GUÍA PRACTICA DE VALIDEZ Y CONFIABILIDAD EN INSTRUMENTOS DE EVALUACIÓN PARA LA INVESTIGACIÓN EN SALUD

"GUIA PRÁCTICA DE VALIDEZ Y CONFIABILIDAD EN INSTRUMENTOS DE EVALUACIÓN PARA LA INVESTIGACIÓN EN SALUD"

TM. Mg. Marco Jiménez Herrera

PhD (c) Sc. Ed.

mjimenezhe@gmail.com

mjimenez@udd.cl

Autor:
Marco Jiménez Hernández
Magíster en Epidemiología
Doctorando en Ciencias de la Educación (PhD(c))
Universidad del Desarrollo, Chile
mjimenez@udd.cl

Santiago de Chile, 2025

Primera edición, Santiago de Chile, 2025.

PRESENTACIÓN DEL AUTOR

Marco Jiménez Herrera es Tecnólogo Médico en Radiología, Magíster en Educación y candidato a Doctor en Ciencias de la Educación. Con una sólida trayectoria como docente universitario e investigador en metodologías educativas en salud, ha desarrollado proyectos enfocados en la validación de instrumentos de evaluación, impresión 3D de modelos anatomo-patológicos y formación de competencias investigativas. Este manual representa la síntesis de su experiencia aplicada en educación superior en salud, combinando rigurosidad metodológica con herramientas prácticas para estudiantes, docentes e investigadores.

USO SUGERIDO DEL MANUAL

Este manual ha sido diseñado como una guía práctica para investigadores, docentes y estudiantes que necesiten diseñar, validar y analizar la confiabilidad de instrumentos de medición aplicados en contextos educativos y de salud. Puede utilizarse como material de apoyo en cursos de metodología de la investigación, evaluación educativa, diseño de instrumentos y análisis psicométrico. Los ejemplos incluidos están basados en casos reales y adaptados al software libre JASP, lo que facilita su replicación en entornos académicos sin costo adicional.

Se recomienda utilizar este documento como una hoja de ruta metodológica, desde la construcción del constructo hasta el análisis estadístico final, considerando las particularidades de cada tipo de instrumentos.

PRÓLOGO

La elaboración de este manual nace desde una necesidad concreta surgida en el desarrollo de mi línea de investigación, centrada en el diseño y validación de modelos anatomo-patológicos impresos en 3D con fines educativos. En ese proceso, descubrí que más allá del valor innovador de los modelos físicos, se requería rigurosidad metodológica para demostrar su validez pedagógica, tanto en términos de representación anatómica como en su impacto en el aprendizaje.

Este camino no solo me llevó a explorar con profundidad las técnicas de validación de contenido y confiabilidad, sino también a reflexionar sobre el rol de los instrumentos en investigación educativa. Entendí que cada prueba, rúbrica o cuestionario aplicado en el aula o en el laboratorio debe ser más que una herramienta: debe ser un dispositivo técnicamente sólido, éticamente responsable y pedagógicamente útil.

Así, este manual surge con un doble propósito. Por un lado, busca ser una guía práctica y accesible para quienes necesitan construir instrumentos válidos y confiables, especialmente en contextos educativos y de salud. Por otro, pretende ser un aporte a la formación de tutores e investigadores en competencias metodológicas, particularmente en el acompañamiento de tesis y proyectos de investigación en entornos universitarios.

Este documento sintetiza aprendizajes acumulados a lo largo de diversas etapas de mi trayectoria como docente, investigador y profesional de la salud. Espero que sirva como punto de partida para quienes inician en este camino, y como recurso de consolidación para quienes ya transitan procesos de evaluación más avanzados.

TM. Mg. Marco Jiménez Herrera
PhD (c) Ciencias de la Educación

TABLA DE CONTENIDO

INTRODUCCIÓN

El siguiente Manual es una propuesta metodológica y estadística para realizar la validación de instrumentos en dos escenarios distintos y particulares. La primera es la validación de un atributo, que en el ejemplo que se plantea es el conocimiento en una determinada área, y el segundo ejemplo es la validación de un instrumento interactivo para la educación en salud. Estas dos propuestas tienen objetivos distintos, sin embargo, existen etapas y pruebas estadísticas que podrían tener algo en común.

Este trabajo está realizado bajo una exhaustiva revisión de la literatura, que respalda la metodología y la estadística propuesta, recogiendo ejemplos publicados en la literatura y adaptándolos educativamente en este manual. El análisis estadístico se realiza a través del programa JASP 0.19 disponible gratuitamente en https://jasp-stats.org/ (*JASP - A Fresh Way to Do Statistics*, n.d.).

La validación de una encuesta, que a su vez es un instrumento que va a servir para medir una variable Intangible como es un atributo, consta de dos grandes procesos. El primero es crear una encuesta que responda a ciertos objetivos planteados y la segunda es el proceso de validación. No obstante, a pesar de que existen metodologías para realizar la validación de un instrumento, no siempre es la misma en todos los casos y se debe de adecuar las pruebas necesarias para cada caso y cada objetivo propuesto. Es por ello que este manual es una propuesta metodológica de validación para determinados ejemplos y podría variar dependiendo del propósito del investigador.

Se debe de aclarar desde un comienzo que la validez no es una propiedad del instrumento, si no, la interpretación que se hace de los resultados, dentro de un contexto y propósito específico (Zamanzadeh et al., 2015)

Para crear la encuesta se deben de construir dimensiones, subdimensiones e ítems. Por ello hay que tener claro que es lo que se quiere validar, que objetivo va a responder cada pregunta que vamos a construir.

Para crear las preguntas hay que buscar el fundamento teórico del tema que estamos abordando en la investigación. Hay que indagar si es que existen encuestas ya

validadas del tema. Si es que ya está validada y es lo que queremos para nuestro estudio se puede aplicar directamente, pero si faltan contenidos, podemos usarla como ejemplo para crear una encuesta propia y validarla posteriormente.

Hay que tener presente que la validación de instrumentos es una línea de investigación, por ello, evaluar muy bien los tiempos en que se realiza el proyecto. Que etapas hay que cumplir para lograr el propósito final del estudio.

MARCO CONCEPTUAL SOBRE VALIDEZ Y CONFIABILIDAD

Los instrumentos de evaluación en investigación educativa y en salud permiten medir variables no observables directamente, como conocimientos, actitudes o percepciones. Para que estos instrumentos sean técnicamente útiles, deben ser válidos y confiables.

Validez se refiere al grado en que un instrumento mide lo que pretende medir. No es una propiedad intrínseca del instrumento, sino de las interpretaciones que se hacen a partir de sus resultados, en un contexto y con un propósito específicos (American Educational Research Association et al., 2018). Las formas más comunes de validez son:

- **Validez de contenido**: grado en que los ítems representan adecuadamente el dominio de contenido.
- **Validez de constructo**: grado en que el instrumento refleja una estructura teórica subyacente.
- **Validez de criterio**: grado en que se relaciona con otras medidas externas que evalúan el mismo fenómeno.

Confiabilidad, por otro lado, se refiere a la consistencia o estabilidad de los resultados obtenidos por el instrumento. Existen distintas formas de evaluarla:

- **Consistencia interna** (alfa de Cronbach, KR-20): mide qué tan coherentes son los ítems entre sí.
- **Estabilidad temporal** (test–retest): mide la repetibilidad del instrumento en el tiempo.
- **Equivalencia entre formas o evaluadores** (kappa, correlación interjueces).

Un instrumento puede ser confiable sin ser válido, pero no puede ser válido si no es confiable. Por eso, ambos criterios deben evaluarse de manera complementaria

durante el proceso de desarrollo y aplicación de un instrumento (Haladyna, 2004; Prieto Adánez & Delgado González, 2010).

RESUMEN DE PASOS PARA LA VALIDACIÓN DE UN INSTRUMENTO

Paso	Etapa del proceso	Objetivo principal
1	Definición del constructo	Identificar qué se desea medir y en qué contexto
2	Revisión de literatura y teoría	Fundamentar teóricamente el constructo
3	Elaboración de ítems y dimensiones	Crear preguntas alineadas al constructo y sus dimensiones
4	Juicio de expertos	Evaluar validez de contenido de cada ítem
5	Análisis de validez estadística	Aplicar V de Aiken, CVC u otros indicadores
6	Evaluación de confiabilidad	Medir consistencia interna (KR-20, alfa, etc.)
7	Análisis por ítem	Calcular índices de dificultad y discriminación
8	Aplicación definitiva del instrumento	Usar la versión validada y confiable en población real

1. PRIMERA ETAPA: ELABORACIÓN DEL CONSTRUCTO.

Existen variables que por su naturaleza no se pueden medir directamente, es el caso de atributos como la inteligencia, el conocimiento, la satisfacción, etc. En el caso que queramos investigar estos atributos debemos de aterrizar esta variable y hacerla medible, a esto le denominamos constructo.

Estos constructos obedecen a dimensiones e indicadores que son evaluados a través de ítems o preguntas.

Nos podemos encontrar con algunos escenarios a la hora de estudiar nuestro constructo. Puede ser que el concepto este totalmente identificado y exista el instrumento listo para ser aplicado, en este caso hay que tener en cuenta la población o el idioma la cual fue dirigido y quizás debemos "re-validar el instrumento" a nuestra realidad. El otro escenario es que el constructo este medianamente conocido y se sepan algunas características de él, en este caso debemos de realizar los pasos necesarios para su validación en la población objetivo, y es el escenario que generalmente nos encontramos cuando estamos construyendo un cuestionario o instrumento que medirá ese atributo. Y el otro escenario y más complicado es que no se conozca nada del constructo y debemos partir de cero. En este contexto existen algunas estrategias para abordar esta problemática, como es, someter los temas propuestos a un comité de jueces expertos para debatir los contenidos pertinentes para armar el constructo, y también existe una metodología bastante robusta, que nos permitiría identificar esos ítems que desconocemos llamado método *DELPHI* que en esta ocasión no indagaremos a profundidad.

Sea cual sea el nivel de conocimiento que se tiene del constructo, debemos de realizar una profunda revisión de la literatura como primer paso para la elaboración de los ítems que integrara nuestro cuestionario que medirá el atributo elegido, o sea, es una etapa totalmente cualitativa.

Se recomienda realizar una revisión de la literatura de manera sistemática siguiendo las directrices PRISMA 2020, con el objetivo de identificar y documentar las definiciones del constructo (Maldonado Suárez & Santoyo Telles, 2024).

Si por ejemplo estamos pensando en validar un instrumento para medir conocimiento, ¿será igual que hacer una prueba de evaluación académica universitaria?. Aunque ambos tipos de pruebas buscan medir el conocimiento, hay diferencias metodológicas y estadísticas importantes entre la creación de un instrumento de investigación (como un cuestionario validado para tesis o estudios científicos) y una prueba universitaria usada para calificar estudiantes en una asignatura.

Al final del primer ejemplo abordaremos con más detención la diferencia de ambas metodologías.

A continuación se mencionarán los conceptos básicos para definir el constructo estudiado.

- **Definición del constructo.**

El punto de partida para diseñar cualquier instrumento válido y confiable es la delimitación precisa del constructo que se desea medir. Un constructo es una entidad conceptual no observable directamente, como conocimientos, habilidades, actitudes o percepciones, que puede ser inferida a través de la conducta, respuestas o desempeño de las personas evaluadas.

Este paso es esencial, ya que define la naturaleza del fenómeno a medir y orienta todo el proceso de construcción del instrumento, incluyendo el tipo de ítems, el formato de respuesta, los métodos de validación y las técnicas estadísticas.

- **Definición del propósito del instrumento**

Antes de construir el constructo, es necesario precisar qué se quiere medir y para qué. Por ejemplo:

¿Se busca medir un aprendizaje posterior a una intervención?

¿Se desea comparar grupos?

¿Se pretende evaluar una competencia en el marco de un perfil profesional?

Esto determinará si el constructo será evaluado de manera diagnóstica, formativa, sumativa o experimental.

- **Fundamentar el constructo teóricamente**

Toda definición de constructo debe estar respaldada por:

  - Revisión de literatura científica actualizada (artículos, libros, guías clínicas o curriculares).
  - Documentos oficiales (normativas profesionales, competencias de egreso, etc.).
  - Consensos de expertos en el área temática.

Esto permite elaborar una definición conceptual clara, acotada y defendible, que evite ambigüedad y oriente las dimensiones que se evaluarán.

- **Identificar dimensiones del constructo**

Muchos constructos complejos pueden descomponerse en dimensiones observables o componentes, que permiten operacionalizar el fenómeno. Cada dimensión debe:

  - Representar un aspecto clave del constructo.
  - Ser medible a través de uno o varios ítems.
  - Alinearse con el objetivo del instrumento.

Ejemplo: Un constructo como “nivel de conocimiento anatómico” podría dividirse en:

  - Reconocimiento de estructuras normales
  - Identificación de patologías
  - Aplicación en contexto clínico

- **Operacionalizar el constructo**

Consiste en transformar cada dimensión en comportamientos, respuestas o productos observables que puedan ser evaluados mediante ítems. Este proceso permite generar indicadores específicos para cada componente.

Ejemplo de operacionalización:

- Dimensión: Reconocimiento de estructuras
- Indicador: "Selecciona correctamente la vértebra anómala en una imagen axial."

- **Seleccionar el formato de medición**

Significa decidir cómo se va a recopilar las respuestas de los participantes para medir el constructo. No es lo mismo medir conocimientos que actitudes o habilidades. Cada tipo de constructo requiere una estructura distinta de preguntas o actividades, llamada formato de medición.

Tabla 1.

Formato de medición

Tipo de constructo	¿Qué estás midiendo?	¿Cómo se mide en la práctica?	Ejemplo real
Cognitivo	Saber algo (teoría, hechos, conceptos)	Con **preguntas cerradas** con respuestas correctas o incorrectas	Un test de conocimientos en protección radiológica con ítems como: "¿Cuál de las siguientes radiaciones tiene mayor poder de penetración?" (a–b–c–d)
Actitudinal	Qué piensa, siente o cree una persona	Con **escalas tipo Likert** (grado de acuerdo)	"Estoy de acuerdo con que la protección radiológica es una prioridad en la práctica clínica" → 1 (Totalmente en desacuerdo) a 5 (Totalmente de acuerdo)
Habilidades	Saber hacer algo (práctica o desempeño observable)	Con **rúbricas de evaluación** que describen niveles de desempeño	Observar a un estudiante usar un modelo anatómico 3D y evaluarlo en una rúbrica con criterios como: "Identifica correctamente estructuras patológicas" (Nivel 1 al 4)
Juicio de expertos	Valoración de calidad, claridad o pertinencia de algo por parte de evaluadores	Con **escalas ordinales** (por ejemplo, 1 a 4) en una ficha de validación	Expertos evalúan ítems de una encuesta: "¿Este ítem es claro?" → 1 (Nada claro) a 4 (Totalmente claro)

¿Cómo se define?

Primero se debe definir el tipo de constructo, y luego elegir el formato adecuado para medirlo. Por ejemplo:

- Si vas a medir si los estudiantes aprendieron anatomía → prueba de opción múltiple.

- Si quieres saber cómo perciben la utilidad del modelo 3D → escala tipo Likert.
- Si los estudiantes deben demostrar que saben identificar estructuras en un modelo → rúbrica de desempeño.
- Si validas los ítems con expertos → escala ordinal para juicio de expertos.

2. SEGUNDA ETAPA: VALIDEZ

La **validez** hace referencia al **grado en que un instrumento mide efectivamente lo que pretende medir** (Carvajal et al., 2011). Es un criterio de calidad esencial en cualquier proceso de evaluación, ya que asegura que los resultados del instrumento sean **interpretables y útiles** en relación con el objetivo para el cual fue construido.

Existen diferentes tipos de validez, entre ellos:

Tabla 2
Tipos de validez

Tipo de validez	¿Qué evalúa?
Validez de contenido	Que los ítems cubran adecuadamente el dominio del conocimiento evaluado.
Validez de constructo	Que el instrumento refleje adecuadamente un constructo teórico subyacente.
Validez de criterio	Que el instrumento se relacione con un criterio externo (ej.: desempeño real).

2.1 VALIDEZ DE CONTENIDO.

La validez de contenido se define como «el grado en que los elementos de un instrumento de evaluación son relevantes y representativos del constructo objetivo para un propósito de evaluación particular» (Haynes et al., 1995). Otra definición es en que tan adecuado es el muestreo que hace que una prueba del universo de posibles conductas, de acuerdo con lo que se pretende medir (Cohen et al., 2001).

Existen tres formas de conocer la validez de contenido: validez racional, validez de respuesta y juicio de expertos (Maldonado Suárez & Santoyo Telles, 2024).
La validez racional implica realizar una exhaustiva revisión de la literatura para identificar los elementos que responderán a nuestro constructo, en determinadas

ocasiones no se conoce lo bastante el tema, por lo que se aconseja realizar la validez de respuesta que consta de la realización de entrevistas a la población de interés, con el objetivo de identificar esos elementos que necesitamos para construir el instrumento. Escobar-Pérez & Cuervo-Martínez, (2008) mencionan que en determinadas ocasiones es aconsejable optar por ambos métodos de validez, la racional y de respuesta, para obtener una versión más completa de los ítems que conforman el instrumento.

Y la otra manera de realizar la validez de contenido es a través de juicio de expertos, quienes deben de ser idóneas para evaluar la propuesta de los ítems que medirá el constructo. Estas personas deben de considerarse expertas en el tema en cuestión, por ejemplo, si estamos hablando de medir nivel de conocimiento en una área de la radiología, deben de detener una vasta experiencia en el rubro y también conocimientos académicos.

Para describir el proceso de validación de contenido, se presentará un ejemplo donde desarrollaremos el proceso de validación por expertos.

2.2 EJEMPLO 1: CUESTIONARIO PARA MEDIR CONOCIMINETO

A continuación, se presenta la primera etapa de este ejemplo, que es una parte de la validación de contenido por jueces expertos sobre una encuesta que mida el nivel de conocimiento en radiodiagnóstico.

Suponiendo que se indago en la literatura nacional e internacional de manera sistemática acerca de cómo otros investigadores abordaron esta problemática, se llegó al siguiente esquema de dimensiones, subdimensiones e ítems que deben de ser evaluados por los jueces expertos.

Tabla 3

Ejemplo sobre la construcción de la encuesta.

DIMENSION	SUBDIMENSION	ITEMS
Conceptos básicos de radiación ionizante	• Definición y Diferencias entre radiación ionizante y no ionizante. • Fuentes naturales y artificiales de radiación. • Unidades de medida de radiación (Sievert, Gray, Becquerel).	¿Qué es la radiación ionizante y en qué se diferencia de la radiación no ionizante? ¿Cuáles son ejemplos de fuentes naturales y artificiales de radiación ionizante? ¿Cuáles son las unidades de medida para la dosis de radiación y qué significan (ej., Sievert, Gray)?
Principios de interacción de la radiación con la materia	• Mecanismos de interacción entre la radiación y los tejidos humanos. • Concepto de dosis absorbida y dosis efectiva. • Factores que influyen en la absorción de radiación en tejidos específicos.	¿Cómo afecta la radiación ionizante a los tejidos del cuerpo humano? ¿Qué diferencia hay entre dosis absorbida y dosis efectiva? ¿Qué factores determinan la cantidad de radiación absorbida en un órgano específico durante un procedimiento radiológico?
Efectos biológicos de la radiación ionizante	• Diferencias entre efectos estocásticos y deterministas. • Umbrales de dosis para efectos inmediatos y a largo plazo. • Riesgos para diferentes sistemas del cuerpo (sistema nervioso, reproductivo, etc.).	¿Cuál es la diferencia entre efectos estocásticos y efectos deterministas de la radiación? ¿Cuáles son los umbrales de dosis para efectos inmediatos (como quemaduras) y efectos a largo plazo (como cáncer)? ¿Qué sistemas del cuerpo son más vulnerables a los efectos de la radiación ionizante?
Principios de protección radiológica	• Principios de ALARA (As Low As Reasonably Achievable). • Importancia de la distancia, el tiempo y el blindaje en la protección.	¿Qué significa el principio de ALARA (As Low As Reasonably Achievable) y por qué es importante? ¿Cuáles son las tres principales medidas de protección radiológica (distancia, tiempo y blindaje) y cómo

	• Diferencia entre protección para el paciente y para el personal técnico.	se aplican? ¿Cuál es la diferencia entre la protección radiológica para el paciente y la protección para el personal técnico?
Normativas y prácticas de seguridad en radiodiagnóstico	• Normas internacionales y nacionales sobre seguridad radiológica. • Uso correcto de equipos de protección (chalecos plomados, dosímetros). • Procedimientos de seguridad en el manejo de equipos de radiodiagnóstico.	¿Qué normativas nacionales o internacionales existen para regular la seguridad en el uso de radiación en medicina? ¿Cuál es el uso adecuado de equipos de protección personal, como los chalecos plomados y los dosímetros? ¿Qué procedimientos de seguridad deben seguirse al manejar equipos de radiodiagnóstico?
Conocimiento de tecnologías de radiodiagnóstico	• Diferencias en dosis y aplicaciones de técnicas como radiografía, tomografía computarizada, fluoroscopia y mamografía. • Riesgos específicos asociados a cada tipo de examen. • Procedimientos para optimizar la calidad de la imagen con la menor dosis posible.	¿Cuáles son las principales diferencias en cuanto a dosis de radiación y propósito entre una radiografía, una tomografía computarizada, una fluoroscopia y una mamografía? ¿Qué riesgos específicos están asociados con cada tipo de examen radiológico y cómo se minimizan? ¿Qué procedimientos de optimización de imagen ayudan a obtener la mejor calidad de imagen con la menor dosis de radiación posible?

Nota: tabla ejemplificadora sobre cómo se debe de construir las preguntas según los objetivos planteados para contenido que se quiere abordar.

En la siguiente etapa se debe de construir el instrumento para los jueces expertos. Ellos evaluarán cada pregunta si está acorde a la dimensión planteada y complementarán si es que hace falta, de acuerdo con una escala. En esta oportunidad evaluarán según una esca de Likert que consta de lo siguiente:

1 = Muy en desacuerdo
2 = En desacuerdo
3 = Neutro
4 = De acuerdo
5 = Muy de acuerdo

Así valorarán cada pregunta de acuerdo a la pertinencia del objetivo propuesto. Por lo tanto a través de escala de Likert, evaluaran los siguientes indicadores a cada pregunta o Ítems propuesto tal cual lo menciona Escobar-Pérez y Cuervo-Martínez, (2008) y Hernández-Nieto, (2002).

1.- Claridad: Evaluadores valoran si la pregunta se entiende claramente, si se usan términos que los participantes reconocen y si el contenido es comprensible sin requerir explicaciones adicionales.
2.- Relevancia: La pertinencia de la pregunta en relación con el objetivo de la medición y la importancia del conocimiento evaluado.
3.- Coherencia : Se evalúa si la pregunta tiene una relación lógica con la temática de la protección radiológica y si el contenido está alineado con los objetivos del estudio
4.- escala: Evalúa si la pregunta es apropiada en nivel de dificultad y profundidad para el grupo de estudio, permitiendo discriminar entre diferentes niveles de conocimiento sin ser ni demasiado básica ni demasiado avanzada

De acuerdo a estos conceptos los jueces evaluaran cada pregunta de acuerdo al objetivo planteado. La tabla a presentar a cada juez quedará de la siguiente manera:

Tabla 4
Ejemplo de tabla a entregar a cada juez validador

DIMENSION	SUBDIMENSION	ITEMS	INDICADORES	JUEZ1	OBS.JUEZ1
Conceptos básicos de radiación ionizante	Definición y Diferencias entre radiación ionizante y no ionizante.	¿Qué es la radiación ionizante y en qué se diferencia de la radiación no ionizante? (colocar alternativas)	claridad		
			relevancia		
			coherencia		
			escala		

Conceptos básicos de radiación ionizante	Fuentes naturales y artificiales de radiación.	¿Cuáles son ejemplos de fuentes naturales y artificiales de radiación ionizante? (colocar alternativas)	claridad		
			relevancia		
			coherencia		
			escala		

Nota: de esta manera se debe de complementar la tabla completa. Considerar las alternativas de respuestas o si va a ser con respuesta corta o larga, también debe de ir en esta tabla.

Después de recoger toda la información se debe proceder a sacar el coeficiente de validez del ítems y el coeficiente de validez total.

Para el cálculo de la validez existen varios test estadísticos, el **Coeficiente de Validez de Contenido (CVC) de Hernández-Nieto(2002), Indice de Validez de Contenido (IVC) de Lawshe (1975)**, el **Coeficiente de Validez de Contenido de Aiken (V Aiken)**(Aiken, 1980)**.** El índice de Lawshe necesita alrededor de 9 validadores, mientras que el de Hernández-Nieto y la V de Aiken solo se necesitan 3 (Pedrosa et al., 2013).

Los más utilizados son el de Hernández Nieto y la V de Aiken, aunque algunos artículos han reportado ambos índices (Arango-Ramírez et al., 2023).

Definición de V de Aiken y CVC de Hernández-Nieto:

V de Aiken: Evalúa si cada ítem del instrumento es relevante y adecuado según el juicio de expertos. Se usa cuando queremos saber si un ítem específico representa bien el contenido que queremos medir.

CVC de Hernández-Nieto: También evalúa validez de contenido, pero más como un promedio general de las opiniones de los jueces. Por eso, a veces se dice que tiene un enfoque más "global", y ayuda a ver si todo el instrumento está bien construido, no solo cada ítem por separado.

El coeficiente de validez tiene la siguiente interpretación:

1.- Menor a 0,6= validez inaceptable

2.- Igual o mayor que 0,6 y menor o igual a 0,7= validez

deficiente

3.- Mayor que 0,71 y menor o igual que 0,8= validez aceptable

4.- Mayor que 0,8 y menor o igual que 0,9= validez buena

5.- Mayor que 0,9 = validez excelente

Diferencias entre V de Aiken y CVC de Hernández Nieto:

Tabla 5.
Diferencias entre V de Aiken y CVC de Hernández Nieto

Aspecto	V de Aiken	CVC de Hernández-Nieto
Fórmula	$V = \frac{\sum(s)}{n(c-1)}$ donde $s = r - lo$	$\mathrm{CVC} = \frac{\bar{X}}{c}$ (media de puntuaciones entre el valor máximo de escala)
Datos requeridos	Calificaciones por jueces sobre **un solo criterio** (aunque se puede aplicar a varios).	Calificaciones por jueces sobre **varios criterios**, evaluados uno a uno.
Escala usada	Ordinal (usualmente de 1 a 5 o 1 a 4).	Cualquiera, pero requiere conocer el **valor máximo de la escala**.
Requiere puntuación mínima	Sí, debe definirse el **mínimo valor posible (lo)**.	No lo requiere. Solo usa la media de evaluaciones.

Interpretación:

Tabla 6
Comparación de valoración

Aspecto	V de Aiken	CVC de Hernández-Nieto
Valoración crítica	≥ 0.70 se considera aceptable; >0.80 se considera alta validez.	≥ 0.80 suele considerarse aceptable, aunqu depende del contexto.
Identifica ítems problemáticos	Sí, por ítem y por criterio.	También, pero con menor precisión discriminativa.

Ejemplo:

Supongamos que validamos el siguiente ítem del cuestionario sobre radiación:

Ítem 3: “El blindaje adecuado en salas de rayos X debe cumplir con las normas de protección radiológica indicadas por la autoridad sanitaria.”

Este ítem fue evaluado por 4 jueces en 4 criterios (claridad, relevancia, coherencia y adecuación de la escala) con puntajes de 1 a 5 (escala Likert).

Tabla 7.
Comparación entre los valores con el mismo puntaje

Criterio	Puntuaciones (Jueces)	V de Aiken	CVC Hernández-Nieto
Claridad	3, 3, 2, 3	0.438	0.275
Relevancia	3, 4, 4, 3	0.625	0.35
Coherencia	3, 4, 3, 3	0.562	0.325
Escala	3, 3, 3, 3	0.500	0.300

Conclusión de validez de contenido Ítem 3:

Los resultados del análisis de validez de contenido para el Ítem 3 evidencian puntuaciones bajas tanto en la V de Aiken como en el coeficiente de validez de contenido (CVC) de Hernández-Nieto. Específicamente, las puntuaciones de V de Aiken se encuentran por debajo del umbral mínimo aceptable de 0.70 en todos los criterios evaluados (claridad = 0.438, relevancia = 0.625, coherencia = 0.562, escala = 0.500). Asimismo, los valores de CVC no superan el estándar mínimo recomendado de 0.80, con promedios que oscilan entre 0.275 y 0.350.

Estos resultados indican que el ítem no cumple con los niveles esperados de calidad en su redacción ni en su adecuación al constructo medido. Se recomienda revisar su formulación, especialmente en términos de claridad del lenguaje técnico utilizado, la alineación con los objetivos de evaluación y la adecuación de su dificultad para el público objetivo. Si las modificaciones no logran mejorar sus indicadores en una segunda ronda de validación, se sugiere considerar su reemplazo o eliminación.

¿Cuándo se debe eliminar un ítem definitivamente?

1. Bajo puntaje en la V de Aiken (menos de 0.70): Si el ítem obtiene una V de Aiken < 0.70 en más de un criterio (por ejemplo, claridad y relevancia), indica que no cumple con los estándares mínimos de validez aceptados en la literatura.
Si ninguna de sus dimensiones alcanza al menos 0.75, es fuerte indicio de que el ítem es débil o inadecuado (Aiken, 1980).

2. CVC de Hernández-Nieto muy bajo (menor a 0.75): Si el ítem presenta un índice de validez de contenido (CVC) < 0.75, especialmente cuando se basa en el promedio de evaluaciones de varios jueces, no justifica su permanencia en la versión final del instrumento, por lo tanto debe de ser modificado o eliminado (Hernández-Nieto, 2002).

3. Falta de mejora tras revisión: Si el ítem fue reformulado y re-evaluado en una segunda ronda, y aún no mejora sus indicadores de validez, se recomienda suprimirlo (Escobar-Pérez & Cuervo-Martínez, 2008).

4. Incoherencia teórica: Aun cuando los valores sean aceptables, si el ítem no guarda coherencia con los objetivos del instrumento o no aporta información relevante al constructo, puede eliminarse por criterio conceptual (Rubio et al., 2003).

5. Juicio cualitativo de expertos: Si dos o más expertos indican que el ítem es confuso, ambiguo, o irrelevante, y estas observaciones se repiten, es evidencia cualitativa suficiente para su eliminación, especialmente si coinciden con malos indicadores cuantitativos (Rubio et al., 2003).

Conclusión:

Un ítem debe ser eliminado cuando presenta puntajes bajos en las medidas de validez de contenido (V de Aiken < 0.70 y CVC < 0.75) de forma sostenida, especialmente en

dimensiones críticas como claridad y relevancia. Además, si tras su reformulación no se observan mejoras significativas, o si su contenido no se alinea con los objetivos del instrumento, su exclusión se considera necesaria. Esta decisión debe basarse tanto en evidencia cuantitativa como en el juicio cualitativo de los expertos.

Posterior a este trabajo se debe seguir con la validez de constructo, validez de criterio y confiabilidad del instrumento y se debe de probar en una muestra piloto para posteriormente implementarlo en la población objetivo.
La prueba piloto debe de ser representativa de la población objetivo.

3. TERCERA ETAPA: PRUEBA PILOTO, VALIDEZ DE CONSTRUCTO, VALIDEZ DE CRITERIO Y CONFIABILIDAD

Luego que ya validamos la encuesta por tres jueces expertos, previo haberle hechos las modificaciones que nos señalaron los jueces, debemos de realizar una prueba piloto a una muestra representativa de la población de estudio.

La prueba piloto tiene como finalidad evaluar la comprensión, claridad y consistencia interna del cuestionario. Según recomendaciones metodológicas, un número mínimo de 20 a 30 participantes es suficiente para una validación preliminar en esta etapa (Hertzog, 2008). Además, para estimaciones de confiabilidad como el coeficiente KR-20 o alfa de Cronbach, se sugiere una muestra no menor a 30 sujetos (Perneger et al., 2015). Para análisis factorial exploratorio, se requiere una proporción de entre 5 a 10 participantes por ítem del cuestionario (Hair et al., 2010), por lo que, en caso de contar con 12 ítems, se estima una muestra mínima de 60 participantes. Esta estrategia permitirá evaluar la viabilidad del instrumento antes de su aplicación definitiva en la muestra total del estudio.

A continuación se presenta una tabla a modo de ejemplo, de la recogida de datos de los encuestados. Hay que tener presente que la encuesta se trata de preguntas y alternativas donde solo una es la correcta por lo tanto el dato es dicotómico (respuesta correcta o incorrecta), por lo tanto, se asignará 1 a si está correcto y 0 si está incorrecta.

Tabla 8.
Base de datos de los encuestados (base)

Participantes	Pregunta 1	Pregunta 2	Pregunta 3	Pregunta …n
1	0	1	1	0
2	0	0	1	1
3	1	1	1	0

3.1 Validez de constructo.

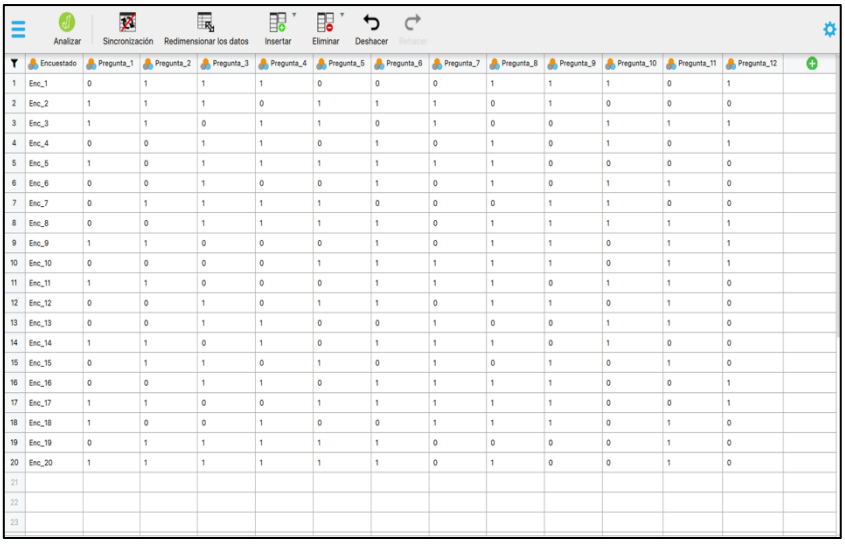
La validez de constructo se utiliza para saber si las preguntas se agrupan en dimensiones o miden el mismo concepto. Si **ya se sabe** las dimensiones y están predefinidas se debe de realizar un **Análisis factorial Confirmatorio (AFC)**, pero si estas dimensiones **no** se conocen se debe de realizar un **Análisis Factorial Exploratorio (AFE)**.

3.1.1 Análisis Factorial Exploratorio.

Supongamos que tenemos 12 preguntas a 20 encuestados como prueba piloto. Estas preguntas se responden a través de alternativas, o sea es dicotómica o está correcta (=1) o está incorrecta (=0).

Para ejemplificarlo se creará una base de datos simulada y todos los análisis se harán con el programa JASP 0.19.2.

Figura 1.
Base de datos en JASP

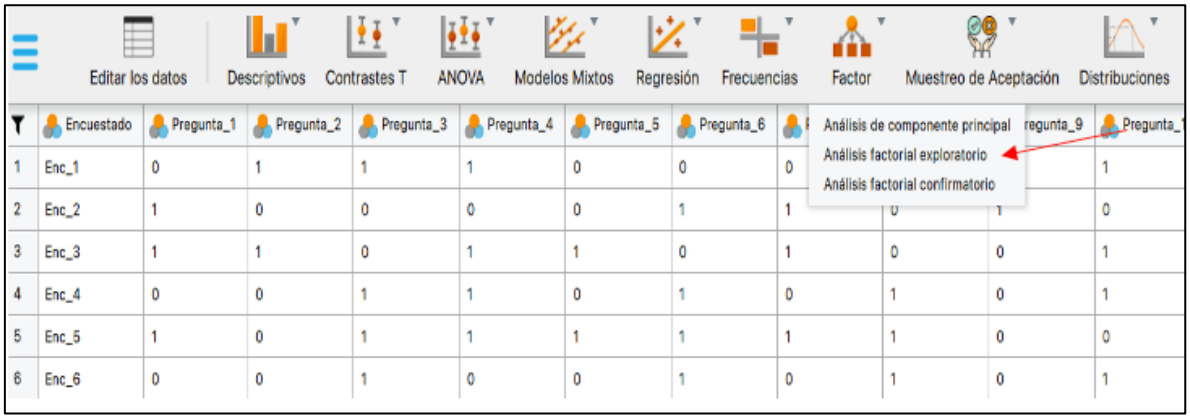


	Encuestado	Pregunta_1	Pregunta_2	Pregunta_3	Pregunta_4	Pregunta_5	Pregunta_6	Pregunta_7	Pregunta_8	Pregunta_9	Pregunta_10	Pregunta_11	Pregunta_12
1	Enc_1	0	1	1	1	0	0	0	1	1	1	0	1
2	Enc_2	1	1	1	0	1	1	1	0	1	0	0	0
3	Enc_3	1	1	0	1	1	0	1	0	0	1	1	1
4	Enc_4	0	0	1	1	0	1	0	1	0	1	0	1
5	Enc_5	1	0	1	1	1	1	1	1	0	0	0	0
6	Enc_6	0	0	1	0	0	1	0	1	0	1	1	0
7	Enc_7	0	1	1	1	1	0	0	0	1	1	0	0
8	Enc_8	0	0	1	1	1	1	0	1	1	1	1	1
9	Enc_9	1	1	0	0	0	1	0	1	1	0	1	1
10	Enc_10	0	0	0	0	1	1	1	1	1	0	1	1
11	Enc_11	1	1	0	0	0	1	1	1	0	1	1	0
12	Enc_12	0	0	1	0	1	1	0	1	1	0	1	0
13	Enc_13	0	0	1	1	0	0	1	0	0	1	1	0
14	Enc_14	1	1	0	1	0	1	1	1	0	1	0	0
15	Enc_15	0	1	1	0	1	0	1	0	1	0	1	0
16	Enc_16	0	0	1	1	0	1	1	1	1	0	0	1
17	Enc_17	1	1	0	0	1	1	1	1	1	0	0	1
18	Enc_18	1	0	0	1	0	0	1	1	1	0	1	0
19	Enc_19	0	1	1	1	1	1	0	0	0	0	1	0
20	Enc_20	1	1	1	1	1	1	0	1	0	0	1	0
21													
22													
23													

Nota: base de datos con 20 encuestados y 12 preguntas que miden conocimiento con respuesta dicotómica (0-1).

A continuación, se describen los pasos para determinar el AFE con sus interpretaciones. Se debe de ir a análisis factorial exploratorio como se indica en la siguiente figura.

Figura 2.
Análisis Factorial Exploratorio (AFE) en JASP



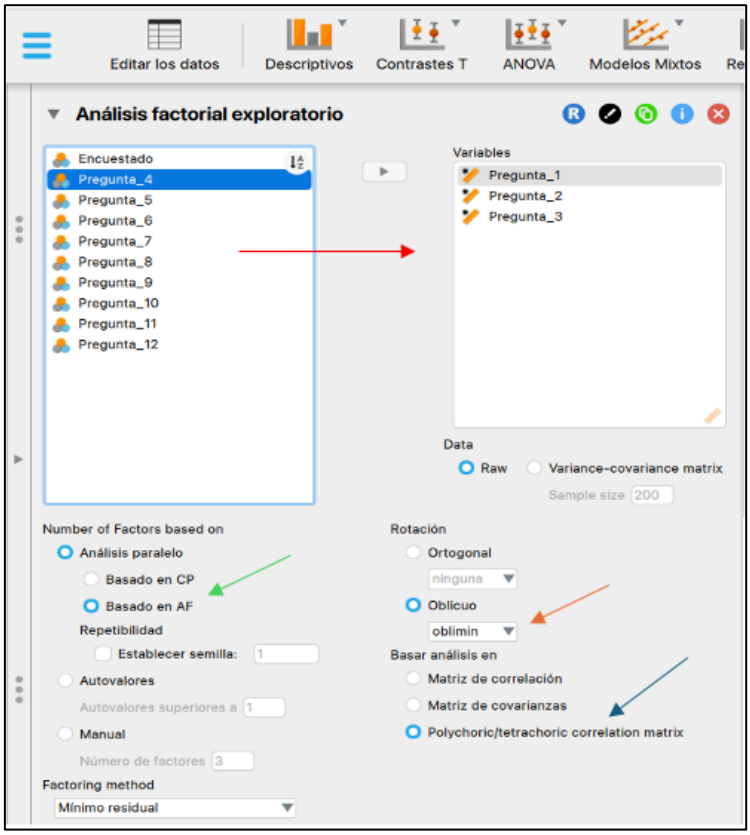
	Encuestado	Pregunta_1	Pregunta_2	Pregunta_3	Pregunta_4	Pregunta_5	Pregunta_6				
1	Enc_1	0	1	1	1	0	0	0			1
2	Enc_2	1	0	0	0	0	1	1	0	1	0
3	Enc_3	1	1	0	1	1	0	1	0	0	1
4	Enc_4	0	0	1	1	0	1	0	1	0	1
5	Enc_5	1	0	1	1	1	1	1	1	0	0
6	Enc_6	0	0	1	0	0	1	0	1	0	1

Nota: se debe de desplegar el análisis donde dice factor (flecha roja) y elegir Análisis factorial exploratorio.

Luego hay que seleccionar algunos parámetros importantes, como la matriz tetracórica (ya que son datos dicotómicos), y tenemos dos opciones de análisis en JASP: decirle que agrupe los factores el, o si creemos que las preguntas se agrupan en un número determinado de factores (dimensiones) (ejemplo factor 1: conocimiento básico, factor 2: conocimiento de unidades y factor 3: conocimiento de medidas).

A continuación, se esquematiza los pasos que se deben de seguir en JASP para el análisis exploratorio sin determinación de factores manualmente.

Figura 3.
Parámetros de análisis para JASP



Nota: primero hay que trasladar las preguntas al cuadro donde dice variables (flecha roja), luego en análisis paralelo marcar Basado en AF (flecha verde), luego en Rotación elegir oblicuo oblimin (flecha naranja) y por último donde dice Basar análisis en, elegir Polychoric/tetrachoric matrix (flecha azul)

La salida de JASP es la siguiente:

Tabla 9.
Salida de JASP de resultados (cargas) de factores y Unicidad

Cargas de los Factores

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Unicidad
Pregunta_1	0.949							0.023
Pregunta_12	0.614							0.553
Pregunta_9	0.438	0.513						0.193
Pregunta_4		0.943						0.009
Pregunta_2		0.490					0.407	0.645
Pregunta_6			0.985					0.020
Pregunta_5			0.455					0.681
Pregunta_10				0.959				0.000
Pregunta_11					0.974			0.012
Pregunta_3						0.896		0.078
Pregunta_8							0.911	0.116
Pregunta_7								0.769

Nota: fíjese que las preguntas están ordenadas de acuerdo con el factor que JASP determino como relevante.

Para interpretar estos datos debemos de saber que una carga es alta cuando es mayor a 0.5. Con respecto a la unicidad, esta indica cuanta varianza de un ítem o pregunta no es explicada por los factores, por lo tanto, necesitamos unicidad baja y se interpreta de la siguiente manera:

Valor de Unicidad	Interpretación práctica
0.00 – 0.30	Muy bien explicado. Ideal.
0.30 – 0.50	Aceptable.
0.50 – 0.70	Dudas. Revisar el ítem.
> 0.70	Problema. Ítem débil o irrelevante.

De acuerdo con estas definiciones la interpretación definitiva a la salida de JASP es la siguiente:

Tabla 10.
Interpretación de los resultados con 7 factores que determinó el análisis de JASP

Ítem	Factor más alto	Interpretación	Unicidad
Pregunta_1	Factor 1 (0.949)	Carga excelente	0.023 (muy baja unicidad)
Pregunta_12	Factor 1 (0.614)	Carga aceptable	0.553 (moderada)
Pregunta_9	Factor 1 (0.438), Factor 3 (0.513)	**Cargas cruzadas** (confusión)	0.193 (moderada)
Pregunta_4	Factor 1 (0.943)	Carga excelente	0.009 (casi toda varianza explicada)
Pregunta_2	Factor 1 (0.490)	Carga baja/moderada	0.645 (alta unicidad)
Pregunta_6	Factor 2 (0.985)	Carga excelente	0.020 (muy baja unicidad)
Pregunta_5	Factor 2 (0.455)	Carga baja/moderada	0.681 (alta unicidad)
Pregunta_10	Factor 4 (0.959)	Carga excelente	0.000 (perfecto, toda varianza explicada)
Pregunta_11	Factor 5 (0.974)	Carga excelente	0.012 (muy baja unicidad)
Pregunta_3	Factor 6 (0.896)	Carga excelente	0.078 (muy baja unicidad)
Pregunta_8	Factor 7 (0.911)	Carga excelente	0.116 (muy baja unicidad)
Pregunta_7	Factor 7 (0.769)	Carga aceptable	0.231 (moderadamente buena)

Nota: la pregunta 9 tiene una carga cruzada, o sea, tiene para factor 1 y factor 3, esto indica que la pregunta se debe de revisar o eliminar.

Interpretación práctica

Items con cargas muy fuertes (excelentes):

- Pregunta_1, Pregunta_4, Pregunta_6, Pregunta_10, Pregunta_11, Pregunta_3, Pregunta_8.: Estas preguntas se explican muy bien por su respectivo factor (unicidad casi nula o muy baja).

Items con problemas leves:

- Pregunta_2 y Pregunta_5: cargas más bajas y unicidades altas (>0.60), revisarlas o eliminarlas si se quiere un modelo más limpio.
- Pregunta_ 7: carga buena y unicidad aceptable. Mantener con observación
- Pregunta _12: carga aceptable, unicidad moderada. Mantener con observació

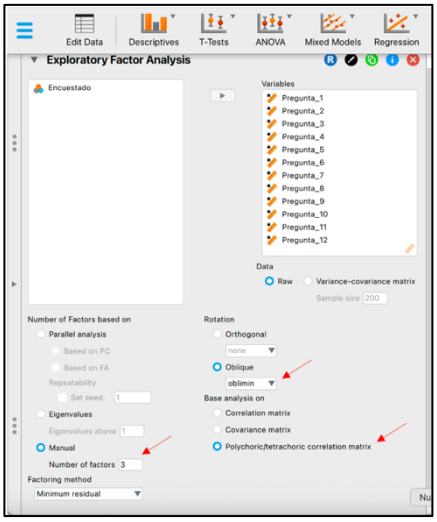
Items con cargas cruzadas:

- Pregunta_9: carga en dos factores (Factor 1 y Factor 3). Esto genera ambigüedad, podría considerarse su reformulación o eliminación si se quiere pureza de factores, aunque la carga mayor es en el factor 3.

En este ejemplo el programa de AFE de JASP agrupó las preguntas en 7 categorías. Sin embargo si nosotros conocemos o tenemos indicios de cuantas dimensiones o factores se agruparían estas preguntas podemos decírselo al programa.

A continuación se presenta los pasos para realizar el AFE pero asignando los factores.

Figura 4.
AFE en JASP con número de factores manual



Nota: flecha roja indica los parámetros del programa para determinar manualmente el número de factores. Note que el número de factores en este ejemplo es 3.

La salida de JASP es la siguiente:

Tabla 11
Salida de JASP para los 3 factores

	Factor 1	Factor 2	Factor 3	Unicidad
Pregunta_1	0.999			0.043
Pregunta_2	0.709			0.503
Pregunta_8	0.693			0.382
Pregunta_12	0.643	0.497		0.389
Pregunta_5	0.612	0.703		0.022
Pregunta_10	0.443	0.491		0.532
Pregunta_7	0.419			0.800
Pregunta_11		1.000		0.009
Pregunta_3		0.465		0.619
Pregunta_9			0.935	0.128
Pregunta_6			0.837	0.282
Pregunta_4				0.835

Nota: el programa arroja las respuestas de las preguntas de acuerdo con el número de factores que nosotros le indicamos. La interpretación de los datos es igual que en el ejemplo anterior, por lo tanto podemos observar que la pregunta _1 tiene un alta carga (0.99) y una baja unicidad (0.043) excelente para mantener en el cuestionario, mientras que la pregunta_2 a pesar de tener una buena carga factorial (0.709), tiene una alta unicidad, en este caso debemos de analizar la pregunta y argumentar si la mantenemos o no .Recordar que la unicidad mayor a 0.5 se debe de revisar la pregunta, y si es mayor a 0.7 la pregunta es débil o irrelevante.

JASP entrega otra tabla que indica las características de los factores incluidos tanto para el ejemplo anterior como en este.

Tabla 12

Características de los Factores

	Solución no rotada				Solución rotada		
	Autovalores	Sumas de cargas al cuadrado	Proporción varianza.	Acumulativo	Sumas de cargas al cuadrado	Proporción varianza.	Acumulativo
Factor 1	3.541	3.303	0.275	0.275	3.169	0.264	0.264
Factor 2	2.793	2.492	0.208	0.483	2.473	0.206	0.470
Factor 3	2.022	1.788	0.149	0.632	1.918	0.160	0.630

Nota: en este cuadro de muestran los valores de los factores cuando la solución no está rotando y cuando lo está.

Explicación de la tabla:

Esta rotación es una transformación matemática que distribuye de mejor manera las cargas de cada pregunta y que quede asignado a un solo factor, evitando el entrecruzamiento y que una pregunta cargue en dos factores. El objetivo es simplificar e interpretar mejor el modelo.

Sobre los autovalores: Todos los factores tienen autovalores mayores a 1 Según la regla de Kaiser, son factores importantes que deben retenerse. Factor 1: autovalor 3.541 es el factor más dominante.

Sobre la varianza explicada: antes de rotar el factor 1 explica el 27% de la varianza total del modelo, factor 2 explica un 20.8%, y el factor 3 explica un 14.9%, la suma total de la varianza es un 63.2%, esto es muy bueno porque se busca que explicar al menos un 60% de la varianza total del modelo.

Después de rotar la varianza se distribuye más equilibrada entre los factores, factor 1 con 26.4%, factor 2 un 20.6% y el factor 3 un 16%. La varianza acumulada es del 63%.

Interpretación general: El análisis factorial exploratorio reveló tres factores con autovalores mayores a 1, los cuales explican conjuntamente el 63% de la varianza total. Tras la rotación (Oblimin), la varianza se distribuyó más equilibradamente entre

los factores, mejorando la interpretación de los componentes latentes del instrumento.

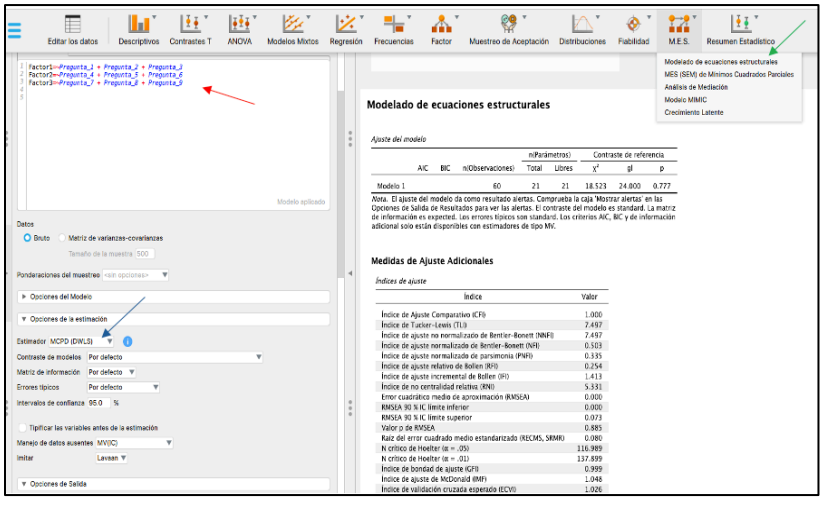
3.1.2 Análisis Factorial confirmatorio (AFC)

Luego de explorar la agrupación de las preguntas en factores, y determinar en cuantos factores se agruparon las preguntas, se debe de realizar el Análisis Factorial Confirmatorio (AFC). Este análisis forma parte de los Modelos de Ecuaciones Estructurales (MES), específicamente en su uso para medir la estructura de un cuestionario. En el caso del ejemplo 1, podríamos ir directamente al AFC ya que se saben las dimensiones y donde se agrupan las preguntas.

Vamos a suponer que tenemos un cuestionario con 12 preguntas y 60 evaluados, los factores están claros y son conocidos, el factor 1 agrupa las preguntas 1, 2 y 3, el factor 2 agrupa las preguntas 4,5 y 6 y el factor 3 agrupa las preguntas 7, 8 y 9. Por lo tanto ahora vamos al AFC para evaluar el ajuste del modelo.

A continuación se presenta el paso a paso para realizar AFC con MES en el programa JASP.

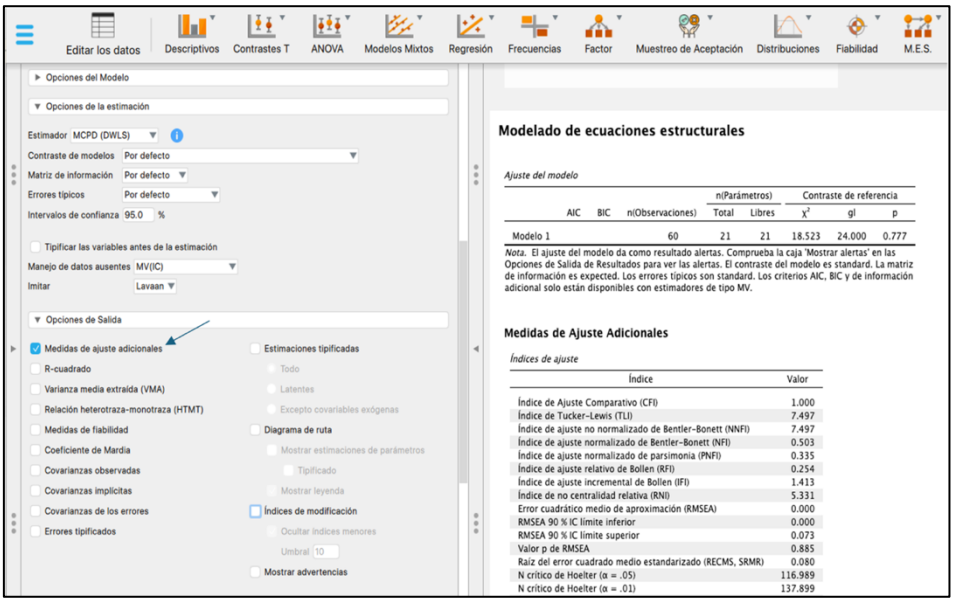
Figura 5
Seleccionando MES en JASP



Nota: la flecha verde indica la selección de Modelos de Ecuaciones Estructurales (M.E.S). la flecha azul indica el estimador DWLS, que es el estimador correcto para variables dicotómicas. Por último flecha roja indicando el modelo de ecuación que es el siguiente: **Factor=~Pregunta + Pregunta + Pregunta**

Si seguimos bajando en los parámetros de análisis se debe de indicar las medidas de ajuste adicionales como se muestra a continuación.

Figura 6
Medidas de ajuste adicionales



Nota: flecha azul indica Medidas de ajuste adicionales

A continuación se describe la salida de JASP a los resultados de los parámetros analíticos y las medidas de ajuste adicionales.

Primero se describirá el ajuste del modelo y posteriormente las Medidas de ajuste adicionales.

Tabla 13.
Ajuste del modelo

Ajuste del modelo

				n(Parámetros)		Contraste de referencia		
	AIC	BIC	n(Observaciones)	Total	Libres	χ^2	gl	p
Modelo 1			60	21	21	18.523	24.000	0.777

Nota: el modelo 1 es el modelo que colocamos en la caja de modelo de JASP. Los primeros parámetros que aparecen son AIC y BIC, estos parámetros no están reportados dado que es cuando se elige un estimador de Máxima Verosimilitud (MVS), y en este caso estamos usando DWLS, que es el indicado para variables dicotómicas (recordemos que lo dicotómico son las respuestas de los encuestados o evaluados). Al lado aparece n(Observaciones) que son los encuestados o evaluados. Luego sigue n(Parámetros) que es la cantidad de parámetros estimados en el modelo. Chi cuadrado (X^2) es la medida de diferencia entre el modelo observado y esperado. Los grados de libertad (gl) son los datos disponibles para estimar. Y el p es el p-value (0.77) si es mayor a 0.05 confirma que no existen diferencias entre el modelo observado y el esperado, por lo tanto es un buen ajuste.

Luego JASP entrega los valores de ajuste adicionales, entre varios parámetros que entrega, veremos los que son relevantes para este ejemplo.

Tabla 14.
Medidas de Ajuste Adicionales

Índices de ajuste

Índice	Valor
Índice de Ajuste Comparativo (CFI)	1.000
Índice de Tucker-Lewis (TLI)	7.497
Índice de ajuste no normalizado de Bentler-Bonett (NNFI)	7.497
Índice de ajuste normalizado de Bentler-Bonett (NFI)	0.503
Índice de ajuste normalizado de parsimonia (PNFI)	0.335
Índice de ajuste relativo de Bollen (RFI)	0.254
Índice de ajuste incremental de Bollen (IFI)	1.413
Índice de no centralidad relativa (RNI)	5.331
Error cuadrático medio de aproximación (RMSEA)	0.000
RMSEA 90 % IC límite inferior	0.000
RMSEA 90 % IC límite superior	0.073
Valor p de RMSEA	0.885
Raíz del error cuadrado medio estandarizado (RECMS, SRMR)	0.080
N crítico de Hoelter (α = .05)	116.989
N crítico de Hoelter (α = .01)	137.899
Índice de bondad de ajuste (GFI)	0.999
Índice de ajuste de McDonald (IMF)	1.048
Índice de validación cruzada esperado (ECVI)	1.026

Nota: no todos los valores representados en esta tabla son los que hay que reportar, los siguientes valores son los más adecuados para este ejemplo:

Tabla 15.
Indicadores pertinentes a variables dicotómicas

Índice	Valor obtenido	Criterio esperado	Interpretación
Índice de ajuste comparativo (CFI)	1.000	> 0.95	Ajuste perfecto del modelo
Índice de Tucker-Lewis (TLI)	0.9	0.90 – 0.95	Excelente ajuste
RMSEA (Error cuadrático medio de aproximación)	0.000	< 0.08 (ideal < 0.05)	Ajuste excelente
RMSEA IC 90% (límite inferior – superior)	0.000 – 0.073	Inferior ≈ 0, superior < 0.08	Intervalo dentro del rango aceptable
Valor p de RMSEA	0.885	> 0.05	No hay evidencia de mal ajuste
SRMR (Residuos estandarizados)	0.080	< 0.08	Justo en el límite aceptable
GFI (Índice de bondad de ajuste)	0.999	> 0.90	Ajuste excelente

Nota: Los resultados del análisis de ecuaciones estructurales muestran un ajuste excelente del modelo a los datos. Se obtuvo un CFI de 1.000, un RMSEA de 0.000 con intervalo de confianza del 90% entre 0.000 y 0.073, y un valor p del RMSEA de 0.885, lo que indica ausencia de mal ajuste. El índice SRMR fue 0.080, ubicado en el límite de aceptabilidad. El TLI mostró un valor de (0.9) asociado a un excelente ajuste del modelo. El GFI fue 0.999, lo cual confirma la solidez estructural del modelo propuesto.

¿ Qué significa cada uno de estos indicadores?

1. CFI : Índice de ajuste comparativo

Qué hace: Compara el modelo propuesto con uno que asume que todas las variables están no relacionadas.

Rango ideal: > 0.90 (excelente si > 0.95).

Interpretación: Si tu CFI es 1.000 → el modelo ajusta perfectamente comparado con uno nulo.

2. TLI : Índice de Tucker-Lewis

Qué hace: Similar al CFI pero penaliza si el modelo es muy complejo.

Rango ideal: > 0.90.

Interpretación: Si es > 1, hay problemas de identificación o saturación. No se puede interpretar correctamente.

3. RMSEA: Error cuadrático medio de aproximación

Qué hace: Mide cuán bien se aproxima el modelo a la población.

Rango ideal:

< 0.05 = excelente

0.05–0.08 = aceptable

0.10 = pobre

Interpretación: RMSEA = 0.000 → ajuste perfecto.

4. p-value de RMSEA

Qué hace: Evalúa si el RMSEA es significativamente diferente de 0.

Rango Ideal: p > 0.05

Interpretación: p = 0.885 → modelo no difiere del ajuste perfecto, ¡excelente!

5. SRMR : Residuos estandarizados medios

Qué hace: Mide la media de las diferencias entre correlaciones observadas y predichas.

Rango ideal: < 0.08.

Interpretación: SRMR = 0.080 → justo en el límite aceptable.

6. GFI : Índice de bondad de ajuste

Qué hace: Evalúa cuánta varianza y covarianza explica el modelo.

Rango Ideal: > 0.90.

Interpretación: GFI = 0.999 → ajuste excelente.

La siguiente salida de JASP es una tabla que muestra los **índices de ajuste de tamaño T**, que se utilizan para evaluar la calidad del ajuste del modelo **en función del tamaño muestral**, considerando un nivel de significancia de α = 0.05.

Tabla 16.
Índice de ajuste de tamaño T

	IAC (CFI)	REACM (RMSEA)
Estimar	0.000	0.073
Límite deficiente–aceptable	0.708	0.142

	IAC (CFI)	REACM (RMSEA)
Límite aceptable–próximo	0.796	0.121

Nota. Los estadísticos de tamaño T se calculan para $\alpha = 0.05$.

Interpretación de la tabla:

- IAC (CFI: Comparative Fit Index)

Valor estimado del modelo: 0.000 → extremadamente bajo, indicando ajuste muy deficiente del modelo comparado con un modelo nulo (sin relaciones entre variables).

El valor mínimo para considerar un ajuste "aceptable" es cercano a 0.90, y >0.95 se considera muy bueno. Aquí, ni siquiera se alcanza el límite inferior de 0.708, lo que indica que el modelo no se ajusta bien a los datos.

- REACM (RMSEA: Root Mean Square Error of Approximation)

Valor estimado del modelo: 0.073

Aunque el valor de corte clásico para un buen ajuste es < 0.05, un valor entre 0.05 y 0.08 se considera ajuste razonable o aceptable. Por lo tanto, este RMSEA de 0.073 indica que el modelo tiene un ajuste moderadamente aceptable, aunque no excelente. En esta tabla, el valor estimado está por debajo del límite deficiente–aceptable (0.142) y del límite aceptable–próximo (0.121), lo que podría indicar que los criterios fueron definidos de manera diferente al estándar (quizás por tamaño muestral o complejidad del modelo).

Nota aclaratoria: "Los estadísticos de tamaño T se calculan para α = 0.05": se refiere al nivel de significancia usado para las pruebas estadísticas que dan origen a estos índices. Un valor típico en estudios sociales y de salud.

- Conclusión

El modelo evaluado no muestra un ajuste adecuado según el CFI (0.000), lo cual es crítico.

El RMSEA es moderadamente aceptable (0.073), aunque no ideal.

Esto sugiere que el modelo necesita mejoras, posiblemente ajustando la especificación de las relaciones entre variables o incluyendo/modificando indicadores.

3.2. VALIDEZ DE CRITERIO

La validez de criterio evalúa el grado en que las puntuaciones obtenidas por un instrumento se relacionan con un criterio externo considerado como referencia válida del mismo constructo. Es una forma de validez empírica que busca comprobar si el instrumento predice, explica o se asocia con una medida independiente que representa el fenómeno que se desea evaluar (Carvajal et al., 2011; AERA, APA & NCME, 2014).

“La validez de criterio es el grado en que las puntuaciones de un test se relacionan con un criterio externo que representa el comportamiento o atributo que se desea medir” (Anastasi & Urbina, 1997, p. 114).

Tipos de validez de criterio

Concurrente: el criterio se mide al mismo tiempo que el instrumento.

Predictiva: el instrumento busca anticipar un resultado futuro relacionado con el constructo.

Por ejemplo, si un cuestionario de conocimiento en radioprotección se compara con los resultados de una prueba de certificación oficial en el mismo tema, estaríamos evaluando validez de criterio concurrente.

- **EJEMPLO DE VALIDEZ DE CRITERIO CONCURRENTE.**

Objetivo: Evaluar si los puntajes de un cuestionario de conocimientos en radiación ionizante se correlacionan con un criterio externo de desempeño académico (nota final del módulo de radiología), lo que evidenciaría validez de criterio concurrente.

Supongamos que tenemos 15 estudiantes con las siguientes puntuaciones:

Tabla 17.

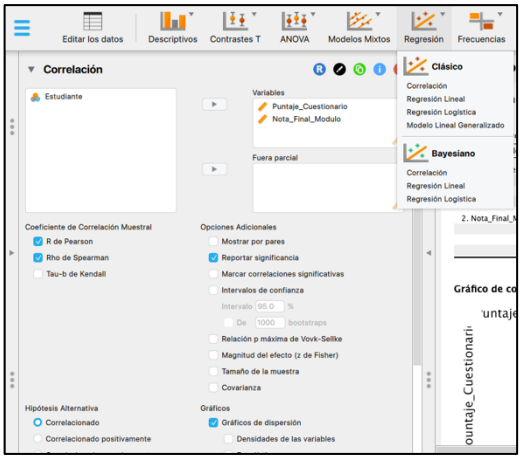
Base de datos para estimar validez de criterio concurrente

Estudiante	Puntaje Cuestionario	Nota Final Modulo
E1	10	6.5
E2	9	6.2
E3	7	5.8
E4	8	6.0
E5	6	5.5
E6	11	6.7
E7	9	6.4
E8	5	5.0
E9	6	5.3
E10	7	5.8
E11	8	6.1
E12	10	6.6
E13	4	4.7
E14	7	5.9
E15	10	6.5

La salida es JASP es la siguiente:

Figura 7.

Parámetros de JASP para el análisis de correlación entre cuestionario y nota final del módulo



Nota: Están seleccionados la R de Pearson y Rho de Spearman, además del grafico de dispersión. Hay que considerar las características de los datos en este ejemplo después de usar la prueba de Shapiro Wilk arrojo un p-valor de 0.64 por lo tanto los valores se distribuyen de manera normal y se usara la R de Pearson, en caso contrario se usaría la Rho de Spearman.

Los resultados de JASP son los siguientes:

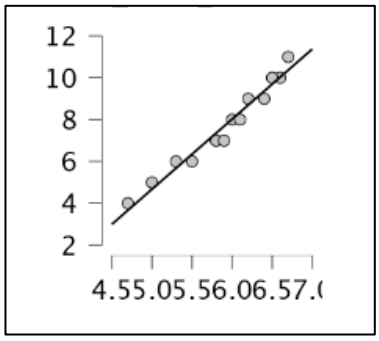
Tabla 18.
Análisis de correlación entre puntaje de cuestionario vs nota final del módulo

Tabla de Correlación

Variable		Puntaje Cuestionario	Nota Final Modulo
1. Puntaje Cuestionario	R de Pearson	—	
	Valor p	—	
	Rho de Spearman	—	
	Valor p	—	
2. Nota Final Modulo	R de Pearson	0.982	—
	Valor p	< .001	—
	Rho de Spearman	0.992	—
	Valor p	< .001	—

Nota: la R de Pearson es de 0.982 y el Valor p es <0.001, por lo tanto indica una **correlación fuerte y significativa** entre el puntaje del cuestionario y el rendimiento en el módulo de radiología, lo que **evidencia una buena validez de criterio concurrente**.

Figura 8.
Gráfico de dispersión entre las dos evaluaciones



Nota: en el eje x son los valores del módulo de radiología y en el eje y los valores de la prueba de conocimiento, se observa una correlación lineal entre ambos indicadores**.**

Conclusión: Para evaluar la validez de criterio del cuestionario de conocimientos en radiación ionizante, se compararon los puntajes obtenidos con la nota final del módulo de radiología, considerada un criterio externo relacionado. Se observó una correlación de Pearson de **r = 0.982 (p < 0.001)**, lo que indica una excelente asociación entre ambos resultados y evidencia una adecuada validez de criterio concurrente.

- **EJEMPLO DE VALIDEZ DE CRITERIO PREDICTIVO**

Es el grado en que las puntuaciones obtenidas en un instrumento permiten predecir un resultado futuro relacionado con el mismo constructo.

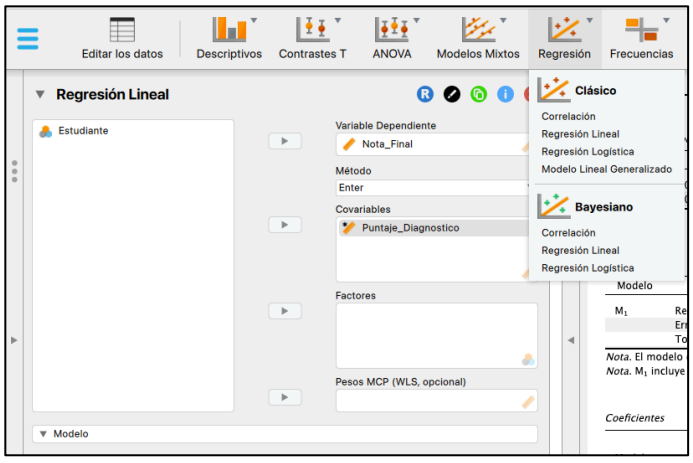
“La validez predictiva se refiere al grado en que una prueba predice un desempeño futuro relacionado con la habilidad o conocimiento que evalúa” (Anastasi & Urbina, 1997).

Siguiendo con el ejemplo de conocimiento en radiación ionizante se podría aplicar validez predictiva de la siguiente manera:

Administra el cuestionario de conocimientos en radiación al inicio del semestre (como prueba diagnóstica). Al final del semestre se recoge : Nota final del módulo, puntaje en un examen práctico y evaluación de desempeño clínico. Luego, se correlaciona los puntajes iniciales con los resultados futuros, para ver si el cuestionario predice el rendimiento académico posterior.

En JASP se realiza de la siguiente manera:

Figura 9.
Parámetros de JASP para el análisis de regresión



Nota: como variable dependiente esta la nota final y como independiente está el puntaje de diagnóstico final.

La salida de JASP es la siguiente:

Tabla 19.
Resumen del Modelo – Nota Final

Modelo	R	R^2	R^2 Ajustado	RMSE
M_0	0.000	0.000	0.000	0.598
M_1	0.982	0.964	0.961	0.118

Nota. M_1 incluye Puntaje Diagnóstico

R = 0.982: Correlación entre el puntaje diagnóstico y la nota final → muy alta.

R^2 = 0.964: El cuestionario predice el 96.4% de la varianza en la nota final → excelente validez predictiva.

R^2 ajustado = 0.961: Corrige el R^2 por el número de predictores y tamaño muestral → sigue siendo alto.

RMSE = 0.118: Error estándar de predicción; más bajo = mejor precisión.

Tabla 20.
Cálculo de ANOVA del modelo

Modelo		Suma de Cuadrados	gl	Cuadrado Medio	F	p
M_1	Regresión	4.833	1	4.833	348.158	< .001
	Error	0.180	13	0.014		
	Total	5.013	14			

Nota. El modelo de la constante se omite, ya que no se puede mostrar información importante.
Nota. M_1 incluye Puntaje Diagnostico

F = 348.158, p < 0.001: El modelo de regresión es altamente significativo, es decir, el puntaje diagnóstico tiene un efecto predictivo real sobre la nota final.

Tabla 21
Cálculo de los coeficientes de la regresión

Coeficientes

Modelo		No tipificado	Error Típico	Tipificado	t	p	IC del 95% Inferior	IC del 95% Superior
M_0	(Constante)	5.933	0.155		38.401	< .001	5.602	6.265
M_1	(Constante)	3.689	0.124		29.744	< .001	3.422	3.957
	Puntaje Diagnostico	0.288	0.015	0.982	18.659	< .001	0.254	0.321

Constante = 3.689: Valor base de la nota cuando el puntaje es 0.

B = 0.288: Por cada punto extra en el cuestionario diagnóstico, la nota final sube 0.288 puntos en promedio.

$p < 0.001$ para el predictor: La relación es estadísticamente significativa.

IC 95% [0.254, 0.321]: Intervalo estrecho, refuerza la precisión del efecto.

Conclusión: El análisis de regresión lineal mostró una fuerte validez de criterio predictiva del cuestionario de conocimientos en radiación ionizante. Se observó una **correlación de R = 0.982 y un coeficiente de determinación $R^2 = 0.964$**, lo que indica que **el cuestionario explicó el 96.4% de la varianza en el rendimiento académico final**. **El efecto del predictor fue significativo (p < .001)**, y **cada punto adicional en el cuestionario aumentó en promedio 0.288** puntos la nota final, lo que confirma su potencial como herramienta diagnóstica para **predecir** el desempeño futuro.

3.3. CONFIABILIDAD DE CONSISTENCIA INTERNA.

Existen varios métodos para calcular la confiabilidad del instrumento, dos de ellos son, Kuder-Richardson y Alpha de Cronbach.

Si la recogida de datos es dicotómica, respuesta correcta o incorrecta (0=incorrecta, 1=correcta), se debe de utilizar la Kuder-Richardson (KR-20). Cuando las respuestas son politómicas se debe de utilizar Alpha de Cronbach. No obstante, estos métodos no son recomendables para medir conocimiento. Campo-Arias y Oviedo, (2008) mencionan: *"Este tipo de coeficientes (alfa de Cronbach y KR-20) sólo se puede calcular apropiadamente a escalas que miden atributos o características y no el conocimiento sobre un tópico particular, es decir, no se puede determinar la consistencia interna a una prueba de conocimiento que se aplica en un colegio o universidad, o sea, que necesitan entrenamiento o conocimiento previo en un tópico particular."*

Los coeficientes de consistencia interna, como el alfa de Cronbach y el KR-20, se basan en la correlación entre ítems. Es decir, suponen que todos los ítems deben medir el mismo constructo o dimensión psicológica (como ansiedad, autoestima, actitudes, etc.), pero en una prueba de conocimiento, esto no se cumple necesariamente.

Por ejemplo en una escala de ansiedad, es esperable que Ítems como "me siento nervioso", "tengo dificultad para dormir", "siento palpitaciones" estén correlacionados, entonces el alfa de Cronbach mide la consistencia de un solo constructo, pero en una prueba de conocimiento sobre radio protección, un estudiante puede saber qué es el tiempo de exposición pero no saber qué es el efecto estocástico y sí saber qué equipo usa rayos X.

Aquí, la falta de correlación entre ítems no implica que la prueba esté mal construida, sino que refleja diferencias reales en el conocimiento.

En vez de estos test estadísticos se aconseja reportar el índice de facilidad del item y el de discriminación, aunque técnicamente no son pruebas de confiabilidad en sentido estricto, estos índices permiten analizar la calidad de los ítems, y son condición necesaria para una buena confiabilidad del instrumento.

En test de conocimiento, estos indicadores sustituyen o complementan la consistencia interna, porque se enfocan en la precisión y rendimiento individual de cada ítem.

Según (Thorndike & Thorndike-Christ, 2010), *“Aunque los análisis de dificultad y discriminación no constituyen medidas de confiabilidad en el sentido técnico del término, son fundamentales para evaluar la calidad de los ítems en las pruebas de rendimiento. Estos análisis permiten determinar qué tan adecuados son los ítems para diferenciar entre examinados con distintos niveles de conocimiento, contribuyendo así a la precisión global del test”*

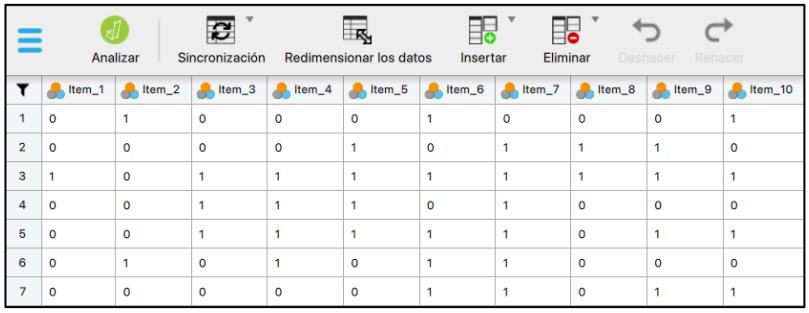
Siguiendo esta lógica, aunque en pruebas de conocimiento no se espera necesariamente alta homogeneidad entre ítems, se puede calcular el coeficiente KR-20 como una estimación de la confiabilidad general del instrumento. Este valor se

debe de interpretar con cautela, en conjunto con el análisis del índice de dificultad, índice de discriminación y juicio de expertos, tal como recomiendan Haladyna (2004) y AERA et al. (2018) (American Educational Research Association et al., 2018; Haladyna, 2004).

En JASP el cálculo de Alpha de Cronbach tiene una particularidad. En JASP, el cálculo del coeficiente KR-20 no aparece con ese nombre directamente en el menú, pero se puede obtener usando el análisis de confiabilidad (Reliability) con ítems dicotómicos, ya que KR-20 es un caso especial del alfa de Cronbach para ítems con respuestas 0 y 1.

En JASP la base de datos debe de ser según el siguiente formato:

Figura 10.
Disposición de la base de datos para el cálculo de confiabilidad



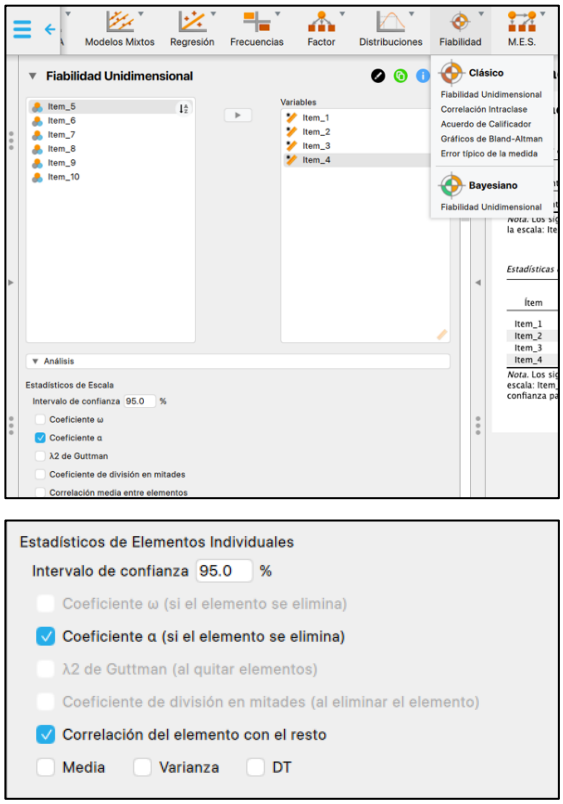
	Item_1	Item_2	Item_3	Item_4	Item_5	Item_6	Item_7	Item_8	Item_9	Item_10
1	0	1	0	0	0	1	0	0	0	1
2	0	0	0	0	1	0	1	1	1	0
3	1	0	1	1	1	1	1	1	1	1
4	0	0	1	1	1	0	1	0	0	0
5	0	0	1	1	1	1	1	0	1	1
6	0	1	0	1	0	1	1	0	0	0
7	0	0	0	0	0	1	1	0	1	1

El análisis puede ser total o por dimensiones según el análisis anterior si es que los ítems se agrupan en una o en varias dimensiones.

En JASP simularemos que son los primeros 4 ítems para una dimensión en particular.

Figura 11.

Parámetros de análisis para el cálculo de confiabilidad



La salida de JASP es la siguiente:

Tabla 22.

Estadísticas de confiabilidad

			IC del 95%	
Coeficiente	Estimar	Error típico	Lower	Upper
Coeficiente α	0.194	0.177	−0.153	0.542

Nota. Los siguientes ítems se correlacionan negativamente con la escala: Item_2, Item_4.

Interpretación:

Un alfa de 0.194 es extremadamente bajo, lo que indica muy baja consistencia interna entre los ítems.

Además, el intervalo de confianza cruza por valores negativos, lo cual no es válido estadísticamente: esto suele significar que:

Hay ítems mal redactados o contradictorios, o que los ítems no están midiendo el mismo constructo (no unidimensional).

Nota: también se indica que Item_2 e Item_4 se correlacionan negativamente con la escala, lo cual puede estar distorsionando el α total.

Tabla 23.

Estadística de confiablidad si el ítem se elimina

Estadísticas de confiabilidad de ítems individuales frecuentes

	Coeficiente α (si ítem eliminado)			Correlación del elemento con el resto		
Ítem	Estimar	IC inferior al 95%	IC superior 95%	Estimar	IC inferior al 95%	IC superior 95%
Item_1	−0.023	−0.516	0.470	0.211		
Item_2	0.288	−0.060	0.636	−0.008		
Item_3	0.129	−0.197	0.455	0.110		
Item_4	0.181	−0.116	0.478	0.073		

Nota. Los siguientes ítems se correlacionan negativamente con la escala: Item_2, Item_4. No está disponible el intervalo analítico de confianza para la correlación elemento–resto.

Tabla 24.

Explicación de los resultados de la tabla anterior

Ítem	α si se elimina	Correlación ítem-total
Item_1	-0.023	0.211
Item_2	0.288	**-0.008** ❌
Item_3	0.129	0.110
Item_4	0.181	**0.073 (bajo)**

Análisis:

Item_2 tiene correlación negativa con el total: esto significa que, a mayor puntaje en ese ítem, menor puntaje total → está invirtiendo el sentido del constructo. Podría estar mal redactado o requiere invertir su puntuación.

Item_4 también tiene correlación muy baja (0.073).

Item_1 y Item_3 tienen correlaciones bajas pero positivas.

El alfa no mejora significativamente al eliminar ninguno de los ítems, aunque Item_2 parece ser el más problemático.

Una vez determinado el coeficiente de confiabilidad interna del instrumento, como el KR-20 (en pruebas de ítems dicotómicos) o el alfa de Cronbach (en pruebas con ítems politómicos), es necesario avanzar hacia un nivel de análisis más específico: el análisis individual por ítem. Este análisis incluye el cálculo del índice de facilidad (o dificultad inversa) y el índice de discriminación, ambos fundamentales para el refinamiento técnico de una prueba.

Aunque los índices de dificultad y discriminación no son coeficientes de confiabilidad en sentido estricto, sí son herramientas esenciales para evaluar la calidad psicométrica de cada ítem. Tal como señalan Haladyna (2004) y (Brookhart & Nitko, 2015), un instrumento puede presentar un coeficiente de confiabilidad global aceptable, pero incluir ítems que no cumplen con criterios mínimos de calidad técnica. Por lo tanto, no basta con calcular la confiabilidad total, sino que se debe garantizar que cada ítem aporte positivamente al propósito de medición de la prueba.

El índice de dificultad permite identificar ítems demasiado fáciles o difíciles, lo que puede limitar la capacidad de discriminar entre niveles de conocimiento. El índice de discriminación, por su parte, indica en qué medida un ítem distingue adecuadamente entre estudiantes con alto y bajo rendimiento total en la prueba.

Estos análisis contribuyen a mejorar la validez interna del instrumento y a optimizar su capacidad para identificar verdaderamente diferencias en el conocimiento entre los sujetos evaluados, incluso si no afectan directamente al coeficiente de confiabilidad global.

Análisis psicométrico por ítem

Una vez evaluada la confiabilidad global del instrumento, es necesario analizar el comportamiento individual de cada ítem mediante dos indicadores clave: el **índice de dificultad** y el **índice de discriminación**. Estos análisis no forman parte directa de la confiabilidad total, pero son fundamentales para decidir si un ítem debe conservarse, revisarse o eliminarse del instrumento final.

Indice de dificultad

El índice de dificultad (también llamado índice de facilidad) corresponde a la proporción de estudiantes que responden correctamente un ítem. Se calcula como:

F = (número de respuestas correctas / total de respuestas)

- F cercano a 1 → ítem muy fácil (todos responden bien)
- F cercano a 0 → ítem muy difícil (nadie responde bien)

Se recomienda conservar ítems con dificultad **entre 0.30 y 0.80** (Nitko & Brookhart, 2015).

Indice de discriminación

Este índice mide la capacidad del ítem para distinguir entre estudiantes con alto y bajo desempeño global. Se calcula como la diferencia entre el porcentaje de aciertos en el grupo superior y el grupo inferior (usualmente los 27% mejores y peores puntajes).

D = (proporción de aciertos en el grupo alto – proporción de aciertos en el grupo bajo)

- D > 0.40 → excelente discriminación
- D entre 0.20 y 0.39 → aceptable
- D < 0.20 → baja discriminación
- D < 0 → ítem problemático (mejores estudiantes fallan más que los peores)

Ejemplo interpretativo:

Ítem	Índice de dificultad	Índice de discriminación	Interpretación
1	0.65	0.42	Ítem adecuado y discriminativo
2	0.92	0.10	Muy fácil, poco útil
3	0.30	−0.15	Ítem problemático, revisar

Según Haladyna (2004), todo ítem debe analizarse con estos dos indicadores antes de ser incluido en la versión definitiva del instrumento. Esto garantiza que cada pregunta no solo mida el contenido esperado, sino que lo haga de forma clara, útil y diferenciadora.

Algunos de los criterios que corrientemente se utilizan en la toma de decisiones, en relación con la facilidad y discriminación de los ítems, son los siguientes:

1. Seleccionar ítems que sean moderadamente fáciles (entre 41 y 60 por ciento) y que, al mismo tiempo, tengan una discriminación moderada o muy alta (índices entre 0,41 y 1.0).
2. Seleccionar ítems con índices de facilidad y discriminación moderados (índices entre 41 y 60 por ciento).
3. Seleccionar ítems con índices de facilidad alto, moderado y bajo, manteniendo un nivel de discriminación entre bajo y moderado.
4. Seleccionar ítems con índices de facilidad baja (reactivos difíciles) y que, al mismo tiempo, tienen un nivel de discriminación moderado.

En relación con estos criterios, es importante señalar que ninguno de ellos es mejor o peor que otro; la decisión acerca de cuál de ellos se debe utilizar en momento dado, depende del tipo de instrumento y del propósito del investigador.

3.3.1. CONFIABILIDAD TEMPORAL: TEST – RETEST

Evalúa la estabilidad temporal de un instrumento, es decir, si produce resultados similares cuando se aplica en dos momentos diferentes, bajo condiciones equivalentes.

"El método test-retest evalúa la confiabilidad de un instrumento midiendo su capacidad de producir resultados estables a lo largo del tiempo" (Hernández-Sampieri y Mendoza, 2018).

Se calcula mediante una correlación (por lo general, Pearson o Spearman) entre los puntajes obtenidos por los mismos sujetos en dos aplicaciones sucesivas del instrumento.

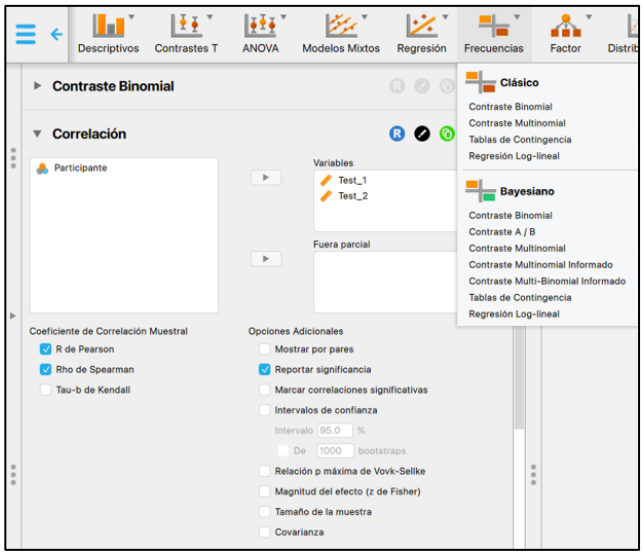
- **Ejemplo de confiabilidad test-retest en JASP**

Objetivo del análisis: Evaluar la estabilidad temporal del cuestionario sobre conocimientos en radiación ionizante, aplicándolo dos veces (Test 1 y Test 2) a los mismos estudiantes con una semana de diferencia.

Este cálculo está en el contexto de medir correlación de estos dos test en tiempos distintos, por lo tanto, se debe de usar el test de Pearson si los datos son **paramétricos** y Spearman si los datos son **no paramétricos u ordinales.**

A continuación, se detallan los pasos para el cálculo de correlación de Pearson o Spearman en JASP

Figura 12.
Parámetros de JASP para el análisis de correlación



La salida en JASP es la siguiente:

Tabla 25.
Tabla de Correlación

Variable		Test_1	Test_2
1. Test_1	R de Pearson	—	
	Valor p	—	
	Rho de Spearman	—	
	Valor p	—	
2. Test_2	R de Pearson	0.939	—
	Valor p	< .001	—
	Rho de Spearman	0.944	—
	Valor p	< .001	—

JASP hace el cruce entre Test 1 y Test 2 arrojando una correlación muy alta entre ambas pruebas. Como el valor de R de Pearson es de 0.939 con un Valor p < .001, la confiabilidad test-retest es excelente, lo cual sugiere que el cuestionario mide de manera estable y reproducible el nivel de conocimiento.

“Una correlación test-retest ≥ 0.80 es considerada buena confiabilidad temporal en contextos educativos” (Carvajal et al., 2011).

Conclusión test – retest: Se aplicó el cuestionario en dos momentos diferentes con una semana de intervalo a 10 estudiantes. El análisis test-retest mostró una correlación de Pearson de $r = 0.94$ ($p < 0.001$), lo que indica alta estabilidad temporal. Este resultado respalda la confiabilidad del cuestionario como instrumento para medir conocimientos sobre radiación ionizante en contextos educativos.

Habíamos mencionado en el primer apartado acerca de las diferencias entre la metodología y las medidas psicométricas entre construir un instrumento para medir conocimiento y una prueba institucional o académica.

- **DIFERENCIAS ENTRE LA CONSTRUCCION DE UN CUESTIONARIO Y UNA PRUEBA INSTUTUCIONAL.**

¿Miden los mismo?

Aspecto	Instrumento de investigación	Prueba universitaria tradicional
Propósito	Medir una variable (constructo) para generar conocimiento científico o tomar decisiones educativas.	Evaluar el aprendizaje del contenido de una asignatura.
Enfoque	Validación rigurosa del instrumento como herramienta de medición.	Validación práctica y curricular del contenido evaluado.
Usuarios finales	Investigadores, docentes, comunidad académica.	Docentes y estudiantes.

- **Diferencias metodológicas**

Dimensión	Instrumento de investigación	Prueba universitaria
Fundamentación teórica	Requiere revisión de literatura y definición explícita del constructo.	Basada en el programa de la asignatura y los contenidos vistos.

Dimensión	Instrumento de investigación	Prueba universitaria
Diseño de ítems	Ítems representativos de dimensiones del constructo, validados por expertos.	Ítems alineados con objetivos de clase o contenidos de la unidad.
Juicio de expertos	Obligatorio para asegurar validez de contenido (Aiken, CVC, etc.).	Puede usarse, pero no siempre se aplica formalmente.
Prueba piloto	Requiere piloto con análisis psicométrico antes de su aplicación definitiva.	Rara vez se hace; puede aplicarse directamente.
Estandarización	Alta: se espera aplicación, corrección e interpretación estandarizada.	Moderada o baja: varía entre docentes o instituciones.

- **Diferencias estadísticas**

Análisis	Instrumento de investigación	Prueba universitaria
Confiabilidad	α de Cronbach, KR-20, test-retest, análisis de ítem.	Se calcula a veces, pero no siempre.
Validez	Evidencia de contenido, constructo, criterio.	Validez basada en juicio del docente o rendimiento esperado.
Análisis de ítems	Discriminación, dificultad, índice de homogeneidad.	Pocas veces se aplica formalmente.
Normas de interpretación	Baremos o criterios científicos.	Escalas de calificación institucional (ej. 1 a 7, 0 a 100).

¿Se pueden usar ambos tipos de pruebas en investigación?

Sí, pero una prueba universitaria no puede considerarse automáticamente un instrumento válido y confiable para investigación si no ha pasado por los procedimientos mínimos (validez, confiabilidad, análisis de ítems). Si se desea usar una prueba universitaria como parte de un estudio, se deben adaptar sus ítems y someterlos a validación formal.

"*Los estándares establecen que un instrumento válido debe estar respaldado por evidencia empírica de validez, confiabilidad, uso apropiado y documentación técnica. Esta rigurosidad no se exige en evaluaciones curriculares comunes, cuyo propósito principal es pedagógico, no científico*" (American Educational Research Association et al., 2018).

"*Una diferencia crítica entre pruebas educativas y pruebas científicas radica en el tratamiento de la confiabilidad. Mientras en las primeras basta con que cumplan una*

función evaluativa formativa, en las segundas es indispensable demostrar empíricamente que los resultados son consistentes y replicables" (Anastasi & Urbina, 1998).

"Una prueba construida por un docente para fines pedagógicos no puede considerarse válida para investigación si no ha pasado por procesos de validación empírica, ya que sus resultados podrían ser sesgados, poco consistentes o irrelevantes para el constructo teórico" (Hernandez-Sampieri & Mendoza, 2018).

A diferencia del primer ejemplo, centrado en la validación de un instrumento conceptual para medir un atributo cognitivo (el conocimiento), este segundo caso explora la validación de una herramienta tangible: modelos anatomo-patológicos impresos en 3D con fines educativos.

En este escenario, el proceso de validación requiere combinar la evaluación por expertos en contenido (validación de estructura y fidelidad anatómica) con métodos de concordancia interjueces y análisis estadísticos adaptados a objetos físicos. Esta sección presenta una ruta metodológica distinta, que incluye indicadores como la V de Aiken, el coeficiente de correlación intraclase (ICC) y pruebas no paramétricas para la comparación entre modelos.

Este abordaje muestra cómo aplicar criterios de validez y confiabilidad a recursos de enseñanza no tradicionales, con foco en su pertinencia anatómica, claridad morfológica y utilidad pedagógica.

4. EJEMPLO 2: VALIDACIÓN DE UNA HERRAMIENTA INTERACTIVA PARA EL APRENDIZAJE

Revisamos anteriormente como se construye y valida un constructo para medir una variable que no se puede medir directamente como es el “conocimiento” y se propuso una metodología según la literatura revisada. En este contexto de la validación de instrumentos, podemos encontrarnos en otro situación y por lo tanto, no necesariamente seguir las mismas pautas y análisis estadístico revisado anteriormente.

En un escenario donde estamos fabricando modelos anatómicos impresos en 3D, en una primera etapa debemos de hacer una validación de contenido, esto significa validar los modelos anatómicos desde el punto de vista que representen lo que realmente se imprimió. No obstante, como el propósito final es enseñar anatomía patológica con estos modelos, una vez validado como modelos anatómicos debemos de crear una metodología de aprendizaje con los modelos. Para posteriormente evaluarlos con un cuestionario de preguntas y estas también deben de ser validadas tal cual como se describe en el ejemplo 1, ya que, en esta etapa se estaría midiendo el conocimiento que compete el propósito de los modelos impresos.

La situación es la siguiente: se quieren imprimir en 3D modelos anatomo-patológicos de columna lumbar con el objetivo de contribuir al aprendizaje de anatomía patológica en estudiantes de la salud.

Luego de identificar las patologías a través de Tomografía Computada, que eligieron los investigadores, se procederá a imprimir estos modelos en 3D.

Como toda investigación lo primero que se debe de realizar es una búsqueda profunda de la evidencia existente para indagar acerca de cómo se ha validado este tipo de instrumentos y se llega a la siguiente metodología:

A través de juicio por expertos (3) se evaluaron los modelos con preguntas generales a todos los modelos y preguntas específicas para cada patología impresa. Los jueces deben evaluar según escala de Likert: 1 = Muy bajo cumplimiento | 2 = Bajo

cumplimiento | 3 = Alto cumplimiento | 4 = Muy alto cumplimiento

4.1 PRUEBAS ESTADÍSTICAS PARA LA VALIDACIÓN DE MODELOS 3D

- **Media y desviación estándar por ítem o modelo**

Objetivo: Conocer la tendencia central y dispersión en las valoraciones.

Pregunta	Media	Desviación estándar
¿Agujeros raquídeos correctos? (Modelo 1)	3.7	0.58

Esto te permite identificar ítems "controvertidos" (mucha variabilidad entre jueces).

- **Coeficiente V de Aiken**

Objetivo: Estimar la validez de contenido ponderando la percepción de los jueces sobre la relevancia de cada ítem (el grado de acuerdo entre jueces sobre la relevancia de los ítems).

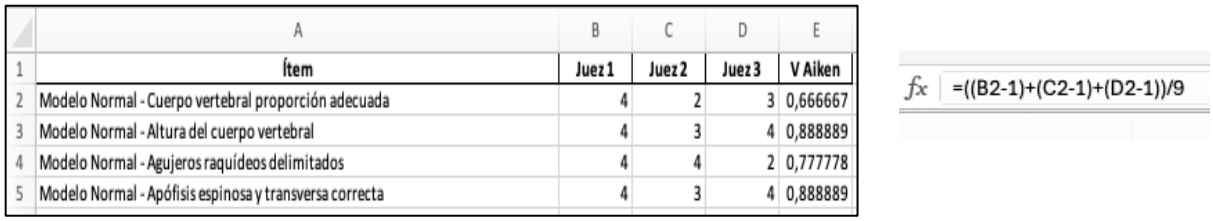
Se calcula así:

$$VC = \frac{\sum s}{n(c-1)}$$

Donde:

- s = r - l = diferencia entre la puntuación dada por el juez y el valor mínimo de la escala.
- n = número de jueces.
- c = número de categorías de la escala (por ejemplo, 4 si se usa de 1 a 4).

Figura 13.
representación del cálculo de la V de Aiken en Excel .



	A	B	C	D	E
1	Ítem	Juez 1	Juez 2	Juez 3	V Aiken
2	Modelo Normal - Cuerpo vertebral proporción adecuada	4	2	3	0,666667
3	Modelo Normal - Altura del cuerpo vertebral	4	3	4	0,888889
4	Modelo Normal - Agujeros raquídeos delimitados	4	4	2	0,777778
5	Modelo Normal - Apófisis espinosa y transversa correcta	4	3	4	0,888889

¿Cómo de interpreta los valores de V de Aiken?

Tabla 26.

Interpretación de V de Aiken

Valor de V de Aiken	Interpretación de validez de contenido
≥ 0.90	Muy alta validez
0.80 – 0.89	Alta validez
0.70 – 0.79	Validez aceptable con reservas
< 0.70	Baja validez; ítem debe ser revisado

4.2. Concordancia entre jueces (consistencia), Coeficiente de Correlación Intraclase (ICC) ¿Qué mide el ICC en evaluaciones por jueces expertos?

El coeficiente de correlación intraclase (ICC) es un índice estadístico que cuantifica el grado de acuerdo o concordancia entre diferentes evaluadores cuando califican los mismos objetos o sujetos.

En el contexto de modelos anatómicos 3D evaluados por jueces expertos, el ICC refleja qué tan consistentes son las puntuaciones otorgadas por los distintos jueces a cada modelo. Un ICC elevado indica que los expertos tienden a dar calificaciones similares a un mismo modelo (alta concordancia), mientras que un ICC bajo sugiere que existen discrepancias importantes entre las evaluaciones de los jueces (baja concordancia o fiabilidad). En otras palabras, el ICC mide la proporción de la variabilidad total de las evaluaciones que se debe a diferencias reales entre los

modelos (p. ej., modelos mejores vs. peores) frente a la variabilidad atribuible a las diferencias entre evaluadores o al error de medida

. Es importante destacar que el ICC está especialmente diseñado para situaciones con múltiples observadores y datos cuantitativos. A diferencia de una correlación de Pearson simple (útil solo para comparar dos evaluadores), el ICC puede involucrar a más de dos jueces y evalúa directamente la concordancia en las puntuaciones, no solo la asociación lineal.

Por ello, el ICC es una de las herramientas más recomendadas para cuantificar la fiabilidad Inter observador en evaluaciones subjetivas realizadas por expertos

En resumidas cuentas, en la validación de modelos 3D por jueces, el ICC indica la consistencia con la que los expertos coinciden en sus juicios sobre la calidad, exactitud o valor educativo de los modelos.

Interpretación de los valores del ICC

El valor del ICC oscila teóricamente entre 0 y 1 (aunque puede obtener valores ligeramente negativos si la discordancia entre evaluadores es muy alta). Un valor próximo a 1.0 implica concordancia casi perfecta entre los jueces, mientras que un valor cercano a 0 indica que la concordancia no es mejor que la obtenida al azar (es decir, cada juez califica de forma independiente sin un patrón común). En casos extremos, un ICC negativo sugiere que las diferencias entre evaluadores superan incluso lo esperable por azar, reflejando desacuerdo sistemático. Para interpretar concretamente el ICC en términos de calidad de la fiabilidad entre evaluadores, suelen emplearse categorías estándar. Por ejemplo, según criterios frecuentemente citados (como los de Cicchetti o los de Koo y Li), se pueden definir rangos de interpretación así:

ICC < 0.50: Fiabilidad pobre o muy baja concordancia entre jueces (inaceptable).

ICC entre 0.50 y 0.75: Concordancia moderada (podría considerarse aceptable en ciertos contextos, pero no óptima).

ICC entre 0.75 y 0.90: Concordancia buena o alta fiabilidad Inter evaluador.

ICC > 0.90: Concordancia excelente, evaluaciones prácticamente coincidentes entre los expertos.

En el contexto de evaluación de modelos 3D, un ICC alto significa que los expertos tienen un criterio uniforme: los modelos que un juez considera de alta calidad (anatómica o educativa) también reciben calificaciones altas de los demás jueces, y lo mismo ocurre con los modelos menos logrados. Por el contrario, un ICC bajo sugiere que los jueces no están de acuerdo en sus valoraciones – por ejemplo, un mismo modelo podría ser calificado como excelente por un experto pero regular por otro, lo que implica falta de consenso.

Valores aceptables de ICC en la evaluación de modelos 3D

En estudios de validación con jueces expertos, ¿qué tan alto debe ser el ICC para considerarse adecuado? Si bien depende del contexto y de lo que se esté evaluando, generalmente se espera que la fiabilidad Inter evaluador sea lo suficientemente alta para afirmar que existe acuerdo sustancial entre los expertos. En el ámbito educativo y anatómico, donde se busca asegurar que un modelo 3D es evaluado consistentemente, suele aspirarse al menos a un nivel "bueno" de acuerdo. Por tanto, valores de ICC por encima de 0.75 se interpretan como indicativos de una concordancia satisfactoria entre los jueces.

Autores consideran que un ICC ≥ 0.80 representa ya un nivel aceptable de fiabilidad en la mayoría de aplicaciones(Koo & Li, 2016). En cambio, valores bajos de ICC (por ejemplo < 0.5) no serían aceptables en una validación de contenido seria, ya que evidenciarían que los expertos no concuerdan en sus juicios. Un ICC en rango pobre o incluso moderado podría sugerir problemas, quizá los criterios de evaluación no están claros, los jueces necesitan entrenamiento adicional, o los modelos generan interpretaciones variadas. Por ejemplo, un ICC de 0.38 se consideraría muy limitado y

pondría en duda la confiabilidad del proceso evaluativo (en algunos estudios se calificaría de "acuerdo pobre-flojo" alrededor de 0.4)

En suma, para considerar válido (en términos de contenido) y confiable el juicio experto sobre los modelos 3D, típicamente se exige un ICC lo suficientemente alto (idealmente en el rango bueno a excelente). Si el ICC obtenido fuera bajo, sería difícil defender la consistencia de las evaluaciones y, por ende, la uniformidad de criterios en la validación de esos modelos.

Justificación:

Al llevar a cabo una validación de contenido (por ejemplo, validar modelos anatómicos impresos en 3D como recursos educativos), es fundamental no solo reunir la opinión de expertos, sino también demostrar cuantitativamente que esas opiniones son consistentes entre sí. Aquí es donde el ICC juega un papel clave. Su uso se justifica por varias razones:

Evidencia de fiabilidad entre expertos: El ICC proporciona una medida numérica de cuánto concuerdan los jueces en sus evaluaciones. En un estudio de validación de contenido, esto aporta evidencia de que la valoración del material (el modelo 3D) no varía arbitrariamente de un experto a otro, sino que hay un criterio común. Un alto ICC sustenta que el instrumento o recurso evaluado es robusto a los cambios de evaluador, es decir, que diferentes expertos llegan a conclusiones similares.

Esta estabilidad en las calificaciones refuerza la confianza en que los modelos están bien diseñados en términos anatómicos y educativos, ya que múltiples especialistas independientes coinciden en su apreciación.

Reducción del error y rigor metodológico: Si todos los jueces aplican criterios semejantes, el error de medición debido a la subjetividad humana se reduce. Al reportar el ICC, los investigadores demuestran que han cuantificado ese aspecto de consistencia, lo cual añade rigor metodológico al estudio. Esto convierte al ICC en una

elección respaldada por la literatura para estudios donde intervienen evaluaciones subjetivas de expertos (Bobak et al., 2018; Koo & Li, 2016). El siguiente cuadro resume las características entre la V de Aiken y el ICC.

¿Cuáles son las diferencias entre la v de Aiken y el ICC?

Tabla 27.

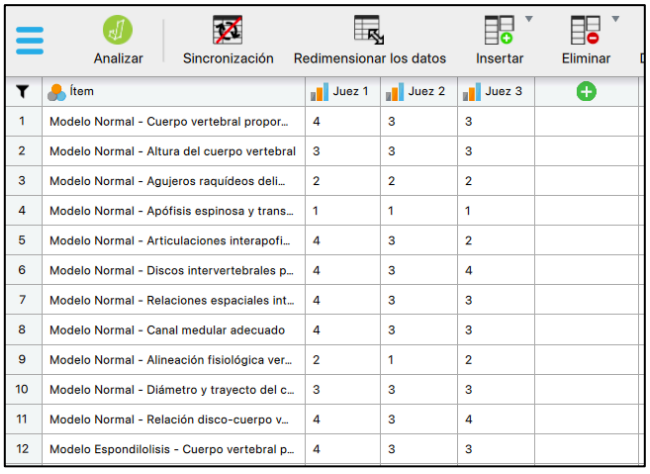
Diferencias en la V de Aiken e ICC

Característica	V de Aiken	ICC (Intraclass Correlation Coefficient)
¿Qué mide?	**Validez de contenido** por ítem: qué tan relevante, clara, coherente o suficiente es una pregunta.	**Concordancia** entre evaluadores: cuánto coinciden en sus puntuaciones generales.
Tipo de análisis	Evaluación **de cada ítem** por parte de jueces.	Evaluación **del nivel de acuerdo global entre jueces**.
Unidad de análisis	**Ítem individual**.	**Conjunto completo de ítems o evaluaciones**.
Variabilidad	No evalúa dispersión entre jueces, sino puntuación respecto al ideal.	Sí evalúa variabilidad entre jueces y dentro de los ítems.
Requiere transformación?	Sí. Convierte la puntuación dada en una razón estandarizada.	No. Trabaja directamente con los valores numéricos.
Interpretación	Valor entre 0 y 1. V ≥ 0.80 indica buena validez de contenido.	Valor entre 0 y 1. ICC ≥ 0.75 indica buena fiabilidad entre jueces.
Cuántos jueces requiere?	3 o más. Idealmente 5 o más.	2 o más, pero mejor con 3 o más.
Uso más frecuente	En validación de instrumentos, ítems de encuestas, rúbricas, etc.	En estudios de confiabilidad Inter evaluador, análisis de repetibilidad.
¿Qué informa?	Si cada pregunta es válida y debe ser conservada o modificada.	Si los jueces son coherentes entre sí en sus evaluaciones.

4.2.1 Cálculo del ICC en JASP

Figura 14.

Base de datos en JASP para el cálculo de ICC



	Ítem	Juez 1	Juez 2	Juez 3
1	Modelo Normal - Cuerpo vertebral propor...	4	3	3
2	Modelo Normal - Altura del cuerpo vertebral	3	3	3
3	Modelo Normal - Agujeros raquídeos deli...	2	2	2
4	Modelo Normal - Apófisis espinosa y trans...	1	1	1
5	Modelo Normal - Articulaciones interapofi...	4	3	2
6	Modelo Normal - Discos intervertebrales p...	4	3	4
7	Modelo Normal - Relaciones espaciales int...	4	3	3
8	Modelo Normal - Canal medular adecuado	4	3	3
9	Modelo Normal - Alineación fisiológica ver...	2	1	2
10	Modelo Normal - Diámetro y trayecto del c...	3	3	3
11	Modelo Normal - Relación disco-cuerpo v...	4	3	4
12	Modelo Espondilolisis - Cuerpo vertebral p...	4	3	3

La salida de JASP es la siguiente:

Tabla 28.

Valores de Correlación intraclase (ICC)

Tipo	Estimación por Punto	IC inferior al 95%	IC superior 95%
ICC1,1	0.460	0.298	0.613

Nota. 55 sujetos y 3 calificadores/medidas. Tipo de CCI según la referencia de Shrout & Fleiss (1979).

V de Aiken: se aplica a **cada pregunta**, como: "¿La representación del canal vertebral es clara?" Resultado: V = 0.92 → válida.

ICC: se aplica al **conjunto de evaluaciones completas** de los jueces. Resultado: ICC = 0.88 → los jueces concuerdan entre sí → fiabilidad alta.

Redacción para la investigación:

Se calculó la V de Aiken por ítem para determinar la validez de contenido de los modelos anatomo-patológicos en impresión 3D, y el coeficiente de correlación intraclase (ICC) para estimar la consistencia Inter evaluador entre los jueces expertos. Ambos índices mostraron valores superiores a 0.80, lo que indica alta validez del contenido y buena concordancia entre evaluadores, respectivamente.

4.2.2. Comparaciones entre modelos

Si se quiere comparar si un modelo fue mejor evaluado que otro:
Pruebas no paramétricas:

- **Friedman** (cuando los mismos jueces evalúan varios modelos).
- **Contraste de Conover (prueba post hoc)** (si agrupas ítems por tipo de modelo y quieres ver si hay diferencias significativas).

La prueba de Friedman es una prueba de ANOVA no paramétrica y en este caso es para saber si existen diferencias significativas de la evaluación entre modelos. Para ello debemos de crear una nueva plantilla, colocando los modelos en las columnas como se muestra a continuación (al revés del cálculo de la V de Aiken):

Figura 15.

Base de datos para el cálculo del test de Friedman en JASP

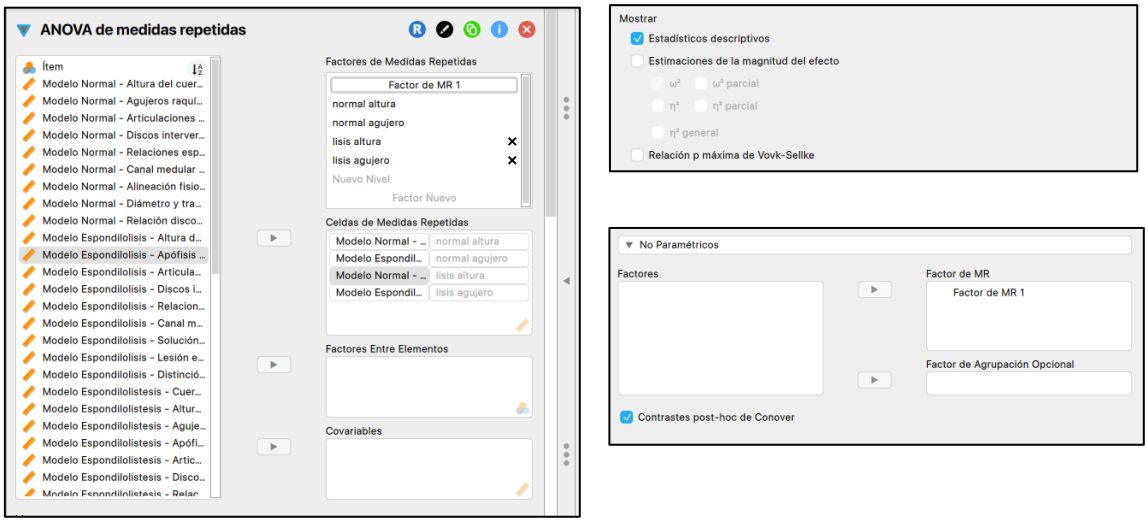
Ítem	Modelo Norm	Modelo Norm	Modelo Norm	Modelo Norm	Modelo Norm	Modelo Norm	Modelo Norm	Modelo Norm	Modelo Norm
Juez 1	4	4	4	4	4	4	4	4	4
Juez 2	2	3	4	3	3	3	3	3	3
Juez 3	3	4	2	4	4	4	4	4	4

Esto se lleva a JASP para el cálculo del test de Friedman.

En JASP se debe de seleccionar ANOVA de medidas repetidas

Figura 16.

Parámetros de análisis del test de Friedman en JASP



Luego JASP arrojará los siguientes resultados:

4.2.3 ANOVA de medidas repetidas

Tabla 29.

Comparación de modelos ANOVA

Efectos Dentro de los Sujetos

Casos	Suma de Cuadrados	gl	Cuadrado Medio	F	p
Factor de MR 1	1.000	3	0.333	4.000	0.070
Residuals	0.500	6	0.083		

Nota. Suma de Cuadrados Tipo III. El valor p es 0.07 (mayor que 0.05) por lo tanto no existen diferencias significativas entre los modelos.

4.2.4. Descriptivos

Tabla 30.

Estadísticos descriptivos

Factor de MR 1	N	Media	DT	ET	Coeficiente de variación
normal altura	3	3.000	1.000	0.577	0.333
normal agujero	3	3.667	0.577	0.333	0.157
lisis altura	3	3.667	0.577	0.333	0.157
lisis agujero	3	3.667	0.577	0.333	0.157

Nota: Podemos observar la media de los modelos, por lo tanto podemos ver los mejores y peores evaluados, con su respectivo coeficiente de variación (variabilidad de los puntajes de los jueces)

4.2.5. No Paramétricos

Tabla 31.

Test de Friedman (modelo no paramétrico de ANOVA)

Contraste de Friedman

Factor	X^2_F	gl	p	W de Kendall
Factor de MR 1	6.000	3	0.112	0.667

El valor o es de 0.112 por lo tanto no hay diferencias significativas entre los modelos. No obstante, podemos hacer una prueba post hoc para ver si entre cada modelo hay diferencias significativas.

Contraste de Conover

Tabla 32.

Comparación de modelos prueba post hoc de Conover

Comparaciones Post-hoc de Conover - Factor de MR 1

		Estadístico-T	gl	W_i	W_j	r_{rb}	p	p_{Bonf}	p_{Holm}
normal altura	normal agujero	2.828	6	4.500	8.500	-1.000	0.030	0.180	0.180
	lisis altura	2.828	6	4.500	8.500	-1.000	0.030	0.180	0.180
	lisis agujero	2.828	6	4.500	8.500	-1.000	0.030	0.180	0.180
normal agujero	lisis altura	0.000	6	8.500	8.500	NaN	1.000	1.000	1.000
	lisis agujero	0.000	6	8.500	8.500	NaN	1.000	1.000	1.000
lisis altura	lisis agujero	0.000	6	8.500	8.500	NaN	1.000	1.000	1.000

¿Qué muestra esta tabla?

Es una **comparación por pares** entre diferentes modelos 3D evaluados. Cada fila indica si hay una diferencia significativa entre dos modelos.

Columna	Significado
Estadístico-T	Valor de la prueba de comparación entre pares
p	Valor p crudo (sin corrección)
pBonf	Valor p ajustado por Bonferroni (controla error tipo I por múltiples pruebas)
pHolm	Valor p ajustado por Holm (más potente que Bonferroni)

¿Hay diferencias significativas?

Busca valores **donde pBonf o pHolm sean menores a 0.05.**

En la tabla:

Comparación	p	pBonf	pHolm	¿Diferencia significativa?
normal altura vs. normal agujero	0.030	0.180	0.180	❌ No (ajustado > 0.05)
normal altura vs. lisis altura	0.030	0.180	0.180	❌ No
normal altura vs. lisis agujero	0.030	0.180	0.180	❌ No
Las demás comparaciones	1.000	1.000	1.000	❌ No

Resultado:

Aunque algunas comparaciones tienen **p = 0.030**, **ninguna es significativa** después de aplicar correcciones por múltiples comparaciones (**Bonferroni o Holm**).

¿Entonces, cuál modelo fue el mejor?

1. Ir a los **rangos promedio** que entregó JASP en el análisis descriptivo.
2. El modelo con el **rango promedio más alto** es el que fue **mejor evaluado** por los jueces.
3. La prueba de Conover solo dice si **esas diferencias fueron estadísticamente significativas**.

¿Qué se puede concluir con esta tabla?

"Aunque el análisis post hoc de Conover mostró diferencias iniciales con valores p < 0.05 en algunas comparaciones (por ejemplo, entre el modelo de altura normal y otros), dichas diferencias no resultaron estadísticamente significativas tras aplicar corrección por Bonferroni y Holm. Por lo tanto, no se evidencian diferencias estadísticamente concluyentes entre los modelos evaluados."

4.2.6. Resumen de las pruebas estadísticas

Tabla 33.

Comparación entre coeficientes

Prueba / Índice	¿Qué mide?	¿Para qué sirve en modelos 3D?	Cuándo usarla
V de Aiken	Validez de contenido por ítem: grado de relevancia según jueces expertos	Verificar si cada criterio o ítem evaluado sobre el modelo es considerado adecuado	Para cada pregunta evaluada en escala Likert
ICC (Coef. intraclase)	Concordancia entre jueces (fiabilidad interevaluador)	Saber si los expertos evaluaron de forma consistente los modelos	Para validar confiabilidad del juicio experto
W de Kendall (Friedman)	Concordancia entre jueces en rangos (consistencia ordinal)	Indica el nivel de acuerdo global entre evaluaciones en escalas ordinales	Cuando aplicas Friedman y tienes varios jueces
Prueba de Friedman	Si hay diferencias significativas entre modelos evaluados por los mismos jueces	Determinar si uno o más modelos fueron mejor o peor evaluados que los otros	Comparar ≥ 3 modelos evaluados por los mismos jueces
Post hoc de Conover/Wilcoxon	Comparaciones por pares entre modelos tras Friedman	Identificar entre qué modelos existen diferencias significativas específicas	Solo si Friedman muestra $p < 0.05$

Conclusión y Recomendaciones Finales.

Este manual práctico de validez y confiabilidad ha sido desarrollado como una guía metodológica integral para investigadores, docentes y estudiantes que enfrentan el desafío de construir y validar instrumentos de medición en contextos educativos y de salud. A través de dos ejemplos contrastantes, un cuestionario sobre conocimiento en radiación ionizante y un conjunto de modelos anatomo-patológicos impresos en 3D, se ha ejemplificado cómo adaptar los principios de validez y confiabilidad a diferentes tipos de instrumentos: uno conceptual y otro físico-interactivo.

A lo largo del manual, se han abordado las principales formas de validez (contenido, constructo y criterio) y confiabilidad (consistencia interna y temporal), así como los análisis complementarios necesarios para garantizar la calidad del instrumento, tales como el índice de facilidad, el poder de discriminación y el coeficiente de concordancia entre jueces (ICC). Las herramientas propuestas están pensadas para ser accesibles, replicables y alineadas con los estándares internacionales de investigación, utilizando software gratuito como JASP.

Recomendaciones finales para los usuarios del manual.

- Adapte cada instrumento a su contexto: la validación es siempre dependiente del propósito, la población y el entorno de aplicación.
- Evite la aplicación directa de pruebas no validadas: incluso si provienen de fuentes reconocidas, es esencial validar en la propia población objetivo.
- No confunda confiabilidad con validez: un instrumento puede ser confiable pero no válido. Ambos aspectos deben ser asegurados de manera complementaria.
- Considere siempre una prueba piloto antes de aplicar el instrumento en su fase definitiva.

- No descarte la evidencia cualitativa: los juicios de expertos y las observaciones cualitativas enriquecen y complementan el análisis estadístico.

Este manual no solo pretende entregar herramientas, sino fomentar una cultura de rigurosidad metodológica en el diseño y uso de instrumentos en investigación educativa y en salud. La validación no es un paso aislado, sino un proceso continuo que evoluciona junto con los propósitos del investigador y la complejidad del constructo estudiado.

RECOMENDACIONES FINALES PARA LA CONSTRUCCIÓN DE INSTRUMENTOS VÁLIDOS Y CONFIABLES

A continuación, se presentan algunas recomendaciones prácticas para investigadores, docentes y estudiantes que elaboren instrumentos de evaluación en contextos educativos o de salud:

1. **Delimitar con precisión el constructo**: antes de redactar ítems, es imprescindible tener claridad sobre qué se quiere medir, en qué población, y con qué propósito.
2. **Fundamentar cada dimensión del instrumento** en teorías, literatura científica o estándares disciplinares. Esto fortalecerá la validez de contenido y permitirá justificar cada parte del instrumento.
3. **Redactar ítems claros, unidimensionales y alineados** a los objetivos del estudio. Evitar ambigüedad, doble negación o redacción compleja.
4. **Aplicar juicio de expertos antes de aplicar la encuesta piloto**, usando criterios de claridad, pertinencia, coherencia y nivel de dificultad. Usar escalas estructuradas para asegurar objetividad.
5. **Calcular indicadores estadísticos de validez de contenido** como la V de Aiken o el CVC, y revisar ítems que no cumplan con los umbrales establecidos.

6. **Evaluar la confiabilidad del instrumento mediante KR-20 o alfa de Cronbach**, dependiendo del tipo de ítems, y analizar qué preguntas debilitan la consistencia interna.
7. **Aplicar análisis por ítem**, evaluando índices de dificultad y discriminación, para decidir qué preguntas conservar, revisar o eliminar.
8. **Contextualizar siempre la interpretación de resultados**, recordando que la validez depende del uso que se hace del instrumento, no de una propiedad inherente.
9. **Documentar cada etapa del proceso de validación**, lo que facilitará la replicabilidad, revisión por pares y eventual publicación del instrumento.
10. **Incluir en los reportes finales una sección de limitaciones** y sugerencias de mejora, especialmente si se planea aplicar el instrumento en otras poblaciones o contextos.

Estas prácticas no solo aseguran mayor rigor en la medición, sino que contribuyen al desarrollo de instrumentos útiles, éticos y adaptados a los desafíos reales de la evaluación educativa.

REFERENCIAS

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). Estándares para Pruebas Educativas y Psicológicas. *Estándares Para Pruebas Educativas y Psicológicas*. https://www.aera.net/Publications/-Online-Store/Books-Publications/BKctl/ViewDetails/SKU/AERWSTDEPTSP

Anastasi, Anne., & Urbina, Susana. (1998). *Tests psicológicos*. 729.

Arango-Ramírez, P. M., González-Rosales, V. M., Leyva-Hernández, S. N., Galván-Mendoza, O., Arango-Ramírez, P. M., González-Rosales, V. M., Leyva-Hernández, S. N., & Galván-Mendoza, O. (2023). Validez y fiabilidad de un instrumento de medición de tipos de residentes desde el enfoque de las representaciones sociales. *Estudios Sociales. Revista de Alimentación Contemporánea y Desarrollo Regional*, *33*(62). https://doi.org/10.24836/ES.V33I62.1351

Bobak, C. A., Barr, P. J., & O'Malley, A. J. (2018). Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. *BMC Medical Research Methodology*, *18*(1), 1–11. https://doi.org/10.1186/S12874-018-0550-6/FIGURES/6

Brookhart, S. M. ., & Nitko, A. J. . (2015). *Educational assessment of students*. 530. https://books.google.com/books/about/Educational_Assessment_of_Students.html?hl=es&id=ySA2ngEACAAJ

Campo-Arias, A., & Oviedo, H. C. (2008). Propiedades Psicométricas de una Escala: la Consistencia Interna. *Revista de Salud Pública*, *10*(5), 831–839. https://www.redalyc.org/articulo.oa?id=42210515

Carvajal, A., Centeno, C., Watson, R., Martínez, M., & Sanz Rubiales, Á. (2011). ¿Cómo validar un instrumento de medida de la salud? *Anales Del Sistema Sanitario de Navarra*, *34*(1), 63–72. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1137-66272011000100007&lng=es&nrm=iso&tlng=es

Cohen, R. Jay., Swerdlik, M. E. ., Velázquez Arellano, J. Alberto., & López Carrasco, M. A. (2001). *Pruebas y evaluación psicológicas : introducción a las pruebas y a la medición*. 807. https://books.google.com/books/about/Pruebas_y_evaluaci%C3%B3n_psicol%C3%B3gicas.html?hl=es&id=yvJISgAACAAJ

Escobar-Pérez, J., & Cuervo-Martínez, A. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. *Avances En Medición*, *6*(1), 27–36.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Logistic Regression: Regression with a Binary Dependent Variable. *Multivariate Data Analysis*, 313–

340. https://books.google.com/books/about/Multivariate_Data_Analysis.html?hl=es&id=VvXZnQEACAAJ

Haladyna, T. M. (2004). *Developing and validating multiple … - Google Books*. 306. https://books.google.com/books/about/Developing_and_Validating_Multiple_choic.html?hl=es&id=4fJ2YXMLTrsC

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment*, *7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Hernández-Nieto, R. (2002). Contributions to Statistical Analysis: The Coefficients of Proportional Variance, Content Validity and Kappa. *Mérida: Universidad de Los Andes*, 228. https://books.google.com/books/about/Contributions_to_Statistical_Analysis.html?hl=es&id=yk_2PAAACAAJ

Hernandez-Sampieri, R., & Mendoza, C. (2018). *METODOLOGÍA DE LA INVESTIGACIÓN, Las rutas cuantitativas, cualitativas y mixta*. McGraw-Hill, Interamericana.

Hertzog, M. A. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing and Health*, *31*(2), 180–191. https://doi.org/10.1002/NUR.20247;PAGE:STRING:ARTICLE/CHAPTER

JASP - A Fresh Way to Do Statistics. (n.d.). Retrieved May 28, 2025, from https://jasp-stats.org/

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/J.JCM.2016.02.012

Maldonado Suárez, N., & Santoyo Telles, F. (2024). Validez de contenido por juicio de expertos: Integración cuantitativa y cualitativa en la construcción de instrumentos de medición. *REIRE: Revista d'innovació i Recerca En Educació, ISSN-e 2013-2255, Vol. 17, Nº. 2, 2024, Págs. 1-19*, *17*(2), 1–19. https://dialnet.unirioja.es/servlet/articulo?codigo=9622062&info=resumen&idioma=ENG

Pedrosa, I., Suárez Álvarez, J., & García Cueto, E. (2013). Evidencias sobre la Validez de Contenido: Avances Teóricos y Métodos para su Estimación. *Acción Psicológica, ISSN 1578-908X, Vol. 10, Nº. 2, 2013 (Ejemplar Dedicado a: Validación de Contenido Desde Metodologías Cualitativas y Cuantitativas), Págs. 3-18*, *10*(2), 3–18. https://dialnet.unirioja.es/servlet/articulo?codigo=4758265&info=resumen&idioma=SPA

Perneger, T. V., Courvoisier, D. S., Hudelson, P. M., & Gayet-Ageron, A. (2015). Sample size for pre-tests of questionnaires. *Quality of Life Research*, *24*(1), 147–151. https://doi.org/10.1007/S11136-014-0752-2/METRICS

Prieto Adánez, G., & Delgado González, A. R. (2010). Fiabilidad y validez. *Papeles Del Psicólogo, ISSN-e 1886-1415, ISSN 0214-7823, Vol. 31, Nº. 1, 2010 (Ejemplar*

Dedicado a: Metodología al Servicio Del Psicólogo), Págs. 67-74, *31*(1), 67–74. https://dialnet.unirioja.es/servlet/articulo?codigo=3150828&info=resumen&idioma=SPA

Rubio, D. M. G., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*(2), 94–104. https://doi.org/10.1093/SWR/27.2.94

Thorndike, R. M., & Thorndike-Christ, T. M. (2010). Measurement and Evaluation in Psychology and Education, 8th Edition. *Pearson*, 509.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A.-R. (2015). Design and Implementation Content Validity Study: Development of an instrument for measuring Patient-Centered Communication. *Journal of Caring Sciences*, *4*(2), 165. https://doi.org/10.15171/JCS.2015.017