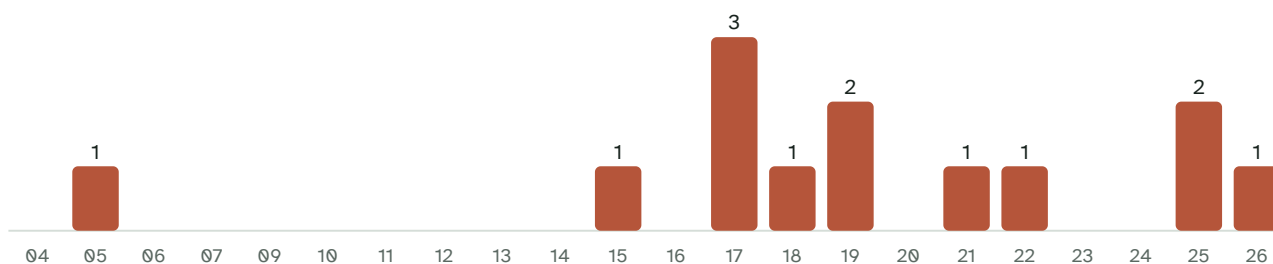


Sesgos, validez y precisión

Todo estudio se equivoca de dos maneras: por azar (error aleatorio, que resta precisión) o de forma sistemática (sesgo, que resta validez). El EIR pregunta sobre todo nombres y definiciones: qué es un error sistemático, qué significa validez interna, cómo se llama cada sesgo (selección, autoselección, detección, Hawthorne, adelanto diagnóstico) y qué distingue la fiabilidad de la validez de un instrumento, además de la escala del índice kappa. En el banco hay 13 preguntas válidas de este tema entre 2005 y 2026 (y una de 2006, anulada).

Apuntes para el EIR · versión 2026-09-27 · **sin revisión enfermera**. Todas las preguntas son reales, de exámenes EIR, con su año y su número; ninguna está escrita por IA. Las respuestas están en el solucionario, al final.

Cómo cae en el EIR



13 preguntas de este tema en 22 exámenes (2004-2026, ordinarias, clasificación segura).

Lo más repetido

- Fiabilidad de un instrumento: resultados consistentes en repetidas aplicaciones (2 exámenes)
- Sesgo de (auto)selección por participación voluntaria (2 exámenes)

Índice

Resumen para repasar	3
Puntos clave	5
1. Error aleatorio y error sistemático: precisión y exactitud	6
2. Validez del estudio: interna y externa	8
3. Sesgo de selección	10
4. Sesgo de información o de medición	12
5. Confusión	14
6. Sesgos de los programas de cribado	15
7. Calidad de un instrumento de medida: fiabilidad y validez	17
8. Concordancia: índice kappa y coeficiente de correlación intraclase	20
Todas las preguntas del tema (13)	22
Preguntas que también podrían ser de este tema (1)	26
Solucionario	27
Fuentes	28

Resumen para repasar

1. Error aleatorio y error sistemático: precisión y exactitud

- **Error aleatorio:** desviación **debida solo al azar**; fuentes: variación biológica, error de muestreo y error de medida; **nunca se elimina del todo**; el de muestreo se reduce **aumentando el tamaño** del estudio ^{F1}.
- **Error sistemático = sesgo:** los resultados se apartan **de forma sistemática** del valor real; poco error sistemático = **exactitud alta**; **no depende del tamaño muestral** ^{F1}.
- Un instrumento es **preciso** si da resultados **consistentes** al repetirlo en las mismas circunstancias ^{F19}.
- Sesgos principales: **selección y medición (clasificación)** ^{F1}.

2. Validez del estudio: interna y externa

- Un estudio es **válido** si sus resultados corresponden a la verdad: **sin error sistemático** y con el menor error aleatorio posible ^{F1}.
- **Validez interna:** resultados **correctos para el grupo estudiado**; la amenazan **todas** las fuentes de error sistemático y mejora con **buen diseño**; se valoraba como «**calidad metodológica**», hoy como **riesgo de sesgo** ^{F1,F10}.
- **Validez externa = generalización** a personas que no estaban en el estudio; la interna es **necesaria pero no suficiente** para la externa ^{F1}.

3. Sesgo de selección

- **Selección:** los participantes **difieren sistemáticamente de la población de interés** ^{F2}.
- **Autoselección / voluntarios:** los que se ofrecen difieren de la población general; es un sesgo de **selección** ^{F1,F3}.
- **Berkson:** estudiar pacientes **hospitalizados** crea asociaciones espurias ^{F4}. **Neyman (prevalencia-incidencia):** se pierden los casos **fatales, breves o leves** por mirar tarde ^{F5}.
- **Trabajador sano:** los trabajadores tienen que estar sanos para trabajar ^{F1}. **No respuesta y desgaste** (pérdidas desiguales) ^{F6,F7}.

4. Sesgo de información o de medición

- **Información:** errores sistemáticos al **recoger, recordar, registrar o manejar** los datos ^{F8}; si es **no diferencial**, según la OMS casi siempre **infraestima** la asociación, aunque sus efectos pueden ser impredecibles ^{F1,F12}.
- **Recuerdo:** casos y controles recuerdan distinto la exposición ^{F1,F9}. **Observador:** se evita con **ciego o doble ciego** ^{F1}.
- **Detección:** diferencias sistemáticas en **cómo se determinan los desenlaces** ^{F11}.
- **Efecto Hawthorne:** los participantes **cambian su conducta al saberse observados** ^{F13}.

5. Confusión

- Un **factor de confusión** es **factor de riesgo** de la enfermedad **y** está **asociado a la exposición** (café, tabaco y cardiopatía) ^{F1,F14}.
- Control en el **diseño: aleatorización, restricción, emparejamiento**; en el **análisis: estratificación y modelos multivariantes** ^{F1}.

6. Sesgos de los programas de cribado

- **Adelanto diagnóstico (lead time)**: la supervivencia **parece más larga** porque se cuenta desde un diagnóstico **adelantado**, aunque el curso no cambie ^{F15,F17}.
- **Duración (length time)**: el cribado detecta preferentemente las formas de **progresión lenta** ^{F16}.
- **Sobrediagnóstico**: cánceres que **nunca se habrían manifestado** clínicamente ^{F17}.
- El NCI atribuye la mejora de supervivencia sin descenso de la mortalidad al **adelanto** y también al **sobrediagnóstico** ^{F17}.

7. Calidad de un instrumento de medida: fiabilidad y validez

- **Fiabilidad**: medir de forma **constante**; resultados comparables al aplicarlo en **distintas ocasiones** ^{F18}.
- **Validez**: el instrumento **mide lo que quiere medir**; puede ser fiable y no válido ^{F18}.
- Fiabilidad: **consistencia interna** (relación entre los ítems; **alfa de Cronbach**), **estabilidad** (test-retest), **equivalencia** y **armonía interjueces** ^{F18}.
- Validez: **aparente, de contenido** (todas las dimensiones), **de criterio** (frente al **estándar de oro**; concomitante y predictiva) y **de constructo** ^{F18}.

8. Concordancia: índice kappa y coeficiente de correlación intraclase

- **Kappa** (Cohen, 1960): variables **categorías**; acuerdo observado **más allá del azar** respecto al **máximo posible más allá del azar**; 1 = acuerdo perfecto, 0 = el esperado por azar ^{F20,F19}.
- Escala de Cohen: 0,21-0,40 **fair** (discreto); 0,41-0,60 **moderado**; 0,61-0,80 **sustancial**; 0,81-1 **casi perfecto**; otras escalas son más exigentes (McHugh: < 0,60 inadecuado) ^{F21}.
- **CCI**: equivalente del kappa para variables **continuas**; **Pearson no mide concordancia** ^{F19}.

Puntos clave

- 1 El error aleatorio se debe al azar y se reduce aumentando el tamaño del estudio; el sistemático (sesgo) no depende del tamaño muestral **F1**.
- 2 Un error que aparta todas las mediciones del valor real de forma sistemática (por ejemplo, un aparato mal calibrado) es un error sistemático o sesgo **F1,F19**.
- 3 Validez interna: resultados correctos en los sujetos estudiados, amenazada por todos los sesgos y mejorada por el buen diseño (antes «calidad metodológica», hoy «riesgo de sesgo»); validez externa: generalización **F1,F10**.
- 4 Sesgo de selección: la población estudiada difiere sistemáticamente de la población de interés; la participación voluntaria produce autoselección, un sesgo de selección **F1,F2,F3**.
- 5 Sesgo de detección: diferencias sistemáticas en cómo se determinan los desenlaces **F11**.
- 6 Efecto Hawthorne: los participantes cambian su conducta al saberse observados **F13**.
- 7 Confusor: factor de riesgo de la enfermedad y asociado a la exposición; se controla con aleatorización, restricción, emparejamiento, estratificación o modelos multivariantes **F1**.
- 8 Sesgo de adelanto diagnóstico: el cribado alarga la supervivencia aparente desde el diagnóstico sin cambiar el curso de la enfermedad **F15,F17**.
- 9 Fiabilidad = resultados constantes en distintas ocasiones; validez = medir lo que se pretende; consistencia interna = relación entre los ítems (alfa de Cronbach) **F18**.
- 10 Kappa: concordancia de variables categóricas corregida por el azar; en la escala de Cohen, 0,61-0,80 es sustancial y 0,21-0,40 «fair» (discreto) **F20,F21**.

1. Error aleatorio y error sistemático: precisión y exactitud

En un estudio epidemiológico hay dos grandes tipos de error. El **error aleatorio** se produce cuando el valor medido en la muestra se aparta del verdadero valor de la población **debido solo al azar**, y produce medidas de asociación inexactas ^{F1}. Sus tres fuentes principales son la **variación biológica individual**, el **error de muestreo** y el **error de medida** ^{F1}.

El error aleatorio **nunca puede eliminarse por completo**, porque solo se estudia una muestra de la población ^{F1}. El error de muestreo suele deberse a que una muestra pequeña no representa todas las variables de la población, y **la mejor forma de reducirlo es aumentar el tamaño del estudio** ^{F1}. El error de medida se reduce con protocolos estrictos y haciendo cada medición lo más precisa posible ^{F1}. La precisión de un estudio también mejora si los grupos que se comparan tienen un tamaño relativo adecuado ^{F1}.

El **error sistemático o sesgo** se produce cuando los resultados **difieren de forma sistemática de los valores verdaderos** ^{F1}. Un estudio con poco error sistemático tiene **alta exactitud**, y **la exactitud no se ve afectada por el tamaño de la muestra** ^{F1}. Se han descrito más de 30 tipos concretos de sesgo; los principales son el **sesgo de selección** y el **sesgo de medición (o de clasificación)** ^{F1}.

Aplicado a un instrumento, se dice que es **preciso** si sus resultados son **consistentes cuando se aplica más de una vez al mismo individuo bajo las mismas circunstancias**; su precisión depende de la variación propia del instrumento (calidad y calibrado del equipo) y de la variación del examinador ^{F19}. Epidat pone el ejemplo de dos básculas, **una de las cuales pesa sistemáticamente 1 kg más que la otra**: la correlación entre ambas puede ser casi 1 aunque la concordancia sea nula ^{F19}.

	Error aleatorio	Error sistemático (sesgo)
Origen	El azar ^{F1}	Se aparta del valor real de forma sistemática ^{F1}
Qué afecta	La precisión ^{F1}	La exactitud y la validez ^{F1}
Tamaño de la muestra	El error de muestreo disminuye al aumentar el tamaño del estudio ^{F1}	No se corrige aumentando la muestra ^{F1}
Se puede eliminar	Nunca del todo ^{F1}	Se reduce con buen diseño y atención al detalle ^{F1}

Diana

Aleatorio = dardos dispersos alrededor del centro (poco precisos). Sistemático = dardos agrupados pero lejos del centro (precisos pero inexactos). Un aparato que se aparta del valor real siempre en el mismo sentido y en la misma cantidad comete un error sistemático, no aleatorio.

Error sistemático

En el estudio que está usted realizando sobre el aumento de peso entre lactantes con lactancia materna exclusiva, lactancia mixta y lactancia exclusiva con fórmulas de leche preparada, la báscula que usted utiliza indica siempre un peso de 200 gramos superior al peso real. Esto causaría un error de tipo:

1. Sistemático.
2. De selección.
3. Aleatorio.
4. De falseamiento.

2. Validez del estudio: interna y externa

La **validez** expresa el grado en que una prueba es capaz de medir lo que pretende medir. Un **estudio es válido si sus resultados corresponden a la verdad**: no debe haber error sistemático y el error aleatorio debe ser lo más pequeño posible ^{F1}. Una alta fiabilidad (repetibilidad) de las mediciones **no garantiza la validez**, porque todas pueden estar lejos del valor real ^{F1}. Hay dos tipos de validez: **interna y externa** ^{F1}.

La **validez interna** es el grado en que los resultados de una observación son **correctos para el grupo concreto de personas estudiado** ^{F1}. Para que un estudio sirva de algo **tiene que ser internamente válido** ^{F1}. La validez interna puede verse amenazada por **todas las fuentes de error sistemático**, pero **puede mejorarse con un buen diseño y atención al detalle** ^{F1}. El Manual Cochrane explica que las herramientas actuales para valorar la **validez interna** de los resultados se centran en el **riesgo de sesgo**, mientras que las anteriores se centraban en la noción más amplia de **«calidad metodológica»** ^{F10}. Validez interna y calidad metodológica hablan, pues, del rigor con que se hizo el estudio, no de su aplicabilidad a otros pacientes ^{F1,F10}.

La **validez externa o generalizabilidad** es el grado en que los resultados de un estudio **se aplican a personas que no participaron en él** ^{F1}. La validez interna es **necesaria para la externa, pero no la garantiza**, y es más fácil de conseguir ^{F1}. Juzgar la validez externa exige un juicio sobre hasta qué punto se pueden extrapolar los resultados (por ejemplo, si el efecto de bajar el colesterol en varones sirve para mujeres), y se refuerza si estudios en otras poblaciones encuentran resultados similares ^{F1}.

	Validez interna	Validez externa
Pregunta que responde	¿Son correctos los resultados en los sujetos estudiados ? ^{F1}	¿Se pueden aplicar a otras personas ? ^{F1}
Depende de	Ausencia de error sistemático; buen diseño ^{F1}	Juicio sobre la extrapolación ; resultados similares en otras poblaciones ^{F1}
Relación	Necesaria para la externa y más fácil de lograr ^{F1}	Requiere la interna, que no basta ^{F1}



Lo que cae

Validez interna = resultados correctos en la muestra estudiada, amenazada por **todos** los sesgos y mejorada por el **buen diseño**; antes se valoraba como **«calidad metodológica»** y hoy como **riesgo de sesgo**. Validez externa = **aplicabilidad o generalización** ^{F1,F10}.

Validez interna: calidad metodológica

Cuando se dice que un estudio de investigación tiene validez interna, se refiere a que tiene:

1. Pertinencia clínica.
2. Aplicabilidad.
3. Calidad metodológica.
4. Pocos sesgos de confusión.

3. Sesgo de selección

El **sesgo de selección** se produce cuando los individuos o grupos de un estudio **difieren sistemáticamente de la población de interés**, lo que lleva a un error sistemático en una asociación o un resultado ^{F2}. La OMS lo describe como una **diferencia sistemática entre las características de las personas seleccionadas para un estudio y las de las que no lo son** ^{F1}. Por ejemplo, los participantes de un ensayo de vacuna antigripal pueden ser adultos jóvenes sanos, mientras que quienes recibirán la vacuna en la práctica son mayores con comorbilidad: la muestra **no representa** a la población a la que se quiere aplicar el resultado ^{F2}.

Si las personas que entran o permanecen en un estudio tienen características distintas de las que no son seleccionadas o de las que abandonan antes de terminar, se obtiene una **estimación sesgada de la asociación entre exposición y efecto**; todos los diseños epidemiológicos deben tener en cuenta el sesgo de selección ^{F1}.

EIR 2017 · 208

P9 · respuesta en el solucionario

Sesgo de selección: la población de estudio no representa a la diana

El sesgo de selección sucede cuando:

1. El error se comete en la recogida de datos.
2. El análisis es inapropiado.
3. La población de estudio no representa a la población diana.
4. Hay un error por la relación que mantienen variables con la exposición y el efecto dentro de la población base.

Una fuente evidente de sesgo de selección es que **los participantes se seleccionen a sí mismos** para un estudio, porque se encuentran mal o porque les preocupa especialmente una exposición ^{F1}. Por ejemplo, quienes responden a una invitación para un estudio sobre tabaco fuman de forma distinta que quienes no responden, que suelen ser más fumadores ^{F1}. Es el **sesgo de autoselección o del voluntario**: los participantes que **se ofrecen voluntariamente** tienen características intrínsecamente distintas de la población general de interés, y la muestra corre el riesgo de no ser representativa ^{F3}. Puede aparecer en todas las fases, desde el reclutamiento hasta el seguimiento, y las diferencias no se limitan a factores sociodemográficos, sino también a actitudes hacia el estudio ^{F3}.



Voluntariedad = selección

Cuando los sujetos **deciden por sí mismos si participan** (o la voluntariedad es criterio de inclusión), el sesgo es de **autoselección** (voluntario), que es un tipo de **sesgo de selección**, no de información ni de confusión ^{F1,F3}.

Sesgo de (auto)selección por participación voluntaria

Los sesgos que se producen cuando en un estudio los sujetos seleccionados toman decisiones propias en relación a su participación, se denominan:

1. Sesgo o paralogismo de Berkson.
2. Sesgo de detección de la enfermedad.
3. Sesgo de prevalencia / incidencia (paralogismo de Nyman).
4. Sesgo de autoselección.
5. Sesgo del trabajador sano.

Sesgo de (auto)selección por participación voluntaria

Si en un estudio epidemiológico, la población participante tiene como criterio de inclusión la voluntariedad, la validez de los resultados están sometidos a un sesgo de:

1. Información.
2. Confusión.
3. Selección.
4. Interpretación.
5. Publicación.

Otras variantes del sesgo de selección que aparecen como distractores **F1,F4,F5,F6,F7**:

Sesgo	Qué ocurre
Berkson (de admisión o ingreso)	La combinación de exposición y enfermedad hace más probable el ingreso hospitalario ; en un estudio de casos y controles con pacientes hospitalizados aparece una asociación espuria entre enfermedad y factor F4
Neyman (prevalencia-incidencia)	Por el momento en que se incluyen los casos: una mirada tardía pierde los episodios mortales o breves y los casos leves o silentes ; excluir a los fallecidos hace parecer la enfermedad menos grave F5
Trabajador sano	Los trabajadores tienen que estar lo bastante sanos para trabajar y los enfermos graves suelen quedar fuera del empleo; es un sesgo de selección de la epidemiología laboral F1
No respuesta	Diferencias sistemáticas entre quienes responden y quienes no F6
Desgaste (attrition)	Pérdidas desiguales de participantes entre los grupos de un ensayo F7

4. Sesgo de información o de medición

El **sesgo de información** es cualquier diferencia sistemática respecto a la verdad que surge en la **recogida, el recuerdo, el registro o el manejo de la información** de un estudio, incluido el tratamiento de los datos que faltan ^{F8}. Sus tipos principales son el sesgo de **clasificación errónea**, el del **observador**, el de **recuerdo** y el de notificación ^{F8}. La OMS lo llama **sesgo de medición**: las mediciones o clasificaciones de enfermedad o exposición son inexactas, es decir, **no miden correctamente lo que deben medir** ^{F1}.

La **clasificación errónea** es **no diferencial** cuando la probabilidad de clasificar mal es igual en todos los grupos y **diferencial** cuando difiere entre ellos ^{F12}. Según la OMS, si el sesgo de medición ocurre por igual en los grupos comparados, **casi siempre infraestima** la verdadera fuerza de la asociación ^{F1}; el Catalogue of Bias matiza que la clasificación errónea puede tener **efectos impredecibles**, aumentando o disminuyendo la asociación observada ^{F12}.

Sesgo	Qué ocurre	Cómo se previene
Recuerdo	Casos y controles recuerdan de forma distinta la exposición pasada; típico de los estudios de casos y controles retrospectivos ^{F1,F9}	Preguntar a todos del mismo modo ^{F9}
Observador	Conocer la exposición influye en las mediciones del investigador, del técnico o del participante ^{F1}	Medir a ciego o doble ciego ^{F1}
Detección	Diferencias sistemáticas entre grupos en cómo se determinan los desenlaces (cómo se recoge o verifica la información del resultado) ^{F11}	Evaluadores del desenlace cegados ^{F11}

Es, por tanto, el sesgo que se produce **en el momento de establecer si ha aparecido el desenlace**, y afecta tanto a ensayos como a estudios observacionales: en una cohorte de supervivientes de cáncer de mama, las mujeres recibían más o menos cribado según los medicamentos que tomaban, lo que podía sesgar las asociaciones observadas ^{F11}. El sesgo de detección puede sobreestimar o infraestimar el efecto: en ensayos aleatorizados, los evaluadores del desenlace no cegados exageraron las odds ratio de media un 36 % ^{F11}. Tampoco es lo mismo que el **sesgo de realización** (performance): diferencias sistemáticas en la **atención** que reciben los grupos, aparte de la intervención estudiada, por conocer la asignación ^{F22}. El de detección no debe confundirse con el sesgo de selección (quién entra en el estudio) ni con el de desgaste (quién se pierde durante el seguimiento) ^{F2,F7}.

Sesgo de detección

En estudios no aleatorizados, la evaluación de errores sistemáticos en el momento temporal en el que se establece la presencia del desenlace de interés supone la detección de:

1. Sesgo de selección.
2. Sesgo de detección.
3. Sesgo de confusión.
4. Sesgo de desgaste.

El **efecto Hawthorne** se produce cuando las personas **modifican algún aspecto de su conducta porque saben que están siendo observadas** ^{F13}. Puede afectar a hábitos como la dieta o la higiene y contaminar un estudio de intervención si uno de los grupos se observa con más frecuencia que el otro ^{F13}. El nombre viene del estudio original en la fábrica **Hawthorne Works**, donde la atención prestada a los trabajadores aumentó temporalmente la producción ^{F13}. En un estudio sobre higiene de manos del personal, el cumplimiento fue un 55 % mayor cuando sabían que los observaban ^{F13}. Se previene con **observación oculta** ^{F13}.

Efecto Hawthorne

En un estudio de investigación en el que la unidad de observación son las personas, ¿cómo se llama al sesgo en el que los participantes que están siendo observados, alteran su comportamiento?:

1. Sesgo del entrevistador.
2. Falacia de Neyman.
3. Falacia de Berkson.
4. Efecto Hawthorne.

5. Confusión

La **confusión** es una distorsión de la asociación entre una exposición y un efecto porque **un factor está asociado de forma independiente con la exposición y con el efecto** ^{F14}. Puede **sugerir una asociación donde no la hay o enmascarar una asociación real** ^{F14}. Para que una variable sea un factor de confusión debe cumplir dos condiciones: **ser por sí misma determinante de la enfermedad (factor de riesgo)** y **estar asociada a la exposición** estudiada ^{F1}.

El ejemplo de la OMS: la relación entre **beber café** y la cardiopatía coronaria puede deberse a que los bebedores de café **fuman más**, y el tabaco causa cardiopatía ^{F1}. La edad y la clase social son confusores frecuentes ^{F1}. Según la OMS, la confusión puede parecer un sesgo, pero en realidad no resulta de un error sistemático del diseño, sino de la distribución no aleatoria de los factores de riesgo en la población de origen ^{F1}. Es frecuente en los estudios observacionales, aunque también puede darse en los aleatorizados ^{F14}.

Fase	Métodos de control de la confusión ^{F1}
Diseño	Aleatorización (ideal en estudios experimentales), restricción (p. ej., solo no fumadores) y emparejamiento (cada caso con un control de igual edad y sexo; si es excesivo, sobreemparejamiento)
Análisis	Estratificación y modelos estadísticos multivariantes

No confundir los tres errores

Selección = **quién** entra o queda en el estudio. Información = **cómo** se miden exposición o efecto.

Confusión = una **tercera variable** relacionada con exposición y efecto ^{F1,F2,F8,F14}.

6. Sesgos de los programas de cribado

El **sesgo de adelanto diagnóstico (lead time)** es una distorsión que **sobreestima el tiempo aparente de supervivencia** con una enfermedad por **adelantar el momento de su diagnóstico** ^{F15}. En las personas cribadas, la supervivencia medida desde el diagnóstico parecerá más larga **aunque la detección precoz no tenga ningún efecto sobre el curso de la enfermedad** ^{F15}. Según el NCI, este sesgo alarga la supervivencia estimada de los cánceres detectados por cribado porque el cálculo **incluye el tiempo previo a lo que habría sido el diagnóstico clínico** sin cribado ^{F17}. Por eso, **una supervivencia más larga no prueba por sí sola el beneficio de la detección precoz** ^{F15}.

El **sesgo de duración (length time)** produce un beneficio aparente de supervivencia en los casos detectados por cribado porque éste **detecta preferentemente las formas de progresión lenta** y se le escapan las de progresión rápida ^{F16}. El **sobrediagnóstico** se produce cuando el cribado **detecta cánceres que nunca se habrían manifestado clínicamente** sin él; aumenta los casos y mejora la supervivencia sin modificar la mortalidad por la causa ^{F17}. El NCI cita la **PSA** como ejemplo de alta tasa de sobrediagnóstico ^{F17}.

Sesgo del cribado	Mecanismo
Adelanto diagnóstico (lead time)	Se cuenta la supervivencia desde un diagnóstico adelantado ; el reloj empieza antes, aunque la muerte llegue igual ^{F15,F17}
Duración (length time)	Se detectan más las formas lentas , de mejor pronóstico ^{F16}
Sobrediagnóstico	Se detectan cánceres que nunca habrían dado clínica ^{F17}

EIR2026-050: ¿adelanto diagnóstico o sobrediagnóstico?

La pregunta EIR2026-050 describe un cribado con PSA en el que la supervivencia a 5 años **desde el diagnóstico** mejora sin que baje la mortalidad, y pide el sesgo que lo explica **principalmente**. El sesgo de adelanto diagnóstico explica exactamente ese tiempo añadido, porque el cálculo incluye el tiempo previo al diagnóstico clínico ^{F17}, y no es la excepción sino **la regla** en cualquier detección precoz eficaz ^{F15}. Pero el NCI atribuye el mismo patrón también al **sobrediagnóstico**, frecuente con la PSA, que mejora la supervivencia sin modificar la mortalidad por la causa ^{F17}, y el sesgo de duración también alarga la supervivencia de los casos cribados ^{F16}. Varias opciones describen, pues, mecanismos reales; la respuesta oficial no se corrige.

Sesgo de adelanto diagnóstico (lead time)

En un programa de cribado de cáncer de próstata mediante Prostate-Specific Antigen (PSA), se observa que la supervivencia a 5 años desde el diagnóstico es del 95 % en pacientes detectados por cribado frente al 70 % en pacientes con diagnóstico clínico. Sin embargo, la mortalidad global por esta causa no se ha reducido. ¿Qué sesgo explica principalmente este fenómeno?:

1. Sesgo de duración.
2. Sesgo de adelanto diagnóstico.
3. Sesgo de sobrediagnóstico.
4. Sesgo de voluntariado.

7. Calidad de un instrumento de medida: fiabilidad y validez

Las dos características métricas esenciales de un instrumento (cuestionario, escala) son la **fiabilidad** y la **validez**; la **sensibilidad** al cambio y la **factibilidad** completan sus propiedades psicométricas ^{F18}. **La fiabilidad hace referencia a medir una variable de manera constante y la validez a que el instrumento mide lo que quiere medir** ^{F18}. No todo instrumento fiable es válido: puede medir de forma constante y aun así no medir el fenómeno que pretende ^{F18}.

Fiabilidad es la propiedad que designa la **constancia y precisión de los resultados que obtiene un instrumento al aplicarlo en distintas ocasiones**; un instrumento es fiable cuando los resultados son **comparables en situaciones similares** ^{F18}. Se expresa como un coeficiente de 0 (ausencia de correlación) a 1 (correlación perfecta), y, según algunos autores, el margen aceptable se sitúa **entre 0,7 y 0,9** ^{F18}. La OMS lo resume así: una prueba es **fiable si da resultados consistentes** ^{F1}.

EIR 2017 · 19

P11 · respuesta en el solucionario

Fiabilidad de un instrumento: resultados consistentes en repetidas aplicaciones

¿Cuándo se considera que un instrumento de medida es fiable?:

1. Cuando lo recomiendan un grupo de expertos acreditados.
2. Cuando mide la variable que pretende medir.
3. Cuando produce resultados consistentes cuando se aplica en diferentes ocasiones: estabilidad o reproductibilidad.
4. Cuando es útil y sensible al problema de salud que tiene que evaluar.

EIR 2019 · 53

P6 · respuesta en el solucionario

Fiabilidad de un instrumento: resultados consistentes en repetidas aplicaciones

Se dice que un cuestionario es fiable cuando:

1. Cuando mide lo que quiere medir.
2. Si contempla todos los aspectos relacionados con el concepto de estudio.
3. Se obtienen resultados consistentes cuando se aplica en diferentes ocasiones.
4. Tiene validez de criterio.

Forma de fiabilidad	Qué mide ^{F18}
Consistencia interna	La homogeneidad de los enunciados del instrumento, indicando la relación entre ellos ; es el método más usado. Técnicas: alfa de Cronbach (grado de correlación interna entre los ítems; el más usado), mitad y mitad, Kuder-Richardson (ítems dicotómicos) y correlación internunciados (cada ítem con la puntuación total)
Estabilidad (test-retest)	La constancia de las respuestas en repetidas ocasiones , en las mismas condiciones y con los mismos sujetos; se suele sugerir un intervalo de 2-4 semanas para variables estables
Equivalencia	Correlación entre dos versiones paralelas del mismo test aplicadas a la vez
Armonía interjueces	La concordancia entre dos o más observadores que miden lo mismo (o del mismo observador en dos ocasiones); Carvajal señala que habitualmente se calcula con Pearson o Spearman, aunque kappa, análisis de varianza o correlación intraclass podrían dar resultados más fiables (Epidat desaconseja Pearson para medir concordancia: ver sección 8)

Según Epidat, un conjunto de ítems (variable sintética) exhibe **consistencia interna cuando hay una alta asociación entre los ítems que la integran**, y la técnica más común para analizarla es el **alfa de Cronbach** ^{F19}. Valores más altos de alfa indican mayor consistencia interna; según George y Mallery, por debajo de 0,5 no es aceptable y por encima de 0,9 es excelente ^{F18}.

EIR 2025 · 35

P3 · respuesta en el solucionario

Consistencia interna: interrelación entre los ítems

En el análisis de las propiedades de medición de un instrumento, ¿a qué se refiere la consistencia interna?:

1. Al grado en que el contenido de un instrumento es un reflejo adecuado del constructo que se va a medir.
2. Al grado de interrelación entre los elementos del instrumento.
3. Al grado en que las puntuaciones de un instrumento son un reflejo adecuado del “estándar de oro”.
4. El grado en que las puntuaciones de un instrumento no cambian para mediciones repetidas por varios evaluadores.

La **validez** explora **en qué grado un instrumento mide lo que debería medir**, es decir, aquello para lo que ha sido diseñado ^{F18}. Se estima de varias formas, cada una de las cuales aporta evidencia a la validación global ^{F18}:

Tipo de validez	Qué evalúa ^{F18}
Aparente	El grado en que los ítems parecen medir lo que se proponen; la parte de «sentido común» de la validez de contenido

Tipo de validez	Qué evalúa ^{F18}
De contenido	Si el cuestionario abarca todas las dimensiones del fenómeno; es válido en su contenido si contempla todos los aspectos relacionados con el concepto . Métodos: Delphi, Fehring, metodología Q, estimación de magnitud
De criterio	Correlación con otra medida de referencia; si está aceptada por los investigadores, estándar o regla de oro . Concomitante (a la vez) y predictiva (medida posterior)
De constructo	Relación del instrumento con la teoría ; convergente-divergente, análisis factorial y validez discriminante

La **sensibilidad** de un instrumento es su capacidad de **detectar cambios** en los sujetos evaluados tras una intervención, y la **factibilidad** mide si es asequible en la práctica (tiempo, sencillez, preguntas sin contestar) ^{F18}.

Fiable frente a válido

Fiable = siempre da lo mismo (constancia, reproducibilidad). **Válido = da lo que tiene que dar** (mide la variable que pretende). Si una opción dice «resultados consistentes en distintas ocasiones», es fiabilidad; si dice «mide lo que pretende medir», es validez.

EIR 2017 · 20

P10 · respuesta en el solucionario

Validez de un instrumento: mide lo que pretende medir

¿Qué es validez de un instrumento de medida?:

1. Es el grado en que un instrumento es sensible a los cambios.
2. Es el grado en que un instrumento mide la variable que pretende medir.
3. Es el grado en que un instrumento produce resultados consistentes.
4. Es el grado en que un instrumento es estable en sus mediciones.

8. Concordancia: índice kappa y coeficiente de correlación intraclase

La variabilidad entre observadores es una fuente importante de error de medida; su estudio distingue el **sesgo entre observadores** (uno da consistentemente valores mayores que otro) y la **concordancia** (hasta qué punto coinciden) ^{F20}. Si los datos son **continuos** se usa el **coeficiente de correlación intraclase**; si son **categoricos**, el estadístico más empleado es el **índice kappa** ^{F20}. Epidat lo aplica a escalas **nominales u ordinales** y a dos o más observadores ^{F19}.

El porcentaje de acuerdo observado es la medida más intuitiva, pero parte de ese acuerdo puede deberse solo al azar ^{F19}. El kappa, elaborado por **Cohen en 1960** ^{F19}, corrige ese problema: es la **razón entre el exceso de concordancia observado más allá del atribuible al azar y el máximo exceso posible** ^{F20}. **Kappa = 1** es la máxima concordancia; **kappa = 0**, la concordancia esperada solo por azar; **kappa < 0**, menos acuerdo del esperado por azar ^{F20}. Para un mismo acuerdo observado, el kappa **depende de la prevalencia** del fenómeno (prevalencias muy altas o muy bajas dan kappas bajos) ^{F19}.

La interpretación del kappa depende de una escala convencional, y hay varias. McHugh recoge la que **sugirió Cohen**, que usa las etiquetas «moderado», «sustancial» y «casi perfecto» ^{F21}:

Kappa	Grado de acuerdo según Cohen (McHugh) ^{F21}	Escala de Fisterra ^{F20}
≤ 0	Sin acuerdo	—
0,01-0,20	Nulo a leve (none to slight)	Pobre (< 0,20)
0,21-0,40	«Fair» (discreto)	Débil
0,41-0,60	Moderado	Moderada
0,61-0,80	Sustancial (substantial)	Buena
0,81-1,00	Casi perfecto (almost perfect)	Muy buena

EIR2022-049: ¿es «insuficiente» un kappa de 0,21-0,40?

La pregunta EIR2022-049 usa la etiqueta «sustancial» para 0,61-0,80, que es la de la escala sugerida por **Cohen**; en ella, 0,21-0,40 se califica de **«fair» (discreto)**, por encima de «nulo a leve», y no de insuficiente ^{F21}. Otras lecturas son más exigentes: McHugh (2012) propone que **cualquier kappa por debajo de 0,60 indica un acuerdo inadecuado** y llama «mínimo» al tramo 0,21-0,39 ^{F21}; Fisterra califica 0,21-0,40 de **débil** ^{F20}, y Epidat atribuye a Landis y Koch (1977) una escala que considera **aceptable un kappa mayor o igual a 0,40**, además de recordar que no hay un valor exacto de buena concordancia ^{F19}. Con esas escalas, 0,21-0,40 sí podría llamarse insuficiente; con la de Cohen, cuyas etiquetas usa la pregunta, no. La respuesta oficial no se corrige.

Índice kappa: grados de acuerdo

Atendiendo al Índice kappa (κ), sería INCORRECTO:

1. El índice kappa (κ) se usa para evaluar la concordancia o reproducibilidad de instrumentos de medida cuyo resultado es categórico (2 o más categorías).
2. El índice kappa (κ) representa la proporción de acuerdos observados más allá del azar, respecto del máximo acuerdo posible más allá del azar.
3. Un resultado entre 0,21-0,40 supone un grado de acuerdo insuficiente.
4. Un resultado entre 0,61-0,80 supone un grado de acuerdo sustancial.

Para variables **continuas**, el **coeficiente de correlación intraclase (CCI)** mide la concordancia entre mediciones repetidas y **puede considerarse el equivalente del kappa**; toma valores de 0 a 1 y se interpreta como medida de **reproducibilidad o fiabilidad** ^{F19}. Según Fleiss, un CCI inferior a 0,4 indica poca reproducibilidad y uno igual o superior a 0,75, reproducibilidad excelente ^{F19}. El **coeficiente de correlación de Pearson no sirve para medir concordancia**: solo cuantifica la asociación lineal, y dos básculas que difieren sistemáticamente 1 kg pueden correlacionar casi perfectamente con concordancia nula ^{F19}. Para comparar dos métodos continuos se usa el **gráfico de Bland y Altman**, con límites de concordancia a $\pm 1,96$ desviaciones estándar de la diferencia media ^{F19}.

Tipo de variable	Medida de concordancia ^{F19}
Nominal u ordinal	Kappa de Cohen (dos o más observadores)
Continua	Coeficiente de correlación intraclase (dos o más observadores)
Continua, dos métodos	Bland-Altman (gráfico)
Consistencia interna de un cuestionario	Alfa de Cronbach

Todas las preguntas del tema (13)

De la más reciente a la más antigua. Las respuestas, en el solucionario.

EIR 2026 · 50

P1 · respuesta en el solucionario

En un programa de cribado de cáncer de próstata mediante Prostate-Specific Antigen (PSA), se observa que la supervivencia a 5 años desde el diagnóstico es del 95 % en pacientes detectados por cribado frente al 70 % en pacientes con diagnóstico clínico. Sin embargo, la mortalidad global por esta causa no se ha reducido. ¿Qué sesgo explica principalmente este fenómeno?:

1. Sesgo de duración.
2. Sesgo de adelanto diagnóstico.
3. Sesgo de sobrediagnóstico.
4. Sesgo de voluntariado.

EIR 2025 · 38

P2 · respuesta en el solucionario

En un estudio de investigación en el que la unidad de observación son las personas, ¿cómo se llama al sesgo en el que los participantes que están siendo observados, alteran su comportamiento?:

1. Sesgo del entrevistador.
2. Falacia de Neyman.
3. Falacia de Berkson.
4. Efecto Hawthorne.

EIR 2025 · 35

P3 · respuesta en el solucionario

En el análisis de las propiedades de medición de un instrumento, ¿a qué se refiere la consistencia interna?:

1. Al grado en que el contenido de un instrumento es un reflejo adecuado del constructo que se va a medir.
2. Al grado de interrelación entre los elementos del instrumento.
3. Al grado en que las puntuaciones de un instrumento son un reflejo adecuado del “estándar de oro”.
4. El grado en que las puntuaciones de un instrumento no cambian para mediciones repetidas por varios evaluadores.

EIR 2022 · 49

P4 · respuesta en el solucionario

Atendiendo al Índice kappa (κ), sería INCORRECTO:

1. El índice kappa (κ) se usa para evaluar la concordancia o reproducibilidad de instrumentos de medida cuyo resultado es categórico (2 o más categorías).
2. El índice kappa (κ) representa la proporción de acuerdos observados más allá del azar, respecto del máximo acuerdo posible más allá del azar.
3. Un resultado entre 0,21-0,40 supone un grado de acuerdo insuficiente.
4. Un resultado entre 0,61-0,80 supone un grado de acuerdo sustancial.

En estudios no aleatorizados, la evaluación de errores sistemáticos en el momento temporal en el que se establece la presencia del desenlace de interés supone la detección de:

1. Sesgo de selección.
2. Sesgo de detección.
3. Sesgo de confusión.
4. Sesgo de desgaste.

Se dice que un cuestionario es fiable cuando:

1. Cuandomide lo que quiere medir.
2. Si contempla todos los aspectos relacionados con el concepto de estudio.
3. Se obtienen resultados consistentes cuando se aplica en diferentes ocasiones.
4. Tiene validez de criterio.

Cuando se dice que un estudio de investigación tiene validez interna, se refiere a que tiene:

1. Pertinencia clínica.
2. Aplicabilidad.
3. Calidad metodológica.
4. Pocos sesgos de confusión.

En el estudio que está usted realizando sobre el aumento de peso entre lactantes con lactancia materna exclusiva, lactancia mixta y lactancia exclusiva con fórmulas de leche preparada, la báscula que usted utiliza indica siempre un peso de 200 gramos superior al peso real. Esto causaría un error de tipo:

1. Sistemático.
2. De selección.
3. Aleatorio.
4. De falseamiento.

El sesgo de selección sucede cuando:

1. El error se comete en la recogida de datos.
2. El análisis es inapropiado.
3. La población de estudio no representa a la población diana.
4. Hay un error por la relación que mantienen variables con la exposición y el efecto dentro de la población base.

¿Qué es validez de un instrumento de medida?:

1. Es el grado en que un instrumento es sensible a los cambios.
2. Es el grado en que un instrumento mide la variable que pretende medir.
3. Es el grado en que un instrumento produce resultados consistentes.
4. Es el grado en que un instrumento es estable en sus mediciones.

¿Cuándo se considera que un instrumento de medida es fiable?:

1. Cuando lo recomiendan un grupo de expertos acreditados.
2. Cuando mide la variable que pretende medir.
3. Cuando produce resultados consistentes cuando se aplica en diferentes ocasiones: estabilidad o reproductibilidad.
4. Cuando es útil y sensible al problema de salud que tiene que evaluar.

Si en un estudio epidemiológico, la población participante tiene como criterio de inclusión la voluntariedad, la validez de los resultados están sometidos a un sesgo de:

1. Información.
2. Confusión.
3. Selección.
4. Interpretación.
5. Publicación.

Los sesgos que se producen cuando en un estudio los sujetos seleccionados toman decisiones propias en relación a su participación, se denominan:

1. Sesgo o paralogismo de Berkson.
2. Sesgo de detección de la enfermedad.
3. Sesgo de prevalencia / incidencia (paralogismo de Nyman).
4. Sesgo de autoselección.
5. Sesgo del trabajador sano.

Preguntas que también podrían ser de este tema (1)

La clasificación automática dudó entre este tema y otro.

EIR 2011 · 13

P14 · respuesta en el solucionario

Cuando se realiza una valoración crítica del diseño de un estudio de investigación, el tamaño y representatividad de la muestra de estudio, la validez de los procedimientos de investigación y la validez de las técnicas de análisis de los datos, se están analizando las dimensiones:

1. Reales y técnicas.
2. Metodológicas.
3. Fundamentales.
4. Interpretativas.
5. Estilísticas.

Solucionario

N°	Pregunta	Respuesta oficial	Notas
P1	EIR 2026 · 50	2	
P2	EIR 2025 · 38	4	
P3	EIR 2025 · 35	2	
P4	EIR 2022 · 49	3	
P5	EIR 2021 · 38	2	
P6	EIR 2019 · 53	3	hoja del archivo oficial
P7	EIR 2019 · 48	3	hoja del archivo oficial
P8	EIR 2018 · 169	1	hoja del archivo oficial
P9	EIR 2017 · 208	3	
P10	EIR 2017 · 20	2	
P11	EIR 2017 · 19	3	
P12	EIR 2015 · 151	3	reconstruida por un tercero
P13	EIR 2005 · 94	4	reconstruida por un tercero
P14	EIR 2011 · 13	2	

Fuentes

- **F1** Organización Mundial de la Salud. Bonita R, Beaglehole R, Kjellström T. Basic epidemiology, 2.^a ed. Ginebra: OMS; 2006 (ISBN 978 92 4 154707 9). Capítulo 3 (Random error, Systematic error, Selection bias, Measurement bias, Confounding, Validity) y capítulo 6 (Screening test). · <https://iris.who.int/handle/10665/43541> · consultada 2026-09-27
- **F2** Catalogue of Bias Collaboration (Centre for Evidence-Based Medicine, Universidad de Oxford). Nunan D, Bankhead C, Aronson JK. Selection bias. Catalogue of Bias; 2017. · <https://catalogofbias.org/biases/selection-bias/> · consultada 2026-09-27
- **F3** Catalogue of Bias Collaboration. Brassey J, Mahtani KR, Spencer EA, Heneghan C. Volunteer bias. Catalogue of Bias; 2017. · <https://catalogofbias.org/biases/volunteer-bias/> · consultada 2026-09-27
- **F4** Catalogue of Bias Collaboration. Spencer EA, Aronson JK, Nunan D, Heneghan C. Berkson's bias (admission rate bias). Catalogue of Bias; 2018. · <https://catalogofbias.org/biases/admission-rate-bias/> · consultada 2026-09-27
- **F5** Catalogue of Bias Collaboration. Spencer EA, Heneghan C. Prevalence-incidence (Neyman) bias. Catalogue of Bias; 2017. · <https://catalogofbias.org/biases/prevalence-incidence-neyman-bias/> · consultada 2026-09-27
- **F6** Catalogue of Bias Collaboration. Turk A, Heneghan C, Nunan D. Non-response bias. Catalogue of Bias; 2019. · <https://catalogofbias.org/biases/non-response-bias/> · consultada 2026-09-27
- **F7** Catalogue of Bias Collaboration. Bankhead C, Aronson JK, Nunan D. Attrition bias. Catalogue of Bias; 2017. · <https://catalogofbias.org/biases/attrition-bias/> · consultada 2026-09-27
- **F8** Catalogue of Bias Collaboration. Bankhead CR, Spencer EA, Nunan D. Information bias. Catalogue of Bias; 2019. · <https://catalogofbias.org/biases/information-bias/> · consultada 2026-09-27
- **F9** Catalogue of Bias Collaboration. Spencer EA, Brassey J, Mahtani K. Recall bias. Catalogue of Bias; 2017. · <https://catalogofbias.org/biases/recall-bias/> · consultada 2026-09-27
- **F10** Boutron I, Page MJ, Higgins JPT, Altman DG, Lundh A, Hróbjartsson A. Chapter 7: Considering bias and conflicts of interest among the included studies (apartado 7.1.1). En: Higgins JPT, Thomas J, et al. (eds.). Cochrane Handbook for Systematic Reviews of Interventions, versión 6.5. Cochrane; 2024. · <https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-07> · consultada 2026-09-27
- **F11** Catalogue of Bias Collaboration. O'Sullivan J, Banerjee A, Pluddemann A. Detection bias. Catalogue of Bias; 2017. · <https://catalogofbias.org/biases/detection-bias/> · consultada 2026-09-27
- **F12** Catalogue of Bias Collaboration. Spencer EA, Mahtani KR, Brassey J, Heneghan C. Misclassification bias. Catalogue of Bias; 2018. · <https://catalogofbias.org/biases/misclassification-bias/> · consultada 2026-09-27
- **F13** Catalogue of Bias Collaboration. Spencer EA, Mahtani K. Hawthorne effect. Catalogue of Bias; 2017. · <https://catalogofbias.org/biases/hawthorne-effect/> · consultada 2026-09-27
- **F14** Catalogue of Bias Collaboration. Aronson JK, Bankhead C, Nunan D. Confounding. Catalogue of Bias; 2018. · <https://catalogofbias.org/biases/confounding/> · consultada 2026-09-27
- **F15** Catalogue of Bias Collaboration (Centre for Evidence-Based Medicine, Universidad de Oxford). Oke J, Fanshawe T, Nunan D. Lead time bias. Catalogue of Bias; 2021. · <https://catalogofbias.org/biases/lead-time-bias/> · consultada 2026-09-27
- **F16** Catalogue of Bias Collaboration. Oke JL, Pluddemann A. Length time bias. Catalogue of Bias; 2025. · <https://catalogofbias.org/biases/length-time-bias/> · consultada 2026-09-27

- **F17** National Cancer Institute (EE. UU.). Cancer Screening Overview (PDQ®) – Health Professional Version. PDQ Screening and Prevention Editorial Board. Última actualización del resumen: 16 de octubre de 2023. · <https://www.cancer.gov/about-cancer/screening/hp-screening-overview-pdq> · consultada 2026-09-27
- **F18** Carvajal A, Centeno C, Watson R, Martínez M, Sanz Rubiales Á. ¿Cómo validar un instrumento de medida de la salud? An Sist Sanit Navar. 2011;34(1):63-72 (tabla 1). · <https://recyt.fecyt.es/index.php/ASSN/article/view/10317> · consultada 2026-09-27
- **F19** Servicio Galego de Saúde (Xunta de Galicia) y OPS-OMS. Epidat 4: Ayuda de Concordancia y consistencia. Octubre de 2014. · https://www.sergas.gal/Saude-publica/Documents/1894/Ayuda_Epidat4_Concordancia_y_consistencia_Octubre2014.pdf · consultada 2026-09-27
- **F20** López de Ullibarri Galparsoro I, Pita Fernández S. Medidas de concordancia: el índice kappa. Fistera (Metodología de la investigación); Cad Aten Primaria 1999; 6: 169-171. Fecha de revisión en Fistera: 01/12/2010. · <https://www.fistera.com/formacion/metodologia-investigacion/medidas-concordancia-indice-kappa/> · consultada 2026-09-27
- **F21** McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82 (tabla 3). doi:10.11613/BM.2012.031. · <https://doi.org/10.11613/BM.2012.031> · consultada 2026-09-27
- **F22** Catalogue of Bias Collaboration. Banerjee A, Pluddemann A, O'Sullivan J, Nunan D. Performance bias. Catalogue of Bias; 2019. · <https://catalogofbias.org/biases/performance-bias/> · consultada 2026-09-27

Pendiente de verificar

- EIR2011-013 (dudosa entre M26-T02 y M26-T06, no intercalada): las «dimensiones» de la valoración crítica de una investigación (sustantivas, metodológicas, éticas, interpretativas, de presentación y estilísticas) proceden de Polit y Hungler, que no está en acceso abierto; M26-T06 la dejó igualmente en pendiente. Verificar en el manual o en una muestra de la editorial.
- Escala de Landis y Koch (Biometrics 1977;33:159-74): no se ha podido leer en acceso abierto; confirmar sus etiquetas y si coinciden con las que McHugh atribuye a Cohen, frente a la lectura de Epidat (aceptable a partir de 0,40).