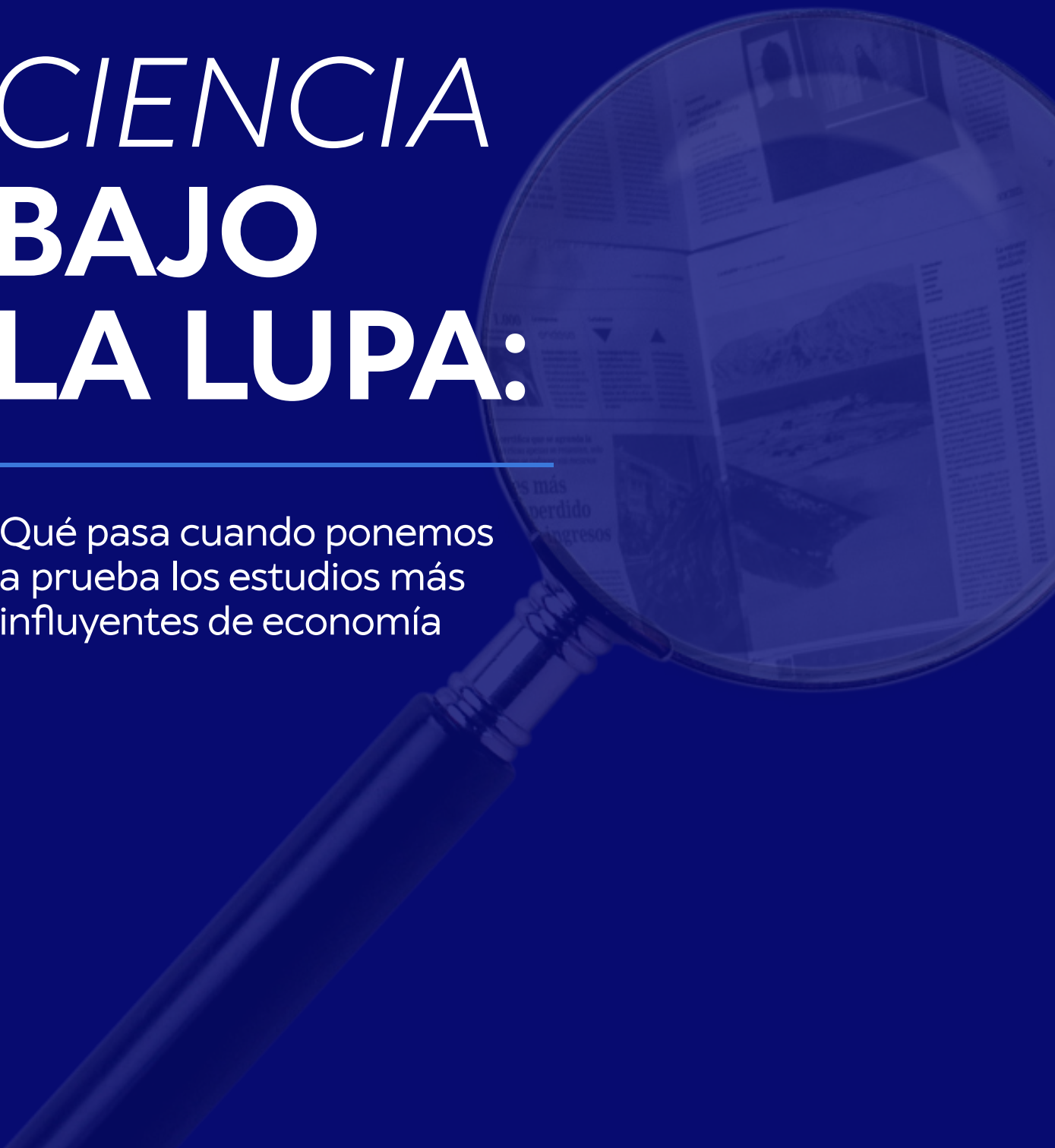


CIENCIA BAJO LA LUPA:

Qué pasa cuando ponemos
a prueba los estudios más
influyentes de economía





Cristina Griffa

Académica FEN U. de Chile.

¿Qué pasaría si sometiéramos los estudios más prestigiosos de economía y ciencia política a una auditoría independiente? Eso es exactamente lo que hizo un equipo de más de 300 investigadores de todo el mundo, con un resultado que sorprende por ambos lados: la mayoría de los estudios se sostiene, pero uno de cada cuatro tiene errores de código que nadie había detectado. El artículo, publicado en Nature en 2026, representa el mayor esfuerzo sistemático de reproducción científica en ciencias sociales hasta la fecha, y plantea preguntas incómodas sobre cuánto confiamos —y cuánto deberíamos confiar— en los resultados que dan forma a políticas públicas y debates académicos.

La académica FEN U. de Chile Cristina Griffa, participó directamente en este proyecto, replicando un estudio sobre incertidumbre de política monetaria publicado en The Economic Journal.

Una auditoría sin precedentes

En abril de 2026, la revista Nature publicó uno de los mayores ejercicios de verificación jamás emprendidos sobre la confiabilidad de la investigación en economía y ciencia política. Liderado por el economista Abel Brodeur, de la Universidad de Ottawa, y coordinado por el Institute for Replication (I4R), el proyecto convocó a cientos de investigadores de universidades de todo el mundo con un objetivo concreto: someter a verificación independiente 110 artículos publicados entre 2022 y 2023 en doce de las revistas más prestigiosas del área, entre las cuales se encuentran el American Economic Review, el Quarterly Journal of Economics, el Journal of Political Economy y el Economic Journal.

La pregunta de fondo era simple pero incómoda: cuando un estudio académico es revisado por pares, aceptado, publicado y citado, ¿qué tan seguros podemos estar de que sus resultados

se sostienen? Más concretamente, si otro equipo toma los datos y el código y trata de obtener los mismos números, ¿lo logra? Y si introduce cambios razonables (usar otra muestra, otro método estadístico, otra definición de la variable), ¿los resultados se mantienen?

Déjenme primero distinguir tres conceptos que utiliza este trabajo: reproducir un estudio es obtener sus mismos números con sus mismos datos y código; comprobar su robustez es ver si el resultado sobrevive a decisiones analíticas distintas pero razonables sobre esos mismos datos; y replicarlo es repetir el hallazgo con datos nuevos. El proyecto se concentró sobre todo en las dos primeras. Para proteger los resultados frente a sesgos, el proyecto estandarizó el procedimiento de antemano y pre-registró sus preguntas de análisis.

"Así como existe la auditoría financiera para verificar que los estados contables de una empresa sean correctos, puede existir una auditoría científica colectiva que fortalezca la credibilidad de lo que publicamos."

Lo que encontramos: el vaso medio lleno

Empecemos con las buenas noticias. Más del 85% de los resultados analizados pudieron reproducirse computacionalmente: cuando los equipos independientes corrieron el código original con los datos originales, obtuvieron los mismos números que el artículo publicado. Esto es mejor de lo que sugerían algunos de los estudios previos más críticos, y refleja en buena medida una transformación silenciosa pero importante: cuando los autores mapearon estas doce revistas año a año, los paquetes de replicación pasaron de acompañar a apenas el 59% de los artículos en 2014 a cerca del 90% en 2021–2023, gracias a políticas editoriales que lo hicieron obligatorio.

El segundo hallazgo también es alentador, pero mide algo distinto: la robustez. Una cosa es que el código original reproduzca los números publicados; otra, que el resultado sobreviva a decisiones analíticas alternativas pero razonables sobre los mismos datos. Cuando los investigadores probaron esos cambios (otra muestra, otro modelo, otra forma de medir una variable), el 72% de los resultados originalmente significativos siguieron siéndolo y en la misma dirección, y la mediana del efecto reproducido fue prácticamente idéntica a la publicada (99%). En otras palabras, los resultados de la mayoría de los estudios son robustos.

Y el vaso medio vacío

Pero también aparecieron cosas que invitan a mejorar. Dejando de lado problemas menores (un paquete que falta, una ruta de archivo que no funciona), en aproximadamente uno de cada cuatro estudios se detectaron errores de código de más relevancia, y algunos estudios acumulaban varios. Entre los más serios: una gran cantidad de observaciones duplicadas, una interacción incompleta en regresiones de diferencias en diferencias, una variable de tratamiento mal codificada y errores de especificación del modelo. Algo también importante es que estaban en artículos ya revisados por pares y publicados en las mejores revistas del mundo. Esto sugiere que los revisores leen el texto y evalúan los argumentos, pero no siempre ejecutan el código línea por línea.

Otro patrón llamativo es que los equipos con más experiencia tendieron a reportar menor robustez. La interpretación que ofrece el propio estudio es que los reproductores más experimentados detectan mejor los resultados poco robustos, como el "ojo entrenado" de un detective que capta pistas sutiles que a otros se les escapan. Esto sugiere que la robustez de toda la muestra probablemente habría sido menor si a cada artículo se le hubiera asignado un equipo igualmente experimentado.



"Incluso en el lugar más exigente del ecosistema científico, una cuarta parte de los estudios tiene errores que nadie había detectado, y casi un tercio de los resultados significativos no resiste cambios analíticos razonables."

Mi participación: reproduciendo la incertidumbre monetaria

Como parte de este esfuerzo colectivo, junto a tres colegas con quienes trabajaba en ese momento en el Reino Unido (Miquel Oliver-Vert, Benjamin Tatlow y Yaolang Zhong) me tocó reproducir un artículo publicado en The Economic Journal por Michael Bauer, Aeimit Lakdawala y Philippe Mueller, titulado "Market-Based Monetary Policy Uncertainty".

El estudio aborda algo que a cualquiera que siga la economía le suena familiar: la incertidumbre sobre qué hará la Fed con las tasas de interés. Los autores construyen una medida novedosa de esa incertidumbre directamente a partir de los precios de mercado, y muestran un patrón notable: cada vez que la Fed anuncia una decisión, la incertidumbre cae de manera abrupta. Tiene sentido: la incertidumbre se resuelve cuando llega la información. Lo difícil es medirlo con precisión y entender por qué importa.

Nuestro trabajo tuvo tres partes. La primera fue verificar la reproducibilidad computacional. Y ahí nos encontramos con un detalle revelador: el código original no corría. No por un error de los autores propiamente, sino por un problema de compatibilidad entre la versión del software que ellos usaron y las versiones disponibles un par de años después. Para alguien sin experiencia técnica, esto podría haber significado dar por perdida la reproducción. Identificamos el problema, hicimos

los ajustes necesarios y logramos reproducir los resultados exactamente.

La segunda parte fue someter el resultado central del artículo a un test estadístico más fino. Los autores muestran gráficamente cómo la incertidumbre cae el día del anuncio. Nosotros corrimos regresiones con ventanas de cinco y diez días alrededor de cada reunión del Comité de Política Monetaria de la Fed. Por un lado, confirmamos la caída en el día del anuncio. Pero encontramos algo adicional, la incertidumbre también tiende a disminuir antes de las reuniones programadas, no solo después. Esto es interesante porque sugiere dos hipótesis: o bien los participantes del mercado están buscando activamente información antes de la reunión, o bien la propia Fed está señalando (implícitamente) la dirección de su decisión antes de comunicar formalmente.

La tercera parte fue una extensión. El artículo original analiza el efecto de la incertidumbre monetaria sobre varios activos financieros. Decidimos ampliar este análisis a otros: el oro, algunos índices bursátiles europeos, el Bitcoin y el franco suizo. Los resultados fueron mixtos. Para el oro y el DAX30 encontramos efectos estadísticamente significativos: la incertidumbre monetaria mueve sus precios de forma sistemática. Para el Bitcoin y el franco suizo los efectos apuntaron en la dirección que cabría esperar, pero sin alcanzar significancia estadística, probablemente por las particularidades de cada mercado.

¿Por qué importa todo esto?

Podría parecer que verificar lo ya verificado es un ejercicio académico, casi redundante. ¿Para qué duplicar el trabajo de revisar lo que las mejores revistas del mundo ya revisaron?

La respuesta está en lo que encontramos. Incluso en el lugar más exigente del ecosistema científico, una cuarta parte de los estudios tiene errores que nadie había detectado, y casi un tercio de los resultados significativos no resiste cambios analíticos razonables. La investigación económica influye en decisiones que afectan la vida de millones de personas: cómo se diseñan los subsidios, los planes sociales, cómo se responde a una crisis inflacionaria, qué políticas se priorizan. La confiabilidad de los resultados que sustentan esas decisiones no es un detalle técnico, es una cuestión pública.

El proyecto del I4R propone algo más profundo que un diagnóstico: una forma institucional de hacer la verificación científica una práctica sistemática y normal. Así como existe la auditoría financiera para verificar que los estados contables de una empresa sean correctos, puede existir (y de hecho está empezando a existir) una auditoría científica colectiva que fortalezca la credibilidad de lo que publicamos.

Para mí, personalmente, la experiencia fue tan formativa como cualquier curso de posgrado. Replicar el trabajo de otros obliga a entender sus decisiones desde adentro, a interrogarlas, y a veces a descubrir cosas que ni los propios autores habían notado. En el fondo, eso es lo que la ciencia debería estar haciendo todo el tiempo.

