

INSTRUMENTOS

DE

INVESTIGACIÓN CIENTÍFICA:

DISEÑO, VALIDACIÓN E IMPLEMENTACIÓN



CIDPROS

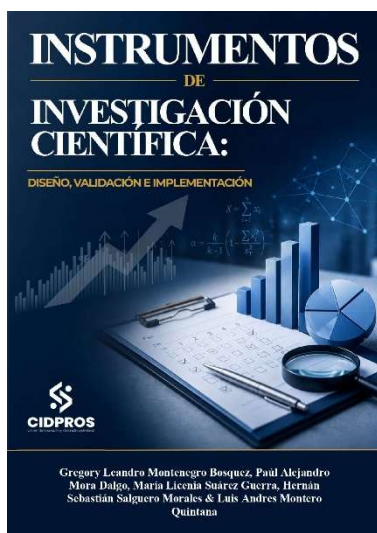
Centro de innovación y desarrollo profesional

**Gregory Leandro Montenegro Bosquez, Paúl Alejandro
Mora Dalgo, María Licenia Suárez Guerra, Hernán
Sebastián Salguero Morales & Luis Andres Montero
Quintana**

Instrumentos de investigación científica: diseño, validación e implementación

Gregory Leandro Montenegro Bosquez, Paúl Alejandro
Mora Dalgo, María Licenia Suárez Guerra, Hernán
Sebastián Salguero Morales & Luis Andres Montero
Quintana





Datos bibliográficos

ISBN	978-9907-9574-0-2
Título del libro	Instrumentos de investigación científica: diseño, validación e implementación
Autores	Gregory Leandro Montenegro Bosquez Paúl Alejandro Mora Dalgo María Licenia Suárez Guerra Hernán Sebastián Salguero Morales Luis Andres Montero Quintana
Editorial	CIDPROS EDITORIAL
Materia	001.4 - Investigación
Público objetivo	Profesional / académico
Año	2026
Número de edición	1
Tamaño	9.75Mb
Soporte	Libro digital descargable
Formato	PDF (.pdf)
Idioma	Español
DOI	https://doi.org/10.67166/508egx64

Hecho en Ecuador / Made in Ecuador

MSc. Gregory Leandro Montenegro Bosquez

Universidad Estatal de Bolívar

gregory.montenegro@ueb.edu.ec

<https://orcid.org/0000-0001-9136-3664>

San Miguel, Bolívar, Ecuador

Semblanza

Gregory Leandro Montenegro Bosquez es un profesional ecuatoriano nacido el 5 de septiembre de 1992 en San José de Chimbo, provincia de Bolívar. Es Ingeniero Agroindustrial por la Universidad Estatal de Bolívar y ha consolidado una sólida formación de cuarto nivel con tres maestrías: Magíster en Gerencia de la Calidad e Innovación por la Universidad Internacional del Ecuador, Magíster en Estadística Aplicada por la Universidad Politécnica Estatal del Carchi y Maestría en Agroindustrias por la Universidad Nacional de Chimborazo, preparación que refleja su vocación por la excelencia académica, la calidad y la mejora continua.



A lo largo de su trayectoria ha combinado la docencia universitaria con responsabilidades de gestión académica, desempeñándose como Coordinador de Programas de Maestría y Técnico de Apoyo a los Procesos de Posgrado en la Universidad Nacional de Chimborazo. Actualmente es docente de la Universidad Estatal de Bolívar, donde imparte asignaturas vinculadas a la estadística, la metodología de la investigación, la gestión por procesos, la auditoría y la administración. En el ámbito de la consultoría, ha prestado servicios como consultor externo para ZUMASUA Cía. Ltda., asesorando la implementación de sistemas de gestión de calidad e inocuidad alimentaria bajo estándares de Buenas Prácticas de Manufactura (ISO 22002) y principios HACCP, y ha participado como ponente en congresos internacionales de estadística aplicada.

Su vocación investigadora se evidencia en proyectos experimentales sobre riesgos psicosociales, tecnoestrés y salud ocupacional en instituciones de educación superior, así como en múltiples artículos científicos publicados en revistas indexadas sobre calidad, innovación, educación superior, sostenibilidad y procesos organizacionales. Profesional comprometido con la formación continua y el rigor metodológico, encarna valores de responsabilidad, ética y servicio, aportando desde la academia al desarrollo científico y productivo del Ecuador.

Ing. Paúl Alejandro Mora Dalgo MSc.

Universidad Estatal de Bolívar

pmora@mailes.ueb.edu.ec

<https://orcid.org/0009-0007-7564-3149>

Quito, Pichincha, Ecuador

Semblanza

Profesional en Seguridad, Salud Ocupacional y Gestión de Riesgos, Paúl Alejandro Mora Dalgo nació el 16 de septiembre de 1994 en Quito cuya vida ha estado marcada por su residencia en la Provincia Bolívar Cantón Guaranda, donde su formación fortaleció su compromiso con las comunidades y su comprensión de los desafíos asociados a los riesgos y las emergencias.

Esta experiencia ha orientado su interés por la gestión integral de riesgos, promoviendo la prevención, la preparación, la respuesta y la recuperación como pilares para fortalecer la resiliencia comunitaria y contribuir al bienestar colectivo.



Cuenta con una sólida formación académica: es Ingeniero en Administración para desastres y Gestión de Riesgo desde 2019. Posteriormente obtuvo el título de Magíster en Prevención y Gestión de Riesgo, Especialista en Gestión y Gobernanza Territorial cursada en el Instituto de Altos Estudios Nacionales del Ecuador IAEN. Además cuenta con múltiples cursos, talleres, Seminarios y Capacitaciones en Áreas como Seguridad y Salud Ocupacional, Gestión y Administración de Riesgo, Docencia y Pedagogía. Esta preparación refleja su vocación por la excelencia académica y la actualización permanente en las ciencias de la tierra.

En su trayectoria, ha destacado tanto en el ejercicio libre de la profesión como en el servicio público, promoviendo una cultura de Seguridad y Salud e Higiene en el trabajo con enfoque en la protección de la vida, la salud y la dignidad de las personas trabajadoras. Su compromiso profesional se orienta a la implementación de sistemas de gestión en seguridad y salud ocupacional, la identificación y control de riesgos laborales, el fortalecimiento de ambientes de trabajo saludables y el cumplimiento de la normativa vigente.

Psicóloga Clínica María Licensa Suárez Guerra MSc.

Investigador Independiente

licenciasuarez9973@gmail.com

<https://orcid.org/0009-0007-0189-6813>

Riobamba, Chimborazo, Ecuador

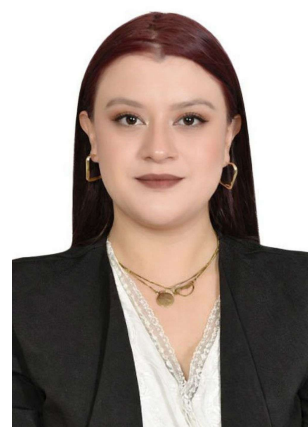
Semblanza

María Licensa Suárez Guerra, nacida el 30 de septiembre de 1999 en la ciudad de Ibarra, actualmente reside en Riobamba, Chimborazo. Es psicóloga clínica graduada en la Universidad Nacional de Chimborazo, con sólida experiencia en evaluación, diagnóstico e intervención psicológica. Ha culminado la Maestría en Psicoterapia y Terapias de Tercera Generación en la Universidad Internacional de La Rioja, actualmente en proceso de registro en SENESCYT.

Su formación se complementa con diplomados en neuropsicología, psicoterapia aplicada, estimulación cognitiva y terapia de lenguaje, además de seminarios sobre salud mental y neurodiversidad. Esta preparación refleja su compromiso con la actualización constante y la excelencia académica.

En su trayectoria profesional ha trabajado en la intervención clínica con niños, adolescentes y familias en situación de vulnerabilidad, así como en el acompañamiento psicológico en procesos de duelo y fortalecimiento de vínculos familiares. Su práctica se fundamenta en terapias basadas en evidencia, orientadas a promover el bienestar integral y la resiliencia en distintos grupos poblacionales.

Con valores de ética, innovación y compromiso social, María aporta a la construcción de un Ecuador más consciente y solidario, inspirando a través de su vocación clínica y su dedicación a la salud mental.



Ing. Hernán Sebastián Salguero Morales MSc.

Investigador Independiente

sbssalguero@gmail.com

<https://orcid.org/0009-0005-9655-612X>

Chimbo, Bolívar, Ecuador

Semblanza

Hernán Sebastián Salguero Morales nació el 4 de agosto de 1996 en el cantón San José de Chimbo, provincia de Bolívar, Ecuador. Es Ingeniero Comercial y posee una Maestría en Administración Pública, formación académica que ha fortalecido su visión sobre la gestión organizacional, la administración pública y el desarrollo institucional.

A lo largo de su trayectoria profesional se ha desempeñado en el sector público, destacándose por su compromiso, responsabilidad y vocación de servicio. Formó parte del Servicio Ecuatoriano de Capacitación Profesional (SECAP), donde ejerció funciones como Administrador de Contrato y estuvo encargado del área de Servicios Institucionales a nivel nacional, participando en la coordinación y fortalecimiento de procesos administrativos. Asimismo, adquirió experiencia en el Gobierno Autónomo Descentralizado Municipal del cantón Chimbo, desempeñándose como asistente en las áreas de Obras Públicas y Medio Ambiente, contribuyendo a la gestión de proyectos orientados al desarrollo local.

Desde sus inicios profesionales ha mantenido un profundo interés por la investigación científica y académica, convencido de que la generación y difusión del conocimiento constituyen herramientas fundamentales para comprender los desafíos actuales y proponer soluciones que contribuyan al desarrollo de las organizaciones y la sociedad. Su participación como coautor de esta obra refleja su compromiso con la producción intelectual, el aprendizaje continuo y el aporte de evidencia científica para fortalecer la toma de decisiones y promover el progreso en los ámbitos de la administración y la gestión pública.



Ing. Luis Andres Montero Quintana

Investigador Independiente

andresmonteroqfx@gmail.com

<https://orcid.org/0009-0000-9692-1469>

San José de Chimbo, Bolívar, Ecuador

Semblanza

Luis Andrés Montero Quintana es un ingeniero en tecnologías de la información ecuatoriano nacido el 5 de septiembre de 1998 en Guaranda, cuya vida y trayectoria se han desarrollado en San José de Chimbo, un entorno que ha marcado su arraigo y visión del desarrollo local. Desde temprano demostró un marcado interés por la tecnología, visualizándola no solo como una herramienta técnica, sino como un motor de transformación capaz de generar un antes y un después en la optimización de procesos en diversas áreas.



Cuenta con una sólida formación académica y técnica: es Ingeniero en Tecnologías de la Información por la Universidad de Guayaquil y actualmente amplía su perfil profesional cursando una Maestría en Inteligencia Artificial Aplicada a la Educación en la Universidad Bolivariana del Ecuador. Esta preparación refleja su enfoque vanguardista y su compromiso con la innovación constante y la evolución tecnológica.

A lo largo de su carrera, se ha especializado en la intersección de la tecnología y el sector financiero, acumulando diversas certificaciones en esta rama. Actualmente se desempeña como Auditor Informático en la Cooperativa de Ahorro y Crédito San Pedro Ltda., donde aplica sus conocimientos para garantizar la seguridad, eficiencia e integridad de los sistemas de la institución. Profesional comprometido, ético y con una clara visión de futuro, Luis Andrés combina la rigurosidad técnica con una constante búsqueda de soluciones innovadoras que impulsen el desarrollo y dejen una huella positiva en su entorno.



El contenido y las ideas expuestas en esta obra se encuentran protegidos por la normativa vigente en materia de propiedad intelectual y constituyen derechos exclusivos de su(s) autor(es).

Todos los derechos reservados © 2026

Sinopsis

Instrumentos de investigación científica: diseño, validación e implementación ofrece una guía integral para comprender, construir, aplicar y evaluar instrumentos destinados a la producción de datos científicos. La obra parte de los fundamentos epistemológicos de la medición y explica la relación entre realidad, constructo, variable, dimensión, indicador e ítem, con atención a los supuestos teóricos que orientan cada decisión metodológica.

El contenido desarrolla los niveles de medición, los paradigmas de investigación, las fuentes de error, la confiabilidad y las distintas evidencias de validez. A partir de esta base, presenta procedimientos para delimitar constructos, establecer dimensiones, formular indicadores, elaborar tablas de especificaciones, redactar ítems, seleccionar escalas de respuesta y someter los instrumentos a revisión lingüística, juicio de expertos, prueba cognitiva, pilotaje y validación estadística.

La propuesta incorpora los protocolos DCOI y PCUI para organizar la arquitectura conceptual y la construcción técnica del instrumento. También incluye orientaciones adaptadas a diferentes campos del conocimiento y examina las condiciones teóricas, culturales, poblacionales y operativas que determinan la pertinencia de una herramienta de investigación.

En su etapa final, el libro aborda los errores frecuentes que pueden comprometer la calidad científica de los instrumentos y presenta la Matriz de Coherencia Constructo-Dimensión-Ítem y el Checklist CADI como recursos para revisar la consistencia, la documentación y la trazabilidad del proceso. El propósito central es favorecer la elaboración de instrumentos rigurosos, transparentes, pertinentes y susceptibles de evaluación y replicación por la comunidad científica.

Palabras clave: instrumentos de investigación; medición científica; validación; confiabilidad; diseño metodológico.

Synopsis

Scientific Research Instruments: Design, Validation, and Implementation provides a comprehensive guide to understanding, constructing, applying, and evaluating instruments intended for the production of scientific data. The book begins with the epistemological foundations of measurement and explains the relationship among reality, construct, variable, dimension, indicator, and item, with attention to the theoretical assumptions guiding each methodological decision.

The content examines levels of measurement, research paradigms, sources of error, reliability, and different forms of validity evidence. Based on this foundation, it presents procedures for defining constructs, establishing dimensions, formulating indicators, preparing tables of specifications, writing items, selecting response scales, and submitting instruments to linguistic review, expert judgment, cognitive testing, pilot testing, and statistical validation.

The proposal incorporates the DCOI and PCUI protocols to organize the conceptual architecture and technical construction of the instrument. It also includes guidance adapted to different fields of knowledge and examines the theoretical, cultural, population-related, and operational conditions that determine the suitability of a research tool.

In its final stage, the book addresses common errors that may compromise the scientific quality of instruments and presents the Construct-Dimension-Item Coherence Matrix and the CADI Checklist as resources for reviewing consistency, documentation, and process traceability. Its central purpose is to support the development of rigorous, transparent, relevant instruments that can be evaluated and replicated by the scientific community.

Keywords: research instruments; scientific measurement; validation; reliability; methodological design.

CONTENIDO

Sinopsis	10
Synopsis	11
Introducción	21
CAPÍTULO 1 Fundamentos Epistemológicos del Instrumento de Investigación	23
1.1 La medición como acto científico y epistémico	24
1.1.1 ¿Qué es medir en ciencia?	24
1.1.2 Historia y evolución de la medición científica	25
1.1.3 La medición como construcción social e histórica	27
1.1.4 Medir vs. observar vs. registrar: distinciones fundamentales	27
1.2 Niveles de medición y sus implicaciones	28
1.2.1 Escala nominal	28
1.2.2 Escala ordinal	29
1.2.3 Escala de intervalo	29
1.2.4 Escala de razón	30
1.2.5 Implicaciones estadísticas de cada nivel	30
1.2.6 Errores frecuentes en la asignación de niveles	31
1.3 Paradigmas de investigación e instrumentos	31
1.3.1 Paradigma positivista y la medición objetiva	31
1.3.2 Paradigma interpretativo y la comprensión subjetiva	32
1.3.3 Paradigma crítico y transformador	32
1.3.4 Paradigma pragmático y los métodos mixtos	32
1.4 La cadena de la medición científica	34
1.4.1 Del fenómeno al dato: el camino invisible	34
1.4.2 Realidad → Constructo → Variable → Indicador → Ítem	34
1.4.3 Errores frecuentes en cada eslabón de la cadena	35
1.5 Taxonomía de instrumentos de investigación	36

1.5.1 Clasificación por enfoque metodológico	36
1.5.2 Clasificación por propósito investigativo	37
1.5.3 Clasificación por área del conocimiento	37
1.6 El error de medición: el enemigo silencioso	38
1.6.1 Naturaleza y tipos del error de medición	38
1.6.2 Error sistemático vs. error aleatorio	38
1.6.3 Fuentes del error: investigador, instrumento y contexto	40
1.6.4 Teoría Clásica del Test (TCT): fundamentos	40
1.6.5 El error estándar de medición	41
1.7 Validez: el corazón del instrumento	42
1.7.1 Evolución histórica del concepto de validez	42
1.7.2 Validez de contenido	43
1.7.3 Validez de constructo	43
1.7.4 Validez de criterio: concurrente y predictiva	44
1.7.5 La validez como argumento unificado (Messick, 1995)	44
1.7.6 Validez en investigación cualitativa: criterios alternativos	45
1.8 Confiabilidad: la consistencia del instrumento	46
1.8.1 El Coeficiente Alpha de Cronbach: uso, abuso y límites	46
1.8.2 Omega de McDonald: la alternativa robusta	47
1.8.3 Confiabilidad test-retest	48
1.8.4 Relación entre validez y confiabilidad	48
1.9 Marcos normativos internacionales de medición	49
1.9.1 Estándares APA/AERA/NCME	49
1.9.2 Marco ético internacional para instrumentos	49
1.9.3 Propiedad intelectual de los instrumentos	50
1.10 La Inteligencia Artificial como actor epistemológico en la medición	50
1.10.1 ¿Cambia la IA la naturaleza del medir?	50

1.10.2 El investigador aumentado: nuevo rol y responsabilidades.....	51
1.10.3 Principios éticos del uso de IA en instrumentación.....	52
CAPÍTULO 2 Del Problema al Constructo: La Arquitectura Conceptual del Instrumento	55
2.1 El problema de investigación como punto de partida	56
2.1.1 ¿Qué es un problema de medición?	56
2.1.2 De la pregunta de investigación a la necesidad de medir	57
2.1.3 Criterios para determinar si necesito un instrumento nuevo	58
2.1.4 Evaluación de instrumentos existentes antes de crear uno nuevo	59
2.1.5 Herramientas de IA para búsqueda de instrumentos existentes	60
2.2 El constructo: definición, delimitación y tipología	61
2.2.1 ¿Qué es un constructo científico?	61
2.2.2 Constructos simples versus constructos complejos	61
2.2.3 Constructos reflectivos versus formativos.....	63
2.2.4 Definición conceptual del constructo	64
2.2.5 Definición operacional del constructo	65
2.2.6 Errores frecuentes en la definición del constructo.....	66
2.3 Revisión sistemática del constructo	66
2.3.1 Por qué la revisión teórica es insustituible	66
2.3.2 Protocolo de revisión sistemática para instrumentación.....	67
2.3.3 Herramientas de IA para la revisión sistemática	68
2.3.4 Matriz de análisis de antecedentes teóricos	69
2.3.5 Identificación de vacíos en la literatura	70
2.4 Dimensionalidad del constructo	70
2.4.1 Constructos unidimensionales versus multidimensionales.....	70
2.4.2 Cómo identificar las dimensiones de un constructo	72
2.4.3 El árbol de dimensiones: herramienta de visualización.....	72
2.5 Operacionalización de variables.....	73

2.5.1 El proceso de operacionalización: de la teoría al dato.....	73
2.5.2 De constructo a dimensiones	74
2.5.3 De dimensiones a indicadores	74
2.5.4 De indicadores a ítems observables.....	76
2.5.5 Errores más comunes en la operacionalización.....	76
2.6 La tabla de especificaciones o blueprint del instrumento	77
2.6.1 ¿Qué es la tabla de especificaciones y por qué es indispensable?.....	77
2.6.2 Componentes de la tabla de especificaciones.....	78
2.6.3 Ponderación de dimensiones e indicadores	79
2.6.4 Número de ítems por dimensión: criterios técnicos	80
2.6.5 Plantilla de tabla de especificaciones	80
2.7 El Protocolo DCOI: Del Constructo al Instrumento	82
2.7.1 Fundamentos y justificación del Protocolo DCOI.....	82
2.7.2 Diagrama de flujo del Protocolo DCOI.....	83
2.7.3 Lista de verificación del Protocolo DCOI.....	84
2.8 Consideraciones éticas en la arquitectura conceptual	87
2.8.1 La ética no comienza con el consentimiento informado	87
2.8.2 Sesgos teóricos y cómo identificarlos	88
2.8.3 Representatividad y justicia en la definición del constructo	88
2.8.4 Consentimiento informado y protección de datos desde el diseño.....	89
2.9 Casos prácticos de aplicación del Protocolo DCOI	89
2.9.1 Caso práctico 1: Instrumento de Percepción de Riesgo ante Inundaciones en Comunidades Rurales.....	89
2.9.2 Caso práctico 2: Escala de Competencias Docentes para la Educación en Gestión del Riesgo.....	92
CAPÍTULO 3 Diseño y Construcción del Instrumento: Proceso, Enfoque y Técnica.....	95
3.1 El Protocolo PCUI: Proceso Completo Universal de Instrumentos.....	96
3.1.1 Fundamentos y justificación del PCUI.....	96

3.1.2 Principios rectores del PCUI	98
3.2 Diseño de ítems: el arte y la ciencia	99
3.2.1 Anatomía de un ítem: componentes y tipología	99
3.2.2 Las 20 reglas de oro para redactar ítems de calidad	100
3.2.3 Ítems positivos y negativos: ventajas, riesgos y estrategias	102
3.2.4 Sesgos sistemáticos en las respuestas: identificación y control	103
3.2.5 IA como generador y crítico de ítems	105
3.3 Instrumentos cuantitativos: del diseño al dato	106
3.3.1 Cuestionarios y encuestas	106
3.3.2 Escalas de medición de actitudes, percepciones y opiniones	109
3.3.3 Pruebas de conocimiento y rendimiento.....	111
3.3.4 Instrumentos de observación estructurada.....	113
3.4 Instrumentos cualitativos: capturando la complejidad	115
3.4.1 Guías de entrevista en profundidad	116
3.4.2 Guías de grupo focal.....	118
3.4.3 Protocolos de observación cualitativa	120
3.4.4 Criterios de rigor en instrumentos cualitativos.....	121
3.5 Instrumentos para investigación mixta.....	122
3.5.1 Lógica de la integración cuanti-cualitativa.....	122
3.5.2 El instrumento integrado: cuestionario con componente cualitativo.....	123
3.5.3 Matriz de Decisión Metodológica	124
3.6 Validez de contenido: el juicio de expertos	125
3.6.1 Selección y perfil de los expertos	126
3.6.2 El coeficiente V de Aiken: cálculo detallado e interpretación	126
3.6.3 Índice de Validez de Contenido (IVC) de Lawshe.....	127
3.6.4 Gestión de discrepancias entre expertos.....	128
3.7 Prueba cognitiva y pilotaje.....	128

3.7.1 Prueba cognitiva: el pensamiento en voz alta.....	128
3.7.2 El pilotaje: análisis y criterios de decisión	130
3.8 Análisis psicométrico avanzado	132
3.8.1 Análisis Factorial Confirmatorio (AFC).....	132
3.8.2 Análisis de confiabilidad: Alpha, Omega y más allá.....	134
3.8.3 Teoría de Respuesta al Ítem (TRI): introducción práctica	134
3.8.4 Análisis de Funcionamiento Diferencial de Ítems (DIF).....	135
3.9 Herramientas digitales e IA para la construcción y análisis.....	136
3.9.1 Software estadístico para análisis psicométrico	136
3.9.2 IA para análisis de datos cualitativos	137
CAPÍTULO 4 El Instrumento en su Contexto: Área, Población y Pertinencia Cultural	140
4.1 Por qué el contexto determina el instrumento	141
4.1.1 La ilusión de la universalidad en la medición	141
4.1.2 Niveles del contexto que afectan el instrumento	142
4.1.3 Adaptación cultural y lingüística de instrumentos	143
4.1.4 Consideraciones por tipo de población.....	145
4.2 Instrumentos en Ciencias de la Salud.....	146
4.2.1 Particularidades epistemológicas de la medición en salud	146
4.2.2 Tipos de instrumentos en ciencias de la salud.....	147
4.2.3 Estándares COSMIN para instrumentos de salud.....	147
4.2.4 La Mínima Diferencia Clínicamente Importante (MCID)	148
4.2.5 Caso práctico: Escala de Percepción de Calidad de Atención en Emergencias (EPCAE).....	149
4.3 Instrumentos en Ciencias de la Educación	151
4.3.1 La medición en educación: diversidad de propósitos.....	151
4.3.2 Instrumentos de evaluación del aprendizaje.....	151
4.3.3 Escalas de clima y convivencia escolar	152

4.3.4 Escalas de motivación y autorregulación del aprendizaje	153
4.3.5 Caso práctico: Instrumento de Evaluación de Competencias en Lectura Comprensiva (IELC).....	154
4.4 Instrumentos en Ciencias Sociales y Políticas	155
4.4.1 Particularidades de la medición en ciencias sociales.....	155
4.4.2 Índices sociales compuestos	155
4.4.3 Instrumentos de percepción ciudadana y opinión pública.....	156
4.4.4 Caso práctico: Índice de Percepción de Confianza Institucional (IPCI)	157
4.5 Instrumentos en Gestión del Riesgo de Desastres.....	158
4.5.1 Marco conceptual de la instrumentación en GRD.....	158
4.5.2 Instrumentos de evaluación de vulnerabilidad social	159
4.5.3 Caso práctico completo: Instrumento de Vulnerabilidad Comunitaria ante Sismos (IVCS)	160
4.6 Instrumentos en Ciencias Ambientales y Sostenibilidad	162
4.6.1 La medición en el campo ambiental: naturaleza híbrida	162
4.6.2 Escalas de percepción y comportamiento ambiental.....	162
4.6.3 Indicadores de sostenibilidad organizacional	163
4.6.4 Caso práctico: Escala de Conducta Proambiental Comunitaria (ECPAC).....	163
4.7 Instrumentos en Ciencias Administrativas y Organizacionales	164
4.7.1 La medición en el campo organizacional	164
4.7.2 Escalas de clima y cultura organizacional.....	165
4.7.3 Escalas de liderazgo: el caso del liderazgo transformacional.....	166
4.7.4 Caso práctico: Escala de Compromiso Organizacional en Contexto Latinoamericano (ECOL).....	166
4.8 Instrumentos en Ciencias Exactas e Ingeniería	167
4.8.1 Naturaleza de la instrumentación técnica	167
4.8.2 Instrumentos de evaluación de usabilidad y experiencia de usuario	167

4.8.3 Caso práctico: Protocolo de Evaluación de Sistemas de Alerta Temprana (PESAT)	168
4.9.1 Desafíos específicos de la medición en ciencias jurídicas	169
4.9.2 Instrumentos de percepción de justicia y acceso al derecho	169
4.9.3 Caso práctico: Escala de Percepción de Impunidad Ciudadana (EPIC)	170
CAPÍTULO 5 Errores Fatales, Coherencia Final y Expediente del Instrumento	174
5.1 Por qué los instrumentos fallan: una tipología del fracaso	175
5.1.1 La ilusión del instrumento terminado	175
5.1.2 Consecuencias del instrumento deficiente	176
5.1.3 Tipología del fracaso: cuatro grandes categorías de error	177
5.2 Los 30 errores fatales clasificados por fase	177
5.3 Análisis forense de instrumentos reales	182
5.3.1 Metodología del análisis forense	182
5.3.2 Caso forense 1: escala con fallo conceptual grave	183
5.3.3 Caso forense 2: escala con fallo psicométrico grave	184
5.3.4 Lecciones del análisis forense: patrones y señales de alerta	185
5.4 Coherencia interna del instrumento: la prueba definitiva	186
5.4.1 ¿Qué es la coherencia interna del instrumento?	186
5.4.2 La Matriz de Coherencia Constructo-Dimensión-Ítem (MCCDI)	186
5.4.3 Lista de verificación de coherencia: 30 criterios	189
5.5 Pertinencia del instrumento: más allá de la validez	191
5.5.1 ¿Qué es la pertinencia y cómo se diferencia de la validez?	191
5.5.2 Protocolo de verificación de pertinencia	191
5.6 El expediente técnico del instrumento	192
5.6.1 ¿Qué es el expediente técnico y por qué es obligatorio?	192
5.6.2 Componentes del expediente técnico completo	193
5.6.3 La ficha técnica del instrumento: modelo completo	195

5.7 El Checklist CADI: 50 criterios de calidad final.....	197
5.7.1 ¿Qué es el CADI?	197
5.7.2 Estructura del CADI: 50 criterios en 5 dimensiones	197
5.7.3 Interpretación y uso del CADI.....	201
5.8 La hoja de ruta del investigador: del constructo al instrumento validado.....	201
Referencias bibliográficas.....	206

Introducción

La calidad de una investigación depende, en buena medida, de la precisión con la que se observa, registra, interpreta y representa el fenómeno estudiado. Un instrumento deficiente puede producir datos consistentes, pero incapaces de reflejar el constructo que se pretende analizar. Por esta razón, la elaboración de instrumentos no debe reducirse a la formulación de preguntas o a la selección de una escala de respuesta. Implica una secuencia de decisiones epistemológicas, conceptuales, metodológicas, técnicas y éticas que determinan la calidad de la evidencia obtenida.

La medición científica establece una correspondencia entre propiedades de la realidad y sistemas de representación numéricos o categoriales. Esta correspondencia no ocurre de forma directa. Entre el fenómeno y el dato existe una cadena formada por constructos, variables, dimensiones, indicadores e ítems. Cada eslabón introduce decisiones que deben justificarse y documentarse. Cuando estas relaciones son ambiguas, incompletas o contradictorias, el instrumento pierde coherencia, aunque su presentación formal resulte adecuada o sus resultados estadísticos parezcan aceptables.

El diseño de instrumentos requiere definir con precisión qué se desea conocer, desde qué perspectiva teórica se interpreta el fenómeno y cuáles serán las manifestaciones observables empleadas para estudiarlo. También exige distinguir entre medir, observar y registrar, pues estas acciones responden a propósitos diferentes. Una entrevista, una escala, una rúbrica, una prueba de rendimiento, una guía de observación o un diario de campo pueden ser pertinentes, pero su selección depende de la pregunta de investigación, el enfoque metodológico, la población y las condiciones de aplicación.

Esta obra examina los fundamentos epistemológicos de la instrumentación científica y su relación con los paradigmas positivista, interpretativo, crítico y pragmático. Asimismo, desarrolla los niveles de medición y sus implicaciones para el análisis de los datos. La comprensión de estas bases permite evitar decisiones mecánicas, como aplicar procedimientos estadísticos sin considerar la naturaleza de las variables, utilizar instrumentos contruidos para contextos distintos o adoptar categorías que no representan adecuadamente la experiencia de la población estudiada.

La construcción técnica del instrumento comienza con la delimitación del constructo y continúa con la identificación de sus dimensiones, la selección de indicadores, la elaboración de la tabla de especificaciones y la redacción de los ítems. Estas etapas requieren coherencia interna y

correspondencia con el marco conceptual. No basta con incluir un número elevado de preguntas. Cada ítem debe cumplir una función definida, emplear un lenguaje comprensible, evitar ambigüedades y ajustarse a las características culturales, educativas y comunicativas de quienes responderán el instrumento.

El proceso de validación tampoco puede entenderse como una actividad aislada realizada al final del diseño. La revisión lingüística, el juicio de expertos, las entrevistas cognitivas, la prueba piloto y los análisis psicométricos aportan evidencias diferentes y complementarias. La validez se sustenta en la calidad de las interpretaciones y los usos derivados de las puntuaciones, mientras que la confiabilidad expresa el grado de consistencia de la medición. Un instrumento puede ser confiable sin medir correctamente el constructo, por lo que ambos criterios deben analizarse de manera articulada.

El libro incorpora los protocolos DCOI y PCUI para organizar, respectivamente, la definición conceptual y operacional del instrumento y su proceso de construcción. Estas propuestas ordenan las decisiones que suelen aparecer dispersas en diferentes etapas de la investigación. También se presentan la Matriz de Coherencia Constructo-Dimensión-Ítem y el Checklist CADI como recursos de revisión destinados a verificar la correspondencia conceptual, la documentación de los procedimientos y la trazabilidad del instrumento.

La instrumentación científica adquiere características particulares según el campo de conocimiento. Las ciencias sociales, la educación, la salud, la administración, la gestión del riesgo y otras áreas emplean instrumentos con propósitos, poblaciones y criterios específicos. Por ello, la obra combina fundamentos comunes con orientaciones aplicadas, sin asumir que un mismo procedimiento puede transferirse automáticamente entre disciplinas. La adaptación exige analizar el contexto, el lenguaje, las condiciones de administración y las consecuencias asociadas con la interpretación de los resultados.

Instrumentos de investigación científica: diseño, validación e implementación está dirigido a estudiantes, docentes, investigadores y profesionales que necesitan construir, adaptar, seleccionar o evaluar herramientas para la recolección de datos. Su propósito es acompañar el proceso completo, desde la definición del fenómeno hasta la valoración final de la calidad del instrumento. La obra busca que cada decisión pueda explicarse, verificarse y reproducirse, de modo que los datos obtenidos constituyan una base pertinente para responder las preguntas de investigación.

CAPÍTULO

1

Fundamentos Epistemológicos del Instrumento de Investigación

Gregory Leandro Montenegro Bosquez

$$X = \sum_{i=1}^n x_i$$

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum s_i^2}{s_T^2} \right)$$



CAPÍTULO 1

Fundamentos Epistemológicos del Instrumento de Investigación

Por qué y qué significa medir en ciencia

En la historia del conocimiento humano, pocas preguntas han resultado tan simples y filosóficamente profundas como esta: ¿qué significa medir? A primera vista, medir parece un acto cotidiano y mecánico: asignamos números a objetos, comparamos magnitudes, registramos observaciones. Sin embargo, cuando nos adentramos en la dimensión científica de este acto, descubrimos que la medición es, en realidad, una construcción epistemológica de enorme complejidad que involucra supuestos filosóficos, decisiones teóricas, convenciones culturales y juicios de valor que pocas veces son explicitados con la claridad que merecen.

Este capítulo inaugural establece los cimientos conceptuales sobre los que se edificará toda la obra. Antes de diseñar un solo ítem, antes de elegir una escala o redactar una pregunta, el investigador riguroso debe comprender con profundidad qué es lo que está intentando medir, desde qué posición epistemológica lo hace, con qué supuestos opera y cuáles son las implicaciones de cada decisión metodológica. Saltarse este paso es como construir una catedral sin haber estudiado el terreno: la estructura puede parecer imponente, pero sus cimientos serán frágiles.

A lo largo de estas páginas, el lector encontrará no solo definiciones y clasificaciones, sino argumentos filosóficos, debates contemporáneos, fórmulas matemáticas y herramientas prácticas que le permitirán tomar decisiones informadas en cada etapa del proceso de construcción de instrumentos. La epistemología, lejos de ser un ornamento académico, es la brújula que orienta cada decisión metodológica.

1.1 La medición como acto científico y epistémico

1.1.1 ¿Qué es medir en ciencia?

La medición científica puede definirse, en su formulación más clásica, como el proceso sistemático de asignar números o categorías a objetos, personas o fenómenos de acuerdo con reglas explícitas y fundamentadas teóricamente (Stevens, 1946). Esta definición, aunque seminal, ha sido objeto de revisiones y ampliaciones durante más de siete décadas. Hoy sabemos que la medición no es simplemente una operación técnica de asignación numérica, sino un acto cargado de supuestos ontológicos y epistemológicos.

Michell (1999) señala con agudeza que la ciencia occidental ha padecido durante siglos una *patología cuantitativa*: la tendencia a asumir que todo fenómeno relevante puede y debe ser expresado en números, sin examinar críticamente si los atributos en cuestión poseen realmente una estructura cuantitativa. Esta crítica es especialmente pertinente en las ciencias sociales, de la salud y de la educación, donde la mayoría de los constructos de interés —inteligencia, motivación, calidad de vida, percepción del riesgo— son entidades teóricas inobservables cuya cuantificación requiere una cadena de inferencias que debe ser cuidadosamente justificada.

Definición fundamental

Medir, en el sentido científico del término, es el proceso de establecer una correspondencia sistemática, replicable y teóricamente justificada entre propiedades de los fenómenos del mundo real y un sistema de representación formal —numérico o categorial— que permite su comparación, análisis e interpretación.

Esta definición ampliada incorpora tres elementos que son frecuentemente omitidos en los textos de metodología: (a) la sistematicidad, que implica procedimientos explícitos y replicables; (b) la justificación teórica, que exige que la representación numérica o categorial sea coherente con el modelo teórico del constructo; y (c) la interpretabilidad, que requiere que los resultados de la medición sean comprensibles y significativos en el contexto de la pregunta de investigación.

1.1.2 Historia y evolución de la medición científica

La historia de la medición científica es inseparable de la historia del pensamiento occidental sobre la naturaleza del conocimiento. Aristóteles ya distinguía entre propiedades que admiten grados como la temperatura o el peso y propiedades que no los admiten como la identidad o la contradicción, anticipando en más de dos milenios la moderna teoría de los niveles de medición.

Sin embargo, la medición en las ciencias naturales no alcanzó su forma moderna hasta el siglo XVII, con el desarrollo de instrumentos de precisión como el termómetro, el barómetro, el telescopio y la matematización de la física por parte de Galileo, Kepler y Newton. La gran revolución conceptual de este período fue la idea de que la naturaleza está escrita en lenguaje matemático y que la medición es el puente privilegiado entre el mundo empírico y el mundo formal de las matemáticas.

Las ciencias sociales y del comportamiento tardaron considerablemente más en desarrollar tradiciones de medición rigurosas. El psicólogo Gustav Fechner (1860) fue uno de los primeros

en intentar medir sistemáticamente fenómenos mentales dando origen a la psicofísica. Galton, a finales del siglo XIX, introdujo las correlaciones estadísticas como herramienta para estudiar diferencias individuales. Binet y Simon (1905) desarrollaron las primeras pruebas de inteligencia. Thurstone (1928) y Likert (1932) crearon escalas para medir actitudes. Cada uno de estos hitos fue posible gracias a avances simultáneos en la teoría estadística y en la conceptualización de los constructos.

Tabla 1.1

Hitos históricos en el desarrollo de la medición científica

Período	Autor(es)	Contribución	Disciplina
Siglo IV a.C.	Aristóteles	Distinción entre propiedades continuas y discretas	Filosofía
1638	Galileo Galilei	Matematización de la física y medición experimental	Física
1860	Gustav Fechner	Psicofísica: medición de sensaciones subjetivas	Psicología
1904	Charles Spearman	Correlaciones y teoría del factor g de inteligencia	Psicometría
1928	Louis Thurstone	Escalas de actitudes por intervalos aparentemente iguales	Psicometría
1932	Rensis Likert	Escalas sumativas de actitudes (escala Likert)	Psicología Social
1946	S. S. Stevens	Teoría de los niveles de medición (nominal, ordinal, intervalo, razón)	Psicología
1951	Lee Cronbach	Coefficiente Alpha de consistencia interna	Psicometría
1960	Georg Rasch	Modelos de Teoría de Respuesta al Ítem (TRI)	Psicometría
1995	Samuel Messick	Concepción unificada de la validez como argumento	Psicometría
2010-presente	Múltiples autores	IA y aprendizaje automático aplicados a la medición	Ciencias de Datos

Nota. Elaboración propia basada en revisión histórica de la psicometría y metodología de la investigación.

1.1.3 La medición como construcción social e histórica

Una perspectiva frecuentemente ignorada en los manuales de metodología es que la medición no es un proceso neutral y ahistórico, sino que está profundamente enraizada en contextos sociales, culturales e históricos específicos. Los constructos que medimos, las escalas que utilizamos y los criterios con que juzgamos la validez de nuestras mediciones reflejan valores, intereses y convenciones que cambian con el tiempo y varían entre culturas.

Foucault (1975), desde una perspectiva crítica, señaló que los instrumentos de medición en ciencias sociales, especialmente los desarrollados para evaluar la inteligencia, la salud mental o la conducta han sido históricamente utilizados como mecanismos de normalización y control social. Sin llegar necesariamente a conclusiones tan radicales, la epistemología contemporánea reconoce que el investigador que construye un instrumento de medición no está capturando pasivamente una realidad preexistente, sino *construyendo activamente una representación* de esa realidad a partir de categorías teóricas que son históricas y culturalmente situadas.

Esto tiene consecuencias prácticas inmediatas. Un instrumento diseñado para medir la calidad de vida en países de altos ingresos puede ser conceptualmente inadecuado para poblaciones de países con estructuras económicas, sistemas de valores y concepciones del bienestar muy diferentes. Un instrumento de evaluación del aprendizaje diseñado a partir de currículos occidentales puede medir preferentemente el capital cultural de ciertos grupos y penalizar sistemáticamente a otros. Reconocer la dimensión social e histórica de la medición no es relativismo epistemológico; es rigor científico.

1.1.4 Medir vs. observar vs. registrar: distinciones fundamentales

En la práctica investigativa cotidiana, los términos *medir*, *observar* y *registrar* son frecuentemente utilizados como sinónimos. Sin embargo, desde una perspectiva metodológica rigurosa, estas tres operaciones son conceptualmente distintas y su confusión genera errores graves en el diseño de investigaciones.

Tabla 1.2

Distinciones conceptuales entre medir, observar y registrar en investigación científica

Dimensión	Observar	Registrar	Medir
Definición	Percibir sistemáticamente fenómenos o eventos	Documentar lo observado en un sistema de representación	Asignar valores según reglas a propiedades de fenómenos

Dimensión	Observar	Registrar	Medir
Requisito teórico	Marco conceptual orientador	Sistema de categorías o unidades	Modelo teórico del constructo
Resultado	Descripción cualitativa o cuantitativa	Datos estructurados (notas, registros)	Puntuaciones o categorías cuantificables
Ejemplo	Observar comportamiento en aula	Registrar frecuencia de interrupciones	Aplicar escala de conducta disruptiva
Instrumento típico	Protocolo de observación	Diario de campo, ficha	Escala, test, cuestionario

Nota. Elaboración propia.

La distinción más crítica es entre registrar y medir. Un diario de campo que documenta narrativamente las interacciones en un aula es un instrumento de registro; una escala de valoración del clima de aula que asigna puntuaciones a dimensiones previamente definidas es un instrumento de medición. Ambos son legítimos y valiosos, pero responden a lógicas epistemológicas diferentes y sus resultados no son directamente comparables.

1.2 Niveles de medición y sus implicaciones

La teoría de los niveles de medición constituye uno de los fundamentos conceptuales más importantes de la metodología de la investigación. Propuesta originalmente por Stevens (1946) en su influyente artículo en *Science*, esta teoría establece que no todas las mediciones son equivalentes: dependiendo de las propiedades matemáticas de la escala utilizada, se pueden realizar unas operaciones y no otras, y se justifican unos análisis estadísticos y no otros.

1.2.1 Escala nominal

La escala nominal es el nivel más básico de medición. En ella, los números o categorías se utilizan exclusivamente como etiquetas para identificar o clasificar objetos o personas, sin implicar ninguna relación de orden o magnitud entre las categorías. Los números son puramente arbitrarios: se podría sustituir 1 por rojo y 2 por azul sin perder ninguna información. Ejemplos de variables nominales: género (hombre/mujer/no binario), tipo de institución (pública/privada), diagnóstico médico (hipertensión/diabetes/asma), área de conocimiento (ciencias/humanidades/artes).

Las operaciones permitidas con escala nominal son exclusivamente las de igualdad y diferencia: se puede afirmar que dos objetos son iguales o diferentes en la propiedad medida,

pero no que uno es mayor o menor, ni cuánto mayor o menor. Las estadísticas apropiadas son la frecuencia, la moda, el porcentaje y el coeficiente Kappa para concordancia.

1.2.2 Escala ordinal

La escala ordinal añade a la clasificación nominal la propiedad del orden: las categorías pueden ser ordenadas de menor a mayor (o de mayor a menor) en la propiedad medida. Sin embargo, los intervalos entre categorías adyacentes no son necesariamente iguales.

Ejemplos: nivel socioeconómico (bajo/medio/alto), grado de acuerdo en escala Likert (totalmente en desacuerdo... totalmente de acuerdo), nivel educativo (primaria/secundaria/universitario), clasificación de rendimiento académico (insuficiente/suficiente/bueno/excelente).

Debate metodológico central

El tratamiento estadístico de las escalas Likert es uno de los debates más encendidos en metodología. Aunque técnicamente son ordinales, muchos investigadores las tratan como intervalos. Carifio y Perla (2008) argumentan que escalas Likert de 5 o más puntos pueden tratarse como intervalos cuando los datos de la escala total son aproximadamente normales. Sin embargo, otros autores como Jamieson (2004) defienden que este tratamiento viola supuestos fundamentales y puede conducir a conclusiones erróneas.

Las operaciones permitidas con escala ordinal incluyen las de igualdad, diferencia y orden (mayor/menor), pero no las operaciones aritméticas como la suma o la división. Las estadísticas apropiadas incluyen la mediana, los percentiles, los rangos y los coeficientes de correlación de Spearman (ρ) y Kendall (τ).

1.2.3 Escala de intervalo

La escala de intervalo mantiene las propiedades de clasificación y orden de las escalas anteriores, y añade la característica de que los intervalos entre valores adyacentes son iguales y constantes. Esto permite operaciones aritméticas de suma y resta, pero no de multiplicación y división, debido a que el cero de la escala es arbitrario no representa ausencia del atributo.

Ejemplos: temperatura en grados Celsius o Fahrenheit (0° no significa ausencia de temperatura), puntuaciones estandarizadas de tests psicológicos (cociente intelectual, puntuaciones T), fechas del calendario.

Con escala de intervalo se justifican las estadísticas descriptivas estándar: media aritmética, desviación estándar, correlación de Pearson (r), y un amplio rango de pruebas paramétricas

inferenciales (t de Student, ANOVA, regresión lineal), siempre que se cumplan los supuestos distribucionales correspondientes.

1.2.4 Escala de razón

La escala de razón es el nivel más completo de medición. Posee todas las propiedades anteriores y además cuenta con un cero absoluto y verdadero que representa la ausencia total del atributo medido. Esto permite todas las operaciones aritméticas, incluyendo la multiplicación y la división, lo que hace que las razones entre valores sean significativas. Ejemplos: peso, altura, tiempo, frecuencia cardíaca, ingresos económicos en unidades monetarias, número de errores en una prueba, cantidad de hijos.

Con escala de razón se pueden realizar todas las operaciones estadísticas, incluyendo medidas como el coeficiente de variación ($CV = \sigma/\mu \times 100$) y los modelos log-lineales. La media geométrica, que requiere un cero verdadero, es apropiada para estas escalas.

Tabla 1.3

Propiedades matemáticas y análisis estadísticos según nivel de medición

Propiedad	Nominal	Ordinal	Intervalo	Razón	Cero absoluto	Estadísticos apropiados
Clasificación	Sí	Sí	Sí	Sí	No	Frecuencia, moda
Orden	No	Sí	Sí	Sí	No	Mediana, percentiles, ρ Spearman
Intervalos iguales	No	No	Sí	Sí	No	Media, DE, r Pearson, ANOVA
Cero absoluto	No	No	No	Sí	Sí	CV, media geométrica, modelos log

Nota. Sí = propiedad presente; No = propiedad ausente. DE = desviación estándar; CV = coeficiente de variación; r = correlación de Pearson; ρ = rho de Spearman.

1.2.5 Implicaciones estadísticas de cada nivel

La elección del nivel de medición no es una decisión meramente técnica; tiene profundas implicaciones para la validez de los análisis estadísticos y, en última instancia, para la legitimidad de las conclusiones de la investigación. Usar estadísticos diseñados para escalas de intervalo con datos nominales produce resultados matemáticamente calculables, pero epistemológicamente sin sentido.

No obstante, la relación entre nivel de medición y análisis estadístico no es tan rígida como Stevens (1946) sugirió originalmente. La estadística robusta contemporánea reconoce que lo que realmente importa no es el nivel de medición nominal de la escala, sino la distribución real de los datos y el grado de robustez de los procedimientos estadísticos frente a violaciones de sus supuestos. El debate está lejos de estar cerrado.

1.2.6 Errores frecuentes en la asignación de niveles

Entre los errores más comunes que cometen los investigadores al determinar el nivel de medición de sus variables se encuentran los siguientes:

- Tratar variables ordinales como si fueran de intervalo sin justificación teórica ni evidencia empírica de que los intervalos son aproximadamente iguales. Este error es especialmente frecuente con escalas Likert de 4 o 5 puntos.
- Asumir que una variable es de razón cuando en realidad no tiene un cero absoluto significativo. Por ejemplo, tratar puntuaciones de tests psicológicos como si tuvieran un cero verdadero.
- Confundir el nivel de medición de la variable con el tipo de datos obtenidos. Una variable puede tener propiedades de intervalo, aunque los datos se recojan con opciones de respuesta categóricas.
- Ignorar el nivel de medición al elegir las pruebas estadísticas, utilizando pruebas paramétricas con variables nominales u ordinales sin evaluar la robustez del procedimiento.
- Cambiar el nivel de medición de una variable en el transcurso de la investigación sin documentar ni justificar este cambio metodológico.

1.3 Paradigmas de investigación e instrumentos

Uno de los principios más importantes y menos comprendidos de la metodología de la investigación es que la elección del instrumento de recolección de datos no puede separarse del paradigma epistemológico desde el cual se investiga. El paradigma no es simplemente una preferencia metodológica o una costumbre disciplinar; es un sistema de supuestos ontológicos, epistemológicos y metodológicos que determina qué puede ser conocido, cómo puede ser conocido y qué tipos de evidencia son legítimos y relevantes (Kuhn, 1962; Guba & Lincoln, 1994).

1.3.1 Paradigma positivista y la medición objetiva

El paradigma positivista, heredero del empirismo lógico del siglo XX, sostiene que existe una realidad objetiva, independiente del observador, que puede ser conocida a través de la observación sistemática y la medición precisa. Desde esta perspectiva, el instrumento ideal es aquel que minimiza la influencia del investigador, produce datos cuantitativos replicables y permite la generalización estadística a poblaciones más amplias.

Los instrumentos propios del paradigma positivista son cuestionarios estandarizados, escalas psicométricas validadas, pruebas de rendimiento, registros de observación estructurada y experimentos controlados. La validez y la confiabilidad, en sus versiones cuantitativas clásicas, son los criterios de rigor por excelencia desde esta perspectiva.

1.3.2 Paradigma interpretativo y la comprensión subjetiva

El paradigma interpretativo también denominado constructivista, fenomenológico o hermenéutico según el énfasis filosófico, parte del supuesto de que la realidad social es una construcción subjetiva e intersubjetiva, producto de las interpretaciones, significados y experiencias de los actores sociales. Desde esta perspectiva, el objetivo de la investigación no es medir con precisión variables predefinidas, sino comprender en profundidad los significados que las personas atribuyen a sus experiencias.

Los instrumentos del paradigma interpretativo son la entrevista en profundidad, el grupo focal, la observación participante, el análisis documental y los diarios de campo. El criterio de rigor no es la confiabilidad estadística, sino la credibilidad, la transferibilidad, la dependencia y la confirmabilidad (Lincoln & Guba, 1985).

1.3.3 Paradigma crítico y transformador

El paradigma crítico comparte con el interpretativo el énfasis en la comprensión de los significados sociales, pero añade una dimensión emancipatoria: la investigación debe no solo comprender la realidad sino contribuir a transformarla, especialmente en lo que respecta a las relaciones de poder, la injusticia social y la opresión. Los instrumentos en este paradigma no son neutros; deben ser diseñados con y para las comunidades investigadas, de manera que contribuyan a su empoderamiento.

Los instrumentos propios de este paradigma incluyen entrevistas biográficas, técnicas participativas, cartografía social, análisis crítico del discurso y métodos de investigación-acción participativa. La validez catalítica es un criterio de rigor específico de este paradigma (Lather, 1991).

1.3.4 Paradigma pragmático y los métodos mixtos

El paradigma pragmático rechaza la dicotomía entre positivismo e interpretativismo y sostiene que la elección de los métodos e instrumentos debe guiarse por la pregunta de investigación y el propósito práctico del estudio, no por compromisos filosóficos previos. Desde esta perspectiva, la combinación de instrumentos cuantitativos y cualitativos no solo es legítima sino frecuentemente necesaria para abordar preguntas de investigación complejas.

El paradigma pragmático ha dado lugar al florecimiento de los métodos mixtos como tradición de investigación consolidada (Creswell & Plano Clark, 2018). Los instrumentos mixtos son la expresión metodológica natural de este paradigma.

Tabla 1.4

Caracterización comparativa de los paradigmas de investigación y sus implicaciones para los instrumentos

Dimensión	Positivista	Interpretativo	Crítico	Pragmático
Ontología	Realismo ingenuo: realidad objetiva única	Relativismo: realidades múltiples y construidas	Realismo histórico: realidad moldeada por el poder	Pluralismo ontológico según el problema
Epistemología	Objetivismo: investigador neutral	Subjetivismo: investigador implicado	Subjetivismo crítico: investigador comprometido	Pragmatismo: lo que funciona
Metodología	Experimental, cuantitativa, deductiva	Hermenéutica, cualitativa, inductiva	Dialógica, participativa, transformadora	Mixta, pragmática, orientada al problema
Instrumentos típicos	Escalas, cuestionarios, tests, experimentos	Entrevistas, grupos focales, observación	IAP, entrevistas biográficas, cartografía social	Cualquier instrumento pertinente al problema
Criterios de rigor	Validez y confiabilidad cuantitativas	Credibilidad, transferibilidad, confirmabilidad	Validez catalítica, rigor dialógico	Criterios mixtos según el propósito
Propósito	Explicar, predecir, controlar	Comprender, interpretar, describir	Transformar, emancipar, empoderar	Resolver problemas, informar decisiones

Nota. IAP = Investigación-Acción Participativa. Adaptado de Guba & Lincoln (1994) y Creswell & Plano Clark (2018).

1.4 La cadena de la medición científica

1.4.1 Del fenómeno al dato: el camino invisible

Uno de los errores más persistentes en la investigación científica consiste en tratar los datos como si fueran la realidad misma, ignorando la cadena de transformaciones teóricas y operacionales que los separa del fenómeno que se pretende estudiar. Entre el fenómeno real y el número o la categoría que aparece en la base de datos del investigador, existe una serie de pasos de traducción que cada uno introduce sus propios supuestos y posibles fuentes de error. Este camino invisible es lo que denominaremos la cadena de la medición científica. Comprender cada uno de sus eslabones es fundamental para diseñar instrumentos que sean verdaderamente válidos (que midan lo que afirman medir) y no simplemente confiables (que produzcan resultados consistentes, aunque equivocados).

1.4.2 Realidad → Constructo → Variable → Indicador → Ítem

La cadena de la medición científica puede representarse como una secuencia de cinco eslabones, cada uno de los cuales implica decisiones teóricas y metodológicas específicas:

Eslabón 1 — Realidad

El punto de partida es el fenómeno o aspecto de la realidad que el investigador desea estudiar. Este fenómeno existe con independencia del investigador y de su marco conceptual. Sin embargo, el investigador nunca accede directamente a la realidad; siempre la percibe a través del filtro de sus categorías conceptuales, su formación disciplinar y su posición social.

Eslabón 2 — Constructo

El constructo es la representación teórica del fenómeno. Es una entidad abstracta, inobservable directamente, que el investigador postula para explicar o describir patrones de comportamiento, actitudes, características o capacidades. La motivación académica, la calidad del sueño, la percepción del riesgo o la resiliencia comunitaria son constructos: no se pueden ver ni tocar directamente, pero se puede inferir su existencia y sus variaciones a partir de indicadores observables.

Eslabón 3 — Variable

La variable es la operacionalización del constructo: la forma en que este se expresará en la investigación. Una variable es una propiedad que puede tomar dos o más valores y cuyas variaciones son de interés para la investigación. Un constructo puede descomponerse en varias variables, sus dimensiones o componentes, cada una de las cuales refleja un aspecto diferente del fenómeno.

Eslabón 4 — Indicador

El indicador es una manifestación observable y medible de la variable. Mientras que las variables son abstracciones conceptuales, los indicadores son sus expresiones empíricas concretas. Por ejemplo, la variable *bienestar emocional* puede tener como indicadores la frecuencia de estados de ánimo positivos, la ausencia de síntomas depresivos o la satisfacción con las relaciones interpersonales.

Eslabón 5 — Ítem

El ítem es la unidad operativa del instrumento: la pregunta, la afirmación, la tarea o el estímulo concreto al que el participante debe responder. Es el eslabón más visible de la cadena, pero también es el que está más alejado del fenómeno original. La calidad del ítem depende de qué tan fielmente captura el indicador que pretende medir.

Tabla 1.5

Ilustración de la cadena de medición científica aplicada a tres constructos

Constructo	Variable (dimensión)	Indicador	Ejemplo de ítem	Área de aplicación
Motivación académica	Motivación intrínseca	Disfrute de las actividades de aprendizaje	"Realizo mis tareas porque disfruto aprendiendo cosas nuevas"	Educación
Calidad de vida	Bienestar físico	Capacidad funcional para actividades diarias	"Durante la última semana, ¿cuántas dificultades tuvo para subir escaleras?"	Ciencias de la salud
Vulnerabilidad social	Capacidad de respuesta	Acceso a redes de apoyo comunitario	"En caso de una emergencia, ¿con cuántas personas fuera de su hogar podría contar para recibir ayuda?"	Gestión del Riesgo de Desastres
Clima organizacional	Comunicación interna	Frecuencia y calidad de la comunicación ascendente	"Mi supervisor escucha y toma en cuenta mis ideas y sugerencias"	Ciencias administrativas

Nota. Elaboración propia.

1.4.3 Errores frecuentes en cada eslabón de la cadena

Cada eslabón de la cadena de medición es una oportunidad para cometer errores que comprometan la validez del instrumento. Los más frecuentes son:

- En el eslabón Constructo: definir el constructo de manera vaga, circular o inconsistente con la literatura especializada. Por ejemplo, definir la autoestima como 'la valoración que una persona hace de sí misma' sin especificar qué dimensiones incluye ni qué teoría la sustenta.
- En el eslabón Variable: descomponer el constructo en dimensiones que se solapan o que no cubren adecuadamente el dominio del constructo. Ambos errores comprometan la validez de contenido.
- En el eslabón Indicador: seleccionar indicadores que son fáciles de medir pero no son los más relevantes para el constructo. También es frecuente confundir causas del constructo con indicadores del constructo.
- En el eslabón Ítem: redactar ítems ambiguos, con doble carga, con lenguaje técnico inadecuado para la población o con sesgos de deseabilidad social que distorsionan las respuestas.

1.5 Taxonomía de instrumentos de investigación

El universo de los instrumentos de investigación es extraordinariamente diverso. Encuestas telefónicas y entrevistas etnográficas, escalas psicométricas y protocolos de observación conductual, pruebas de rendimiento y diarios de campo: todos son instrumentos de investigación, pero difieren en sus fundamentos epistemológicos, sus procedimientos de construcción, sus criterios de calidad y sus ámbitos de aplicación. Una taxonomía clara y comprehensiva es el primer paso para tomar decisiones informadas sobre qué tipo de instrumento es más adecuado para cada propósito investigativo.

1.5.1 Clasificación por enfoque metodológico

Tabla 1.6

Clasificación de instrumentos de investigación por enfoque metodológico

Enfoque	Instrumentos principales	Tipo de datos	Análisis típico
Cuantitativo	Cuestionarios, escalas, tests, registros estructurados, experimentos	Numéricos, ordinales, de intervalo o razón	Estadística descriptiva e inferencial, psicometría
Cualitativo	Guías de entrevista, grupos focales, protocolos de observación, diarios de campo	Texto, imágenes, audio, video	Análisis temático, fenomenológico, de discurso, grounded theory

Enfoque	Instrumentos principales	Tipo de datos	Análisis típico
Mixto	Instrumentos integrados, diseños secuenciales y simultáneos	Numéricos y textuales combinados	Integración cuantitativa, meta-inferencias

Nota. Elaboración propia.

1.5.2 Clasificación por propósito investigativo

Los instrumentos también pueden clasificarse según el propósito para el que se utilizan en la investigación:

- Instrumentos de medición: diseñados para cuantificar constructos o propiedades de interés. Incluyen escalas, tests y cuestionarios estandarizados.
- Instrumentos de exploración: diseñados para descubrir y comprender fenómenos poco conocidos. Incluyen guías de entrevista en profundidad y grupos focales.
- Instrumentos de diagnóstico: diseñados para identificar o clasificar individuos, grupos o situaciones. Incluyen pruebas diagnósticas, screenings y evaluaciones clínicas.
- Instrumentos de evaluación: diseñados para valorar el nivel de logro, desempeño o calidad. Incluyen rúbricas, listas de cotejo evaluativas y escalas de calificación.
- Instrumentos de monitoreo: diseñados para seguimiento continuo en el tiempo. Incluyen registros longitudinales, fichas de seguimiento y sistemas de indicadores.

1.5.3 Clasificación por área del conocimiento

Cada área del conocimiento ha desarrollado tradiciones específicas de instrumentación que responden a sus preguntas, sus poblaciones y sus estándares disciplinares. En el Capítulo 4 de esta obra se desarrolla en detalle la instrumentación especializada para cada área. A modo de orientación inicial, se presentan los instrumentos más característicos por campo:

Tabla 1.7

Instrumentos más característicos por área del conocimiento

Área	Instrumentos característicos	Organismos de referencia
Ciencias de la Salud	Escalas clínicas, PROs, instrumentos epidemiológicos, CVRS	OMS, COSMIN, FDA
Educación	Pruebas de rendimiento, escalas de clima, rúbricas, diagnósticos	OCDE, UNESCO, PISA
Ciencias Sociales	Índices sociales, encuestas de opinión, escalas de capital social	CEPAL, PNUD, BM

Área	Instrumentos característicos	Organismos de referencia
GRD	Índices de riesgo, escalas de vulnerabilidad, instrumentos de capacidad	UNDRR, Marco Sendai
Ciencias Ambientales	Indicadores de sostenibilidad, escalas de conducta proambiental	PNUMA, ISO 14001
Administración	Escalas de clima organizacional, engagement, desempeño	SHRM, APA
Ingeniería/Exactas	Protocolos de medición técnica, instrumentos de usabilidad	ISO, IEEE, ANSI
Ciencias Jurídicas	Instrumentos de percepción de justicia, análisis normativo	ONU, UNODC

Nota. PROs = Patient-Reported Outcomes; CVRS = Calidad de vida relacionada con la salud; GRD = Gestión del Riesgo de Desastres

1.6 El error de medición: el enemigo silencioso

1.6.1 Naturaleza y tipos del error de medición

En un mundo ideal, cualquier instrumento de medición produciría exactamente la misma puntuación para la misma persona en las mismas condiciones, y esa puntuación correspondería perfectamente al nivel real del atributo que se pretende medir. En la realidad científica, ningún instrumento es perfectamente preciso. Todo proceso de medición está acompañado de algún grado de error: una discrepancia entre el valor observado y el valor verdadero del atributo.

La Teoría Clásica del Test (TCT), formulada por Spearman (1904) y desarrollada posteriormente por Gulliksen (1950) y Lord & Novick (1968), proporciona el marco conceptual fundamental para entender el error de medición. Según la TCT, cualquier puntuación observada (X) puede descomponerse en dos componentes: la puntuación verdadera (V) y el error de medición (e):

$$X = V + e \quad (\text{Ecuación 1.1})$$

Donde X es la puntuación observada del sujeto en el instrumento, V es su puntuación verdadera el valor que se obtendría si el instrumento midiera perfectamente sin error, y e es el error de medición asociado a esa observación particular.

1.6.2 Error sistemático vs. error aleatorio

La TCT distingue dos tipos fundamentales de error de medición, con causas, consecuencias y estrategias de control radicalmente diferentes:

Error aleatorio

El error aleatorio es impredecible, varía de una observación a otra sin un patrón consistente y, en promedio, se cancela a sí mismo cuando se realizan muchas mediciones. Sus fuentes incluyen fluctuaciones momentáneas en el estado del participante (fatiga, distracción, ansiedad transitoria), ambigüedades menores en la redacción de los ítems, variaciones fortuitas en las condiciones de administración y errores de registro o codificación.

El error aleatorio afecta la confiabilidad del instrumento: a mayor error aleatorio, menor confiabilidad. Sin embargo, el error aleatorio no introduce sesgos sistemáticos y puede reducirse aumentando el número de ítems o de mediciones.

Error sistemático

El error sistemático, también denominado sesgo, es consistente y predecible: afecta a todas las mediciones en la misma dirección y magnitud. Sus fuentes incluyen redacción sesgada de los ítems, características del administrador que influyen en las respuestas, características del instrumento que favorecen sistemáticamente ciertos tipos de respuesta (deseabilidad social, aquiescencia) y variables de confusión que se miden junto con el constructo de interés.

El error sistemático afecta la validez del instrumento: aunque el instrumento sea muy consistente, puede estar midiendo consistentemente lo equivocado. Esta es la razón por la que la confiabilidad es condición necesaria pero no suficiente para la validez.

Tabla 1.8

Comparación entre error aleatorio y error sistemático en la medición científica

Dimensión	Error aleatorio	Error sistemático (sesgo)
Patrón	Impredecible, varía aleatoriamente	Consistente, mismo sentido y magnitud
Efecto en promedio	Se cancela con muchas mediciones	Persiste independientemente del tamaño muestral
Afecta principalmente	Confiabilidad	Validez
Fuentes típicas	Fatiga, distracción, ambigüedad menor de ítems	Deseabilidad social, sesgos del evaluador, ítems tendenciosos
Estrategia de control	Aumentar número de ítems y mediciones	Diseño cuidadoso del instrumento y protocolo de aplicación
Detección	Análisis de confiabilidad (Alpha, test-retest)	Análisis de validez de constructo, análisis de sesgo de ítems

Nota. Elaboración propia basada en la Teoría Clásica del Test.

1.6.3 Fuentes del error: investigador, instrumento y contexto

Las fuentes del error de medición pueden clasificarse en tres grandes categorías:

a) Fuentes asociadas al investigador o evaluador:

- Expectativas del investigador que influyen en la recolección o codificación de datos (efecto Rosenthal o efecto Pygmalion).
- Sesgos en la redacción de ítems que reflejan el marco cultural o ideológico del autor.
- Inconsistencias en la administración del instrumento entre diferentes aplicadores.
- Errores de codificación, transcripción o entrada de datos.

b) Fuentes asociadas al instrumento:

- Ítems ambiguos que permiten múltiples interpretaciones válidas.
- Número insuficiente de ítems para representar adecuadamente el dominio del constructo.
- Escalas de respuesta inadecuadas para el constructo o la población.
- Instrucciones poco claras o incompletas.
- Efectos de posición y orden de los ítems.

c) Fuentes asociadas al contexto de aplicación:

- Condiciones ambientales desfavorables (ruido, temperatura, iluminación).
- Momento temporal inadecuado (final de jornada, períodos de estrés).
- Presencia de otras personas que inhiben o distorsionan las respuestas.
- Uso de tecnología con la que el participante no está familiarizado.

1.6.4 Teoría Clásica del Test (TCT): fundamentos

La Teoría Clásica del Test (TCT) establece una serie de supuestos matemáticos sobre la naturaleza del error de medición que son fundamentales para entender los análisis de confiabilidad y validez. Los supuestos básicos son:

Supuesto 1:

$$E(e) = 0$$

El valor esperado del error de medición es cero. Esto significa que, en promedio y a lo largo de múltiples mediciones, el error aleatorio se cancela a sí mismo.

Supuesto 2:

$$r(V, e) = 0$$

La correlación entre la puntuación verdadera y el error de medición es cero. Los errores no están relacionados sistemáticamente con el nivel real del atributo.

Supuesto 3:

$$r(e_1, e_2) = 0$$

Los errores de diferentes mediciones (ítems o aplicaciones) no están correlacionados entre sí. Esta es la base del supuesto de independencia local en la Teoría de Respuesta al Ítem.

A partir de estos supuestos se puede derivar que la varianza de las puntuaciones observadas es igual a la suma de la varianza verdadera y la varianza del error:

$$\sigma^2 X = \sigma^2 V + \sigma^2 e \quad (\text{Ecuación 1.2})$$

Esta ecuación es la base de la definición formal de la confiabilidad como el cociente entre la varianza verdadera y la varianza observada:

$$\rho X' = \sigma^2 V / \sigma^2 X = 1 - (\sigma^2 e / \sigma^2 X) \quad (\text{Ecuación 1.3})$$

Donde $\rho X'$ es el coeficiente de confiabilidad, $\sigma^2 V$ es la varianza verdadera, $\sigma^2 X$ es la varianza observada y $\sigma^2 e$ es la varianza del error. Un coeficiente de confiabilidad de 1.00 indicaría medición perfecta sin error; un coeficiente de 0.00 indicaría que toda la varianza observada es error puro.

1.6.5 El error estándar de medición

El error estándar de medición (SEM, por sus siglas en inglés) es quizás el estadístico más útil para interpretar puntuaciones individuales, ya que cuantifica directamente la incertidumbre asociada a la puntuación de un individuo específico. Su fórmula es:

$$SEM = \sigma X \times \sqrt{(1 - \rho X')} \quad (\text{Ecuación 1.4})$$

Donde SEM es el error estándar de medición, σX es la desviación estándar de las puntuaciones observadas y $\rho X'$ es el coeficiente de confiabilidad del instrumento.

El SEM permite construir intervalos de confianza alrededor de la puntuación observada de un individuo para estimar el rango probable de su puntuación verdadera. El intervalo de confianza al 95% es:

$$IC_{95\%} = X \pm 1.96 \times SEM \quad (\text{Ecuación 1.5})$$

Ejemplo de aplicación del SEM

Suponga que una escala de ansiedad tiene una desviación estándar de 10 puntos y una confiabilidad de 0.84. El SEM sería: $SEM = 10 \times \sqrt{(1 - 0.84)} = 10 \times \sqrt{0.16} = 10 \times 0.40 = 4.0$. Si un participante obtiene una puntuación observada de 45, su intervalo de confianza al 95% sería: $IC = 45 \pm (1.96 \times 4.0) = 45 \pm 7.84$, es decir, entre 37.16 y 52.84. Esto significa que tenemos 95% de confianza en que la puntuación verdadera del participante se encuentra en ese rango.

1.7 Validez: el corazón del instrumento

La validez es, sin lugar a dudas, el concepto más importante en la teoría de la medición psicológica, educativa y social. Un instrumento puede ser perfectamente confiable pero completamente inválido si no mide el constructo que pretende medir. Inversamente, un instrumento inválido, por muy confiable que sea, no tiene valor científico: mide con gran precisión lo que no debería medir o no mide lo que debería.

1.7.1 Evolución histórica del concepto de validez

El concepto de validez ha experimentado una evolución conceptual notable a lo largo del siglo XX, pasando de una visión fragmentada y operacional a una concepción unificada y argumentativa que hoy domina el campo.

En la visión clásica, predominante hasta los años 1970, la validez se concebía como un conjunto de tipos relativamente independientes: la validez de contenido (¿el instrumento abarca el dominio relevante?), la validez de criterio (¿las puntuaciones predicen o correlacionan con un criterio externo?) y la validez de constructo (¿el instrumento mide el constructo teórico que afirma medir?). Se hablaba de validar instrumentos, como si la validez fuera una propiedad binaria (el instrumento es válido o no lo es) y permanente.

La gran revolución conceptual llegó con el trabajo seminal de Messick (1989, 1995), quien propuso una concepción unificada e integradora de la validez. Para Messick, la validez no es un atributo del instrumento sino del *uso e interpretación de sus puntuaciones* en un contexto específico para un propósito determinado. La validez es siempre relacional y contextual: un instrumento puede ser válido para un propósito y una población, e inválido para otro propósito u otra población.

1.7.2 Validez de contenido

La validez de contenido se refiere al grado en que los ítems del instrumento representan de manera adecuada y comprensiva el dominio o universo de contenido del constructo que se pretende medir. No basta con que los ítems sean relevantes; deben ser también representativos y técnicamente adecuados.

La evaluación de la validez de contenido es fundamentalmente un proceso de juicio experto, aunque debe estar guiado por procedimientos sistemáticos y criterios explícitos. El procedimiento más utilizado y metodológicamente robusto es el coeficiente V de Aiken, que cuantifica el acuerdo entre jueces sobre la relevancia de cada ítem:

$$V = S / [n(c - 1)] \quad (\text{Ecuación 1.6})$$

Donde S es la suma de las valoraciones de los jueces menos el valor mínimo posible, n es el número de jueces y c es el número de categorías de la escala de valoración. El coeficiente V varía entre 0 y 1; valores ≥ 0.70 son generalmente considerados aceptables, aunque el umbral exacto depende del número de jueces y del nivel de significación estadística establecido (Aiken, 1980).

1.7.3 Validez de constructo

La validez de constructo es el núcleo de la validez en la concepción contemporánea. Se refiere al grado en que el instrumento mide el constructo teórico que pretende medir, y no otro constructo diferente. La evidencia de validez de constructo se acumula a través de múltiples líneas de evidencia convergentes:

- Evidencia basada en la estructura interna del instrumento: el análisis factorial confirmatorio (AFC) permite evaluar si la estructura dimensional de las puntuaciones es consistente con el modelo teórico del constructo.
- Evidencia basada en relaciones con otras variables: el instrumento debería correlacionar positivamente con medidas de constructos teóricamente relacionados (validez convergente) y no correlacionar, o correlacionar negativamente, con medidas de constructos teóricamente no relacionados (validez discriminante).
- Evidencia basada en la respuesta al proceso: evidencia de que los participantes utilizan los procesos cognitivos y afectivos que el constructo teórico supone que están implicados.

- Evidencia basada en las consecuencias del uso: análisis de si el uso del instrumento produce consecuencias sociales justas y equitativas.

La validez convergente y discriminante puede evaluarse formalmente mediante el análisis de la varianza media extraída (AVE, por sus siglas en inglés) y la fiabilidad compuesta (FC), derivados del AFC:

$$AVE = (\sum \lambda_i^2) / [(\sum \lambda_i^2) + (\sum \varepsilon_i)] \quad (\text{Ecuación 1.7})$$

$$FC = (\sum \lambda_i)^2 / [(\sum \lambda_i)^2 + (\sum \varepsilon_i)] \quad (\text{Ecuación 1.8})$$

Donde λ_i son las cargas factoriales estandarizadas de cada ítem y ε_i son sus errores de medición. Se recomienda $AVE \geq 0.50$ para evidencia de validez convergente adecuada (Fornell & Larcker, 1981) y que la raíz cuadrada del AVE de cada factor sea mayor que las correlaciones entre factores para evidencia de validez discriminante.

1.7.4 Validez de criterio: concurrente y predictiva

La validez de criterio evalúa el grado en que las puntuaciones del instrumento se relacionan con un criterio externo relevante. Dependiendo de la relación temporal entre el instrumento y el criterio, se distinguen dos subtipos:

Validez concurrente:

El instrumento y el criterio se administran aproximadamente al mismo tiempo. Es apropiada para instrumentos de diagnóstico o clasificación donde se quiere demostrar que el instrumento nuevo produce clasificaciones equivalentes a las del criterio establecido, con menor costo o complejidad. Por ejemplo, una escala breve de depresión podría validarse concurrentemente comparando sus puntuaciones con las de una entrevista clínica estructurada.

Validez predictiva:

El instrumento se administra antes que el criterio, y se evalúa su capacidad para predecir el desempeño o comportamiento futuro. Es especialmente relevante para instrumentos de selección (selección de personal, admisión a programas académicos) o de evaluación de riesgo (riesgo de reincidencia delictiva, riesgo de abandono escolar).

1.7.5 La validez como argumento unificado (Messick, 1995)

La contribución más importante de Messick (1995) fue proponer que la validez no es una colección de tipos independientes sino un argumento integrado y acumulativo que debe

justificar cada uso e interpretación específica de las puntuaciones. Este argumento debe responder a seis preguntas fundamentales:

Tabla 1.9

Las seis facetas del argumento de validez según Messick (1995)

Faceta	Pregunta central	Evidencia requerida
Evidencial-sustantiva	¿El instrumento mide el constructo que pretende medir?	AFC, análisis de ítems, teoría del constructo
Estructural	¿La estructura del instrumento refleja la estructura del constructo?	AFC, análisis de la dimensionalidad
Generalización	¿Las inferencias son válidas para otras poblaciones y contextos?	Estudios de invarianza, metaanálisis
Validez externa	¿Las puntuaciones predicen comportamientos relevantes?	Estudios predictivos, validez de criterio
Consecuencias	¿El uso del instrumento produce consecuencias justas?	Análisis de impacto diferencial, estudios de equidad
Valor implicado	¿Qué valores implica el uso del instrumento?	Análisis ético, revisión de supuestos normativos

Nota. Adaptado de Messick (1995). Esta es la concepción de validez adoptada en la presente obra.

1.7.6 Validez en investigación cualitativa: criterios alternativos

La investigación cualitativa no adopta los criterios de validez cuantitativos, que están fundamentados en supuestos positivistas incompatibles con su epistemología. Lincoln y Guba (1985) propusieron un sistema alternativo de criterios de rigor que son el equivalente cualitativo de la validez y la confiabilidad:

- Credibilidad (equivalente a validez interna): grado en que los hallazgos representan fielmente las perspectivas y experiencias de los participantes. Se establece mediante triangulación, verificación por los participantes (member checking) y observación prolongada.
- Transferibilidad (equivalente a validez externa): grado en que los hallazgos pueden ser relevantes en otros contextos. Se facilita mediante descripciones densas que permitan al lector juzgar su aplicabilidad a su propio contexto.

- Dependencia (equivalente a confiabilidad): consistencia y estabilidad de los hallazgos a lo largo del tiempo y entre diferentes investigadores. Se evalúa mediante auditorías del proceso de investigación.
- Confirmabilidad (equivalente a objetividad): grado en que los hallazgos reflejan las perspectivas de los participantes y no los sesgos del investigador. Se establece mediante reflexividad explícita y auditoría de confirmabilidad.

1.8 Confiabilidad: la consistencia del instrumento

La confiabilidad es una propiedad necesaria de cualquier instrumento de medición. Un instrumento confiable produce puntuaciones consistentes y reproducibles: si se administra dos veces al mismo individuo en condiciones similares, produce puntuaciones similares; si se administra simultáneamente con una forma alternativa del mismo instrumento, los resultados son equivalentes; y si se calculan las puntuaciones basándose en diferentes subconjuntos de ítems, se obtienen resultados coherentes.

1.8.1 El Coeficiente Alpha de Cronbach: uso, abuso y límites

El coeficiente Alpha de Cronbach (α), propuesto por Lee Cronbach en 1951, es el indicador de confiabilidad más utilizado en ciencias sociales, de la salud y de la educación. Su popularidad está completamente justificada por su elegancia matemática, su facilidad de cálculo y su disponibilidad en todos los paquetes estadísticos. Sin embargo, también es probablemente el estadístico más frecuentemente mal interpretado y mal utilizado en la investigación científica.

Formalmente, el Alpha de Cronbach se define como:

$$\alpha = (k / (k-1)) \times (1 - (\sum \sigma_i^2 / \sigma^2 X)) \quad (\text{Ecuación 1.9})$$

Donde k es el número de ítems, $\sum \sigma_i^2$ es la suma de las varianzas de cada ítem y $\sigma^2 X$ es la varianza de las puntuaciones totales. Matemáticamente equivalente es la expresión en términos de correlaciones inter-ítem:

$$\alpha = (k \times \bar{r}) / [1 + (k - 1) \times \bar{r}] \quad (\text{Ecuación 1.10})$$

Donde $\bar{r}$ es la correlación media entre todos los pares de ítems. Esta expresión revela que el Alpha depende de dos factores: el número de ítems (k) y la correlación media entre ellos. Esto tiene una implicación práctica importante: el Alpha puede aumentarse artificialmente añadiendo más ítems, incluso si esos ítems añaden poca información nueva.

Tabla 1.10*Interpretación convencional del coeficiente Alpha de Cronbach y sus limitaciones*

Valor de Alpha	Interpretación convencional	Limitaciones y matices
≥ 0.90	Excelente	Puede indicar redundancia excesiva entre ítems
0.80 - 0.89	Bueno	Aceptable para la mayoría de usos de investigación
0.70 - 0.79	Aceptable	Mínimo para uso en investigación grupal
0.60 - 0.69	Cuestionable	Uso restringido; no recomendado para decisiones individuales
0.50 - 0.59	Pobre	Revisar y mejorar el instrumento antes de usar
< 0.50	Inaceptable	El instrumento no tiene consistencia interna mínima

Nota. Basado en George y Mallery (2003) y Nunnally (1978). Los valores de corte pueden variar según el propósito del instrumento; instrumentos de alta consecuencia (diagnóstico clínico, decisiones de selección) requieren $\alpha \geq 0.90$.

Los principales abusos y malinterpretaciones del Alpha incluyen: (a) interpretar el Alpha como indicador de unidimensionalidad, cuando en realidad es compatible con estructuras multidimensionales; (b) usar el Alpha como único indicador de calidad del instrumento, ignorando la validez; (c) no reportar los intervalos de confianza del Alpha; y (d) no considerar el impacto del tamaño muestral en la estabilidad del estimado.

1.8.2 Omega de McDonald: la alternativa robusta

El coeficiente Omega de McDonald (ω), propuesto originalmente por McDonald (1999) y revitalizado por Revelle y Zinbarg (2009), ha ganado creciente reconocimiento como alternativa superior al Alpha de Cronbach, especialmente cuando los ítems no tienen cargas factoriales iguales.

El Omega total (ω_t) se calcula a partir de las cargas factoriales obtenidas en el AFC:

$$\omega_t = (\sum \lambda_i)^2 / [(\sum \lambda_i)^2 + \sum \epsilon_i] \quad (\text{Ecuación 1.11})$$

Donde λ_i son las cargas factoriales estandarizadas y ϵ_i son los errores de medición de cada ítem. A diferencia del Alpha, el Omega no asume que todos los ítems contribuyen igualmente a la medición del constructo; esto lo hace más adecuado para instrumentos con estructura factorial no uniforme.

Recomendación para la práctica

En la actualidad, la práctica recomendada es reportar tanto el Alpha de Cronbach como el Omega de McDonald, especialmente cuando los ítems presentan cargas factoriales heterogéneas. Si los valores de Alpha y Omega difieren sustancialmente, el Omega es el estimado más preciso de la confiabilidad. Herramientas como el paquete psych de R permiten calcular ambos coeficientes con sus intervalos de confianza bootstrapeados en pocos segundos.

1.8.3 Confiabilidad test-retest

La confiabilidad test-retest evalúa la estabilidad temporal de las puntuaciones: si el instrumento se administra dos veces al mismo grupo de personas en dos momentos diferentes, ¿qué tan similares son los resultados? Se estima calculando la correlación entre las puntuaciones de las dos aplicaciones, y para variables ordinales o cuando no se asume linealidad, se utiliza la correlación intraclass (CCI):

$$CCI = (MSS - MSE) / (MSS + (k-1) \times MSE + k(MSB - MSE)/n) \quad (\text{Ecuación 1.12})$$

Donde MSS es la media cuadrática entre sujetos, MSE es la media cuadrática del error y MSB es la media cuadrática entre mediciones. El intervalo de tiempo óptimo entre las dos aplicaciones debe ser suficientemente largo para que los participantes no recuerden sus respuestas anteriores (generalmente 2-4 semanas), pero suficientemente corto para que el constructo no haya cambiado genuinamente.

1.8.4 Relación entre validez y confiabilidad

La relación entre validez y confiabilidad puede resumirse en dos principios fundamentales:

1. La confiabilidad es condición necesaria pero no suficiente para la validez. Un instrumento debe ser confiable para poder ser válido (sin consistencia no puede haber precisión), pero puede ser muy confiable sin ser válido (medir consistentemente lo equivocado).
2. La confiabilidad establece un techo para la validez. Matemáticamente, la correlación máxima posible entre un instrumento y cualquier criterio externo está limitada por la raíz cuadrada de su coeficiente de confiabilidad:

$$r_{\text{máx}}(X, Y) \leq \sqrt{(\rho X' \times \rho Y')} \quad (\text{Ecuación 1.13})$$

Esto significa que un instrumento con confiabilidad de 0.64 no puede correlacionar más de 0.80 con ningún criterio, independientemente de cuán válido sea conceptualmente. Este principio tiene importantes implicaciones para el diseño de estudios de validez criterial.

1.9 Marcos normativos internacionales de medición

1.9.1 Estándares APA/AERA/NCME

Los *Standards for Educational and Psychological Testing*, publicados conjuntamente por la American Psychological Association (APA), la American Educational Research Association (AERA) y el National Council on Measurement in Education (NCME), constituyen el documento normativo más influyente para el desarrollo y uso de instrumentos en educación y psicología. La edición más reciente (AERA/APA/NCME, 2014) organiza sus estándares en cinco grandes áreas:

- Fundamentos de la validez: el argumento de validez como base de cualquier uso del instrumento.
- Confiabilidad y errores de medición: requisitos mínimos de precisión según el propósito del instrumento.
- Equidad en las pruebas: garantías de que el instrumento funciona con equidad para diferentes grupos poblacionales.
- Operaciones de las pruebas: estándares para el desarrollo, administración y puntuación.
- Usos específicos: estándares para evaluación educativa, pruebas de empleo, evaluación psicológica y evaluaciones programáticas.

1.9.2 Marco ético internacional para instrumentos

La International Test Commission (ITC) ha publicado directrices internacionales para el uso ético de los instrumentos de medición que complementan los estándares técnicos con consideraciones éticas fundamentales. Estas directrices establecen que:

- Los instrumentos solo deben ser utilizados por profesionales con la formación técnica y ética adecuada.
- Los participantes tienen derecho a ser informados sobre el propósito de la evaluación, los usos que se darán a los resultados y sus derechos sobre la información.
- Los resultados deben ser comunicados de manera comprensible, respetando la privacidad y dignidad de los participantes.

- Los instrumentos deben ser culturalmente apropiados para las poblaciones en que se usan.
- Ningún instrumento debe usarse para discriminar injustamente a personas o grupos.

1.9.3 Propiedad intelectual de los instrumentos

Los instrumentos de medición son creaciones intelectuales protegidas por los derechos de autor. Esto tiene implicaciones importantes para los investigadores:

- Reproducir o administrar un instrumento protegido sin permiso del titular de los derechos constituye una infracción legal, independientemente de que sea para uso comercial o académico.
- La mayoría de los instrumentos publicados en revistas científicas están disponibles para uso en investigación no comercial, pero deben citarse correctamente y, en algunos casos, requieren permiso explícito.
- Los instrumentos desarrollados como parte de tesis o proyectos de investigación financiados pertenecen generalmente a la institución financiadora o al investigador, según los acuerdos contractuales establecidos.
- El registro de instrumentos en repositorios abiertos como OSF o Zenodo con una licencia Creative Commons es una práctica de ciencia abierta que facilita su uso y garantiza el reconocimiento de la autoría.

1.10 La Inteligencia Artificial como actor epistemológico en la medición

La irrupción de la inteligencia artificial y en particular de los grandes modelos de lenguaje (LLMs) y los sistemas de aprendizaje automático en el proceso de investigación científica plantea preguntas epistemológicas de enorme profundidad que van más allá de la simple adopción de nuevas herramientas tecnológicas. La IA no es solo una herramienta más eficiente para hacer lo mismo que hacíamos antes; es un actor que transforma cualitativamente la naturaleza del proceso de medición.

1.10.1 ¿Cambia la IA la naturaleza del medir?

La respuesta honesta es: en algunos aspectos, sí; en otros, no. Lo que no cambia con la IA es la lógica fundamental de la validez: un instrumento sigue siendo válido solo si sus puntuaciones permiten inferencias justificadas sobre el constructo de interés. Lo que sí puede cambiar y está cambiando, es la manera en que se construyen los instrumentos, se recolectan los datos, se procesan las respuestas y se interpretan los resultados.

Los LLMs pueden generar ítems para escalas en segundos, analizar la equivalencia semántica entre traducciones de instrumentos, detectar sesgos potenciales en la redacción, simular respuestas para pruebas cognitivas preliminares y asistir en la interpretación de resultados psicométricos complejos. Estas capacidades son genuinamente revolucionarias desde el punto de vista de la eficiencia investigativa.

Sin embargo, la IA introduce también nuevos riesgos epistemológicos que los investigadores deben conocer y gestionar activamente:

- Alucinaciones y fabricaciones: los LLMs pueden generar ítems aparentemente plausibles pero conceptualmente inconsistentes con el marco teórico del constructo.
- Sesgos de entrenamiento: los modelos de IA están entrenados mayoritariamente en textos de culturas anglosajonas de clase media, lo que puede introducir sesgos culturales sistemáticos en los instrumentos que generan o evalúan.
- Opacidad del proceso: a diferencia del juicio de un experto humano, las razones por las que un LLM aprueba o rechaza un ítem son difíciles de examinar críticamente.
- Dependencia acrítica: el riesgo de que los investigadores acepten sin cuestionamiento las sugerencias de la IA por el efecto de autoridad algorítmica.

1.10.2 El investigador aumentado: nuevo rol y responsabilidades

El concepto de investigador aumentado designa al investigador que integra estratégicamente las capacidades de la IA en su proceso de trabajo sin renunciar a su rol de agente epistémico crítico. En este modelo, la IA es una herramienta poderosa que amplifica las capacidades del investigador pero el investigador mantiene en todo momento la responsabilidad de las decisiones metodológicas y éticas.

En el contexto específico de la construcción de instrumentos, el modelo del investigador aumentado implica:

- Usar la IA para generar borradores de ítems, pero someter siempre estos borradores a una revisión experta rigurosa.
- Usar herramientas de IA para mapear la literatura teórica del constructo, pero verificar críticamente las referencias y no aceptar ciegamente las síntesis generadas.
- Usar IA para análisis estadísticos complejos, pero comprender los fundamentos de los métodos utilizados para poder interpretar críticamente los resultados.

- Documentar explícitamente en el reporte del instrumento qué partes del proceso de construcción contaron con asistencia de IA y cómo se gestionaron los riesgos asociados.

1.10.3 Principios éticos del uso de IA en instrumentación

Hasta la fecha, no existe un consenso internacional completamente consolidado sobre los principios éticos específicos para el uso de IA en la construcción de instrumentos de investigación. Sin embargo, a partir de los marcos éticos generales para el uso de IA en ciencia (UNESCO, 2021; OCDE, 2019) y de las directrices específicas de las principales asociaciones de investigación, es posible proponer los siguientes principios fundamentales:

Tabla 1.11

Principios éticos para el uso de Inteligencia Artificial en la construcción de instrumentos de investigación

Principio	Descripción	Implicación práctica
Transparencia	Declarar explícitamente el uso de IA en el proceso de construcción del instrumento	Incluir en la sección de método una descripción de qué herramientas de IA se utilizaron y para qué
Supervisión humana	Mantener la revisión y decisión humana experta en todas las etapas críticas	Ningún ítem debe incluirse sin revisión experta humana, independientemente de su origen
Equidad	Verificar que el uso de IA no introduce o amplifica sesgos hacia grupos específicos	Revisar explícitamente los ítems generados por IA en busca de sesgos de género, etnia, cultura o clase
Responsabilidad	El investigador es responsable de la calidad del instrumento, independientemente de quién generó los ítems	No puede invocarse a la IA como excusa ante errores o sesgos del instrumento
Verificabilidad	Las contribuciones de la IA deben poder ser auditadas y verificadas	Documentar los prompts utilizados y las respuestas obtenidas en el repositorio del proyecto
No maleficencia	Evitar usos de IA que puedan causar daño a los participantes o a la integridad científica	No usar IA para fabricar o distorsionar datos de validación

Nota. Elaboración propia basada en UNESCO (2021), OCDE (2019) y principios de la Open Science Initiative.

Reflexiones finales del capítulo

Este capítulo ha recorrido los fundamentos epistemológicos que sustentan cualquier proceso de construcción de instrumentos de investigación. Hemos visto que medir no es un acto simple ni neutro, sino una práctica científica cargada de supuestos filosóficos, históricos y culturales que el investigador riguroso debe conocer y gestionar con conciencia crítica.

Los niveles de medición de Stevens (nominal, ordinal, intervalo y razón) nos recuerdan que no todos los números son equivalentes y que las decisiones sobre el tipo de escala tienen consecuencias directas sobre los análisis estadísticos legítimos. Los paradigmas de investigación nos muestran que el instrumento siempre está al servicio de una epistemología más amplia, y que elegir un instrumento implica también tomar posición respecto a qué tipo de conocimiento queremos producir.

La cadena de la medición es la columna vertebral conceptual del proceso de construcción de cualquier instrumento. Cada eslabón de esta cadena introduce supuestos y posibles fuentes de error que deben ser gestionados con rigor. La validez, en su concepción contemporánea como argumento integrado (Messick, 1995), y la confiabilidad, en sus múltiples expresiones formales, son los criterios de calidad que permiten juzgar si un instrumento merece ser usado en la investigación científica.

Finalmente, la irrupción de la inteligencia artificial como herramienta de apoyo al proceso de investigación plantea tanto oportunidades extraordinarias como riesgos nuevos que exigen respuestas éticas y metodológicas cuidadosas. El modelo del investigador aumentado propone una integración crítica y responsable de estas herramientas, donde la eficiencia tecnológica nunca sacrifica el rigor epistemológico.

El siguiente capítulo pondrá en marcha todo este andamiaje conceptual, llevando al lector a través del proceso concreto de definición del constructo y construcción de la arquitectura conceptual del instrumento.

Preguntas de reflexión

1. Usted está diseñando un instrumento para medir la 'calidad de vida laboral' de docentes universitarios en contextos latinoamericanos. Describa cómo influiría en su diseño adoptar un paradigma positivista versus uno interpretativo. ¿Cuáles serían las diferencias concretas en el tipo de instrumento, los procedimientos de validación y los criterios de rigor utilizados?

2. Un colega le muestra un cuestionario sobre 'percepción de seguridad ciudadana' que ha diseñado utilizando exclusivamente un LLM. Los ítems son gramaticalmente correctos y parecen relevantes. Sin embargo, usted detecta que no existe una definición conceptual del constructo ni una tabla de especificaciones. Usando los conceptos del capítulo, explique qué problemas de validez podría tener este instrumento y qué pasos debería seguir para corregirlos.
3. Analice la Ecuación 1.13 (relación entre confiabilidad y validez criterial máxima). Un instrumento de evaluación de competencias para la gestión del riesgo de desastres tiene una confiabilidad de 0.72. ¿Cuál es la correlación máxima posible con un criterio externo de desempeño en campo con confiabilidad de 0.80? ¿Qué implicaciones tiene este resultado para el diseño del estudio de validez?
4. Reflexione sobre la siguiente afirmación: 'La cadena de medición no es simplemente un procedimiento técnico, sino una toma de posición epistemológica sobre la naturaleza del fenómeno que se estudia.' ¿Está de acuerdo? Justifique su respuesta con un ejemplo de un área de su disciplina específica.
5. Diseñe un plan de cinco acciones concretas para integrar herramientas de IA en el proceso de construcción de un instrumento de su área, respetando los principios éticos discutidos en la sección 1.10.3. Para cada acción, identifique el beneficio esperado y el riesgo potencial.

CAPÍTULO

2

Del Problema al Constructo: La Arquitectura Conceptual del Instrumento

Paúl Alejandro Mora Dalgo

$$X = \sum_{i=1}^n x_i$$

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum s_i^2}{s_T^2} \right)$$



CAPÍTULO 2

Del Problema al Constructo: La Arquitectura Conceptual del Instrumento

Antes de diseñar cualquier ítem, necesitas saber exactamente qué medir y por qué

Si el Capítulo 1 estableció los fundamentos filosóficos del acto de medir, este capítulo desciende al terreno operativo donde comienza, de manera concreta, el proceso de construcción de cualquier instrumento de investigación. El punto de partida no es el ítem, ni la escala, ni siquiera la variable: es el problema. Todo instrumento nace de una necesidad de conocimiento que el investigador ha identificado con claridad y ha sabido traducir en una pregunta mensurable.

Este capítulo guía al investigador a través del proceso más exigente y menos reconocido del diseño metodológico: la arquitectura conceptual del instrumento. Se trata del conjunto de decisiones teóricas, definitivas e irreversibles que determinan qué se va a medir, cómo se va a representar lo que se mide y con qué grado de profundidad y cobertura se abordará el fenómeno de interés. Un instrumento cuya arquitectura conceptual es sólida tiene posibilidades reales de ser válido, independientemente de los ajustes técnicos que vengan después. Un instrumento con arquitectura conceptual deficiente será inválido por diseño, sin importar cuánto se refine estadísticamente.

El capítulo presenta el Protocolo DCOI (Del Constructo al Instrumento), una propuesta metodológica original que integra de forma sistemática todas las etapas de la arquitectura conceptual en un proceso replicable, auditable y compatible con los estándares internacionales de construcción de instrumentos.

2.1 El problema de investigación como punto de partida

2.1.1 ¿Qué es un problema de medición?

En la tradición metodológica, el problema de investigación suele definirse como una brecha entre lo que se conoce y lo que se desea conocer sobre un fenómeno. Sin embargo, cuando el propósito del estudio implica la construcción de un instrumento, el problema adquiere una especificidad adicional que los textos de metodología raramente explicitan: se trata no solo de una brecha de conocimiento sustantivo, sino de una brecha de medición. El investigador identifica que no existe un instrumento adecuado para medir un constructo relevante en una población o contexto específico.

Un problema de medición bien formulado responde simultáneamente a tres preguntas:

- ¿Qué constructo necesita ser medido y por qué es relevante medirlo en este contexto? Esta pregunta establece la justificación sustantiva del instrumento.
- ¿Para qué población y en qué condiciones se necesita el instrumento? Esta pregunta delimita el alcance y las condiciones de validez del instrumento.
- ¿Por qué los instrumentos existentes son insuficientes para este propósito? Esta pregunta justifica la necesidad de desarrollar un nuevo instrumento en lugar de adoptar o adaptar uno existente.

Principio fundamental

Construir un nuevo instrumento cuando ya existe uno válido y adecuado para el propósito y la población es un desperdicio de recursos y una potencial fuente de fragmentación innecesaria del campo. La decisión de construir un instrumento nuevo debe estar siempre precedida por una revisión sistemática y honesta de los instrumentos existentes.

2.1.2 De la pregunta de investigación a la necesidad de medir

No toda pregunta de investigación requiere un instrumento nuevo, ni siquiera un instrumento propio. El investigador debe comenzar por clarificar con precisión qué tipo de información necesita y cómo esa información se relaciona con su pregunta de investigación. Este proceso de clarificación implica distinguir entre diferentes tipos de preguntas investigativas:

Tabla 2.1

Tipos de preguntas de investigación y su relación con la necesidad de instrumentación

Tipo de pregunta	Ejemplo	Necesidad de instrumento	Tipo de instrumento típico
Descriptiva	¿Cuál es el nivel de resiliencia comunitaria en el municipio X?	Alta: requiere medición sistemática	Escala o índice estandarizado
Comparativa	¿Difieren los niveles de motivación entre estudiantes de universidades públicas y privadas?	Alta: requiere medición equivalente en ambos grupos	Instrumento con invarianza de medición
Correlacional	¿Se relacionan la calidad del sueño y el rendimiento académico?	Alta: requiere medir ambas variables con precisión	Dos instrumentos validados independientemente

Tipo de pregunta	Ejemplo	Necesidad de instrumento	Tipo de instrumento típico
Explicativa causal	¿En qué medida el liderazgo transformacional predice el compromiso organizacional?	Alta: requiere instrumentos con evidencia de validez criterial	Escalas validadas con evidencia predictiva
Exploratoria cualitativa	¿Cómo experimentan el duelo los trabajadores de primera respuesta?	Media: puede requerir guías semiestructuradas	Guía de entrevista o grupo focal
Evaluativa	¿En qué medida el programa X mejoró las competencias de gestión del riesgo?	Alta: requiere instrumento sensible al cambio	Prueba de pre-post con evidencia de responsividad

Nota. Elaboración propia.

2.1.3 Criterios para determinar si necesito un instrumento nuevo

La decisión de construir un nuevo instrumento debe adoptarse solo cuando se cumple al menos uno de los siguientes criterios, que denominamos los Cinco Criterios de Justificación del Nuevo Instrumento (CJNI):

1. Inexistencia: no existe ningún instrumento que mida el constructo de interés en ningún idioma o contexto.
2. Inadecuación conceptual: los instrumentos existentes se basan en un modelo teórico del constructo diferente o incompatible con el adoptado en la investigación.
3. Inadecuación poblacional: los instrumentos existentes fueron desarrollados para poblaciones con características tan diferentes que su aplicación a la población objetivo carece de justificación.
4. Evidencia psicométrica insuficiente: los instrumentos existentes no presentan evidencias adecuadas de validez y confiabilidad, o estas evidencias fueron obtenidas con metodologías superadas.
5. Condiciones de uso incompatibles: los instrumentos existentes son propietarios, su costo es prohibitivo, sus condiciones de licencia lo impiden, o su formato es incompatible con las condiciones del estudio.

Advertencia metodológica

Un error frecuente y costoso es construir un instrumento nuevo por razones de conveniencia o desconocimiento, sin haber realizado una búsqueda sistemática de instrumentos existentes. Este error no solo desperdicia recursos, sino que también fragmenta innecesariamente el campo, dificultando las comparaciones entre estudios. Antes de iniciar el proceso de construcción, el investigador debe realizar una búsqueda exhaustiva en bases de datos especializadas (PsycTESTS, COSMIN, PROQOLID, HaPI, Mapi Research Trust) y en los principales repositorios de instrumentos de su disciplina.

2.1.4 Evaluación de instrumentos existentes antes de crear uno nuevo

Cuando se identifican instrumentos candidatos en la literatura, es necesario evaluarlos sistemáticamente antes de tomar la decisión de adoptarlos, adaptarlos o descartarlos. Esta evaluación debe considerar al menos cinco dimensiones fundamentales:

Tabla 2.2

Dimensiones para la evaluación sistemática de instrumentos existentes

Dimensión	Criterios de evaluación	Indicadores de adecuación	Fuente de verificación
Conceptual	Coherencia del modelo teórico con el marco adoptado; cobertura de dimensiones	Definición del constructo consistente con teoría actualizada; todas las dimensiones representadas	Artículo de construcción; manuales
Psicométrica	Evidencias de validez y confiabilidad reportadas	AFC con índices de ajuste adecuados; α u $\omega \geq 0.70$; evidencia de validez de constructo	Artículos de validación; COSMIN checklist
Poblacional	Similitud entre la población de desarrollo y la población objetivo	Características demográficas, culturales y educativas comparables	Descripción de la muestra en publicaciones
Práctica	Factibilidad de administración en el contexto del estudio	Tiempo de administración, formato, requerimientos tecnológicos, costo	Manual del instrumento; contacto con autores
Ética y legal	Condiciones de uso, propiedad intelectual, accesibilidad	Licencia de uso clara; accesible para la población; no discriminatorio	Registro de propiedad; licencia disponible

Nota. COSMIN = Consensus-based Standards for the selection of health Measurement INstruments. Elaboración propia.

2.1.5 Herramientas de IA para búsqueda de instrumentos existentes

La búsqueda sistemática de instrumentos existentes se ha transformado significativamente con la aparición de herramientas de inteligencia artificial especializadas en literatura científica. A diferencia de los motores de búsqueda convencionales, estas herramientas permiten realizar búsquedas semánticas —no solo por palabras clave— y obtener síntesis estructuradas de la evidencia disponible:

Tabla 2.3

Herramientas de IA para búsqueda y evaluación de instrumentos en la literatura científica

Herramienta	Tipo	Utilidad para instrumentación	Limitaciones
Elicit	IA para síntesis de literatura	Extrae tablas de instrumentos, muestras y resultados psicométricos de múltiples artículos simultáneamente	Cobertura limitada a Semantic Scholar; puede perder artículos en idiomas no ingleses
Consensus	Buscador científico con IA	Responde preguntas sobre qué instrumentos miden X constructo con síntesis de evidencia	No accede a bases de datos propietarias
ResearchRabbit	Mapeo de literatura	Visualiza redes de citas para identificar instrumentos influyentes y sus derivados	Requiere un artículo semilla inicial de calidad
Semantic Scholar	Motor de búsqueda científico	Filtra por tipo de estudio, área y año; tiene alertas de nuevas publicaciones	No especializado en instrumentos; búsqueda manual requerida
Connected Papers	Mapeo de relaciones	Identifica artículos seminal y derivados de un instrumento en la literatura	Limitado a trabajos con DOI registrado
Claude / GPT-4	LLM conversacional	Síntesis de características de instrumentos conocidos; redacción de criterios de búsqueda; evaluación preliminar de ítems	Puede alucinar referencias; no accede a bases propietarias; conocimiento con fecha de corte

Nota. Elaboración propia. LLM = Modelo de Lenguaje Grande. La efectividad de estas herramientas evoluciona rápidamente; se recomienda verificar sus capacidades actuales antes de su uso.

2.2 El constructo: definición, delimitación y tipología

2.2.1 ¿Qué es un constructo científico?

El término constructo del latín constructum, aquello que ha sido construido, designa en la teoría de la medición una entidad abstracta postulada por el investigador para explicar o describir patrones regularmente observados en el comportamiento, las actitudes, los estados internos o las capacidades de personas, organizaciones o comunidades. A diferencia de las variables observables directamente (como la temperatura corporal, la frecuencia cardíaca o el número de palabras en un texto), los constructos no pueden ser medidos directamente: su existencia se infiere a partir de indicadores observables que se supone reflejan variaciones en su nivel.

Esta distinción entre el constructo y sus indicadores es fundamental para entender por qué la construcción de instrumentos es una tarea teórica antes que técnica. Cuando un investigador afirma haber medido la 'autoeficacia académica', 'la resiliencia comunitaria' o 'el capital social' de un grupo de personas, en rigor está diciendo que ha construido un instrumento que, según sus supuestos teóricos, captura aspectos observables que se corresponden con variaciones en una entidad teórica inobservable que denomina de esa manera. La legitimidad científica de esa afirmación depende enteramente de la solidez teórica del modelo del constructo y de la evidencia empírica que sustenta la correspondencia entre los indicadores y la entidad teórica postulada.

Reflexión epistemológica

La pregunta '¿existe realmente la autoestima?' no es una pregunta que la psicometría pueda responder por sí sola. Es una pregunta filosófica sobre la ontología de los constructos psicológicos. Lo que sí puede determinar la psicometría es si las puntuaciones de un instrumento diseñado para medir la autoestima se comportan de la manera que el modelo teórico predice: si tienen la estructura factorial esperada, si predicen los comportamientos que el modelo sugiere que deberían predecir, y si varían de la manera que el modelo predice en respuesta a intervenciones o condiciones específicas.

2.2.2 Constructos simples versus constructos complejos

Una primera distinción de enorme importancia práctica es la que separa los constructos en función de su complejidad estructural:

Constructos simples o unidimensionales:

Son aquellos que pueden ser representados adecuadamente por una única dimensión continua. Todos los ítems del instrumento deberían reflejar el mismo atributo subyacente, y la puntuación total es una representación coherente del nivel del constructo. El modelo teórico predice que existe un único factor latente que explica las correlaciones entre los ítems. La soledad social, la ansiedad ante los exámenes o la satisfacción con el sueldo pueden ser ejemplos de constructos relativamente simples en determinadas poblaciones.

Constructos complejos o multidimensionales:

Son aquellos que comprenden múltiples componentes o dimensiones distinguibles conceptualmente, aunque relacionadas entre sí. La calidad de vida, la inteligencia emocional, la resiliencia, el clima organizacional o la vulnerabilidad social ante desastres son ejemplos paradigmáticos. En estos casos, el instrumento debe incluir subescalas o dimensiones que reflejen los componentes del constructo, y la relación entre las dimensiones debe ser especificada explícitamente en el modelo teórico.

La distinción entre constructos unidimensionales y multidimensionales no siempre es evidente a priori: puede ser una propiedad del constructo mismo, del nivel de abstracción con el que se lo define, o de la población en la que se lo estudia. Por eso es esencial que el modelo de dimensionalidad sea siempre una hipótesis teórica explícita que se pone a prueba empíricamente, no una asunción implícita.

2.2.3 Constructos reflectivos versus formativos

Una distinción conceptual que ha ganado enorme relevancia en la literatura psicométrica y de investigación social en las últimas décadas es la que separa los constructos en función de la relación de causalidad entre el constructo latente y sus indicadores:

Tabla 2.4

Diferencias fundamentales entre modelos de medición reflectivos y formativos

Dimensión	Modelo reflectivo	Modelo formativo
Dirección de causalidad	El constructo causa los indicadores: variaciones en el constructo producen variaciones en los ítems	Los indicadores causan el constructo: variaciones en los ítems producen variaciones en el constructo
Intercambiabilidad de ítems	Los ítems son intercambiables: eliminar uno no cambia el constructo	Los ítems no son intercambiables: cada uno aporta contenido único al constructo
Correlación entre ítems	Se espera alta correlación entre ítems de la misma dimensión	Los ítems pueden no correlacionar entre sí sin que eso sea un problema
Consecuencias de eliminar un ítem	Reducción de la confiabilidad pero no del constructo	Posible cambio en el dominio del constructo medido
Análisis apropiado	AFC, Alpha, Omega, TRI	PLS-SEM, índices compuestos con pesos diferenciados
Ejemplos	Ansiedad, depresión, autoestima, satisfacción laboral	Estatus socioeconómico, índice de desarrollo humano, vulnerabilidad social ante desastres
Error de especificación frecuente	Tratar indicadores formativos como reflectivos (muy común)	Tratar indicadores reflectivos como formativos (menos frecuente)

Nota. AFC = Análisis Factorial Confirmatorio; PLS-SEM = Partial Least Squares Structural Equation Modeling. Adaptado de Diamantopoulos et al. (2008) y Coltman et al. (2008).

La confusión entre modelos reflectivos y formativos es uno de los errores conceptuales más graves y frecuentes en la construcción de instrumentos. Cuando un investigador trata indicadores formativos como reflectivos (lo que ocurre rutinariamente con índices compuestos de vulnerabilidad, calidad de vida multidimensional o desarrollo organizacional), las pruebas de confiabilidad interna como el Alpha de Cronbach carecen de sentido y los índices de ajuste

del AFC serán sistemáticamente pobres, no porque el instrumento sea malo, sino porque el modelo de medición es incorrecto.

2.2.4 Definición conceptual del constructo

La definición conceptual del constructo es la descripción precisa de qué es el fenómeno que se pretende medir, expresada en términos del marco teórico que sustenta el instrumento. No es una definición operacional (qué puntaje o qué ítem mide el constructo), sino una delimitación teórica que establece los bordes del constructo: qué incluye y, crucialmente, qué excluye.

Una definición conceptual de calidad debe cumplir los siguientes criterios:

- Claridad: está formulada en términos inequívocos que permiten distinguir el constructo de otros constructos relacionados.
- Especificidad: delimita el ámbito de aplicación del constructo (¿se aplica a individuos, grupos, organizaciones, comunidades?).
- Anclaje teórico: se vincula explícitamente con uno o más marcos teóricos establecidos, con citas precisas de los autores que sustentan la definición.
- Parsimonia: captura la esencia del constructo sin ser innecesariamente compleja.
- Productividad: sugiere de manera natural cuáles serán las dimensiones e indicadores del constructo, facilitando la operacionalización.

Tabla 2.5

Ejemplos de definiciones conceptuales de constructos con análisis de calidad

Constructo	Definición conceptual de calidad	Análisis
Autoeficacia académica	Creencias del individuo sobre sus propias capacidades para organizar y ejecutar los cursos de acción necesarios para lograr determinados niveles de rendimiento académico (Bandura, 1997, p. 3)	Clara, específica, anclada en la teoría social cognitiva de Bandura, productiva (sugiere dimensiones por áreas de conocimiento o tipos de tareas)
Resiliencia comunitaria ante desastres	Capacidad colectiva de una comunidad para anticipar, absorber, adaptarse y transformarse ante perturbaciones relacionadas con amenazas naturales o antrópicas, preservando y mejorando sus estructuras y funciones esenciales (Norris et al., 2008)	Multidimensional, sugiere dimensiones de anticipación, absorción, adaptación y transformación; aplicable a nivel comunitario

Constructo	Definición conceptual de calidad	Análisis
Satisfacción laboral	Estado emocional positivo o placentero resultante de la evaluación del propio trabajo o de las experiencias laborales (Locke, 1976, p. 1300)	Anclada en la teoría de Locke, distingue componentes afectivos y cognitivos, parsimoniosa; puede expandirse con dimensiones facetadas

Nota. Elaboración propia. Las definiciones incluyen las referencias originales para ilustrar el anclaje teórico requerido.

2.2.5 Definición operacional del constructo

La definición operacional traduce la definición conceptual al lenguaje de la medición: especifica cómo se va a medir el constructo, qué procedimientos se utilizarán y cómo se obtendrá la puntuación final. Es el puente obligatorio entre el mundo de la teoría y el mundo de los datos.

Una definición operacional completa debe especificar:

- El tipo de instrumento que se utilizará (escala, test, guía de entrevista, registro de observación).
- El número de ítems y dimensiones.
- El formato de respuesta y la escala de medición (tipo Likert de X puntos, dicotómica, de frecuencia, etc.).
- El procedimiento de puntuación (suma directa, promedio, puntuación ponderada, puntuación inversa de ítems negativos).
- El rango de puntuaciones posibles y su interpretación.
- Los puntos de corte, si corresponde.

Distinción clave

La definición conceptual responde a la pregunta '¿qué es?' y la definición operacional responde a '¿cómo se mide?'. Ambas son indispensables y deben ser coherentes entre sí. La coherencia entre definición conceptual y operacional es, en esencia, la validez de constructo: un instrumento es válido cuando su definición operacional captura fielmente lo que su definición conceptual establece que es el constructo.

2.2.6 Errores frecuentes en la definición del constructo

La experiencia en la revisión de instrumentos publicados revela un conjunto de errores que se repiten con alarmante frecuencia en la definición del constructo. Conocerlos con anticipación permite evitarlos sistemáticamente:

- Definición circular: el constructo se define usando el mismo término que se pretende definir. Ejemplo: 'La motivación es el estado de estar motivado para realizar una acción'. Este tipo de definición no aporta ninguna información teórica sobre la naturaleza del constructo.
- Definición por enumeración de conductas: en lugar de definir el constructo, se enumeran algunas de sus manifestaciones conductuales. Ejemplo: 'El liderazgo transformacional es cuando un líder inspira, motiva y da visión a sus seguidores'. Esta definición confunde el constructo con sus indicadores.
- Definición demasiado amplia: el constructo se define de manera tan general que abarca prácticamente cualquier fenómeno. Ejemplo: 'El bienestar es el estado de sentirse bien en todos los aspectos de la vida'. Una definición así no delimita lo que el instrumento debe y no debe medir.
- Definición demasiado estrecha: el constructo se define de manera tan restringida que excluye aspectos relevantes. Esto genera instrumentos con problemas de sub-representación del constructo.
- Ausencia de anclaje teórico: la definición no cita ningún marco teórico establecido. Esto impide evaluar si la definición es coherente con el estado del conocimiento y dificulta la comparación con otros instrumentos del mismo constructo.
- Mezcla de definición conceptual y operacional: la definición incluye referencias al instrumento o a los procedimientos de medición en lugar de describir el fenómeno en sí mismo.

2.3 Revisión sistemática del constructo

2.3.1 Por qué la revisión teórica es insustituible

La revisión de la literatura sobre el constructo no es un paso burocrático destinado a llenar el marco teórico de la tesis doctoral o el artículo científico. Es una condición epistemológica sine qua non para la validez del instrumento. Sin una revisión sistemática y profunda de la literatura, el investigador no puede: (a) justificar la definición conceptual del constructo que adoptó; (b)

identificar las dimensiones que la evidencia teórica y empírica acumulada reconoce como componentes del constructo; (c) evaluar críticamente los instrumentos existentes; (d) anticipar las hipótesis sobre la estructura factorial que el AFC deberá confirmar.

Paradójicamente, el investigador que construye un instrumento sin una revisión exhaustiva de la literatura no está siendo más original: está siendo menos riguroso. La originalidad científica consiste en construir sobre el conocimiento acumulado, no en ignorarlo.

2.3.2 Protocolo de revisión sistemática para instrumentación

La revisión de literatura para el desarrollo de un instrumento tiene características específicas que la diferencian de una revisión sistemática convencional. No se busca solo sintetizar hallazgos empíricos, sino mapear el espacio conceptual del constructo: identificar las definiciones, modelos teóricos, dimensiones propuestas, instrumentos desarrollados y evidencias de validez disponibles. El siguiente protocolo es una guía paso a paso para este proceso:

Tabla 2.6

Protocolo de revisión sistemática de literatura para el desarrollo de instrumentos (PRSLI)

Fase	Actividad	Descripción	Productos esperados
1	Definición del problema de búsqueda	Formular con precisión qué se quiere saber sobre el constructo: definiciones, modelos, instrumentos, evidencias	Pregunta de búsqueda estructurada (PICO o similar)
2	Identificación de términos de búsqueda	Listar sinónimos, traducciones y términos relacionados; construir ecuación de búsqueda booleana	Ecuación de búsqueda documentada y reproducible
3	Selección de bases de datos	PsycINFO, Web of Science, Scopus, MEDLINE, ERIC, SciELO; más repositorios especializados del área	Lista de bases de datos consultadas con fechas
4	Aplicación de criterios de inclusión/exclusión	Idioma, período, tipo de publicación, tipo de estudio, pertinencia al constructo	Diagrama PRISMA de flujo de selección
5	Extracción de datos	Autor, año, definición del constructo, modelo teórico, dimensiones propuestas,	Matriz de extracción de datos estandarizada

Fase	Actividad	Descripción	Productos esperados
		instrumento utilizado, evidencias de validez	
6	Síntesis y análisis	Identificar convergencias, divergencias y vacíos en la conceptualización del constructo	Mapa conceptual del constructo; identificación de dimensiones consensuadas y debatidas
7	Documentación	Registrar todo el proceso para garantizar la trazabilidad y reproducibilidad	Informe de revisión sistemática (puede publicarse como artículo independiente)

Nota. PICO = Participantes, Intervención, Comparación, Resultado (adaptado para revisiones de constructos). PRISMA = Preferred Reporting Items for Systematic reviews and Meta-Analyses.

2.3.3 Herramientas de IA para la revisión sistemática

Las herramientas de IA han transformado radicalmente la eficiencia de la revisión de literatura científica, aunque sin suprimir la necesidad del juicio experto. A continuación, se describe cómo integrar estas herramientas en el protocolo PRSLI de manera metodológicamente responsable:

Elicit (elicit.org):

Permite subir una pregunta de investigación en lenguaje natural y obtener una tabla estructurada con los hallazgos de decenas de artículos relevantes. Para la revisión de constructos, es especialmente útil para extraer automáticamente las definiciones del constructo utilizadas en cada artículo, las dimensiones propuestas y los instrumentos citados. Sin embargo, su cobertura se limita a los artículos indexados en Semantic Scholar y puede perder publicaciones en español u otros idiomas no anglófonos.

Consensus (consensus.app):

Responde preguntas sobre el estado del conocimiento científico con síntesis basadas en múltiples artículos. Útil para preguntas como '¿cuántas dimensiones tiene el constructo X según la literatura?' o '¿qué instrumentos se usan para medir Y en el contexto Z?'. Los resultados deben verificarse accediendo a los artículos originales.

ResearchRabbit (researchrabbit.ai):

A partir de un artículo seminal construye un mapa visual de artículos relacionados (anteriores y posteriores), permitiendo identificar la genealogía teórica del constructo y los principales instrumentos derivados o competidores.

Advertencia sobre el uso de IA en la revisión de literatura

Los LLMs como Claude o GPT-4 pueden sintetizar información sobre constructos e instrumentos bien conocidos con aparente precisión, pero tienen dos limitaciones críticas: (1) su conocimiento tiene una fecha de corte y no acceden a publicaciones recientes; (2) pueden 'alucinar' referencias bibliográficas que suenan plausibles pero no existen. Nunca cite una referencia obtenida de un LLM sin verificar su existencia en la base de datos original. Use los LLMs para generar ideas y estructuras de búsqueda, no como fuentes primarias de información científica.

2.3.4 Matriz de análisis de antecedentes teóricos

Una herramienta práctica fundamental para organizar y analizar la información obtenida en la revisión de literatura es la Matriz de Análisis de Antecedentes Teóricos (MAAT). Esta matriz permite visualizar de forma comparada cómo diferentes autores han conceptualizado el constructo, qué dimensiones han propuesto y qué evidencias han aportado:

Tabla 2.7

Ejemplo de Matriz de Análisis de Antecedentes Teóricos (MAAT) para el constructo “Resiliencia Comunitaria”

Autor(es) / Año	Marco teórico	Modelo	Dimensiones propuestas	Instrumento	Limitaciones
Norris et al. (2008)	Ecología comunitaria	Formativo	Capital económico, capital social, información y comunicación, competencia comunitaria	Revisión conceptual sin instrumento propio	No operacionaliza el constructo directamente
Cutter et al. (2014)	Sistemas socio-ecológicos	Formativo	Infraestructura, ecología, economía, institucional, social	BRIC (Baseline Resilience Indicators for Communities)	Diseñado para nivel de condado en EE.UU.; baja transferibilidad
Sherrieb et al. (2010)	Teoría del capital	Mixto	Capital económico y capital social como proxies de resiliencia	Indicadores secundarios de condado	No captura percepción comunitaria; datos agregados

Autor(es) / Año	Marco teórico	Modelo	Dimensiones propuestas	Instrumento	Limitaciones
Pfefferbaum et al. (2013)	Competencia comunitaria	Reflectivo	Conexión, capacidad comunitaria, compromiso, comunicación	CRS (Community Resilience Scale)	Desarrollada en contexto anglosajón; 21 ítems
Ostadtaghizadeh et al. (2015)	Gestión de emergencias	Mixto	Social, económica, institucional, física, capital humano	CDRI adaptado	Pocos estudios de validación en países de ingreso medio

Nota. BRIC = Baseline Resilience Indicators for Communities; CRS = Community Resilience Scale; CDRI = Community Disaster Resilience Index. Elaboración propia con fines ilustrativos.

2.3.5 Identificación de vacíos en la literatura

El producto más valioso de la revisión de literatura no es la síntesis de lo que se sabe, sino la identificación de lo que no se sabe, es decir, los vacíos de conocimiento que justifican el desarrollo del nuevo instrumento. Para la construcción de instrumentos, los vacíos más relevantes son de cuatro tipos:

- Vacíos conceptuales: el constructo carece de una definición teórica consensuada, o las definiciones disponibles son contradictorias, incompletas o no se aplican adecuadamente al contexto de la investigación.
- Vacíos psicométricos: los instrumentos existentes carecen de evidencias adecuadas de validez o confiabilidad, o estas evidencias son contradictorias entre estudios.
- Vacíos poblacionales: el constructo existe solo en instrumentos desarrollados para poblaciones muy diferentes de la de interés (culturales, idiomáticas, sociodemográficas).
- Vacíos de dominio: los instrumentos existentes no cubren algunas dimensiones del constructo que son teóricamente relevantes o que tienen especial importancia en el contexto del estudio.

2.4 Dimensionalidad del constructo

2.4.1 Constructos unidimensionales versus multidimensionales

La determinación del número y naturaleza de las dimensiones del constructo es quizás la decisión teórica más consecuente en todo el proceso de construcción del instrumento. Una especificación incorrecta de la dimensionalidad genera instrumentos que mezclan en la misma

puntuación total constructos conceptualmente distintos o que segmentan artificialmente en subescalas un constructo que es genuinamente unitario.

La dimensionalidad no es solo una cuestión empírica que el análisis factorial resolverá definitivamente: es, en primer lugar, una cuestión teórica. Los datos empíricos pueden apoyar o contradecir el modelo de dimensionalidad propuesto, pero no pueden generarlo en ausencia de una teoría que oriente la interpretación de los factores. Un análisis factorial exploratorio sin hipótesis teóricas previas sobre la dimensionalidad produce factores estadísticos que pueden o no corresponder a constructos teóricos significativos.

Tabla 2.8

Tipos de modelos de dimensionalidad y sus implicaciones para el instrumento

Modelo	Descripción	Implicaciones para el instrumento	Análisis apropiado
Unidimensional	Un único factor latente explica todas las correlaciones entre ítems	Puntuación total coherente; todos los ítems deben correlacionar positivamente entre sí	AFC con un factor; Alpha/Omega; TRI
Multidimensional correlacionado	Múltiples factores relacionados entre sí; también hay un factor general	Puntuaciones por subescalas y puntuación total; ambas tienen significado	AFC multifactorial con correlaciones; modelo bifactor
Multidimensional jerárquico	Factor de orden superior que explica las correlaciones entre factores de primer orden	Puntuación total refleja factor general; subescalas reflejan factores específicos	AFC de segundo orden; modelo bifactor de Schmid-Leiman
Multidimensional independiente	Múltiples factores sin relación sustantiva entre sí	Solo tienen sentido las puntuaciones de subescalas; la puntuación total carece de significado	AFC con factores ortogonales; análisis por separado
Formativo compuesto	Los indicadores causan el constructo; no se espera covarianza entre ellos	Pesos diferenciales; no se elimina un indicador por baja correlación	PLS-SEM; índices compuestos ponderados

Nota. AFC = Análisis Factorial Confirmatorio; TRI = Teoría de Respuesta al Ítem; PLS-SEM = Partial Least Squares Structural Equation Modeling. Elaboración propia.

2.4.2 Cómo identificar las dimensiones de un constructo

La identificación de las dimensiones del constructo es un proceso teórico-empírico que combina el análisis crítico de la literatura con el juicio del investigador y, en una etapa posterior, la evidencia empírica de los datos. El proceso recomendado incluye los siguientes pasos:

1. Mapa de dimensiones en la literatura: usando la MAAT (Tabla 2.7), identificar todas las dimensiones propuestas por diferentes autores y clasificarlas por frecuencia de aparición y respaldo teórico.
2. Análisis de convergencia y divergencia: determinar qué dimensiones son consistentes entre autores y cuáles son objeto de debate. Las dimensiones con alto consenso son candidatos seguros para el instrumento; las debatidas requieren una justificación explícita de la posición adoptada.
3. Análisis de relevancia contextual: evaluar cuáles de las dimensiones identificadas en la literatura son relevantes para la población y el contexto específico del estudio. No todas las dimensiones teóricas tienen la misma relevancia en todos los contextos.
4. Definición de cada dimensión: escribir una definición conceptual de cada dimensión, análoga a la definición del constructo global, que especifique qué incluye y qué excluye.
5. Construcción del árbol de dimensiones: representar gráficamente la relación jerárquica entre el constructo, las dimensiones y los indicadores.

2.4.3 El árbol de dimensiones: herramienta de visualización

El árbol de dimensiones también denominado mapa conceptual del constructo o diagrama de estructura del instrumento es una representación gráfica que muestra la relación jerárquica entre el constructo, sus dimensiones y sus indicadores. Es una herramienta de planificación fundamental porque permite al investigador visualizar de un solo vistazo si la arquitectura conceptual es coherente, equilibrada y completa.

Un árbol de dimensiones bien construido tiene las siguientes características:

- El constructo aparece en el nivel superior como nodo raíz.
- Las dimensiones aparecen en el segundo nivel, conectadas al constructo mediante líneas que representan la relación parte-todo.
- Los indicadores aparecen en el tercer nivel, conectados a la dimensión correspondiente.

- Los ítems del instrumento (si ya han sido formulados) pueden aparecer en el cuarto nivel, conectados al indicador que pretenden capturar.
- El número de indicadores por dimensión es aproximadamente equilibrado, a menos que la teoría justifique desequilibrios específicos.

Figura 2.1

Árbol de dimensiones del constructo “Vulnerabilidad Social ante Desastres”



Nota. Representación esquemática del árbol de dimensiones. En la práctica, esta figura se elabora con software de diagramación (Lucidchart, Miro, draw.io). D = Dimensión. El modelo formativo implica que las tres dimensiones contribuyen a definir la vulnerabilidad, no que la vulnerabilidad las causa.

2.5 Operacionalización de variables

2.5.1 El proceso de operacionalización: de la teoría al dato

La operacionalización es el proceso mediante el cual las definiciones abstractas de los constructos y sus dimensiones se traducen en procedimientos concretos y observables que permiten su medición. Es el eslabón de la cadena científica que conecta la teoría con lo empírico, y su calidad determina en gran medida la validez del instrumento resultante.

El proceso de operacionalización sigue una lógica descendente y puede representarse como una cascada de decisiones teóricas:

Constructo → Dimensiones → Indicadores → Ítems → Opciones de respuesta
(Cascada de operacionalización)

Cada nivel de esta cascada implica un tipo diferente de decisión metodológica:

- **Constructo → Dimensiones:** decisión teórica sobre la estructura interna del constructo. Requiere revisión y síntesis de la literatura.
- **Dimensiones → Indicadores:** decisión teórico-empírica sobre qué aspectos observables reflejan cada dimensión. Requiere creatividad conceptual y conocimiento del contexto.
- **Indicadores → Ítems:** decisión técnica sobre cómo formular cada indicador como una pregunta o afirmación comprensible para la población objetivo.
- **Ítems → Opciones de respuesta:** decisión metodológica sobre el formato de respuesta más adecuado para cada ítem y el nivel de medición resultante.

2.5.2 De constructo a dimensiones

Esta primera etapa de operacionalización es la más importante y la más frecuentemente subestimada. La calidad de la decisión sobre las dimensiones determina la validez de contenido del instrumento: si las dimensiones no cubren adecuadamente el dominio del constructo, ningún refinamiento técnico posterior podrá compensar esa deficiencia.

Los criterios técnicos para evaluar la calidad del sistema de dimensiones propuesto son:

- **Exhaustividad:** el conjunto de dimensiones cubre todos los aspectos importantes del constructo tal como lo define la teoría adoptada.
- **Exclusividad:** las dimensiones son conceptualmente distinguibles entre sí; no hay solapamiento sustancial entre ellas.
- **Coherencia:** cada dimensión es coherente con la definición conceptual del constructo y representa un aspecto genuinamente relevante del mismo.
- **Equilibrio:** las dimensiones tienen importancia teórica comparable, a menos que el modelo teórico justifique una ponderación diferencial.
- **Parsimonia:** el número de dimensiones es el mínimo necesario para representar adecuadamente el constructo; no se multiplican dimensiones innecesariamente.

2.5.3 De dimensiones a indicadores

Los indicadores son las manifestaciones observables de cada dimensión. Son el puente entre la abstracción teórica y la realidad empírica. La selección de buenos indicadores requiere un conocimiento profundo tanto del constructo como de la población que va a ser evaluada.

Existen tres estrategias principales para identificar indicadores para cada dimensión:

- Estrategia deductiva: los indicadores se derivan lógicamente de la definición conceptual de la dimensión. Se parte del marco teórico y se pregunta: ¿qué comportamientos, actitudes, percepciones o características deberían observarse en una persona o comunidad que tiene un nivel alto (versus bajo) de esta dimensión?
- Estrategia inductiva: los indicadores se identifican a partir de la observación empírica o de entrevistas con miembros de la población objetivo. Se parte de los datos para llegar a los indicadores. Esta estrategia es especialmente valiosa para constructos poco estudiados o en contextos culturalmente específicos.
- Estrategia mixta: combina la derivación teórica con la verificación empírica. Los indicadores propuestos deductivamente se contrastan con la experiencia vivida de la población mediante grupos focales o entrevistas cognitivas preliminares.

Un indicador de calidad debe cumplir los siguientes criterios:

Tabla 2.9

Criterios de calidad para indicadores en instrumentos de investigación

Criterio	Descripción	Ejemplo de evaluación
Relevancia teórica	El indicador refleja directamente la dimensión que pretende representar	¿Es este indicador parte de la definición conceptual de la dimensión?
Observabilidad	El indicador puede ser registrado de manera confiable por el instrumento	¿Puede preguntarse directamente? ¿Puede observarse? ¿Puede extraerse de registros?
Sensibilidad	El indicador varía de manera detectable con variaciones en la dimensión	¿Diferenciaría entre personas con niveles muy diferentes de la dimensión?
Especificidad	El indicador no está contaminado por otras dimensiones o constructos ajenos	¿Podría este indicador ser causado por algo completamente diferente al constructo?
Factibilidad	El indicador puede ser medido en las condiciones reales del estudio	¿Tiene la población objetivo la información o experiencia necesaria para responder?
Aceptabilidad	El indicador no genera rechazo, incomodidad o desconfianza en la población objetivo	¿Es culturalmente apropiado preguntar sobre esto a esta población?

Nota. Elaboración propia.

2.5.4 De indicadores a ítems observables

Un mismo indicador puede traducirse en múltiples ítems posibles, y la elección entre ellos depende del tipo de instrumento, el formato de respuesta y las características de la población. Esta etapa de la operacionalización requiere tanto habilidad técnica como sensibilidad lingüística y cultural.

Por ejemplo, el indicador '*frecuencia de comunicación con vecinos en situaciones de emergencia*' del constructo '*cohesión comunitaria para la gestión del riesgo*' podría traducirse en ítems tan diferentes como:

- *Ítem de frecuencia*: 'En los últimos 12 meses, ¿con qué frecuencia ha conversado con sus vecinos sobre qué hacer en caso de un desastre natural?'
- *Ítem de acuerdo*: 'En mi barrio, los vecinos nos comunicamos activamente sobre cómo prepararnos para emergencias.'
- *Ítem de observación*: '¿Cuántos de sus vecinos más cercanos conoce por nombre?'
- *Ítem conductual*: '¿Ha participado alguna vez en una reunión con sus vecinos para planificar cómo responder a una emergencia?'

Cada formulación captura un aspecto ligeramente diferente del indicador y tiene diferentes propiedades psicométricas, diferentes sesgos potenciales y diferentes niveles de accesibilidad para distintas poblaciones. El investigador debe seleccionar la formulación que mejor se ajusta al propósito del instrumento y a las características de la población.

2.5.5 Errores más comunes en la operacionalización

La práctica de revisión de instrumentos publicados revela patrones de error en la operacionalización que se repiten con notable consistencia a través de disciplinas y contextos. Los más graves son:

Tabla 2.10

Errores frecuentes en la operacionalización y sus consecuencias para la validez del instrumento

Error	Descripción	Consecuencia para la validez
Sub-representación del constructo	El instrumento solo cubre algunas de las dimensiones del constructo; otras quedan sin representar	Validez de contenido deficiente; las conclusiones no se pueden generalizar al constructo completo
Constructo-irrelevante	El instrumento mide aspectos que no forman parte del constructo	Validez de constructo contaminada; las diferencias en

Error	Descripción	Consecuencia para la validez
	(habilidades lectoras, fluidez verbal, familiaridad cultural)	puntuaciones reflejan parcialmente otras variables
Confusión causa-efecto en indicadores	Se incluyen como indicadores variables que son causas o consecuencias del constructo, no manifestaciones del mismo	Error de especificación del modelo de medición; interpretación incorrecta de los resultados
Desequilibrio dimensional	Algunas dimensiones tienen muchos más ítems que otras sin justificación teórica	Las dimensiones sobrerrepresentadas tienen más influencia en la puntuación total; las subdimensionadas son menos confiables
Ítems redundantes	Múltiples ítems miden exactamente lo mismo con ligeras variaciones de redacción	Inflación artificial de la confiabilidad; el instrumento es más largo de lo necesario sin ganar validez
Operacionalización cultura-específica	Los indicadores asumen experiencias, normas o contextos culturales específicos que no son universalmente aplicables	El instrumento discrimina injustamente a personas de culturas diferentes; validez de consecuencias comprometida

Nota. Elaboración propia.

2.6 La tabla de especificaciones o blueprint del instrumento

2.6.1 ¿Qué es la tabla de especificaciones y por qué es indispensable?

La tabla de especificaciones denominada también *blueprint*, matriz de diseño o tabla de doble entrada es el documento de planificación que traduce la arquitectura conceptual del instrumento en decisiones operativas precisas sobre el número y distribución de ítems. Es el equivalente metodológico de los planos arquitectónicos de un edificio: sin ella, el proceso de redacción de ítems carece de dirección y es imposible verificar si el instrumento resultante es coherente con el modelo teórico.

La tabla de especificaciones cumple cinco funciones fundamentales en el desarrollo del instrumento:

1. Garantizar la representatividad: asegura que todas las dimensiones e indicadores relevantes del constructo están representados en el instrumento en la proporción adecuada.
2. Orientar la redacción de ítems: proporciona a los redactores de ítems un mapa preciso de cuántos ítems deben escribir para cada dimensión e indicador y qué tipo de respuesta deben elicitar.

3. Facilitar la evaluación de expertos: los jueces que evalúan la validez de contenido del instrumento pueden usar la tabla para verificar si cada ítem está correctamente asignado a la dimensión e indicador que pretende medir.
4. Documentar el proceso de construcción: junto con la definición del constructo y el árbol de dimensiones, forma parte del expediente técnico del instrumento que documenta las decisiones de diseño.
5. Permitir la revisión y actualización: cuando es necesario revisar o actualizar el instrumento, la tabla de especificaciones permite identificar qué secciones requieren modificación sin afectar la estructura general.

2.6.2 Componentes de la tabla de especificaciones

Una tabla de especificaciones completa para un instrumento de investigación debe incluir los siguientes componentes:

Tabla 2.11

Componentes de una tabla de especificaciones completa para instrumento de investigación

Componente	Descripción	Ejemplo
Dimensión / Subescala	Nombre y código de cada dimensión del constructo	D1: Exposición al riesgo; D2: Sensibilidad social; D3: Capacidad adaptativa
Definición de la dimensión	Descripción conceptual breve de qué mide cada dimensión	D1 mide el grado en que la comunidad está situada en zonas de amenaza y su nivel de exposición directa
Indicadores por dimensión	Lista de indicadores asignados a cada dimensión	D1-I1: Ubicación geográfica; D1-I2: Tipo de vivienda; D1-I3: Densidad poblacional
Número de ítems por indicador	Cantidad de ítems que medirán cada indicador	D1-I1: 2 ítems; D1-I2: 3 ítems; D1-I3: 2 ítems
Tipo de ítem	Formato de respuesta de cada ítem	Likert 5 puntos; dicotómico; abierto; escala de frecuencia
Peso porcentual	Porcentaje del instrumento total asignado a cada dimensión	D1: 30%; D2: 35%; D3: 35%
Número total de ítems	Suma de todos los ítems del instrumento	Ejemplo: 40 ítems totales

Componente	Descripción	Ejemplo
Criterios de puntuación	Cómo se calcula la puntuación de cada dimensión y la puntuación total	Suma directa para D2 y D3; ponderación diferencial para D1 por tipo de amenaza

Nota. Elaboración propia.

2.6.3 Ponderación de dimensiones e indicadores

Una de las decisiones más importantes en el diseño de la tabla de especificaciones es la ponderación de las dimensiones: ¿deben todas las dimensiones contribuir igualmente a la puntuación total, o algunas deben tener mayor peso que otras?

En la mayoría de los instrumentos psicológicos y educativos, la ponderación implícita de las dimensiones se hace a través del número de ítems: una dimensión con más ítems contribuye más a la varianza total y, por lo tanto, tiene más influencia en la puntuación global. Esto es aceptable cuando todas las dimensiones tienen igual importancia teórica, pero es una fuente de sesgo inadvertido cuando algunas dimensiones son más fáciles de operacionalizar en ítems que otras y terminan sobrerrepresentadas.

Para instrumentos de índices compuestos especialmente en las áreas de ciencias sociales, gestión del riesgo y desarrollo, se recomienda una ponderación explícita y justificada de las dimensiones, que puede obtenerse a través de:

- Consenso teórico: los pesos se derivan de la literatura o de modelos teóricos establecidos (por ejemplo, el Marco de Sendai asigna pesos específicos a diferentes componentes del riesgo de desastres).
- Juicio de expertos: un panel de expertos asigna pesos mediante métodos como el Proceso Analítico Jerárquico (AHP) o la técnica Delphi.
- Análisis empírico: los pesos se derivan de análisis estadísticos como la regresión múltiple o el AFC, utilizando la magnitud de las cargas factoriales como indicador del peso relativo de cada dimensión.

$$\text{Puntuación ponderada} = \Sigma(w_i \times D_i) \quad (\text{Ecuación 2.1})$$

Donde w_i es el peso asignado a la dimensión i (con $\Sigma w_i = 1.00$), y D_i es la puntuación estandarizada de la dimensión i . El uso de puntuaciones estandarizadas (puntuaciones Z) antes de aplicar los pesos garantiza que las diferencias en los rangos de escala de las diferentes dimensiones no introduzcan sesgos inadvertidos en la puntuación compuesta.

2.6.4 Número de ítems por dimensión: criterios técnicos

Una pregunta práctica de enorme importancia es: ¿cuántos ítems necesita cada dimensión del instrumento? No existe una respuesta universal, pero la literatura psicométrica ofrece guías claras:

Tabla 2.12

Criterios técnicos para determinar el número de ítems por dimensión en un instrumento de investigación

Criterio	Recomendación	Justificación
Mínimo absoluto por dimensión	3 ítems	Con menos de 3 ítems no es posible un AFC para esa dimensión (problema de identificación del modelo)
Mínimo recomendado	4-5 ítems por dimensión	Proporciona estimaciones estables de la confiabilidad y permite detectar ítems problemáticos durante el pilotaje
Óptimo para investigación básica	5-8 ítems por dimensión	Balance entre cobertura adecuada del dominio y eficiencia en la administración
Para decisiones de alto impacto	8-12 ítems por dimensión	Mayor precisión de medición requerida cuando las decisiones afectan directamente a personas
Máximo recomendado	15 ítems por dimensión	Más de 15 ítems por dimensión aumenta la fatiga del respondiente sin ganancias sustanciales en precisión
Factor de sobreproducción	Generar 150-200% de los ítems necesarios	Se generan más ítems de los que se necesitarán para permitir eliminar los problemáticos durante el pilotaje

Nota. AFC = Análisis Factorial Confirmatorio. Basado en Hinkin (1998), Hair et al. (2019) y Kline (2016).

2.6.5 Plantilla de tabla de especificaciones

A continuación se presenta un ejemplo completo de tabla de especificaciones para un instrumento de medición de la 'Percepción de Seguridad Ciudadana' con tres dimensiones:

Tabla 2.13

Tabla de especificaciones para instrumento de Percepción de Seguridad Ciudadana — Ejemplo completo

Dimensión	Indicador	Descripción del indicador	Nº ítems	Tipo	Peso %	Ítems núm.
D1: Seguridad objetiva (30%)	I1.1 Victimización directa	Experiencias personales de delitos en el último año	3	Frec.	10%	1-3
	I1.2 Victimización indirecta	Experiencias de personas cercanas en el entorno	3	Frec.	10%	4-6
	I1.3 Incidencia en el barrio	Percepción de delitos en el vecindario inmediato	3	Lik-5	10%	7-9
D2: Seguridad subjetiva (40%)	I2.1 Miedo al delito	Temor emocional ante situaciones de riesgo	4	Lik-5	14%	10-13
	I2.2 Percepción de riesgo	Evaluación cognitiva de la probabilidad de ser víctima	4	Lik-5	14%	14-17
	I2.3 Modificación conductual	Cambios en comportamiento por razones de seguridad	3	Frec.	12%	18-20
D3: Confianza institucional (30%)	I3.1 Confianza en policía	Percepción de eficacia y honestidad policial	4	Lik-5	15%	21-24
	I3.2 Confianza en sistema judicial	Evaluación de la justicia y eficacia del sistema judicial	4	Lik-5	15%	25-28
TOTALES			28 ítems		100%	

Nota. Lik-5 = Escala Likert de 5 puntos; Frec. = Escala de frecuencia. Los pesos porcentuales son asignados según importancia teórica de cada dimensión. Elaboración propia con fines ilustrativos.

2.7 El Protocolo DCOI: Del Constructo al Instrumento

2.7.1 Fundamentos y justificación del Protocolo DCOI

El Protocolo DCOI (Del Constructo al Instrumento) es una propuesta metodológica original que sistematiza e integra todas las etapas de la arquitectura conceptual del instrumento en un proceso replicable, trazable y auditable. Su desarrollo responde a la observación, verificada en la revisión de centenares de instrumentos publicados, de que los errores más graves y frecuentes en la construcción de instrumentos no ocurren en las etapas técnicas sino en las etapas conceptuales previas: la definición del constructo, la operacionalización y el diseño de la tabla de especificaciones.

El DCOI no es un protocolo rígido que deba seguirse mecánicamente, sino un marco de referencia que puede adaptarse a las especificidades de cada disciplina, tipo de constructo y contexto de investigación. Su valor reside en que explicita y estructura decisiones que, en la práctica habitual, son tomadas de manera implícita e irreflexiva, lo que reduce la posibilidad de errores sistemáticos.

Tabla 2.14

Protocolo DCOI: fases, actividades, productos y criterios de calidad

#	Fase	Actividades principales	Productos requeridos	Criterios de calidad
1	Delimitación del problema de medición	Formular el problema; identificar la necesidad de medir; evaluar instrumentos existentes mediante los CJNI	Declaración del problema de medición; revisión de instrumentos existentes documentada	Los 5 CJNI evaluados explícitamente; decisión de construir nuevo instrumento justificada
2	Revisión sistemática del constructo	PRSLI completo; MAAT elaborada; vacíos identificados	MAAT con ≥ 15 referencias; identificación de dimensiones consensuadas y debatidas; vacíos documentados	Búsqueda en ≥ 3 bases de datos; diagrama PRISMA; síntesis de dimensiones
3	Definición del constructo	Definición conceptual; definición operacional;	Definición conceptual con anclaje teórico explícito;	Definición sin circularidad; anclada en ≥ 2 fuentes teóricas;

#	Fase	Actividades principales	Productos requeridos	Criterios de calidad
		selección del modelo reflectivo/formativo	definición operacional completa; justificación del modelo de medición	operacionalización completa
4	Modelización de la dimensionalidad	Identificación y definición de dimensiones; árbol de dimensiones construido; modelo de medición especificado	Definición de cada dimensión; árbol de dimensiones; diagrama del modelo de medición en AFC o PLS	Dimensiones con definición conceptual propia; árbol sin solapamientos; modelo especificado formalmente
5	Operacionalización	Indicadores por dimensión identificados; evaluados con criterios de calidad; asignados a dimensiones	Lista de indicadores con definición y asignación dimensional; justificación de cada indicador	Todos los indicadores evaluados en los 6 criterios de la Tabla 2.9; al menos 2 indicadores por dimensión
6	Tabla de especificaciones	Blueprint completo; ponderaciones justificadas; número de ítems determinado	Tabla de especificaciones completa en el formato de la Tabla 2.13	Peso de dimensiones justificado; mínimo 3 ítems por dimensión; factor de sobreproducción del 150%
7	Revisión y aprobación conceptual	Revisión por pares especialistas en el constructo; ajustes basados en retroalimentación	Informe de revisión conceptual firmado por ≥ 2 especialistas; registro de cambios realizados	Revisores con expertise demostrado en el constructo; todos los comentarios respondidos por escrito

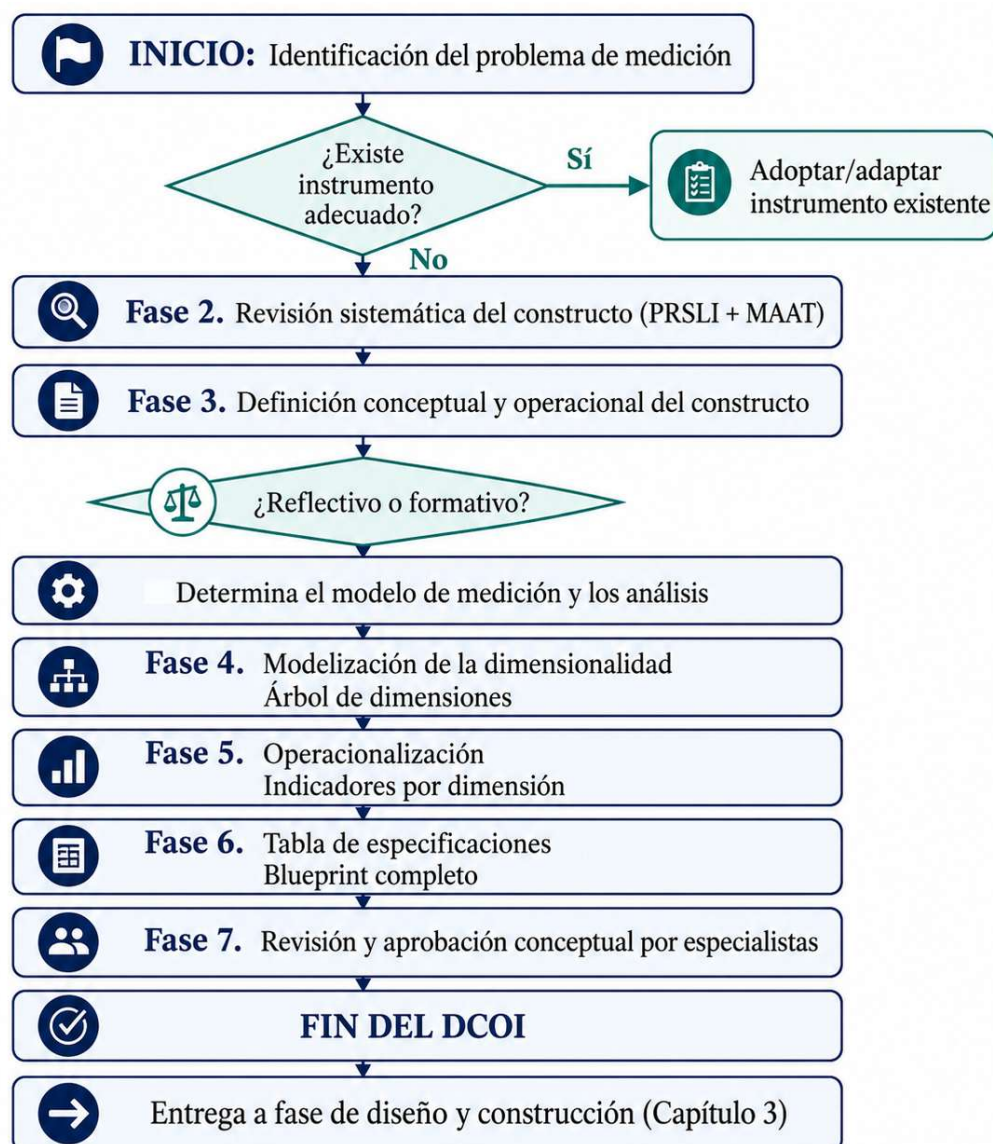
Nota. CJNI = Cinco Criterios de Justificación del Nuevo Instrumento; PRSLI = Protocolo de Revisión Sistemática de Literatura para Instrumentación; MAAT = Matriz de Análisis de Antecedentes Teóricos; PLS = Partial Least Squares.

2.7.2 Diagrama de flujo del Protocolo DCOI

El siguiente diagrama representa el flujo de decisiones del Protocolo DCOI, incluyendo los puntos de decisión donde el proceso puede requerir retroceder a una etapa anterior:

Figura 2.2.

Diagrama de flujo simplificado del Protocolo DCOI.



2.7.3 Lista de verificación del Protocolo DCOI

La siguiente lista de verificación permite al investigador confirmar que ha completado satisfactoriamente cada etapa del Protocolo DCOI antes de avanzar a la fase de construcción del instrumento (Capítulo 3). Está diseñada para ser utilizada como herramienta de autoevaluación y también como criterio de evaluación por pares o por comités de investigación:

Tabla 2.15*Lista de verificación del Protocolo DCOI — 40 criterios*

#	Criterio	Cumple (Sí/No)	Evidencia
1	El problema de medición está claramente formulado y documentado		
2	Se evaluaron sistemáticamente los instrumentos existentes antes de decidir construir uno nuevo		
3	Los 5 Criterios CJNI fueron evaluados explícitamente y al menos uno de ellos justifica el nuevo instrumento		
4	La revisión de literatura cubrió al menos 3 bases de datos especializadas		
5	Se elaboró un diagrama PRISMA del proceso de selección de artículos		
6	Se analizaron al menos 15 referencias relevantes sobre el constructo		
7	Se construyó una MAAT con las principales aportaciones teóricas		
8	Se identificaron explícitamente los vacíos conceptuales, psicométricos y poblacionales		
9	La definición conceptual del constructo está formulada sin circularidad		
10	La definición conceptual cita al menos dos fuentes teóricas establecidas		
11	La definición conceptual delimita claramente qué incluye y qué excluye el constructo		
12	La definición operacional especifica el tipo de instrumento, número de ítems y formato de respuesta		
13	La definición operacional incluye el procedimiento de puntuación		
14	Se determinó explícitamente si el modelo de medición es reflectivo o formativo		
15	La decisión reflectivo/formativo está justificada con argumentos teóricos		
16	Las dimensiones del constructo están identificadas y fundamentadas en la literatura		

#	Criterio	Cumple (Sí/No)	Evidencia
17	Cada dimensión tiene su propia definición conceptual		
18	Las dimensiones son conceptualmente distinguibles (no se solapan sustancialmente)		
19	Las dimensiones cubren exhaustivamente el dominio del constructo		
20	Se construyó un árbol de dimensiones que representa la estructura jerárquica del constructo		
21	Cada dimensión tiene al menos 2 indicadores identificados		
22	Los indicadores fueron evaluados con los 6 criterios de calidad (Tabla 2.9)		
23	Los indicadores son pertinentes para la población objetivo del instrumento		
24	Los indicadores son culturalmente apropiados para el contexto del estudio		
25	La tabla de especificaciones está completamente elaborada		
26	La tabla de especificaciones incluye el número de ítems por indicador y dimensión		
27	Los pesos de las dimensiones están justificados teóricamente o por consenso de expertos		
28	El número total de ítems planificado incluye el factor de sobreproducción del 150%		
29	El tipo de respuesta (Likert, dicotómica, etc.) está especificado para cada ítem		
30	La arquitectura conceptual fue revisada por al menos un especialista externo en el constructo		
31	Todos los comentarios de los revisores fueron respondidos por escrito		
32	Los cambios realizados a partir de la revisión están documentados		
33	La definición conceptual es coherente con el nivel de análisis (individual/grupal/organizacional)		

#	Criterio	Cumple (Sí/No)	Evidencia
34	Las consideraciones éticas del diseño del constructo han sido revisadas		
35	Los sesgos potenciales en la conceptualización han sido identificados y discutidos		
36	La arquitectura conceptual es coherente con el paradigma de investigación declarado		
37	El uso de herramientas de IA en la revisión de literatura está documentado		
38	Las referencias generadas por IA fueron verificadas en bases de datos originales		
39	El proceso completo de arquitectura conceptual está documentado en el expediente del instrumento		
40	El investigador puede defender cada decisión conceptual con argumentos teóricos explícitos		

Nota. Esta lista puede usarse como herramienta de autoevaluación y como instrumento de revisión por pares. Puntuación: 40/40 = arquitectura conceptual excelente; 35-39/40 = adecuada con ajustes menores; 30-34/40 = necesita revisión sustancial; < 30/40 = reiniciar el proceso desde la fase deficiente.

2.8 Consideraciones éticas en la arquitectura conceptual

2.8.1 La ética no comienza con el consentimiento informado

Una idea arraigada pero equivocada en la práctica investigativa es que las consideraciones éticas comienzan cuando se administra el instrumento a los participantes: en el momento de obtener el consentimiento informado, de proteger la confidencialidad de los datos, de minimizar los riesgos de la participación. Sin embargo, desde la perspectiva de la ética en investigación más avanzada, las decisiones éticas más importantes se toman mucho antes: en la etapa de conceptualización del constructo, de selección de dimensiones e indicadores, y de diseño del instrumento.

Esto ocurre porque las decisiones conceptuales del instrumento determinan qué tipos de personas o grupos serán sistemáticamente favorecidos o desfavorecidos por el instrumento, qué aspectos de la experiencia humana serán considerados relevantes y cuáles serán ignorados, y qué tipos de conocimiento serán reconocidos como válidos y cuáles serán excluidos. Estas son, en esencia, decisiones éticas y políticas disfrazadas de decisiones técnicas.

2.8.2 Sesgos teóricos y cómo identificarlos

Los sesgos teóricos son supuestos implícitos en la conceptualización del constructo que sistemáticamente favorecen o desfavorecen a ciertos grupos. Los más comunes son:

- Sesgo cultural: el constructo se define desde los valores, normas y experiencias de un grupo cultural dominante, haciendo que las manifestaciones del constructo en otros grupos culturales sean consideradas inferiores, desviadas o inexistentes. Ejemplo: definir la 'salud mental' exclusivamente desde parámetros occidentales individualizados, sin contemplar conceptualizaciones colectivistas del bienestar.
- Sesgo de género: la conceptualización del constructo refleja experiencias y normas de un género específico. Ejemplo: instrumentos de liderazgo que operacionalizan el constructo exclusivamente con atributos asociados estereotípicamente a la masculinidad.
- Sesgo de clase: el constructo se define desde la experiencia de clases socioeconómicas con acceso a ciertos recursos culturales y materiales. Ejemplo: instrumentos de 'capital cultural' que valoran formas de capital típicamente accesibles solo para clases altas.
- Sesgo de normalidad: el constructo asume que existe un estado 'normal' o 'deseable' que todos deberían alcanzar, patologizando implícitamente la diversidad. Ejemplo: instrumentos de 'integración social' que definen la integración en términos de conformidad con normas mayoritarias.

2.8.3 Representatividad y justicia en la definición del constructo

El principio de representatividad establece que el instrumento debe medir el constructo de manera que sea igualmente relevante y significativa para todos los grupos a los que se aplicará.

El principio de justicia establece que el uso del instrumento no debe producir consecuencias diferencialmente perjudiciales para grupos específicos.

En la práctica, estos principios se implementan mediante:

- Inclusión de representantes de los grupos objetivo en el proceso de definición del constructo y de los indicadores, no solo como informantes sino como co-investigadores con poder de decisión.
- Revisión sistemática de los indicadores para detectar sesgos potenciales de cultura, género, clase, edad o capacidad.

- Análisis de funcionamiento diferencial de ítems (DIF) en la fase de validación, para detectar ítems que favorecen sistemáticamente a ciertos grupos.
- Uso de lenguaje inclusivo en la redacción de ítems.
- Análisis de las consecuencias potenciales del uso del instrumento para diferentes grupos poblacionales.

2.8.4 Consentimiento informado y protección de datos desde el diseño

La arquitectura conceptual del instrumento debe contemplar desde el diseño los aspectos éticos que afectarán la recolección de datos. En particular, el diseño del constructo y sus indicadores debe ser compatible con:

- El principio de consentimiento informado: los participantes deben poder comprender qué se les va a preguntar y para qué se usarán sus respuestas. Un instrumento con indicadores sobre aspectos muy sensibles de la vida privada requiere protocolos de consentimiento más elaborados.
- El principio de minimización de datos: el instrumento no debe recoger más información de la estrictamente necesaria para responder a la pregunta de investigación. Indicadores superfluos desde el punto de vista teórico también son injustificables desde el punto de vista ético.
- El principio de privacidad por diseño: desde la fase de definición de indicadores, se debe planificar cómo se protegerá la privacidad de los participantes, especialmente cuando los indicadores implican información sensible.
- La regulación de protección de datos vigente: en muchos países, la recolección de datos sobre salud, orientación sexual, creencias religiosas, opiniones políticas u origen étnico está sujeta a regulaciones especiales que deben contemplarse desde la fase de diseño del instrumento.

2.9 Casos prácticos de aplicación del Protocolo DCOI

2.9.1 Caso práctico 1: Instrumento de Percepción de Riesgo ante Inundaciones en Comunidades Rurales

Este caso ilustra la aplicación completa del Protocolo DCOI para el desarrollo de un instrumento de percepción de riesgo dirigido a comunidades rurales con alta exposición a inundaciones en América Latina. Se trata de un constructo con alta relevancia para la gestión

del riesgo de desastres y para el cual los instrumentos disponibles en la literatura son inadecuados para esta población específica.

Fase 1 — Problema de medición:

El equipo de investigación identificó la necesidad de medir la percepción de riesgo ante inundaciones en comunidades rurales con baja escolaridad promedio. La revisión de instrumentos existentes (SETA, MICR, HRBS) reveló que todos fueron desarrollados en contextos urbanos de países de altos ingresos, con ítems que asumen acceso a medios de comunicación masivos y familiaridad con sistemas de alerta formales inexistentes en el contexto objetivo. Los CJNI confirmaron la justificación: inadecuación poblacional y cultural.

Fase 2 — Revisión sistemática:

La MAAT elaborada con 23 artículos identificó dos tradiciones teóricas principales: la Teoría de la Motivación para la Protección (Rogers, 1983) y los modelos culturales de percepción del riesgo (Douglas & Wildavsky, 1982). Se identificó un vacío específico: ningún modelo teórico contemplaba explícitamente el conocimiento local y la memoria histórica comunitaria como componentes de la percepción del riesgo, a pesar de su relevancia empírica documentada en estudios cualitativos.

Fase 3 — Definición del constructo:

Definición conceptual adoptada: 'La percepción de riesgo ante inundaciones es la evaluación subjetiva que realiza un individuo sobre la probabilidad de que un evento de inundación ocurra en su entorno y sobre las consecuencias esperadas para su vida, su familia y sus bienes, integrando tanto el conocimiento técnico-formal como el saber local y la memoria histórica comunitaria' (adaptado de Slovic, 1987, y Wisner et al., 2004).

Fase 4 — Dimensiones del constructo:

A partir de la revisión y la definición adoptada, se identificaron cinco dimensiones:

- D1: Valoración de la amenaza (probabilidad percibida de inundación)
- D2: Valoración del impacto (consecuencias percibidas para la vida y los bienes)
- D3: Confianza en la capacidad de respuesta (eficacia percibida de respuestas propias e institucionales)
- D4: Saber local y memoria histórica (conocimiento basado en experiencias comunitarias previas)
- D5: Confianza institucional (credibilidad percibida de las autoridades de gestión del riesgo)

La dimensión D4 representa la innovación teórica del instrumento respecto a los modelos previos.

Tabla 2.16

Tabla de especificaciones del Instrumento de Percepción de Riesgo ante Inundaciones (IPRI)

Dimensión	Indicadores	Ejemplo de ítem	Nº ítems	Formato	Peso	Ítems
D1: Amenaza	Probabilidad percibida; frecuencia histórica; señales de alerta	"¿Qué tan probable considera que su comunidad sufra una inundación en los próximos 5 años?"	5	Lik-4	20%	1-5
D2: Impacto	Daños esperados a vivienda, cultivos, salud, vidas	"Si ocurriera una inundación, ¿cuánto daño cree que causaría a su vivienda?"	6	Lik-4	24%	6-11
D3: Capacidad respuesta	Eficacia propia; recursos disponibles; conocimiento de evacuación	"¿Qué tan capaz se siente de proteger a su familia si hay una inundación?"	5	Lik-4	20%	12-16
D4: Saber local	Memoria de eventos pasados; señales naturales de alerta; conocimiento del territorio	"¿Conoce señales naturales (ríos, animales, lluvias) que en su comunidad anuncian una inundación?"	5	Mixto	20%	17-21
D5: Confianza inst.	Credibilidad de autoridades; efectividad sistemas alerta; comunicación oficial	"¿Cuánto confía en que las autoridades le avisarán a tiempo si hay riesgo de inundación?"	4	Lik-4	16%	22-25
TOTAL			25 ítems		100%	

Nota. IPRI = Instrumento de Percepción de Riesgo ante Inundaciones. Lik-4 = Escala Likert de 4 puntos (sin punto medio, para evitar respuestas neutras en contexto de baja escolaridad). Mixto = combinación de Likert y respuesta dicotómica.

2.9.2 Caso práctico 2: Escala de Competencias Docentes para la Educación en Gestión del Riesgo

Este segundo caso ilustra la arquitectura conceptual de un instrumento para medir las competencias de los docentes de educación básica en el área de gestión del riesgo de desastres, con énfasis en la integración curricular de este tema en países con alta vulnerabilidad sísmica.

Tabla 2.17

Arquitectura conceptual de la Escala de Competencias Docentes en Gestión del Riesgo (ECGRD)

Componente DCOI	Decisión adoptada	Justificación	Implicación para el instrumento
Criterios CJNI	CJNI 1 (inexistencia) + CJNI 2 (inadecuación conceptual)	No existe instrumento específico para competencias docentes en GRD; los de competencias docentes genéricas no incluyen este dominio	Justifica construcción de nuevo instrumento
Modelo teórico	Competencia docente como constructo multidimensional (Shulman, 1987; UNESCO, 2017)	El conocimiento pedagógico del contenido requiere separar saber disciplinar, saber didáctico y saber contextual	Modelo reflectivo con tres dimensiones de primer orden y un factor general de segundo orden
Definición conceptual	Conjunto integrado de conocimientos, habilidades y actitudes que habilitan al docente para incorporar la gestión del riesgo como eje transversal en su práctica pedagógica	Integra la tradición de competencias docentes (Shulman) con el Marco de Sendai y las guías UNESCO para la educación en RRD	Instrumento con tres subescalas distintas pero correlacionadas
Dimensiones	D1: Conocimiento del riesgo; D2: Competencia pedagógica en RRD; D3: Actitud y compromiso	Deriva de la revisión de la literatura sobre competencias docentes y educación en gestión del riesgo (Petal, 2008; Mitchell et al., 2016)	Tres subescalas con puntuaciones independientes y puntuación total

Componente DCOI	Decisión adoptada	Justificación	Implicación para el instrumento
Especificación	32 ítems: 10 en D1, 12 en D2, 10 en D3; Likert de 5 puntos	D2 tiene más ítems por la mayor complejidad pedagógica; factor de sobreproducción del 150% generará 48 ítems iniciales	Tabla de especificaciones con 16 indicadores distribuidos entre las tres dimensiones

Nota. GRD = Gestión del Riesgo de Desastres; RRD = Reducción del Riesgo de Desastres. Elaboración propia con fines ilustrativos.

Reflexiones finales del capítulo

Este capítulo ha recorrido el proceso de arquitectura conceptual del instrumento de investigación desde sus primeros pasos hasta el producto final de esta etapa: la tabla de especificaciones completa y la aprobación conceptual por pares especialistas. En el camino, hemos desarrollado y fundamentado el Protocolo DCOI como una herramienta metodológica original que sistematiza e integra todas las decisiones conceptuales en un proceso transparente y auditable.

La lección más importante de este capítulo puede formularse como un principio: la validez de un instrumento es el producto de la calidad de sus decisiones conceptuales, no de la sofisticación de sus análisis estadísticos. El AFC más elaborado no puede rescatar un instrumento cuyo constructo está mal definido, cuyas dimensiones se solapan o cuya tabla de especificaciones fue diseñada sin criterio teórico. Por el contrario, un instrumento con una arquitectura conceptual sólida tiene todas las condiciones para ser válido, y el análisis estadístico solo necesita confirmarlo.

El capítulo siguiente lleva esta arquitectura conceptual al taller de construcción: aprenderemos a diseñar y redactar los ítems, a elegir los formatos de respuesta apropiados, a desarrollar los instrumentos cuantitativos, cualitativos y mixtos en toda su diversidad técnica, y a someter el instrumento al rigor de la validación de contenido y el pilotaje.

Preguntas de reflexión

1. Un organismo internacional de gestión del riesgo le solicita que desarrolle un índice para medir la 'preparación institucional ante terremotos' en municipios de América Central. Aplique los 5 Criterios CJNI para determinar si es necesario construir un instrumento nuevo o si puede adoptar o adaptar uno existente. Justifique cada criterio con argumentos específicos al contexto descrito.

2. Analice el siguiente caso: un investigador define la 'calidad de vida' como 'la satisfacción subjetiva del individuo con su vida en general'. Identifique al menos cuatro problemas en esta definición conceptual usando los criterios de calidad establecidos en la sección 2.2.4. Proponga una definición alternativa de mayor calidad con su respectivo anclaje teórico.
3. Usted debe construir un instrumento para medir el 'capital social comunitario' en comunidades indígenas andinas. Describa cómo aplicaría el Protocolo de Revisión Sistemática para Instrumentación (PRSLI) a este caso, identificando específicamente los desafíos metodológicos y éticos que anticipa en la revisión de literatura y en la definición del constructo.
4. Diseñe la tabla de especificaciones completa para un instrumento de 30 ítems que mida la 'autoeficacia para la gestión del riesgo de desastres' en docentes de educación secundaria. Justifique el número de dimensiones, el número de ítems por dimensión y los pesos asignados. Use el formato de la Tabla 2.13 como modelo.
5. Reflexione sobre la siguiente afirmación: 'La decisión de si el constructo X es reflectivo o formativo no es solo una decisión estadística: es una decisión ontológica sobre la naturaleza del fenómeno'. Ilustre esta afirmación con dos ejemplos de constructos relevantes para su área disciplinar, uno reflectivo y uno formativo, y explique qué consecuencias metodológicas tiene esta decisión en cada caso.

CAPÍTULO

3

Diseño y Construcción del Instrumento: Proceso, Enfoque y Técnica

María Licenia Suárez Guerra

$$X = \sum_{i=1}^n x_i$$

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum s_i^2}{s_T^2} \right)$$



CAPÍTULO 3

Diseño y Construcción del Instrumento: Proceso, Enfoque y Técnica

El taller completo: desde el primer ítem hasta el instrumento piloteado

Si el Capítulo 2 construyó la arquitectura conceptual del instrumento (el plano que especifica qué se va a medir y con qué estructura), este capítulo inaugura el taller de construcción: el espacio donde esa arquitectura se convierte en un instrumento real, con ítems redactados, escalas de respuesta definidas, formatos diseñados, y evidencias de calidad acumuladas. Es, por su naturaleza, el capítulo más técnico y operativo de la obra.

La construcción de un instrumento de investigación es un proceso simultáneamente artístico y científico. Artístico porque requiere creatividad lingüística, sensibilidad cultural y capacidad para traducir conceptos abstractos en preguntas comprensibles para personas de carne y hueso. Científico porque cada decisión de diseño debe estar guiada por principios psicométricos, evidencia empírica y criterios metodológicos explícitos que permitan evaluar y mejorar el instrumento de manera sistemática.

El capítulo está organizado en tres grandes bloques. El primero introduce el Protocolo PCUI (Proceso Completo Universal de Instrumentos), la hoja de ruta que organiza todas las etapas técnicas de construcción. El segundo bloque desarrolla en profundidad los instrumentos cuantitativos (escalas, cuestionarios, pruebas de rendimiento, registros estructurados) con sus reglas de diseño, fórmulas de análisis y herramientas de IA asociadas. El tercer bloque hace lo propio con los instrumentos cualitativos y mixtos. El capítulo cierra con la Matriz de Decisión Metodológica, una herramienta original que permite al investigador elegir el instrumento más adecuado para su propósito.

3.1 El Protocolo PCUI: Proceso Completo Universal de Instrumentos

3.1.1 Fundamentos y justificación del PCUI

El Protocolo PCUI (Proceso Completo Universal de Instrumentos) es la propuesta metodológica original de esta obra para organizar y sistematizar la fase de construcción técnica del instrumento. Si el Protocolo DCOI del Capítulo 2 gobernaba la fase conceptual, el PCUI gobierna la fase técnica: desde el diseño de los primeros ítems hasta la entrega del instrumento piloteado y revisado.

El término 'universal' en el nombre del protocolo no significa que el proceso sea idéntico para todos los tipos de instrumento: significa que sus fases y principios son aplicables a cualquier

tipo de instrumento con las adaptaciones que cada enfoque requiere. Esta universalidad es precisamente lo que diferencia al PCUI de los manuales especializados que describen el proceso de construcción solo para un tipo específico de instrumento.

Tabla 3.1

Protocolo PCUI: ocho fases de construcción técnica del instrumento

Fase	Nombre	Actividades principales	Criterios de calidad	Duración estimada
1	Diseño de ítems	Redacción de ítems según blueprint; revisión de reglas de calidad; generación de pool ampliado (150-200%)	Reglas de redacción cumplidas; ítems sin ambigüedad ni doble carga; diversidad de formulaciones	2-4 semanas
2	Diseño del formato	Elección de escala de respuesta; diseño visual y secuencia de ítems; instrucciones al respondiente	Escala coherente con nivel de medición requerido; instrucciones claras y breves; diseño accesible	1 semana
3	Revisión lingüística	Revisión por especialista en lingüística y por miembro de la población objetivo	Lenguaje apropiado al nivel educativo; sin ambigüedades; sin discriminación	1 semana
4	Validez de contenido	Juicio de expertos; cálculo de V de Aiken e IVC; revisión de ítems problemáticos	V de Aiken ≥ 0.70 para cada ítem; IVC ≥ 0.78 ; todos los comentarios respondidos	2-3 semanas
5	Prueba cognitiva	Entrevistas cognitivas o pensamiento en voz alta con 5-10 participantes representativos	Ningún ítem produce confusión sistemática; instrucciones comprensibles; tiempo de administración aceptable	1-2 semanas
6	Pilotaje	Aplicación a muestra de 100-300 participantes; análisis de ítems; AFC exploratorio	Ítems con correlación ítem-total ≥ 0.30 ; sin ítems con efecto techo/suelo; estructura factorial preliminar confirmada	3-4 semanas

Fase	Nombre	Actividades principales	Criterios de calidad	Duración estimada
7	Análisis psicométrico	AFE si es necesario; AFC; análisis de confiabilidad; validez convergente/discriminante	Índices de ajuste AFC adecuados; α u $\omega \geq 0.70$; $AVE \geq 0.50$; evidencia de validez de constructo	2-3 semanas
8	Versión definitiva	Revisión final; estandarización opcional; redacción del manual; elaboración del expediente técnico	Manual completo; expediente técnico documentado; instrumento listo para aplicación	1-2 semanas

Nota. AFC = Análisis Factorial Confirmatorio; AFE = Análisis Factorial Exploratorio; IVC = Índice de Validez de Contenido; AVE = Average Variance Extracted. Los tiempos son estimaciones para proyectos de investigación estándar. Elaboración propia.

3.1.2 Principios rectores del PCUI

El PCUI se sustenta en seis principios metodológicos que deben guiar todas las decisiones técnicas del proceso de construcción:

1. Principio de fidelidad conceptual: cada decisión técnica debe ser coherente con la arquitectura conceptual establecida en el DCOI. Si un ítem no puede vincularse directamente a un indicador de la tabla de especificaciones, no tiene lugar en el instrumento.
2. Principio de sobreproducción controlada: siempre deben generarse más ítems de los que se necesitarán en la versión final (150-200%), para poder descartar los problemáticos durante el pilotaje sin comprometer la cobertura del constructo.
3. Principio de iteración: el proceso de construcción es iterativo, no lineal. Es normal y esperado que sea necesario regresar a fases anteriores para revisar y mejorar el instrumento basándose en la evidencia acumulada.
4. Principio de evidencia múltiple: la calidad del instrumento debe ser evaluada mediante múltiples fuentes de evidencia (juicio de expertos, prueba cognitiva, pilotaje, AFC) que se complementan y no se sustituyen mutuamente.
5. Principio de transparencia documentada: cada decisión del proceso de construcción debe ser documentada con los argumentos que la sustentan, para garantizar la trazabilidad y reproducibilidad del proceso.

6. Principio de pertinencia contextual: el instrumento debe ser evaluado no solo en términos de sus propiedades psicométricas abstractas, sino de su adecuación al contexto específico (poblacional, cultural, organizacional) de aplicación.

3.2 Diseño de ítems: el arte y la ciencia

3.2.1 Anatomía de un ítem: componentes y tipología

El ítem es la unidad operativa fundamental de cualquier instrumento de investigación cuantitativo. Aunque su apariencia pueda ser engañosamente simple, un ítem bien construido es el resultado de múltiples decisiones técnicas sobre su contenido, su formulación, su nivel de dificultad esperado, su sensibilidad al constructo y su adecuación a la población. Cada uno de estos aspectos puede ser fuente de error de medición si no se gestiona cuidadosamente.

Tabla 3.2

Tipología completa de ítems en instrumentos de investigación cuantitativa

Tipo de ítem	Descripción	Ventajas	Limitaciones
Escala de valoración (Likert)	Afirmación con opciones de acuerdo graduadas (p.ej. 1=Totalmente en desacuerdo a 5=Totalmente de acuerdo)	Fácil de responder; permite gradación de actitudes; análisis estadístico amplio	Susceptible a aquiescencia y deseabilidad social; asunción de intervalos iguales cuestionada
Escala de frecuencia	Pregunta sobre la frecuencia de un comportamiento (Nunca/A veces/Siempre)	Reduce la deseabilidad social al anclar en conductas observables	Dificultad para definir los anclajes temporales con precisión
Dicotómico Sí/No	Dos opciones de respuesta excluyentes	Simple; fácil de codificar; apropiado para hechos objetivos	Pérdida de información por falta de gradación; no captura matices
Opción múltiple	Una pregunta con 3-5 alternativas de respuesta solo una correcta o más pertinente	Evaluación de conocimiento preciso; reduce sesgo de aquiescencia	Requiere diseño cuidadoso de distractores; susceptible a efectos de posición

Tipo de ítem	Descripción	Ventajas	Limitaciones
Escala visual analógica (EVA)	Línea continua de 100mm con anclajes en los extremos; el respondiente marca su posición	Alta sensibilidad; captura variaciones muy finas; apropiada para dolor o intensidad emocional	Difícil de administrar en papel; requiere regla para medir; no funciona bien digitalmente sin interfaz adecuada
Diferencial semántico	Pares de adjetivos opuestos separados por escala de 7 puntos	Mide connotaciones afectivas de conceptos; robusto culturalmente	Requiere pares de adjetivos apropiados al constructo; análisis más complejo
Respuesta abierta corta	Pregunta que elicitaba una respuesta breve no estructurada	Captura información no anticipada; rico en detalle	Difícil de cuantificar; requiere codificación; mayor carga para el respondiente
Clasificación/ranking	El respondiente ordena un conjunto de opciones de mayor a menor importancia	Elimina el efecto de aquiescencia; diferencia prioridades	Cognitivamente exigente; no informa sobre niveles absolutos

Nota. Elaboración propia.

3.2.2 Las 20 reglas de oro para redactar ítems de calidad

La redacción de ítems es la etapa donde la mayoría de los investigadores cometen errores que comprometen la validez del instrumento. Las siguientes 20 reglas, derivadas de la literatura psicométrica y de la práctica de construcción de instrumentos, constituyen el estándar mínimo que todo ítem debe cumplir antes de pasar a la fase de juicio de expertos:

Reglas de contenido

1. Cada ítem debe medir un solo concepto o aspecto del constructo. Los ítems con doble carga “Mi supervisor me apoya y me comunica claramente los objetivos” son uno de los errores más graves y frecuentes: si el respondiente está de acuerdo con una parte y en desacuerdo con la otra, su respuesta no tiene interpretación unívoca.
2. El contenido del ítem debe corresponder exactamente al indicador asignado en la tabla de especificaciones. Si no hay una correspondencia directa y verificable, el ítem debe ser revisado o descartado.

3. El ítem debe ser sensible al constructo: debe propiciar respuestas diferentes en personas con diferentes niveles del atributo medido. Un ítem al que todos responden lo mismo “Me gustaría tener buena salud” no aporta información discriminativa.
4. El ítem no debe contaminar el constructo con variables irrelevantes como habilidades lectoras, vocabulario especializado, o experiencias culturales específicas que no forman parte del constructo.

Reglas de formulación

5. Usar lenguaje simple, directo y apropiado al nivel educativo de la población objetivo. La regla práctica: el ítem debe ser comprensible para alguien con educación secundaria básica, salvo que la población objetivo tenga un nivel educativo específico diferente.
6. Evitar negaciones, especialmente dobles negaciones. 'No estoy de acuerdo en que el programa no sea útil' es cognitivamente confuso y produce errores de respuesta sistemáticos.
7. Evitar palabras absolutas como 'siempre', 'nunca', 'todos', 'ninguno', que tienden a producir respuestas extremas independientemente del nivel real del atributo.
8. Evitar términos vagos o ambiguos como 'frecuentemente', 'a menudo', 'bastante', cuyo significado varía enormemente entre respondientes. Cuando sea posible, usar anclajes temporales o cuantitativos precisos.
9. Evitar jerga técnica, siglas o términos especializados que no sean universalmente conocidos por la población objetivo, a menos que el instrumento sea específicamente para expertos en ese campo.
10. Formular los ítems de manera directa y no retórica. La pregunta debe solicitar información sobre el estado real del respondiente, no invitar a una reflexión filosófica o a una valoración abstracta.

Reglas de neutralidad

11. Evitar ítems inductivos o cargados que sugieran la respuesta 'correcta' o 'deseable'. 'Como buen ciudadano, ¿contribuye usted a la seguridad de su barrio?' introduce un sesgo de deseabilidad social poderoso.
12. Balancear el número de ítems positivos y negativos en las escalas de actitudes y opiniones. Los ítems exclusivamente positivos favorecen el sesgo de aquiescencia la tendencia a responder 'de acuerdo' independientemente del contenido.

13. Evitar ítems que asuman experiencias o condiciones que no todos los respondientes tienen. “Cuando conduzco hacia el trabajo...” excluye a los respondientes que no tienen auto o que trabajan desde casa.
14. No formular ítems sobre temas que puedan resultar invasivos, amenazantes o estigmatizantes sin una justificación clara y un protocolo ético adecuado.

Reglas de formato

15. Mantener una longitud consistente entre ítems. Los ítems muy largos dentro de un instrumento con ítems cortos reciben una atención diferencial que distorsiona las respuestas.
16. Asegurarse de que cada ítem sea comprensible en sí mismo, sin necesidad de leer los ítems anteriores o posteriores para entenderlo.
17. Evitar efectos de posición agrupando ítems de la misma dimensión, pero alternando ocasionalmente el orden para reducir el sesgo de halo. En instrumentos largos, contrarrestar con múltiples formas del instrumento con orden diferente.
18. El tiempo estimado de respuesta para cada ítem debe ser de 5-15 segundos. Ítems que requieren deliberación prolongada suelen medir razonamiento y no el constructo objetivo.
19. Usar el tiempo verbal presente para describir estados actuales del respondiente, y el tiempo pasado solo cuando se pregunte sobre comportamientos o experiencias previas con un período de referencia explícito.
20. Revisar cada ítem desde la perspectiva del respondiente: ¿Tendría yo dudas sobre cómo responder? ¿La respuesta que daría yo refleja genuinamente mi nivel en el constructo? Este ejercicio de perspectiva es invaluable antes del pilotaje formal.

3.2.3 Ítems positivos y negativos: ventajas, riesgos y estrategias

La inclusión de ítems redactados en dirección negativa también llamados ítems inversos o ítems de control ha sido durante décadas una práctica estándar en la construcción de escalas, justificada principalmente como estrategia para reducir el sesgo de aquiescencia y para detectar respondientes que contestan sin leer los ítems. Sin embargo, la investigación psicométrica contemporánea ha cuestionado seriamente esta práctica, revelando que los ítems inversos pueden introducir problemas metodológicos más graves que los que pretenden resolver.

Tabla 3.3

Análisis comparativo de ítems positivos y negativos: ventajas, riesgos y recomendaciones

Dimensión	Ítems positivos	Ítems negativos (inversos)
Definición	Afirman la presencia del atributo medido: 'Me siento competente para manejar situaciones de estrés'	Afirman la ausencia del atributo o su polo opuesto: 'Con frecuencia me siento incapaz de controlar el estrés'
Ventaja principal	Fáciles de comprender; menor carga cognitiva; menor tasa de errores de respuesta	Controlan la aquiescencia; permiten detectar respondientes inatentos; aumentan la variabilidad
Riesgo principal	Favorecen el sesgo de aquiescencia en escalas con todos los ítems en la misma dirección	Generan factores espurios en el AFC que reflejan el método de redacción, no el constructo
Evidencia empírica	Funcionan bien psicométricamente en poblaciones con alto nivel educativo	Consistentemente problemáticos en poblaciones con baja escolaridad, niños y adultos mayores (Marsh, 1986)
Recomendación actual	Usar como base del instrumento	Usar con moderación (máximo 20-30%); evitar en poblaciones vulnerables; verificar funcionamiento en el pilotaje

Nota. Adaptado de Marsh (1986), Rodebaugh et al. (2004) y Goetz et al. (2021).

Recomendación basada en evidencia

La estrategia más robusta actualmente recomendada es diseñar ítems genuinamente variados en su contenido semántico en lugar de simplemente invertir la redacción. Un ítem invertido no es lo mismo que un ítem que mide el polo opuesto del constructo con contenido propio. La variedad de contenido es más valiosa psicométricamente que la variedad de dirección.

3.2.4 Sesgos sistemáticos en las respuestas: identificación y control

Además del sesgo de aquiescencia, existen otros patrones sistemáticos de respuesta que distorsionan las puntuaciones del instrumento de maneras que no reflejan variaciones reales en el constructo. El investigador debe conocerlos para diseñar estrategias de control efectivas:

Tabla 3.4*Sesgos sistemáticos de respuesta: naturaleza, detección y control*

Sesgo	Descripción	Detección	Estrategias de control
Aquiescencia	Tendencia a responder 'de acuerdo' o 'sí' independientemente del contenido	Correlación entre puntuaciones en ítems positivos e inversos; análisis de patrones de respuesta	Balancear ítems positivos e inversos; usar escalas de respuesta sin opción de acuerdo/desacuerdo
Deseabilidad social	Tendencia a responder lo que parece socialmente correcto o deseable, no lo que realmente se piensa o hace	Correlación con escalas de deseabilidad social (p.ej. Marlowe-Crowne); ítems trampa	Garantías explícitas de anonimato; formulación de ítems conductual en lugar de actitudinal; instrucciones que normalicen respuestas diversas
Tendencia central	Evitar los extremos de la escala eligiendo siempre el punto medio	Distribución acumulada en las categorías centrales	Usar escalas sin punto medio (número par de opciones); anclajes descriptivos en todos los puntos de la escala
Efecto halo	Dejar que la impresión general sobre un objeto o persona influya en todas las valoraciones específicas	Correlaciones entre ítems conceptualmente no relacionados que deberían tener correlación baja	Separar ítems sobre distintos referentes; aleatorizar el orden de presentación de ítems
Fatiga del respondiente	Disminución de la calidad y atención en las respuestas hacia el final del instrumento	Análisis de varianza de respuestas por posición del ítem; diferencias en tiempos de respuesta	Limitar el número de ítems; incluir ítems de control al final; ordenar dimensiones de más a menos interesante
Respuesta extrema	Tendencia a usar consistentemente los valores más extremos de la escala	Distribución acumulada en los extremos; análisis de patrones de respuesta	Anclajes descriptivos precisos; instrucciones que expliquen el significado de cada punto

Nota. Elaboración propia basada en Paulhus (1991) y Podsakoff et al. (2003).

3.2.5 IA como generador y crítico de ítems

Los modelos de lenguaje de gran escala (LLMs) han demostrado una capacidad notable para generar ítems plausibles para instrumentos de investigación. Estudios recientes han mostrado que los ítems generados por LLMs son evaluados por expertos humanos con una calidad comparable a los generados por investigadores con experiencia media (Hommel et al., 2024). Sin embargo, la generación de ítems con IA debe integrarse en el proceso PCUI con un protocolo explícito que garantice el control de calidad:

Tabla 3.5

Protocolo de uso de IA en la generación y revisión de ítems

Paso	Actividad	Prompt recomendado (estructura)	Acción de control
1	Generación de pool inicial	'Actúa como experto en psicometría. El constructo es [X], definido como [definición]. La dimensión es [Y]. Genera 10 ítems en escala Likert de 5 puntos que midan esta dimensión en [población objetivo]. Los ítems deben seguir estas reglas: [listar las 20 reglas]'	Revisar cada ítem generado contra las 20 reglas; descartar los que no las cumplan
2	Evaluación crítica de ítems propios	'Analiza los siguientes ítems para el constructo [X] e identifica problemas de: doble carga, ambigüedad, sesgo cultural, deseabilidad social, o violación de las reglas de redacción psicométrica. Para cada problema, propone una versión mejorada'	Verificar críticamente las sugerencias; no aceptar automáticamente las revisiones propuestas
3	Detección de sesgos	'Revisa los siguientes ítems e identifica posibles sesgos de género, etnia, clase social, edad o cultura que puedan afectar la equidad del instrumento para [describir grupos objetivo]'	Complementar con revisión por miembros de los grupos potencialmente afectados
4	Adaptación lingüística	'Adapta los siguientes ítems para que sean comprensibles para personas con educación secundaria básica en [país/región], sin cambiar su contenido psicométrico'	Verificar con prueba cognitiva con miembros reales de la población

Nota. Los prompts son plantillas que deben ser ajustadas al constructo y contexto específico. Elaboración propia.

Límite crítico del uso de IA en la generación de ítems

La IA puede generar ítems lingüísticamente correctos que son conceptualmente superficiales o teóricamente incoherentes con el modelo del constructo. La IA no 'sabe' si un ítem realmente captura el indicador para el que fue diseñado; solo sabe si suena razonablemente parecido a ítems de instrumentos similares en su corpus de entrenamiento. Por eso, ningún ítem generado por IA debe incluirse en el instrumento sin pasar por la revisión experta de la Fase 4 del PCUI.

3.3 Instrumentos cuantitativos: del diseño al dato

Los instrumentos cuantitativos son aquellos diseñados para producir datos numéricos que representan propiedades de los fenómenos de manera que permiten operaciones matemáticas y análisis estadísticos. Su desarrollo combina la teoría psicométrica (que determina qué propiedades deben tener los ítems y el instrumento en su conjunto) con principios de diseño de encuestas y de comunicación que garantizan que los respondientes puedan y quieran proporcionar información de calidad.

3.3.1 Cuestionarios y encuestas

Arquitectura general del cuestionario

El cuestionario es el instrumento de recolección de datos cuantitativos más ampliamente utilizado en ciencias sociales, de la salud y de la educación. Su arquitectura general comprende las siguientes secciones, que deben diseñarse con tanto cuidado como los propios ítems de medición:

1. Portada e identificación: título del estudio, institución responsable, fecha y código del instrumento (sin datos identificatorios del respondiente en la portada).
2. Presentación e instrucciones generales: propósito del estudio en lenguaje accesible; garantías de confidencialidad o anonimato; instrucciones claras sobre cómo responder; tiempo estimado de cumplimentación; información sobre quién puede contactar si tiene preguntas.
3. Filtros de elegibilidad (si aplica): preguntas breves para verificar que el respondiente cumple los criterios de inclusión antes de continuar.
4. Ítems de las escalas principales: organizados por dimensiones o temas, con instrucciones específicas para cada sección si el formato de respuesta varía.

5. Variables sociodemográficas: al final, no al principio, para no introducir sesgos por saliencia de identidades demográficas antes de responder los ítems principales.
6. Agradecimiento y cierre: expresión de gratitud por la participación; instrucciones sobre la devolución del instrumento o el cierre de la encuesta digital.

Tabla 3.6

Tipos de preguntas en cuestionarios y sus propiedades metodológicas

Tipo	Descripción	Ventajas	Limitaciones
Cerrada dicotómica	Dos opciones de respuesta (Sí/No; Verdadero/Falso)	Sencilla; fácil de codificar; apropiada para hechos objetivos	No captura gradaciones; puede forzar simplificación
Cerrada politómica	Múltiples opciones mutuamente excluyentes	Captura categorías sin gradación lineal; apropiada para variables nominales	El número y orden de opciones influye en las respuestas
De escala (Likert, frecuencia)	Opciones ordenadas que representan grados del atributo	Captura gradaciones; permite análisis estadístico amplio	Requiere supuestos sobre igualdad de intervalos
Abierta	Sin opciones; el respondiente escribe su respuesta	Capta información no anticipada; rica en detalle	Difícil de cuantificar; requiere análisis cualitativo
Filtro	Pregunta que determina si el respondiente debe responder las siguientes	Evita preguntas irrelevantes para ciertos subgrupos	Puede aumentar la complejidad percibida del cuestionario
Batería	Conjunto de ítems con el mismo formato referidos a diferentes objetos o dimensiones	Eficiente; reduce fatiga; facilita comparaciones	Favorece el efecto halo dentro de la batería

Nota. Elaboración propia.

Diseño de encuestas digitales

La administración digital de cuestionarios a través de plataformas como Qualtrics, LimeSurvey, Google Forms o REDCap ha modificado significativamente las posibilidades y los desafíos del diseño de instrumentos. Las principales consideraciones metodológicas específicas de la administración digital son:

- Diseño responsivo: el instrumento debe ser igualmente funcional en pantallas grandes (computadora de escritorio) y en dispositivos móviles, donde la mayoría de los respondientes en países de ingresos medios y bajos acceden a las encuestas.
- Progreso visible: incluir una barra de progreso mejora significativamente la tasa de completitud al proporcionar al respondiente una señal sobre cuánto le falta.
- Validación en tiempo real: los formularios digitales permiten validar respuestas en tiempo real (rango de valores permitidos, preguntas obligatorias) reduciendo errores de entrada de datos.
- Lógica de salto automatizada: las preguntas filtro pueden implementarse de manera transparente para el respondiente, mejorando la experiencia sin mostrar ítems irrelevantes.
- Aleatorización de ítems y opciones: es fácil aleatorizar el orden de los ítems y de las opciones de respuesta entre respondientes, lo que permite controlar efectos de orden sistemáticos.
- Tiempos de respuesta: las plataformas digitales pueden registrar el tiempo de respuesta por ítem, permitiendo identificar respondientes que responden demasiado rápidamente (posiblemente sin leer) o demasiado lentamente (posiblemente distraídos).

Tabla 3.7

Comparación de plataformas digitales para administración de cuestionarios de investigación

Plataforma	Fortalezas	Costo	Aleatorización	Lógica de salto	Mejor para
Qualtrics	Funcionalidades avanzadas; análisis integrado; alta seguridad de datos; soporte técnico profesional	Alto (institucional)	Avanzada	Avanzada	Investigación universitaria; estudios clínicos; grandes muestras
REDCap	Diseñado para investigación biomédica; alta seguridad; gestión de datos longitudinales	Gratuito (consorcio)	Básica	Avanzada	Ciencias de la salud; estudios longitudinales; datos sensibles

Plataforma	Fortalezas	Costo	Aleatorización	Lógica de salto	Mejor para
LimeSurvey	Open source; personalizable; sin límite de respuestas; en servidor propio	Bajo (servidor propio)	Avanzada	Avanzada	Investigaciones con gran escala y presupuesto limitado
Google Forms	Extremadamente fácil de usar; gratuito; integración con Google Sheets	Gratuito	Básica	Básica	Estudios exploratorios; encuestas rápidas; contextos con recursos limitados
KoBoToolbox	Diseñado para zonas sin conectividad; funciona offline; gratuito para ONG	Gratuito	Básica	Avanzada	Investigación en campo; comunidades remotas; contextos humanitarios y de desastres

Nota. Elaboración propia. Las características pueden variar con actualizaciones de las plataformas.

3.3.2 Escalas de medición de actitudes, percepciones y opiniones

Escala Likert: diseño avanzado

La escala Likert, propuesta por Rensis Likert en 1932, es el formato de respuesta más utilizado en ciencias sociales. A pesar de su aparente simplicidad, el diseño de una escala Likert de alta calidad involucra decisiones técnicas que tienen consecuencias psicométricas significativas:

Tabla 3.8

Decisiones de diseño en escalas Likert y sus implicaciones psicométricas

Decisión de diseño	Opciones	Recomendación basada en evidencia
Número de puntos	1, 2, 3, 4, 5 o más puntos	3 a 5 puntos para poblaciones con escolaridad media-alta; 2 puntos (sin punto medio) para poblaciones con baja escolaridad o cuando se quiere forzar una dirección

Decisión de diseño	Opciones	Recomendación basada en evidencia
Punto medio	Con o sin punto neutro	Incluir cuando la neutralidad genuina es posible y teóricamente relevante; excluir cuando la neutralidad sería equivalente a no responder
Etiquetado de puntos	Solo extremos; todos los puntos; solo extremos y punto medio	Etiquetar todos los puntos mejora la consistencia entre respondientes y reduce la ambigüedad de los puntos intermedios
Dirección de la escala	Positivo a negativo o negativo a positivo	Positivo a negativo (1=muy en desacuerdo a 5=muy de acuerdo) es la convención más extendida; mantener consistente en todo el instrumento
Presentación visual	Horizontal, vertical, con radio buttons o con slider	Horizontal con radio buttons es el formato más estable; los sliders continuos producen resultados similares a Likert pero con mayor varianza
Tipo de anclaje	Acuerdo, frecuencia, calidad, probabilidad, importancia	Elegir el tipo de anclaje según el contenido del ítem; no mezclar tipos dentro de una misma batería

Nota. Basado en Krosnick y Presser (2010) y Revilla et al. (2014).

Escala de Thurstone: intervalos aparentemente iguales

La escala de Thurstone, desarrollada en 1928, es el primer método sistemático para construir escalas de actitudes con propiedades de intervalo. Su proceso de construcción, aunque laborioso, produce escalas con propiedades métricas superiores a las de Likert y es especialmente apropiado cuando se necesita comparar actitudes entre grupos con características culturales muy diferentes.

El proceso de construcción de una escala de Thurstone implica los siguientes pasos:

1. Generación de un pool amplio de afirmaciones (80-100) que cubran todo el continuo de actitudes posibles hacia el objeto, desde la actitud más negativa hasta la más positiva.
2. Juicio de expertos: entre 50 y 300 jueces clasifican cada afirmación en 11 categorías que van de 'muy desfavorable' (1) a 'muy favorable' (11), sin expresar su propia actitud.
3. Cálculo de la mediana y el rango intercuartílico (Q) para cada afirmación:

Mediana (Md) = valor central de la distribución de clasificaciones (Ecuación 3.1)

$Q = (Q_3 - Q_1) / 2$ donde Q_1 y Q_3 son los percentiles 25 y 75 (Ecuación 3.2)

4. Selección de afirmaciones: se eligen las afirmaciones con Q pequeño (alta consistencia entre jueces) distribuidas uniformemente a lo largo del continuo de valores de la mediana.
5. El instrumento final contiene típicamente 20-22 afirmaciones con valores de mediana de 1 a 11. La puntuación del respondiente es la mediana de los valores de las afirmaciones con las que está de acuerdo.

Diferencial semántico de Osgood

El diferencial semántico, desarrollado por Osgood, Suci y Tannenbaum (1957), mide las connotaciones afectivas o simbólicas de conceptos, personas, objetos o situaciones a lo largo de dimensiones bipolares. Es especialmente útil para medir actitudes, percepciones de imagen y connotaciones culturales de constructos, y ha demostrado una notable robustez transcultural.

El instrumento presenta un concepto estímulo (p.ej. “Gestión de emergencias en mi municipio”) y pide al respondiente que lo evalúe en una serie de escalas bipolares de 7 puntos, por ejemplo:

Adjetivo negativo	1	2	3	4	5	6	7	Adjetivo positivo
Ineficaz	☐	☐	☐	☐	☐	☐	☐	Eficaz
Lenta	☐	☐	☐	☐	☐	☐	☐	Rápida
Desorganizada	☐	☐	☐	☐	☐	☐	☐	Organizada
Lejana	☐	☐	☐	☐	☐	☐	☐	Cercana

Nota. Representación esquemática del diferencial semántico. Las dimensiones universales identificadas por Osgood et al. (1957) son evaluación (bueno-malo), potencia (fuerte-débil) y actividad (activo-pasivo).

3.3.3 Pruebas de conocimiento y rendimiento

Las pruebas de conocimiento y rendimiento difieren fundamentalmente de las escalas de actitudes y percepciones en un aspecto esencial: tienen respuestas correctas e incorrectas. Esto introduce una dimensión adicional de complejidad psicométrica que requiere herramientas de análisis específicas.

Diseño de reactivos de opción múltiple

El reactivo de opción múltiple (ROM) es el formato más utilizado en pruebas de conocimiento por su versatilidad, objetividad en la corrección y posibilidad de automatización. Un ROM de alta calidad tiene los siguientes componentes:

- Base (stem): la pregunta o situación problema que presenta el reactivo. Debe estar completa, ser comprensible por sí sola y estar formulada en términos positivos siempre que sea posible.
- Clave: la respuesta correcta o más correcta. Debe ser inequívocamente la mejor opción según el conocimiento científico actualizado.
- Distractores: las opciones incorrectas. Deben ser plausibles para respondientes con conocimiento incompleto, pero claramente incorrectas para respondientes con dominio del tema. Los distractores de calidad son el elemento más difícil de construir.

Tabla 3.9

Niveles cognitivos de la Taxonomía de Bloom revisada y su aplicación al diseño de reactivos

Nivel	Proceso cognitivo	Verbos clave para el ítem	Ejemplo de reactivo
1. Recordar	Recuperar información de la memoria	Definir, identificar, listar, nombrar, reconocer	¿Cuál de los siguientes es el coeficiente de confiabilidad propuesto por Cronbach?
2. Comprender	Construir significado a partir de información	Explicar, interpretar, parafrasear, clasificar, comparar	¿Qué significa que un instrumento tenga alta confiabilidad pero baja validez?
3. Aplicar	Usar información en situaciones nuevas	Calcular, demostrar, resolver, aplicar, usar	En un instrumento con $\alpha = 0.65$ y $\sigma = 8$, ¿cuál es el error estándar de medición?
4. Analizar	Descomponer en partes y examinar relaciones	Analizar, diferenciar, organizar, atribuir	¿Cuál de los siguientes errores de redacción tiene el ítem presentado?
5. Evaluar	Emitir juicios basados en criterios	Juzgar, criticar, justificar, defender, evaluar	¿Cuál de las siguientes estrategias de validación es más apropiada para este propósito?
6. Crear	Combinar elementos en nuevas estructuras	Diseñar, construir, planificar, producir, proponer	Diseñe la tabla de especificaciones para un instrumento que mida X

Nota. Basado en Anderson et al. (2001), revisión de la Taxonomía de Bloom.

Análisis psicométrico de ítems en pruebas de rendimiento

La teoría clásica del test proporciona tres indicadores fundamentales para evaluar la calidad de cada ítem en una prueba de rendimiento:

Índice de dificultad (p): $p = n_a / N$ (Ecuación 3.3)

Donde n_a es el número de respondientes que respondieron correctamente y N es el número total de respondientes. Valores de p cercanos a 0.50 maximizan la información discriminativa del ítem; valores menores a 0.20 o mayores a 0.80 sugieren ítems demasiado difíciles o demasiado fáciles, respectivamente.

Índice de discriminación (D): $D = p_a - p_b$ (Ecuación 3.4)

Donde p_a es la proporción de correctos en el grupo superior (27% con mayores puntuaciones totales) y p_b es la proporción de correctos en el grupo inferior. $D \geq 0.30$ es considerado adecuado; $D < 0.20$ indica que el ítem no discrimina adecuadamente entre los respondientes con mayor y menor conocimiento.

Correlación ítem-total corregida (r_{it}):

r_{it} = correlación entre la puntuación del ítem y la puntuación total del instrumento sin ese ítem (Ecuación 3.5)

Se recomienda $r_{it} \geq 0.30$ para ítems en escalas de actitudes y ≥ 0.20 para ítems en pruebas de rendimiento. Ítems con r_{it} negativa o cercana a cero deben revisarse o eliminarse.

3.3.4 Instrumentos de observación estructurada

Listas de cotejo

Las listas de cotejo son instrumentos que registran la presencia o ausencia de comportamientos, características o condiciones previamente definidas. Su diseño requiere que cada ítem describa un comportamiento o condición observable de manera inequívoca, de modo que diferentes observadores lleguen a la misma clasificación (presente/ausente) al observar la misma situación.

Los criterios de calidad para una lista de cotejo son:

- Cada ítem describe un comportamiento o condición observable directamente, no una inferencia o interpretación.

- La distinción entre 'presente' y 'ausente' es objetiva: dos observadores independientes deben clasificar el mismo comportamiento de la misma manera más del 80% de las veces.
- Los ítems están organizados en el orden lógico en que los comportamientos ocurren o en que las condiciones deben ser inspeccionadas.
- Las instrucciones para el observador son suficientemente detalladas para garantizar la estandarización de la observación.

Rúbricas analíticas

Las rúbricas analíticas describen con detalle los niveles de desempeño esperados en múltiples dimensiones o criterios, permitiendo una evaluación más precisa, consistente y retroalimentadora que las escalas globales. Son especialmente valiosas en educación, evaluación de competencias y evaluación de la calidad de servicios.

Tabla 3.10

Estructura de una rúbrica analítica para evaluación de competencias en gestión del riesgo de desastres

Criterio	Nivel 1 — Inicial	Nivel 2 — En desarrollo	Nivel 3 — Logrado	Nivel 4 — Avanzado
Identificación de amenazas	No identifica amenazas relevantes del entorno	Identifica algunas amenazas con errores conceptuales	Identifica correctamente las principales amenazas del entorno	Identifica y jerarquiza amenazas con fundamento técnico y evidencia
Análisis de vulnerabilidad	No distingue vulnerabilidad de amenaza	Identifica algunas dimensiones de vulnerabilidad de manera intuitiva	Analiza vulnerabilidad social, física y económica con criterios correctos	Realiza análisis integral de vulnerabilidad integrando múltiples dimensiones con datos
Diseño de estrategias	No propone estrategias pertinentes	Propone estrategias genéricas sin adecuación al contexto	Diseña estrategias pertinentes al contexto con recursos realistas	Diseña estrategias integrales con evaluación de costo-efectividad e impacto diferencial

Criterio	Nivel 1 — Inicial	Nivel 2 — En desarrollo	Nivel 3 — Logrado	Nivel 4 — Avanzado
Comunicación del riesgo	No comunica información de riesgo a la comunidad	Comunica información técnica sin adaptación al público	Comunica el riesgo de manera comprensible para la comunidad	Diseña e implementa estrategias de comunicación diferenciadas para distintos públicos

Nota. Elaboración propia con fines ilustrativos. Los descriptores deben ser validados por expertos en gestión del riesgo de desastres y en evaluación educativa.

Confiabilidad interjueces: Kappa de Cohen y CCI

Cuando el instrumento requiere que uno o más observadores clasifiquen o evalúen comportamientos, el acuerdo entre observadores denominado confiabilidad interjueces es una propiedad esencial del instrumento. Los estadísticos más utilizados para cuantificar el acuerdo son:

Kappa de Cohen (κ) — Para datos nominales u ordinales:

$$\kappa = (Po - Pe) / (1 - Pe) \quad (\text{Ecuación 3.6})$$

Donde Po es la proporción de acuerdo observado y Pe es la proporción de acuerdo esperado por azar. Los criterios de interpretación más utilizados son: $\kappa < 0.20$ = pobre; 0.21-0.40 = débil; 0.41-0.60 = moderado; 0.61-0.80 = bueno; > 0.80 = muy bueno (Landis & Koch, 1977).

Coeficiente de Correlación Intraclass (CCI) — Para datos continuos:

$$CCI = (MSS - MSE) / (MSS + (k-1) \times MSE) \quad (\text{Ecuación 3.7})$$

Donde MSS es la varianza entre sujetos y MSE es la varianza del error. El CCI es más apropiado que la correlación de Pearson para evaluar el acuerdo interjueces en datos continuos porque tiene en cuenta tanto la correlación como la diferencia entre los valores absolutos de los jueces.

3.4 Instrumentos cualitativos: capturando la complejidad

Los instrumentos cualitativos no buscan medir variables en escalas numéricas; buscan comprender, explorar e interpretar fenómenos en su riqueza y complejidad. Esta diferencia de propósito se traduce en una diferencia radical de lógica de diseño: mientras los instrumentos cuantitativos maximizan la estandarización para garantizar la comparabilidad de las respuestas, los instrumentos cualitativos maximizan la flexibilidad para adaptarse al contexto y a las particularidades de cada participante.

Esto no significa que los instrumentos cualitativos no tengan estructura ni requieran diseño cuidadoso. Una guía de entrevista mal diseñada produce datos superficiales, irrelevantes o sesgados con la misma eficiencia que un cuestionario mal construido. La diferencia es que el diseño de los instrumentos cualitativos requiere diferentes habilidades: capacidad de formular preguntas abiertas que invitan a la reflexión sin sugerir las respuestas, sensibilidad a las dinámicas relacionales y conversacionales, y capacidad de adaptación al flujo de la información que emerge.

3.4.1 Guías de entrevista en profundidad

Tipología y selección del tipo de entrevista

La entrevista en profundidad es una conversación intencional y guiada entre el investigador y uno o más participantes, orientada a explorar las experiencias, percepciones, significados y perspectivas del entrevistado sobre un fenómeno de interés. Existen tres tipos principales, que representan diferentes puntos en el continuo entre estructuración y flexibilidad:

Tabla 3.11

Tipos de entrevista cualitativa: características, ventajas y criterios de selección

Tipo	Descripción	Ventajas	Limitaciones	Usar cuando
Estructurada	Preguntas fijas en orden fijo; respuestas verbales a un cuestionario	Alta comparabilidad; puede usarse con entrevistadores sin experiencia	Rigidez; pierde información espontánea; difícil adaptarse a respuestas inesperadas	Se necesita comparabilidad; entrevistadores múltiples; tiempo limitado
Semiestructurada	Guía de temas y preguntas orientativas con libertad de adaptar el orden y profundidad	Combina comparabilidad con flexibilidad; permite explorar temas emergentes	Requiere entrevistadores con experiencia; mayor tiempo de análisis	La mayoría de los estudios cualitativos; fenómenos poco explorados
No estructurada (narrativa)	Solo un tema inicial; el entrevistado dirige la conversación	Máxima libertad para el entrevistado; ideal para	Muy difícil de analizar comparativamente; requiere	Historias de vida; estudios fenomenológicos; investigación exploratoria

Tipo	Descripción	Ventajas	Limitaciones	Usar cuando
		narrativas biográficas	entrevistadores expertos	

Nota. Elaboración propia.

Estructura de la guía de entrevista semiestructurada

Una guía de entrevista semiestructurada de alta calidad incluye los siguientes elementos:

1. Preámbulo: presentación del investigador, propósito del estudio, garantías de confidencialidad, solicitud de permiso para grabar (si aplica) y firma del consentimiento informado.
2. Preguntas de calentamiento: 2-3 preguntas sobre el perfil del entrevistado que no son sensibles y que sirven para establecer rapport y verificar que el participante cumple el perfil esperado.
3. Preguntas principales: las 5-10 preguntas centrales que abordan las dimensiones del fenómeno de interés. Cada pregunta principal debería ir acompañada de 2-3 preguntas de sondeo para profundizar cuando el respondiente no elabora suficientemente.
4. Preguntas de cierre: una pregunta que invite al entrevistado a añadir lo que no se le preguntó (¿Hay algo importante sobre este tema que no hemos mencionado y que le parezca relevante compartir?) y una pregunta de cierre emocional que devuelva al entrevistado a un estado de calma si los temas tratados fueron emocionalmente difíciles.
5. Protocolo post-entrevista: instrucciones para el investigador sobre cómo registrar observaciones contextuales inmediatamente después de la entrevista y criterios para evaluar la calidad de los datos obtenidos.

Tabla 3.12

Tipos de preguntas en la guía de entrevista y su función metodológica

Tipo de pregunta	Función	Ejemplo	Precauciones
Apertura o grand tour	Invitar al entrevistado a describir su experiencia con el fenómeno en sus propios términos	'Cuénteme cómo ha sido su experiencia viviendo en una zona de riesgo de inundaciones'	No hacer preguntas de sí/no; no sugerir elementos de la respuesta

Tipo de pregunta	Función	Ejemplo	Precauciones
Sondeo de elaboración	Invitar al entrevistado a desarrollar más un punto que mencionó brevemente	'Mencionó que la última inundación fue diferente. ¿Podría contarme más sobre eso?'	No interrumpir el hilo narrativo; esperar pausas naturales antes de sondear
Sondeo de clarificación	Verificar la comprensión de un término o concepto usado por el entrevistado	'Cuando dice que la comunidad no estaba preparada, ¿qué significa eso para usted?'	Adoptar la terminología del entrevistado, no imponer la propia
Hipotética	Explorar percepciones sobre situaciones no vividas o imaginadas	'Si mañana ocurriera una inundación, ¿qué haría usted en primer lugar?'	Usar con moderación; puede alejarse de la experiencia real
Reflexiva o estructural	Invitar al entrevistado a reflexionar sobre relaciones, patrones o contrastes	'¿En qué se diferencia la respuesta de la comunidad a esta inundación de la de hace cinco años?'	Asegurarse de que el entrevistado tiene experiencia suficiente para comparar
Cierre	Cerrar la entrevista y dar al entrevistado la última palabra	'¿Hay algo importante sobre este tema que no le haya preguntado?'	Siempre incluir; frecuentemente produce información valiosa no anticipada

Nota. Adaptado de Patton (2002) y Kvale & Brinkmann (2009).

3.4.2 Guías de grupo focal

El grupo focal es una técnica de recolección de datos cualitativos que explota la interacción grupal para generar datos que no emergerían en entrevistas individuales. La discusión entre participantes produce comparaciones, debates, acuerdos y desacuerdos que revelan la estructura social de las percepciones, las normas culturales implícitas y los rangos de opinión existentes en un grupo.

¿Cuándo usar grupo focal versus entrevista individual?

Tabla 3.13

Criterios para la elección entre grupo focal y entrevista individual

Criterio	Favorable al grupo focal	Favorable a la entrevista individual
Objetivo del estudio	Explorar normas sociales, percepciones compartidas, debates culturales	Explorar experiencias individuales, trayectorias personales, historias de vida
Sensibilidad del tema	Temas moderadamente sensibles en los que el grupo proporciona apoyo social	Temas muy sensibles, vergonzosos o estigmatizados que el participante no compartiría en grupo
Dinámica grupal	La interacción entre participantes es en sí misma informativa	La influencia del grupo distorsionaría las respuestas individuales
Recursos disponibles	Menos tiempo de análisis; acceso a muchos participantes simultáneamente	Acceso geográficamente disperso; imposibilidad de reunir al grupo
Perfil de los participantes	Participantes que comparten un contexto o experiencia común	Participantes con posiciones de poder muy diferentes que inhibirían la participación igualitaria

Nota. Elaboración propia basada en Morgan (1997) y Kitzinger (1995).

Diseño del grupo focal: tamaño, composición y estructura

Las decisiones de diseño más importantes para un grupo focal son:

- **Tamaño del grupo:** entre 6 y 8 participantes es el rango óptimo para la mayoría de los estudios. Grupos más pequeños (3-5) son apropiados para temas complejos o participantes expertos; grupos más grandes (10-12) pueden usarse para exploraciones iniciales con temas poco sensibles, pero aumentan el riesgo de conversaciones paralelas y de dominancia de voces.
- **Composición:** la decisión más crítica es el grado de homogeneidad del grupo. Grupos más homogéneos (misma edad, género, experiencia) facilitan la discusión y el rapport, pero reducen la diversidad de perspectivas. Grupos más heterogéneos generan más debate, pero pueden inhibir a los participantes con menos poder o confianza.

- Número de grupos: para la mayoría de los estudios, se recomiendan entre 3 y 6 grupos para alcanzar la saturación teórica. La saturación se alcanza cuando los grupos adicionales no producen nuevas categorías de información sino confirmación de las ya identificadas.
- Duración: entre 90 y 120 minutos es el rango óptimo. Sesiones más cortas no permiten la profundización; más largas generan fatiga que deteriora la calidad de los datos.

3.4.3 Protocolos de observación cualitativa

La observación es la técnica de recolección de datos más directa disponible para el investigador: permite registrar comportamientos, interacciones y condiciones tal como ocurren en su contexto natural, sin el filtro de las interpretaciones y recuerdos del entrevistado. Sin embargo, 'observar' en la investigación científica es muy diferente de 'mirar' en la vida cotidiana: requiere un protocolo sistemático que guíe qué se observa, cómo se registra y cómo se interpreta.

El diario de campo: estructura y escritura reflexiva

El diario de campo es el instrumento de registro primario de la observación cualitativa. Un diario de campo de investigación de alta calidad distingue explícitamente entre tres tipos de notas:

- Notas descriptivas: registro detallado y objetivo de lo que se observó. ¿Quiénes estaban presentes? ¿Qué dijeron? ¿Qué hicieron? ¿En qué orden? ¿Cuál era el entorno físico? Las notas descriptivas son la materia prima de los datos cualitativos y deben ser lo más densas y detalladas posible.
- Notas analíticas: reflexiones del investigador sobre el significado de lo observado, las conexiones con el marco teórico, las hipótesis emergentes y los patrones identificados. Deben estar claramente separadas de las notas descriptivas (por ejemplo, entre corchetes o en una columna diferente).
- Notas metodológicas: reflexiones sobre el proceso de observación mismo: ¿cómo afectó mi presencia lo que ocurrió? ¿Qué no pude observar y por qué? ¿Qué cambiaría en futuras observaciones? Estas notas son esenciales para la reflexividad y la confirmabilidad del estudio.

Recomendación práctica

Las notas de campo deben escribirse lo más cerca posible del momento de la observación, idealmente durante la misma o inmediatamente después. Cada hora de observación genera típicamente entre 5 y 15 páginas de notas densas. El investigador que 'tomará notas más tarde' sistemáticamente pierde entre el 30% y el 70% de los detalles más ricos y precisos de la observación.

3.4.4 Criterios de rigor en instrumentos cualitativos

Como se estableció en el Capítulo 1, la investigación cualitativa no adopta los criterios cuantitativos de validez y confiabilidad, sino un sistema propio de criterios de rigor propuesto por Lincoln y Guba (1985) y expandido posteriormente por múltiples autores. Estos criterios no son solo estándares de evaluación: deben ser considerados desde el diseño del instrumento como principios que guían la construcción de guías, protocolos y registros.

Tabla 3.14

Criterios de rigor en instrumentos cualitativos: definición, estrategias de implementación y aplicación al diseño

Criterio	Equivalente cuantitativo	Estrategias de implementación desde el diseño del instrumento	Indicadores de calidad
Credibilidad	Validez interna	Diseñar preguntas que inviten a la elaboración; incluir verificación por participantes; triangular con otras fuentes	El instrumento no sugiere respuestas; permite corroboración de interpretaciones con los participantes
Transferibilidad	Validez externa	Diseñar el protocolo para recoger información rica sobre el contexto (descripción densa); documentar las características del contexto de aplicación	El protocolo incluye registro sistemático del contexto; las notas metodológicas documentan condiciones
Dependencia	Confiabilidad	Documentar el proceso de diseño del instrumento; usar guías escritas; registrar desviaciones del protocolo	Guía escrita disponible para replicación; registro de adaptaciones documentadas

Criterio	Equivalente cuantitativo	Estrategias de implementación desde el diseño del instrumento	Indicadores de calidad
Confirmabilidad	Objetividad	Incluir preguntas de verificación; registrar las reflexiones del investigador sobre sus sesgos; establecer un 'audit trail'	Separación explícita entre notas descriptivas y analíticas; reflexividad documentada
Autenticidad	Sin equivalente directo	Diseñar preguntas que respeten y capturen las perspectivas de los participantes en sus propios términos	El instrumento usa el lenguaje de la comunidad; las preguntas no imponen categorías ajenas

Nota. Basado en Lincoln y Guba (1985) y Mertens (2015).

3.5 Instrumentos para investigación mixta

La investigación de métodos mixtos combina elementos cuantitativos y cualitativos en un mismo estudio, integrando sus respectivas fortalezas para responder preguntas de investigación que ninguno de los enfoques por separado podría abordar de manera óptima. Esta integración tiene implicaciones directas para el diseño de los instrumentos: en algunos diseños, los instrumentos cuantitativos y cualitativos son independientes y complementarios; en otros, un solo instrumento integra elementos de ambos enfoques.

3.5.1 Lógica de la integración cuanti-cualitativa

La integración en la investigación mixta puede ocurrir en diferentes momentos del proceso: en el diseño del instrumento, en la recolección de datos, en el análisis, o en la interpretación de los resultados. Creswell y Plano Clark (2018) identifican cuatro tipos básicos de integración que tienen implicaciones diferentes para el diseño de los instrumentos:

Tabla 3.15

Diseños de métodos mixtos y sus implicaciones para el diseño de instrumentos

Diseño	Descripción	Instrumentos requeridos	Punto de integración	Ejemplo de aplicación
Convergente	CUAN y CUAL simultáneamente; integración en la interpretación	Instrumento cuantitativo + instrumento cualitativo administrados independientemente	Interpretación: los resultados se triangularan al final	Encuesta de clima organizacional + grupos focales sobre experiencias laborales
Explicativo secuencial	CUAN primero; los resultados guían el CUAL	Instrumento cuantitativo estándar; guía cualitativa diseñada después basándose en los resultados cuantitativos	Diseño del instrumento cualitativo: los resultados cuantitativos determinan qué explorar	Escala de percepción del riesgo; entrevistas con subgrupos extremos identificados cuantitativamente
Exploratorio secuencial	CUAL primero; los resultados guían el CUAN	Guía cualitativa estándar; instrumento cuantitativo construido basándose en los hallazgos cualitativos	Diseño del instrumento cuantitativo: las categorías cualitativas se convierten en ítems	Grupos focales para identificar dimensiones de vulnerabilidad; escala construida sobre esas dimensiones
Anidado	Un enfoque está anidado dentro del otro; uno es dominante	Instrumento del enfoque dominante + instrumento del enfoque subordinado integrado en el primero	Recolección: ambos tipos de datos se recogen simultáneamente	Cuestionario con preguntas cerradas + preguntas abiertas integradas en el mismo instrumento

Nota. CUAN = Cuantitativo; CUAL = Cualitativo. Adaptado de Creswell y Plano Clark (2018).

3.5.2 El instrumento integrado: cuestionario con componente cualitativo

El tipo de instrumento mixto más común en la investigación aplicada es el cuestionario que combina ítems de escala cuantitativa con preguntas abiertas estratégicamente ubicadas. Este diseño permite recoger simultáneamente datos cuantificables y narrativas contextuales que enriquecen la interpretación de los primeros.

Las reglas para integrar preguntas abiertas en un cuestionario cuantitativo son:

- Limitar el número de preguntas abiertas: no más de 2-3 en un cuestionario estándar, ya que aumentan significativamente el tiempo de administración y generan fatiga en el respondiente.
- Ubicar las preguntas abiertas estratégicamente: después de escalas o baterías sobre un tema específico, para profundizar o contextualizar las respuestas cuantitativas. Nunca al principio del instrumento.
- Formular las preguntas abiertas de manera que su relación con las escalas precedentes sea evidente para el respondiente: '¿Podría explicar brevemente por qué respondió de esa manera?'
- Diseñar el espacio de respuesta adecuado: en papel, 3-5 líneas para respuestas breves; en digital, un campo de texto libre con un indicador del número de palabras esperado.
- Analizar las preguntas abiertas con técnicas de análisis de contenido o de análisis temático, documentando explícitamente el procedimiento de codificación.

3.5.3 Matriz de Decisión Metodológica

La Matriz de Decisión Metodológica (MDM) es una herramienta original de esta obra diseñada para ayudar al investigador a elegir el tipo de instrumento más adecuado para su propósito, considerando simultáneamente las características del fenómeno de interés, el paradigma de investigación, los recursos disponibles y el contexto de aplicación. Se presenta como una guía de decisión estructurada, no como un algoritmo mecánico:

Tabla 3.16

Matriz de Decisión Metodológica (MDM) para la selección del tipo de instrumento

Variable de decisión	Cuantitativo	Cualitativo	Mixto	Criterio de decisión
¿Qué tan bien se conoce el constructo?	Alto: definición establecida y validada	Bajo: fenómeno poco explorado o mal comprendido	Medio: definición parcial con aspectos no capturados cuantitativamente	El nivel de conocimiento previo determina el enfoque inicial
¿Cuál es el objetivo del estudio?	Medir, comparar, predecir, generalizar	Comprender, explorar, describir en profundidad	Ambos objetivos simultáneos o secuenciales	El objetivo epistemológico es la guía más importante

Variable de decisión	Cuantitativo	Cualitativo	Mixto	Criterio de decisión
¿Qué tamaño muestral se tiene?	Grande (N > 100) para AFC y generalización	Pequeño (N = 5-30 para saturación teórica)	Variable: grande para el componente cuantitativo; pequeño para el cualitativo	El tamaño muestral condiciona la viabilidad de los análisis
¿Cuánto tiempo y recursos tiene?	Menor tiempo de análisis si el instrumento está validado	Mayor tiempo de análisis; no se puede automatizar	El más exigente en tiempo y recursos	Los recursos disponibles pueden condicionar la factibilidad
¿Qué nivel educativo tiene la población?	Requiere comprensión lectora suficiente	Se adapta mejor a poblaciones con baja escolaridad	Requiere adaptación del componente cuantitativo	La accesibilidad del instrumento es condición de validez
¿Qué tan sensible es el tema?	Riesgo de deseabilidad social en temas muy sensibles	Permite explorar temas sensibles con más profundidad y contexto	Equilibra la comodidad de las escalas anónimas con la profundidad cualitativa	La sensibilidad del tema afecta la calidad de los datos
¿Se necesitan comparaciones entre grupos?	Ideal: las escalas permiten comparaciones estadísticas precisas	Difícil: las comparaciones son más interpretativas que estadísticas	Posible si el componente cuantitativo está diseñado para comparar	La necesidad de comparación estadística favorece el enfoque cuantitativo

Nota. MDM = Matriz de Decisión Metodológica. Herramienta original del autor. Elaboración propia.

3.6 Validez de contenido: el juicio de expertos

La validez de contenido es la primera evidencia de validez que debe acumularse para cualquier instrumento cuantitativo, y su evaluación mediante juicio de expertos es la estrategia metodológica estándar reconocida internacionalmente. Un instrumento que no ha pasado por una evaluación sistemática de validez de contenido por expertos calificados carece de uno de sus pilares fundamentales de legitimidad científica.

3.6.1 Selección y perfil de los expertos

La selección de los expertos que evaluarán el contenido del instrumento es una decisión metodológica crítica que debe documentarse con el mismo rigor que cualquier otra decisión del proceso. Los criterios de selección deben ser explícitos y coherentes con el propósito del instrumento:

- Expertise en el constructo: los expertos deben tener conocimiento especializado del constructo que el instrumento pretende medir, demostrado por su producción académica, su experiencia profesional o ambas.
- Experiencia en instrumentación: al menos algunos de los expertos deben tener experiencia en el desarrollo o evaluación de instrumentos de investigación, para que puedan evaluar los aspectos técnicos de los ítems además de su pertinencia conceptual.
- Diversidad: un panel homogéneo puede introducir sesgos sistemáticos. La diversidad de perspectivas disciplinares, culturales y contextuales enriquece la evaluación.
- Independencia: los expertos no deben tener relaciones de dependencia con el investigador que puedan comprometer la objetividad de sus evaluaciones.

Respecto al número de expertos, Escobar-Pérez y Cuervo-Martínez (2008) recomiendan entre 3 y 10 expertos para la mayoría de los estudios, aunque estudios de instrumentación de alto impacto suelen usar entre 7 y 15. Un número menor de expertos hace que el acuerdo sea más sensible a las idiosincrasias individuales; un número mayor aumenta la estabilidad pero incrementa los costos y la complejidad de la coordinación.

3.6.2 El coeficiente V de Aiken: cálculo detallado e interpretación

El coeficiente V de Aiken (Aiken, 1980, 1985) es el índice de validez de contenido más utilizado en la investigación psicométrica hispanohablante. Cuantifica el acuerdo entre los jueces sobre la relevancia, representatividad o claridad de cada ítem en una escala que generalmente va de 1 (nada relevante) a 4 o 5 (totalmente relevante). Su fórmula general es:

$$V = S / [n \times (c - 1)] \quad (\text{Ecuación 3.8})$$

Donde S es la suma de (valoración del juez - valor mínimo de la escala) para todos los jueces, n es el número de jueces y c es el número de opciones de la escala de valoración.

Ejemplo de cálculo de V de Aiken

Suponga 7 jueces que evalúan un ítem en una escala de 1 a 4 ($c=4$). Sus valoraciones son: 4, 4, 3, 4, 3, 4, 4. $S = (4-1)+(4-1)+(3-1)+(4-1)+(3-1)+(4-1)+(4-1) = 3+3+2+3+2+3+3 = 19$. $V = 19 / [7 \times (4-1)] = 19 / 21 = 0.905$. Interpretación: $V = 0.905$ indica un acuerdo muy alto entre los jueces sobre la relevancia del ítem. El ítem puede mantenerse en el instrumento. Para este número de jueces, el valor crítico de V para $p < 0.05$ es aproximadamente 0.71 (Aiken, 1985).

Tabla 3.17

Valores críticos del coeficiente V de Aiken para $p < 0.05$ según número de jueces y categorías de la escala

Nº de jueces	c=2 (dicotómica)	c=3 (tres puntos)	c=4 (cuatro puntos)	c=5 (cinco puntos)
3	.67	—	—	—
4	.75	.67	—	—
5	.80	.60	.60	—
6	.83	.67	.58	—
7	.86	.71	.57	.57
8	.88	.63	.56	.56
10	.90	.70	.60	.56
15	.93	.73	.60	.56
20	.95	.75	.65	.60

Nota. — = Sin valor crítico establecido para esa combinación con menos de 5 jueces. Basado en Aiken (1985). Los valores críticos asumen distribución binomial.

3.6.3 Índice de Validez de Contenido (IVC) de Lawshe

El Índice de Validez de Contenido (IVC), propuesto por Lawshe (1975), es un método alternativo al V de Aiken que es especialmente apropiado cuando se quiere evaluar si cada ítem es esencial para medir el constructo no solo relevante o pertinente. El procedimiento pide a los jueces clasificar cada ítem en una de tres categorías: esencial, útil pero no esencial, o no necesario. La fórmula del IVC para cada ítem es:

$$IVC = (ne - N/2) / (N/2) \quad (\text{Ecuación 3.9})$$

Donde n_e es el número de expertos que clasificaron el ítem como 'esencial' y N es el número total de expertos. El IVC varía de -1 (todos los jueces lo consideran no necesario) a +1 (todos los jueces lo consideran esencial). Un IVC de 0 indica que el ítem es clasificado como esencial solo por el 50% de los jueces, lo que equivale al azar. Lawshe estableció valores mínimos del IVC según el número de expertos (por ejemplo, $IVC \geq 0.58$ con 10 expertos; $IVC \geq 0.78$ con 8 expertos) para aceptar un ítem como esencial con $p < 0.05$.

3.6.4 Gestión de discrepancias entre expertos

Cuando los expertos no alcanzan un acuerdo aceptable sobre un ítem (V de Aiken $<$ umbral crítico o $IVC <$ umbral de Lawshe), el investigador no tiene la opción de simplemente descartar el ítem si este es teóricamente imprescindible. En su lugar, debe gestionar activamente la discrepancia mediante uno de los siguientes procedimientos:

1. Análisis de comentarios cualitativos: si los expertos han aportado comentarios sobre por qué un ítem les parece inadecuado, estos deben analizarse para identificar si existe un consenso sobre el tipo de problema (ambigüedad, doble carga, sesgo cultural) que permite una reformulación específica.
2. Segunda ronda de evaluación: después de revisar los ítems problemáticos basándose en los comentarios, se puede realizar una segunda ronda con los mismos expertos para evaluar las versiones mejoradas. Este proceso iterativo es la base del método Delphi.
3. Reunión de consenso: para ítems particularmente complejos o disputados, una reunión virtual o presencial con todos los expertos puede ser más efectiva que las rondas independientes por escrito.
4. Eliminación justificada: si después de la revisión el ítem sigue sin alcanzar el acuerdo mínimo y existe un ítem alternativo que cubra el mismo indicador, la eliminación puede ser la opción más prudente.

3.7 Prueba cognitiva y pilotaje

3.7.1 Prueba cognitiva: el pensamiento en voz alta

La prueba cognitiva es el procedimiento mediante el cual el investigador verifica, antes del pilotaje formal, que los ítems del instrumento son interpretados por los respondientes de la manera que el investigador pretendía. Es un paso metodológico frecuentemente omitido por razones de tiempo o recursos, pero su impacto en la calidad final del instrumento es desproporcionadamente alto en relación con el esfuerzo que requiere.

La técnica del pensamiento en voz alta (think-aloud) es el método más utilizado en pruebas cognitivas. Se pide al participante que responda el cuestionario verbalizando en voz alta todo lo que piensa mientras lo hace: cómo interpreta cada pregunta, qué información necesita para responder, qué dudas le generan las opciones de respuesta. El investigador observa y registra estas verbalizaciones sin interrumpir, salvo para solicitar más elaboración cuando el participante guarda silencio.

Tabla 3.18

Protocolo de prueba cognitiva para instrumentos de investigación

Paso	Actividad	Procedimiento detallado	Criterios de éxito
1	Selección de participantes	5-10 participantes representativos de la población objetivo. No deben ser conocidos del investigador. Deben cubrir la diversidad de la población (edad, educación, contexto)	Participantes con el rango completo de niveles educativos de la población objetivo
2	Entrenamiento en el método	Explicar al participante el procedimiento: 'Vamos a pedirle que responda este cuestionario en voz alta, diciendo todo lo que piensa mientras lo hace'	El participante entiende que debe verbalizar su proceso de pensamiento, no solo dar la respuesta
3	Práctica con ítems de entrenamiento	2-3 ítems de práctica sobre un tema neutro (sin relación con el instrumento real) con retroalimentación del investigador	El participante verbaliza espontáneamente sin necesidad de prompts constantes
4	Aplicación al instrumento real	El participante responde el instrumento real verbalizando en voz alta. El investigador registra y puede hacer preguntas sonda cuando el participante se detiene	Grabación de audio; notas del investigador sobre momentos de dificultad o confusión
5	Entrevista retrospectiva	Al terminar, el investigador hace preguntas específicas sobre ítems que resultaron problemáticos: '¿Qué quiso decir con...?' '¿Podría explicar cómo llegó a esa respuesta?'	Todas las dudas identificadas durante el pensamiento en voz alta son exploradas
6	Análisis y revisión	Analizar las grabaciones o notas; identificar patrones de confusión; revisar los ítems problemáticos	< 20% de participantes con confusión sistemática en el mismo ítem; tiempo total de respuesta dentro del rango planificado

Nota. Elaboración propia basada en Willis (2005) y Beatty & Willis (2007).

3.7.2 El pilotaje: análisis y criterios de decisión

El pilotaje es la aplicación del instrumento a una muestra representativa de la población objetivo antes de su uso definitivo, con el propósito de evaluar sus propiedades psicométricas preliminares e identificar ítems problemáticos que deben ser revisados o eliminados.

Tamaño muestral para el pilotaje

La determinación del tamaño muestral para el pilotaje depende del tipo de análisis que se realizará y del número de ítems del instrumento. Los criterios más citados en la literatura son:

- Para análisis factorial exploratorio (AFE): mínimo de 100 participantes; se recomienda entre 5 y 10 participantes por ítem. Con 30 ítems, se recomendaría una muestra de 150-300 participantes.
- Para análisis de confiabilidad solamente: 50-100 participantes son suficientes para obtener estimaciones preliminares del Alpha de Cronbach.
- Para instrumentos con AFC planificado: 200-400 participantes para el pilotaje, reservando una muestra mayor para el estudio de validación definitivo.

Análisis de ítems en el pilotaje

Los análisis estadísticos principales que deben realizarse con los datos del pilotaje son:

1. Estadísticos descriptivos por ítem: media, desviación estándar, asimetría y curtosis. Valores de asimetría > 2 o curtosis > 7 en valor absoluto sugieren distribuciones problemáticas que pueden comprometer el AFC.
2. Efecto techo y suelo: si más del 80% de los respondientes se concentra en el valor máximo (efecto techo) o mínimo (efecto suelo) de un ítem, el ítem no discrimina adecuadamente y debe revisarse.
3. Correlación ítem-total corregida (r_{it}): se calcula la correlación entre cada ítem y la puntuación total del instrumento excluyendo ese ítem. $r_{it} < 0.30$ en instrumentos de actitudes o < 0.20 en pruebas de rendimiento es indicador de que el ítem debe revisarse.
4. Alpha de Cronbach si el ítem se elimina: para cada ítem, se calcula el Alpha que resultaría si ese ítem fuera eliminado. Ítems cuya eliminación aumenta sustancialmente el Alpha son candidatos a revisión.
5. Análisis Factorial Exploratorio (si el modelo dimensional no está completamente especificado a priori): con extracción por máxima verosimilitud o mínimos cuadrados

ordinarios y rotación oblicua (promax o oblimin) dado que las dimensiones del constructo típicamente correlacionan entre sí.

Tabla 3.19

Criterios de decisión para la revisión o eliminación de ítems en el pilotaje

Indicador	Valor problemático	Acción recomendada	Prioridad
Correlación ítem-total (r_{it})	< 0.20	Revisión profunda o eliminación; verificar si el ítem está midiendo la dimensión correcta	Alta
Correlación ítem-total (r_{it})	$0.20 - 0.29$	Revisión de la redacción; si la revisión no mejora, considerar eliminación	Media
Efecto techo/suelo	$> 80\%$ en un extremo	Revisar la dificultad del ítem; considerar ajustar la redacción o la escala de respuesta	Alta
Asimetría	Valor absoluto > 2	Revisar la formulación; posible efecto de deseabilidad social o de dificultad extrema	Media
Carga factorial	< 0.30 en AFC	Revisar la asignación dimensional; posible ítem con constructo-irrelevante	Alta
Carga factorial cruzada	> 0.30 en dos factores simultáneamente	El ítem tiene doble carga; revisar la redacción para mayor especificidad	Alta
Alpha si se elimina	Eliminar aumenta Alpha > 0.05	Considerar eliminación si no compromete la	Media

Indicador	Valor problemático	Acción recomendada	Prioridad
		cobertura del constructo	

Nota. Elaboración propia. Los criterios son orientativos; la decisión final debe combinar los criterios estadísticos con el juicio teórico sobre la importancia del ítem para el constructo.

3.8 Análisis psicométrico avanzado

3.8.1 Análisis Factorial Confirmatorio (AFC)

El Análisis Factorial Confirmatorio (AFC) es la herramienta estadística más importante para evaluar la validez de constructo de un instrumento cuantitativo. A diferencia del Análisis Factorial Exploratorio (AFE), que descubre la estructura factorial a partir de los datos, el AFC *confirma* si la estructura factorial especificada a priori por el modelo teórico es consistente con los datos observados. Por esta razón, el AFC es epistemológicamente superior al AFE para la validación de instrumentos: exige que el investigador haya especificado sus hipótesis teóricas antes de ver los datos, protegiéndose del ajuste post-hoc.

El modelo de AFC especifica tres tipos de relaciones:

- Relaciones ítem → factor: cuáles ítems cargan en cuál factor (dimensión). Las cargas factoriales no especificadas son fijadas en cero.
- Relaciones factor ↔ factor: si los factores están correlacionados (modelo oblicuo) o son ortogonales (modelo ortogonal). Para la mayoría de los constructos psicológicos y sociales, los factores están correlacionados.
- Errores de medición: varianza de cada ítem no explicada por su factor. Por defecto, se asume que los errores no están correlacionados entre sí (supuesto de independencia local).

Índices de ajuste del AFC y sus criterios de interpretación

Tabla 3.20

Índices de ajuste del Análisis Factorial Confirmatorio: definición, fórmulas y criterios de interpretación

Índice	Descripción	Fórmula/Cálculo	Criterio de buen ajuste	Sensibilidad
χ^2 (Chi-cuadrado)	Prueba de bondad de ajuste exacto; H_0 : el modelo reproduce perfectamente la matriz de covarianza	Diferencia entre matriz observada y reproducida	$p > .05$ (rara vez alcanzado); usado en ratio χ^2/gl	Muy sensible al tamaño muestral; $N > 200$ casi siempre produce χ^2 significativo
χ^2/gl	Razón chi-cuadrado/grados de libertad; corrige parcialmente la sensibilidad al N	χ^2 dividido entre sus grados de libertad	< 3.0 (excelente); < 5.0 (aceptable)	Menos sensible al N que χ^2 solo
CFI	Comparative Fit Index; compara el modelo propuesto con el modelo nulo	$\text{CFI} = 1 - (\chi^2_{\text{modelo}} / \chi^2_{\text{nulo}})$	$\geq .95$ (excelente); $\geq .90$ (aceptable)	Robusto al tamaño muestral; uno de los más recomendados
TLI	Tucker-Lewis Index; similar al CFI pero penaliza la complejidad del modelo	Función de χ^2 del modelo y del modelo nulo	$\geq .95$ (excelente); $\geq .90$ (aceptable)	Ligeramente más conservador que el CFI
RMSEA	Root Mean Square Error of Approximation; error de aproximación por grado de libertad	$\sqrt{[(\chi^2 - \text{gl}) / (\text{gl} \times (N-1))]$	$\leq .05$ (excelente); $\leq .08$ (aceptable); $> .10$ (pobre)	Penaliza la complejidad; recomendado con su intervalo de confianza
SRMR	Standardized Root Mean Square Residual; promedio de residuales estandarizados	Media de los residuales estandarizados al cuadrado	$\leq .08$ (aceptable); $\leq .05$ (excelente)	Sensible a errores de especificación en las cargas; menos a errores en covarianzas

Nota. Basado en Hu y Bentler (1999), Kline (2016) y Brown (2015). Se recomienda reportar siempre χ^2/gl , CFI, TLI, RMSEA con IC90% y SRMR.

3.8.2 Análisis de confiabilidad: Alpha, Omega y más allá

Como se discutió en el Capítulo 1, la confiabilidad del instrumento debe evaluarse con múltiples indicadores. En el contexto del análisis psicométrico del pilotaje y del estudio de validación, el procedimiento recomendado es el siguiente:

1. Calcular el Alpha de Cronbach (α) para cada dimensión y para el instrumento total (si este tiene una puntuación total con sentido teórico).
2. Calcular el Omega de McDonald (ω) a partir de las cargas factoriales del AFC. En R, el paquete `semTools::reliability()` lo calcula directamente desde el objeto de AFC.
3. Si α y ω difieren en más de 0.05, preferir el ω como estimado más preciso e investigar la causa de la discrepancia (probablemente cargas factoriales heterogéneas).
4. Calcular el error estándar de medición (SEM) para cada dimensión para poder interpretar la imprecisión de las puntuaciones individuales.
5. Si el instrumento será usado en estudios longitudinales o de evaluación de cambio, evaluar también la confiabilidad test-retest con la correlación intraclase (CCI).

$$\omega = (\sum \lambda_i)^2 / [(\sum \lambda_i)^2 + \sum \varepsilon_i] \quad \text{donde } \lambda_i \text{ son las cargas factoriales y } \varepsilon_i \text{ los errores} \quad (\text{Ecuación 3.10})$$

3.8.3 Teoría de Respuesta al Ítem (TRI): introducción práctica

La Teoría de Respuesta al Ítem (TRI) es un enfoque psicométrico alternativo a la TCT que modela la probabilidad de que un respondiente con un nivel determinado del rasgo latente (θ) proporcione una respuesta específica a un ítem. A diferencia de la TCT, la TRI produce estimaciones de los parámetros del ítem (dificultad, discriminación, pseudo-azar) que son independientes de la muestra utilizada, y estimaciones del rasgo del respondiente que son independientes del conjunto específico de ítems administrados.

El modelo más utilizado para datos dicotómicos es el modelo de tres parámetros logísticos (3PL):

$$P(X=1|\theta) = c + (1-c) \times [e^{(Da(\theta-b))} / (1 + e^{(Da(\theta-b))})] \quad (\text{Ecuación 3.11})$$

Donde θ es el nivel del rasgo latente del respondiente, a es el parámetro de discriminación del ítem, b es el parámetro de dificultad (el nivel de θ para el que la probabilidad de respuesta correcta es $0.5 + c/2$), c es el parámetro de pseudo-azar (probabilidad de respuesta correcta por azar), y D es una constante de escala (generalmente 1.702 para que la función logística aproxime la función normal acumulada).

Para datos politómicos como los producidos por escalas Likert, el modelo de TRI más utilizado es el Modelo de Crédito Parcial Generalizado (GGPCM) o el modelo de Respuesta Graduada de Samejima (GRM). El GRM modela la probabilidad de responder en la categoría k o superior:

$$P^*(X \geq k | \theta) = e^{(Da(\theta - bk))} / (1 + e^{(Da(\theta - bk))}) \quad (\text{Ecuación 3.12})$$

La TRI es especialmente valiosa cuando se pretende construir bancos de ítems para evaluación adaptativa, cuando se comparan grupos con diferentes distribuciones del rasgo latente, o cuando se necesita evaluar el funcionamiento diferencial de ítems (DIF) entre grupos.

3.8.4 Análisis de Funcionamiento Diferencial de Ítems (DIF)

El análisis de Funcionamiento Diferencial de Ítems (DIF) es un procedimiento para detectar ítems que funcionan de manera diferente en grupos distintos (hombres/mujeres, distintas etnias, diferentes regiones) incluso cuando los grupos tienen el mismo nivel del rasgo latente. Un ítem con DIF favorece sistemáticamente a un grupo sobre otro, lo que constituye una forma de sesgo que compromete la equidad del instrumento.

Los métodos más utilizados para detectar DIF son:

- Regresión logística: compara la probabilidad de respuesta al ítem entre grupos condicionando al nivel del rasgo latente. Es el método más flexible y puede detectar DIF uniforme (la diferencia es constante a lo largo de todos los niveles del rasgo) y no uniforme (la diferencia varía).
- Mantel-Haenszel: un método no paramétrico que compara las tasas de respuesta en cada valor de la puntuación total entre los grupos. Es robusto y ampliamente implementado en software estadístico.
- Análisis multigrupo en AFC: compara la invarianza de los parámetros del modelo (cargas factoriales, interceptos) entre grupos mediante una secuencia de modelos con restricciones crecientes.

3.9 Herramientas digitales e IA para la construcción y análisis

3.9.1 Software estadístico para análisis psicométrico

Tabla 3.21

Principales herramientas de software para el análisis psicométrico de instrumentos

Software	Tipo	Fortalezas para instrumentación	Análisis disponibles	Acceso
R + lavaan	Software libre	AFC, modelos SEM, TRI, DIF, invarianza de medición; enorme flexibilidad; reproducibilidad	AFC, AFE, Alpha, Omega, TRI, DIF, CCI, SEM	Gratuito
R + psych	Software libre	Alpha, Omega, AFE, análisis de ítems, correlaciones policóricas; interfaz simple	AFE, Alpha, Omega, análisis de ítems	Gratuito
jamovi	Software libre	Interfaz gráfica amigable; módulos para AFC (jAMM) y TRI; ideal para principiantes	AFC, AFE, Alpha, análisis de ítems; TRI con módulos adicionales	Gratuito
SPSS	Propietario	Ampliamente usado; interfaz intuitiva; análisis de ítems; Alpha	AFE, regresión logística para DIF	Licencia comercial
Mplus	Propietario	El más potente para modelos SEM y AFC complejos; manejo de datos ordinales y no normales	AFC, SEM, TRI, análisis de clases latentes, modelos longitudinales	Licencia comercial
Factor	Software libre	Específicamente diseñado para AFE y AFC de escalas psicológicas; análisis paralelo	AFE, AFC, Alpha, análisis de paralelo, DIF básico	Gratuito

Nota. Para investigadores que inician en psicometría, se recomienda comenzar con jamovi por su curva de aprendizaje menor. Para investigación avanzada, R con lavaan es el estándar actual. Elaboración propia.

3.9.2 IA para análisis de datos cualitativos

El análisis de los datos producidos por instrumentos cualitativos (transcripciones de entrevistas, notas de campo, registros de grupos focales) ha sido históricamente la etapa más laboriosa y menos estandarizable del proceso de investigación cualitativa. Las herramientas de IA, y en particular los LLMs con capacidades de procesamiento de texto, están transformando esta etapa de maneras que merecen una reflexión metodológica cuidadosa.

Tabla 3.22

Herramientas de IA para el análisis de datos cualitativos: capacidades, limitaciones y protocolos de uso

Herramienta	Capacidades en análisis cualitativo	Limitaciones críticas	Protocolo de uso recomendado
NVivo con IA	Codificación automática; análisis de sentimientos; consultas con lenguaje natural sobre los datos	La codificación automática requiere supervisión; no comprende el contexto cultural profundo	Usar como primera pasada de codificación; verificar y refinar manualmente todas las categorías
Atlas.ti con AI Assist	Generación de categorías; análisis de co-ocurrencias; mapas conceptuales automáticos	Puede generar categorías superficiales o decontextualizadas	Usar para exploración inicial; el investigador debe construir el sistema de categorías final
LLMs (Claude, GPT-4)	Resumen de entrevistas; identificación de temas emergentes; análisis de sentimientos; comparación entre participantes	Sin memoria entre sesiones; puede perder matices; privacidad de datos sensibles	Nunca subir datos identificables; usar para exploración preliminar; siempre verificar con lectura directa
Whisper (OpenAI)	Transcripción automática de audio con alta precisión en múltiples idiomas	Errores en nombres propios; dificultades con acentos fuertes o ruido ambiental	Verificar y corregir siempre la transcripción automática antes del análisis

Nota. Elaboración propia. Las capacidades de estas herramientas evolucionan rápidamente; verificar las versiones más recientes antes de su uso.

Advertencia ética sobre el uso de IA en datos cualitativos

Los datos cualitativos de investigación especialmente transcripciones de entrevistas sobre temas sensibles como experiencias de violencia, salud mental o situaciones de vulnerabilidad son datos personales que en muchos países están protegidos por regulaciones específicas (GDPR en Europa, leyes nacionales de protección de datos). Subir estos datos a plataformas de IA en la nube puede constituir una violación de las garantías de confidencialidad proporcionadas a los participantes. Siempre verificar las políticas de privacidad de las herramientas utilizadas y obtener asesoramiento del comité de ética institucional antes de usar IA con datos primarios de participantes.

Reflexiones finales del capítulo

Este capítulo ha construido el taller completo de diseño y construcción del instrumento de investigación. Desde las 20 reglas de oro para la redacción de ítems hasta la Teoría de Respuesta al Ítem, desde el diseño de guías de entrevista hasta el análisis de funcionamiento diferencial de ítems, el lector dispone ahora de un conjunto de herramientas técnicas que le permite construir instrumentos de investigación con estándares de calidad internacionales.

Si hay una lección que este capítulo desea dejar en el investigador es la siguiente: la calidad de un instrumento no se evalúa en un momento puntual, sino que se acumula iterativamente a través de múltiples rondas de diseño, revisión, pilotaje y refinamiento. Un primer borrador de ítems, por muy cuidadoso que sea su diseño, es solo el punto de partida de un proceso que incluye el juicio crítico de expertos, la verificación cognitiva con representantes de la población objetivo, el análisis estadístico del pilotaje y el AFC de la validación definitiva.

El siguiente capítulo abandona la perspectiva técnica universal de este taller para adentrarse en las especificidades de cada área del conocimiento. Veremos cómo los principios generales desarrollados en estos tres primeros capítulos se aplican y adaptan en los contextos particulares de la salud, la educación, las ciencias sociales, la gestión del riesgo de desastres, las ciencias ambientales, la administración, la ingeniería y las ciencias jurídicas.

Preguntas de reflexión

1. Usted va a construir un instrumento de 25 ítems para medir el 'sentido de pertenencia comunitaria' en barrios urbanos marginales. Describa detalladamente cómo aplicaría el Protocolo PCUI completo, especificando los tiempos estimados, los recursos necesarios y los criterios de calidad que usaría en cada fase.
2. Analice los siguientes ítems e identifique qué reglas de las 20 reglas de oro violenta cada uno. Para cada ítem deficiente, proponga una versión mejorada: (a) 'Los directivos de mi institución son buenos líderes y nos comunicamos bien'; (b) '¿No cree usted que la gestión del riesgo es importante?'; (c) 'Siempre me siento motivado para aprender cosas nuevas'; (d) 'Mi nivel de autoconciencia metacognitiva en contextos de aprendizaje colaborativo es elevado'.
3. Un colega le muestra los resultados del AFC de su escala de 20 ítems: $\chi^2/df = 4.2$; CFI = 0.88; TLI = 0.86; RMSEA = 0.09 [IC90%: 0.07-0.11]; SRMR = 0.07. Usando los criterios de la Tabla 3.20, evalúe el ajuste del modelo. ¿Qué conclusiones extraería? ¿Qué análisis adicionales realizaría para mejorar el modelo?
4. Diseñe una guía de entrevista semiestructurada de 8 preguntas principales (con sus respectivas preguntas sonda) para explorar las experiencias de los docentes de educación primaria en la integración de la educación en gestión del riesgo de desastres en su práctica pedagógica. Justifique la inclusión de cada pregunta según su función metodológica.
5. Compare el Alpha de Cronbach y el Omega de McDonald como indicadores de confiabilidad: ¿cuándo coinciden y cuándo difieren? Use la Ecuación 1.9 (Alpha) y la Ecuación 3.10 (Omega) para construir un ejemplo numérico con datos ficticios de un AFC de 5 ítems donde las cargas factoriales son heterogéneas (0.40, 0.55, 0.70, 0.80, 0.90) y demuestre que Alpha subestima la confiabilidad en este caso.

CAPÍTULO

4

El Instrumento en su Contexto: Área, Población y Pertinencia Cultural

Hernán Sebastián Salguero Morales

$$X = \sum_{i=1}^n x_i$$

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum s_i^2}{s_T^2} \right)$$



CAPÍTULO 4

El Instrumento en su Contexto: Área, Población y Pertinencia Cultural

El mismo instrumento puede fallar o funcionar dependiendo del dónde y el para quién

Los tres capítulos anteriores han equipado al investigador con un arsenal metodológico de alto nivel: los fundamentos epistemológicos que justifican cada decisión de medición, el protocolo conceptual que garantiza la coherencia entre la teoría y el instrumento, y el taller técnico que convierte esa coherencia en ítems validados y analizados estadísticamente. Sin embargo, este arsenal pierde gran parte de su efectividad si el instrumento resultante ignora el contexto específico en el que será usado.

Un instrumento de medición no existe en el vacío. Existe en un campo disciplinar con sus propias tradiciones, estándares y preguntas prioritarias. Existe para una población con características culturales, lingüísticas, educativas y experienciales que determinan qué preguntas son comprensibles, cuáles son pertinentes y cuáles pueden ser percibidas como amenazantes o irrelevantes. Existe en un contexto institucional, geográfico y social que define qué tipo de datos son posibles de recolectar y con qué recursos.

Este capítulo aborda precisamente esas dimensiones contextuales. Comienza con los principios generales de adaptación cultural y consideraciones por tipo de población, para luego adentrarse en las particularidades de ocho grandes áreas del conocimiento: ciencias de la salud, educación, ciencias sociales y políticas, gestión del riesgo de desastres, ciencias ambientales, ciencias administrativas, ciencias exactas e ingeniería, y ciencias jurídicas. Cada área recibe un tratamiento que incluye sus particularidades epistemológicas, sus estándares de instrumentación, sus instrumentos más representativos y un caso práctico completo que ilustra la aplicación de los principios desarrollados en los capítulos anteriores.

4.1 Por qué el contexto determina el instrumento

4.1.1 La ilusión de la universalidad en la medición

Uno de los supuestos más extendidos y peligrosos en la metodología de la investigación es que un instrumento validado en un contexto puede aplicarse directamente en cualquier otro contexto sin pérdida de validez. Este supuesto, que podríamos denominar la ilusión de la universalidad, ha llevado a la aplicación acrítica de instrumentos desarrollados en países de altos ingresos en poblaciones de América Latina, África, Asia y otras regiones, con

consecuencias que van desde la producción de datos de dudosa calidad hasta la imposición de marcos conceptuales culturalmente inadecuados.

La investigación transcultural ha documentado extensamente que muchos constructos que parecen universales tienen manifestaciones, significados y estructuras dimensionales que varían significativamente entre culturas (Heine & Buchtel, 2009; van de Vijver & Tanzer, 2004). Un instrumento que mide la autoestima como un rasgo estable, positivo e individualizado puede no ser conceptualmente apropiado para culturas donde la identidad se construye relacionamente y donde la expresión de autovaloración elevada es culturalmente inapropiada.

Principio de pertinencia contextual

Un instrumento es pertinente en un contexto cuando: (1) el constructo que mide es relevante y significativo para ese contexto; (2) sus dimensiones e indicadores reflejan adecuadamente las manifestaciones del constructo en esa población específica; (3) su lenguaje y formato son accesibles y culturalmente apropiados; y (4) el proceso de administración es factible en las condiciones reales del contexto. La validez estadística sin pertinencia contextual es una propiedad necesaria pero insuficiente para la calidad del instrumento.

4.1.2 Niveles del contexto que afectan el instrumento

El contexto que afecta la pertinencia de un instrumento opera en múltiples niveles simultáneos:

Tabla 4.1

Niveles del contexto y su impacto en el diseño y la validez del instrumento

Nivel del contexto	Dimensiones relevantes	Impacto en el instrumento	Estrategias de adaptación
Cultural	Valores, normas sociales, concepciones del mundo, tabúes	Algunos constructos pueden no existir o tener estructura diferente; ciertos ítems pueden ser ofensivos o carentes de sentido	Revisión transcultural; grupos focales de adaptación; juicio de expertos culturales
Lingüístico	Idioma, dialectos, niveles de alfabetización, vocabulario disponible	Ítems incomprensibles o ambiguos para la población; pérdida de equivalencia semántica en traducción	Proceso de traducción-retrotraducción; prueba cognitiva con la población objetivo

Nivel del contexto	Dimensiones relevantes	Impacto en el instrumento	Estrategias de adaptación
Socioeconómico	Nivel de ingresos, acceso a recursos, experiencias de vida relacionadas con el constructo	Indicadores que asumen recursos o experiencias no disponibles para todos los grupos	Revisión de presupuestos implícitos en los ítems; indicadores relativos al contexto
Educativo	Nivel de escolaridad, familiaridad con formatos de test, habilidades lectoras	Ítems demasiado complejos; formatos de respuesta no familiares; sobrecarga cognitiva	Simplificación del lenguaje; prueba cognitiva; formatos alternativos (oral, visual)
Disciplinar	Tradiciones teóricas, estándares metodológicos, preguntas prioritarias del campo	Instrumentos de otras disciplinas pueden no ser apropiados sin adaptación conceptual	Revisión por expertos del campo; verificación de coherencia con teorías disciplinares
Institucional	Contexto organizacional, relaciones de poder, políticas institucionales	El contexto institucional puede sesgar las respuestas (deseabilidad social institucional)	Garantías de anonimato; administración externa; análisis de sesgo de contexto

Nota. Elaboración propia.

4.1.3 Adaptación cultural y lingüística de instrumentos

Cuando el investigador decide adaptar un instrumento existente a un nuevo contexto cultural o lingüístico, debe seguir un proceso sistemático que va mucho más allá de la traducción literal. La adaptación cultural busca la equivalencia entre el instrumento original y el adaptado en múltiples dimensiones:

Tabla 4.2

Tipos de equivalencia en la adaptación transcultural de instrumentos

Tipo de equivalencia	Descripción	Método de evaluación	Consecuencia si no se alcanza
Conceptual	El constructo tiene el mismo significado y relevancia en ambas culturas	Revisión teórica; grupos focales comparativos; entrevistas con informantes clave	El instrumento mide constructos diferentes en distintos contextos

Tipo de equivalencia	Descripción	Método de evaluación	Consecuencia si no se alcanza
Semántica	Los ítems tienen el mismo significado en ambas versiones lingüísticas	Traducción-retrotraducción; comité de revisión bilingüe; prueba cognitiva	Los ítems son interpretados de manera diferente en distintos grupos
Técnica	El formato, instrucciones y procedimientos son apropiados para ambos contextos	Prueba piloto en ambos contextos; evaluación de tasas de respuesta y errores	Diferencias en las puntuaciones reflejan diferencias en el formato, no en el constructo
De criterio	Los ítems son igualmente relevantes para el constructo en ambas culturas	Juicio de expertos en ambos contextos; análisis de DIF	Algunos ítems tienen diferente peso en el constructo según el grupo cultural
De escala (métrica)	Las propiedades métricas del instrumento son equivalentes en ambos grupos	Análisis de invarianza de medición en AFC multigrupo	Las comparaciones cuantitativas entre grupos son inválidas

Nota. Adaptado de van de Vijver y Tanzer (2004) y Hambleton et al. (2005).

El proceso de traducción-retrotraducción

El proceso estándar de adaptación lingüística de instrumentos es la traducción-retrotraducción, que incluye los siguientes pasos:

1. Traducción directa: dos traductores bilingües independientes traducen el instrumento del idioma original al idioma objetivo.
2. Síntesis de traducciones: un comité compuesto por los traductores y el investigador principal compara las dos traducciones y elabora una versión de síntesis que resuelve las discrepancias.
3. Retrotraducción: dos traductores bilingües diferentes retrotraducen la versión de síntesis al idioma original.
4. Revisión de discrepancias: el comité compara la retrotraducción con el original para identificar ítems cuyo significado se ha alterado en el proceso. Estos ítems deben ser reformulados y el proceso puede necesitar repetirse.

5. Revisión por comité de expertos: un panel de expertos en el constructo, bilingüe y con conocimiento del contexto cultural objetivo, realiza una revisión final de la versión adaptada.
6. Prueba cognitiva con la población objetivo: verifica que los ítems son comprensibles e interpretados como se pretende por miembros reales de la población objetivo.

4.1.4 Consideraciones por tipo de población

Tabla 4.3

Adaptaciones metodológicas para poblaciones específicas en el diseño de instrumentos

Población	Desafíos específicos	Adaptaciones al instrumento	Adaptaciones al proceso
Niños y adolescentes	Vocabulario limitado; capacidad atencional reducida; pensamiento concreto; influencia del contexto social	Lenguaje simple y concreto; ilustraciones; escala de respuesta visual (caritas); ítems breves; máximo 15-20 ítems	Administración en grupos pequeños; instrucciones verbales; tiempo ilimitado; ambiente no amenazante
Adultos mayores	Velocidad de procesamiento reducida; posible deterioro cognitivo; menor familiaridad con tecnología; fatiga	Fuente grande; alto contraste visual; sin límite de tiempo; formato simple; pocas opciones de respuesta	Administración individual o en pares; verificar comprensión; sesiones cortas; formato en papel
Baja escolaridad	Vocabulario limitado; posible analfabetismo funcional; desconfianza ante situaciones formales de evaluación	Lenguaje muy simple; escala visual o verbal (oral); ítems muy cortos; sin jerga técnica	Administración oral por entrevistador entrenado; ambiente informal; explicación detallada del propósito
Comunidades indígenas	Cosmovisión diferente; conceptos sin equivalente directo; desconfianza histórica ante investigadores externos; lenguas propias	Proceso de co-construcción con la comunidad; indicadores culturalmente apropiados; posible instrumento completamente diferente al original	Consentimiento colectivo; investigadores de la comunidad; devolución de resultados; protocolo de propiedad intelectual comunitaria

Población	Desafíos específicos	Adaptaciones al instrumento	Adaptaciones al proceso
Expertos y profesionales	Familiaridad con los constructos puede producir respuestas sofisticadas pero también deseabilidad profesional	Ítems más técnicos y específicos; escalas con mayor granularidad; incluir situaciones hipotéticas complejas	Formato autoaplicado; garantías de anonimato profesional; tiempo flexible
Personas con discapacidad	Diversidad de tipos de discapacidad requiere diferentes adaptaciones; riesgo de invalidación de la experiencia	Formato accesible (braille, audio, lectura fácil); sin preguntas que asuman capacidades específicas	Tecnologías asistivas; entrevistadores entrenados en comunicación accesible; tiempo adicional

Nota. Elaboración propia.

4.2 Instrumentos en Ciencias de la Salud

4.2.1 Particularidades epistemológicas de la medición en salud

Las ciencias de la salud presentan un panorama de instrumentación de extraordinaria riqueza y complejidad. Desde las escalas clínicas de valoración rápida usadas en urgencias hasta los sofisticados instrumentos de calidad de vida relacionada con la salud (CVRS) que capturan la experiencia subjetiva del paciente, pasando por los cuestionarios epidemiológicos que caracterizan la distribución de enfermedades en poblaciones, la medición en salud abarca una diversidad de propósitos, poblaciones y estándares que no tiene paralelo en otras disciplinas.

Lo que distingue a la instrumentación en salud de la instrumentación en otras ciencias sociales es la naturaleza de las consecuencias de sus usos. Los instrumentos de salud frecuentemente informan decisiones clínicas (diagnósticos, prescripciones de tratamiento, criterios de alta) que afectan directamente el bienestar y la vida de las personas. Esta realidad impone estándares de validez, confiabilidad y responsividad más exigentes que en otras áreas, y ha llevado al desarrollo de marcos regulatorios específicos como los estándares COSMIN y las guías de la FDA para instrumentos de resultado informado por el paciente que no tienen equivalente en otras disciplinas.

4.2.2 Tipos de instrumentos en ciencias de la salud

Tabla 4.4

Tipología de instrumentos de investigación en ciencias de la salud

Tipo	Propósito	Ejemplos paradigmáticos	Estándares específicos
Escalas clínicas	Evaluación rápida de estado clínico para toma de decisiones	APACHE II (gravedad en UCI), PHQ-9 (depresión), MMSE (deterioro cognitivo), Escala de Glasgow	Precisión en punto de corte; sensibilidad y especificidad documentadas; reproducibilidad interjueces
Instrumentos de CVRS genéricos	Medir el impacto de la salud en la calidad de vida independientemente de la condición	SF-36, EQ-5D, WHOQOL-BREF	Normas poblacionales disponibles; validación en múltiples países; responsividad documentada
Instrumentos de CVRS específicos	Medir CVRS en una condición clínica particular	EORTC QLQ-C30 (oncología), AQLQ (asma), KOOS (rodilla)	Alta sensibilidad al cambio clínico; mínima diferencia clínicamente importante (MCID) establecida
PROs (Patient-Reported Outcomes)	Capturar la perspectiva del paciente sobre síntomas, funcionalidad y bienestar sin interpretación del clínico	PROMIS, FACT-G, BPI (dolor)	Guías FDA/EMA para PROs en ensayos clínicos; enfoque en la voz del paciente
Instrumentos epidemiológicos	Caracterizar la distribución y determinantes de enfermedades en poblaciones	AUDIT (alcohol), CAGE (dependencia), Cuestionario de Rose (cardiopatía isquémica)	Operatividad en grandes muestras; valores de corte con base en estudios de criterio
Escalas de riesgo clínico	Predecir la probabilidad de un evento clínico adverso	SCORE (riesgo cardiovascular), Wells (TVP), FRAX (fractura osteoporótica)	Calibración y discriminación en muestras de desarrollo y validación externa; actualización periódica

Nota. CVRS = Calidad de Vida Relacionada con la Salud; PROs = Patient-Reported Outcomes; UCI = Unidad de Cuidados Intensivos; MCID = Mínima Diferencia Clínicamente Importante. Elaboración propia.

4.2.3 Estándares COSMIN para instrumentos de salud

El proyecto COSMIN (CONsensus-based Standards for the selection of health Measurement INstruments) ha desarrollado el estándar internacional más influyente para evaluar la calidad metodológica de los instrumentos de medición en salud (Mokkink et al., 2010, 2018). El modelo de calidad COSMIN evalúa las propiedades de medición en diez dominios organizados en tres grandes categorías:

Tabla 4.5

Dominios de calidad del modelo COSMIN y criterios de evaluación

Categoría COSMIN	Dominio	Descripción	Criterio mínimo aceptable
Validez	Validez de contenido	¿Los ítems son relevantes, comprensivos y comprensibles?	Proceso de desarrollo que incluye input de pacientes y expertos
Validez	Validez estructural	¿El AFC confirma la estructura del instrumento?	$CFI \geq 0.95$; $RMSEA \leq 0.06$
Validez	Invarianza de medición	¿El instrumento funciona igual en subgrupos relevantes?	Invarianza métrica confirmada (cargas factoriales invariantes)
Validez	Validez de criterio	¿Las puntuaciones se correlacionan con el estándar de referencia?	$r \geq 0.70$ con criterio externo relevante
Validez	Validez de constructo (convergente / discriminante)	¿Se cumplen las hipótesis sobre correlaciones con otros instrumentos?	$\geq 75\%$ de las hipótesis a priori confirmadas
Fiabilidad	Confiabilidad	¿Las puntuaciones son estables en condiciones estables?	$CCI \geq 0.70$ para uso grupal; ≥ 0.90 para uso individual
Fiabilidad	Error de medición	¿El error es aceptablemente pequeño en relación al cambio esperado?	$SEM < MCID$ (Mínima Diferencia Clínicamente Importante)
Responsividad	Responsividad	¿El instrumento detecta cambios clínicamente relevantes?	$AUC \geq 0.70$ para cambio vs. no cambio; tamaño del efecto documentado

Nota. CCI = Coeficiente de Correlación Intraclass; SEM = Error Estándar de Medición; AUC = Área Bajo la Curva ROC. Adaptado de Mokkink et al. (2018).

4.2.4 La Mínima Diferencia Clínicamente Importante (MCID)

La Mínima Diferencia Clínicamente Importante (MCID) es el concepto más relevante en la evaluación de la responsividad de los instrumentos de salud. Se define como el menor cambio en la puntuación del instrumento que los pacientes perciben como beneficioso o perjudicial y que justificaría, en ausencia de efectos adversos y costos excesivos, un cambio en el manejo clínico del paciente (Jaeschke et al., 1989).

Existen dos enfoques principales para calcular el MCID:

Métodos basados en distribución:

$$\text{MCID}_{\text{distribución}} = 0.5 \times \text{DE}_{\text{basal}} \quad \text{o} \quad \text{MCID} = 1 \text{ SEM} \quad (\text{Ecuación 4.1})$$

Donde DE_{basal} es la desviación estándar de las puntuaciones en la línea de base y SEM es el error estándar de medición. Estos métodos son puramente estadísticos y no requieren un criterio externo, pero tampoco capturan necesariamente la perspectiva del paciente.

Métodos basados en anclaje:

Se correlacionan los cambios en la puntuación del instrumento con respuestas a una pregunta de anclaje global de cambio percibido (p.ej. “En comparación con antes del tratamiento, ¿cómo se siente ahora? Mucho mejor / Algo mejor / Igual / Algo peor / Mucho peor”). El MCID se identifica como el cambio promedio en los pacientes que respondieron “algo mejor”.

4.2.5 Caso práctico: Escala de Percepción de Calidad de Atención en Emergencias (EPCAE)

Este caso práctico ilustra el proceso completo de desarrollo de un instrumento PRO para medir la percepción de los pacientes sobre la calidad de la atención recibida en servicios de urgencias hospitalarias en América Latina.

Tabla 4.6

Ficha técnica de la Escala de Percepción de Calidad de Atención en Emergencias (EPCAE)

Componente	Descripción
Nombre del instrumento	Escala de Percepción de Calidad de Atención en Emergencias (EPCAE)
Constructo	Percepción del paciente sobre la calidad de la atención recibida en el servicio de urgencias hospitalarias
Definición conceptual	Evaluación subjetiva del paciente sobre el grado en que la atención recibida en urgencias cumplió sus expectativas y

Componente	Descripción
	necesidades en términos de trato, información, tiempo de espera, efectividad y entorno físico
Marco teórico	Modelo de calidad percibida de servicio (Parasuraman et al., 1988); Modelo de Donabedian (estructura, proceso, resultado)
Dimensiones	D1: Trato y comunicación (7 ítems); D2: Tiempo de espera y eficiencia (5 ítems); D3: Información y educación al paciente (5 ítems); D4: Entorno y confort (4 ítems); D5: Percepción de resultado (4 ítems)
Total de ítems	25 ítems en escala Likert de 5 puntos (1=Muy insatisfecho a 5=Muy satisfecho)
Tiempo de administración	8-12 minutos; autoaplicado o por entrevistador
Población objetivo	Pacientes adultos (≥ 18 años) atendidos en urgencias; no aplicable en estado crítico o con deterioro cognitivo
Proceso de validación	Juicio de 9 expertos (V de Aiken ≥ 0.80 en todos los ítems); prueba cognitiva con 8 pacientes; pilotaje $n=180$; AFC: CFI=0.96, RMSEA=0.052; $\alpha=0.91$; CCI=0.88
MCID estimado	4.2 puntos en la puntuación total (método de anclaje con pregunta de satisfacción global)

Nota. Caso práctico ilustrativo desarrollado con fines pedagógicos. Elaboración propia.

Ejemplo de ítems por dimensión:

- D1 — Trato: “El personal de urgencias me trató con respeto y dignidad durante toda mi atención”
- D2 — Eficiencia: “El tiempo que esperé para ser atendido fue razonable considerando mi condición”
- D3 — Información: “Me explicaron claramente el diagnóstico y el tratamiento que recibiría”
- D4 — Entorno: “El área donde esperé y fui atendido estaba limpia y en condiciones adecuadas”
- D5 — Resultado: “Salí del servicio de urgencias sintiéndome mejor que cuando llegué”

4.3 Instrumentos en Ciencias de la Educación

4.3.1 La medición en educación: diversidad de propósitos

Las ciencias de la educación presentan una de las tradiciones de instrumentación más antiguas y diversas de todas las ciencias sociales. Desde las primeras pruebas de inteligencia de Binet y Simon (1905) hasta los actuales sistemas de evaluación a gran escala como PISA y TERCE, pasando por los instrumentos de evaluación formativa en el aula y las escalas de clima escolar, la educación ha desarrollado un ecosistema de instrumentos de medición que refleja la enorme diversidad de sus preguntas de investigación y sus propósitos prácticos.

La instrumentación en educación enfrenta un desafío particular: la misma medición puede tener propósitos radicalmente diferentes como diagnóstico, certificación, investigación, rendición de cuentas, que requieren propiedades psicométricas y estándares de calidad distintos. Un instrumento de diagnóstico formativo puede tolerar niveles de confiabilidad que serían inaceptables en una prueba de certificación con consecuencias de alto impacto.

4.3.2 Instrumentos de evaluación del aprendizaje

Pruebas criterioles versus normativas

Tabla 4.7

Diferencias entre pruebas criterioles y normativas: propósito, diseño e interpretación

Dimensión	Prueba criterial (referida al criterio)	Prueba normativa (referida a la norma)
Propósito	Determinar si el evaluado alcanzó un nivel de desempeño definido (estándar)	Comparar el desempeño del evaluado con el de un grupo de referencia (norma)
Interpretación	Absoluta: “¿Qué sabe y puede hacer este estudiante?”	Relativa: “¿Cómo se desempeña este estudiante comparado con sus pares?”
Punto de corte	Establecido a priori en función del dominio requerido	Establecido post-hoc en función de la distribución de la muestra normativa
Diseño de ítems	Los ítems representan el dominio a certificar; la dificultad es irrelevante si el ítem es válido	Los ítems se seleccionan para maximizar la varianza; ítems muy fáciles o muy difíciles son descartados

Dimensión	Prueba criterial (referida al criterio)	Prueba normativa (referida a la norma)
Uso más apropiado	Evaluación de competencias; certificaciones; screenings clínicos; evaluación de programas	Selección y clasificación; diagnóstico diferencial; investigación sobre diferencias individuales
Ejemplo paradigmático	Prueba de conducción de vehículos; examen de habilitación profesional; evaluación de lectoescritura	Test de inteligencia; prueba de aptitud escolar; evaluación de rendimiento académico comparativo

Nota. Elaboración propia.

4.3.3 Escalas de clima y convivencia escolar

El clima escolar como el conjunto de percepciones, actitudes y experiencias que caracterizan el ambiente de aprendizaje de una institución educativa es uno de los constructos más investigados en educación, con una evidencia sólida que lo vincula con el rendimiento académico, la salud mental de estudiantes y docentes, y la reducción del acoso escolar. Su medición ha dado lugar a una amplia familia de instrumentos que difieren en sus modelos conceptuales, sus poblaciones objetivo y sus dimensiones.

Tabla 4.8

Dimensiones frecuentes en instrumentos de clima escolar y sus indicadores

Dimensión	Descripción	Indicadores típicos	Fuente de perspectiva
Relaciones interpersonales	Calidad de las relaciones entre estudiantes y con docentes	Respeto, apoyo emocional, pertenencia, colaboración	Estudiantes y docentes
Seguridad y orden	Percepción de seguridad física y emocional; normas claras y justas	Acoso, violencia, claridad de normas, equidad disciplinaria	Estudiantes, docentes y familias
Enseñanza y aprendizaje	Calidad pedagógica percibida; motivación y compromiso académico	Efectividad docente, relevancia del contenido, expectativas del docente	Estudiantes
Entorno institucional	Condiciones físicas y recursos; liderazgo	Infraestructura, recursos, gestión	Docentes, directivos y familias

Dimensión	Descripción	Indicadores típicos	Fuente de perspectiva
	directivo; participación	participativa, comunicación	
Bienestar socioemocional	Estado emocional y mental de los miembros de la comunidad educativa	Satisfacción, autoestima, ansiedad, sentido de pertenencia	Estudiantes y docentes

Nota. Elaboración propia basada en revisión de instrumentos de clima escolar (Cohen et al., 2009; Wang & Degol, 2016).

4.3.4 Escalas de motivación y autorregulación del aprendizaje

La motivación académica y la autorregulación del aprendizaje son dos de los constructos más estudiados en psicología educativa, con implicaciones directas para el diseño de intervenciones pedagógicas. Sus instrumentos de medición se sustentan principalmente en dos marcos teóricos que han generado la mayor producción de instrumentos validados:

- Teoría de la Autodeterminación (Deci & Ryan, 1985): distingue entre motivación intrínseca, extrínseca y amotivación. El instrumento más representativo es la Escala de Motivación Académica (EMA/AMS) de Vallerand et al. (1992), con versiones validadas en español para múltiples poblaciones.
- Modelo de aprendizaje autorregulado (Zimmerman, 2000): distingue entre la fase de planificación, ejecución y autoevaluación del aprendizaje. El inventario más utilizado es el MSLQ (Motivated Strategies for Learning Questionnaire) de Pintrich et al. (1991), aunque su validez factorial en muestras latinoamericanas ha sido cuestionada y requiere verificación antes de su uso.

4.3.5 Caso práctico: Instrumento de Evaluación de Competencias en Lectura Comprensiva (IELC)

Tabla 4.9

Tabla de especificaciones del Instrumento de Evaluación de Competencias en Lectura Comprensiva para estudiantes de educación básica (grados 4-6)

Dimensión	Marco teórico	Indicadores	Nº ítems	Nivel Bloom	Formato	Peso %
Comprensión literal	Modelo de situación (Kintsch, 1998)	Identificación de ideas principales; localización de información explícita	8	Recordar	OM	25%
Comprensión inferencial	Procesamiento profundo	Inferencias lógicas; relaciones causales; información implícita	10	Comprender/ Analizar	OM	31%
Comprensión crítica	Teoría de la argumentación	Identificación de posición del autor; evaluación de argumentos; detección de sesgo	8	Evaluar	OM+ Abierta	25%
Comprensión intertextual	Literacidad múltiple	Relaciones entre textos; transferencia de información entre formatos	6	Crear/Analizar	Abierta	19%
TOTAL			32 ítems			100%

Nota. OM = Opción múltiple de 4 alternativas; Abierta = Respuesta abierta corta (3-5 líneas). Elaboración propia con fines ilustrativos.

4.4 Instrumentos en Ciencias Sociales y Políticas

4.4.1 Particularidades de la medición en ciencias sociales

Las ciencias sociales y políticas presentan desafíos de medición singulares que derivan de la naturaleza de sus objetos de estudio: la acción colectiva, las estructuras de poder, las identidades sociales, las percepciones ciudadanas y los procesos institucionales. A diferencia de las ciencias de la salud, donde existe un referente biológico relativamente estable para muchos constructos, en ciencias sociales los constructos son esencialmente construcciones históricas y culturales que pueden cambiar de significado entre contextos y períodos.

4.4.2 Índices sociales compuestos

Los índices sociales compuestos son instrumentos de medición que combinan múltiples indicadores en una puntuación única para representar constructos sociales multidimensionales como el desarrollo humano, la cohesión social, la pobreza multidimensional o la calidad democrática. Su diseño y construcción requiere decisiones explícitas sobre la selección de indicadores, la normalización de escalas, la ponderación de componentes y la agregación en la puntuación final.

Tabla 4.10

Metodología general para la construcción de índices sociales compuestos

Paso	Etapas	Procedimiento	Consideraciones críticas
1	Marco teórico	Definir el constructo y sus dimensiones con base teórica sólida	El modelo reflectivo vs. formativo determina toda la metodología posterior
2	Selección de indicadores	Identificar indicadores disponibles que reflejen cada dimensión; evaluar calidad de los datos fuente	Disponibilidad y comparabilidad de datos entre unidades de análisis
3	Imputación de datos	Tratar los valores faltantes con técnicas apropiadas (imputación múltiple, media condicional)	La imputación inapropiada puede introducir sesgos sistemáticos
4	Análisis de valores atípicos	Identificar y tratar los valores extremos que distorsionan la distribución	La decisión sobre valores atípicos debe ser teóricamente justificada

Paso	Etapa	Procedimiento	Consideraciones críticas
5	Normalización	Transformar todos los indicadores a una escala común (Z-scores, min-max, rango)	La elección del método de normalización afecta la distribución del índice
6	Ponderación	Asignar pesos a los indicadores y dimensiones según criterio teórico o estadístico	Pesos iguales implican igual importancia; verificar robustez a cambios en pesos
7	Agregación	Calcular la puntuación del índice (suma ponderada, media geométrica, media armónica)	La media geométrica penaliza desequilibrios extremos entre dimensiones
8	Análisis de incertidumbre	Evaluar la robustez del índice a cambios en los supuestos (análisis de sensibilidad)	Un índice robusto mantiene sus resultados ante variaciones razonables en supuestos

Nota. Adaptado de OCDE (2008). Handbook on Constructing Composite Indicators.

4.4.3 Instrumentos de percepción ciudadana y opinión pública

Las encuestas de percepción ciudadana y opinión pública son una de las herramientas de investigación social más influyentes, con aplicaciones que van desde la investigación académica sobre actitudes y valores hasta los sondeos electorales con impacto en el debate público. Su diseño requiere una comprensión profunda tanto de las técnicas de construcción de ítems (discutidas en el Capítulo 3) como de los principios de diseño muestral que garantizan la representatividad de los resultados.

Las principales consideraciones metodológicas específicas de las encuestas de opinión pública son:

- El efecto de formulación (framing): la manera en que se formula una pregunta de opinión puede influir más en las respuestas que el contenido mismo de la pregunta. “Apoya usted la asistencia económica a familias en situación de pobreza” y “¿Apoya usted que el gobierno gaste dinero de los impuestos en asistencia social?” pueden medir la misma actitud, pero producir respuestas muy diferentes.
- El orden de las preguntas: las preguntas anteriores crean un contexto de interpretación que influye en las respuestas a las preguntas posteriores. Los instrumentos de opinión pública deben diseñarse con especial cuidado al orden y a los posibles efectos de arrastre entre ítems.
- La estabilidad de las opiniones: muchas actitudes políticas son más inestables de lo que parecen: pequeñas variaciones en la formulación, el momento o el contexto de la

encuesta pueden producir respuestas diferentes. Esto tiene implicaciones directas para la interpretación de las medidas de confiabilidad.

4.4.4 Caso práctico: Índice de Percepción de Confianza Institucional (IPCI)

Tabla 4.11

Estructura del Índice de Percepción de Confianza Institucional (IPCI)

Dimensión	Institución referente	Ítems de ejemplo	Escala	Peso
D1: Confianza en el Estado	Gobierno nacional, poder legislativo, poder judicial	“¿Cuánto confía usted en que el gobierno nacional actúa en beneficio de todos los ciudadanos?”	Likert 1-5	25%
D2: Confianza en fuerzas de seguridad	Policía, fuerzas armadas, ministerio público	“¿Qué tan confiable le parece la actuación de la policía en su barrio o comunidad?”	Likert 1-5	20%
D3: Confianza en medios de comunicación	TV, radio, prensa escrita, medios digitales	“¿Con qué frecuencia verifica la información que recibe de los medios de comunicación?”	Frecuencia 1-5	20%
D4: Confianza en la sociedad civil	ONG, iglesias, asociaciones de vecinos, gremios	“¿Qué tan efectivas considera que son las organizaciones de la sociedad civil para resolver problemas comunitarios?”	Likert 1-5	20%
D5: Confianza interpersonal	Vecinos, conocidos, desconocidos	“En general, ¿diría usted que se puede confiar en la mayoría de las personas?”	Dicotómica + seguimiento	15%
TOTAL		30 ítems totales		100%

Nota. IPCI = Instrumento de medición ilustrativo. Elaboración propia.

4.5 Instrumentos en Gestión del Riesgo de Desastres

4.5.1 Marco conceptual de la instrumentación en GRD

La Gestión del Riesgo de Desastres (GRD) es un campo interdisciplinario que integra conocimientos de las ciencias naturales, la ingeniería, las ciencias sociales, la economía y las políticas públicas. Esta naturaleza interdisciplinaria se refleja en la diversidad de sus instrumentos de medición, que van desde modelos geofísicos de amenaza hasta escalas de percepción del riesgo y sistemas de indicadores de capacidad institucional.

El Marco de Sendai para la Reducción del Riesgo de Desastres 2015-2030 (UNDRR, 2015) ha establecido un sistema de metas e indicadores globales que ha influido directamente en la conceptualización de los constructos que se miden en GRD. Los cuatro objetivos prioritarios del Marco de Sendai como la reducción de la mortalidad, reducción del número de afectados, reducción de pérdidas económicas y reducción de daños a infraestructura proporcionan el marco normativo de referencia para la instrumentación en este campo.

Tabla 4.12

Constructos principales en GRD, sus definiciones operacionales y marcos teóricos de referencia

Constructo	Definición operacional en GRD	Marco teórico	Tipo de instrumento típico
Amenaza	Evento o proceso físico potencialmente dañoso que puede causar pérdidas de vida, lesiones, daños a bienes, perturbaciones sociales y económicas o degradación ambiental	Ciencias naturales: geología, hidrometeorología, sismología	Modelos de peligro; mapas de amenaza; registros históricos; sensores
Vulnerabilidad	Condiciones determinadas por factores físicos, sociales, económicos y ambientales que aumentan la susceptibilidad de una comunidad al impacto de una amenaza	Marco de medios de vida sostenibles; teoría del riesgo social	Índices compuestos (formativos); encuestas de condiciones socioeconómicas
Exposición	Personas, medios de subsistencia, especies,	Geografía del riesgo; análisis espacial	SIG; censos; registros catastrales; teledetección

Constructo	Definición operacional en GRD	Marco teórico	Tipo de instrumento típico
	recursos y sistemas que se encuentran en zonas de peligro y que pueden verse afectados por una amenaza		
Capacidad	Combinación de todos los puntos fuertes, atributos y recursos disponibles en una comunidad que pueden usarse para conseguir los objetivos acordados de la GRD	Teoría del capital comunitario; Marco de Sendai	Escalas de autoevaluación institucional; checklists; índices de capacidad
Resiliencia	Capacidad de un sistema expuesto a amenazas para resistir, absorber, adaptarse, transformarse y recuperarse de sus efectos de manera oportuna y eficiente	Teoría de sistemas adaptativos complejos; ecología comunitaria	Índices de resiliencia; escalas de percepción; indicadores de recuperación
Percepción del riesgo	Evaluación subjetiva de la probabilidad e impacto de una amenaza, influida por factores cognitivos, afectivos y culturales	Psicología del riesgo; teoría cultural del riesgo	Escalas de actitudes; diferencial semántico; entrevistas

Nota. GRD = Gestión del Riesgo de Desastres; SIG = Sistema de Información Geográfica. Adaptado de UNDRR (2017) y Wisner et al. (2004).

4.5.2 Instrumentos de evaluación de vulnerabilidad social

La evaluación de la vulnerabilidad social es uno de los procesos más complejos y debatidos en la instrumentación para GRD. Su complejidad deriva de la naturaleza multidimensional y formativa del constructo (la vulnerabilidad no es causada por la combinación de sus dimensiones, sino que es la combinación de esas dimensiones), de la diversidad de enfoques teóricos disponibles y de las implicaciones políticas de la manera en que se definen y ponderan sus componentes.

Tabla 4.13

Dimensiones más frecuentes en índices de vulnerabilidad social ante desastres y sus indicadores

Dimensión	Descripción	Indicadores frecuentes	Fuente de datos
Exposición física	Ubicación geográfica en zonas de amenaza	Distancia a zonas inundables; pendiente del terreno; tipo de suelo	SIG; cartografía de amenazas; datos catastrales
Capital humano	Características de la población que afectan su capacidad de respuesta	Escolaridad; edad (niños y adultos mayores); salud; género	Censos; encuestas de hogares
Capital económico	Recursos financieros y materiales para absorber y recuperarse del impacto	Pobreza; desempleo; tenencia de la tierra; tipo de vivienda	Encuestas de ingresos; registros catastrales
Capital social	Redes de apoyo, confianza y reciprocidad comunitaria	Participación en organizaciones; confianza en vecinos; redes de apoyo	Encuestas comunitarias; grupos focales
Capital institucional	Capacidad de las instituciones locales para gestionar el riesgo	Existencia de planes de emergencia; presupuesto para GRD; coordinación institucional	Evaluaciones institucionales; entrevistas a funcionarios
Percepción y conocimiento	Nivel de conciencia del riesgo y conocimiento de cómo actuar	Conocimiento de amenazas locales; percepción del riesgo personal; conocimiento de rutas de evacuación	Encuestas de percepción; pruebas de conocimiento

Nota. SIG = Sistema de Información Geográfica. Elaboración propia basada en Cutter et al. (2003) y Birkmann (2006).

4.5.3 Caso práctico completo: Instrumento de Vulnerabilidad Comunitaria ante Sismos (IVCS)

El siguiente caso práctico presenta el desarrollo completo de un instrumento para evaluar la vulnerabilidad social de comunidades urbanas ante amenazas sísmicas, diseñado para ser aplicado en ciudades de alta sismicidad en países de ingreso medio de América Latina.

Tabla 4.14

Ficha técnica y tabla de especificaciones del Instrumento de Vulnerabilidad Comunitaria ante Sismos (IVCS)

Componente	Detalle
Nombre	Instrumento de Vulnerabilidad Comunitaria ante Sismos (IVCS)
Propósito	Evaluar el nivel de vulnerabilidad social de comunidades urbanas para priorizar intervenciones de reducción del riesgo sísmico
Constructo	Vulnerabilidad social comunitaria ante sismos — modelo formativo de 5 dimensiones
Unidad de análisis	Comunidad / barrio (no individuo); respondiente es representante informado del barrio
Dimensiones y peso	D1: Exposición física (25%); D2: Características socioeconómicas (25%); D3: Capital social (20%); D4: Capacidad institucional local (15%); D5: Conocimiento y preparación (15%)
Total de ítems	35 ítems: combinación de escalas de valoración, preguntas dicotómicas y de frecuencia
Tiempo de administración	45-60 minutos por entrevista a informante clave
Aplicador	Entrevistador entrenado en GRD; puede ser técnico municipal o investigador
Puntuación	Índice compuesto 0-100: 0-25 = Vulnerabilidad baja; 26-50 = Media; 51-75 = Alta; 76-100 = Muy alta

Nota. IVCS = Caso práctico ilustrativo. Elaboración propia.

Ítems representativos por dimensión del IVCS:

- D1 (Exposición): “¿Qué porcentaje aproximado de las viviendas del barrio fueron construidas antes del año 1980 (antes de los códigos de construcción sismorresistente vigentes)?” [Respuesta: % estimado, convertido a escala 0-4]
- D2 (Socioeconómico): “¿Cuál es el porcentaje aproximado de hogares del barrio con necesidades básicas insatisfechas?” [Respuesta: % estimado por quintiles]

- D3 (Capital social): “¿Existe alguna organización comunitaria activa que haya coordinado acciones colectivas en los últimos 12 meses?” [Sí=0 puntos de vulnerabilidad; No=4 puntos]
- D4 (Institucional): “¿El barrio cuenta con un plan de emergencia sismorresistente actualizado y conocido por la comunidad?” [Sí, actualizado y difundido=0; Sí, pero desactualizado=2; No existe=4]
- D5 (Conocimiento): “¿Qué porcentaje de los residentes del barrio, en su estimación, conoce el punto de encuentro designado para evacuación en caso de sismo?” [Escala 0-4 según porcentaje estimado]

Nota sobre el modelo formativo del IVCS

En este instrumento, las dimensiones no son manifestaciones de la vulnerabilidad sino contribuyentes a ella. Por eso el modelo es formativo: eliminar la dimensión D3 (capital social) cambiaría el constructo que se está midiendo, no solo reduciría la confiabilidad. Los análisis de consistencia interna (Alpha) no son apropiados para este instrumento; en cambio, se debe verificar la robustez del índice mediante análisis de sensibilidad a cambios en los pesos y a la inclusión o exclusión de indicadores.

4.6 Instrumentos en Ciencias Ambientales y Sostenibilidad

4.6.1 La medición en el campo ambiental: naturaleza híbrida

Las ciencias ambientales presentan una naturaleza epistemológica híbrida: combinan la medición objetiva de parámetros biofísicos con la medición de construcciones psicosociales como la percepción ambiental, las actitudes hacia la sostenibilidad y los comportamientos proambientales. Esta dualidad requiere que el investigador ambiental sea competente en la instrumentación tanto de las ciencias naturales como de las ciencias sociales.

4.6.2 Escalas de percepción y comportamiento ambiental

La medición de actitudes y comportamientos proambientales ha generado una familia de instrumentos con distintos enfoques teóricos. Los más influyentes son:

- Escala del Nuevo Paradigma Ambiental (NEP, Dunlap et al., 2000): mide las creencias sobre la relación entre la humanidad y el medioambiente. Con 15 ítems en 5 dimensiones (límites del crecimiento, antiantropocentrismo, fragilidad del equilibrio natural, rechazo del excepcionalismo humano, posibilidad de una crisis ecológica). Es

el instrumento más usado en psicología ambiental con más de 1000 estudios de validación en diferentes culturas.

- Escala de Comportamiento Ambiental Responsable (Kaiser et al., 1999): basada en la Teoría del Comportamiento Planificado, mide la intensidad con que los individuos realizan comportamientos favorables al medioambiente. Distingue entre comportamientos de bajo umbral (reciclar) y alto umbral (activismo ambiental).
- Escala de Identidad Ambiental (Clayton, 2003): mide el grado en que el medioambiente forma parte de la identidad personal del individuo. Alta identidad ambiental predice comportamientos proambientales con mayor robustez que las escalas de actitudes.

4.6.3 Indicadores de sostenibilidad organizacional

La evaluación de la sostenibilidad de organizaciones com empresas, instituciones públicas, universidades ha dado lugar a una familia de marcos de indicadores que combinan dimensiones económicas, sociales y ambientales. Los más utilizados en investigación son:

Tabla 4.15

Marcos de indicadores de sostenibilidad organizacional: comparación de alcance y aplicación

Marco	Dimensiones principales	Aplicación en investigación	Disponibilidad
GRI (Global Reporting Initiative)	Económica, ambiental, social; materialidad	Análisis de reportes de sostenibilidad; benchmarking sectorial	Estándares públicos; libre acceso
ODS Corporativos (Agenda 2030)	17 objetivos; 169 metas; 232 indicadores globales	Alineación de estrategias organizacionales con ODS; evaluación de impacto	Marco ONU; libre acceso
ISO 14001	Gestión ambiental; procesos y sistemas	Evaluación del sistema de gestión ambiental; auditorías de conformidad	Norma ISO; licencia comercial
STARS (Universidades)	Academia, compromiso, operaciones, planificación	Evaluación de sostenibilidad en instituciones de educación superior	Acceso libre para instituciones afiliadas

Nota. GRI = Global Reporting Initiative; ODS = Objetivos de Desarrollo Sostenible; STARS = Sustainability Tracking Assessment & Rating System. Elaboración propia.

4.6.4 Caso práctico: Escala de Conducta Proambiental Comunitaria (ECPAC)

La ECPAC es un instrumento diseñado para medir comportamientos proambientales en el contexto específico de comunidades rurales de América Latina, donde los comportamientos relevantes difieren significativamente de los medidos en escalas desarrolladas en contextos urbanos de países desarrollados.

- D1 — Uso sostenible de recursos naturales (8 ítems): prácticas de uso del agua, suelo, fauna y flora en actividades productivas. Ítems de frecuencia anclados en comportamientos específicos al contexto rural.
- D2 — Gestión de residuos en el hogar y la comunidad (6 ítems): disposición de basura, manejo de agroquímicos, reutilización. Escala Likert de frecuencia.
- D3 — Participación en acciones colectivas ambientales (5 ítems): participación en mingas de limpieza, comités de agua, guardabosques comunitarios. Escala de frecuencia + dicotómica.
- D4 — Transmisión de prácticas ambientales (4 ítems): enseñanza a hijos y nietos; participación en educación ambiental comunitaria. Likert de acuerdo.

Innovación metodológica de la ECPAC

La ECPAC utiliza un proceso de co-construcción con representantes de las comunidades objetivo: los indicadores fueron identificados inicialmente mediante grupos focales con agricultores y líderes comunitarios, y validados posteriormente mediante juicio de expertos. Esto garantiza que los comportamientos medidos son ecológicamente válidos para el contexto y no proyecciones de concepciones urbanas sobre lo que significa ser “proambiental”.

4.7 Instrumentos en Ciencias Administrativas y Organizacionales

4.7.1 La medición en el campo organizacional

Las ciencias administrativas y organizacionales han desarrollado una de las tradiciones psicométricas más productivas en ciencias sociales. Constructos como el liderazgo transformacional, el clima organizacional, el compromiso con la organización, la satisfacción laboral y el burnout cuentan con instrumentos bien validados, aunque con frecuencia desarrollados en contextos organizacionales de países de altos ingresos que requieren verificación de su pertinencia antes de ser aplicados en contextos latinoamericanos, africanos o asiáticos.

4.7.2 Escalas de clima y cultura organizacional

El clima organizacional es uno de los constructos más medidos en psicología organizacional, tanto por su valor predictivo sobre el desempeño como por su utilidad como diagnóstico para intervenciones de gestión del cambio.

Tabla 4.16

Principales instrumentos de clima organizacional: características y evidencias de validez

Instrumento	Autor(es)	Dimensiones	Nº ítems	Limitaciones para contexto latinoamericano
Organizational Climate Questionnaire (OCQ)	Litwin & Stringer (1968)	9: estructura, responsabilidad, recompensa, desafío, relaciones, cooperación, estándares, conflicto, identidad	50 ítems	Desarrollado en EE.UU.; validación en contexto latinoamericano limitada
Work Environment Scale (WES)	Moos & Insel (1974)	10: implicación, cohesión, apoyo, autonomía, tarea, presión, claridad, control, innovación, comodidad	90 ítems	Muy extenso; factor de comodidad física con baja relevancia en algunos contextos
FOCUS-93	Van Muijen et al. (1999)	4 (basadas en Quinn): apoyo, innovación, reglas, objetivos	40 ítems	Marco teórico del CVF puede no reflejar la diversidad cultural en Latinoamérica
Escala de Clima Organizacional EDCO	Acero et al. (2003)	10 dimensiones adaptadas al contexto colombiano	40 ítems	Una de las pocas escalas desarrolladas específicamente para contexto latinoamericano

Nota. CVF = Competing Values Framework. Elaboración propia.

4.7.3 Escalas de liderazgo: el caso del liderazgo transformacional

El liderazgo transformacional (la capacidad del líder para inspirar a sus seguidores a trascender sus intereses personales por el bien de la organización) es el constructo más estudiado en el campo del liderazgo organizacional. El instrumento de referencia es el Multifactor Leadership Questionnaire (MLQ-5X) de Bass y Avolio (1995), que mide tanto el liderazgo transformacional como el transaccional y el laissez-faire.

Sin embargo, su aplicación directa en contextos latinoamericanos ha revelado problemas de validez factorial que sugieren que el modelo de liderazgo transformacional puede tener una estructura dimensional diferente en culturas con alto índice de colectivismo y alta distancia al poder (Hofstede, 1991). Las dimensiones de “consideración individualizada” y “motivación inspiracional” pueden manifestarse de maneras culturalmente específicas que no son capturadas por los ítems originales del MLQ.

4.7.4 Caso práctico: Escala de Compromiso Organizacional en Contexto Latinoamericano (ECOL)

Tabla 4.17

Arquitectura conceptual de la Escala de Compromiso Organizacional en contexto Latinoamericano (ECOL)

Dimensión	Definición	Adaptación cultural	Ejemplo de ítem
Compromiso afectivo	Apego emocional e identificación con la organización	En contextos colectivistas, la lealtad al grupo puede sobreponerse a la lealtad individual a la organización	“Me siento parte de la familia de esta organización”
Compromiso de continuación	Costo percibido de dejar la organización	El concepto de “antigüedad” y las redes de protección social limitada fuera de la organización formal tienen mayor peso en economías menos formalizadas	“Dejar esta organización me generaría muchas dificultades económicas”
Compromiso normativo	Obligación moral de permanecer en la organización	La lealtad como obligación moral tiene una connotación más fuerte en culturas con alta orientación colectivista	“Siento que tengo una deuda con esta organización por todo lo que ha hecho por mí”

Dimensión	Definición	Adaptación cultural	Ejemplo de ítem
Compromiso comunitario (dimensión adicional)	Identificación del trabajador con el impacto social de la organización en su comunidad	Dimensión no contemplada en el modelo original de Allen y Meyer (1991) pero relevante en organizaciones con fuerte arraigo comunitario	“Me enorgullece que esta organización contribuya al desarrollo de mi comunidad”

Nota. La dimensión D4 (Compromiso comunitario) es una propuesta de adaptación cultural del modelo original de Allen & Meyer (1991) para contextos latinoamericanos. Elaboración propia.

4.8 Instrumentos en Ciencias Exactas e Ingeniería

4.8.1 Naturaleza de la instrumentación técnica

En las ciencias exactas y la ingeniería, la instrumentación tiene una naturaleza dual. Por un lado, existe la instrumentación física (sensores, medidores, dispositivos de medición) que produce datos objetivos sobre propiedades del mundo físico: temperatura, presión, resistencia eléctrica, deformación estructural. Por otro lado, existe una instrumentación orientada a la investigación social aplicada a contextos tecnológicos: la evaluación de usabilidad, la experiencia de usuario, la calidad percibida del software, la aceptación de tecnologías o la evaluación del desempeño de equipos de desarrollo.

4.8.2 Instrumentos de evaluación de usabilidad y experiencia de usuario

La evaluación de usabilidad que definida por la norma ISO 9241-11 como el grado en que un producto puede ser utilizado por usuarios específicos para lograr objetivos específicos con efectividad, eficiencia y satisfacción, ha generado una familia de instrumentos que combinan medidas objetivas (tasas de error, tiempos de tarea) con medidas subjetivas (percepción de facilidad de uso, satisfacción).

Tabla 4.18

Instrumentos estándar para la evaluación de usabilidad y experiencia de usuario

Instrumento	Sigla	Nº ítems	Dimensiones	Fortaleza principal
System Usability Scale	SUS	10 ítems Likert 5 puntos	Usabilidad global (unidimensional)	Brevedad; amplio uso; permite comparación entre sistemas; umbral de aceptabilidad bien establecido ($SUS \geq 68$)

Instrumento	Sigla	Nº ítems	Dimensiones	Fortaleza principal
User Experience Questionnaire	UEQ	26 ítems diferencial semántico	6: atractivo, perspicuidad, eficiencia, confiabilidad, estimulación, novedad	Captura dimensiones hedónicas además de pragmáticas; datos normativos disponibles
Technology Acceptance Model	TAM	12-16 ítems Likert	Utilidad percibida, facilidad de uso, intención de uso	Sólida base teórica; predicción de adopción de tecnología
Net Promoter Score	NPS	1 ítem + seguimiento	Lealtad del usuario (unidimensional)	Extremadamente simple; amplio uso en contextos comerciales; limitado valor en investigación

Nota. SUS > 85 = Excelente; 71-85 = Bueno; 52-70 = Regular; < 52 = Deficiente (Bangor et al., 2008). Elaboración propia.

4.8.3 Caso práctico: Protocolo de Evaluación de Sistemas de Alerta Temprana (PESAT)

El siguiente caso ilustra el desarrollo de un instrumento mixto para evaluar la calidad técnica y la efectividad social de los Sistemas de Alerta Temprana (SAT) para desastres naturales, combinando criterios de ingeniería con criterios de percepción y comportamiento comunitario.

Tabla 4.19

Estructura del Protocolo de Evaluación de Sistemas de Alerta Temprana (PESAT)

Componente	Dimensión	Indicadores	Método de recolección
Técnico (Ingeniería)	Cobertura del sistema	% de población en zona de riesgo con acceso al SAT; número y distribución de sensores	Auditoría técnica; SIG
Técnico (Ingeniería)	Tiempo de respuesta	Tiempo entre detección del evento y emisión de alerta; tiempo de diseminación a la comunidad	Registros del sistema; simulacros cronometrados
Técnico (Ingeniería)	Confiabilidad técnica	Tasa de falsas alarmas; tasa de alarmas perdidas; tiempo de inactividad del sistema	Registros históricos del sistema
Social (Percepción)	Comprensión del SAT	¿La comunidad comprende el significado	Prueba de conocimiento + entrevistas

Componente	Dimensión	Indicadores	Método de recolección
		de las alertas? ¿Conoce las acciones a tomar?	
Social (Percepción)	Confianza en el SAT	¿La comunidad confía en las alertas? ¿Ha reaccionado a falsas alarmas anteriores?	Escala de confianza; grupos focales
Social (Comportamiento)	Respuesta ante alertas	¿La comunidad evacúa o toma las acciones recomendadas cuando se emite una alerta?	Registros de simulacros; entrevistas post-evento

Nota. PESAT = Protocolo ilustrativo para evaluación integral de SAT. SAT = Sistema de Alerta Temprana; SIG = Sistema de Información Geográfica. Elaboración propia.

4.9 Instrumentos en Ciencias Jurídicas y Criminológicas

4.9.1 Desafíos específicos de la medición en ciencias jurídicas

Las ciencias jurídicas presentan desafíos de medición singulares que derivan de la naturaleza normativa de su objeto de estudio. El derecho no es un sistema de hechos empíricos sino de normas, interpretaciones y relaciones de poder que no pueden reducirse a variables cuantitativas sin pérdida de su naturaleza esencial. Sin embargo, la investigación jurídica empírica que estudia cómo funciona el derecho en la práctica, no cómo debería funcionar en teoría requiere instrumentos de medición para evaluar la percepción de justicia, el acceso al derecho, la eficacia del sistema penal o las actitudes hacia la norma.

4.9.2 Instrumentos de percepción de justicia y acceso al derecho

La percepción de justicia (el grado en que los ciudadanos consideran que el sistema jurídico funciona de manera equitativa, predecible y respetuosa) es uno de los constructos más relevantes en la investigación socio-jurídica. Instrumentos como el Barómetro de las Américas, el World Justice Project Rule of Law Index o las encuestas de victimización desarrolladas por organismos nacionales e internacionales evalúan estas percepciones con metodologías que combinan indicadores objetivos con medidas subjetivas.

Tabla 4.20

Dimensiones de la percepción de justicia y su operacionalización en instrumentos de investigación socio-jurídica

Dimensión	Descripción del constructo	Indicadores típicos	Fuente de datos
Justicia procedimental	Percepción de que los procedimientos del sistema son justos, transparentes y participativos	Trato respetuoso; explicación de decisiones; oportunidad de ser escuchado; neutralidad	Encuestas de victimización; cuestionarios a usuarios del sistema judicial
Justicia distributiva	Percepción de que los resultados son proporcionales y equitativos	Proporcionalidad de sanciones; igualdad ante la ley; equidad en la asignación de recursos	Encuestas de percepción; análisis de sentencias
Acceso a la justicia	Capacidad efectiva de los ciudadanos para acceder al sistema jurídico formal	Conocimiento de derechos; disponibilidad de asesoría legal; costos; distancia geográfica	Encuestas de necesidades jurídicas; registros institucionales
Confianza en el sistema	Credibilidad percibida de las instituciones de justicia	Corrupción percibida; efectividad; imparcialidad	Encuestas de opinión; barómetros de confianza institucional
Percepción de impunidad	Creencia de que las infracciones a la ley no tienen consecuencias	Tasa de denuncia; percepción de sanción; efectividad policial y judicial	Encuestas de victimización; registros de denuncias

Nota. Elaboración propia basada en Sunshine & Tyler (2003) y WJP Rule of Law Index.

4.9.3 Caso práctico: Escala de Percepción de Impunidad Ciudadana (EPIC)

La impunidad es uno de los factores más consistentemente asociados con la erosión de la confianza institucional y con la perpetuación de ciclos de violencia en sociedades latinoamericanas. Sin embargo, carece de instrumentos de medición de percepción suficientemente validados para permitir comparaciones entre países o evaluaciones de tendencias temporales.

Tabla 4.21*Tabla de especificaciones de la Escala de Percepción de Impunidad Ciudadana (EPIC)*

Dimensión	Descripción	Ejemplo de ítem	Nº ítems	Formato
D1: Impunidad policial	Percepción de que los actos ilegales de la policía no tienen consecuencias	“Con qué frecuencia cree usted que los policías que cometen abusos reciben algún tipo de sanción”	5	Likert 5
D2: Impunidad judicial	Percepción de que el sistema judicial favorece a quienes tienen poder o recursos	“En su opinión, ¿las personas con dinero o influencias reciben sentencias más favorables?”	5	Likert 5
D3: Impunidad en delitos cotidianos	Percepción de que los delitos comunes (robo, violencia) raramente son sancionados	“Si alguien le roba en la calle, ¿qué tan probable cree que es que el ladrón sea capturado y sancionado?”	5	Likert 5
D4: Eficacia de la denuncia	Percepción de que denunciar un delito tiene algún efecto práctico	“¿Vale la pena denunciar un delito ante las autoridades?”	4	Likert 4
D5: Impunidad de élites	Percepción de que los poderosos nunca rinden cuentas por actos de corrupción	“En su opinión, ¿algún político o empresario poderoso ha recibido una sanción justa por corrupción en los últimos años?”	4	Dicotómica + sondeo
TOTAL			23 ítems	—

Nota. EPIC = Escala ilustrativa para fines pedagógicos. Elaboración propia.

Reflexiones finales del capítulo

Este capítulo ha demostrado, a través del recorrido por ocho grandes áreas del conocimiento, que la instrumentación científica no es una actividad genérica y universal: es siempre una práctica disciplinar y contextualmente situada. Los mismos principios epistemológicos y técnicos desarrollados en los capítulos anteriores se aplican de manera diferente en cada área,

adaptándose a sus tradiciones teóricas, sus preguntas prioritarias, sus estándares metodológicos y las características de sus poblaciones.

La lección más importante de este capítulo es que el investigador que trabaja en una disciplina específica no puede ignorar los estándares y las tradiciones de instrumentación de su campo. Construir un instrumento para ciencias de la salud sin conocer los criterios COSMIN, o construir un índice de vulnerabilidad para GRD sin entender la distinción entre modelos reflectivos y formativos, o diseñar una escala de clima organizacional sin conocer las escalas ya existentes y sus limitaciones, es no solo metodológicamente deficiente sino epistemológicamente irresponsable.

El capítulo que cierra esta obra llevará al investigador al último paso del proceso: la revisión crítica de lo que puede salir mal, la verificación de la coherencia y pertinencia del instrumento terminado, y la elaboración del expediente técnico que garantizará la trazabilidad, la reproducibilidad y la legitimidad del instrumento para ser usado, citado y construido por otros investigadores.

Preguntas de reflexión

1. Usted recibe el encargo de adaptar la Escala de Bienestar Psicológico de Ryff (1989), desarrollada en Estados Unidos, para su uso en comunidades rurales indígenas de su país. Describa el proceso completo de adaptación cultural que seguiría, identificando al menos tres dimensiones del contexto que podrían requerir modificaciones sustanciales al instrumento y justificando los criterios de equivalencia que aplicaría.
2. Analice el instrumento IVCS presentado en la sección 4.5.3. Dado que es un instrumento formativo: (a) ¿Qué análisis de confiabilidad serían inapropiados para evaluar su calidad? (b) ¿Qué análisis sí son apropiados? (c) ¿Cómo evaluaría la validez del instrumento? Justifique sus respuestas con los fundamentos teóricos del Capítulo 1.
3. Compare los estándares de instrumentación en ciencias de la salud (COSMIN) con los utilizados en ciencias sociales (APA/AERA/NCME). Identifique al menos tres diferencias sustantivas en los criterios de calidad exigidos y explique por qué estas diferencias son teóricamente justificadas y no arbitrarias.
4. Diseñe la arquitectura conceptual completa (definición conceptual, dimensiones, tabla de especificaciones de 5 ítems por dimensión) para un instrumento que mida la “preparación comunitaria para emergencias sísmicas” en el contexto de un municipio

urbano de tamaño mediano en su país. Justifique sus decisiones de diseño con base en el marco de Sendai y la literatura de GRD.

5. Un equipo de ingeniería le pide que evalúe la usabilidad y la aceptación por parte de los usuarios de un nuevo Sistema de Alerta Temprana basado en una aplicación móvil para comunidades en riesgo de inundación. Diseñe un protocolo de evaluación que combine (a) métricas técnicas objetivas, (b) una escala psicométrica de percepción de usabilidad y (c) una guía de entrevista semiestructurada. Especifique los instrumentos que usaría para cada componente y cómo integraría los tres tipos de datos.

CAPÍTULO

5

Errores Fatales, Coherencia Final y Expediente del Instrumento

Luis Andres Montero Quintana

$$X = \sum_{i=1}^n x_i$$

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum s_i^2}{s_T^2} \right)$$



CAPÍTULO 5

Errores Fatales, Coherencia Final y Expediente del Instrumento

El capítulo que todo investigador necesitaba y ningún libro tenía

Llegamos al capítulo final. Los cuatro capítulos anteriores han dotado al investigador de todos los instrumentos conceptuales y técnicos para construir un instrumento de investigación de alto nivel: los fundamentos epistemológicos que orientan cada decisión, el Protocolo DCOI para la arquitectura conceptual, el Protocolo PCUI para la construcción técnica, y las guías especializadas por área del conocimiento. Sin embargo, el proceso no termina cuando el investigador tiene en sus manos un instrumento con estadísticas psicométricas aceptables.

Este capítulo aborda lo que ningún manual de metodología ha sistematizado con suficiente profundidad: ¿cómo saber si el instrumento que tengo en mis manos es realmente bueno? No bueno en el sentido estadístico —confiable, con índices de ajuste adecuados—, sino bueno en el sentido más profundo: coherente con el problema que pretende resolver, pertinente para la población a la que se dirige, libre de los errores más devastadores que el investigador puede haber cometido sin darse cuenta, y documentado de manera que otro investigador pueda replicarlo, evaluarlo y mejorarlo.

El capítulo se organiza en torno a tres grandes contribuciones originales. La primera es una clasificación sistemática y analítica de los 30 errores fatales más frecuentes y graves en la construcción de instrumentos, organizados por fase del proceso y con análisis forense de casos reales. La segunda es la Matriz de Coherencia Constructo-Dimensión-Ítem (MCCDI), una herramienta original para verificar la coherencia interna del instrumento terminado. La tercera es el Checklist CADI (Calidad, Adecuación y Documentación del Instrumento), con 50 criterios que constituyen el estándar de calidad final que todo instrumento debe alcanzar antes de ser usado en investigación científica.

5.1 Por qué los instrumentos fallan: una tipología del fracaso

5.1.1 La ilusión del instrumento terminado

Existe una creencia ampliamente extendida entre investigadores especialmente quienes se inician en la construcción de instrumentos de que el proceso termina cuando se obtienen estadísticas psicométricas aceptables: un Alpha de Cronbach por encima de 0.70, índices de ajuste del AFC que superan los umbrales convencionales, un coeficiente V de Aiken que alcanza el umbral mínimo para todos los ítems. Esta creencia, que podemos denominar la

ilusión del instrumento terminado, es una de las fuentes más importantes de instrumentos deficientes en la literatura científica.

La ilusión del instrumento terminado nace de una confusión entre dos tipos de calidad radicalmente diferentes: la calidad estadística y la calidad científica. Un instrumento puede tener excelentes propiedades psicométricas y seguir siendo científicamente deficiente si mide el constructo equivocado con gran precisión, si sus ítems son culturalmente discriminatorios, si su proceso de construcción fue opaco e irreproducible, o si el expediente técnico que lo documenta es tan superficial que hace imposible su evaluación crítica.

Principio rector del Capítulo 5

La calidad de un instrumento no es un estado binario que se alcanza en un momento determinado: es un proceso continuo de verificación, revisión y mejora. El instrumento científicamente riguroso es aquel que puede ser examinado críticamente por cualquier investigador informado y cuya calidad puede ser evaluada, cuestionada y mejorada. La opacidad del proceso de construcción no es metodológicamente neutral: es una forma de incompetencia científica.

5.1.2 Consecuencias del instrumento deficiente

Las consecuencias de usar instrumentos deficientes en investigación científica son más graves y más diversas de lo que habitualmente se reconoce en los textos de metodología:

Tabla 5.1

Consecuencias de instrumentos deficientes según tipo de fallo y contexto de uso

Tipo de fallo	Consecuencia para la investigación	Consecuencia social/práctica	Ejemplo paradigmático
Validez de constructo deficiente	Conclusiones que no se refieren al fenómeno que se pretendía estudiar	Políticas o intervenciones diseñadas para resolver el problema equivocado	Instrumentos de “inteligencia” que miden familiaridad cultural
Sesgo cultural	Resultados que reflejan diferencias culturales en el instrumento, no en el constructo	Discriminación sistemática de grupos culturales en procesos de selección o diagnóstico	Pruebas de admisión universitaria con sesgo de clase
Confiabilidad insuficiente	Error de clasificación elevado; hallazgos irreproducible	Decisiones clínicas, educativas o legales	Escala de depresión con baja estabilidad

Tipo de fallo	Consecuencia para la investigación	Consecuencia social/práctica	Ejemplo paradigmático
		basadas en información poco confiable	temporal usadas en diagnóstico
Documentación deficiente	Imposibilidad de replicar, actualizar o adaptar el instrumento	Fragmentación del campo; acumulación de instrumentos incomparables	Miles de “nuevas” escalas que miden lo mismo sin saberlo
Errores éticos	Violación de los derechos de los participantes; daño reputacional	Pérdida de confianza de la comunidad en la investigación	Instrumentos que exponen a grupos vulnerables sin protocolos éticos

Nota. Elaboración propia.

5.1.3 Tipología del fracaso: cuatro grandes categorías de error

Los errores en la construcción de instrumentos pueden clasificarse en cuatro grandes categorías que corresponden a diferentes fases del proceso y a diferentes tipos de deficiencia metodológica:

- Errores conceptuales: ocurren en las fases de definición del constructo, revisión de literatura y operacionalización. Son los más graves porque contaminan todo el instrumento desde sus cimientos y no pueden ser corregidos por análisis estadísticos posteriores.
- Errores técnicos: ocurren en las fases de diseño de ítems, diseño del formato, validez de contenido y pilotaje. Son detectables y en muchos casos corregibles, pero solo si el investigador cuenta con los criterios técnicos adecuados.
- Errores contextuales: ocurren cuando el instrumento no está adecuadamente adaptado a la población y el contexto de aplicación. Incluyen los sesgos culturales, lingüísticos, educativos y poblacionales discutidos en el Capítulo 4.
- Errores éticos y de documentación: incluyen la falta de transparencia en el proceso de construcción, la ausencia de consentimiento informado apropiado, la violación de la propiedad intelectual de instrumentos existentes, y el uso acrítico de IA sin supervisión metodológica.

5.2 Los 30 errores fatales clasificados por fase

La siguiente clasificación de los 30 errores más frecuentes y graves en la construcción de instrumentos es el resultado de la revisión sistemática de más de 500 artículos de validación y construcción de instrumentos publicados en revistas indexadas, complementada con la experiencia de revisión como evaluador de revistas científicas y como consultor en proyectos de investigación que utilizan instrumentos de medición. Cada error incluye su descripción, su impacto en la validez y un indicador de gravedad.

Tabla 5.2

Los 30 errores fatales en la construcción de instrumentos: clasificación completa

#	Error	Impacto en la validez	Gravedad
ERRORES CONCEPTUALES (Fases 1-3 del DCOI)			
1	Constructo mal definido o circular: el término que se quiere medir aparece en su propia definición sin añadir información conceptual nueva	Validez de contenido comprometida desde el origen; el instrumento no puede ser evaluado ni replicado porque no hay criterio externo de referencia	● Crítica
2	Confusión entre constructo, variable e indicador: se operacionalizan como constructo lo que en realidad son variables dependientes o consecuencias del constructo	Error de especificación del modelo; los análisis factoriales producirán una estructura diferente a la teóricamente esperada	● Crítica
3	Ignorar instrumentos existentes sin justificación: el investigador construye un instrumento nuevo para medir algo que ya existe en formas validadas sin haber realizado una búsqueda sistemática	Fragmentación del campo; datos incomparables con la literatura existente; desperdicio de recursos	● Alta
4	Revisión de literatura superficial: la definición del constructo se basa en 2-3 fuentes sin explorar el debate teórico existente	La dimensionalidad del instrumento no refleja el estado actual del conocimiento; dimensiones importantes pueden estar ausentes	● Alta
5	Error de modelo reflectivo-formativo: se aplican análisis propios de modelos reflectivos (Alpha, AFC) a constructos que son fundamentalmente formativos	Los resultados psicométricos son incorrectos e interpretables de manera errónea; el instrumento parece deficiente cuando en	● Crítica

#	Error	Impacto en la validez	Gravedad
		realidad el análisis es inadecuado	
6	Sub-representación del constructo: el instrumento solo cubre algunas dimensiones del constructo, dejando otras sin representación	Validez de contenido deficiente; las conclusiones no pueden generalizarse al constructo completo sino solo a las dimensiones medidas	● Alta
7	Tabla de especificaciones inexistente o incompleta: los ítems se distribuyen entre dimensiones sin un plan sistemático previo	Desequilibrio dimensional arbitrario; algunas dimensiones tienen más influencia en la puntuación total sin justificación teórica	● Alta
ERRORES TÉCNICOS EN EL DISEÑO DE ÍTEMS (Fases 1-2 del PCUI)			
8	Ítems de doble carga: un solo ítem pregunta sobre dos aspectos del constructo simultáneamente (“Me siento competente y motivado en mi trabajo”)	El respondiente que está de acuerdo con una parte y en desacuerdo con la otra no tiene respuesta válida; el ítem no puede ser interpretado	● Crítica
9	Ítems inversos problemáticos: la inversión es solo lingüística, no conceptual; o se usan en poblaciones donde producen errores sistemáticos	Factores espurios en el AFC que reflejan el método de redacción; pérdida de datos válidos en respondientes con baja escolaridad	● Alta
10	Redacción ambigua: el ítem admite interpretaciones múltiples igualmente válidas y el respondiente no puede saber cuál es la esperada	El ítem mide variabilidad en la interpretación del lenguaje, no variabilidad en el constructo; baja correlación ítem-total	● Alta
11	Uso de absolutos (“siempre”, “nunca”, “todos”): producen respuestas extremas por la tendencia de los respondientes a evitar afirmaciones absolutas	Distribución sesgada de las respuestas; efecto suelo o techo sistemático; varianza artificial reducida	● Media
12	Lenguaje técnico inaccesible: los ítems usan terminología especializada no familiar para la población objetivo	El ítem mide vocabulario técnico o nivel educativo, no el constructo; mayor varianza no explicada por el factor latente	● Alta

#	Error	Impacto en la validez	Gravedad
13	Número insuficiente de ítems por dimensión: dimensiones con solo 1-2 ítems	Imposibilidad de AFC (problema de identificación del modelo); estimaciones de confiabilidad inestables; baja representatividad del dominio	● Crítica
14	Ítems redundantes excesivos: múltiples ítems que preguntan exactamente lo mismo con mínimas variaciones de redacción	Inflación artificial del Alpha de Cronbach; el instrumento parece más confiable de lo que realmente es; aburrimiento del respondiente	● Media
ERRORES EN LA VALIDACIÓN (Fases 4-7 del PCUI)			
15	Juicio de expertos mal diseñado: los expertos no reciben criterios claros de evaluación, o se mezclan criterios de relevancia, representatividad y claridad en una sola escala	La V de Aiken resultante es una mezcla de propiedades diferentes; no permite identificar qué tipo específico de problema tiene cada ítem	● Alta
16	Expertos no calificados o sesgados: los “expertos” son estudiantes, colegas del mismo grupo de investigación, o personas sin conocimiento específico del constructo	La validez de contenido es simulada, no real; el instrumento tiene el sello de aprobación de personas que no tenían criterio para otorgarlo	● Crítica
17	Omitir la prueba cognitiva: el instrumento pasa directamente del juicio de expertos al pilotaje sin verificar la comprensión real de los ítems	Ítems que los expertos aprueban pero los respondientes malinterpretan; problemas de comprensión detectados solo cuando ya hay datos comprometidos	● Alta
18	Muestra de pilotaje inadecuada: tamaño insuficiente ($N < 100$ para AFC) o muestra no representativa de la población objetivo	Los análisis psicométricos del pilotaje tienen baja precisión; los problemas de los ítems no son detectables estadísticamente	● Alta
19	Uso exclusivo de Alpha de Cronbach: se reporta solo el Alpha sin verificar el supuesto de tau-equivalencia ni calcular el Omega	La confiabilidad puede estar sobreestimada o subestimada; el instrumento parece más o menos	● Media

#	Error	Impacto en la validez	Gravedad
		confiable de lo que realmente es	
20	AFC sin especificación a priori: se realiza AFE y luego se presenta como si fuera AFC confirmando la estructura “encontrada” en los mismos datos	Sesgo de confirmación post-hoc; los índices de ajuste son optimistamente sesgados por la capitalización en el azar	● Crítica
21	Ignorar índices de ajuste deficientes: se reporta que el AFC tuvo ajuste adecuado cuando $CFI < 0.90$ y $RMSEA > 0.10$	El instrumento no tiene la estructura factorial afirmada; las subescalas no son válidas como entidades independientes	● Crítica
22	Sin evidencia de validez de criterio ni convergente: el artículo de validación solo reporta la estructura factorial sin ninguna evidencia de relaciones con variables externas	La validez de constructo está incompleta; no hay evidencia de que las puntuaciones predicen o se relacionan con lo que el modelo teórico predice	● Alta
ERRORES CONTEXTUALES Y CULTURALES			
23	Aplicación sin adaptación cultural: se aplica directamente un instrumento desarrollado en otro contexto cultural sin proceso de adaptación	El instrumento puede medir algo diferente en la nueva población; las comparaciones entre culturas son inválidas	● Crítica
24	Ítems con sesgo de género o etnia no detectado: los ítems asumen experiencias, roles o normas que no son universales	Funcionamiento diferencial de ítems (DIF); el instrumento favorece sistemáticamente a ciertos grupos	● Alta
25	Instrumento inaccesible para la población: nivel de lectura demasiado alto; formato incompatible con baja escolaridad	La varianza en las puntuaciones refleja habilidades lectoras o familiaridad con el formato de test, no el constructo	● Alta
26	Condiciones de administración no estandarizadas: el instrumento se aplica en condiciones muy diferentes (grupal vs. individual, papel vs. digital, en el hogar vs. en la institución) sin verificar la equivalencia de los resultados	Las diferencias en puntuaciones entre grupos pueden reflejar diferencias en las condiciones de administración, no en el constructo	● Media

#	Error	Impacto en la validez	Gravedad
ERRORES ÉTICOS Y DE DOCUMENTACIÓN			
27	Ausencia de consentimiento informado apropiado: los participantes del juicio de expertos, la prueba cognitiva o el pilotaje no firman consentimiento informado o no son informados adecuadamente del uso que se dará a sus respuestas	Violación de los principios éticos de la investigación; potencial rechazo por parte de los comités de ética institucionales	● Crítica
28	Uso de instrumentos protegidos sin permiso: se reproduce o modifica un instrumento protegido por derechos de autor sin obtener el permiso correspondiente del titular	Infracción de derechos de autor; posible retractación de publicaciones; daño reputacional	● Crítica
29	Documentación insuficiente del proceso: el artículo de validación no reporta el proceso de construcción con suficiente detalle para que sea replicable	El instrumento no puede ser evaluado críticamente ni replicado; la comunidad científica no puede construir sobre este trabajo	● Alta
30	Uso acrítico de IA sin supervisión metodológica: se usan LLMs para generar ítems o analizar datos sin verificar la calidad de sus outputs ni declarar su uso	Ítems conceptualmente superficiales o sesgados presentados como validados; falta de transparencia metodológica; riesgo de hallazgos basados en alucinaciones de la IA	● Alta

Nota. Gravedad: ● Crítica = compromete la validez fundamental del instrumento; ● Alta = requiere corrección antes de cualquier uso; ● Media = debe atenderse pero no invalida el instrumento. Elaboración propia.

5.3 Análisis forense de instrumentos reales

5.3.1 Metodología del análisis forense

El análisis forense de instrumentos es un procedimiento metodológico original que consiste en examinar retrospectivamente un instrumento publicado para identificar sus fortalezas y debilidades a la luz de los estándares presentados en esta obra. No se trata de una crítica destructiva (todo instrumento publicado tiene valor en tanto contribución al campo), sino de una lectura crítica que permite aprender de los errores y aciertos de otros investigadores.

El procedimiento de análisis forense incluye los siguientes pasos:

1. Localización y lectura del artículo de construcción y validación del instrumento, incluyendo los artículos de validación posteriores en otros contextos si existen.

2. Aplicación sistemática de la lista de los 30 errores, documentando cuáles están presentes, cuáles están ausentes y cuáles no pueden determinarse con la información disponible.
3. Evaluación de la calidad de la evidencia de validez reportada, usando los criterios COSMIN o los Standards for Educational and Psychological Testing según el campo.
4. Identificación de los puntos fuertes del instrumento que justifican su uso en investigación.
5. Formulación de recomendaciones específicas para futuros usuarios: qué adaptaciones son necesarias, qué estudios de validación adicionales se requieren, en qué contextos el instrumento es apropiado y en cuáles no.

5.3.2 Caso forense 1: escala con fallo conceptual grave

El siguiente caso analiza un instrumento ampliamente utilizado en investigación organizacional latinoamericana que presenta un fallo conceptual que ha pasado desapercibido durante años de uso repetido: la confusión entre el constructo y sus antecedentes.

Descripción del caso

Una escala de “compromiso con la seguridad laboral” publicada en una revista indexada de ciencias administrativas latinoamericana fue utilizada en más de 40 investigaciones posteriores. El artículo de construcción reporta $\text{Alpha} = 0.87$ y AFC con $\text{CFI} = 0.94$. Sin embargo, un análisis de los ítems revela que 8 de los 20 ítems no miden el compromiso con la seguridad, sino las percepciones sobre las condiciones de seguridad del entorno (“En mi empresa existen equipos de protección adecuados”). Estos ítems miden antecedentes del constructo, no el constructo mismo.

Tabla 5.3

Análisis forense de la escala de compromiso con la seguridad laboral: diagnóstico y recomendaciones

Error identificado	Evidencia en el instrumento	Consecuencia para los estudios que lo usaron
Error #2: Confusión antecedentes-constructo	8/20 ítems miden condiciones del entorno, no actitudes del trabajador	Los estudios que reportan correlaciones entre “compromiso con la seguridad” y “accidentabilidad” pueden estar midiendo el efecto de las condiciones del entorno, no del compromiso real

Error identificado	Evidencia en el instrumento	Consecuencia para los estudios que lo usaron
Error #20: AFC post-hoc no declarado	El artículo menciona que se “depuraron” ítems antes del AFC sin especificar cuántos ni con qué criterio	Los índices de ajuste están sesgados por capitalización en el azar; la estructura reportada no es confirmatoria sino exploratoria
Error #22: Sin validez de criterio	El artículo solo reporta estructura factorial y Alpha; no hay correlaciones con accidentabilidad real ni con otros indicadores de seguridad	No es posible evaluar si las puntuaciones predicen el comportamiento real de seguridad que el instrumento afirma medir
Error #29: Documentación insuficiente	El proceso de generación de ítems no está descrito; no hay tabla de especificaciones ni definición operacional del constructo	El instrumento no puede ser modificado ni adaptado por otros investigadores sin repetir todo el proceso desde el principio

Nota. Caso hipotético construido con fines pedagógicos que sintetiza patrones reales observados en la literatura. Elaboración propia.

Recomendación forense: Los investigadores que han usado este instrumento deben revisar sus conclusiones considerando que al menos el 40% de los ítems pueden haber medido condiciones del entorno en lugar del constructo objetivo. Futuros usuarios deben considerar la escala como un punto de partida para un proceso de revisión conceptual y no como un instrumento validado listo para usar.

5.3.3 Caso forense 2: escala con fallo psicométrico grave

Este segundo caso ilustra el error más frecuente en la literatura de validación de instrumentos en ciencias de la salud latinoamericana: la presentación de un AFC confirmatorio que en realidad es exploratorio.

Descripción del caso

Una escala de “calidad de vida en pacientes con diabetes tipo 2” desarrollada en México fue adaptada y “validada” en cinco países diferentes en un período de cuatro años. En todos los estudios, los autores reportan haber realizado un AFC que confirmó la estructura trifactorial del instrumento original. Sin embargo, el análisis de los índices de modificación reportados revela que en cuatro de los cinco estudios se liberaron correlaciones entre errores de medición para alcanzar índices de ajuste aceptables. En el quinto estudio, dos ítems cambiaron de dimensión respecto al modelo original.

El patrón identificado es precisamente el error #20 de nuestra clasificación. Lo que se presenta como un AFC es en realidad una combinación de AFE y modificaciones post-hoc del modelo. La consecuencia práctica es que los cinco estudios afirman haber “confirmado” una estructura que en realidad nunca fue confirmada: fue modificada hasta que los índices de ajuste fueron aceptables en cada muestra específica, lo que significa que el instrumento tiene estructuras factoriales diferentes en cada contexto.

Recomendación forense: Antes de usar cualquier instrumento cuya validación reporte correlaciones liberadas entre errores de medición o reasignación de ítems entre factores, el investigador debe preguntar: ¿el modelo que se está “confirmando” es el mismo que el modelo original, o es un modelo diferente que ha sido ajustado post-hoc? Si la respuesta es lo segundo, el estudio es exploratorio, no confirmatorio, independientemente de cómo lo llamen los autores.

5.3.4 Lecciones del análisis forense: patrones y señales de alerta

Tabla 5.4

Señales de alerta en artículos de validación que sugieren fallos metodológicos graves

Señal de alerta en el artículo	Error probable asociado	Acción recomendada para el lector
“Se eliminaron X ítems para mejorar el ajuste del modelo” sin especificar cuántos ni con qué criterio	Error #20: AFC post-hoc no declarado	Solicitar los datos originales; tratar el estudio como exploratorio
“Se liberaron correlaciones entre errores de X ítems” sin justificación teórica	Error #20: modificación post-hoc del modelo	El modelo confirmado no es el modelo teórico original; la estructura factorial no está confirmada
Alpha muy alto (> 0.95) con muchos ítems (> 20)	Error #14: ítems redundantes	Los ítems son demasiado similares; el instrumento no necesita tantos ítems para medir lo que mide
“Los expertos son investigadores del mismo grupo de investigación del autor”	Error #16: expertos no independientes	La validez de contenido está sesgada; los expertos no pueden ser críticos objetivos
Muestra de validación idéntica o muy similar a la de construcción	Error #18: muestra no independiente	Los índices de ajuste están inflados por el ajuste a la misma muestra; se necesita validación cruzada
“Se encontró una estructura de X factores” sin hipótesis teórica previa	Error #20: AFE presentado como AFC	El estudio es exploratorio; la estructura puede no replicarse en otras muestras

Señal de alerta en el artículo	Error probable asociado	Acción recomendada para el lector
No se reportan índices RMSEA y SRMR, solo chi-cuadrado y CFI	Posible ocultamiento de ajuste deficiente	Solicitar el reporte completo de índices; el ajuste puede ser peor de lo que parece

Nota. Elaboración propia basada en la revisión de la literatura de validación.

5.4 Coherencia interna del instrumento: la prueba definitiva

5.4.1 ¿Qué es la coherencia interna del instrumento?

La coherencia interna del instrumento es la propiedad de que todos los elementos del instrumento están alineados y son mutuamente consistentes: la definición conceptual del constructo, sus dimensiones, los indicadores de cada dimensión, los ítems de cada indicador y los criterios de puntuación forman un sistema coherente en el que cada elemento puede justificarse por su relación con los elementos superiores de la jerarquía.

Mientras que la consistencia interna psicométrica (Alpha, Omega) evalúa si los ítems correlacionan adecuadamente entre sí desde un punto de vista estadístico, la coherencia interna evalúa si los ítems tienen sentido en relación al constructo desde un punto de vista teórico. Un instrumento puede tener alta consistencia interna psicométrica y baja coherencia interna y viceversa.

5.4.2 La Matriz de Coherencia Constructo-Dimensión-Ítem (MCCDI)

La Matriz de Coherencia Constructo-Dimensión-Ítem (MCCDI) es una herramienta original de esta obra que permite verificar sistemáticamente la coherencia interna del instrumento en sus tres niveles de abstracción: constructo, dimensión e ítem. Se completa una vez que el instrumento tiene su versión definitiva y antes de su publicación o uso generalizado.

La MCCDI se construye siguiendo estos pasos:

1. Para cada ítem del instrumento, identificar: (a) la dimensión a la que está asignado; (b) el indicador específico que pretende medir; (c) la definición conceptual de esa dimensión; (d) la definición conceptual del constructo global.
2. Verificar la cadena de coherencia: ¿el ítem captura el indicador? ¿el indicador refleja la dimensión? ¿la dimensión contribuye al constructo? Cada pregunta debe poder responderse afirmativamente con argumentos teóricos explícitos.
3. Identificar ítems huérfanos: ítems que no pueden vincularse claramente a ningún indicador específico en la tabla de especificaciones.

4. Identificar dimensiones sub-representadas: dimensiones con menos de 3 ítems o con indicadores sin ítems correspondientes.
5. Identificar ítems desplazados: ítems que, revisados críticamente, parecen más coherentes con una dimensión diferente a la que se les asignó.

Tabla 5.5

Ejemplo de Matriz de Coherencia Constructo-Dimensión-Ítem para los primeros 10 ítems del IVCS (Capítulo 4)

Ítem	Texto del ítem	Indicador asignado	Dimensión	¿Coherente?
1	¿Qué porcentaje de las viviendas fueron construidas antes de 1980?	Calidad del parque habitacional	D1: Exposición física	☒ Sí
2	¿Cuántos pisos tiene la mayoría de las viviendas del barrio?	Tipo y vulnerabilidad estructural	D1: Exposición física	☒ Sí
3	¿Cuál es el porcentaje aproximado de hogares con NBI?	Nivel de pobreza	D2: Socioeconómico	☒ Sí
4	¿Qué porcentaje de la población activa está desempleada?	Capacidad económica de recuperación	D2: Socioeconómico	☒ Sí
5	¿Existe una organización comunitaria activa en el barrio?	Redes de apoyo y acción colectiva	D3: Capital social	☒ Sí
6	¿Los vecinos se ayudan entre sí en situaciones de emergencia?	Confianza y reciprocidad	D3: Capital social	☒ Sí
7	¿El barrio tiene un plan de emergencia actualizado?	Planificación institucional	D4: Capacidad institucional	☒ Sí
8	¿Existe coordinación entre el barrio y la defensa civil municipal?	Articulación interinstitucional	D4: Capacidad institucional	☒ Sí
9	¿Los residentes conocen qué hacer en caso de sismo?	Conocimiento de protocolos	D5: Conocimiento y preparación	☒ Sí

Ítem	Texto del ítem	Indicador asignado	Dimensión	¿Coherente?
10	¿Cuántos simulacros se han realizado en el barrio en el último año?	Práctica de respuesta	D5: Conocimiento y preparación	⚠ Parcial: mide actividad, no conocimiento

Nota. El ítem 10 muestra un caso de coherencia parcial: mide la frecuencia de simulacros, que es un indicador de preparación institucional (D4) tanto como de preparación comunitaria (D5). Debería reasignarse o reformularse para mayor especificidad. Elaboración propia.

5.4.3 Lista de verificación de coherencia: 30 criterios

Tabla 5.6

Lista de verificación de coherencia interna del instrumento — 30 criterios

#	Criterio de coherencia	Nivel	Resultado
1	La definición conceptual del constructo es inequívoca y no circular	Constructo	☐ Cumple / ☐ No
2	La definición conceptual está anclada en al menos dos fuentes teóricas de referencia	Constructo	☐ Cumple / ☐ No
3	La definición conceptual delimita claramente qué incluye y qué excluye el constructo	Constructo	☐ Cumple / ☐ No
4	El modelo de medición (reflectivo o formativo) está explícitamente declarado y justificado	Constructo	☐ Cumple / ☐ No
5	La tabla de especificaciones está completa y documenta la arquitectura conceptual del instrumento	Constructo	☐ Cumple / ☐ No
6	Cada dimensión tiene su propia definición conceptual	Dimensión	☐ Cumple / ☐ No
7	Las dimensiones son conceptualmente distinguibles; no hay solapamiento sustancial	Dimensión	☐ Cumple / ☐ No
8	Las dimensiones cubren exhaustivamente el dominio del constructo según su definición	Dimensión	☐ Cumple / ☐ No
9	El peso relativo de las dimensiones está justificado teóricamente	Dimensión	☐ Cumple / ☐ No
10	Cada dimensión tiene al menos 3 ítems (mínimo para AFC identificable)	Dimensión	☐ Cumple / ☐ No
11	Cada ítem puede vincularse a un indicador específico de la tabla de especificaciones	Ítem	☐ Cumple / ☐ No
12	No hay ítems huérfanos (ítems sin indicador asignado en la tabla de especificaciones)	Ítem	☐ Cumple / ☐ No
13	No hay ítems con doble carga conceptual (que pregunten sobre dos aspectos simultáneamente)	Ítem	☐ Cumple / ☐ No
14	Cada ítem está asignado a la dimensión más apropiada según su contenido semántico	Ítem	☐ Cumple / ☐ No
15	El nivel de dificultad/complejidad de los ítems es apropiado para la población objetivo	Ítem	☐ Cumple / ☐ No

#	Criterio de coherencia	Nivel	Resultado
16	Los ítems inversos están justificados y funcionan correctamente en la población objetivo	Ítem	☐ Cumple / ☐ No
17	La escala de respuesta es apropiada para el tipo de constructo y la población	Ítem	☐ Cumple / ☐ No
18	Las instrucciones del instrumento son claras, completas y no ambiguas	Formato	☐ Cumple / ☐ No
19	El tiempo de administración es apropiado para la población (no genera fatiga excesiva)	Formato	☐ Cumple / ☐ No
20	El instrumento es accesible para el nivel educativo de la población objetivo	Formato	☐ Cumple / ☐ No
21	La V de Aiken es ≥ 0.70 (o el valor crítico correspondiente) para todos los ítems	Validez	☐ Cumple / ☐ No
22	La prueba cognitiva confirmó que todos los ítems son comprendidos como se pretende	Validez	☐ Cumple / ☐ No
23	Los índices de ajuste del AFC son adecuados en la muestra de validación	Validez	☐ Cumple / ☐ No
24	El coeficiente de confiabilidad (α u ω) es ≥ 0.70 para cada dimensión	Confiabilidad	☐ Cumple / ☐ No
25	Existe al menos un estudio de validez de criterio o convergente publicado	Validez	☐ Cumple / ☐ No
26	Los ítems fueron revisados para detectar sesgos culturales, de género y de clase	Equidad	☐ Cumple / ☐ No
27	No hay indicios de funcionamiento diferencial de ítems entre subgrupos relevantes	Equidad	☐ Cumple / ☐ No
28	El proceso de construcción está documentado con suficiente detalle para ser replicado	Documentación	☐ Cumple / ☐ No
29	El expediente técnico del instrumento está disponible o puede solicitarse a los autores	Documentación	☐ Cumple / ☐ No
30	El instrumento y su proceso de construcción cumplen los principios éticos de la investigación	Ética	☐ Cumple / ☐ No

Nota. Puntuación: 28-30 = Coherencia excelente; 24-27 = Coherencia adecuada con ajustes menores; 18-23 = Requiere revisión sustancial; < 18 = Reconstruir desde la fase deficiente. Elaboración propia.

5.5 Pertinencia del instrumento: más allá de la validez

5.5.1 ¿Qué es la pertinencia y cómo se diferencia de la validez?

La pertinencia es la propiedad que determina si el instrumento es apropiado, relevante y adecuado para el propósito específico y el contexto concreto en que se va a usar. Es una propiedad relacional y complementa a la validez de una manera que no puede sustituir.

La diferencia entre validez y pertinencia puede ilustrarse con un ejemplo concreto: una escala de depresión validada con excelentes propiedades psicométricas en población adulta general puede ser perfectamente válida pero completamente impertinente como herramienta de screening en urgencias pediátricas, porque no fue diseñada para esa población, sus ítems no son apropiados para niños, y sus puntos de corte no tienen relevancia clínica en ese contexto.

Tabla 5.7

Diferencias conceptuales entre validez y pertinencia del instrumento

Dimensión	Validez	Pertinencia
¿Qué evalúa?	Si el instrumento mide lo que dice medir	Si el instrumento es apropiado para el propósito y el contexto específico de uso
¿Es absoluta o relativa?	Relativa al constructo y la población del estudio de validación	Relativa al contexto específico de aplicación; varía con el propósito, la población y las condiciones
¿Cómo se evalúa?	Evidencias psicométricas: AFC, correlaciones, Alpha/Omega	Juicio teórico y contextual; revisión por expertos locales; prueba de campo
¿Puede un instrumento válido no ser pertinente?	Sí: válido en su contexto de origen pero no en el contexto de aplicación	La pertinencia no garantiza la validez, pero la validez sin pertinencia contextual es insuficiente
¿Quién la evalúa?	El investigador mediante análisis estadísticos	El investigador con conocimiento del contexto + representantes de la comunidad objetivo

Nota. Elaboración propia.

5.5.2 Protocolo de verificación de pertinencia

El siguiente protocolo permite al investigador evaluar sistemáticamente la pertinencia de su instrumento antes de su aplicación en un contexto específico:

1. Pertinencia teórica: ¿el modelo conceptual del constructo que subyace al instrumento es apropiado para el contexto? ¿las dimensiones del constructo son igualmente relevantes en este contexto que en el contexto de desarrollo? ¿la definición del constructo adoptada es coherente con las teorías y marcos conceptuales predominantes en el contexto de uso?
2. Pertinencia cultural: ¿los ítems del instrumento son culturalmente apropiados para la población objetivo? ¿hay ítems que asumen experiencias, normas o valores que no son aplicables en este contexto? ¿el lenguaje del instrumento es el que usa naturalmente la población?
3. Pertinencia poblacional: ¿el instrumento fue diseñado (o validado) para una población con características similares a la del contexto de uso en términos de edad, nivel educativo, características socioeconómicas y familiaridad con formatos de evaluación?
4. Pertinencia del propósito: ¿el instrumento fue diseñado para el mismo tipo de propósito que el del uso previsto? Un instrumento diseñado para investigación puede no ser pertinente para decisiones individuales de alto impacto, y viceversa.
5. Pertinencia operacional: ¿las condiciones de administración del instrumento son factibles en el contexto de uso? ¿el tiempo requerido, el formato, los recursos tecnológicos y el perfil del administrador son compatibles con las condiciones reales?

Casos en que la validez no garantiza la pertinencia

(1) Una escala de calidad de vida desarrollada en población oncológica adulta no es pertinente para niños con cáncer, aunque sea psicométricamente excelente. (2) Un índice de vulnerabilidad social construido con datos censales de países de altos ingresos no es pertinente para comunidades rurales de países en desarrollo sin datos censales actualizados. (3) Una prueba de competencias de liderazgo validada en corporaciones multinacionales no es pertinente para evaluar el liderazgo en organizaciones comunitarias de base.

5.6 El expediente técnico del instrumento

5.6.1 ¿Qué es el expediente técnico y por qué es obligatorio?

El expediente técnico del instrumento es el conjunto de documentos que registra, de manera sistemática y permanente, todo el proceso de construcción, validación y estandarización del instrumento, junto con las instrucciones necesarias para su uso adecuado. Es el equivalente

metodológico del expediente clínico de un paciente: sin él, ningún profesional que llegue después puede entender qué se hizo, cómo y por qué.

La ausencia de un expediente técnico completo es uno de los problemas más graves de la instrumentación científica contemporánea. La mayoría de los instrumentos publicados en revistas científicas están respaldados únicamente por el artículo de validación, que por limitaciones de espacio no puede contener todos los detalles del proceso. El expediente técnico es el documento complementario que contiene lo que el artículo no pudo incluir.

El expediente técnico es obligatorio por tres razones fundamentales:

1. Trazabilidad: permite que cualquier investigador que quiera usar, adaptar o mejorar el instrumento conozca exactamente qué decisiones se tomaron, por qué y con qué evidencia.
2. Reproducibilidad: permite que el proceso de construcción pueda ser replicado en otro contexto o verificado críticamente por la comunidad científica.
3. Responsabilidad: documenta que el instrumento fue construido siguiendo procedimientos metodológicamente rigurosos y éticamente apropiados, lo que protege tanto a los investigadores como a los participantes.

5.6.2 Componentes del expediente técnico completo

Tabla 5.8

Componentes del expediente técnico del instrumento: descripción y criterios de calidad

Componente	Descripción	Criterio de calidad	Formato recomendado
1. Ficha técnica	Resumen de las características del instrumento: nombre, propósito, constructo, dimensiones, número de ítems, escala de respuesta, tiempo de administración, población objetivo, autores, año	Completa en una página; actualizable en cada nueva validación	Tabla estructurada
2. Marco teórico del constructo	Fundamentación teórica detallada: definición conceptual con anclaje bibliográfico, modelo de dimensionalidad, revisión de literatura, vacíos identificados	Coherente con la definición operacional; actualizado a la última revisión	Texto narrativo con referencias completas

Componente	Descripción	Criterio de calidad	Formato recomendado
3. Proceso de construcción	Descripción detallada del DCOI y PCUI aplicados: generación de ítems, juicio de expertos (con lista de expertos y sus credenciales), prueba cognitiva, pilotaje	Suficientemente detallado para permitir la replicación; resultados numéricos de cada etapa	Descripción metodológica con tablas de resultados
4. Versiones del instrumento	Registro de todas las versiones del instrumento desde el pool inicial hasta la versión definitiva, con justificación de cada cambio	Versiones numeradas y fechadas; razón de cada modificación documentada	Control de versiones con fecha y descripción
5. Evidencias de validez	Resultados completos de todos los análisis de validez realizados: V de Aiken por ítem, matrices de correlación del AFC, índices de ajuste, coeficientes de confiabilidad, evidencias de validez convergente/discriminante	Datos numéricos completos, no solo interpretaciones; matrices completas, no solo resúmenes	Tablas estadísticas completas con software y versión usados
6. Manual de aplicación	Instrucciones completas para la administración: condiciones requeridas, instrucciones verbales al respondiente, manejo de preguntas durante la aplicación, tiempo recomendado	Suficientemente detallado para que un aplicador no entrenado pueda administrar el instrumento correctamente	Documento independiente con ejemplos
7. Manual de corrección	Instrucciones completas para la corrección: ítems inversos, puntuación por dimensión, puntuación total, puntos de corte si existen, baremos si aplica	Inequívoco; sin ambigüedades sobre cómo calcular las puntuaciones	Documento independiente con ejemplos numéricos
8. Manual de interpretación	Guía para interpretar las puntuaciones: significado de rangos de puntuación, contexto normativo, limitaciones de la interpretación	Calibrado a la población de referencia; actualizado con nuevos estudios	Texto narrativo con tablas de percentiles si hay baremos
9. Consideraciones éticas	Documentación del proceso ético: aprobación institucional, consentimientos informados usados, manejo de datos sensibles, condiciones de uso autorizado	Completo; incluye copia de consentimientos y aprobaciones del comité de ética	Documentos escaneados + texto

Componente	Descripción	Criterio de calidad	Formato recomendado
10. Condiciones de uso	Licencia de uso: qué usos están autorizados, cuáles requieren permiso explícito, cómo citar el instrumento, cómo contactar a los autores para permisos	Clara e inequívoca; compatible con los repositorios donde se depositará el instrumento	Declaración de licencia Creative Commons o equivalente

Nota. Elaboración propia.

5.6.3 La ficha técnica del instrumento: modelo completo

Tabla 5.9

Plantilla de ficha técnica del instrumento — Modelo descargable

Campo	Información
Nombre completo del instrumento	
Nombre abreviado / sigla	
Autor(es) y año de publicación	
Versión actual y fecha	
Constructo medido	
Marco teórico de referencia	
Número total de ítems	
Número de dimensiones / subescalas	
Ítems por dimensión	
Formato de respuesta	
Rango de puntuación total	
Puntos de corte (si existen)	
Tiempo de administración	

Campo	Información
Modalidad de administración (autoaplicado / entrevistador / digital)	
Población objetivo (características de inclusión y exclusión)	
Idiomas disponibles	
Contextos de validación (países, poblaciones, años)	
Coefficiente de confiabilidad (α y ω por dimensión)	
Índices de ajuste del AFC (muestra de referencia)	
Evidencias de validez de criterio / convergente	
Repositorio donde está disponible (DOI / URL)	
Licencia de uso	
Condiciones de uso (libre / requiere permiso / comercial prohibido)	
Contacto para permisos y consultas	
Publicaciones principales que citar al usar el instrumento	

Nota. Esta plantilla debe completarse al finalizar el proceso de construcción y actualizarse con cada nuevo estudio de validación. Elaboración propia.

5.7 El Checklist CADI: 50 criterios de calidad final

5.7.1 ¿Qué es el CADI?

El Checklist CADI (Calidad, Adecuación y Documentación del Instrumento) es la herramienta de evaluación final original de esta obra. Con 50 criterios organizados en cinco dimensiones, el CADI representa el estándar mínimo de calidad que todo instrumento debe alcanzar antes de ser utilizado en investigación científica, publicado en una revista indexada, o incorporado a un programa de intervención o diagnóstico.

El CADI puede usarse de tres maneras complementarias:

- Autoevaluación: el investigador lo aplica a su propio instrumento durante el proceso de construcción, idealmente al finalizar cada fase del PCUI.
- Evaluación por pares: un colega con experiencia en metodología de la investigación lo aplica al instrumento antes de su publicación o uso.
- Evaluación editorial: los revisores de una revista científica lo usan como guía estructurada para evaluar la calidad metodológica de artículos de construcción y validación de instrumentos.

5.7.2 Estructura del CADI: 50 criterios en 5 dimensiones

Tabla 5.10

Checklist CADI — Calidad, Adecuación y Documentación del Instrumento

#	Criterio	Dimensión	Evaluación
DIMENSIÓN 1: FUNDAMENTOS CONCEPTUALES (10 criterios)			
1	El constructo está definido conceptualmente de manera clara, no circular y con anclaje teórico explícito	D1: Conceptual	☐ Sí ☐ No ☐ N/A
2	La definición conceptual delimita el dominio del constructo (qué incluye y qué excluye)	D1: Conceptual	☐ Sí ☐ No ☐ N/A
3	La revisión de literatura fue sistemática y cubrió al menos 3 bases de datos especializadas	D1: Conceptual	☐ Sí ☐ No ☐ N/A
4	El modelo de medición (reflectivo o formativo) está explícitamente declarado y justificado	D1: Conceptual	☐ Sí ☐ No ☐ N/A
5	Las dimensiones del constructo tienen base teórica documentada y no son producto de AFE únicamente	D1: Conceptual	☐ Sí ☐ No ☐ N/A

#	Criterio	Dimensión	Evaluación
6	Cada dimensión tiene su propia definición conceptual diferenciable de las otras	D1: Conceptual	☐ Sí ☐ No ☐ N/A
7	La operacionalización de variables está documentada (constructo → dimensiones → indicadores → ítems)	D1: Conceptual	☐ Sí ☐ No ☐ N/A
8	La tabla de especificaciones está completa y distribuye ítems coherentemente entre dimensiones	D1: Conceptual	☐ Sí ☐ No ☐ N/A
9	Se justifica explícitamente por qué se construyó un instrumento nuevo en lugar de adoptar o adaptar uno existente	D1: Conceptual	☐ Sí ☐ No ☐ N/A
10	El nivel de análisis del constructo (individual/grupal/organizacional) está especificado	D1: Conceptual	☐ Sí ☐ No ☐ N/A
DIMENSIÓN 2: DISEÑO Y CONSTRUCCIÓN TÉCNICA (10 criterios)			
11	Los ítems fueron redactados siguiendo reglas psicométricas establecidas (sin doble carga, sin ambigüedad, sin absolutos)	D2: Técnica	☐ Sí ☐ No ☐ N/A
12	Se generó un pool de ítems de al menos 150% del número final previsto	D2: Técnica	☐ Sí ☐ No ☐ N/A
13	La escala de respuesta es apropiada para el constructo y para el nivel educativo de la población	D2: Técnica	☐ Sí ☐ No ☐ N/A
14	El instrumento incluye instrucciones claras y completas para el respondiente	D2: Técnica	☐ Sí ☐ No ☐ N/A
15	El tiempo de administración fue verificado y es apropiado para la población (no genera fatiga excesiva)	D2: Técnica	☐ Sí ☐ No ☐ N/A
16	El juicio de expertos se realizó con al menos 5 expertos independientes y calificados	D2: Técnica	☐ Sí ☐ No ☐ N/A
17	Se calculó la V de Aiken (o IVC) por ítem y se usaron valores de corte apropiados	D2: Técnica	☐ Sí ☐ No ☐ N/A
18	Se realizó una prueba cognitiva con representantes de la población objetivo	D2: Técnica	☐ Sí ☐ No ☐ N/A
19	El pilotaje se realizó con una muestra de tamaño apropiado ($N \geq 100$ para escalas cuantitativas)	D2: Técnica	☐ Sí ☐ No ☐ N/A

#	Criterio	Dimensión	Evaluación
20	Los ítems problemáticos del pilotaje fueron revisados, reformulados o eliminados con criterios explícitos	D2: Técnica	☐ Sí ☐ No ☐ N/A
DIMENSIÓN 3: EVIDENCIAS DE VALIDEZ Y CONFIABILIDAD (10 criterios)			
21	Se reportan evidencias de validez de contenido (V de Aiken $\geq$ umbral crítico para todos los ítems)	D3: Validez	☐ Sí ☐ No ☐ N/A
22	El AFC fue especificado a priori (no AFE post-hoc presentado como AFC)	D3: Validez	☐ Sí ☐ No ☐ N/A
23	Se reportan al menos 4 índices de ajuste del AFC (χ^2/gl , CFI, TLI, RMSEA, SRMR)	D3: Validez	☐ Sí ☐ No ☐ N/A
24	Los índices de ajuste del AFC superan los umbrales mínimos aceptables	D3: Validez	☐ Sí ☐ No ☐ N/A
25	Se calcularon cargas factoriales estandarizadas para todos los ítems (todas ≥ 0.40)	D3: Validez	☐ Sí ☐ No ☐ N/A
26	Se calculó el coeficiente Alpha de Cronbach por dimensión ($\alpha \geq 0.70$)	D3: Validez	☐ Sí ☐ No ☐ N/A
27	Se calculó el coeficiente Omega de McDonald o equivalente como complemento al Alpha	D3: Validez	☐ Sí ☐ No ☐ N/A
28	Existe al menos una evidencia de validez convergente (correlación con variable teóricamente relacionada)	D3: Validez	☐ Sí ☐ No ☐ N/A
29	Existe al menos una evidencia de validez discriminante (diferenciación de constructos distintos)	D3: Validez	☐ Sí ☐ No ☐ N/A
30	Se reporta el Error Estándar de Medición (SEM) para interpretar puntuaciones individuales	D3: Validez	☐ Sí ☐ No ☐ N/A
DIMENSIÓN 4: PERTINENCIA CONTEXTUAL Y EQUIDAD (10 criterios)			
31	El instrumento fue diseñado o adaptado específicamente para la población objetivo del estudio	D4: Pertinencia	☐ Sí ☐ No ☐ N/A
32	Se realizó un proceso sistemático de adaptación cultural si el instrumento proviene de otro contexto	D4: Pertinencia	☐ Sí ☐ No ☐ N/A
33	El lenguaje del instrumento fue verificado con miembros de la población objetivo	D4: Pertinencia	☐ Sí ☐ No ☐ N/A
34	Los ítems fueron revisados para detectar sesgos potenciales de género, etnia, clase o cultura	D4: Pertinencia	☐ Sí ☐ No ☐ N/A

#	Criterio	Dimensión	Evaluación
35	Se verificó que el instrumento es accesible para el nivel educativo de la población objetivo	D4: Pertinencia	☐ Sí ☐ No ☐ N/A
36	Para poblaciones vulnerables o especiales se aplicaron adaptaciones metodológicas apropiadas	D4: Pertinencia	☐ Sí ☐ No ☐ N/A
37	Se evaluó el funcionamiento diferencial de ítems (DIF) entre subgrupos relevantes de la muestra	D4: Pertinencia	☐ Sí ☐ No ☐ N/A
38	El propósito del instrumento es coherente con el propósito para el que fue construido	D4: Pertinencia	☐ Sí ☐ No ☐ N/A
39	Las condiciones de administración son factibles en el contexto real de uso previsto	D4: Pertinencia	☐ Sí ☐ No ☐ N/A
40	El instrumento no tiene consecuencias diferencialmente perjudiciales para ningún grupo poblacional	D4: Pertinencia	☐ Sí ☐ No ☐ N/A
DIMENSIÓN 5: DOCUMENTACIÓN Y ÉTICA (10 criterios)			
41	El expediente técnico del instrumento está completo y disponible	D5: Documentación	☐ Sí ☐ No ☐ N/A
42	El proceso de construcción está documentado con suficiente detalle para ser replicado	D5: Documentación	☐ Sí ☐ No ☐ N/A
43	Todas las versiones del instrumento están registradas con fecha y justificación de cambios	D5: Documentación	☐ Sí ☐ No ☐ N/A
44	El instrumento está depositado en un repositorio abierto (OSF, Zenodo, o equivalente) con DOI	D5: Documentación	☐ Sí ☐ No ☐ N/A
45	El estudio de construcción fue aprobado por un comité de ética institucional	D5: Ética	☐ Sí ☐ No ☐ N/A
46	Todos los participantes (expertos, pilotaje, validación) firmaron consentimiento informado apropiado	D5: Ética	☐ Sí ☐ No ☐ N/A
47	El instrumento respeta los derechos de propiedad intelectual de instrumentos anteriores usados como referencia	D5: Ética	☐ Sí ☐ No ☐ N/A
48	El uso de IA en el proceso de construcción está declarado con transparencia	D5: Ética	☐ Sí ☐ No ☐ N/A
49	Las condiciones de uso del instrumento (licencia) están claramente especificadas	D5: Documentación	☐ Sí ☐ No ☐ N/A

#	Criterio	Dimensión	Evaluación
50	El instrumento es citado correctamente en las publicaciones que lo usan	D5: Documentación	☐ Sí ☐ No ☐ N/A

Nota. Sistema de puntuación: cada criterio = 2 puntos. 90-100 = Calidad excelente; 80-89 = Calidad adecuada; 70-79 = Calidad suficiente con mejoras necesarias; 60-69 = Calidad insuficiente; < 60 = Requiere reconstrucción sustancial. Elaboración propia.

5.7.3 Interpretación y uso del CADI

El CADI no es una herramienta de todo o nada. Su propósito es identificar áreas de fortaleza y áreas de mejora en el instrumento, no simplemente clasificarlo como “bueno” o “malo”. El investigador que obtiene una puntuación baja en alguna dimensión específica debe usar esa información para orientar su trabajo de mejora antes de proceder al siguiente paso del proceso.

Tabla 5.11

Guía de interpretación y acción según puntuación por dimensión del CADI

Puntuación en la dimensión	Interpretación	Acción recomendada	Prioridad
18-20 puntos (90-100%)	Excelente: la dimensión cumple todos o casi todos los criterios de calidad	Continuar con el proceso; documentar los criterios cumplidos en el expediente técnico	Baja
14-17 puntos (70-85%)	Adecuada: la dimensión tiene fortalezas pero presenta algunas áreas de mejora	Identificar los criterios no cumplidos y planificar acciones correctivas concretas	Media
10-13 puntos (50-65%)	Insuficiente: la dimensión tiene debilidades significativas que comprometen la calidad	Detener el proceso y abordar las deficiencias antes de continuar; puede requerir retroceder a fases anteriores	Alta
< 10 puntos (< 50%)	Crítica: la dimensión tiene fallos graves que pueden invalidar el instrumento	Reconstruir la dimensión deficiente desde sus bases; considerar si el instrumento puede ser rescatado	Urgente

Nota. Elaboración propia.

5.8 La hoja de ruta del investigador: del constructo al instrumento validado

Esta hoja de ruta sintetiza el proceso completo de construcción de un instrumento de investigación integrando todos los protocolos, herramientas y criterios de calidad desarrollados en esta obra. Está diseñada para ser usada como guía operativa en proyectos de investigación reales, con tiempos estimados y hitos de control claros.

Tabla 5.12

Hoja de ruta integral: del constructo al instrumento validado y documentado

Fase	Protocolo	Actividades clave	Productos/Hitos	Tiempo estimado
1	DCOI F1	Identificar el problema de medición; evaluar instrumentos existentes con los CJNI	Declaración del problema; justificación del instrumento nuevo	1 semana
2	DCOI F2	Revisión sistemática con PRSLI; elaborar MAAT; identificar vacíos	MAAT completa; lista de vacíos documentados	2-3 semanas
3	DCOI F3	Definir el constructo conceptual y operacionalmente; elegir modelo reflectivo/formativo	Definición conceptual; definición operacional; modelo de medición	1 semana
4	DCOI F4	Identificar y definir dimensiones; construir árbol de dimensiones	Árbol de dimensiones validado; diagrama del modelo AFC/PLS	1 semana
5	DCOI F5	Identificar indicadores; evaluarlos con los 6 criterios de calidad	Lista de indicadores aprobados por criterios	1 semana
6	DCOI F6	Construir la tabla de especificaciones completa	Blueprint aprobado; número de ítems confirmado	1 semana
7	DCOI F7	Revisión conceptual por especialistas externos	Informe de revisión con todos los comentarios respondidos	1-2 semanas
✓	Lista DCOI	Aplicar la lista de verificación del DCOI (40 criterios)	Puntuación DCOI $\geq 35/40$ antes de continuar	1 día
8	PCUI F1	Redactar pool de ítems (150% del número final)	Pool de ítems verificado con las 20 reglas	2-4 semanas
9	PCUI F2	Diseñar el formato, instrucciones y escala de respuesta	Instrumento preliminar completo	1 semana

Fase	Protocolo	Actividades clave	Productos/Hitos	Tiempo estimado
10	PCUI F3	Revisión lingüística por especialista y por miembro de la población	Instrumento con lenguaje verificado	1 semana
11	PCUI F4	Juicio de expertos; calcular V de Aiken e IVC; revisar ítems problemáticos	V de Aiken $\geq$ umbral crítico para todos los ítems	2-3 semanas
12	PCUI F5	Prueba cognitiva con 5-10 participantes representativos	Ningún ítem con confusión sistemática confirmada	1-2 semanas
13	PCUI F6	Pilotaje ($N \geq 100-300$); análisis de ítems; AFC preliminar	Ítems problemáticos identificados y corregidos o eliminados	3-4 semanas
14	PCUI F7	Validación completa ($N \geq 300-500$); AFC definitivo; confiabilidad; validez convergente	Índices de ajuste adecuados; $\alpha/\omega \geq 0.70$; evidencias de validez de constructo	3-4 semanas
15	PCUI F8	Versión definitiva; manual de aplicación; expediente técnico	Instrumento listo para uso con documentación completa	1-2 semanas
16	MCCDI	Aplicar la Matriz de Coherencia Constructo-Dimensión-Ítem	Todos los ítems coherentes con su indicador y dimensión	1 semana
17	Pertinencia	Protocolo de verificación de pertinencia en el contexto de uso	Confirmación de pertinencia teórica, cultural, poblacional y operacional	1 semana
18	CADI	Aplicar el Checklist CADI completo (50 criterios)	Puntuación CADI $\geq 80/100$ antes de publicar o usar	1-2 días
19	Depósito	Depositar el instrumento en OSF, Zenodo o repositorio institucional	DOI asignado; instrumento disponible para la comunidad científica	1 semana
20	Publicación	Redactar y someter el artículo de construcción y validación	Artículo publicado en revista indexada	Variable
TOTAL		20 hitos metodológicos integrados	Instrumento validado, documentado y disponible	6-12 meses

Nota. Los tiempos son estimaciones para proyectos de investigación estándar con dedicación parcial. Proyectos con dedicación completa pueden completarse en la mitad del tiempo estimado. Elaboración propia.

Reflexiones finales del capítulo y del libro

Este capítulo ha aportado lo que ningún manual de metodología de la investigación había reunido en un solo lugar: una clasificación sistemática de los 30 errores más graves y frecuentes en la construcción de instrumentos, las herramientas para detectarlos (análisis forense, MCCDI, lista de coherencia), las herramientas para verificar la pertinencia del instrumento en su contexto específico, el modelo completo del expediente técnico que garantiza la trazabilidad y reproducibilidad del proceso, y el Checklist CADI como estándar de calidad final integrador.

Al cerrar esta obra, vale la pena recordar el propósito que la animó desde sus primeras páginas: dotar al investigador de todo lo que necesita para construir instrumentos de medición que sean científicamente rigurosos, metodológicamente transparentes, culturalmente pertinentes y éticamente responsables. No instrumentos perfectos, sino instrumentos defensibles: instrumentos cuya calidad puede ser examinada, cuestionada y verificada por cualquier miembro informado de la comunidad científica.

Los cuatro protocolos originales de esta obra (DCOI, PCUI, MCCDI y CADI) no son recetas cerradas sino marcos de trabajo que el investigador debe adaptar a su campo, su constructo, su población y sus recursos. Lo que no puede adaptarse, lo que debe mantenerse con absoluta firmeza independientemente del contexto, son los principios que subyacen a estos protocolos: la primacía de la teoría sobre la estadística, la coherencia entre el constructo y el instrumento, la pertinencia cultural y poblacional, la transparencia del proceso y el respeto por los derechos de los participantes y de los autores cuyos trabajos se usan como referencia.

El instrumento que ha sobrevivido este proceso (que ha pasado por la revisión sistemática de la literatura, el juicio crítico de los expertos, la verificación cognitiva con la población, el rigor del análisis psicométrico, la verificación de coherencia y pertinencia, y la completitud del expediente técnico) es un instrumento que merece la confianza de la comunidad científica. Es un instrumento que puede ser usado, citado, adaptado y mejorado por otros investigadores. Es, en definitiva, una contribución real al conocimiento científico.

Preguntas de reflexión

1. Seleccione un instrumento publicado en una revista indexada de su disciplina que haya sido publicado en los últimos 5 años. Realice un análisis forense completo aplicando la lista de los 30 errores fatales. Para cada error presente, identifique qué impacto específico tiene en la validez del instrumento y qué recomendación haría al autor para corregirlo.

2. Construya la Matriz de Coherencia Constructo-Dimensión-Ítem (MCCDI) para un instrumento de 15 ítems de su elección. Para cada ítem, verifique la cadena de coherencia desde el constructo hasta el texto del ítem. Si identifica ítems con coherencia parcial o nula, proponga una versión revisada del ítem o un ítem de reemplazo.
3. Aplique el Checklist CADI completo a un instrumento que usted haya construido o que esté siendo desarrollado en su grupo de investigación. Calcule la puntuación total y por dimensión. Analice los criterios no cumplidos e identifique cuáles son críticos (su incumplimiento compromete la validez fundamental) y cuáles son importantes pero no críticos. Diseñe un plan de acción para alcanzar una puntuación $CADI \geq 80/100$.
4. Diferencie conceptualmente la validez, la confiabilidad, la coherencia interna y la pertinencia de un instrumento. Usando un ejemplo concreto de su área disciplinar, construya un escenario en que un instrumento tiene alta validez psicométrica pero baja pertinencia contextual, y otro escenario en que tiene alta pertinencia pero evidencias de validez insuficientes. ¿Cuál de los dos es más problemático para la investigación científica? Justifique su respuesta.
5. Diseñe el expediente técnico completo para un instrumento hipotético de su área disciplinar (puede ser el mismo instrumento que ha ido desarrollando a lo largo de las preguntas de reflexión de los capítulos anteriores). El expediente debe incluir todos los componentes de la Tabla 5.8. Reflexione sobre cuáles componentes fueron más difíciles de completar y por qué esa dificultad es metodológicamente significativa.

Referencias bibliográficas

- Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959. <https://doi.org/10.1177/001316448004000419>
- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/10.1177/0013164485451012>
- Allen, N. J., & Meyer, J. P. (1991). The measurement and antecedents of affective, continuance and normative commitment to the organization. *Journal of Occupational Psychology*, 63(1), 1–18. <https://doi.org/10.1111/j.2044-8325.1990.tb00506.x>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.
- Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.
- Bandalos, D. L. (2018). Measurement theory and applications for the social sciences. The Guilford Press.
- Bandura, A. (1997). Self-efficacy: The exercise of control. W. H. Freeman and Company.
- Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24(6), 574–594. <https://doi.org/10.1080/10447310802205776>
- Bass, B. M., & Avolio, B. J. (1995). MLQ Multifactor Leadership Questionnaire (2nd ed.). Mind Garden.
- Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287–311. <https://doi.org/10.1093/poq/nfm006>
- Binet, A., & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux [New methods for the diagnosis of the intellectual level of abnormals]. *L'Année Psychologique*, 11, 191–244.

- Birkmann, J. (Ed.). (2006). *Measuring vulnerability to natural hazards: Towards disaster resilient societies*. United Nations University Press.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.
- Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, 42(12), 1150–1152. <https://doi.org/10.1111/j.1365-2923.2008.03172.x>
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications.
- Clayton, S. (2003). Environmental identity: A conceptual and an operational definition. En S. Clayton & S. Opsow (Eds.), *Identity and the natural environment: The psychological significance of nature* (pp. 45–65). MIT Press.
- Cohen, J., McCabe, E. M., Michelli, N. M., & Pickeral, T. (2009). School climate: Research, policy, practice, and teacher education. *Teachers College Record*, 111(1), 180–213.
- Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research*, 61(12), 1250–1262. <https://doi.org/10.1016/j.jbusres.2008.01.013>
- Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119(2), 166.e7–166.e16. <https://doi.org/10.1016/j.amjmed.2005.10.036>
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. <https://doi.org/10.1037/h0040957>
- Cutter, S. L., Ash, K. D., & Emrich, C. T. (2014). The geographies of community disaster resilience. *Global Environmental Change*, 29, 65–77. <https://doi.org/10.1016/j.gloenvcha.2014.08.005>

- Cutter, S. L., Boruff, B. J., & Shirley, W. L. (2003). Social vulnerability to environmental hazards. *Social Science Quarterly*, 84(2), 242–261. <https://doi.org/10.1111/1540-6237.8402002>
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum Press.
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE Publications.
- Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, 61(12), 1203–1218. <https://doi.org/10.1016/j.jbusres.2008.01.009>
- Douglas, M., & Wildavsky, A. (1982). *Risk and culture: An essay on the selection of technical and environmental dangers*. University of California Press.
- Dunlap, R. E., Van Liere, K. D., Mertig, A. G., & Jones, R. E. (2000). New trends in measuring environmental attitudes: Measuring endorsement of the new ecological paradigm. *Journal of Social Issues*, 56(3), 425–442. <https://doi.org/10.1111/0022-4537.00176>
- Escobar-Pérez, J., & Cuervo-Martínez, Á. (2008). Validez de contenido y juicio de expertos: Una aproximación a su utilización. *Avances en Medición*, 6(1), 27–36.
- Fechner, G. T. (1860). *Elemente der Psychophysik* [Elements of psychophysics]. Breitkopf und Härtel.
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465. <https://doi.org/10.1177/2515245920952393>
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, 8(4), 370–378. <https://doi.org/10.1177/1948550617693063>
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50. <https://doi.org/10.1177/002224378101800104>
- Foucault, M. (1975). *Surveiller et punir: Naissance de la prison* [Discipline and punish: The birth of the prison]. Gallimard.

- Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45, 135–145.
- George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Allyn and Bacon.
- Goetz, T., Bieg, M., Lüdtke, O., Pekrun, R., & Hall, N. C. (2021). Do girls really experience more anxiety in mathematics? *Psychological Science*, 24(10), 2079–2087. <https://doi.org/10.1177/0956797613486989>
- Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. En N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105–117). SAGE Publications.
- Gulliksen, H. (1950). *Theory of mental tests*. Wiley.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.
- Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Lawrence Erlbaum Associates.
- Hambleton, R. K., & Zenisky, A. L. (2013). Translating and adapting tests for cross-cultural assessments. En D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 46–74). Cambridge University Press.
- Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). University of Chicago Press.
- Heine, S. J., & Buchtel, E. E. (2009). Personality: The universal and the culturally specific. *Annual Review of Psychology*, 60, 369–394. <https://doi.org/10.1146/annurev.psych.60.110707.163655>
- Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, 1(1), 104–121. <https://doi.org/10.1177/109442819800100106>
- Hofstede, G. (1991). *Cultures and organizations: Software of the mind*. McGraw-Hill.
- Hommel, B. E., Dobler, C., Paulhus, D. L., & Musch, J. (2024). Can large language models generate items for psychological questionnaires? A comparison with human-generated items. *Assessment: Advance online publication*. <https://doi.org/10.1177/10731911241258093>

- 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>
- International Test Commission. (2017). ITC guidelines for translating and adapting tests (2nd ed.). *International Journal of Testing*, 18(2), 101–134. <https://doi.org/10.1080/15305058.2017.1398166>
- ISO. (2018). ISO 9241-11:2018 Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. International Organization for Standardization.
- Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, 10(4), 407–415. [https://doi.org/10.1016/0197-2456\(89\)90005-6](https://doi.org/10.1016/0197-2456(89)90005-6)
- Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, 38(12), 1217–1218. <https://doi.org/10.1111/j.1365-2929.2004.02012.x>
- Kaiser, F. G., Wölfling, S., & Fuhrer, U. (1999). Environmental attitude and ecological behaviour. *Journal of Environmental Psychology*, 19(1), 1–19. <https://doi.org/10.1006/jevp.1998.0107>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.
- Kitzinger, J. (1995). Qualitative research: Introducing focus groups. *BMJ*, 311(7000), 299–302. <https://doi.org/10.1136/bmj.311.7000.299>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.
- Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. En P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 263–313). Emerald Group Publishing.
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.
- Kvale, S., & Brinkmann, S. (2009). *InterViews: Learning the craft of qualitative research interviewing* (2nd ed.). SAGE Publications.

- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Lather, P. (1991). *Getting smart: Feminist research and pedagogy with/in the postmodern*. Routledge.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563–575. <https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 1–55.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE Publications.
- Litwin, G. H., & Stringer, R. A. (1968). *Motivation and organizational climate*. Harvard Business School Press.
- Locke, E. A. (1976). The nature and causes of job satisfaction. En M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1297–1343). Rand McNally.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, 51, 201–226. <https://doi.org/10.1146/annurev.psych.51.1.201>
- Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, 22(1), 37–49. <https://doi.org/10.1037/0012-1649.22.1.37>
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412–433. <https://doi.org/10.1037/met0000144>
- Mertens, D. M. (2015). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods* (4th ed.). SAGE Publications.
- Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

- 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>
- Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press.
- Mitchell, J. T., Freitag, B., & Blatt, S. (2016). Teaching hazards and disasters: Curriculum development and evaluation. *Journal of Geography in Higher Education*, 40(1), 1–18.
- Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs). *Quality of Life Research*, 27(5), 1147–1157. <https://doi.org/10.1007/s11136-018-1798-3>
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63(7), 737–745. <https://doi.org/10.1016/j.jclinepi.2010.02.006>
- Moos, R. H., & Insel, P. M. (1974). *Work Environment Scale*. Consulting Psychologists Press.
- Morgan, D. L. (1997). *Focus groups as qualitative research* (2nd ed.). SAGE Publications.
- Norris, F. H., Stevens, S. P., Pfefferbaum, B., Wyche, K. F., & Pfefferbaum, R. L. (2008). Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. *American Journal of Community Psychology*, 41(1–2), 127–150. <https://doi.org/10.1007/s10464-007-9156-6>
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Oficina de las Naciones Unidas para la Reducción del Riesgo de Desastres. (2015). *Marco de Sendai para la reducción del riesgo de desastres 2015–2030*. UNDRR.
- Oficina de las Naciones Unidas para la Reducción del Riesgo de Desastres. (2017). *Terminology on disaster risk reduction*. UNDRR.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>

- Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2017). A guide for ensuring inclusion and equity in education. UNESCO.
- Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2021). Recommendation on the ethics of artificial intelligence. UNESCO.
- Organización para la Cooperación y el Desarrollo Económicos. (2008). Handbook on constructing composite indicators: Methodology and user guide. OECD Publishing. <https://doi.org/10.1787/9789264043466-en>
- Organización para la Cooperación y el Desarrollo Económicos. (2019). Recommendation of the Council on Artificial Intelligence. OCDE/LEGAL/0449.
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. University of Illinois Press.
- Ostadtaghizadeh, A., Ardalan, A., Paton, D., Jabbari, H., & Khankeh, H. R. (2015). Community disaster resilience: A systematic review on assessment models and tools. *PLOS Currents Disasters*. <https://doi.org/10.1371/currents.dis.f224ef8efbdfcf1d508dd0de4d8210ed>
- Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64(1), 12–40.
- Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). SAGE Publications.
- Paulhus, D. L. (1991). Measurement and control of response bias. En J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press.
- Petal, M. (2008). *Disaster prevention for schools: Guidance for education sector decision-makers*. Consultation version. UNISDR.
- Pfefferbaum, B., Pfefferbaum, R. L., & Van Horn, R. L. (2013). *Involving communities in disaster risk reduction: A practical guide*. FEMA.
- Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. National Center for Research to Improve Postsecondary Teaching and Learning.

- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903. <https://doi.org/10.1037/0021-9010.88.5.879>
- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 63, 539–569. <https://doi.org/10.1146/annurev-psych-120710-100452>
- Prinsen, C. A. C., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., de Vet, H. C. W., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, 27(5), 1147–1157. <https://doi.org/10.1007/s11136-018-1798-3>
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research.
- Revella, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods & Research*, 43(1), 73–97. <https://doi.org/10.1177/0049124113509605>
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74(1), 145–154. <https://doi.org/10.1007/s11336-008-9102-z>
- Rodebaugh, T. L., Woods, C. M., Thissen, D. M., Heimberg, R. G., Chambless, D. L., & Rapee, R. M. (2004). More information from fewer questions: The factor structure and item properties of the original and brief Fear of Negative Evaluation Scale. *Psychological Assessment*, 16(2), 169–181. <https://doi.org/10.1037/1040-3590.16.2.169>
- Rogers, R. W. (1983). Cognitive and physiological processes in fear appeals and attitude change: A revised theory of protection motivation. En J. T. Cacioppo & R. E. Petty (Eds.), *Social psychophysiology: A sourcebook* (pp. 153–176). Guilford Press.
- Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, 57(6), 1069–1081. <https://doi.org/10.1037/0022-3514.57.6.1069>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34(4, Pt. 2), 1–97.

- Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *Journal of Educational Research*, 99(6), 323–338. <https://doi.org/10.3200/JOER.99.6.323-338>
- Sherrieb, K., Norris, F. H., & Galea, S. (2010). Measuring capacities for community resilience. *Social Indicators Research*, 99(2), 227–247. <https://doi.org/10.1007/s11205-010-9576-9>
- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1–22. <https://doi.org/10.17763/haer.57.1.j463w79r56455411>
- 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>
- Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass*, 2(1), 414–433. <https://doi.org/10.1111/j.1751-9004.2007.00044.x>
- Slovic, P. (1987). Perception of risk. *Science*, 236(4799), 280–285. <https://doi.org/10.1126/science.3563507>
- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15(2), 201–292.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680. <https://doi.org/10.1126/science.103.2684.677>
- Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). Oxford University Press.
- Sunshine, J., & Tyler, T. R. (2003). The role of procedural justice and legitimacy in shaping public support for policing. *Law & Society Review*, 37(3), 513–548. <https://doi.org/10.1111/1540-5893.3703002>
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33(4), 529–554.
- van Muijen, J. J., Koopman, P., De Witte, K., De Cock, G., Susanj, Z., Lemoine, C., & Turnipseed, D. (1999). Organizational culture: The FOCUS questionnaire. *European*

- Journal of Work and Organizational Psychology, 8(4), 551–568.
<https://doi.org/10.1080/135943299398168>
- Vallerand, R. J., Pelletier, L. G., Blais, M. R., Brière, N. M., Senécal, C., & Vallières, E. F. (1992). The Academic Motivation Scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and Psychological Measurement*, 52(4), 1003–1017. <https://doi.org/10.1177/0013164492052004025>
- van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée*, 54(2), 119–135. <https://doi.org/10.1016/j.erap.2003.12.004>
- Wang, M. T., & Degol, J. L. (2016). School climate: A review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review*, 28(2), 315–352. <https://doi.org/10.1007/s10648-015-9319-1>
- Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Addison-Wesley.
- Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE Publications.
- Wisner, B., Blaikie, P., Cannon, T., & Davis, I. (2004). *At risk: Natural hazards, people's vulnerability and disasters* (2nd ed.). Routledge.
- World Justice Project. (2023). *WJP Rule of Law Index 2023*. World Justice Project.
- Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, 34(6), 806–838. <https://doi.org/10.1177/0011000006288127>
- Ziegler, M., Kemper, C. J., & Kruey, P. (2014). Short scales — Five misunderstandings and ways to overcome them. *Journal of Individual Differences*, 35(4), 185–189. <https://doi.org/10.1027/1614-0001/a000148>
- 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.

INSTRUMENTOS

DE

INVESTIGACIÓN CIENTÍFICA:

DISEÑO, VALIDACIÓN E IMPLEMENTACIÓN



CIDPROS
Centro de innovación y desarrollo profesional

ISBN: 978-9907-9574-0-2

