcanzado	Articulación entre hallazgos, calidad de medición, fidelidad e implicaciones pedagógicas	La tesis ofrece una interpretación integrada y no solo un reporte estadístico
Objetivo específico 6: formular propuesta derivada de la evidencia	Alcanzado	Capítulo 5, arquitectura IA+RV, fases, instrumentos, recursos y evaluación de efectividad	La investigación se traduce en una propuesta doctoral aplicable, auditable y escalable bajo condiciones
Objetivo específico 7: validar la propuesta mediante juicio de expertos	Alcanzado	Delphi modificado de dos rondas; valoración experta de pertinencia, validez, factibilidad, aplicabilidad y generalización.	La propuesta fue revisada, validada y refinada mediante juicio experto, fortaleciendo su coherencia interna, viabilidad institucional y potencial de aplicación en contextos educativos comparables.
Hipótesis principales sobre mejora diferencial IA+RV	Confirmadas por la evidencia principal	Resultados favorables en Comunicación y Solución de problemas	La arquitectura IA+RV se asoció con mejoras superiores frente a la comparación
Hipótesis o expectativas sobre dosis y moderadores	No confirmadas en términos estrictos o apoyadas solo parcialmente	Análisis complementarios de dosificación y variabilidad	La evidencia no permite sostener una relación lineal de dosis ni efectos moderadores robustos
Hipótesis de plausibilidad instruccional	Apoyada por convergencia de evidencia	Fidelidad, rúbricas, <i>feedback</i> , RV, IA y <i>debriefing</i>	El mecanismo plausible se ubica en la arquitectura didáctica integrada

Nota. La tabla sintetiza la trazabilidad entre objetivos, evidencia e interpretación. La confirmación de los efectos principales no implica que todos los mecanismos secundarios hayan sido demostrados con el mismo nivel de certeza; por ello, las conclusiones distinguen entre resultados confirmados, resultados plausibles y líneas pendientes de investigación.

RECOMENDACIONES

Las recomendaciones derivadas de esta investigación se organizan en dos planos complementarios: investigación futura e implementación institucional. En ambos casos, se formulan como acciones derivadas de los hallazgos, los límites empíricos y las condiciones de efectividad identificadas en el estudio. Su propósito es orientar la continuidad académica de la línea IA+RV y favorecer una adopción responsable, evaluable y pedagógicamente justificada en educación superior.

Recomendaciones académicas y metodológicas para investigaciones futuras

Se recomienda que futuras investigaciones repliquen la arquitectura IA+RV en otras instituciones, disciplinas, programas y niveles educativos, preservando sus componentes pedagógicos esenciales: rúbricas analíticas, práctica situada, retroalimentación criterial, mediación docente y cierre reflexivo. Esta línea permitiría estimar con mayor precisión la validez externa del modelo, identificar condiciones contextuales que favorecen o limitan su funcionamiento y valorar si los efectos observados se mantienen en escenarios curriculares distintos al contexto original del estudio.

También se recomienda profundizar en el análisis de los mecanismos que explican el cambio observado. Para ello, futuros estudios podrían incorporar diseños con niveles de exposición deliberadamente diferenciados, mediciones más finas de la calidad del *feedback*, profundidad del *debriefing*, patrones de revisión posterior a la retroalimentación y trazas de interacción durante la práctica. Esta estrategia permitiría distinguir con mayor claridad si la mejora depende principalmente de la práctica inmersiva, de la retroalimentación asistida por IA, de la mediación docente o de la articulación completa de la arquitectura instruccional.

Finalmente, se recomienda ampliar la evaluación hacia diseños longitudinales y comparativos. La evidencia futura debería examinar si las mejoras en Comunicación y Solución de problemas se sostienen en el tiempo, si se transfieren a tareas nuevas y si la arquitectura IA+RV resulta más eficiente que otras alternativas pedagógicas, como tutoría intensiva, aprendizaje basado en problemas sin RV o laboratorios presenciales.

Esta línea sería especialmente útil para orientar decisiones institucionales sobre inversión, escalamiento, sostenibilidad y equidad.

Recomendaciones prácticas para implementación institucional responsable

Para una implementación institucional responsable, se recomienda adoptar la arquitectura IA+RV de manera gradual, comenzando por cursos o módulos donde las competencias puedan evaluarse mediante tareas auténticas y rúbricas analíticas. El escalamiento no debe depender de la disponibilidad de tecnología, sino de evidencia mínima de funcionamiento: aplicación completa del ciclo instruccional, uso consistente de criterios de desempeño, registro de exposición efectiva, documentación de incidentes y revisión de la calidad de la retroalimentación.

Asimismo, se recomienda establecer reglas claras de gobernanza para el uso de IA en actividades formativas. La IA debe operar como apoyo al aprendizaje y no como sustituto del juicio docente, de la autoría estudiantil o de la evaluación académica. Para ello, la institución debe definir lineamientos sobre privacidad, minimización de datos, trazabilidad de configuraciones, supervisión humana, integridad académica y límites de uso en actividades evaluables, en concordancia con marcos recientes de adopción responsable de IA en educación (UNESCO, 2023; Varsik & Vosberg, 2024).

También se recomienda fortalecer la formación docente asociada al modelo. Esta formación no debe limitarse al uso técnico de plataformas, visores o herramientas de IA, sino incluir diseño instruccional, construcción y aplicación de rúbricas, conducción de *briefing* y *debriefing*, interpretación pedagógica del *feedback* automatizado, gestión de carga cognitiva y atención a contingencias durante experiencias inmersivas. La calidad de la arquitectura depende de que el docente conserve la mediación académica y pueda integrar la tecnología sin perder control pedagógico sobre la experiencia.

Finalmente, toda implementación de la arquitectura IA+RV debe contemplar condiciones explícitas de equidad, seguridad y sostenibilidad, incluyendo alternativas equivalentes para estudiantes que no puedan participar en experiencias inmersivas, protocolos ante malestar o ciber mareo, registro de incidencias, soporte técnico y evaluación periódica de dosis, fidelidad, calidad del *feedback*, desempeño y seguridad.

Bajo este enfoque, la tesis demuestra que una arquitectura didáctica criterial, cíclica y auditable, que integra inteligencia artificial como apoyo formativo y realidad virtual como entorno de desempeño, puede asociarse con mejoras sustantivas en competencias complejas de Comunicación y Solución de problemas en educación superior. Su aporte no consiste en afirmar que la tecnología mejora el aprendizaje por sí misma, sino en mostrar que su valor pedagógico emerge cuando se articula con criterios claros, práctica situada, retroalimentación accionable, mediación docente, evaluación consistente y gobernanza responsable. Al mismo tiempo, los hallazgos deben interpretarse con prudencia, pues la evidencia respalda el efecto principal sobre desempeño observable, pero no permite afirmar con igual certeza una relación lineal entre dosis y mejora, efectos moderadores robustos o persistencia de largo plazo sin nuevas verificaciones empíricas. En consecuencia, el valor final de esta investigación radica en establecer condiciones bajo las cuales una innovación tecnopedagógica puede implementarse, evaluarse, protegerse y proyectarse con responsabilidad educativa, entendiendo la IA y la RV no como fines en sí mismos, sino como componentes de una arquitectura formativa institucionalmente sostenible.

BIBLIOGRAFÍA

- Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(1), 4–23.
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. SAGE.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. <https://www.testingstandards.net/open-access-files.html>
- Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, 57(5), 13–18.
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.
- Association of American Colleges and Universities. (n.d.). *VALUE rubrics (Valid Assessment of Learning in Undergraduate Education)*. Retrieved December 25, 2025, from <https://projects.ncsu.edu/project/clinast/Files/ValueRubrics.pdf>
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.
- Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, 52, 1–26. doi:10.1146/annurev.psych.52.1.1
- Barzola, V. E. (2020). *Designing feedback for collocated teams using multimodal learning analytics* (Doctoral dissertation, University of Technology Sydney). ProQuest Dissertations & Theses Global. <https://www.proquest.com/openview/6743ebc67c8a31c767c8b61b3fbfe132/1?pq-origsite=gscholar&cbl=2026366&diss=y>

- Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., Czajkowski, S., & the NIH Behavior Change Consortium. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology, 23*(5), 443–451.
- Biocca, F., Harms, C., & Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators and Virtual Environments, 12*(5), 456–480.
doi:10.1162/105474603322761270
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.
<https://doi.org/10.1191/1478088706qp063oa>
- Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research (2nd ed.)*. Guilford Press.
- Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press.
- Cámara de Diputados del H. Congreso de la Unión. (2014). *Reglamento de la Ley General de Salud en Materia de Investigación para la Salud*. Diario Oficial de la Federación.
https://www.diputados.gob.mx/LeyesBiblio/regley/Reg_LGS_MIS.pdf
- Cámara de Diputados del H. Congreso de la Unión. (2025). *Ley Federal de Protección de Datos Personales en Posesión de los Particulares*. Diario Oficial de la Federación. <https://www.diputados.gob.mx/LeyesBiblio/pdf/LFPDPPP.pdf>
- Cámara de Diputados del H. Congreso de la Unión. (2026). *Ley Federal del Derecho de Autor*. Diario Oficial de la Federación.
<https://www.diputados.gob.mx/LeyesBiblio/pdf/LFDA.pdf>

- Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317–372.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.
- Campbell, M. K., Piaggio, G., Elbourne, D. R., & Altman, D. G. (2012). CONSORT 2010 statement: Extension to cluster randomised trials. *BMJ*, 345, e5661. <https://doi.org/10.1136/bmj.e5661>
- Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. John Wiley & Sons.
- Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, 2, 40.
- Çeken, B., & Taşkın, N. (2024). Examination of multimedia learning principles in augmented reality and virtual reality learning environments. *Journal of Computer Assisted Learning*. <https://doi.org/10.1111/jcal.13097>
- Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, 31(2), 199–218. <https://doi.org/10.1177/1098214010366173>
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), 464–504. <https://doi.org/10.1080/10705510701301834>
- Chen, S., & Lee, S. (2024). Enhancing university students' soft skills: A quantitative study on problem-solving and social competence. *Journal of Ecohumanism*, 3(7). <https://doi.org/10.62754/joe.v3i7.4554>
- Cheng, A., Grant, V., Robinson, T., Catena, H., Lachapelle, K., Kim, J., & Eppich, W. (2014). The Promoting Excellence and Reflective Learning in Simulation

- (PEARLS) approach to health care debriefing: A faculty development guide. *Clinical Simulation in Nursing*, 10(8), e401–e406.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5
- Cinar, O. E. (2024). *Augmented reality, virtual reality, and artificial intelligence to enhance programming comprehension* (Tesis doctoral, Queen's University Belfast). URL: https://pure.qub.ac.uk/files/617361950/Augmented_reality_virtual_reality_and_artificial_intelligence_to_enhance_programming_comprehension.pdf
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Contaldo, M., Pasceri, G., Vignati, G., Bracchi, L., Triggiani, S., & Carrafiello, G. (2024). AI in radiology: Navigating medical responsibility. *Diagnostics*, 14(14), 1506. <https://doi.org/10.3390/diagnostics14141506>
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29.
- Cummings, J. J., & Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychology*, 19(2), 272–309. <https://doi.org/10.1080/15213269.2015.1015740>
- Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, 9(3), 458–467.

- Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*(1), 23–45.
- DeVellis, R. F. (2017). *Scale development: Theory and applications (4th ed.)*. SAGE.
- Díaz-Muñoz, G. (2020). Metodología del estudio piloto. *Revista Chilena de Radiología, 26*(3), 100–104. <https://doi.org/10.4067/S0717-93082020000300100>
- Di Palma, R., Beausaert, S., Mahr, D., Heller, J., & Hilken, T. (2025). Does using virtual reality to enhance students' presentation skills actually work? The role of feedback awareness and usefulness. *Journal of Computer Assisted Learning*. <https://doi.org/10.1111/jcal.70097>
- Ding, M. (2023). *A mixed-method study of the impact of high-immersion virtual reality on foreign language speaking anxiety* (Doctoral dissertation, University of Glasgow). <https://theses.gla.ac.uk/84029/1/2023DingPhD.pdf>
- Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. Arnold.
- Dubiel, A., Kamińska, D., Zwoliński, G., Ramić-Brkić, B., Agostini, D., & Zancanaro, M. (2025). Virtual reality for the training of soft skills for professional education: trends and opportunities. *Interactive Learning Environments, 1*–21. <https://doi.org/10.1080/10494820.2025.2450634>
- Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412. <https://doi.org/10.1111/bjop.12046>
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3–4), 327–350.
- Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

- Ericsson, K. A. (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge University Press.
- Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine*, 15(11), 988–994.
- Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3), 363–406. doi:10.1037/0033-295X.100.3.363
- Etienne, É. (2025). *Virtual reality and analytics to enhance communication skills: A multidisciplinary approach* (Doctoral dissertation, University of Liège). <https://orbi.uliege.be/bitstream/2268/332627/1/Corpus%20Th%C3%A8se%20H%C3%A9l%C3%A9ne%20ETIENNE.pdf>
- Fanning, R. M., & Gaba, D. M. (2007). The role of debriefing in simulation-based learning. *Simulation in Healthcare*, 2(2), 115–125.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304–317. <https://doi.org/10.1504/IJTEL.2012.051816>
- Ferreira, C., Gabriel, B., Valente, R., Andrade-Campos, A., Dias-de-Oliveira, J., Neto, V., & Figueiredo, C. (2024). In search of a more balanced engineering curriculum: the perspective of students, teachers, alumni and employers. *Trends in Higher Education*, 3(1), 142–154. <https://doi.org/10.3390/higheredu3010008>
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). Wiley.

- Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. University of South Florida, Louis de la Parte Florida Mental Health Institute.
- Gómez Portes, C. (2023). *Artificial intelligence and immersive systems techniques for precision physical rehabilitation* (Doctoral dissertation, University of Castilla-La Mancha). URL: https://www.esi.uclm.es/www/cglez/downloads/students/phd/2023_cgomez.pdf
- Gómez Tobías, R. (2026). *AI and VR integration for communication and problem-solving in higher education (Behavioral dataset) [Data set]*. Open Science Framework. <https://doi.org/10.17605/OSF.IO/56GPE>
- Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? *Prevention Science*, 8(3), 206–213.
- Hamilton, E. R., Rosenberg, J. M., & Akcaoglu, M. (2016). The Substitution Augmentation Modification Redefinition (SAMR) model: A critical review and suggestions for its use. *TechTrends*, 60(5), 433–441. <https://doi.org/10.1007/s11528-016-0091-y>
- Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing*, 32(4), 1008–1015.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach (2nd ed.)*. Guilford Press.
- Hayes, R. J., & Moulton, L. H. (2017). *Cluster randomised trials (2nd ed.)*. CRC Press. <https://doi.org/10.4324/9781315370286>
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128.

- Hernández Sampieri, R., Fernández-Collado, C., & Baptista-Lucio, P. (2022). *Metodología de la investigación (7ª ed.)*. McGraw-Hill.
- Hoffman, H. R. (2023). *The impact of immersive VR versus desktop VR technologies on learning attainment and sense of presence* (Doctoral dissertation, University of South Dakota).
<https://red.library.usd.edu/cgi/viewcontent.cgi?article=1116&context=diss-thesis>
- Holmes, W., Bialik, M., & Fadel, C. (2023). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.
- Hsu, C.-C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research, and Evaluation*, 12(10), 1–8.
<https://doi.org/10.7275/pdz9-th90>
- Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- INEGI (2021). *Encuesta nacional sobre acceso y permanencia en la educación (ENAPE)*. URL: <https://www.inegi.org.mx/programas/enape/2021/>
- INEGI (2024). *Encuesta nacional sobre disponibilidad y uso de tecnologías de la información en los hogares (ENDUTIH)*. URL: https://www.inegi.org.mx/contenidos/saladerensa/boletines/2025/endutih/ENDUTIH_24.pdf
- Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies*, 23(4), 1515–1529. <https://doi.org/10.1007/s10639-017-9676-0>
- Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R^2 to random slopes models. *Methods in Ecology and Evolution*, 5(9), 944–946.

- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144.
<https://doi.org/10.1016/j.edurev.2007.05.002>
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.
<https://doi.org/10.1016/j.lindif.2023.102274>
- Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3), 983–997.
<https://doi.org/10.2307/2533558>
- Kim, S. (2024). *Harnessing AI for writing development: Exploring emerging multilinguals' ChatGPT usage patterns* (Doctoral dissertation, The Ohio State University). OhioLINK ETD Center.
https://etd.ohiolink.edu/acprod/odb_etd/ws/send_file/send?accession=osu1732498316837348&disposition=inline
- Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice-Hall.
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42–78.
<https://doi.org/10.3102/0034654315581420>

- Kumaravel, B. T. (2022). *Interactive cross-dimensional media for collaboration and guidance in mixed reality environments* (Doctoral dissertation, University of California, Berkeley). ProQuest Dissertations & Theses Global.
<https://www.proquest.com/openview/a3a36c0f3426369c5868b142784369f6/1?pq-origsite=gscholar&cbl=2026366&diss=y>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Langley, G. J., Moen, R. D., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance* (2nd ed.). Jossey-Bass.
- LaViola, J. J. (2000). A discussion of cybersickness in virtual environments. *ACM SIGCHI Bulletin*, 32(1), 47–56.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563–575.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.
<https://doi.org/10.1038/nature14539>
- Linstone, H. A., & Turoff, M. (1975). *The Delphi method: Techniques and applications*. Addison-Wesley.
- Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198–1202.
- Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.

- Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, 49 (4), 1494–1502. <https://doi.org/10.3758/s13428-016-0809-y>
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382–385.
- Makransky, G., & Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review*, 33, 937–958.
<https://doi.org/10.1007/s10648-020-09586-2>
- Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, 60, 225–236.
<https://doi.org/10.1016/j.learninstruc.2017.12.007>
- Martínez-Domínguez, M., & Fierros-González, I. (2022). Determinants of internet use by school-age children: The challenges for Mexico during the COVID-19 pandemic. *Telecommunications Policy*, 46(1), 102241.
<https://doi.org/10.1016/j.telpol.2021.102241>
- Martínez-Domínguez, M., & Mora-Rivera, J. (2020). Internet adoption and usage patterns in rural Mexico. *Technology in Society*, 60, 101226.
<https://doi.org/10.1016/j.techsoc.2019.101226>
- Mayer, R. E. (2021). *Multimedia learning (3rd ed.)*. Cambridge University Press.
<https://doi.org/10.1017/9781108859019>
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114(2), 376–390.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46.
<https://doi.org/10.1037/1082-989X.1.1.30>

- Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K–12 and higher education: A meta-analysis. *Computers & Education*, 70, 29–40. <https://doi.org/10.1016/j.compedu.2013.07.033>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems*, E77-D(12), 1321–1329.
- Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017–1054. <https://doi.org/10.1111/j.1467-9620.2006.00684.x>
- Morris, S. B. (2008). Estimating effect sizes from pretest–posttest–control group designs. *Organizational Research Methods*, 11(2), 364–386. <https://doi.org/10.1177/1094428106291059>
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. Oxford University Press.
- Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2), 133–142. <https://doi.org/10.1111/j.2041-210x.2012.002>
- National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1)*. U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- Ngo, T. (2024). The importance of soft skills for academic performance and career development—from the perspective of university students. *International Journal of Engineering Pedagogy (IJEP)*, 14(3), 53–68. <https://doi.org/10.3991/ijep.v14i3.45425>

- Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: An example, design considerations and applications. *Information & Management*, 42(1), 15–29. <https://doi.org/10.1016/j.im.2003.11.002>
- Otzen, T., & Manterola, C. (2017). Técnicas de muestreo sobre una población a estudio. *International Journal of Morphology*, 35(1), 227–232. <https://doi.org/10.4067/S0717-95022017000100037>
- Ouhaichi, K. (2024). *A new multimodal learning analytics framework for designing personalized learning experiences: The eMMLA framework* (Doctoral dissertation, Uppsala University). <https://www.diva-portal.org/smash/get/diva2:1807660/FULLTEXT01.pdf>
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129–144. <https://doi.org/10.1016/j.edurev.2013.01.002>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.
- Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology*, 110(6), 785–797. <https://doi.org/10.1037/edu0000241>
- Pelaez-Sanchez, I. C., Glasserman-Morales, L. D., & Rocha-Feregrino, G. (2024, October). Exploring digital competencies in higher education: design and validation of instruments for the era of Industry 5.0. In *Frontiers in Education* (Vol. 9, p. 1415800). Frontiers Media SA.
- Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060), 1226–1227.
- Piaget, J. (1970). *The science of education and the psychology of the child*. Viking Press.
- Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489–497. <https://doi.org/10.1002/nur.20147>

- Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., Griffey, R., & Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(2), 65–76.
- PuenteDura, R. (2010). *SAMR and TPCK: Intro to Advanced Practice*.
http://hippasus.com/resources/sweden2010/SAMR_TPCK_IntroToAdvancedPractice.pdf
- Pupic, N., Ghaffari-zadeh, A., Hu, R., Singla, R., Darras, K., Karwowska, A., & Forster, B. (2023). An evidence-based approach to artificial intelligence education for medical students: a systematic review. *PLOS Digital Health*, 2(11), e0000255.
<https://doi.org/10.1371/journal.pdig.0000255>
- Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education. *Computers & Education*, 147, 103778. <https://doi.org/10.1016/j.compedu.2019.103778>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Sage.
- Robles, M. M. (2012). Executive perceptions of the top 10 soft skills needed in today's workplace. *Business Communication Quarterly*, 75(4), 453-465.
<https://doi.org/10.1177/1080569912460400>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.
- Rudolph, J. W., Simon, R., Dufresne, R. L., & Raemer, D. B. (2006). There's no such thing as "nonjudgmental" debriefing: A theory and method for debriefing with good judgment. *Simulation in Healthcare*, 1(1), 49–55.
<https://doi.org/10.1097/01266021-200600110-00006>
- Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach (4th ed.)*. Pearson.

- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78. doi:10.1037/0003-066X.55.1.68
- Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, 85 (1), 112–118. <https://doi.org/10.1037/0021-9010.85.1.112>
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2 (6), 110–114. <https://doi.org/10.2307/3002019>
- Schmidt, F. L., Oh, I.-S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology*, 59 (2), 281–305. <https://doi.org/10.1111/j.1744-6570.2006.00065.x>
- Schmucker, R. (2024). *Sequence-modeling for assessments and interventions in intelligent tutoring systems* (Doctoral dissertation, Carnegie Mellon University). https://www.cs.cmu.edu/research/phd-dissertation-pdfs/rschmuck_phd_mld-2024.pdf
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, c332. <https://doi.org/10.1136/bmj.c332>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Shumeiko, N., & Osadcha, K. (2024). Application of artificial intelligence in higher education institutions for developing soft skills of future specialists in the sphere

- of information technology. *Journal of Physics Conference Series*, 2871(1), 012027. <https://doi.org/10.1088/1742-6596/2871/1/012027>
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189.
- Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, 46(5), 30–32.
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107–120.
<https://doi.org/10.1007/s11336-008-9101-0>
- Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991–1006.
- Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3, 74.
<https://doi.org/10.3389/frobt.2016.00074>
- Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 6(6), 603–616.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.)*. Sage.
- Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, 42(4), 73–93.
- Suresh, A. (2022). *Automating feedback to improve teachers' effective use of instructional discourse in K–12 mathematics classrooms* (Doctoral dissertation, University of Colorado Boulder). ProQuest Dissertations & Theses Global.
<https://www.proquest.com/openview/15e25cc3c1547a2e1390d398d1ea42fd/1.pdf?pq-origsite=gscholar&cbl=18750&diss=y>

- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. doi:10.1207/s15516709cog1202_4
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer.
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. doi:10.1023/A:1022193728205
- Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics (7th ed.)*. Pearson.
- Tashakkori, A., & Teddlie, C. (2010). *SAGE handbook of mixed methods in social & behavioral research (2nd ed.)*. SAGE.
- Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
- Trisnawati, W., Putra, R., & Balti, L. (2023). The impact of artificial intelligence in education toward 21st century skills: a literature review. *PIJED*, 2(2), 501-513. <https://doi.org/10.59175/pijed.v2i2.152>
- Twisk, J. W. R. (2013). *Applied longitudinal data analysis for epidemiology: A practical guide (2nd ed.)*. Cambridge University Press.
- UNESCO. (2021). *Recomendación sobre la ética de la inteligencia artificial*. <https://www.unesco.org/es/articles/recomendacion-sobre-la-etica-de-la-inteligencia-artificial>
- UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO.
- Universidad de Investigación e Innovación de México. (s. f.-a). *Guía de estudio 7: Seminario de Tesis Doctoral III* [Material del curso]. Doctorado en Educación e Innovación.
- Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of*

- Clinical Epidemiology*, 59(9), 920–925.
<https://doi.org/10.1016/j.jclinepi.2006.02.007>
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221.
<https://doi.org/10.1080/00461520.2011.611369>
- Varsik, S., & Vosberg, L. (2024). *The potential impact of artificial intelligence on equity and inclusion in education* (OECD Artificial Intelligence Papers No. 23). OECD Publishing. <https://doi.org/10.1787/15df715b-en>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30).
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality*, 64(1), 1–48. <https://doi.org/10.1111/j.1467-6494.1996.tb00813.x>
- White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), 377–399.
- Wiggins, G. (2012). Seven keys to effective feedback. *Educational Leadership*, 70(1), 10–16.
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594–604.
- Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., Haddock, S. H. D., Huff, K. D., Mitchell, I. M., Plumbley, M. D., Waugh, B., White, E. P., & Wilson, P. (2014). Best practices for scientific computing. *PLOS Biology*, 12(1), e1001745. <https://doi.org/10.1371/journal.pbio.1001745>

- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 3087. <https://doi.org/10.3389/fpsyg.2019.03087>
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data (2nd ed.)*. MIT Press.
- Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach (6th ed.)*. Cengage Learning.
- Woolf, B. P. (2010). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. <https://doi.org/10.1186/s41239-019-0171-0>
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. En M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70.

APÉNDICE A

INSTRUMENTOS, CONTROL DE CALIDAD Y DOCUMENTACIÓN DE CAMPO

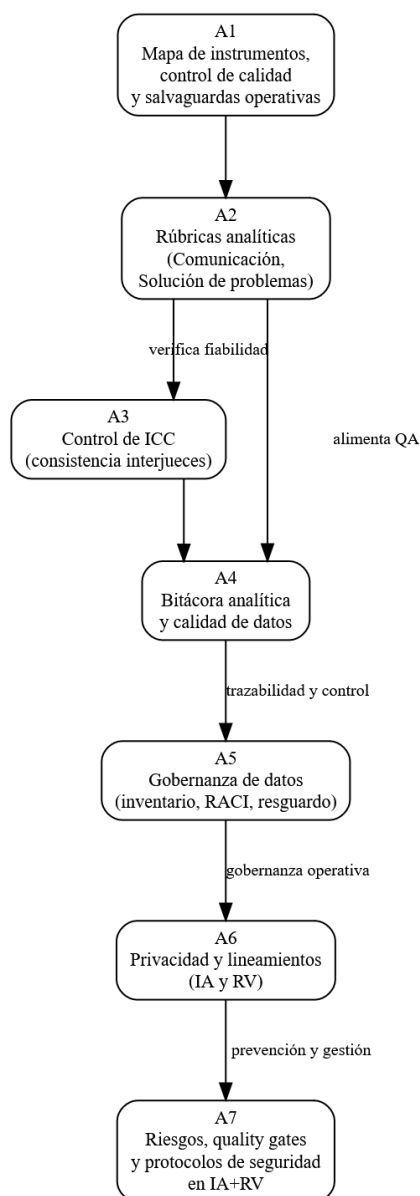
Este apéndice consolida la evidencia operativa que soporta la validez interna del estudio, la consistencia de la medición y el cumplimiento de criterios mínimos de gobernanza de datos en un entorno de intervención con inteligencia artificial y realidad virtual. Su propósito no es ampliar el marco teórico ni duplicar la metodología, sino documentar, de manera auditable, los instrumentos utilizados, los procedimientos de calibración interjueces, los controles de integridad de datos y los protocolos de operación segura. En conjunto, los elementos A1 a A7 permiten rastrear la secuencia completa desde la autorización y consentimiento, hasta la generación de evidencias evaluables, su aseguramiento de calidad y las salvaguardas de privacidad, integridad académica, trazabilidad y seguridad en sesiones inmersivas. Esta documentación se presenta con un criterio de mínimo suficiente, priorizando claridad, replicabilidad y control de riesgos para su eventual escalamiento institucional.

A1. Síntesis de instrumentos, control de calidad y salvaguardas operativas

El presente apartado introduce la arquitectura documental del Apéndice A y sintetiza la relación funcional entre los instrumentos de medición, los controles de calidad y las salvaguardas operativas del estudio. Su propósito es ofrecer una vista integrada de los componentes que sostienen la validez interna, la consistencia de la medición, la integridad de los datos y la operación segura de la intervención con inteligencia artificial y realidad virtual. En esta lógica, el mapa organiza la secuencia que conecta las rúbricas analíticas, la verificación de consistencia interjueces, la bitácora analítica y los controles de gobernanza, privacidad, riesgos y operación inmersiva, mostrando cómo cada elemento contribuye a la trazabilidad y auditabilidad del estudio sin duplicar la exposición metodológica del cuerpo principal de la tesis.

Figura A1.1

Mapa de instrumentos, control de calidad y salvaguardas operativas



Nota. La figura resume la estructura del Apéndice A: instrumentos de medición (A2), control de calidad de la calificación (A3), trazabilidad y de calidad de datos (A4), y salvaguardas de gobernanza, privacidad, riesgos y operación segura en IA+RV (A5 a A7), organizados a partir del mapa integrador presentado en A1. Su propósito es ofrecer una visión auditable del soporte metodológico y operativo del estudio sin duplicar la metodología principal.

A2. Rúbricas y guías de aplicación

Las rúbricas analíticas constituyeron el instrumento central para evaluar el desempeño en las dos competencias objetivo del estudio: comunicación oral y solución de problemas/decisión. Su función fue traducir constructos complejos en criterios observables, con anclajes explícitos de desempeño y una escala común que permitiera comparabilidad entre momentos de medición y entre evaluadores. En este sentido, las rúbricas no solo operaron como herramienta de calificación, sino también como mecanismo de estandarización metodológica, al fijar el lenguaje común de calidad que orientó tanto la evaluación como la retroalimentación.

La inclusión de dos rúbricas diferenciadas responde a la naturaleza multidimensional de las variables dependientes. La primera capturó aspectos de claridad, estructura, evidencia, adecuación a la audiencia, fluidez y lenguaje no verbal en presentaciones orales. La segunda evaluó la capacidad para definir problemas, generar alternativas, ponderar consecuencias, justificar decisiones, contextualizar soluciones e integrar principios e impacto social. En ambos casos, la escala 0–4 permitió conservar sensibilidad diagnóstica a nivel de criterio, mientras que el reescalamiento a 0–100 facilitó la interpretación comparativa y el análisis cuantitativo posterior.

Dado que la validez del estudio depende no solo de la pertinencia de los criterios sino también de la consistencia con que fueron aplicados, la documentación de las rúbricas se complementa en el siguiente apartado con el control de doble calificación ciega y la verificación de consistencia interjueces mediante ICC. De este modo, el apéndice no presenta únicamente los instrumentos, sino también la evidencia mínima de que su aplicación fue estable, trazable y metodológicamente defendible.

Tabla A2.1*Rúbrica analítica de comunicación oral (escala 0–4; puntaje global 0–100)*

Criterio	0 – Insuficiente	1 – Básico	2 – Aceptable	3 – Bueno	4 – Excelente	Peso (%)
Claridad y coherencia	Idea principal ausente o confusa; digresiones frecuentes.	Idea general perceptible pero inconsistente.	Secuencia comprensible con algunas rupturas.	Flujo lógico con mínimas rupturas.	Discurso cristalino; hilo conductor sostenido.	20
Estructura argumental	Sin apertura/cierre; no hay tesis ni soporte.	Apertura/cierre débiles; soporte escaso.	Tesis clara; soporte moderado.	Estructura completa; nexos claros.	Estructura impecable; transiciones precisas.	20
Uso de evidencia	Afirmaciones sin sustento.	Evidencia poco pertinente o débil.	Evidencia pertinente; referencias parciales.	Evidencia variada y pertinente.	Evidencia sólida y bien integrada (datos/citas).	20
Adecuación a la audiencia	Registro inadecuado; desconexión con público.	Registro parcialmente adecuado.	Registro adecuado con ajustes mínimos.	Ajustes claros a necesidades del público.	Excelente adaptación (términos, ejemplos, tono).	15
Fluidez y manejo de muletillas	Pausas excesivas; muletillas reiteradas.	Fluidez irregular; muletillas frecuentes.	Fluidez suficiente; muletillas ocasionales.	Fluidez constante; escasas muletillas.	Fluidez sobresaliente; muletillas mínimas.	15
Lenguaje no verbal	Contacto visual nulo; postura cerrada.	Contacto visual limitado; gestos pobres.	Contacto visual y gestos adecuados.	No verbal consistente y de apoyo.	No verbal estratégico que potencia el mensaje.	10

Nota. El puntaje global se calculó en escala 0–100 como: $[\sum(w_i \times c_i) / 4] \times 100$, donde w_i representa el peso proporcional del criterio i y la suma ponderada produce un índice en rango 0–4 reescalado a 0–100. La rúbrica, adaptada de marcos VALUE, se aplicó en PRE y POST con doble calificación ciega; el puntaje final corresponde al promedio entre jueces. La consistencia interjueces se estimó mediante ICC (2, k) bajo modelo de efectos aleatorios y acuerdo absoluto, con umbral mínimo de aceptación $\geq .75$ (preferible $\geq .80$). $Puntaje = (\sum_{i=1}^6 nivel_i(0 - 4) \times w_i) \times 25$

Tabla A2.2*Rúbrica analítica de solución de problemas (escala 0–4; puntaje global 0–100)*

Criterio	0 – Insuficiente	1 – Básico	2 – Aceptable	3 – Bueno	4 – Excelente	Peso (%)
Definición del problema	Problema mal definido o equivocado.	Definición parcial; omite condiciones.	Definición adecuada con supuestos básicos.	Definición precisa con supuestos explícitos.	Definición robusta con límites y contexto.	15
Generación de alternativas	Una alternativa o triviales.	Pocas alternativas; escasa originalidad.	Varias alternativas plausibles.	Alternativas diversas y factibles.	Portafolio creativo y viable.	15
Evaluación de alternativas	Sin criterios; selección arbitraria.	Criterios implícitos o pobres.	Criterios explícitos básicos.	Criterios relevantes y ponderados.	Evaluación rigurosa con trade-offs.	15
Consideración de consecuencias	No anticipa efectos.	Efectos parciales y superficiales.	Efectos principales identificados.	Efectos de corto/mediano plazo.	Efectos sistémicos y de largo plazo.	15
Justificación de la decisión	Sin justificación.	Justificación débil.	Justificación coherente.	Justificación sólida con evidencia.	Justificación robusta y replicable.	15
Contextualización	Ignora restricciones reales.	Considera pocas restricciones.	Considera restricciones básicas.	Integra restricciones y recursos.	Integra contexto, actores y viabilidad.	15
Principios/impacto social	No considera ética/impacto.	Menciona ética superficialmente.	Considera principios mínimos.	Analiza riesgos/beneficios sociales.	Integra ética/impacto o en decisión.	10

Nota. El puntaje global se calculó como promedio ponderado de los siete criterios (pesos en %), reescalado a 0–100. La calificación se realizó mediante doble evaluación ciega; el puntaje final corresponde al promedio entre jueces. Se estableció como criterio de calidad de medición un umbral mínimo de fiabilidad interjueces ICC (2, k) $\geq .75$. Los anclajes conductuales por nivel (0–4) y ejemplos de desempeño se documentaron en una guía de evaluación para estandarizar la aplicación de la rúbrica y apoyar la calibración entre evaluadores.

A3. Control de calidad de la calificación e ICC (consistencia interjueces)

Este apéndice documenta tanto el formato operativo utilizado para registrar la doble calificación ciega y las discrepancias entre jueces, como el resumen de los índices de consistencia interjueces que sustentan la aceptación de las rúbricas para fines analíticos. La primera tabla funciona como hoja de control de calificación y resolución de discrepancias; la segunda sintetiza los valores observados de fiabilidad interjueces mediante ICC (2,2), con su respectivo intervalo de confianza y criterio de decisión.

Tabla A3.1

Formato de control de doble calificación y discrepancias entre jueces

ID producto	Criterio	Juez A (0–100)	Juez B (0–100)	Δ (abs)	$\Delta \geq 10$? (sí/no)	Observación
	Claridad/coherencia					
	Estructura argumental					
	Evidencia/soporte					
	Adecuación a audiencia					
	Fluidez/muletillas					
	No verbal					

Nota. Formato operativo utilizado para registrar la doble calificación ciega, identificar discrepancias absolutas entre jueces y activar, cuando correspondió, el protocolo de revisión o recalibración.

Tabla A3.2*Resumen de consistencia interjueces e índice ICC (2,2)*

Instrumento	n / observaciones (PRE+POST)	ICC (2,2)	IC95%	Criterio de aceptación	Decisión
Rúbrica de Comunicación	320	0.903	[0.882, 0.920]	≥ 0.75	Aceptada
Rúbrica de Solución de problemas/decisión	320	0.921	[0.907, 0.934]	≥ 0.75	Aceptada

Nota. El valor $n = 320$ corresponde al número total de observaciones calificadas en la muestra analítica per-protocol, derivado de 160 estudiantes con dos momentos de medición (PRE y POST). Cada observación fue evaluada por dos jueces independientes y ciegos; por ello, se reporta ICC (2,2) como estimador principal de acuerdo absoluto, dado que el puntaje operativo utilizado en el análisis corresponde al promedio de ambos evaluadores. El criterio mínimo de aceptación se fijó en 0.75, en consonancia con recomendaciones metodológicas para estudios aplicados. Los valores observados sustentan la estabilidad del proceso de calificación y la idoneidad de las rúbricas para su uso en la inferencia principal.

A4. Bitácora analítica, integridad y de calidad de datos

Este apartado documenta el componente de aseguramiento de calidad asociado a la evaluación del desempeño mediante doble calificación ciega. Su propósito es demostrar que la aplicación de las rúbricas no dependió de juicios aislados o idiosincráticos, sino de un procedimiento verificable de consistencia entre evaluadores.

Tabla A4.1

Checklist de integridad y coherencia de datos por corte temporal (PRE/POST)

Corte	Integridad campos obligatorios (%)	Rango válido (min IA/RV)	Productos con 1 juez (%)	Faltantes PRE/POST (%)	Acción
PRE					
POST					

Nota. La tabla resume los controles mínimos de integridad y coherencia aplicados por corte temporal (PRE y POST) antes de la construcción de la base analítica. La integridad de campos obligatorios se refiere al porcentaje de registros con variables esenciales completas; rango válido verifica que los valores observados en minutos de IA/RV se ubiquen dentro de los dominios permitidos; productos con 1 juez identifica casos con doble calificación incompleta; y faltantes PRE/POST resume la proporción de datos ausentes por corte. La columna Acción documenta la decisión operativa adoptada en cada caso, por ejemplo, validación, corrección, exclusión técnica o envío a revisión. Esta *checklist* funcionó como control previo de calidad para asegurar consistencia entre captura, calificación y procesamiento.

Tabla A4.2

Registro de imputación múltiple (si MAR; $m=20$)

Fecha	Variables incluidas	Método	Seed	Diagnóstico convergencia	Observación
	PRE, grupo, programa, dosificación	MICE		OK / Revisar	

Nota. La imputación múltiple se aplicó únicamente cuando el patrón de faltantes fue consistente con MAR y la tasa de faltantes superó el umbral operativo definido; en caso contrario, se reportó análisis de casos completos con verificación de sensibilidad.

Tabla A4.3

Plantilla de cambios al plan analítico y justificación (prerregistro vs. ejecución)

Fecha	Cambio	Justificación	Impacto en resultados

Nota. La bitácora garantizó trazabilidad (quién, qué, cuándo, por qué) de cada decisión analítica.

A5. Gobernanza de datos, roles y resguardo

La gobernanza de datos se operacionalizó como un conjunto mínimo de reglas y responsabilidades para asegurar integridad analítica, control de accesos y trazabilidad durante todo el ciclo de vida del estudio. Dado que la intervención incorpora componentes digitales (IA y RV) y produce tanto mediciones de desempeño (rúbricas) como registros operativos y de práctica, se adoptó un enfoque de minimización y seudonimización: se recolectó únicamente lo necesario para responder H1 a H4 y para documentar calidad de implementación, evitando capturar información personal no indispensable. En este marco, la evidencia de gobernanza se presenta como inventario de datos y clasificación (Tabla A5.1), matriz de responsabilidades RACI (Tabla A5.2), reglas de resguardo y trazabilidad (Tabla A5.3) y un registro mínimo de accesos recomendado para escenarios de escalamiento institucional (Tabla A5.4).

La Tabla A5.1 establece la arquitectura de datos del estudio desde una perspectiva de riesgo y control. Para cada conjunto de datos se explicita la fuente, el tipo de variables, el nivel de sensibilidad y la condición de identificabilidad, junto con el rol autorizado y el periodo de retención. Esta representación reduce ambigüedad operativa y permite verificar el principio de mínimo privilegio, al tiempo que clarifica qué insumos pueden reproducirse en formato agregado o anonimizado para auditoría académica.

Tabla A5.1*Inventario de datos y clasificación*

Conjunto de datos	Fuente	Contenido/ variables	Nivel de sensibilidad	Identificabilidad	Acceso (rol)	Retención
Rúbricas (PRE/POST)	Evaluadores (2 jueces) + rúbrica analítica	Puntajes por criterio y total (0–100)	Alta	Seudonimizado (ID de participante)	IP; Evaluadore s; Analista	Hasta cierre de tesis + 12 meses
Escalas breves (presencia, utilidad, autoeficacia)	Cuestionarios en línea/papel	Ítems, puntaje total, metadatos de aplicación	Media	Seudonimizado (ID)	IP; Analista	Hasta cierre de tesis + 12 meses
Registros de práctica (VirtualSpeech)	Plataforma	Minutos de práctica, sesiones, métricas de desempeño y reportes	Media	Seudonimizado / Agregado	IP; Analista	Mientras dure el análisis; luego agregado
Bitácoras docentes y checklists de fidelidad	Docente/Obs ervador	<i>briefing/debri efing</i> , rúbrica, señalización IA/RV, incidencias	Media	No identificable (por clase) / Seudonimizado	IP; Docente	Hasta cierre de tesis + 12 meses
Entrevistas/observaci ón (cualitativo)	Investigador	Notas o transcripción (sin datos sensibles)	Alta	Seudonimizado; redacción sin PII	IP	Hasta cierre de tesis + 12 meses

Nota. La tabla siguiente resume los conjuntos de datos utilizados, su sensibilidad y los controles de acceso aplicados durante el ciclo de vida del estudio.

La Tabla A5.2 traduce la gobernanza a asignaciones verificables de responsabilidad. Se priorizó una estructura simple que separa roles críticos: evaluación (jueces), análisis (analista o IP) y resguardo (IP), evitando solapamientos que suelen generar errores de control de versiones, accesos no documentados o incertidumbre sobre aprobaciones. Esta matriz funciona como control preventivo y como evidencia de quién ejecuta, quién aprueba y quién debe estar informado en tareas sensibles.

Tabla A5.2*Matriz RACI de gobernanza y responsabilidades*

Actividad crítica	R (Responsable)	A (Aprobador)	C (Consultado)	I (Informado)
Seudonimización y llave de IDs	IP	IP	TI/Soporte (si aplica)	Evaluadores
Resguardo y control de versiones	IP	IP	TI/Soporte (si aplica)	Evaluadores
Calificación con rúbricas (doble ciego)	Evaluadores	IP	—	—
Limpieza, QA e imputación	Analista / IP	IP	—	—
Atención de solicitudes ARCO (si aplican)	IP	Responsable de datos institucional (si aplica)	Jurídico/Protección de datos (si aplica)	Participantes
Publicación de resultados y materiales	IP	IP	Asesor/Comité (si aplica)	Institución

Nota. La gobernanza operativa se organizó con una asignación mínima de responsabilidades (RACI) para evitar ambigüedad durante captura, resguardo, análisis y publicación.

Las reglas de resguardo y trazabilidad se definieron para sostener tres objetivos: (i) reducir el riesgo de reidentificación mediante separación de llaves y minimización de variables, (ii) asegurar integridad y reproducibilidad con control de versiones y bitácora de cambios, y (iii) acotar el periodo de conservación a lo estrictamente necesario para cierre académico, revisión y réplica interna. En conjunto, estas reglas operan como un sistema de aseguramiento interno que permite auditar decisiones analíticas y operativas sin trasladar datos identificables a terceros ni ampliar innecesariamente el perímetro de riesgo.

Tabla A5.3*Reglas de resguardo, retención y trazabilidad*

Acción	Descripción
Separación de llaves: la correspondencia	ID ₁ identidad (si existió) se mantuvo separada del se analítico.
Control de versiones	Archivos analíticos y outputs se gestionaron con convenciones de nombre (fecha–versión) y bitácora de cambios.
Minimización	Se recolectaron únicamente variables necesarias para responder H1–H4 y para QA/fidelidad.
Retención:	Se define un periodo acotado; al término, se conserva únicamente evidencia agregada o anonimizada para auditoría académica.
Transferencias	No se realizan transferencias a terceros fuera del ecosistema institucional, salvo licencias/plataformas educativas utilizadas para la práctica.

Nota. Estas reglas operan como salvaguardas de gobernanza y control interno: reducen el riesgo de reidentificación, sostienen la integridad del flujo analítico y permiten auditoría de decisiones (ingesta → depuración → modelado → salidas) mediante trazabilidad documental. La retención se limita al periodo estrictamente necesario para cierre académico, revisión y réplica interna; concluido dicho periodo, se conserva únicamente evidencia agregada o anonimizada (p. ej., tablas, estimaciones y bitácoras) y se eliminan insumos identificables conforme al protocolo de resguardo. Las transferencias quedan restringidas al ecosistema institucional y a plataformas educativas usadas para la práctica, sin cesión de datos a terceros para fines comerciales o distintos de los autorizados.

Aunque el estudio se ejecutó en un ecosistema institucional controlado, se incorpora una plantilla de registro mínimo de accesos como mecanismo recomendado para escalamiento. Su inclusión enfatiza la trazabilidad de consultas, ediciones o exportaciones de archivos analíticos y outputs, y apoya la rendición de cuentas cuando la implementación involucra múltiples cursos, laboratorios o equipos docentes. En la tesis, este instrumento se presenta como buena práctica de gobernanza aplicable en contextos de adopción ampliada.

Tabla A5.4*Formato de registro mínimo de accesos (plantilla)*

Fecha	Repositorio/ archivo	Rol/usuario	Acción (lectura/edición/exportación)	Motivo	Observaciones
_____	_____	_____	_____	_____	_____

Nota. Esta plantilla documenta el registro mínimo de accesos a repositorios y archivos con información del estudio (datos, bitácoras, outputs y materiales analíticos), con fines de trazabilidad, control de versiones y rendición de cuentas. Se completa cada vez que un rol autorizado consulta, edita o exporta información, especificando el motivo y cualquier observación relevante (p. ej., corrección de error, actualización de versión, extracción de tabla para reporte). En la tesis, este nivel de registro se mantiene como mecanismo de gobernanza recomendado para escenarios de escalamiento institucional; su aplicación prioriza el principio de mínimo privilegio y el seguimiento de cambios que puedan afectar reproducibilidad, integridad y cumplimiento de lineamientos de privacidad.

A6. Avisos de privacidad y lineamientos de uso (IA y RV)

La incorporación de IA y RV en un entorno educativo exige delimitar con precisión el alcance del tratamiento de datos, el rol de la tecnología y las reglas de integridad académica. En esta intervención, la IA se definió como apoyo formativo para retroalimentación, mientras que la evaluación sumativa se mantuvo bajo control humano mediante rúbricas analíticas aplicadas por jueces independientes. En paralelo, se establecieron lineamientos de privacidad y uso responsable para reducir riesgos de exposición de datos, minimizar sesgos no detectados y evitar prácticas de sustitución de desempeño. La evidencia de cumplimiento se presenta como aviso de privacidad simplificado (Tabla A6.1), lineamientos para estudiantes (Tabla A6.2), *checklist* de comprensión al inicio del curso (Tabla A6.3) y declaración breve para materiales y syllabus (Tabla A6.4).

El aviso de privacidad simplificado se diseñó para comunicar, en lenguaje claro, cuatro elementos operativos: responsable del tratamiento, finalidades específicas, categorías de datos y condiciones generales de resguardo y transferencias. En coherencia con el principio de minimización, se enfatiza la preferencia por seudonimización y el acceso restringido por rol. Cuando se emplean plataformas educativas para práctica, el aviso delimita que dichas plataformas pueden procesar registros de uso conforme a sus propios términos, sin que ello implique una transferencia deliberada de datos personales por parte del estudio fuera del marco institucional.

Tabla A6.1

Aviso de privacidad simplificado para participantes (plantilla)

Concepto	Descripción
Responsable del tratamiento:	Tecnológico de Monterrey (unidad académica responsable del curso/laboratorio).
Finalidad:	(i) Evaluar el desempeño PRE–POST mediante rúbricas; (ii) analizar percepciones de presencia, utilidad del <i>feedback</i> y autoeficacia; (iii) documentar fidelidad de implementación; y (iv) generar reportes académicos y productos de investigación.
Datos tratados	Identificadores académicos mínimos (si aplican), respuestas a instrumentos, puntajes de rúbricas y registros de práctica; los datos se trabajan preferentemente seudonimizados
Transferencias:	No se comparten datos personales con terceros; las plataformas educativas pueden procesar registros de uso bajo sus propios términos.
Derechos ARCO	El participante puede solicitar Acceso, Rectificación, Cancelación u Oposición conforme a la LFPDPPP, mediante el canal institucional indicado.
Contacto:	[correo institucional del investigador / unidad de protección de datos].

Nota. Plantilla de aviso de privacidad simplificado para informar a las y los participantes, en lenguaje claro, sobre el responsable del tratamiento, las finalidades específicas del estudio, las categorías de datos utilizadas y las condiciones generales de transferencia y resguardo. El aviso se entrega junto con el consentimiento informado y se enfatiza que los datos se tratan preferentemente seudonimizados, bajo criterios de minimización (solo lo necesario para los objetivos H1–H4 y QA/fidelidad), acceso restringido y conservación acotada. Cuando se empleen plataformas educativas externas para la práctica, se aclara que dichas plataformas pueden procesar registros de uso conforme a sus propios términos, sin que ello implique compartir datos personales por parte del estudio fuera del ecosistema institucional. Se incluye el canal para ejercer derechos ARCO y el contacto institucional para dudas o aclaraciones.

Los lineamientos de uso responsable se establecieron como condición de participación para proteger integridad académica y claridad pedagógica. El criterio central es separar apoyo formativo de evaluación: la IA puede orientar iteraciones, clarificar criterios y ofrecer retroalimentación, pero no reemplaza la autoría del estudiante ni genera productos evaluables. Esta delimitación reduce ambigüedad sobre uso permitido y uso indebido, y se complementa con trazabilidad mínima de versiones para sustentar reflexión y *debriefing*, sin expandir la recolección de datos.

Tabla A6.2

Lineamientos de uso responsable de IA (estudiantes)

Lineamiento	Descripción
Uso de IA	La IA se usa exclusivamente como apoyo formativo (<i>feedback</i>), no para generar productos evaluables ni sustituir desempeño
Se prohíbe	Se prohíbe copiar/pegar texto generado por IA como respuesta final de actividades evaluadas; el uso indebido se gestiona conforme a políticas de integridad académica.
Transparencia	Cuando el estudiante utilice IA fuera del entorno de <i>feedback</i> (p. ej., para clarificar criterios), deberá declararlo si la actividad lo solicita
Privacidad	No ingresar datos personales, sensibles o identificables de terceros en <i>prompts</i> o herramientas externas.
Trazabilidad	conservar evidencia mínima del ciclo (versión del trabajo + <i>feedback</i> recibido + cambios realizados) para apoyar reflexión y <i>debriefing</i> .

Nota. Los lineamientos se entregaron y explicaron a los estudiantes como condición de participación, con el fin de asegurar que la IA funcionara solo como andamiaje formativo y no como sustituto del desempeño evaluable. Se consideró uso permitido el apoyo para comprender criterios, recibir retroalimentación y orientar iteraciones; y uso indebido, presentar contenido generado por IA como producto final o eludir la autoría del estudiante. Cuando la actividad lo requirió, se exigió declarar apoyos externos. En todos los casos se mantuvo el principio de minimización de datos, evitando ingresar información personal o identificable en herramientas externas. La trazabilidad solicitada se limitó a evidencia mínima de versión inicial, *feedback* recibido y cambios realizados, únicamente para sustentar reflexión y *debriefing*.

El *checklist* de comprensión opera como evidencia de comunicación efectiva de reglas básicas y como control preventivo ante incidentes o incumplimientos. Su propósito no es evaluar aprendizaje, sino documentar que el estudiante recibió inducción operativa sobre seguridad en RV, privacidad y uso permitido de IA. La evidencia se conserva de forma mínima, únicamente para verificar comunicación de lineamientos y apoyar acciones correctivas cuando sea necesario.

Tabla A6.3*Inducción operativa (IA+RV) al inicio del curso: checklist de comprensión*

Item	Comprendido (Sí/No)	Dudas / comentarios	Firma (iniciales)
Uso de visores (ajuste, pausas, higiene) y protocolo ante ciber-mareo.	_____	_____	_____
Propósito del uso de IA: <i>feedback</i> formativo (no generación de productos evaluables).	_____	_____	_____
Reglas de privacidad: no capturar ni compartir información identificable.	_____	_____	_____
Evidencia de trabajo: conservar versiones y retroalimentación para <i>debriefing</i> .	_____	_____	_____
Canal de reporte de incidentes o problemas técnicos.	_____	_____	_____

Nota. Formato para documentar que el estudiante recibió inducción operativa y comprende reglas básicas (seguridad, privacidad e integridad).

La declaración breve se incorpora en materiales del curso como recordatorio recurrente del rol tecnológico y del esquema human-in-the-loop. Su presencia reduce el riesgo de malentendidos, refuerza el principio de privacidad (no ingresar datos personales en herramientas externas) y sostiene consistencia comunicativa entre syllabus, Canvas y actividades, lo que facilita fidelidad de implementación y cumplimiento institucional.

Tabla A6.4

Declaración breve para materiales del curso

“En este curso utilizaremos realidad virtual (RV) y un tutor basado en IA como apoyo formativo. La IA se emplea únicamente para retroalimentación inmediata; la evaluación final la realiza el profesorado con rúbricas analíticas. No ingreses datos personales en herramientas externas y respeta las reglas de integridad académica.”

Nota. Texto modelo para incorporar en syllabus, Canvas o comunicados del curso, con el propósito de asegurar transparencia sobre el rol de la RV y la IA, delimitar su uso a apoyo formativo (retroalimentación inmediata) y reafirmar que la evaluación sumativa se realiza por el profesorado mediante rúbricas analíticas. La declaración funciona como recordatorio de privacidad y minimización (no ingresar datos personales en herramientas externas), de integridad académica (uso permitido/prohibido) y de la condición *human-in-the-loop* en decisiones de alto impacto académico. Puede colocarse al inicio del curso y repetirse previo a actividades evaluables, como control preventivo y evidencia de comunicación de lineamientos.

A7. Matriz de riesgos y *quality gates* y operación segura en IA+RV

La gestión de riesgos se integró como componente operativo del aseguramiento de calidad, con énfasis en privacidad, integridad académica, seguridad en realidad virtual y trazabilidad de configuración del tutor basado en inteligencia artificial. En lugar de tratar el riesgo como un apéndice abstracto, se construyó un registro mínimo de riesgos con controles y evidencia asociada, y se definieron *quality gates* como umbrales verificables para decidir continuidad, ajuste o escalamiento. Este enfoque permite evaluar la adopción con evidencia, evitando decisiones basadas únicamente en percepción o disponibilidad tecnológica.

De manera complementaria, la realidad virtual introduce riesgos operativos y de bienestar que exigen procedimientos explícitos para preservar seguridad, continuidad pedagógica y comparabilidad de implementación. En este estudio, los protocolos se definieron como controles operativos y no como instrumentos de medición. Su función fue estandarizar condiciones mínimas de uso, reducir incidentes, documentar la dosis efectiva de exposición y asegurar alternativas equivalentes cuando el uso del visor no fuera viable por malestar, accesibilidad o fallas técnicas.

En conjunto, la evidencia de este apartado se organiza en dos planos complementarios. El primero corresponde al control preventivo y de decisión para la adopción del modelo, mediante la matriz de riesgos y los *quality gates* de implementación. El segundo corresponde a la operación segura de las sesiones inmersivas, mediante inducción breve, *checklist* por sesión, protocolo ante ciber-mareo, formato de incidente y lineamientos de higiene y cuidado del equipo. Esta integración permite documentar, en una misma secuencia, tanto la lógica de aseguramiento previo al escalamiento como las salvaguardas concretas de operación en aula.

Tabla A7.1

Matriz de riesgos (registro mínimo)

ID	Riesgo	Categoría	P	I	Nivel	Control/Mitigación	Evidencia
R1	Exposición de datos personales en <i>prompts</i> o plataformas	Privacidad	2	4	8	Lineamientos de minimización; inducción; seudonimización; prohibición de PII en <i>prompts</i>	A6.2–A6.3; A5.1
R2	Sesgo en <i>feedback</i> automatizado que afecte equidad	Equidad/Sesgo	2	4	8	Human-in-the-loop; rúbrica como criterio; revisión docente en <i>debriefing</i> ; monitoreo por cohorte	A6.2; A7.2
R3	Uso de IA para producir evidencias evaluables (integridad)	Integridad académica	3	4	12	Reglas explícitas; trazabilidad de versiones; evaluación humana; sanciones institucionales	A6.2; A6.4
R4	Ciber-mareo o malestar por RV	Seguridad/Salud	3	3	9	Protocolo RV (pausas, suspensión, alternativas); bitácora de incidentes	A7.5–A7.6
R5	Falla técnica (conectividad/headsets) afecta dosis y comparabilidad	Operación/Calidad	3	3	9	<i>Checklist</i> pre-sesión; plan B (desktop o alternativa equivalente); registro de dosis real	A7.4; A5.1
R6	Captura accidental de imágenes/audio identificable en aula	Privacidad	2	3	6	Política de no grabación; difuminado; consentimiento específico si aplica	Anexo 1; Anexo 2.1
R7	Deriva de <i>prompts</i> /configuración del tutor IA (cambios no documentados)	Trazabilidad	2	4	8	Registro de versión; control de configuración; bitácora de cambios	A5.3; A4.3
R8	Dependencia del proveedor (cambios de licencia, métricas)	Sostenibilidad	3	3	9	Contrato/licencia institucional; respaldo de rúbricas; rutas alternativas	Anexo 2.1; Sección 4.2.4
R9	Accesibilidad limitada para estudiantes con restricciones médicas	Equidad/Accesibilidad	2	4	8	Acomodaciones razonables y alternativas equivalentes; criterios de suspensión	A7.3; A7.5
R10	Interpretación errónea del <i>feedback</i> sin mediación pedagógica	Pedagógico	2	3	6	<i>debriefing</i> docente; rúbrica como lenguaje común; meta- <i>feedback</i>	A6.2; A2.1–A2.2

Nota. Registro de riesgos y controles para operar IA+RV con seguridad, equidad y calidad; incorpora *quality gates* para decisiones de escalamiento basadas en evidencia. Escala: Probabilidad (P) 1–5; Impacto (I) 1–5; Nivel = P×I. Los controles listados corresponden a mitigaciones implementadas o planificadas para escalamiento.

Los *quality gates* se diseñaron como una capa de decisión basada en umbrales mínimos, con evidencia requerida y acciones explícitas. Cada gate corresponde a una

dimensión crítica: operación y seguridad, fidelidad pedagógica, calidad de medición, integridad académica y equidad de acceso. El objetivo es convertir requisitos cualitativos en criterios verificables antes de ampliar el modelo a más cohortes o cursos.

Tabla A7.2

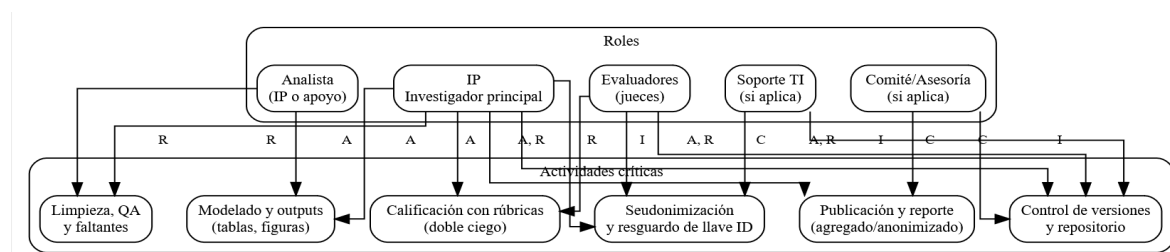
Quality gates para escalamiento (umbral–evidencia–decisión)

Gate	Criterio	Umbral mínimo	Evidencia requerida	Decisión
G1 (Operación)	Seguridad y logística RV	0 incidentes severos; $\geq 90\%$ <i>checklist</i> pre-sesión completo	Bitácora de incidentes	Continuar / ajustar
G2 (Fidelidad)	Ciclo <i>briefing</i> –simulación– <i>debriefing</i>	$\geq 80\%$ sesiones con <i>briefing</i> y <i>debriefing</i> documentados	<i>Checklists</i> por sesión; evidencia de <i>debriefing</i>	Continuar / reforzar capacitación
G3 (Medición)	Fiabilidad de rúbricas	ICC(2,k) $\geq .75$ y estabilidad por criterio	Hoja ICC + reporte de calibración	Aprobar medición / recalibrar
G4 (Integridad)	Uso responsable de IA	0 casos confirmados de sustitución de desempeño en evidencias evaluables	Declaraciones + revisión docente + trazabilidad de versiones	Escalar / reforzar reglas
G5 (Equidad)	Acceso y acomodaciones	100% estudiantes con alternativa equivalente cuando RV no sea viable	Registro de acomodaciones; plan alternativo	Escalar / corregir brecha

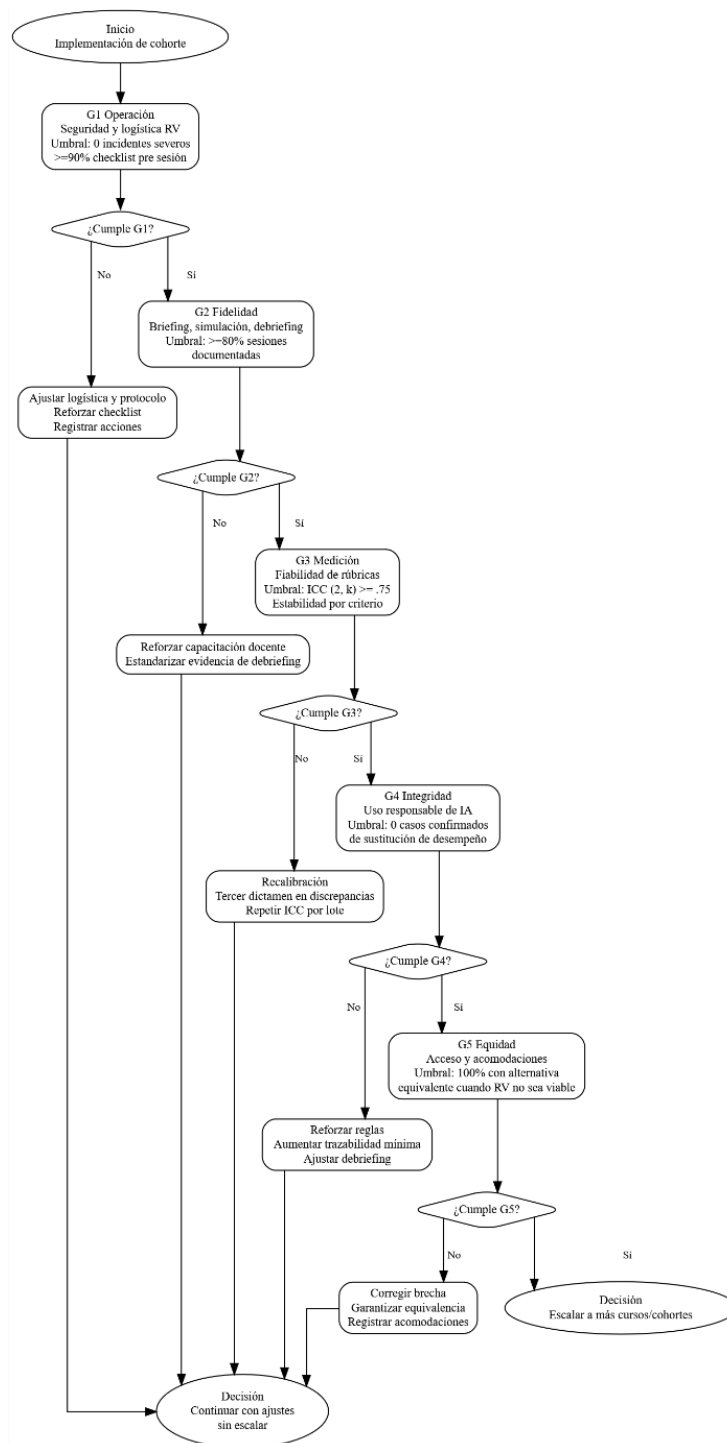
Nota. Los *quality gates* operacionalizan el criterio adopción con fidelidad y evidencia y definen umbrales mínimos antes de ampliar el modelo a más cursos/cohortes.

Figura A7.1

Gobernanza operativa y responsabilidades RACI en el ciclo de datos del estudio



Nota. La figura resume la asignación mínima de responsabilidades para evitar ambigüedad operativa y sostener trazabilidad. El IP conserva aprobación y control de resguardo y publicación; los evaluadores son responsables de la calificación doble ciego; el analista ejecuta QA, modelado y generación de outputs bajo aprobación del IP; y el soporte TI y la asesoría se incorporan solo como roles consultados cuando aplica.

Figura A7.2*Quality gates para continuidad y escalamiento del modelo IA y RV*

Nota. El diagrama convierte requisitos de adopción en umbrales verificables y evidencia mínima requerida. La lógica es conservadora: si un gate no cumple, se prioriza ajuste y mejora de fidelidad, seguridad, medición, integridad o equidad antes de escalar a más cohortes, con documentación explícita de acciones correctivas.

La inducción breve fija reglas de seguridad y criterios de pausa o suspensión. Este control reduce variabilidad por uso incorrecto del equipo y prioriza el bienestar del estudiante sin afectar la continuidad académica. En congruencia con el enfoque *human-in-the-loop*, el docente media la experiencia mediante pausas, monitoreo de malestar y activación de alternativas equivalentes.

Tabla A7.3

Inducción breve al inicio de práctica

Concepto	Descripción
Objetivo	Explicar uso seguro del visor y criterios de pausa/suspensión.
Antes de iniciar	Ajuste del visor, limpieza, postura segura y espacio libre.
Regla de seguridad	Si aparece mareo, náusea, sudoración, dolor de cabeza o desorientación, detener inmediatamente y levantar la mano.
Pausas programadas	Micro-pausa cada 10–15 minutos (o antes si el estudiante lo solicita)
Acomodaciones	Quien no pueda usar visor podrá realizar alternativa equivalente (desktop/simulador) sin penalización.

Nota. Protocolos de inducción, seguridad y registro de incidentes para sesiones con visores; incluye *checklist* operativo y alternativas equivalentes.

El *checklist* operativo por sesión documenta condiciones mínimas pre, durante y post, con el objetivo de reducir variaciones no deseadas en la implementación. Se registran verificación del entorno físico, funcionamiento técnico, pausas y monitoreo, así como el tiempo efectivo de práctica (dosis). Las observaciones capturan incidencias relevantes (por ejemplo, calibración, interrupciones o fallas) y permiten mejora iterativa sin convertir el *checklist* en una medición de desempeño.

Tabla A7.4

Checklist operativo por sesión (pre–durante–post)

Momento	Verificación	Cumple (Sí/No)	Observaciones
Pre	Espacio libre y señalizado; cableado seguro; limpieza de visor/controladores	—	_____
Pre	Audio/visual correcto; calibración; cuenta/plataforma funcional	—	_____
Durante	Pausas realizadas; monitoreo de malestar	—	_____
Durante	Registro de tiempo efectivo (dosis) y <i>debriefing</i> planificado	—	_____
Post	Limpieza; resguardo; registro de incidentes (si aplica)	—	_____

Nota. Lista de verificación operativa utilizada en cada sesión para estandarizar condiciones de implementación y reducir riesgos de seguridad y variabilidad no deseada. El *checklist* documenta controles mínimos antes, durante y después de la práctica inmersiva (condiciones físicas, funcionamiento técnico, pausas/monitoreo y registro de dosis), y sirve como evidencia de fidelidad operativa y de gestión preventiva. Las Observaciones capturan incidencias relevantes (p. ej., fallas técnicas, ajustes de calibración, interrupciones, malestar reportado) y apoyan la mejora iterativa del procedimiento sin convertir el *checklist* en un instrumento de medición de resultados.

El protocolo ante ciber-mareo estandariza la respuesta inmediata cuando el estudiante presenta síntomas compatibles con ciberenfermedad. Se prioriza interrupción segura, recuperación breve y registro mínimo de síntomas y disparadores, seguido de una decisión documentada sobre reincorporación controlada o suspensión. La alternativa equivalente se activa para sostener la continuidad pedagógica sin afectar evaluación.

Tabla A7.5*Protocolo ante ciber-mareo (ciberenfermedad)*

Paso	Acción
Paso 1	Detener la práctica y retirar el visor en un área segura.
Paso 2	Sentar al estudiante y ofrecer agua; evitar reincorporación inmediata.
Paso 3	Registrar síntomas, duración y disparadores (escena, movimiento, tiempo).
Paso 4	Evaluar si procede reincorporación con modo reducido (movimiento mínimo) o suspensión definitiva de la sesión.
Paso 5	Activar alternativa equivalente (desktop / actividad fuera de RV) para no afectar evaluación.

Nota. Procedimiento estandarizado para respuesta inmediata ante síntomas compatibles con ciber-mareo/ciberenfermedad durante prácticas en RV. Su objetivo es proteger la salud del estudiante, contener el riesgo y asegurar continuidad pedagógica mediante una alternativa equivalente (sin penalización académica) cuando la sesión de RV deba reducirse o suspenderse. El protocolo prioriza: (i) interrupción segura, (ii) recuperación y observación breve, (iii) registro de síntomas y posibles disparadores, (iv) decisión documentada sobre reincorporación controlada o suspensión, con trazabilidad para seguimiento y prevención y (v) activar la decisión tomada.

El formato de reporte de incidente constituye evidencia mínima de auditabilidad operativa. Se limita a la documentación estrictamente necesaria para gestionar el incidente y prevenir recurrencias, y no se utiliza para análisis estadístico de resultados ni para evaluar desempeño del estudiante. Su valor es de aseguramiento de proceso: permite identificar patrones de fallas técnicas o condiciones que incrementan malestar y, con ello, ajustar la logística o la escena.

Tabla A7.6*Formato de reporte de incidente (plantilla)*

Fecha y hora	_____
Curso/clase	_____
Descripción del incidente	_____
Acción tomada (pausa/suspensión/alternativa)	_____
Resultado/seguimiento	_____
Responsable que reporta (nombre y firma)	_____

Nota. Plantilla de reporte utilizada para registrar incidentes operativos o de seguridad asociados a la implementación (p. ej., malestar, fallas de equipo, interrupciones críticas o ajustes de seguridad). El formato estandariza la documentación mínima necesaria para auditabilidad, análisis de recurrencias y activación de medidas correctivas (cambios en escena, pausas, calibración, logística o alternativas). La información se limita a lo estrictamente necesario para gestión del incidente y mejora del proceso; no se emplea para análisis estadístico de resultados ni para evaluar desempeño del estudiante.

Los lineamientos de higiene y cuidado del equipo reducen riesgos sanitarios y preservan la vida útil del visor y controladores. También sostienen condiciones comparables entre sesiones. Cuando la higiene o el resguardo no puedan garantizarse, se activa una alternativa equivalente que preserve el objetivo instruccional sin penalización, manteniendo coherencia con el principio de equidad de acceso.

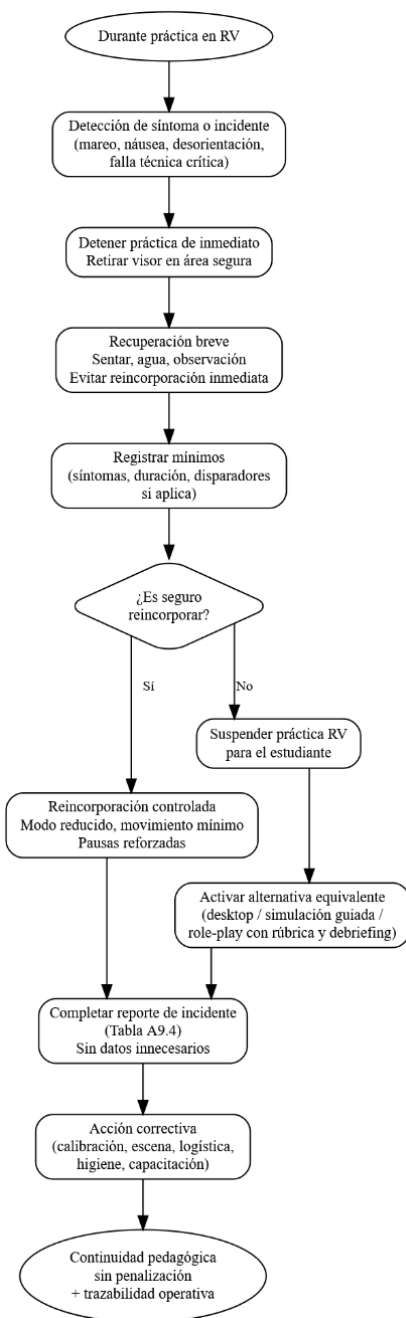
Tabla A7.7*Lineamientos de higiene y cuidado del equipo*

Componente / Control	Lineamiento operativo (higiene y cuidado del equipo)
Limpieza entre usuarios	Limpieza de superficies de contacto entre usos con insumos autorizados por laboratorio.
Barrera sanitaria (contacto facial)	Uso de cubiertas faciales desechables o sanitizables cuando esté disponible.
Resguardo y mantenimiento post-sesión	Resguardo en estuche y verificación de baterías/carga posterior a sesión.

Nota. Estos lineamientos se aplican antes y después de cada sesión para reducir riesgos sanitarios y preservar la vida útil del equipo (visor y controladores). El responsable de la sesión verifica su cumplimiento y registra incidencias cuando exista falta de insumos, daño del equipo o cualquier condición que impida la limpieza o el resguardo conforme al protocolo institucional. Cuando el uso de visor no sea viable (condición médica, malestar, accesibilidad o falla técnica), se activa una alternativa equivalente que preserve el objetivo instruccional y el criterio de evaluación (p. ej., versión desktop del escenario, simulación guiada o *role-play* con rúbrica y *debriefing*).

Figura A7.3

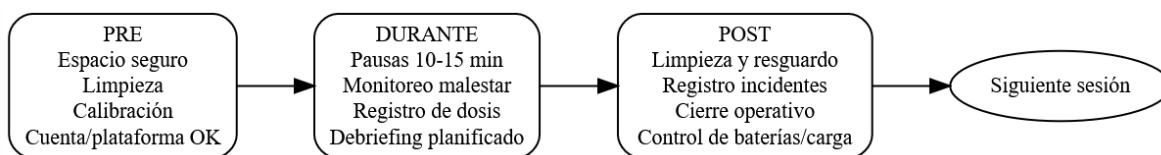
Diagrama de respuesta ante incidente durante práctica en Realidad Virtual y activación de alternativa equivalente



Nota. El mapa estandariza la respuesta ante incidentes para proteger bienestar, contener riesgo y asegurar continuidad pedagógica mediante una alternativa equivalente. La documentación se limita a lo indispensable para auditabilidad operativa y mejora del proceso.

Figura A7.4

Ciclo operativo pre–durante–post para estandarización de sesiones con RV



Nota. Representación del ciclo operativo mínimo para reducir variabilidad de implementación y documentar controles clave por sesión (seguridad, funcionamiento técnico, pausas, dosis, cierre e incidentes).

APÉNDICE B

CUADERNO DE ANÁLISIS REPRODUCIBLE (PYTHON/COLAB)

Este apéndice documenta el componente de reproducibilidad computacional del estudio. Su objetivo es asegurar que las transformaciones de datos, las verificaciones de calidad, los modelos estadísticos y la generación de tablas y figuras puedan ejecutarse nuevamente bajo un entorno controlado, con dependencias explícitas y un flujo de trabajo estable. El cuaderno reproducible se organiza como un *pipeline* analítico que inicia con la carga y validación de datos, continúa con controles de integridad y medición (incluida la fiabilidad e ICC), y culmina en la estimación de efectos y la exportación de productos en formato publicable. La trazabilidad se refuerza mediante bitácora mínima de ejecución y huella de corrida (hash), con el fin de distinguir entre cambios sustantivos y variaciones menores del entorno. En términos de gobernanza académica, este apéndice funciona como evidencia de transparencia metodológica y facilita tanto auditoría interna como replicación por terceros bajo las restricciones éticas y de privacidad aplicables.

B1. Entorno computacional

Este apartado especifica el entorno técnico en el que se ejecutó el análisis reproducible. Su propósito es dejar constancia de las versiones de Python, librerías, dependencias y configuraciones mínimas requeridas para regenerar las salidas del estudio con un grado razonable de consistencia. Al explicitar el entorno computacional, se reduce la ambigüedad asociada a diferencias entre plataformas, se facilita la réplica interna o externa y se fortalece la trazabilidad de los resultados reportados en el manuscrito.

Tabla B1.1

Paquetes y versiones (entorno Python)

Paquete	Versión
Python	3.10.x
numpy	1.26.x
pandas	2.2.x
scipy	1.11.x
statsmodels	0.14.x
semopy	2.x
scikit-learn	1.4.x
pingouin (opcional)	0.5.x
matplotlib	3.8.x

Nota. Las versiones se reportan como rangos (x) para reflejar la variación menor permitida en el entorno de ejecución (Google Colab). La reproducción exacta de resultados se garantiza mediante el cuaderno de análisis y el archivo de dependencias asociado (*requirements/lockfile*), disponible en el repositorio del proyecto.

B2. Estructura del proyecto y rutas

Esta sección documenta la estructura mínima del proyecto reproducible y las rutas utilizadas en el cuaderno de análisis en Python (Google Colab). El propósito es garantizar que un tercero pueda ejecutar el *pipeline* con los mismos insumos, replicar salidas intermedias y finales y verificar la trazabilidad de cada artefacto (tablas, figuras y bitácora). Para evitar dependencia de rutas locales, el cuaderno utiliza variables de ruta (paths) y una convención de carpetas estable, separando claramente: datos crudos, datos analíticos, *scripts*, outputs publicables y metadatos de ejecución. La estructura propuesta se alinea con buenas prácticas de reproducibilidad: separación de llaves y datos sensibles (cuando existan), control de versiones, minimización de datos y generación automática de outputs en ubicaciones deterministas.

Figura B2.1

Estructura mínima del proyecto reproducible y organización de rutas del pipeline analítico

```

/Tesis_IA_RV/
├─ data/
│   ├── raw/           # datos crudos (no se suben a OSF si son sensibles)
│   └─ processed/      # datos listos (seudonimizados)
├─ notebooks/
│   └─ Tesis_IA_RV_Analisis_Reproducible_UIIX_FULL.ipynb
├─ src/
│   ├── utils.py        # funciones de apoyo
│   ├── psicometria.py  #  $\alpha$ ,  $\omega$ , CFA, invariancia
│   ├── modelos.py      # multinivel, dppc2, ANCOVA
│   └─ graficas.py      # figuras APA
├─ outputs/
│   ├── CSV/            # tablas .csv
│   ├── APA_Tables.xlsx  # libro consolidado (cap. 3-4)
│   ├── figs/           # .png/.pdf de figuras
│   └─ 99_bitacora.txt   # bitácora de ejecución
├─ LICENSE
├─ CITATION.cff
├─ README.md
└─ requirements.txt

```

Nota. La figura muestra la estructura mínima de carpetas y archivos utilizada para organizar el proyecto reproducible de la tesis. La carpeta `data/` separa datos crudos y datos procesados; `notebooks/` contiene el cuaderno principal de análisis; `src/` agrupa funciones auxiliares para psicometría, modelos y gráficas; y `outputs/` concentra tablas, figuras y bitácoras generadas por el pipeline. La organización permite distinguir insumos, código, resultados y metadatos de ejecución, favoreciendo reproducibilidad, trazabilidad y auditoría técnica. Los datos crudos no deben compartirse públicamente si contienen información sensible o identificable; para repositorios abiertos debe emplearse únicamente la versión seudonimizada o desidentificada.

La estructura se diseñó para minimizar ambigüedad entre insumos y resultados. Los datos crudos se mantienen como solo lectura y los conjuntos analíticos se generan mediante pasos documentados del cuaderno. Las tablas publicables se exportan a un archivo consolidado para facilitar integración al manuscrito, y las figuras se almacenan en una carpeta dedicada para garantizar consistencia de nombres y referencias. La bitácora de ejecución y la huella (*hash*) se registran al final del *pipeline*, permitiendo auditoría técnica de la corrida sin exponer información identificable. La Figura B2.1 sintetiza esta organización y muestra la separación funcional entre datos, código, salidas y archivos de documentación reproducible.

Tabla B2.1

Estructura del proyecto y contenido esperado por carpeta

Carpeta	Propósito	Contenido típico	Regla operativa
data/00_raw/	Resguardo de insumos originales	CSV originales, exportaciones de plataforma, instrumentos en formato no editable	Solo lectura. No se edita manualmente.
data/10_interim/	Salidas intermedias del preprocesamiento	Datos depurados preliminares, recodificaciones parciales	Se regenera desde notebook.
data/20_processed/	<i>Dataset</i> analítico final	<i>Dataset</i> PRE y POST armonizado, diccionario de variables	Fuente única para modelado.
docs/	Documentación y metadatos	<i>Codebook</i> , notas metodológicas, mapeo de variables	Sin resultados numéricos.
notebooks/	Cuadernos ejecutables	Notebook principal y auxiliares	No contiene datos embebidos.
src/	Funciones reutilizables	Funciones de limpieza, QA, modelos, gráficas	Importable desde notebook.
outputs/APA/	Tablas publicables	APA_Tables.xlsx y tablas complementarias	Generación automática.
outputs/figs/	Figuras	PNG para reporte y manuscrito	Nombres estandarizados.
outputs/logs/	Bitácora y huella	99_bitacora.txt, run_hash.txt	Se actualiza por corrida.
configs/	Configuración	Parámetros del <i>pipeline</i> , seeds, thresholds	Controla reproducibilidad.

Nota. La separación 00_raw (solo lectura) versus 20_processed (analítico final) evita cambios no trazables. Las carpetas de outputs/ se consideran artefactos regenerables desde el cuaderno.

Tabla B2.2*Variables de rutas y parámetros mínimos (Google Colab)*

Variable	Descripción	Ejemplo de valor
PROJECT_ROOT	Ruta raíz del proyecto (Drive o entorno local Colab)	/content/drive/MyDrive/OSF_AI_RV/
PATH_RAW	Carpeta de datos crudos	PROJECT_ROOT + "data/00_raw/"
PATH_INTERIM	Carpeta de datos intermedios	PROJECT_ROOT + "data/10_interim/"
PATH_PROCESSED	Carpeta de datos procesados	PROJECT_ROOT + "data/20_processed/"
PATH_OUTPUT_APA	Carpeta para tablas publicables	PROJECT_ROOT + "outputs/APA/"
PATH_OUTPUT_FIGS	Carpeta para figuras	PROJECT_ROOT + "outputs/figs/"
PATH_LOGS	Carpeta de bitácoras y huella	PROJECT_ROOT + "outputs/logs/"
SEED	Semilla global para reproducibilidad	2026
M_IMPUTATIONS	Número de imputaciones (si aplica)	20
ICC_THRESHOLD	Umbral mínimo de ICC para aceptación	0.75
ALPHA	Nivel de significancia	0.05

Nota. Las rutas se definen como variables para permitir ejecución consistente tanto en Google Drive como en un entorno local. La semilla, umbrales y parámetros quedan centralizados para reducir divergencia entre corridas.

B3. Flujo de ejecución en Python (Google Colab)

El cuaderno reproducible implementa un *pipeline* secuencial y determinista que ejecuta, en orden fijo, controles de calidad, construcción de variables, estimación de fiabilidad e ICC, validación de medición (CFA e invariancia) y modelos principales de efecto (Tiempo $\times$ Grupo), junto con análisis de dosis-respuesta, moderaciones y sensibilidad. Para asegurar reproducibilidad, la ejecución se controla mediante parámetros centralizados (semilla, umbrales y rutas) y genera automáticamente las salidas en carpetas predefinidas, sin intervención manual. Cada corrida registra una bitácora mínima y una huella de ejecución (hash), permitiendo trazabilidad técnica y auditoría de consistencia.

Tabla B3.1

Flujo de ejecución del cuaderno reproducible en Python (Google Colab): etapas del pipeline analítico

Paso	Proceso
1	Carga y validación de datos
2	Integridad y control de faltantes (MCAR/MAR)
3	Construcción de variables y escalas
4	Fiabilidad (α , ω)
5	ICC interjueces y anclajes
6	CFA e invariancia PRE/POST
7	Modelos multinivel (efecto Tiempo $\times$ Grupo) y d_{ppc2}
8	Dosis-respuesta
9	Moderaciones (experiencia tecnológica, programa)
10	Gráficas APA y tablas
11	Sensibilidad/robustez
12	Exportación a APA_Tables.xlsx
13	Figuras a outputs/figs/
14	Bitácora y huella (hash)

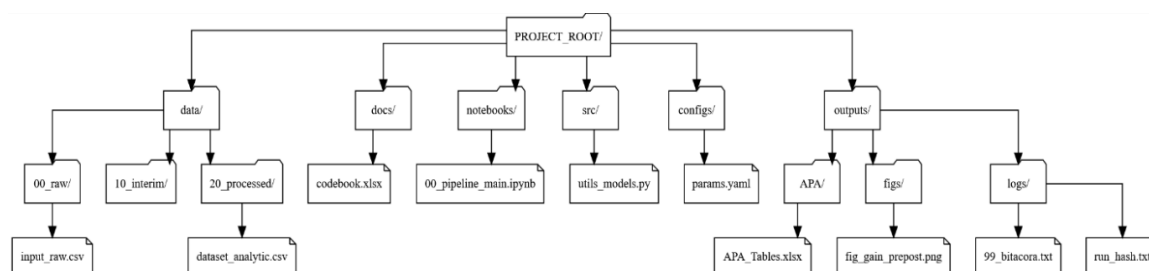
Nota. La tabla resume la ejecución del cuaderno reproducible. Cada etapa genera salidas registradas en el Apéndice B4. El proceso se aplica a PRE y POST y culmina en la estimación de β_3 , d_{ppc2} y los controles de sensibilidad y trazabilidad.

La determinación de seeds se aplica a los procesos estocásticos relevantes (por ejemplo, imputación múltiple cuando procede y procedimientos de remuestreo si se

emplean en sensibilidad). Asimismo, se restringe el origen de datos del modelado al *dataset* analítico en `data/20_processed/`, garantizando que toda tabla y figura se derive del mismo conjunto armonizado PRE y POST. Las salidas publicables se consolidan en un archivo único de tablas (APA) y un conjunto de figuras exportadas, con nomenclatura estandarizada para evitar discrepancias al integrar resultados en el manuscrito.

Figura B.3.1

Estructura de carpetas del proyecto reproducible y ubicación de salidas analíticas

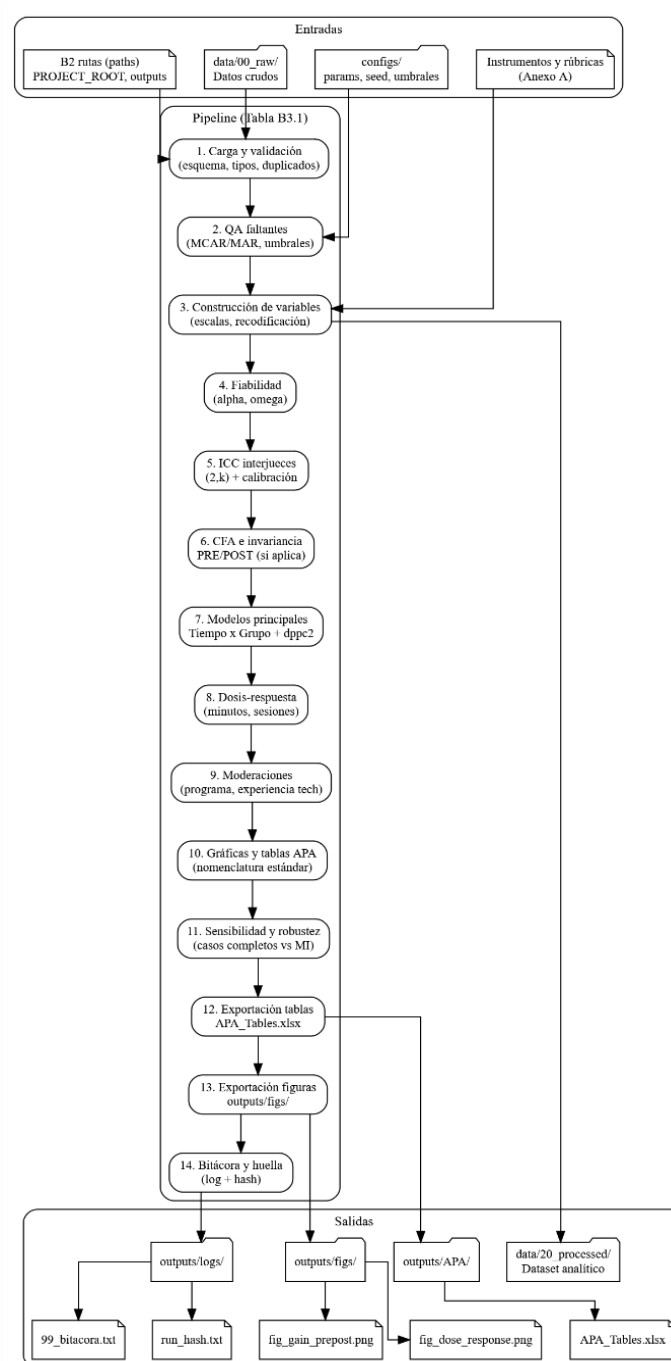


Nota. La estructura separa datos crudos, datos analíticos y artefactos regenerables. Las tablas publicables se consolidan en `outputs/APA/` y las figuras se exportan a `outputs/figs/`. La bitácora y la huella de ejecución se almacenan en `outputs/logs/` para trazabilidad.

Con esta estructura, el flujo descrito en la Tabla B3.1 se implementa como un *pipeline* que lee insumos desde `data/00_raw/`, genera *datasets* analíticos en `data/20_processed/` y exporta resultados publicables a `outputs/`, manteniendo parámetros y rutas centralizados para reproducibilidad.

Figura B.3.2

Pipeline analítico reproducible del cuaderno en Python (Colab): etapas, controles y artefactos de salida



Nota. La figura resume el orden de ejecución del cuaderno reproducible y explicita la relación entre entradas (datos crudos, instrumentos y configuración) y salidas deterministas (*dataset* analítico, tablas APA, figuras y registro de ejecución). La bitácora y la huella permiten verificar consistencia de corridas sin incorporar información identificable.

B3.1 Instrucciones mínimas de reproducción del análisis

La reproducción del análisis se diseñó para ejecutarse en un entorno estándar de Google Colab, utilizando el cuaderno principal del proyecto y la estructura de rutas descrita en B2. En primer término, debe verificarse que el directorio raíz del proyecto contenga las carpetas mínimas requeridas (*data/20_processed/*, *notebooks/*, *configs/*, *outputs/*) y que el *dataset* analítico seudonimizado se encuentre disponible en la ruta definida para modelado. En segundo término, debe abrirse el cuaderno principal *00_master_pipeline.ipynb*, cargar las dependencias especificadas para el entorno Python y confirmar que las variables de ruta y parámetros globales (por ejemplo, semilla, umbrales operativos y carpetas de salida) coincidan con la configuración declarada en el proyecto.

Una vez validado el entorno, la ejecución debe realizarse de forma secuencial, sin omitir celdas, siguiendo el flujo documentado en la Tabla B3.1. El *pipeline* ejecuta controles de integridad, construcción de variables, fiabilidad, ICC, validaciones de medición, modelos principales y análisis complementarios, y posteriormente exporta las salidas a las carpetas de *outputs/*. La corrida se considera satisfactoria cuando genera, al menos, los siguientes artefactos: (i) el archivo consolidado de tablas publicables en formato APA, (ii) las figuras finales en formato PNG, y (iii) la evidencia mínima de trazabilidad, compuesta por la bitácora de ejecución y la huella de corrida (hash).

La verificación de reproducción se completa comprobando que los *outputs* existan en las rutas previstas y que la bitácora no registre fallas críticas. Así, la replicación depende de la correspondencia verificable entre insumos, parámetros, ejecución y salidas reportadas en la tesis.

En concordancia con el Capítulo 4, el paquete reproducible distingue entre la cohorte de implementación ($N = 180$), empleada para cobertura operativa e incidencias, y la base analítica per-protocol ($n = 160$), utilizada para la inferencia principal. En consecuencia, el cuaderno ejecuta los modelos principales sobre el *dataset* con evidencia PRE y POST comparable, mientras que la cohorte ampliada se conserva para descriptivos y análisis complementarios de implementación, dosis y fidelidad.

B4. Depósito en Open Science Framework (OSF)

El repositorio OSF concentra el paquete reproducible completo del estudio, organizado conforme a la estructura descrita en B2. En particular, el depósito incluye: (i) la base de datos anonimizada o seudonimizada en el nivel requerido para replicación académica, (ii) el cuaderno ejecutable en Python (Google Colab) y, cuando aplica, el código modular en src/, (iii) la documentación mínima de variables y decisiones analíticas (*codebook* y bitácora), y (iv) los artefactos generados por la ejecución (tablas y figuras). Esta organización permite que un revisor o lector técnico pueda reconstruir el flujo de análisis desde insumos hasta outputs publicables, verificando tanto los resultados principales.

DOI OSF del paquete reproducible: DOI: 10.17605/OSF.IO/56GPE

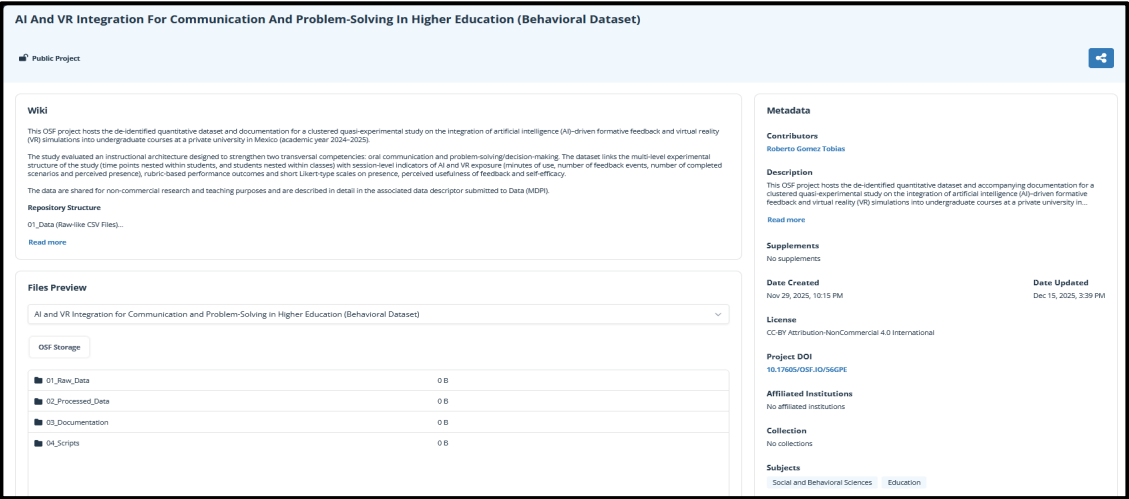
Título: *AI and VR Integration for Communication and Problem-Solving in Higher Education (Behavioral Dataset)*

Enlace de Acceso: <https://osf.io/56gpe/>

Licencia: Creative Commons Atribución–No Comercial 4.0 (CC BY-NC 4.0).

Figura B.4.1

Depósito del paquete reproducible de la tesis en Open Science Framework (OSF)



Nota. La figura muestra la página del proyecto depositado en Open Science Framework (OSF), correspondiente al paquete reproducible asociado a la tesis.

Tabla B4.1

Inventario mínimo del paquete reproducible en OSF: artefactos, propósito y condición de privacidad

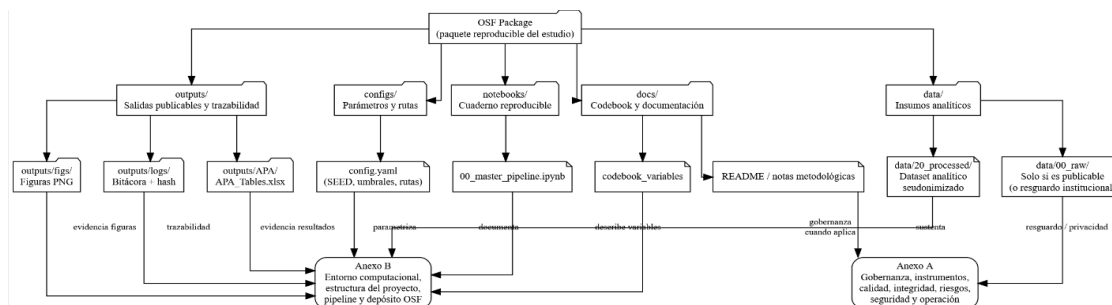
Componente	Ruta/Nombre (en repositorio)	Propósito	Generado por	Condición de privacidad / notas
Dataset analítico	data/20_processed/ (archivo(s) analítico(s))	Base única para todas las estimaciones (PRE/POST, grupo, covariables, puntajes)	<i>Pipeline</i> (B3)	Anonimizado o seudonimizado según protocolo; sin PII
Datos crudos (si aplica)	data/00_raw/ (solo si es publicable)	Trazabilidad del origen y auditoría interna	Captura/Export	Si no es publicable, sustituir por no depositado; resguardo institucional
Codebook	docs/codebook_variables.pdf o docs/codebook.csv	Definición de variables, escalas, codificación y rangos	Autor/Analista	Sin PII; describe estructura y transformaciones
Cuaderno principal	notebooks/00_master_pipeline.ipynb	Reproducción integral del flujo B3	Autor/Analista	Ejecutable en Colab; fija rutas y parámetros
Scripts auxiliares	src/ (funciones)	Modularidad (limpieza, QA, modelos, figuras)	Autor/Analista	Código sin datos
Configuración	configs/config.yaml (o equivalente)	Semilla, umbrales, rutas, switches de sensibilidad	<i>Pipeline</i>	Permite corridas consistentes entre entornos
Tablas publicables	outputs/APA/APA_Tables.xlsx	Tablas finales en formato integrable al manuscrito	<i>Pipeline</i>	Derivadas del <i>dataset</i> analítico; sin identificadores
Figuras	outputs/figs/ (PNG)	Figuras finales (ganancia PRE-POST, dosis-respuesta, etc.)	<i>Pipeline</i>	Derivadas; sin identificadores
Bitácora de ejecución	outputs/logs/99_bitacora.txt	Evidencia mínima de corrida (fechas, versiones, checks)	<i>Pipeline</i>	No contiene PII; documenta decisiones y eventos
Huella (hash)	outputs/logs/run_hash.txt	Verificación de consistencia de corrida y artefactos	<i>Pipeline</i>	Recomendado para auditoría técnica
Licencia y citación	LICENSE y CITATION.cff (si se incluye)	Claridad de uso, atribución y citación	Autor	Consistente con CC BY-NC 4.0
Readme del paquete	README.md	Guía de reproducción paso a paso	Autor	Enfatiza rutas, prerequisites y outputs esperados

Nota. La tabla explicita el inventario mínimo del paquete reproducible y su función dentro de la lógica de replicación: un *dataset* analítico único, un *pipeline* ejecutable y salidas deterministas (tablas y figuras).

Con el fin de presentar de manera legible la organización del paquete reproducible y su vínculo con los apéndices metodológicos de la tesis, la Figura B.3 sintetiza en formato vertical la estructura depositada en OSF, distinguiendo los insumos analíticos, el cuaderno ejecutable, la configuración, la documentación y las salidas publicables, así como su correspondencia funcional con los Apéndices A y B.

Figura B.4.2

Estructura del paquete reproducible en OSF y correspondencia con apéndices de la tesis



Nota. Esta estructura permite auditoría académica y replicación técnica sin exponer información identificable.

APÉNDICE C

VALIDACIÓN COMPLEMENTARIA DE LA PROPUESTA MEDIANTE DELPHI MODIFICADO DE DOS RONDAS

La validación complementaria de la propuesta se realizó mediante un Delphi modificado de dos rondas, con el propósito de valorar la claridad conceptual, la pertinencia pedagógica, la factibilidad técnica y operativa, la viabilidad institucional, la calidad de las salvaguardas de gobernanza y la originalidad de la arquitectura IA+RV propuesta. Este proceso no tuvo como finalidad sustituir la validación empírica inicial derivada del estudio, sino complementarla mediante juicio experto orientado al refinamiento del diseño y a su potencial de adopción institucional.

Dado que el Capítulo 5 presenta la síntesis analítica e interpretativa del Delphi, este apéndice conserva la documentación metodológica y operativa del proceso. En consecuencia, su propósito no es repetir la argumentación desarrollada en el cuerpo de la tesis, sino documentar los criterios de selección del panel, el procedimiento seguido, los instrumentos aplicados, los resultados agregados por ronda, la matriz de ajustes introducidos entre rondas y la evidencia documental mínima necesaria para asegurar trazabilidad y auditabilidad.

Los resultados del Delphi se reportan de forma agregada y anonimizada. Las medias, rangos, variaciones y comentarios consolidados corresponden a las valoraciones del panel experto, sin asociar respuestas individuales con nombres personales. Este tratamiento permite preservar la confidencialidad de los participantes y, al mismo tiempo, mantener la transparencia metodológica del proceso de validación.

C1. Criterios de selección y caracterización del panel de expertos

El panel se integró mediante muestreo intencional por criterio, atendiendo a la experiencia académica y profesional de los participantes en áreas directamente vinculadas con la propuesta. Se buscó incorporar perfiles con conocimiento en innovación educativa, diseño instruccional, evaluación por competencias y rúbricas, inteligencia artificial en educación, realidad virtual o simulación educativa, así como gobernanza y ética del uso de tecnología en educación superior.

La selección del panel respondió a la necesidad de valorar la propuesta desde distintas perspectivas complementarias. En particular, se consideró relevante contar con expertos capaces de revisar la coherencia pedagógica del modelo, la congruencia entre competencias e instrumentos de evaluación, la factibilidad técnica del componente inmersivo, el papel formativo de la IA, la viabilidad institucional de adopción gradual, las salvaguardas de privacidad y seguridad, y las condiciones de acompañamiento docente requeridas para su implementación.

Tabla C1.1

Criterios de inclusión del panel de expertos

Criterio	Descripción
Formación académica	Doctorado o maestría con trayectoria consolidada en el área de especialidad.
Experiencia profesional o académica	Mínimo ocho años en docencia, investigación, innovación educativa, gestión académica, tecnología educativa o áreas afines.
Pertinencia temática	Experiencia demostrable en al menos una de las áreas centrales de la propuesta: diseño instruccional, evaluación por competencias, IA educativa, RV, innovación educativa, gobernanza digital o formación docente.
Producción o trayectoria	Publicaciones, proyectos, coordinación académica, consultoría especializada, desarrollo institucional o experiencia relevante en implementación educativa.
Disponibilidad	Disposición para participar en dos rondas sucesivas dentro del calendario previsto.
Confidencialidad	Aceptación de que los resultados fueran reportados de manera agregada y anonimizada mediante códigos alfanuméricos.

Tabla C1.2*Caracterización del panel participante*

Código del experto	Grado académico	Área principal de expertise	Años de experiencia	Ámbito institucional principal	Aporte esperado a la validación
E1	Doctorado	Diseño instruccional	18	Universidad	Valorar la coherencia del diseño pedagógico y la secuencia formativa de la propuesta
E2	Doctorado	Evaluación por competencias	15	Universidad	Valorar congruencia entre resultados, criterios de desempeño y mecanismos de evaluación
E3	Doctorado	Inteligencia artificial en educación	12	Investigación universitaria	Valorar la integración de la IA como apoyo formativo y sus implicaciones pedagógicas
E4	Maestría	Realidad virtual y simulación educativa	30	Tecnología educativa	Valorar factibilidad técnica y pertinencia del componente inmersivo en contextos universitarios
E5	Doctorado	Innovación educativa	20	Gestión académica	Valorar viabilidad institucional, escalabilidad y adopción progresiva de la propuesta
E6	Doctorado	Gobernanza y ética digital	14	Universidad y consultoría	Valorar salvaguardas de privacidad, seguridad, equidad y gobernanza tecnológica
E7	Maestría	Formación y desarrollo docente	16	Universidad	Valorar factibilidad operativa y condiciones de acompañamiento para implementación docente

Nota. Los integrantes del panel se reportan mediante códigos alfanuméricos, de E1 a E7, para preservar una identificación funcional y homogénea dentro del proceso Delphi. La composición del panel buscó asegurar diversidad de expertise en dimensiones directamente vinculadas con la propuesta: diseño instruccional, evaluación, inteligencia artificial en educación, realidad virtual, innovación educativa, gobernanza digital y formación docente.

C2. Procedimiento general del Delphi modificado

El Delphi se aplicó en dos rondas. En la primera, se envió a cada experto un resumen ejecutivo de la propuesta, acompañado de una matriz estructurada de valoración con escala de cuatro puntos y espacios para comentarios abiertos. Las respuestas se procesaron mediante estadística descriptiva simple y síntesis temática de observaciones cualitativas.

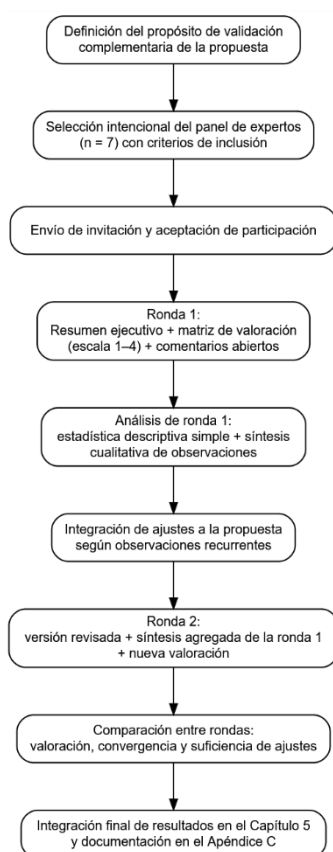
Posteriormente, se ajustó la propuesta con base en las sugerencias recurrentes del panel. La versión revisada se envió nuevamente en una segunda ronda, acompañada de una síntesis agregada de los resultados de la primera. El propósito de esta segunda aplicación fue estimar convergencia, valorar la suficiencia de los ajustes y fortalecer la claridad y consistencia interna del modelo.

El proceso se estructuró de la siguiente manera:

1. Definición de criterios de selección del panel experto.
2. Integración del panel mediante muestreo intencional por criterio.
3. Envío del resumen ejecutivo de la propuesta.
4. Aplicación del instrumento de valoración experta de la ronda 1.
5. Sistematización de resultados cuantitativos y cualitativos.
6. Identificación de observaciones recurrentes.
7. Ajuste de la propuesta doctoral.
8. Envío de la versión revisada y síntesis agregada de la ronda 1.
9. Aplicación del instrumento de valoración experta de la ronda 2.
10. Comparación entre rondas y cierre interpretativo del proceso.

Figura C2.1

Procedimiento de validación complementaria mediante Delphi modificado de dos rondas



Nota. El diagrama resume la secuencia del Delphi modificado de dos rondas utilizada para la validación complementaria de la propuesta doctoral. La primera ronda permitió obtener valoración experta inicial e identificar áreas de mejora; la segunda permitió valorar la suficiencia de los ajustes y estimar convergencia en torno a la versión revisada de la propuesta.

C3. Formato de invitación a expertos

Figura C3.1

Modelo de invitación enviada al panel de expertos

Estimado/a Dr./Dra. [Nombre]:

Por este medio solicito atentamente su participación como experto/a en un proceso de validación complementaria mediante Delphi modificado de dos rondas, vinculado con mi tesis doctoral. El propósito de este proceso es valorar la claridad, pertinencia, factibilidad, gobernanza, escalabilidad y aporte doctoral de una propuesta de arquitectura didáctica IA+RV para el desarrollo de competencias transversales en educación superior.

La propuesta deriva de un estudio empírico sobre la integración de inteligencia artificial y realidad virtual en cursos universitarios, orientado a fortalecer el desempeño observable en Comunicación y Solución de problemas. Su participación permitirá revisar la propuesta desde una perspectiva experta y aportar observaciones que contribuyan a su refinamiento académico, metodológico e institucional.

La participación consistirá en revisar un resumen ejecutivo de la propuesta y responder una matriz de valoración estructurada con escala de cuatro puntos, acompañada de espacios para observaciones cualitativas. Posteriormente, se compartirá una versión ajustada de la propuesta, junto con una síntesis agregada de la primera ronda, para emitir una segunda valoración sobre la suficiencia de los ajustes realizados.

La información será utilizada exclusivamente con fines académicos. Los resultados se reportarán de forma agregada y anonimizada; los participantes serán identificados mediante códigos alfanuméricos, sin asociar nombres personales a respuestas individuales.

Agradezco de antemano su tiempo, disposición y contribución experta al fortalecimiento de esta propuesta doctoral.

Atentamente,
Roberto Gómez Tobías
Doctorando del Programa de Doctorado en Educación e Innovación
Universidad de Investigación e Innovación de México

C4. Resumen ejecutivo enviado a expertos

El resumen ejecutivo remitido al panel de expertos incluyó los siguientes apartados:

1. Problema que atiende la propuesta
2. Hallazgo central del estudio empírico
3. Propósito general de la arquitectura IA+RV
4. Tesis de diseño y mecanismo instruccional
5. Componentes no negociables de la propuesta
6. Sistema mínimo viable
7. Estrategia de implementación por fases
8. Criterios de gobernanza, seguridad, equidad y escalamiento responsable

El apartado no sustituyó el capítulo completo de propuesta. Su finalidad fue ofrecer al panel una versión sintética, homogénea y suficiente del problema, el hallazgo empírico, la tesis de diseño, los componentes centrales, el sistema mínimo viable, la estrategia por fases y los criterios de gobernanza. Esto permitió que todos los expertos valoraran la misma versión base de la propuesta durante la primera ronda.

C5. Instrumento de valoración experta. Ronda 1

C5.1 Instrucciones generales

Se solicitó a los expertos valorar cada criterio en una escala de cuatro puntos, donde:

- 1 = Bajo
- 2 = Aceptable con debilidades importantes
- 3 = Adecuado
- 4 = Alto

Después de cada criterio se pidió incluir observaciones o sugerencias concretas de mejora. El instrumento combinó valoración cuantitativa y retroalimentación cualitativa, con el fin de identificar tanto el nivel de aceptación inicial de la propuesta como los aspectos que requerían mayor precisión antes de una eventual adopción institucional gradual.

Tabla C5.1

Matriz de valoración experta de la propuesta. Ronda 1

Dimensión	Criterio de valoración	Escala (1–4)	Observaciones / sugerencias
Claridad conceptual	La propuesta presenta con claridad su propósito, alcance y lógica general		
Coherencia empírica	La propuesta deriva de manera consistente de los hallazgos del estudio		
Pertinencia pedagógica	El mecanismo instruccional IA+RV está pedagógicamente bien justificado		
Coherencia interna	Los componentes de la propuesta guardan consistencia entre sí		
Factibilidad técnica	La propuesta puede implementarse con condiciones tecnológicas razonables		
Factibilidad operativa	La propuesta resulta operativamente viable en contexto universitario		
Escalabilidad institucional	La propuesta puede adoptarse gradualmente sin perder coherencia		
Gobernanza y seguridad	La propuesta incorpora salvaguardas suficientes de ética, privacidad y seguridad		
Equidad y accesibilidad	La propuesta contempla condiciones razonables de inclusión y alternativas equivalentes		
Originalidad y aporte doctoral	La propuesta representa una contribución doctoral pertinente y novedosa		

C5.2 Preguntas abiertas de la ronda 1

1. ¿Qué componente considera más sólido de la propuesta?
2. ¿Qué componente considera más débil o menos desarrollado?
3. ¿Qué ajuste recomendaría antes de considerar la propuesta como lista para adopción institucional gradual?
4. ¿Existen riesgos no suficientemente explicitados?
5. Comentarios generales.

C6. Síntesis de resultados de la primera ronda

La primera ronda permitió identificar una valoración favorable de la propuesta, especialmente en coherencia empírica, pertinencia pedagógica y originalidad del aporte doctoral. Las dimensiones que recibieron observaciones de mejora se relacionaron principalmente con claridad conceptual, factibilidad técnica y operativa, escalabilidad institucional, y gobernanza y seguridad. Las observaciones no cuestionaron el núcleo pedagógico de la propuesta, sino que solicitaron mayor explicitación de condiciones de adopción, criterios de salida, recursos mínimos y responsabilidades institucionales.

Tabla C6.1

Síntesis agregada de resultados de la ronda 1

Dimensión	Media	Mediana	Rango	Comentario dominante	Acción de ajuste
Claridad conceptual	3.3	3	3–4	Reforzar definición del alcance y secuencia.	Se clarificaron la denominación, el mecanismo instruccional y los objetivos de la propuesta.
Coherencia empírica	3.8	4	3–4	Muy bien anclada a los hallazgos.	No se realizó ajuste mayor; se mantuvo el anclaje en resultados empíricos.
Pertinencia pedagógica	3.9	4	3–4	Mecanismo sólido.	No se realizó ajuste mayor; se conservó la lógica de práctica situada, rúbricas y <i>feedback</i> criterial.
Factibilidad técnica y operativa	3.1	3	2–4	Precisar recursos mínimos y soporte.	Se reforzaron el sistema mínimo viable, los recursos necesarios y la factibilidad institucional.
Escalabilidad institucional	3.0	3	2–4	Hacer más visibles fases y criterios.	Se fortaleció la estrategia por fases, los criterios de salida y el escalamiento responsable.
Gobernanza y seguridad	3.2	3	2–4	Delimitar mejor responsabilidades.	Se ampliaron las salvaguardas de privacidad, seguridad, supervisión humana, equidad y trazabilidad.
Originalidad y aporte doctoral	3.7	4	3–4	Aporte doctoral claro.	Se hizo más explícita la originalidad del mecanismo integrado IA+RV.

Nota. Los resultados se presentan de forma agregada para preservar la confidencialidad del panel. La primera ronda mostró alta valoración de la pertinencia pedagógica y coherencia empírica, junto con áreas de mejora relacionadas con factibilidad, gobernanza y escalabilidad.

C7. Matriz de ajustes realizados entre rondas

A partir de los resultados de la primera ronda, se realizaron ajustes orientados a fortalecer la claridad conceptual, la factibilidad institucional, los criterios de adopción, el papel de la IA, la seguridad, la gobernanza y la explicitación del aporte doctoral. Los ajustes no modificaron el núcleo de la propuesta, pero sí incrementaron su precisión, auditabilidad y potencial de adopción responsable.

Tabla C7.1

Observaciones recurrentes de la ronda 1 y ajustes incorporados en la propuesta

Observación recurrente de expertos	Ajuste incorporado	Sección modificada
Mayor claridad sobre las condiciones de escalamiento.	Se reforzó la estrategia por fases y los criterios de salida para avanzar entre etapas.	5.3.3 y 5.4.3
Necesidad de hacer visible la factibilidad institucional.	Se fortaleció el apartado de factibilidad técnica, operativa e institucional, así como la definición de recursos mínimos.	5.3.4 y 5.4.2
Criterios de adopción poco visibles.	Se consolidó el apartado de adopción y escalamiento responsable con evidencias mínimas y decisiones asociadas.	5.4.3
Necesidad de delimitar mejor el papel de la IA.	Se reforzó la noción de IA como tutor formativo criterial con supervisión humana y no como evaluador final autónomo.	5.1.4, 5.2.1 y 5.2.3
Mayor claridad sobre gobernanza y seguridad.	Se ampliaron los criterios de privacidad, equidad, trazabilidad, alternativas equivalentes y seguridad en RV.	5.2.3 y 5.4.3
Claridad del aporte doctoral.	Se hizo más explícita la originalidad del mecanismo integrado IA+RV y su contribución como arquitectura auditable.	5.1.1 y 5.5

Nota. La matriz muestra la trazabilidad entre observaciones expertas, ajustes incorporados y secciones modificadas. Su finalidad es documentar que el Delphi funcionó como proceso de refinamiento de diseño y no únicamente como valoración externa de la propuesta.

C8. Instrumento de valoración experta. Ronda 2

C8.1 Instrucciones generales

En la segunda ronda se envió a los expertos una versión ajustada de la propuesta, acompañada de la síntesis agregada de la primera ronda. Se solicitó revalorar la propuesta con base en los cambios realizados y emitir un juicio sobre la suficiencia del ajuste.

La segunda ronda tuvo tres propósitos: valorar si los ajustes atendieron las observaciones principales de la primera ronda, estimar convergencia en torno a la versión revisada de la propuesta y determinar si el modelo podía considerarse suficientemente claro, pertinente y viable para una adopción gradual bajo evaluación continua.

Tabla C8.1

Matriz de valoración experta. Ronda 2

Dimensión	Valoración ronda 1 (promedio)	Nueva valoración (1–4)	¿Considera suficiente el ajuste? (Sí/No)	Comentario final
Dimensión	Media ronda 1	Media agregada ronda 2	¿Considera suficiente el ajuste?	Comentario final
Claridad conceptual	3.3	3.7	Sí	La propuesta quedó más clara en propósito, alcance y secuencia.
Coherencia empírica	3.8	3.9	Sí	La relación entre hallazgos y propuesta se mantiene sólida.
Pertinencia pedagógica	3.9	4.0	Sí	El mecanismo instruccional está bien fundamentado.
Factibilidad técnica y operativa	3.1	3.6	Sí	La viabilidad es suficiente bajo condiciones mínimas explícitas.
Escalabilidad institucional	3.0	3.5	Sí	La estrategia por fases mejora la claridad de adopción gradual.
Gobernanza y seguridad	3.2	3.7	Sí	Las salvaguardas y responsabilidades quedaron mejor delimitadas.

C8.2 Preguntas de cierre de la ronda 2

1. ¿Considera que la propuesta, en su versión revisada, es suficientemente clara y pertinente para una adopción gradual bajo evaluación continua?
2. ¿Recomendaría su aprobación académica como propuesta doctoral?
3. Observaciones finales.

C9. Resultados finales del Delphi modificado

La comparación entre la ronda 1 y la ronda 2 muestra una mejora general en la valoración de la propuesta. Las dimensiones con mayor incremento fueron factibilidad técnica y operativa, escalabilidad institucional, y gobernanza y seguridad. Esto indica que los ajustes incorporados entre rondas respondieron a las principales observaciones del panel.

Tabla C9.1

Comparación entre la ronda 1 y la ronda 2

Dimensión	Media ronda 1	Media ronda 2	Variación	Lectura
Claridad conceptual	3.3	3.7	+0.4	Mejora evidente
Coherencia empírica	3.8	3.9	+0.1	Alta estabilidad
Pertinencia pedagógica	3.9	4.0	+0.1	Consenso alto
Factibilidad técnica y operativa	3.1	3.6	+0.5	Mejora clara
Escalabilidad institucional	3.0	3.5	+0.5	Mejora clara
Gobernanza y seguridad	3.2	3.7	+0.5	Mejora clara
Originalidad y aporte doctoral	3.7	3.9	+0.2	Muy alta valoración

Nota. La segunda ronda mostró una mejora general en la valoración de la propuesta respecto de la primera, especialmente en claridad conceptual, factibilidad técnica y operativa, escalabilidad institucional, gobernanza y explicitación de criterios de adopción. Aunque algunas dimensiones mantuvieron una valoración moderada, los expertos consideraron que los ajustes incorporados fueron suficientes para sostener la pertinencia y viabilidad de la propuesta como modelo de adopción gradual bajo evaluación continua.

En conjunto, el Delphi modificado de dos rondas aportó evidencia complementaria de que la propuesta es clara, pertinente y viable bajo criterios académicos e institucionales razonables. La mejora observada entre rondas sugiere que

el proceso no solo sirvió para obtener valoración experta, sino también para refinar la propuesta y aumentar su consistencia interna.

Esta validación experta no sustituye la validación empírica inicial de la arquitectura IA+RV, sino que la complementa desde la perspectiva de su calidad de diseño y su potencial de adopción responsable. En términos metodológicos, el Delphi permitió fortalecer la trazabilidad entre hallazgos, propuesta, criterios de implementación, gobernanza y escalamiento. En términos doctorales, contribuyó a consolidar la arquitectura IA+RV como una propuesta auditable, pertinente y susceptible de transferencia bajo condiciones explícitas de fidelidad, seguridad, equidad y evaluación continua.

ANEXO 1

FORMATO INSTITUCIONAL DE CONSENTIMIENTO INFORMADO Y AVISO DE PRIVACIDAD

Consentimiento Informado

Introducción

Queremos hacerle una invitación a participar voluntariamente en un proyecto de investigación educativa. Este proyecto tiene como objetivo analizar si la integración de herramientas de inteligencia artificial y realidad virtual, utilizadas como apoyo didáctico en actividades regulares del curso, contribuye al desarrollo de las competencias de comunicación y solución de problemas en estudiantes de pregrado.

Antes de que usted acepte participar en este estudio, le pedimos su consentimiento informado. Es de suma importancia que lea cuidadosamente el propósito y objetivos del estudio. Si usted tiene preguntas, puede hacerlas directamente al investigador responsable, quien le ayudará a resolver cualquier inquietud. Su decisión de participar es voluntaria, lo que significa que usted es totalmente libre de ingresar o no en el estudio. Podrá retirar su consentimiento en cualquier momento y sin tener que explicar las razones.

Propósito del Estudio

El propósito de este estudio es evaluar una intervención educativa que integra inteligencia artificial y realidad virtual para fortalecer las competencias de comunicación y solución de problemas en estudiantes de educación superior. El estudio busca identificar si estas actividades generan cambios en el desempeño académico frente a la metodología habitual del curso, con el fin de mejorar el diseño de estrategias didácticas basadas en tecnologías emergentes.

Objetivo

Analizar la influencia de una arquitectura didáctica basada en inteligencia artificial y realidad virtual en el desarrollo de las competencias de comunicación y solución de problemas en estudiantes de pregrado del Tecnológico de Monterrey durante el ciclo académico 2024–2025.

¿Cuánto durará la participación?

Su participación se realizará durante el periodo académico del curso e incluirá actividades regulares de aprendizaje, una medición inicial y una medición final. En algunos casos, se podrá solicitar responder cuestionarios breves o participar en espacios de retroalimentación sobre la experiencia de aprendizaje.

La participación no implicará una carga sustancial adicional a las actividades académicas normales del curso.

¿Qué sucederá durante el estudio de investigación?

Durante el estudio, los estudiantes realizarán actividades académicas relacionadas con comunicación, análisis de problemas y toma de decisiones. Algunas actividades podrán incluir inteligencia artificial para retroalimentación formativa y realidad virtual para práctica en escenarios simulados. Se recopilarán productos académicos, resultados de rúbricas, registros de participación, cuestionarios breves y observaciones generales del proceso. En caso de malestar durante el uso de realidad virtual, la actividad podrá suspenderse. La inteligencia artificial no sustituirá la evaluación del profesor ni tomará decisiones académicas. Toda la información será tratada de forma confidencial y los resultados se reportarán sin identificar a los participantes.

Este consentimiento ha sido revisado por el Comité de Ética en Investigación del Instituto Tecnológico y de Estudios Superiores de Monterrey.

Si tiene alguna preocupación o queja acerca de este estudio o sobre cómo se está realizando, o alguna pregunta con respecto a sus derechos, usted puede comunicarse al (01) 81 88882107 o al correo electrónico roberto.gt@tec.mx

Acepto participar

Sí: _____

No: _____

Nombre del participante: _____

Firma del participante: _____

Fecha: _____

Disponible en:

<https://comiteinstitucionaletica.tec.mx/sites/g/files/vgjovo1911/files/Consentimiento-Informado-simple.pdf>

ANEXO 2

LICENCIAS, ATRIBUCIONES Y MATERIALES DE TERCEROS

La implementación de una intervención con componentes tecnológicos suele requerir el uso de materiales de terceros (por ejemplo, capturas de pantalla de plataformas o imágenes referenciales de interfaz). Para preservar integridad académica, claridad interpretativa y cumplimiento de derechos de autor, se estableció un registro mínimo de licencias y atribuciones para cualquier elemento no producido íntegramente por el autor. El criterio rector es distinguir evidencia empírica del estudio (derivada de datos propios y análisis estadísticos) de materiales ilustrativos (empleados únicamente para contextualización operativa). En este Anexo se documentan los elementos de terceros, su fuente, el tipo de licencia o permiso aplicable y el texto de atribución estándar, evitando cualquier publicación de datos identificables o inferencias no sustentadas.

Tabla AN2.1

Registro de materiales de terceros utilizados

Elemento (figura/tabla)	Material	Fuente/propietario	Tipo de licencia/permiso	Atribución	Restricciones
Figura 3.3B / 3.3C	Captura de pantalla referencial de VirtualSpeech (<i>dashboard/skills</i>)	© VirtualSpeech Ltd.	Uso educativo bajo licencia institucional (ver contrato/licencia interna)	“Captura de pantalla de VirtualSpeech (© VirtualSpeech Ltd.), usada con fines educativos y únicamente ilustrativos; no contiene datos del estudio.”	No publicar datos identificables; no inferir resultados empíricos

Nota. Cada figura o material no producido íntegramente por el autor.

La Tabla anterior funciona como inventario de cumplimiento: identifica cada figura o tabla basada en material de terceros, especifica el propietario, el marco de uso educativo y las restricciones pertinentes. Esta documentación evita ambigüedad sobre autoría y licencia, y al mismo tiempo protege la validez del reporte al declarar

explícitamente que dichos elementos son ilustrativos y no forman parte de los resultados del estudio.

Tabla AN2.2

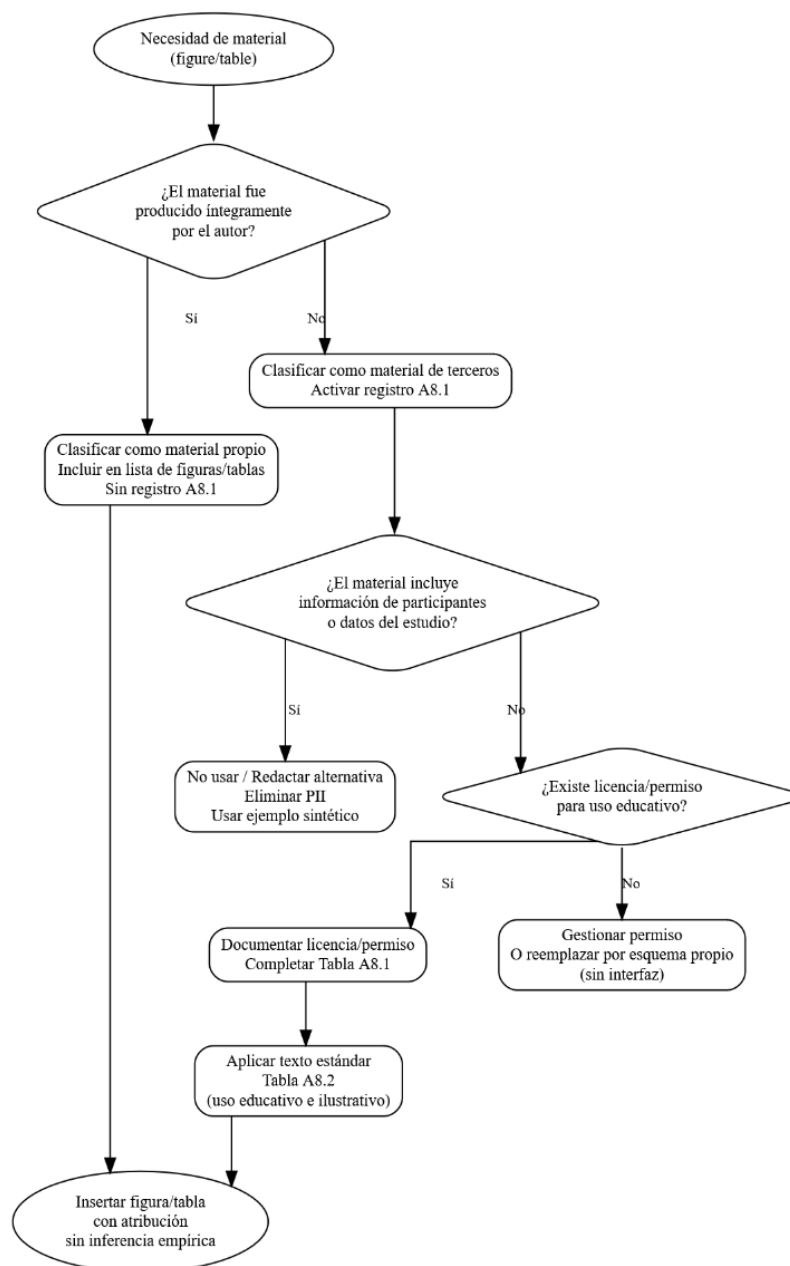
Texto estándar de atribución para figuras de plataforma

“Captura de pantalla de [NOMBRE DE PLATAFORMA] (© [PROPIETARIO]), usada con fines educativos y únicamente ilustrativos. La imagen no corresponde a resultados de participantes del estudio ni forma parte de los análisis estadísticos.”

Nota. Texto estándar de atribución y delimitación de uso para figuras basadas en capturas de pantalla de plataformas. Su objetivo es: (i) reconocer titularidad y derechos del propietario (©), (ii) establecer que el uso es educativo e ilustrativo (no como evidencia empírica), y (iii) dejar explícito que la imagen no representa datos de participantes ni se integra a los análisis estadísticos.

Figura Anexo 2.1

Diagrama mínimo de cumplimiento para uso de materiales de terceros (licencias y atribución)



Nota. El mapa formaliza un control mínimo para distinguir material propio de material de terceros y asegurar que cualquier elemento externo sea ilustrativo, con licencia o permiso aplicable, atribución consistente y sin inclusión de datos del estudio ni información identificable.