RECURSOS EDUCATIVOS EN ABIERTO - DOCTA COMPLUTENSE

"MANUAL DE USO DE LA ESCALA PEDro PARA TRABAJOS ACADÉMICOS DEL GRADO EN FISIOTERAPIA"

ASIGNATURA DE BASES CIENTÍFICAS EN FISIOTERAPIA

GRADO EN FISIOTERAPIA

UNIVERSIDAD COMPLUTENSE MADRID

Juan Antonio Valera Calero PT, PhD

Profesor y Coordinador de la Asignatura

Departamento de Fisioterapia

Universidad Complutense de Madrid

PROPÓSITO DEL DOCUMENTO

Este documento explica cómo utilizar la escala PEDro para leer de forma crítica ensayos clínicos de fisioterapia. La escala convierte preguntas metodológicas concretas en una puntuación sencilla, pero su uso correcto exige entender qué pregunta responde cada criterio y qué información no proporciona. El objetivo no es aprender a sumar diez puntos de memoria, sino saber localizar la información en un artículo, distinguir entre lo que está claramente descrito y lo que solo parece probable, y relacionar los resultados con una decisión clínica.

Al terminar el tema, deberías poder…	Evidencia práctica de aprendizaje
Explicar qué significa calidad metodológica.	Diferenciarla de riesgo de sesgo, significación estadística y relevancia clínica.
Reconocer la estructura de la escala PEDro.	Identificar por qué hay 11 criterios, pero la puntuación final va de 0 a 10.
Aplicar cada criterio a un artículo.	Localizar la información relevante y decidir si el criterio está claramente satisfecho.
Distinguir conceptos que suelen confundirse.	Diferenciar aleatorización, ocultación de la asignación y cegamiento; separar seguimiento e intención de tratar.
Interpretar la puntuación total con prudencia.	Revisar qué criterios faltan y no utilizar la puntuación como un veredicto sobre el tratamiento.
Usar la escala dentro de la práctica basada en la evidencia.	Combinar calidad, tamaño del efecto, precisión, aplicabilidad y preferencias de la persona.

Para trabajar con esta guía, lee primero la reflexión y la estructura general; después aplica las explicaciones a un artículo completo y resuelve el caso práctico. En cada respuesta, anota una justificación. El caso y sus cifras son inventados con finalidad docente. Las reglas de puntuación se basan en las instrucciones originales de PEDro, recogidas en la bibliografía.

UNIVERSIDAD COMPLUTENSE MADRID

> La puntuación de la escala PEDro es un punto de partida para analizar un artículo, pero no es un sustituto de la lectura crítica.

LA IMPORTANCIA DE ANALIZAR LA CALIDAD METODOLÓGICA

La práctica basada en la evidencia no consiste en encontrar un artículo y aplicar automáticamente su conclusión. Una decisión de fisioterapia integra la mejor investigación disponible, la experiencia profesional y los valores, objetivos y contexto de la persona. Para que esa integración sea responsable, primero hay que preguntarse si el diseño del estudio podía producir resultados sin sesgos importantes, suficientemente precisos y reproducibles.

La publicación de un artículo, el tamaño de su muestra o la presencia de un valor p inferior a 0,05 no garantizan por sí solos que sus resultados sean fiables. Los errores en la asignación de los participantes, la medición de los resultados, el seguimiento o el análisis pueden desplazar la estimación del efecto. Además, la forma de valorar la metodología puede cambiar la confianza que una revisión sistemática deposita en los estudios que incluye. Por eso, analizar la calidad metodológica informa tanto la síntesis de la evidencia como la decisión clínica individual.

Situación	Pregunta que conviene hacerse
El artículo informa de un resultado positivo.	¿La forma de diseñar y analizar el estudio puede haber favorecido ese resultado?
El estudio tiene muchos participantes.	¿El tamaño compensa una asignación inadecuada o una medición poco protegida frente a errores?
El resultado es estadísticamente significativo.	¿El efecto es suficientemente grande para importar a la persona y a su función?
La intervención parece similar a la de mi consulta.	¿La población, la dosis, el comparador y las medidas se parecen a mi situación clínica?
El ensayo obtiene una puntuación elevada.	¿Qué criterios faltan y qué otros datos necesito para valorar beneficios, daños y costes?

Adaptado de Sackett et al. (1996), Maher et al. (2003) y Morton (2009)

Durante el grado, los apuntes y las recomendaciones del profesorado ayudan a organizar el conocimiento. El siguiente paso es aprender a examinar las fuentes con autonomía. En la práctica encontrarás evidencia incompleta, contradictoria o poco aplicable a tu paciente. Citar un artículo no basta para justificar una decisión: hay que explicar qué aporta, qué limitaciones tiene y qué incertidumbre permanece. Esta capacidad permite seguir aprendiendo a lo largo de la vida profesional.

> Es importante ir más allá de qué resultados obtuvo el estudio y reflexionar sobre cuánta confianza merece ese resultado y si puede aplicarse a la persona que tenemos delante.

QUÉ ES LA ESCALA PEDRO

PEDro significa Physiotherapy Evidence Database. Esta base de datos reúne ensayos clínicos, revisiones sistemáticas y guías de práctica clínica de fisioterapia. Su escala ayuda a identificar ensayos con elementos favorables para la validez interna e información estadística suficiente para interpretar los resultados. Cada criterio se responde con sí o no; solo se concede el punto cuando se cumple claramente. El criterio 1 se registra, pero no se suma.

La escala PEDro ayuda a…	La escala PEDro no puede decir por sí sola…
Ordenar preguntas importantes sobre el diseño, la conducción y la presentación de un ensayo.	Si el tratamiento será útil para cualquier paciente.
Detectar elementos que pueden afectar a la validez interna.	Si el tamaño del efecto es clínicamente relevante.
Comprobar si existen datos suficientes para interpretar el efecto y su variabilidad.	Si los beneficios superan a los daños, la carga o el coste.
Comparar de forma estructurada la información metodológica de varios ensayos relacionados.	Si el resultado se aplica a una población, dosis o contexto distintos.

La escala está pensada para estudios experimentales. Los ensayos clínicos, al tener un diseño en los que se comparan varios grupos, obtendrán puntuaciones mayores que diseños cuasiexperimentales donde no existen esos grupos. Además, hay otros elementos que puntúan favorablemente dentro de los ensayos clínicos (si hay asignación aleatoria, cegamiento…). No es la herramienta adecuada para valorar revisiones sistemáticas, estudios de diagnóstico, cohortes o casos clínicos.

Calidad metodológica y riesgo de sesgo son conceptos relacionados, pero no equivalentes. La calidad metodológica abarca características del diseño y la realización de una investigación; PEDro valora una selección de ellas e información estadística. El riesgo de sesgo se centra en la posibilidad de que un error sistemático distorsione un resultado concreto. Por eso, PEDro no sustituye a una herramienta específica de riesgo de sesgo cuando se necesita esa valoración.

> Un estudio puede estar bien diseñado y, aun así, evaluar una intervención con un efecto pequeño, poco importante o difícil de aplicar en la práctica.

UNIVERSIDAD COMPLUTENSE MADRID

PROPIEDADES PSICOMÉTRICAS DE LA ESCALA

Desde el punto de vista psicométrico, la escala PEDro presenta principalmente evidencia de fiabilidad y validez.

La **fiabilidad** indica hasta qué punto una herramienta produce resultados estables y similares cuando la aplican diferentes evaluadores. En el caso de la PEDro, la fiabilidad entre evaluadores se ha estudiado mediante el coeficiente kappa para los criterios individuales, que se responden como sí o no. Maher et al. (2003) observaron valores de kappa de 0,36 a 0,80 en las evaluaciones individuales y de 0,50 a 0,79 cuando los evaluadores alcanzaban un consenso. Estos resultados indican que la coincidencia entre evaluadores varía según el criterio: algunos son fáciles de interpretar, mientras que otros, como la ocultación de la asignación, el seguimiento o la intención de tratar, pueden generar más dudas. Para la puntuación total, la fiabilidad se valoró mediante el coeficiente de correlación intraclase, con valores de 0,56 en evaluaciones individuales y de 0,68 tras alcanzar un consenso, lo que representa una fiabilidad moderada. Además, la mayoría de las puntuaciones obtenidas por distintos evaluadores se situaron a una distancia de uno o dos puntos.

La **validez** se refiere a si la escala mide razonablemente aquello que pretende medir. En este sentido, de Morton (2009) observó que los criterios mantenían una jerarquía coherente y que la puntuación ordinal total se correlacionaba estrechamente con una puntuación transformada mediante un análisis de Rasch (r = 0,99), lo que apoya el uso de la suma de los diez criterios puntuables como indicador global de calidad metodológica. Sin embargo, estos resultados no significan que la escala mida directamente la eficacia del tratamiento, su relevancia clínica o la aplicabilidad de los resultados a cualquier paciente. Por ello, la puntuación total debe interpretarse junto con los criterios concretos que se cumplen o se incumplen y con una valoración del tamaño del efecto, su precisión, los posibles daños, la aplicabilidad y las preferencias de la persona.

UNIVERSIDAD COMPLUTENSE MADRID

ESTRUCTURA DE LA ESCALA

Los 11 criterios se pueden entender mejor si se agrupan según la pregunta general que responden. El primer criterio se relaciona con la posibilidad de aplicar los resultados a otras personas y no se suma a la puntuación. Los criterios 2 a 9 se centran principalmente en la validez interna. Los criterios 10 y 11 comprueban si el artículo ofrece comparaciones y datos suficientes para interpretar el efecto.

> La escala tiene 11 preguntas, pero la puntuación PEDro tiene 10 puntos. El criterio 1 se conserva para valorar la aplicabilidad y no se incluye en la suma.

Bloque	Ítems	Qué pregunta general responde	¿Suma a la puntuación?
Aplicabilidad inicial	1	¿Se explica quién podía participar y cómo se definió la población?	No
Asignación y grupos	2 a 4	¿La comparación entre grupos se protegió al inicio?	Sí
Cegamiento	5 a 7	¿Las expectativas pudieron influir menos en participantes, terapeutas o evaluadores?	Sí
Seguimiento y análisis	8 y 9	¿Se conservaron y analizaron adecuadamente los participantes asignados?	Sí
Información estadística	10 y 11	¿Se compararon los grupos y se describieron el tamaño y la variabilidad del efecto?	Sí

UNIVERSIDAD COMPLUTENSE
MADRID

ÍTEM 1: CRITERIOS DE ELEGIBILIDAD

Este criterio comprueba si el informe describe la procedencia de los participantes y los criterios utilizados para decidir quién podía entrar en el estudio. Se necesitan ambos elementos: la fuente de reclutamiento y los criterios de elegibilidad. Su función principal es ayudar a juzgar la aplicabilidad de la muestra.

Qué debe aparecer	Ejemplo sencillo	Error frecuente
Fuente de los participantes	Personas atendidas en consultas de fisioterapia de dos centros.	Suponerlo porque aparece el nombre de un hospital.
Criterios de inclusión	Adultos con dolor cervical de más de tres meses y limitación funcional.	Usar la tabla de características basales como si fuera la lista de elegibilidad.
Criterios de exclusión, cuando se establezcan	Cirugía reciente, signos neurológicos progresivos o enfermedad sistémica relevante.	Exigir dos listas separadas de inclusión y exclusión: la escala pide criterios de elegibilidad.

> El criterio 1 describe la población y la selección de participantes. No evalúa si los grupos fueron comparables ni se suma a la puntuación final.

UNIVERSIDAD COMPLUTENSE MADRID

ÍTEMS 2 Y 3: ALEATORIZACIÓN Y OCULTACIÓN DE LA ASIGNACIÓN

Estos criterios protegen momentos distintos del ensayo. La aleatorización determina cómo se asigna cada participante a un grupo. La ocultación de la asignación evita que la persona que decide si alguien cumple los criterios de entrada pueda anticipar qué grupo recibirá. Confundirlos es uno de los errores más habituales al aplicar la escala.

Criterio	Pregunta	Evidencia que permite puntuar	Lo que no basta
2. Asignación aleatoria	¿La pertenencia al grupo se decidió mediante un procedimiento aleatorio?	El artículo dice que los participantes fueron asignados al azar o mediante una secuencia aleatoria generada por ordenador, moneda o dados.	Alternar participantes, utilizar el número de historia, la fecha de nacimiento o el día de la consulta.
3. Asignación oculta	¿La persona que decide la inclusión desconocía la próxima asignación?	Sobres opacos y sellados, o contacto con un responsable externo que custodia la secuencia y comunica la asignación después de incluir a la persona.	Decir simplemente que el estudio fue aleatorizado o que el evaluador estaba ciego.

Regla específica del criterio 2: basta con que el informe afirme que la asignación fue aleatoria; PEDro no exige que detalle cómo se generó la secuencia. Si describe un método predecible, como la alternancia, no cumple. En un ensayo cruzado, en el que cada persona recibe varios tratamientos, se valora si el orden se asignó al azar.

Un artículo puede describir una secuencia aleatoria y no explicar cómo se protegió la asignación antes de incluir a cada participante. En ese caso, el criterio 2 podría estar satisfecho y el criterio 3 no. La respuesta debe seguir la información disponible, aunque el método real del estudio quizá haya sido mejor de lo que el artículo permite saber.

UNIVERSIDAD COMPLUTENSE

> Aleatorizar deja la asignación al azar. Ocultarla impide anticipar el grupo cuando se decide incluir al participante. El cegamiento se refiere a lo que se conoce después de asignarlo.

ÍTEM 4: SIMILARIDAD DE LOS GRUPOS AL INICIO DEL ESTUDIO

Este criterio pregunta si los grupos eran suficientemente parecidos al comienzo para que una diferencia posterior no pueda explicarse solo por una desigualdad inicial importante. En estudios de tratamiento, el informe debe presentar al menos una medida de la gravedad del problema y otra medida de resultado clave diferente, ambas recogidas al inicio. También se revisan otras variables pronósticas relevantes.

Qué revisar	Ejemplo	Cómo interpretarlo
Gravedad del problema	Dolor, discapacidad o limitación funcional inicial.	Permite saber si un grupo empezó con un problema claramente más grave.
Resultado clave al inicio	Por ejemplo, movilidad al inicio si la gravedad se ha descrito mediante dolor.	Debe ser una medida diferente de la utilizada para describir la gravedad.
Variables pronósticas	Edad, duración, comorbilidades o factores que puedan cambiar la recuperación.	Se valora la relevancia clínica de las diferencias, no solo si p es mayor de 0,05.

Una prueba estadística de comparación basal puede aportar información (si p>0,05), pero no sustituye la inspección de los valores y de su importancia clínica. Una diferencia pequeña puede ser estadísticamente significativa en una muestra grande y una diferencia relevante puede no alcanzar significación en una muestra pequeña. Según la regla de PEDro, este criterio puede cumplirse aunque solo se presenten los datos basales de quienes completaron el estudio. La calidad del seguimiento se examina por separado en el criterio 8.

> No se exige que los grupos sean idénticos. Se exige que no exista una diferencia inicial clínicamente importante que, por sí sola, pueda explicar el resultado.

UNIVERSIDAD COMPLUTENSE MADRID

ÍTEMS 5 A 7: CEGAMIENTO

El cegamiento, también llamado enmascaramiento, intenta reducir la influencia de las expectativas sobre la experiencia de los participantes, la administración del tratamiento o la medición de los resultados. Las etiquetas doble ciego o ciego simple pueden ser ambiguas: identifica quién desconocía la asignación. En participantes y terapeutas, además, debe ser razonable que no pudieran distinguir las intervenciones.

Quién	Qué tendría que ocurrir	Ejemplo y precaución
5. Participantes	Todos desconocen su asignación y no pueden distinguir razonablemente las intervenciones.	Puede ser posible con una intervención simulada creíble. Es difícil si se compara ejercicio con terapia manual.
6. Terapeutas	Todos los que administran la terapia desconocen la asignación y no pueden distinguir las intervenciones.	Es difícil o imposible en muchas intervenciones activas. La imposibilidad no permite conceder el punto.
7. Evaluadores	Todos los evaluadores de al menos un resultado clave desconocen la asignación.	Ser independiente no basta: debe estar ciego. Si hay varios evaluadores de ese resultado, todos deben cumplirlo.

Que el cegamiento del terapeuta no sea viable no convierte automáticamente en inútil un ensayo de fisioterapia. Sí indica una posible fuente de influencia que debe considerarse junto con el cegamiento de los evaluadores, la naturaleza del comparador, la medida utilizada y el resto del diseño. La ausencia de un punto debe interpretarse dentro de ese contexto.

> En una medida autoinformada, como el dolor indicado por el paciente, PEDro considera ciego al evaluador si el propio participante está ciego. No basta con que otra persona recoja el cuestionario sin conocer el grupo.

UNIVERSIDAD COMPLUTENSE MADRID

CRITERIOS 8 Y 9 SEGUIMIENTO E INTENCIÓN DE TRATAR

ÍTEMS 8 Y 9: SEGUIMIENTO E INTENCIÓN DE TRATAR

Estos criterios se refieren a lo que ocurre después de la asignación. El criterio 8 comprueba si se obtuvo al menos un resultado clave en más del 85 % de las personas inicialmente asignadas. El criterio 9 se cumple si todas las personas con datos recibieron lo asignado o, cuando hubo incumplimientos, se analizaron en su grupo original para al menos un resultado clave.

Criterio	Pregunta práctica	Ejemplo que sí cumple	Error que suele aparecer
8. Más del 85 %	¿Se indica cuántas personas fueron asignadas y cuántas aportaron un resultado clave?	Se asignaron 80, se obtuvo el resultado en 72 y (72/80) × 100 = 90 %.	Contar como observados los datos que se estimaron para sustituir valores ausentes.
9. Intención de tratar	¿El análisis respeta el grupo inicial cuando hubo incumplimientos o cambios?	Una persona asignada a ejercicio recibe otra terapia, pero sus datos se analizan en el grupo de ejercicio.	Excluir del análisis a quienes incumplen, aunque tengan datos, o cambiarlos de grupo según lo que recibieron.

El umbral del criterio 8 es estricto: 68 de 80 equivale al 85 % y no cumple; 69 de 80 equivale al 86,25 % y sí cumple. Si hay varios momentos de seguimiento, basta con superar el 85 % en al menos uno para un resultado clave. Esto no asegura un seguimiento adecuado a largo plazo.

Para el criterio 9, aplica la definición concreta de PEDro. No exige por sí mismo imputar los datos ausentes, es decir, estimar valores para sustituir los que no se recogieron. Tampoco exige las palabras intención de tratar si se describe un análisis equivalente o se afirma que todos recibieron lo asignado. Un número total de personas analizadas, sin esa información, no basta.

UNIVERSIDAD COMPLUTENSE

> El seguimiento responde a cuántas personas aportaron datos. La intención de tratar responde a cómo se utilizaron esos datos en relación con la asignación inicial.

ÍTEMS 10 Y 11: INFORMACIÓN ESTADÍSTICA

Los dos últimos criterios no juzgan si el resultado fue positivo o negativo. Comprueban si el informe permite ver la comparación entre grupos y el tamaño del efecto con su variabilidad. Esta información es necesaria para valorar tanto la dirección del resultado como la precisión de la estimación.

Criterio	Qué debe aparecer	Ejemplo	Lo que no basta
10. Comparación entre grupos	Una comparación estadística de un grupo con otro para al menos un resultado clave.	Diferencia entre grupos al final, diferencia de cambios o interacción grupo por tiempo, con p o intervalo de confianza.	Comparar solo el cambio antes-después dentro de cada grupo.
11. Medida puntual y variabilidad	Una medida del tamaño del efecto y una medida de dispersión o incertidumbre.	Media y desviación estándar en cada grupo; diferencia entre grupos e intervalo de confianza; mediana y rango intercuartílico en cada grupo.	Presentar únicamente p, porcentajes sin denominadores o una afirmación de mejora.

Por ejemplo, que el grupo de ejercicio mejore de forma significativa respecto a su propio valor inicial no demuestra que sea superior al grupo control. Para responder a esa pregunta hay que examinar la comparación entre grupos. Después conviene valorar si la diferencia es clínicamente importante y si su intervalo de confianza incluye efectos demasiado pequeños para cambiar la decisión.

También se admiten datos en figuras si se identifica qué representan los puntos y las barras de error. En resultados categóricos, como mejoró o no mejoró, el criterio 11 se cumple si se presenta el número de personas de cada categoría en cada grupo. No es obligatorio añadir una media o una desviación estándar a ese tipo de resultado.

> La desviación estándar describe la dispersión de los valores entre personas. El intervalo de confianza describe la precisión de una estimación. Ambas medidas pueden satisfacer el criterio 11, pero responden a preguntas distintas.

DISTINCIONES QUE HAY QUE TENER CLARAS

La aplicación de la escala mejora cuando se separan conceptos que suelen aparecer juntos en los artículos, pero que responden a preguntas diferentes. La tabla siguiente funciona como resumen para repasar antes de puntuar.

No es lo mismo	Primera idea	Segunda idea	Consecuencia para PEDro
Aleatorización y ocultación	La aleatorización genera la secuencia de grupos.	La ocultación protege la decisión de incluir y asignar.	Pueden puntuar de forma independiente en los criterios 2 y 3.
Cegamiento y ocultación	El cegamiento ocurre después de la asignación y afecta a una persona concreta.	La ocultación ocurre antes de conocer la asignación.	Un evaluador ciego no demuestra que la asignación estuviera oculta.
Seguimiento e intención de tratar	El seguimiento cuenta cuántos aportan datos.	La intención de tratar indica cómo se analizan según el grupo original.	Los criterios 8 y 9 pueden tener respuestas distintas.
Significación y relevancia clínica	La significación depende de la prueba estadística y de la incertidumbre.	La relevancia depende del tamaño del cambio y de lo que importa a la persona.	El criterio 10 no decide si el cambio es útil.
Calidad metodológica y riesgo de sesgo	La primera resume características metodológicas seleccionadas.	El segundo estima la posibilidad de error sistemático en un resultado.	PEDro informa una parte de la lectura crítica y no sustituye todas las herramientas.
Puntuación total y perfil de criterios	La suma facilita una visión rápida.	Los criterios ausentes pueden tener consecuencias distintas.	Dos ensayos con la misma puntuación no son necesariamente equivalentes.

UNIVERSIDAD COMPLUTENSE MADRID

CÓMO CALCULAR E INTERPRETAR LA PUNTUACIÓN

La puntuación se obtiene sumando los criterios 2 a 11. Cada criterio aporta 1 punto cuando está claramente satisfecho y 0 cuando no lo está o el informe no permite confirmarlo. El resultado final va de 0 a 10. El criterio 1 se registra por separado porque aporta información sobre la población, pero no forma parte de la suma.

Parte de la valoración	Qué permite decir	Qué no permite decir
Criterio individual	Si el artículo describe un elemento metodológico concreto.	Que todo el ensayo sea bueno o malo por un único criterio.
Puntuación 0 a 10	Cuántos de los diez criterios puntuables están claramente presentes.	Que el tratamiento funcione, que el efecto sea importante o que el estudio sea aplicable.
Perfil de criterios	Qué fortalezas y limitaciones metodológicas conviene tener presentes.	Una clasificación universal válida para todas las intervenciones y áreas de fisioterapia.

No cambies el denominador porque el cegamiento sea inviable, la puntuación continúa siendo sobre 10. Si una publicación utiliza etiquetas como calidad alta o baja, revisa qué umbral ha elegido; esas categorías no sustituyen la valoración de los ítems. PEDro desaconseja comparar la calidad de ensayos de áreas terapéuticas distintas solo mediante la suma, porque las posibilidades de cegamiento pueden ser muy diferentes.

Combina la puntuación con la pregunta clínica, la aplicabilidad, la magnitud y precisión del efecto, los posibles daños, los costes y las preferencias de la persona. Una puntuación alta puede acompañar a un efecto pequeño; una más baja exige examinar qué limitaciones reducen la confianza en el resultado.

> No existe una puntuación total que convierta automáticamente un ensayo en "bueno" o "malo". Aunque la metodología sea consistente, la escala no evalúa como se interpretan los resultados obtenidos.

¿PUEDE LA INTELIGENCIA ARTIFICIAL HACER LA EVALUACIÓN DE FORMA PRECISA Y FIABLE?

La inteligencia artificial (IA) puede ayudar a aplicar la escala PEDro e incluso generar una evaluación completa. Sin embargo, que un sistema sea capaz de responder a los once ítems y calcular una puntuación no significa que sus decisiones sean correctas. La distinción fundamental es entre utilizar la IA como apoyo para la lectura crítica y delegarle, sin comprobación, la evaluación metodológica de un estudio.

¿QUÉ MUESTRAN LAS INVESTIGACIONES?

Los resultados disponibles varían según el sistema utilizado, las instrucciones proporcionadas y los artículos evaluados. Por tanto, no existe una única respuesta válida para todas las herramientas de IA.

Referencia	¿Qué se evaluó?	Resultado principal
Hao et al. (2025)	ChatGPT-4 evaluó 50 ensayos, utilizando sus textos completos y las instrucciones oficiales de PEDro. Se comparó con un consenso de evaluadores humanos.	La coincidencia completa, incluyendo las decisiones por ítem, se produjo en 23 ensayos (el 46 %). El acuerdo por ítem osciló entre el 74 % y el 100 %, con mayores dificultades en la ocultación de la asignación y el seguimiento.
Valera-Calero et al. (2026)	Se compararon las evaluaciones de GPT-5 con las puntuaciones publicadas en diez revisiones sistemáticas, para 227 ensayos.	El acuerdo por ítem osciló entre el 69,2 % y el 94,7 %. El menor acuerdo correspondió al ítem 1, que no se suma al total. Persistieron discrepancias que requirieron revisar los artículos y sus justificaciones.
Ju et al. (2026)	Se desarrolló y evaluó una herramienta específica, GPT-PEDro ET, en 125 ensayos de rehabilitación del ictus.	Obtuvo una elevada repetibilidad de la puntuación total (CCI de 1,00) y un CCI de 0,83 frente a la referencia. Son resultados prometedores, circunscritos a esa herramienta y a la muestra estudiada.

UNIVERSIDAD COMPLUTENSE MADRID

Estos resultados permiten entender dos diferencias importantes. La primera es que repetir una respuesta no equivale a acertar. Un sistema puede ofrecer una puntuación muy estable y, aun así, mantener una interpretación equivocada. La segunda es que coincidir en la puntuación total no garantiza coincidir en cada criterio: dos evaluaciones pueden sumar seis puntos y haber identificado fortalezas y limitaciones diferentes. Además, un coeficiente de correlación intraclase de 0,83 no significa que el sistema haya acertado el 83 % de las evaluaciones.

Tampoco puede concluirse que un modelo sea superior a otro comparando directamente estos estudios, porque utilizaron muestras, procedimientos y referencias diferentes. Los resultados de una herramienta diseñada específicamente para PEDro no se trasladan automáticamente a cualquier conversación con un asistente de IA.

¿POR QUÉ SIGUE SIENDO NECESARIA LA REVISIÓN HUMANA?

Aplicar PEDro no consiste únicamente en localizar palabras. Que un artículo indique que la asignación fue aleatoria no demuestra que estuviera oculta. Del mismo modo, mencionar el cegamiento no basta: hay que identificar quién desconocía la intervención asignada y comprobar si se cumplen las condiciones del criterio correspondiente. Estas distinciones están recogidas en las instrucciones oficiales de PEDro.

La IA puede localizar información relevante, pero también omitirla, interpretarla incorrectamente o atribuir al estudio procedimientos que el texto no permite confirmar. Una explicación extensa y convincente no sustituye la comprobación de su contenido.

La evaluación humana tampoco es infalible. En el trabajo de Valera-Calero et al. se revisaron los desacuerdos para determinar qué interpretación estaba respaldada por el artículo; discrepar de una revisión publicada no implicaba necesariamente que la IA estuviera equivocada. Su posible utilidad reside, precisamente, en señalar información que merezca una segunda lectura, manteniendo la decisión final bajo responsabilidad humana.

¿CÓMO PUEDE UTILIZARSE DE FORMA RESPONSABLE TRABAJOS ACADÉMICOS?

Cuando utilices la escala PEDro para evaluar ensayos clínicos en una tarea, una revisión bibliográfica o tu Trabajo de Fin de Grado (TFG), la IA puede ayudarte a localizar información y contrastar interpretaciones. Su uso debe ajustarse a las indicaciones de la asignatura o del tutor. En cualquier caso, debes poder explicar y justificar cada puntuación incluida en tu trabajo, aunque hayas utilizado una herramienta de IA como apoyo. Para incorporarla de forma adecuada, puedes seguir estas orientaciones:

1. **Define para qué vas a utilizarla**: Diferencia tareas concretas, como aclarar el significado de un criterio, localizar información sobre el seguimiento o contrastar una valoración. Esto te permitirá explicar qué parte del proceso has realizado con ayuda de la IA.
2. **Realiza una primera evaluación del artículo**: Lee el texto completo, consulta las instrucciones oficiales de PEDro y anota tu propuesta para cada ítem. Disponer de una valoración propia facilita comparar argumentos y detectar posibles errores cuando consultes la respuesta de la IA.
3. **Solicita propuestas que puedas comprobar**: Proporciona el artículo completo y las instrucciones de la escala. Para cada criterio, pide una puntuación provisional, una justificación y el fragmento del artículo que la respalda, indicando su página, apartado o tabla. Una puntuación aislada aporta poca información para elaborar una evaluación académica defendible.
4. **Comprueba cada decisión antes de incorporarla al trabajo**: Verifica que los fragmentos citados existen y que permiten cumplir el criterio correspondiente. Si hay desacuerdo con tu valoración, vuelve al artículo y a las instrucciones; consulta con el tutor cuando la duda persista. Si, tras revisar el texto, el cumplimiento no queda suficientemente claro, no se concede el punto. Esto no

UNIVERSIDAD COMPLUTENSE MADRID

demuestra necesariamente que el procedimiento se realizara mal, sino que su cumplimiento no puede confirmarse con la información publicada.

5. **Conserva un registro de la evaluación**: Para cada estudio, guarda las puntuaciones finales, las evidencias que las justifican y las dudas resueltas. Aplica un procedimiento uniforme a todos los artículos. Si el protocolo del trabajo exige dos evaluadores humanos independientes, el uso de IA no sustituye esa segunda evaluación.
6. **Describe su uso con transparencia**: Registra la herramienta, la versión disponible, la fecha y las instrucciones empleadas. Explica su función en el apartado que corresponda según las normas del trabajo y conserva las respuestas relevantes. La descripción debe reflejar el procedimiento que realmente has seguido.

> Incluir una puntuación PEDro en tu trabajo implica asumir la responsabilidad de su justificación. Ante una pregunta del docente o del tribunal, debes poder señalar qué información del artículo respalda cada decisión y explicar cómo has aplicado el criterio.

UNIVERSIDAD COMPLUTENSE
MADRID

CASO PRÁCTICO

Lee la descripción como si fuera el apartado de métodos y resultados de un ensayo. Después intenta responder cada criterio antes de consultar la solución comentada. El caso es hipotético y está diseñado para practicar la lógica de la escala, no para enseñar un tratamiento concreto.

Información del informe	Lo que describe el artículo
Población	En dos consultas de fisioterapia se reclutaron 80 adultos de 25 a 65 años con dolor cervical inespecífico de más de tres meses y limitación funcional. Se excluyeron cirugía reciente, radiculopatía progresiva y enfermedad sistémica relevante.
Asignación	Una secuencia generada por ordenador asignó a 40 personas a ejercicio terapéutico y a 40 a terapia manual de forma aleatoria. Una unidad externa preparó sobres opacos, cerrados y numerados que se abrieron en orden después de confirmar la inclusión.
Medidas y situación inicial	El resultado principal fue la rotación cervical medida en grados. Se recogieron también dolor y discapacidad. La tabla basal presenta estas medidas, edad, sexo y duración de los síntomas, sin diferencias clínicamente importantes entre grupos.
Cegamiento	Participantes y terapeutas sabían qué intervención se administraba. Un único evaluador, independiente y ciego a la asignación, midió la rotación cervical. El dolor y la discapacidad fueron autoinformados.
Seguimiento	A las ocho semanas se midió la rotación cervical de 72 de las 80 personas asignadas. No se obtuvo esa medición en las otras ocho.
Análisis	Seis de las 72 personas evaluadas no cumplieron el tratamiento asignado. Se conservaron sus datos en el grupo original. Para las ocho personas sin medición se imputaron datos; el análisis incluyó a las 80 según la asignación inicial.
Resultados	Se comparó la rotación cervical de ambos grupos a las ocho semanas. Se informó la diferencia de medias entre grupos con su intervalo de confianza del 95 %, y la media y desviación estándar de cada grupo.

UNIVERSIDAD COMPLUTENSE MADRID

Resuelve el caso en una hoja con cuatro columnas: criterio, respuesta, frase que la justifica y punto. Utiliza sí, no o no está claro para razonar, pero recuerda que, en los criterios puntuables, tanto no como no está claro aportan 0. Calcula la suma y explica las dos limitaciones que consideres más importantes.

> Antes de mirar la solución, anota para cada criterio: sí, no o no está claro. Escribe también qué frase del caso justifica tu decisión.

SOLUCIÓN COMENTADA

La solución muestra cómo se transforma la descripción del caso en una valoración justificable. Fíjate en que los criterios 5 y 6 no puntúan, aunque esa ausencia sea difícil de evitar en una comparación entre ejercicio y terapia manual.

Ítem	Decisión	Justificación	Puntuación
1	Sí	Se describen dos consultas como fuente y los criterios de elegibilidad. Se registra, pero no se suma.	No suma
2	Sí	Se informa de una secuencia generada por ordenador para asignar a los grupos de forma aleatoria.	1
3	Sí	Los sobres opacos y cerrados se abrieron después de confirmar la inclusión.	1
4	Sí	Se describen una medida de gravedad, un resultado clave y variables pronósticas sin diferencias iniciales clínicamente importantes.	1
5	No	Los participantes podían distinguir ejercicio de terapia manual y conocían su grupo.	0
6	No	Los terapeutas sabían qué intervención administraban. En este caso no era razonable esperar cegamiento.	0
7	Sí	El único evaluador de la rotación cervical, resultado principal, estaba ciego. Esto no implica cegamiento para dolor o discapacidad.	1
8	Sí	Se midió realmente a 72/80: 90 %, que supera el 85 %. Los ocho valores imputados no cuentan como mediciones obtenidas.	1
9	Sí	Los seis incumplidores con datos se analizaron en su grupo original. La imputación no es lo que determina este punto.	1
10	Sí	Se informó una comparación entre grupos con un intervalo de confianza del 95 por ciento.	1
11	Sí	Se presentaron medias y desviaciones estándar, además de la diferencia entre grupos.	1
Total		Puntuación PEDro: suma de los ítems 2 a 11.	8/10

LISTA DE COMPROBACIÓN PARA EL ESTUDIO

1. ¿Puedo explicar qué es la escala PEDro y para qué tipo de estudios se utiliza?
2. ¿Sé por qué la escala tiene 11 criterios, pero la puntuación final va de 0 a 10?
3. ¿Puedo distinguir calidad metodológica, riesgo de sesgo, significación estadística y relevancia clínica?
4. ¿Sé separar asignación aleatoria y ocultación de la asignación?
5. ¿Puedo identificar quién tendría que estar ciego en los criterios 5, 6 y 7?
6. ¿Sé decidir si un seguimiento supera el 85 por ciento usando el número inicial y el número evaluado?
7. ¿Puedo explicar qué significa analizar a los participantes según su asignación original?
8. ¿Sé diferenciar una comparación entre grupos de una comparación antes-después dentro de un solo grupo?
9. ¿Puedo reconocer una medida puntual y una medida de variabilidad?
10. ¿Concedo puntos solo cuando el criterio está claramente satisfecho en el informe?
11. ¿Reviso qué criterios faltan en vez de interpretar solo la suma total?
12. ¿Puedo explicar por qué una puntuación alta no demuestra que un tratamiento sea clínicamente útil?
13. ¿Sé relacionar la lectura metodológica con el PICO, la aplicabilidad y las preferencias de la persona?

> Una buena valoración PEDro siempre deja dos huellas, una puntuación y una justificación que permite entender cómo se obtuvo.

UNIVERSIDAD COMPLUTENSE MADRID

BIBLIOGRAFÍA

Albanese, E., Bütikofer, L., Armijo-Olivo, S., Ha, C., & Egger, M. (2020). Construct validity of the Physiotherapy Evidence Database (PEDro) quality scale for randomized trials: Item response theory and factor analyses. *Research Synthesis Methods, 11*(2), 227–236. doi:10.1002/jrsm.1385.

Armijo-Olivo, S., da Costa, B. R., Cummings, G. G., Ha, C., Fuentes, J., Saltaji, H., & Egger, M. (2015). PEDro or Cochrane to assess the quality of clinical trials? A meta-epidemiological study. *PLOS ONE, 10*(7), e0132634. doi:10.1371/journal.pone.0132634.

Bhogal, S. K., Teasell, R. W., Foley, N. C., & Speechley, M. R. (2005). The PEDro scale provides a more comprehensive measure of methodological quality than the Jadad scale in stroke rehabilitation literature. *Journal of Clinical Epidemiology, 58*(7), 668–673. doi:10.1016/j.jclinepi.2005.01.002.

Cashin, A. G., & McAuley, J. H. (2020). Clinimetrics: Physiotherapy Evidence Database (PEDro) Scale. *Journal of Physiotherapy, 66*(1), 59. doi:10.1016/j.jphys.2019.08.005.

de Morton, N. A. (2009). The PEDro scale is a valid measure of the methodological quality of clinical trials: A demographic study. *Australian Journal of Physiotherapy, 55*(2), 129–133. doi:10.1016/S0004-9514(09)70043-1.

Elkins, M. R., Herbert, R. D., Moseley, A. M., Sherrington, C., & Maher, C. (2010). Rating the quality of trials in systematic reviews of physical therapy interventions. *Cardiopulmonary Physical Therapy Journal, 21*(3), 20–26. Texto completo.

Foley, N. C., Bhogal, S. K., Teasell, R. W., Bureau, Y., & Speechley, M. R. (2006). Estimates of quality and reliability with the Physiotherapy Evidence-Based Database Scale to assess the methodology of randomized controlled trials of pharmacological and nonpharmacological interventions. *Physical Therapy, 86*(6), 817–824. doi:10.1093/ptj/86.6.817.

Gómez-Conesa, A., García-Moreno, J. M., Calvo-Muñoz, I., Suarez-Serrano, C. M., & López-López, J. A. (2026). Use of the Spanish version of the Physiotherapy Evidence Database (PEDro) scale: A meta-research study of systematic reviews. *Physiotherapy Research International, 31*(4), e70334. doi:10.1002/pri.70334.

Hao, J., Yao, Z., Remis, A., & Yu, X. (2025). Evaluating evidence with AI: ChatGPT's performance on the PEDro scale in physical therapy research. *Journal of Evidence-Based Medicine, 18*, e70074. doi:10.1111/jebm.70074.

Ju, Y.-J., Wang, Y.-C., Chou, F., Hou, W.-H., Chen, S.-M., & Hsieh, C.-L. (2026). Development and performance validation of an automated generative pretrained transformer-based evaluation tool for the Physiotherapy Evidence Database scale. *Archives of Physical Medicine and Rehabilitation*. Publicación anticipada en línea. doi:10.1016/j.apmr.2026.02.011.

Macedo, L. G., Elkins, M. R., Maher, C. G., Moseley, A. M., Herbert, R. D., & Sherrington, C. (2010). There was evidence of convergent and construct validity of Physiotherapy Evidence Database quality scale for physiotherapy trials. *Journal of Clinical Epidemiology, 63*(8), 920–925. doi:10.1016/j.jclinepi.2009.10.005.

Maher, C. G., Sherrington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy, 83*(8), 713–721. doi:10.1093/ptj/83.8.713.

Moseley, A. M., & Pinheiro, M. B. (2022). Research Note: Evaluating risk of bias in randomised controlled trials. *Journal of Physiotherapy, 68*(2), 148–150. doi:10.1016/j.jphys.2022.02.003.

Moseley, A. M., Rahman, P., Wells, G. A., Zadro, J. R., Sherrington, C., Toupin-April, K., & Brosseau, L. (2019). Agreement between the Cochrane risk of bias tool and Physiotherapy Evidence Database (PEDro) scale: A meta-epidemiological study of randomized controlled trials of physical therapy interventions. *PLOS ONE, 14*(9), e0222770. doi:10.1371/journal.pone.0222770.

Paci, M., Bianchini, C., & Baccini, M. (2022). Reliability of the PEDro scale: Comparison between trials published in predatory and non-predatory journals. *Archives of Physiotherapy, 12*, artículo 10. doi:10.1186/s40945-022-00133-6.

UNIVERSIDAD COMPLUTENSE MADRID

Physiotherapy Evidence Database. (2026). *Escala PEDro*. Versión española y notas de administración. Recurso oficial.

Shiwa, S. R., Costa, L. O. P., Costa, L. C. M., Moseley, A., Hespanhol Junior, L. C., Venâncio, R., Ruggero, C., Sato, T. O., & Lopes, A. D. (2011). Reproducibility of the Portuguese version of the PEDro Scale. *Cadernos de Saúde Pública, 27*(10), 2063–2068. doi:10.1590/S0102-311X2011001000019.

Valera-Calero, J. A., Monclús-Díez, G., Fernández-López, I., López-Redondo, M., Sánchez-Jorge, S., Buffet-García, J., Usta-Ayanoğlu, K. Y., & Varol, U. (2026). Analysis of item-level disagreement between researchers and ChatGPT on the PEDro scale for assessing the methodological quality of experimental studies. *Journal of Evidence-Based Medicine, 19*, e70145. doi:10.1111/jebm.70145.

Verhagen, A. P., de Vet, H. C., de Bie, R. A., Kessels, A. G., Boers, M., Bouter, L. M., & Knipschild, P. G. (1998). The Delphi list: A criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *Journal of Clinical Epidemiology, 51*(12), 1235–1241. doi:10.1016/S0895-4356(98)00131-0.

Yamato, T. P., Maher, C., Koes, B., & Moseley, A. (2017). The PEDro scale had acceptably high convergent validity, construct validity, and interrater reliability in evaluating methodological quality of pharmaceutical trials. *Journal of Clinical Epidemiology, 86*, 176–181. doi:10.1016/j.jclinepi.2017.03.002.

UNIVERSIDAD COMPLUTENSE
MADRID